РОССИЙСКАЯ АКАДЕМИЯ НАУК
ОТДЕЛЕНИЕ МАТЕМАТИЧЕСКИХ НАУК РАН
ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР ИМ. А. А. ДОРОДНИЦЫНА РАН
при поддержке
РОССИЙСКОГО ФОНДА ФУНДАМЕНТАЛЬНЫХ ИССЛЕДОВАНИЙ
КОМПАНИИ FORECSYS

Математические методы распознавания образов

ММРО-15

Доклады 15-й Всероссийской конференции

МАКС ПРЕСС
Москва, 2011

УДК 004.85+004.89+004.93+519.2+519.25+519.7+519.85
ББК 22.1:32.973.26-018.2
М34

Математические методы распознавания образов: 15-я Всероссийская конференция,
М34 г. Петрозаводск, 11–17 сентября 2011 г.: Сборник докладов. — М.: МАКС Пресс, 2011. — 618 с.
ISBN 978-5-317-03787-1

В сборнике представлены доклады 15-й Всероссийской конференции «Математические методы распознавания образов», проводимой Вычислительным центром им. А. А. Дородницына Российской академии наук при финансовой и организационной поддержке РФФИ и компании «Форексис».

Конференция регулярно проводится один раз в два года, начиная с 1983 г., и является самым представительным российским научным форумом в области распознавания образов и анализа изображений, интеллектуального анализа данных, машинного обучения, обработки сигналов, математических методов прогнозирования.

УДК 004.85+004.89+004.93+519.2+519.25+519.7+519.85
ББК 22.1:32.973.26-018.2

ISBN 978-5-317-03787-1

Оргкомитет

Председатель: Журавлев Юрий Иванович, *академик РАН*
Ученый секретарь: Чехович Юрий Викторович, *к.ф.-м.н.*

Матросов Виктор Леонидович, *академик РАН*
Донской Владимир Иосифович, *д.ф.-м.н.*
Дюкова Елена Всеволодовна, *д.ф.-м.н.*
Дедус Флоренц Федорович, *д.т.н.*
Мазалов Владимир Викторович, *д.ф.-м.н.*
Немирко Анатолий Павлович, *д.ф.-м.н.*
Устинин Михаил Николаевич, *д.ф.-м.н.*
Ивахненко Андрей Александрович, *к.ф.-м.н.*
Инякин Андрей Сергеевич, *к.ф.-м.н.*
Песков Николай Владимирович, *к.ф.-м.н.*

Программный комитет

Председатель: Рудаков Константин Владимирович, *чл.-корр. РАН*
Ученый секретарь: Воронцов Константин Вячеславович, *д.ф.-м.н.*

Сойфер Виктор Александрович, *чл.-корр. РАН*
Местецкий Леонид Моисеевич, *д.т.н.*
Моттль Вадим Вячеславович, *д.ф.-м.н.*
Пытьев Юрий Петрович, *д.ф.-м.н.*
Рязанов Владимир Васильевич, *д.ф.-м.н.*
Рейер Иван Александрович, *к.т.н.*

Технический комитет

Председатель: Громов Андрей Николаевич

Помазкова Евгения Владимировна
Быкова Елена Анатольевна
Каневский Даниил Юрьевич
Лисица Андрей Валерьевич
Толстихин Илья Олегович
Рябенко Евгений Алексеевич
Фролова Надежда Сергеевна

Краткое оглавление

Характеристики сжатия недоопределенных данных*

Шоломов Л. А.
sholomov@isa.ru
Москва, Институт системного анализа РАН

С недоопределенными данными имеют дело во многих разделах информатики, в числе которых распознавание образов. Статья носит обзорный характер. В ней рассмотрен вопрос, в каком виде результаты классической теории информации о сжатии последовательностей символов переносятся на недоопределенные данные и какие новые эффекты здесь возникают.

Введение

Пусть имеется последовательность цифр, написанная нечетко. В ней может, например, встретиться символ, похожий на 3 и на 5, либо символ, похожий на 1, 4 и 7. Такие символы считаются недоопределенными и обозначаются $a_{3,5}$ и $a_{1,4,7}$, а сами цифры $0, 1, \ldots, 9$ рассматриваются как основные символы. В общем случае недоопределенный символ ассоциируется с множеством основных символов, одним из которых он может быть замещен (доопределен). Результат решения задачи распознавания нечетко написанной последовательности дает некоторую последовательность основных символов, доопределяющую исходную.

Помимо задач распознавания, с недоопределенными данными имеют дело в задачах синтеза управляющих систем, принятия решений, управления, генетики. Поэтому целесообразно изучить недоопределенные данные в качестве самостоятельного объекта подобно тому, как это делается в теории информации для полностью определенных данных.

Данная работа носит обзорный характер. Более подробное изложение содержащихся в ней (и других) фактов имеется в [1, 2].

Энтропия недоопределенных данных

Пусть $M = \{0, 1, \ldots, m-1\}$ — некоторое множество и каждому непустому подмножеству $T \subseteq M$ сопоставлен символ a_T. Алфавит всех символов a_T обозначается через A, а его подалфавит $\{a_0, a_1, \ldots, a_{m-1}\}$, символы которого соответствуют элементам множества M, — через A_0. Символы из A_0 называются *основными*, из A — *недоопределенными*. *Доопределением* символа $a_T \in A$ считается всякий основной символ a_i, $i \in T$, а доопределением последовательности в алфавите A — любая последовательность в алфавите A_0, полученная из исходной заменой всех ее символов некоторыми доопределениями. Символ a_M, доопределимый любым основным символом, называется *неопределенным* и обозначается $*$.

Работа выполнена при финансовой поддержке ОНИТ РАН по программе фундаментальных исследований, проект 1.1 «Теория и методы эффективного использования недоопределенных данных».

Пусть имеется источник S, порождающий символы $a_T \in A$ независимо с вероятностями $p_T \geqslant 0$, $\sum_T p_T = 1$. *Энтропией источника* S назовем величину

$$\mathcal{H}(S) = \min_Q \Big\{ - \sum_{T \subseteq M} p_T \log \sum_{i \in T} q_i \Big\} \qquad (1)$$

(логарифмы двоичные), где минимум берется по наборам $Q = (q_i)_{i \in M}$, $q_i \geqslant 0$, $\sum_i q_i = 1$. Вместо $\mathcal{H}(S)$ будем также использовать обозначение $\mathcal{H}(P)$, где $P = (p_T)_{T \subseteq M}$.

Для полностью определенного источника S, порождающего символы a_i с вероятностями p_i, $i = 0, \ldots, m-1$, величина $\mathcal{H}(S)$ совпадает с энтропией Шеннона $H(p_0, \ldots, p_{m-1}) = -\sum_i p_i \log p_i$, а для источника, алфавит которого состоит лишь из основных символов a_i и неопределенного символа $*$, энтропия задается выражением

$$\begin{aligned} \mathcal{H}(S) &= (1 - p_*) H\left(\frac{p_0}{1 - p_*}, \ldots, \frac{p_{m-1}}{1 - p_*} \right) = \\ &= (1 - p_*) \log (1 - p_*) - \textstyle\sum_i p_i \log p_i. \end{aligned} \qquad (2)$$

В общем случае энтропия $\mathcal{H}(S)$ не допускает явного представления, но имеется простой и быстро сходящийся алгоритм ее вычисления.

Свойства энтропии

На $\mathcal{H}(S)$ переносятся (в модифицированном виде) основные свойства энтропии Шеннона [3]. Приведем некоторые из них.

Утверждение 1. $\mathcal{H}(S) \geqslant 0$, *причем*

$$\mathcal{H}(S) = 0 \iff \bigcap_{T:\, p_T > 0} T \neq \varnothing.$$

Это означает, что $\mathcal{H}(S) = 0$, лишь если S порождает последовательности, доопределимые одинаковыми символами. Для полностью определенного источника равенство $H(S) = 0$ имеет место, лишь если S выдает последовательность одинаковых символов.

Утверждение 2. $\mathcal{H}(S) \leqslant \log m - \sum_{1 \leqslant t \leqslant m} p(t) \log t$, *где*

$$p(t) = \sum_{T:\, |T| = t} p_T, \quad 1 \leqslant t \leqslant m,$$

— распределение числа t доопределений символов источника S.

Если S полностью определен, то $p(t) = 0$ для $t \geqslant 2$, и это неравенство превращается в известное соотношение $H(S) \leqslant \log m$.

Утверждение 3. *$\mathcal{H}(SS') \leqslant \mathcal{H}(S) + \mathcal{H}(S')$, а если S и S' статистически независимы, имеет место равенство.*

В отличие от обычной энтропии H, независимость здесь не является необходимой для равенства.

Следующее свойство энтропии не имеет аналога в классической теории информации, поскольку связано с наличием неопределенных символов. Пусть $P = (p_T)_{T\subseteq M}$ и набор $P' = (p'_T)_{T\subset M}$ получен из P отбрасыванием компоненты $p_M = p_*$ и пересчетом остальных компонент:

$$p'_T = \frac{p_T}{1-p_*}, \quad T \subset M.$$

Утверждение 4. *Имеет место равенство*

$$\mathcal{H}(P) = (1-p_*)\mathcal{H}(P').$$

Смысл этого соотношения будет прояснен позже. Его частным случаем является равенство (2).

Сжатие недоопределенных данных

Задача кодирования недоопределенных источников отличается от обычной задачи кодирования источников [3] тем, что по коду последовательности следует восстановить какое-либо ее доопределение, а не саму последовательность. Качество кодирования, как обычно, характеризуется *средней длиной кода* $\bar{l}$ на символ последовательности. Теорема кодирования полностью определенных источников обобщается на недоопределенный случай.

Утверждение 5. *При любом способе блокового кодирования недоопределенного источника S выполнено*

$$\bar{l} \geqslant \mathcal{H}(S)$$

и существует блоковое кодирование, для которого

$$\bar{l} \leqslant \mathcal{H}(S) + O\Big(\frac{\log n}{n}\Big),$$

где n — длина блока.

Рассмотрим теперь задачу сжатия в детерминированной постановке. Для набора натуральных чисел $\boldsymbol{l} = (l_T)_{T\subseteq M}$, $\sum_T l_T = n$, обозначим через $\mathcal{K}_n(\boldsymbol{l})$ множество всех последовательностей длины n в алфавите A, в которых символ a_T, $T \subseteq M$, встречается l_T раз. Обозначим через $N_n(\boldsymbol{l})$ минимальную мощность множества последовательностей в алфавите A_0, содержащего доопределение каждой последовательности из $\mathcal{K}_n(\boldsymbol{l})$. Величину $\log N_n(\boldsymbol{l})$ назовем *комбинаторной энтропией* класса $\mathcal{K}_n(\boldsymbol{l})$. С классом $\mathcal{K}_n(\boldsymbol{l})$ свяжем функционал

$$h_n(\boldsymbol{l}) = n\mathcal{H}(\boldsymbol{l}/n).$$

Утверждение 6. *Справедливы оценки комбинаторной энтропии*

$$h_n(\boldsymbol{l}) - c\log n \leqslant \log N_n(\boldsymbol{l}) \leqslant h_n(\boldsymbol{l}) + c\log n,$$

где $c = c(m)$ — некоторая константа.

Это означает, что $h_n(\boldsymbol{l})$ с точностью до $O(\log n)$ совпадает с минимальной длиной кода, достаточной для представления последовательностей класса $\mathcal{K}_n(\boldsymbol{l})$, и характеризует сжимаемость последовательностей этого класса.

Кодирование с заданным критерием верности

Рассмотрим теперь случай, когда при доопределении последовательностей в алфавите A допустимы ошибки. При этом символ a_T не обязательно воспроизводится некоторым символом a_i, $i \in T$, но имеется ограничение на число ошибок или другой параметр, связанный с ошибками.

Начнем с вероятностной модели. Пусть $r_{Ti} = p(a_T, a_i)$ — совместная вероятность того, что источник S порождает символ a_T и он заменяется на a_i, $\sum_i r_{Ti} = p_T$. Пусть задано множество R допустимых совместных распределений $r = (r_{Ti})_{T\subseteq M}^{i\in M}$. Введем функцию

$$\mathcal{H}_R(P) = \min_{r=(r_{Ti})\in R} \sum_{T,i} r_{Ti} \log \frac{r_{Ti}}{p_T \sum\limits_{T'} r_{T'i}}. \qquad (3)$$

Утверждение 7. *В случае, когда ошибки доопределения не допустимы, т. е множество R образовано всеми распределениями, для которых $r_{Ti} > 0$ лишь если $i \in T$, имеет место равенство*

$$\mathcal{H}_R(P) = \mathcal{H}(P).$$

Рассмотрим класс последовательностей $\mathcal{K}_n(\boldsymbol{l})$. Пусть последовательность $\boldsymbol{a}$ этого класса воспроизводится последовательностью $\dot{\boldsymbol{a}}$ в алфавите A_0. Обозначим через $w_{Ti}(\boldsymbol{a}, \dot{\boldsymbol{a}})$ число позиций, в которых символ a_T воспроизводится как a_i и введем набор

$$w(\boldsymbol{a}, \dot{\boldsymbol{a}}) = \big(w_{Ti}(\boldsymbol{a}, \dot{\boldsymbol{a}})\big)_{T\subset M}^{i\in M}.$$

Условие допустимости задается множеством W наборов w: считается, что воспроизведение $\dot{\boldsymbol{a}}$ вместо $\boldsymbol{a}$ допустимо, если $w(\boldsymbol{a}, \dot{\boldsymbol{a}}) \in W$.

Обозначим через $N_{n,W}(\boldsymbol{l})$ минимальную мощность множества последовательностей длины n в алфавите A_0, в котором для каждой последовательности из $\mathcal{K}_n(\boldsymbol{l})$ имеется W-допустимая последовательность. Величину $\log N_{n,W}(\boldsymbol{l})$ будем называть *комбинаторной W-энтропией* класса $\mathcal{K}_n(\boldsymbol{l})$.

С классом $\mathcal{K}_n(\boldsymbol{l})$ и условием допустимости W свяжем функционал

$$h_{n,W}(\boldsymbol{l}) = n\mathcal{H}_{W/n}(\boldsymbol{l}/n),$$

где используется функция из (3) для множества $R = W/n$, образованного всеми наборами $r = w/n$, $w \in W$.

Утверждение 8. *Комбинаторная W-энтропия класса $\mathcal{K}_n(\boldsymbol{l})$ удовлетворяет оценкам*

$$h_{n,W}(\boldsymbol{l}) - c\log n \leqslant \log N_{n.W}(\boldsymbol{l}) \leqslant h_{n,W}(\boldsymbol{l}) + c\log n,$$

где $c = c(m)$ — константа.

Таким образом, величина $h_{n,W}(\boldsymbol{l})$ с точностью до $O(\log n)$ совпадает с минимальной длиной кода, достаточной для представления последовательностей класса $\mathcal{K}_n(\boldsymbol{l})$ при точности воспроизведения W, и ее будем использовать в качестве характеристики W-сжимаемости последовательностей этого класса.

Устойчивость характеристик сжатия к изменению числа неопределенных символов

Обозначим через $\mathcal{K}_n(l)$, $l \leqslant n$, класс всех последовательностей длины n с l булевыми символами и $n - l$ неопределенными символами $*$, а через $\log N_n(l)$ — его комбинаторную энтропию. Э. И. Нечипорук [4] доказал, что

$$\log N_n(l) = l + O(\log n).$$

Этот факт допускают следующую интерпретацию, которую будем называть *эффектом Нечипорука*. Недоопределенные последовательности и последовательности меньшей длины, полученные из них удалением неопределенных символов, могут быть представлены кодами одинаковой с точностью до $O(\log n)$ длины.

Эффект Нечипорука обобщается для недоопределенных последовательностей общего вида.

Утверждение 9. *Если класс $\mathcal{K}_{n'}(\boldsymbol{l}')$, $n' = n - l_*$, образован из класса $\mathcal{K}_n(\boldsymbol{l})$ удалением в его последовательностях символов $*$, то*

$$\log N_n(\boldsymbol{l}) = \log N_{n'}(\boldsymbol{l}') + O(\log n).$$

Это соотношение вытекает из утверждений 6 и 4, поскольку в применении к классам $\mathcal{K}_n(\boldsymbol{l})$ и $\mathcal{K}_{n'}(\boldsymbol{l}')$ утверждение 4 дает равенство

$$h_n(\boldsymbol{l}) = n\mathcal{H}(\boldsymbol{l}/n) = n'\mathcal{H}(\boldsymbol{l}'/n') = h_{n'}(\boldsymbol{l}').$$

Оно проясняет комбинаторный смысл утверждения 4.

Эффект Нечипорука распространяется и на кодирование с заданным критерием верности.

Неопределенный символ $*$ при воспроизведении может быть заменен любым символом. Это означает, что если воспроизведение последовательности $\dot{\boldsymbol{a}}$ вместо $\boldsymbol{a}$ допустимо и последовательность $\dot{\boldsymbol{a}}_1$ отличается от $\dot{\boldsymbol{a}}$ лишь в разрядах, где $\boldsymbol{a}$ содержит символы $*$, то воспроизведение $\dot{\boldsymbol{a}}_1$ вместо $\boldsymbol{a}$ также допустимо.

Пусть условие допустимости для класса $\mathcal{K}_n(\boldsymbol{l})$ задается множеством $W = \{w\}$. *Укороченным* для набора $w = (w_{Ti})_{T\subseteq M}^{i\in M}$ назовем набор w', образованный из w удалением всех компонент w_{*i}. Очевидно, что допустимость последовательностей $\dot{\boldsymbol{a}}$ и наборов w полностью определяется укороченными наборами w'. Обозначим через W' множество всех укороченных наборов w' для $w \in W$.

Применительно к кодированию с заданным критерием верности эффект Нечипорука приобретает следующий вид.

Утверждение 10. *Если класс $\mathcal{K}_{n'}(\boldsymbol{l}')$, $n' = n - l_*$, образован из $\mathcal{K}_n(\boldsymbol{l})$ удалением в его последовательностях символов $*$, то*

$$\log N_{n,W}(\boldsymbol{l}) = \log N_{n',W'}(\boldsymbol{l}') + O(\log n).$$

Мы видели, что утверждение 4 можно рассматривать как теоретико-информационный аналог комбинаторного утверждения 9. Утверждение 10 также имеет теоретико-информационный аналог.

Пусть источник S порождает символы a_T алфавита A с вероятностями p_T, образующими набор P, и пусть допустимая точность воспроизведения S в алфавите A_0 задается множеством R совместных распределений $r = (r_{Ti})$. Рассмотрим источник S', порождающий символы из $A\backslash\{*\}$ с набором вероятностей P' из утверждения 4. Каждому совместному распределению $(r_{Ti})_{T\subseteq M}^{i\in M} \in R$ сопоставим распределение $(r'_{Ti})_{T\subseteq M}^{i\in M}$, образованное из (r_{Ti}) удалением всех компонент r_{*i} и нормировкой остальных компонент $r'_{Ti} = \dfrac{r_{Ti}}{1 - p_*}$. Множество полученных совместных распределений (r'_{Ti}) обозначим через R'.

Утверждение 11. *Имеет место равенство*

$$\mathcal{H}_R(P) = (1 - p_*)\mathcal{H}_{R'}(P').$$

Оно приводит к соотношению

$$h_{n,W}(\boldsymbol{l}) = h_{n',W'}(\boldsymbol{l}'),$$

из которого в силу утверждения 8 следует утверждение 10.

Всюду в данном пункте речь шла об устранении всех неопределенных символов. Аналогичные факты справедливы и при устранении части неопределенных символов, а также при добавлении неопределенных символов. Таким образом, характеристики сжатия оказываются устойчивыми к изменению числа неопределенных символов. Это справедливо и в случае, когда искажения не допускаются, и в случае кодирования с заданным критерием верности.

Структура лучших доопределений

Один из подходов к решению задач распознавания и классификации состоит в том, что в качестве решающего правила берется правило, согласованное с заданной (обучающей) информацией и имеющее наименьшую сложность (в условленном смысле). Поскольку сложность тесно связана с энтропией, некоторым основанием для такого подхода служит результат Вапника и Червоненкиса [5] о соотношении размера обучающей выборки, энтропии класса решающих правил и точности результата. В данном пункте рассматриваются наилучшие по сложности доопределения. Вначале результаты излагаются для недоопределенных последовательностей, затем они распространяются на недоопределенные булевы функции.

Для недоопределенной последовательности $\boldsymbol{a}$ длины n и ее доопределения $\dot{\boldsymbol{a}}$ обозначим через w_{Ti} число позиций, в которых символ a_T доопределен символом a_i. Структуру доопределения $\dot{\boldsymbol{a}}$ последовательности $\boldsymbol{a}$ будем характеризовать набором $r(\boldsymbol{a},\dot{\boldsymbol{a}})$ частот

$$r_{Ti} = \frac{w_{Ti}}{n}, \quad T \subseteq M,\ i \in M.$$

Пусть $K(\dot{\boldsymbol{a}})$ означает колмогоровскую сложность [6] доопределения $\dot{\boldsymbol{a}}$. Под колмогоровской сложностью $K(\boldsymbol{a})$ недоопределенной последовательности $\boldsymbol{a}$ будем понимать $\min K(\dot{\boldsymbol{a}})$ по всем ее доопределениям. Доопределение, на котором достигается $K(\boldsymbol{a})$, назовем *лучшим*.

Рассмотрим класс последовательностей $\mathcal{K}_n(\boldsymbol{l})$. Будем считать, что n растет и

$$\frac{l_T}{n} \to p_T, \quad T \subseteq M. \tag{4}$$

Введем набор $P = (p_T)_{T\subseteq M}$ и рассмотрим функцию $\mathcal{H}(P)$ (1) для этого набора. При достаточно слабых условиях минимум в (1) достигается на единственном наборе Q. Будем считать, что имеет место этот случай, который будем называть *невырожденным*. Единственную точку минимума будем обозначать $\hat{Q} = (\hat{q}_i)_{i\in M}$. Положим

$$\hat{r}_{Ti} = \frac{p_T \hat{q}_i}{\sum_{j\in T} \hat{q}_j} \tag{5}$$

и образуем набор $\hat{r} = (\hat{r}_{Ti})_{T\subseteq M}^{i\in M}$. Расстояние между наборами $r' = (r'_{Ti})$ и $r'' = (r''_{Ti})$ будем измерять величиной

$$d(r', r'') = \max_{T.i} |r'_{Ti} - r''_{T,i}|.$$

Следующий результат показывает, что лучшие доопределения почти всех последовательностей класса $\mathcal{K}_n(\boldsymbol{l})$ имеют почти одинаковую структуру.

Утверждение 12. *Если выполнены условия (4) и имеет место невырожденный случай, то для любого $\varepsilon > 0$ доля последовательностей $\boldsymbol{a} \in \mathcal{K}_n(\boldsymbol{l})$, для которых все лучшие доопределения $\dot{\boldsymbol{a}}$ удовлетворяют условию*

$$d\big(r(\boldsymbol{a},\dot{\boldsymbol{a}}), \hat{r}\big) < \varepsilon,$$

стремится к 1 при $n \to \infty$.

Утверждение 12 остается справедливым, если вместо лучших рассматривать асимптотически наилучшие доопределения.

Полученные результаты допускают модификацию применительно к реализации систем $F(\tilde{x}) = \big(f_1(\tilde{x}), \ldots, f_k(\tilde{x})\big)$, $\tilde{x} = (x_1, \ldots, x_n)$, частичных булевых функций (считаем k фиксированным, n — растущим). Систему F будем характеризовать параметрами $l_{\tilde{\alpha}}$, $\tilde{\alpha} \in \{0,1,*\}^k$, где $l_{\tilde{\alpha}}$ — число наборов $\tilde{x}$, на которых $F(\tilde{x}) = \tilde{\alpha}$. Введем класс $\mathcal{F}_n(\boldsymbol{l})$ систем F с набором параметров $\boldsymbol{l} = (l_{\tilde{\alpha}},\ \tilde{\alpha} \in \{0,1,*\}^k)$. Система F может быть описана в терминах недоопределенных последовательностей длины 2^n, если занумеровать наборы $\tilde{\sigma} \in \in \{0,1\}^k$ символами a_i, а каждому $\tilde{\alpha} \in \{0,1,*\}^k$ сопоставить символ a_T, где T — множество номеров доопределений $\tilde{\sigma}$ набора $\tilde{\alpha}$. Условию (4) соответствует $l_{\tilde{\alpha}} 2^{-n} \to p_{\tilde{\alpha}}$.

Будем рассматривать реализацию систем F схемами в произвольном конечном базисе и лучшим считать доопределение, реализуемое минимальной схемой. Справедлив аналог утверждения 12, состоящий в том, что в невырожденном случае лучшие доопределения почти всех систем из $\mathcal{F}_n(\boldsymbol{l})$ имеют почти одинаковую структуру, которая может быть найдена подобно (5). То же справедливо для доопределений, возникающих при асимптотически наилучших для $\mathcal{F}_n(\boldsymbol{l})$ методах синтеза.

Литература

[1] *Шоломов Л. А.* Элементы теории недоопределенной информации // Прикладная дискретная математика. Приложение № 2. — 2009. — С. 18–42.

[2] *Шоломов Л. А.* О правиле сложения энтропий для недоопределенных данных // Дискр. анализ и исслед. опер. — 2010. — Т. 17, № 5. — С. 67–90.

[3] *Галлагер Р.* Теория информации и надежная связь. — М.: Советское радио, 1974.

[4] *Нечипорук Э. И.* О сложности вентильных схем, реализующих булевские матрицы с неопределенными элементами // ДАН СССР. — 1965. — Т. 163, № 1. — С. 40–42.

[5] *Вапник В. Н., Червоненкис А. Я.* Теория распознавания образов: статистические проблемы обучения. — М.: Наука, 1974.

[6] *Колмогоров А. Н.* Алгоритм, информация, сложность. — М.: Знание, 1991.

Математическое моделирование неполноты знания модели объекта исследования

Пытьев Ю. П.
yurii.pytyev@gmail.com
Москва, МГУ им. М. В. Ломоносова

В научной, инженерной, технической и прочей творческой деятельности невозможно исключить использование неясной, неполной и недостоверной информации, ассоциированной с опытом, практикой и с полученными ранее знаниями. Рассмотрен метод математического моделирования подобной информации, выраженной в форме субъективных суждений, возможно — коллективных. Основой метода является конструкция *неопределенного случайного элемента*, заданного на произведении пространств: вероятностного $(\Omega, \mathcal{A}, \Pr(\cdot; x))$, известного с точностью до значения параметра $x \in X$ и моделирующего объект исследования, и пространства $(X, \mathcal{P}(X), \mathrm{P}l, \mathrm{Be}l)$ с мерами правдоподобия $\mathrm{P}l$ и доверия $\mathrm{Be}l$, моделирующего субъективные суждения модельера-исследователя о возможных значениях $x \in X$ и его «диалог с моделью» объекта. Показано, что такая модель «диалога модельера-исследователя с моделью объекта» позволяет оценить правдоподобие и доверие его суждений о *любых свойствах объекта*, обусловленных его моделью $(\Omega, \mathcal{A}, \Pr(\cdot; x))$, $x \in X$.

Вероятностные методы применяются при моделировании многих аспектов как неясности и неопределенности, отражающих неполноту знаний, их недостоверность, так и случайности, нечеткости и неточности, относящихся к их содержанию. Модель случайности и нечеткости обычно считается вероятностной или возможностной [1, 2], а неясность и неопределенность ассоциируются с неполным знанием последней. Как правило, «неполное знание» вероятностной модели $(\Omega, \mathcal{A}, \Pr(\cdot; x))$ обусловлено её зависимостью от *неизвестного* параметра $x \in X$.

В работе неизвестный параметр охарактеризован как неопределенный элемент (н. э.) $\widetilde{x}$, моделирующий субъективные суждения модельера об истинности каждого значения $x \in X$ значениями *мер правдоподобия* равенства $\widetilde{x} = x$ и *доверия* неравенства $\widetilde{x} \neq x$ [3].

В данном случае $\widetilde{x}$ — н. э., канонический для пространства с правдоподобием и доверием $(X, \mathcal{P}(X), \mathrm{P}l^{\widetilde{x}}, \mathrm{Be}l^{\widetilde{x}})$ (т. е. определяющий меры $\mathrm{P}l^{\widetilde{x}}$ и $\mathrm{Be}l^{\widetilde{x}}$), в котором X — множество значений $\widetilde{x}$, $\mathcal{P}(X)$ — класс всех подмножеств X, меры правдоподобия $\mathrm{P}l^{\widetilde{x}}(\cdot) : \mathcal{P}(X) \to [0,1]$ и доверия $\mathrm{Be}l^{\widetilde{x}}(\cdot) : \mathcal{P}(X) \to [0,1]$ определены равенствами:

$$\left.\begin{array}{ll} \mathrm{P}l^{\widetilde{x}}(E) \stackrel{\mathrm{def}}{=} & \mathrm{P}l^{\widetilde{x}}(\widetilde{x} \in E) = \sup\limits_{x \in E} t^{\widetilde{x}}(x); \\ \mathrm{Be}l^{\widetilde{x}}(E) \stackrel{\mathrm{def}}{=} & \mathrm{Be}l^{\widetilde{x}}(\widetilde{x} \in E) = \inf\limits_{x \in X \setminus E} s^{\widetilde{x}}(x), \end{array}\right\} \quad (1)$$

в которых $E \in \mathcal{P}(X)$, $x \in X$,

$$t^{\widetilde{x}}(x) \stackrel{\mathrm{def}}{=} \mathrm{P}l^{\widetilde{x}}(\widetilde{x} = x); \quad s^{\widetilde{x}}(x) \stackrel{\mathrm{def}}{=} \mathrm{Be}l^{\widetilde{x}}(\widetilde{x} \neq x), \quad (2)$$

и использован тот факт, что $E = \bigcup_{x \in E}\{x\} \equiv \equiv \bigcap_{x \in X \setminus E}(X \setminus \{x\})$. Функция $t^{\widetilde{x}}(\cdot) : X \to [0,1]$ ($s^{\widetilde{x}}(\cdot) : X \to [0,1]$) называется распределением правдоподобий (доверий) значений $\widetilde{x}$, её значение $t^{\widetilde{x}}(x)$ ($s^{\widetilde{x}}(x)$) в (2) определяет правдоподобие равенства $\widetilde{x} = x \in X$ (доверие неравенства $\widetilde{x} \neq x \in X$), значение $\mathrm{P}l^{\widetilde{x}}(E)$ ($\mathrm{Be}l^{\widetilde{x}}(E)$) — правдоподобие включения $\widetilde{x} \in E$ (доверие включения $\widetilde{x} \in E$) [3].

Меры $\mathrm{P}l^{\widetilde{x}}$ и $\mathrm{Be}l^{\widetilde{x}}$ формально эквивалентны мерам возможности и необходимости [2], в частности численные значения $\mathrm{P}l^{\widetilde{x}}(E)$ и $\mathrm{Be}l^{\widetilde{x}}(E)$, $E \in \mathcal{P}(X)$, в (1), отличные от нуля и единицы, не могут быть содержательно истолкованы, существенна лишь их упорядоченность. Более того, меры $\mathrm{P}l^{\widetilde{x}}(\cdot)$ и $\mathrm{P}l'^{\widetilde{x}}(\cdot)$ ($\mathrm{Be}l^{\widetilde{x}}(\cdot)$ и $\mathrm{Be}l'^{\widetilde{x}}(\cdot)$) эквивалентны, если $\exists\, \gamma(\cdot) \in \Gamma$ $\forall\, E \in \mathcal{P}(X)$ $\gamma(\mathrm{P}l^{\widetilde{x}}(E)) = \mathrm{P}l'^{\widetilde{x}}(E)$ ($\exists\, \gamma(\cdot) \in \Gamma$ $\forall\, E \in \mathcal{P}(X)$ $\gamma(\mathrm{Be}l^{\widetilde{x}}(E)) = \mathrm{Be}l'^{\widetilde{x}}(E)$), где Γ — класс непрерывных, строго монотонных функций $\gamma(\cdot) : [0,1] \to [0,1]$, $\gamma(0) = 0$, $\gamma(1) = 1$, являющийся группой относительно групповой операции «$\circ$», $\gamma \circ \gamma'(a) \stackrel{\mathrm{def}}{=} \gamma(\gamma'(a))$, $a \in [0,1]$. Γ определяет группу автоморфизмов шкал $\mathcal{L} = ([0,1], \leqslant, +, \cdot) \equiv ([0,1], \leqslant, \max, \min)$ и $\widetilde{\mathcal{L}} = ([0,1], \widetilde{\leqslant}, \widetilde{+}, \widetilde{\cdot}) \equiv \equiv ([0,1], \widetilde{\leqslant}, \min, \max)$ значений мер $\mathrm{P}l(\cdot) : \mathcal{P}(X) \to \to \mathcal{L}$ и $\mathrm{Be}l(\cdot) : \mathcal{P}(X) \to \widetilde{\mathcal{L}}$, а именно: $\forall\, \gamma(\cdot) \in \Gamma$ $\gamma([a,b]) = [a,b]$, $\gamma(a * b) = \gamma(a) * \gamma(b)$, где $*$ — символ любой из операций сложения $+$, $\widetilde{+}$, умножения $\cdot$, $\widetilde{\cdot}$ и сравнения $\leqslant$, $\widetilde{\leqslant}$, а тот факт, что $a + b = a \widetilde{\cdot} b = \max\{a,b\}$, $a \cdot b = a \widetilde{+} b = \min\{a,b\}$, $a, b \in [0,1]$, следует из условий непрерывности $* : [0,1]^2 \to [0,1]$, коммутативности $a * b = b * a$, $a \leqslant b \Leftrightarrow b \widetilde{\leqslant} a$ и свойств $a + 0 = a \widetilde{+} 1 = a$, $a + 1 = a \widetilde{\cdot} 1 = 1$, $a \cdot 0 = a \widetilde{+} 0 = 0$ и $a \cdot 1 = a \widetilde{+} 1 = a$ нейтральных элементов 0 и 1 шкал $\mathcal{L}$ и $\widetilde{\mathcal{L}}$ и группы Γ [2, 3]. $\mathcal{L}$ и $\widetilde{\mathcal{L}}$ суть дуальные дистрибутивные решетки относительно решеточных операций $\vee \sim +$, $\wedge \sim \cdot$ и $\vee \sim \widetilde{+}$, $\wedge \sim \widetilde{\cdot}$.

Неопределенный элемент как высказывательная переменная

Если семейство вероятностных пространств $(\Omega, \mathcal{A}, \Pr(\cdot; x))$, $x \in X$, характеризует объект исследования как его *неопределенная вероятностная модель*, то модельер-исследователь может, основываясь на своих недостоверных априорных знаниях свойств объекта, считать $x \in X$ значениями н. э. $\widetilde{x}$

и предположить, насколько, по его мнению, *относительно правдоподобны* равенства $\widetilde{x} = x$, $x \in X$, и насколько *следует относительно доверять* неравенствам $\widetilde{x} \neq x$, $x \in X$. В работе показано, как на основе такой «модели диалога модельера с моделью объекта исследования» определить распределения правдоподобий и доверий любых характеристик объекта исследования как функций н. э. $\widetilde{x}$, таких, например, как неопределенные вероятности $\widetilde{\Pr}(A) \stackrel{\text{def}}{=} \Pr(A;\widetilde{x})$, $A \in \mathcal{A}$, неопределенные числовые характеристики случайных величин, заданных на $(\Omega, \mathcal{A}, \Pr(\cdot;\widetilde{x}))$, решения задачи статистической идентификации состояния неопределенного стохастического объекта и т. д., и как минимизировать риск потерь, сопутствующий решению модельера о значении н. э.

В таком контексте естественно считать, что н. э. $\widetilde{x}$ моделирует *высказывания* о его значениях, истинность и ложность которых не абсолютны. Такая модель н. э. основана на теоретико-множественном представлении логики высказываний, согласно которому в $(X, \mathcal{P}(X), \mathrm{P}l^{\widetilde{x}}, \mathrm{Bel}^{\widetilde{x}})$ X — множество элементарных высказываний, каждое высказывание a взаимно однозначно представлено множеством $A \in \mathcal{P}(X)$ элементарных высказываний $x \in X$, каждое из которых влечет a: $a \leftrightarrow A = \bigcup_{\substack{x \in X \\ x \to a}} \{x\} \equiv \{x \in X,\ x \to a\}$, где $\leftrightarrow$ — символ взаимно однозначного соответствия, $\to$ — символ логической импликации. Наконец, каждое элементарное высказывание $x \in X$ представлено в X одноточечным множеством $\{x\}$, $x \leftrightarrow \{x\}$, и выделяется среди всех высказываний тем, что любое элементарное высказывание $x \in X$ не следует ни из какого высказывания, кроме x и всегда ложного высказывания $\mathbf{0}$.

При таком представлении, если $a \leftrightarrow A$, $b \leftrightarrow B$, то конъюнкция $a \,\&\, b \leftrightarrow A \cap B$, дизъюнкция $a \vee b \leftrightarrow A \cup B$, отрицание $\neg a \leftrightarrow X \setminus A$, импликация $a \to b \equiv (\neg a) \vee b \leftrightarrow (X \setminus A) \cup B$, всегда истинное высказывание $\mathbf{1} \leftrightarrow X$, всегда ложное $\mathbf{0} \leftrightarrow \varnothing$.

В этой связи правдоподобие $t^{\widetilde{x}}(x) = \mathrm{P}l^{\widetilde{x}}(\widetilde{x} = x)$ $(\mathrm{P}l^{\widetilde{x}}(\widetilde{x} \in E))$ будем интерпретировать как правдоподобие истинности высказывания $x \leftrightarrow \{x\}$ $(e \leftrightarrow E)$ и условимся называть правдоподобием (истинности) высказывания, согласно которому $\widetilde{x} = x$ $(\widetilde{x} \in E)$, доверие $s^{\widetilde{x}}(x) = \mathrm{Bel}^{\widetilde{x}}(\widetilde{x} \neq x)$ $(\mathrm{Bel}^{\widetilde{x}}(\widetilde{x} \in E))$ будем интерпретировать как доверие истинности высказывания $\neg x \leftrightarrow X \setminus \{x\}$ $(e \leftrightarrow E)$ и называть доверием (истинности) высказывания, согласно которому $\widetilde{x} \neq x$ $(\widetilde{x} \in E)$, $x \in X$, иногда опуская «истинности».

Если $\varphi(\cdot) : X \to Y$ — некоторая функция, задающая $\widetilde{y} = \varphi(\widetilde{x})$, то $\widetilde{y}$ — н. э., канонический для $(Y, \mathcal{P}(Y), \mathrm{P}l^{\widetilde{y}}, \mathrm{Bel}^{\widetilde{y}})$, и

$$t^{\widetilde{y}}(y) \stackrel{\text{def}}{=} \mathrm{P}l^{\widetilde{y}}(\{y\}) \stackrel{\text{def}}{=} \mathrm{P}l^{\widetilde{x}}(\varphi(\widetilde{x}) = y) = \sup_{\substack{x \in X \\ \varphi(x) = y}} t^{\widetilde{x}}(x), \quad (3)$$

$$s^{\widetilde{y}}(y) \stackrel{\text{def}}{=} \mathrm{Bel}^{\widetilde{y}}(Y \setminus \{y\}) = \mathrm{Bel}^{\widetilde{x}}(\varphi(\widetilde{x}) \neq y) = \inf_{\substack{x \in X, \\ \varphi(x) = y}} s^{\widetilde{x}}(x)$$

— правдоподобие высказывания, согласно которому $\varphi(\widetilde{x}) = y$, и соответственно доверие высказывания, согласно которому $\varphi(\widetilde{x}) \neq y$, $y \in Y$.

Неопределенный случайный элемент

Введем пару пространств: вероятностное $(Y, \mathcal{B}, \Pr^{\eta})$ и с правдоподобием и доверием $(X, \mathcal{P}(X), \mathrm{P}l^{\widetilde{x}}, \mathrm{Bel}^{\widetilde{x}})$, в которых η — канонический для $(Y, \mathcal{B}, \Pr^{\eta})$ случайный элемент, определяющий $\Pr^{\eta}(A) \stackrel{\text{def}}{=} \stackrel{\text{def}}{=} \Pr^{\eta}(\eta \in A)$, $A \in \mathcal{A}$, $\widetilde{x}$ — канонический для $(X, \mathcal{P}(X), \mathrm{P}l^{\widetilde{x}}, \mathrm{Bel}^{\widetilde{x}})$ н. э., определяющий $\mathrm{P}l^{\widetilde{x}}$ и $\mathrm{Bel}^{\widetilde{x}}$ согласно формулам (1), (2), и рассмотрим отображение $q(\cdot,\cdot) : (Y, \mathcal{B}) \times (X, \mathcal{P}(X)) \to (\Omega, \mathcal{A})$, $\mathcal{B}, \mathcal{A}$-измеримое при каждом $x \in X$: $\forall\, x \in X\ \forall\, A \in \mathcal{A}$ $\{y \in Y,\ q(y,x) \in A\} \in \mathcal{B}$. *Функцию* $\widetilde{\xi} = q(\eta, \widetilde{x})$ *случайного* η *и неопределенного* $\widetilde{x}$ *элементов назовем неопределенным случайным элементом, заданным на* $(Y, \mathcal{B}, \Pr^{\eta}) \times (X, \mathcal{P}(X), \mathrm{P}l^{\widetilde{x}}, \mathrm{Bel}^{\widetilde{x}})$ *со значениями в* $(\Omega, \mathcal{A})$. При $\widetilde{x} = x$ $\widetilde{\xi}\Big|_{\widetilde{x}=x} \stackrel{\text{def}}{=} q(\eta, x)$ — случайный элемент, определенный на $(Y, \mathcal{B}, \Pr^{\eta})$ со значениями в $(\Omega, \mathcal{A}, \Pr(\cdot;x))$, $x \in X$, где $\forall\, A \in \mathcal{A}$

$$\Pr(A;x) \stackrel{\text{def}}{=} \Pr^{\eta}(q(\eta,x) \in A) = \int\limits_{\substack{y \in Y, \\ q(y,x) \in A}} \Pr^{\eta}(dy)$$

— вероятность, определяющая семейство $(\Omega, \mathcal{A}, \Pr(\cdot;x))$, $x \in X$, и *неопределенную вероятность* $\widetilde{\Pr}(A) \stackrel{\text{def}}{=} \Pr(A;\widetilde{x})$, $A \in \mathcal{A}$. Поэтому согласно (3)

$$t^{\widetilde{\Pr}(A)}(\mathrm{pr}) \stackrel{\text{def}}{=} \mathrm{P}l^{\widetilde{x}}(\Pr(A;\widetilde{x}) = \mathrm{pr}) = \\ = \sup\{t^{\widetilde{x}}(x) \mid x \in X,\ \Pr(A;x) = \mathrm{pr}\}$$

— правдоподобие высказывания, согласно которому неопределенная вероятность $\widetilde{\Pr}(A)$ события A равна pr; это высказывание для каждого $A \in \mathcal{A}$ представлено множеством $\{x \in X,\ \Pr(A;x) = \mathrm{pr}\}$ элементарных высказываний, $\mathrm{pr} \in [0,1]$. Соответственно

$$s^{\widetilde{\Pr}(A)}(\mathrm{pr}) \stackrel{\text{def}}{=} \mathrm{Bel}^{\widetilde{x}}(\Pr(A;\widetilde{x}) \neq \mathrm{pr}) = \\ = \inf\{s^{\widetilde{x}}(x) \mid x \in X,\ \Pr(A,x) = \mathrm{pr}\}$$

— доверие высказывания, согласно которому $\widetilde{\Pr}(A) \neq \mathrm{pr}$, $\mathrm{pr} \in [0,1]$.

Пример 1. Рассмотрим семейство дискретных вероятностей $\Pr(\cdot;x) : \mathcal{P}(\Omega) \to [0,1]$, $x \in X = (0,\infty)$, где $\Omega = \{\omega_1, \omega_2, \ldots\}$, и

$$\Pr(\{\omega_i\};x) \stackrel{\text{def}}{=} \mathrm{pr}_i(x) = x/(1+x)^i,\ i = 1,2,\ldots, \quad (4)$$

где x — значение неопределенного элемента $\widetilde{x}$ со значениями в $(0,\infty)$.

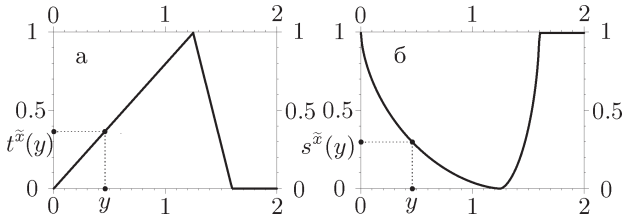


Рис. 1. Распределение правдоподобий значений $\widetilde{x}$ в (6) (*а*) и доверий значений $\widetilde{x}$, $s^{\widetilde{x}}(\cdot) = \theta_\alpha \circ t^{\widetilde{x}}(\cdot)$, $\theta_\alpha(a) = (1 - a^\alpha)^{1/\alpha}$, $a \in [0,1]$, $\alpha = 0{,}63$, в (7) (*б*)

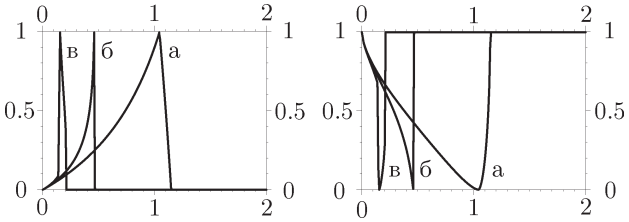


Рис. 2. Графики зависимостей: $\mathrm{Pl}^{\widetilde{x}}(\mathsf{E}_{\widetilde{x}}\xi_\lambda(\cdot) = m)$, (6) (слева), $\mathrm{Bel}^{\widetilde{x}}(\mathsf{E}_{\widetilde{x}}\xi_\lambda(\cdot) \neq m)$, (7) (справа), $m \in (0,\infty)$, при значениях $\lambda = 0{,}11$ (*а*), $\lambda = 1{,}58$ (*б*), $\lambda = 3{,}55$ (*в*)

Рассмотрим высказывание, согласно которому $\forall\, i = 1,2,\ldots \widetilde{\mathrm{pr}}_i \stackrel{\mathrm{def}}{=} \mathrm{pr}_i(\widetilde{x}) = \mathrm{pr}_i$. Его правдоподобие $\mathrm{Pl}^{\widetilde{x}}(\forall\, i \in \{1,2,\ldots\}\ \widetilde{\mathrm{pr}}_i = \mathrm{pr}_i)$ не равно нулю лишь на семействе $\mathrm{pr}_i = \mathrm{pr}_i(y) = y/(1+y)^i$, $i = 1,2,\ldots$, $y \in (0,\infty)$, поэтому $\mathrm{Pl}^{\widetilde{x}}(\forall\, i \in \{1,2,\ldots\}\ \widetilde{\mathrm{pr}}_i = \mathrm{pr}_i(y)) = \mathrm{Pl}^{\widetilde{x}}(\widetilde{x} = y) = t^{\widetilde{x}}(y)$ и, соответственно, $\mathrm{Bel}^{\widetilde{x}}(\exists\, i \in \{1,2,\ldots\}\ \widetilde{\mathrm{pr}}_i \neq \mathrm{pr}_i(y)) = \mathrm{Bel}^{\widetilde{x}}(\widetilde{x} \neq y) = s^{\widetilde{x}}(y)$ суть правдоподобие высказывания, согласно которому $\forall\, i \in \{1,2,\ldots\}\ \widetilde{\mathrm{pr}}_i = \mathrm{pr}_i(y)$, т. е. $\widetilde{\mathrm{Pr}}(\cdot) = \mathrm{Pr}(\cdot; y)$, и соответственно — доверие высказывания, согласно которому $\widetilde{\mathrm{Pr}}(\cdot) \neq \mathrm{Pr}(\cdot; y)$, $y \in (0,1)$.

Рассмотрим неопределенное математическое ожидание $\widetilde{\mathsf{E}}\xi_\lambda \stackrel{\mathrm{def}}{=} \mathsf{E}_{\widetilde{x}}\xi_\lambda(\cdot)$ случайной величины, зависящей от параметра $\lambda > 0$, заданной равенствами

$$\xi_\lambda(\omega_i) = \lambda^{i-1}e^{-\lambda}/(i-1)!,\ i = 1,2,\ldots$$

Так как

$$\mathsf{E}_x\xi_\lambda(\cdot) = \sum_{i=1}^{\infty} \tfrac{\lambda^{i-1}}{(i-1)!}e^{-\lambda}\tfrac{x}{(x+1)^i} = \tfrac{x}{x+1}\exp\left(-\tfrac{\lambda x}{x+1}\right), \quad (5)$$

то согласно (3)

$$t^{\widetilde{\mathsf{E}}\xi_\lambda}(m) \stackrel{\mathrm{def}}{=} \mathrm{Pl}^{\widetilde{x}}(\mathsf{E}_{\widetilde{x}}\xi_\lambda(\cdot) = m) = \\ = \sup\{t^{\widetilde{x}}(x) \mid x \in (0,\infty), \mathsf{E}_x\xi_\lambda(\cdot) = m\}, \quad (6)$$

$m \in (0,\infty)$, — распределение правдоподобий высказываний, откуда следует $\widetilde{\mathsf{E}}\xi_\lambda = m$. Соответственно

$$s^{\widetilde{\mathsf{E}}\xi_\lambda}(m) \stackrel{\mathrm{def}}{=} \mathrm{Bel}^{\widetilde{x}}(\mathsf{E}_{\widetilde{x}}\xi_\lambda(\cdot) \neq m) = \\ = \inf\{s^{\widetilde{x}}(x) \mid x \in (0,\infty),\ \mathsf{E}_x\xi_\lambda(\cdot) = m\}, \quad (7)$$

— распределение доверий высказываний, согласно которым $\widetilde{\mathsf{E}}\xi_\lambda \neq m$, $m \in (0,\infty)$, для каждого $\lambda > 0$ (рис. 1, 2).

Иногда н. э., в частности $\widetilde{\mathsf{E}}\xi_\lambda$, удобнее охарактеризовать не распределениями $t^{\widetilde{\mathsf{E}}\xi_\lambda}(\cdot)$ (6) и $s^{\widetilde{\mathsf{E}}\xi_\lambda}(\cdot)$ (7), а указав его, в известном смысле, оптимальное значение. В таком случае исследователь должен определить семейство пространств $(\mathrm{L}, \mathcal{P}(\mathrm{L}), \mathrm{Pl}^{\widetilde{l}}(\cdot|m,m'),\ \mathrm{Bel}^{\widetilde{l}}(\cdot|m,m'))$, $(m,m') \in (0,\infty)^2$, в котором L — пространство элементарных потерь, $\mathcal{P}(\mathrm{L})$ — класс всех подмножеств L, $\mathrm{Pl}^{\widetilde{l}}(\cdot|\cdot,\cdot)$, $\mathrm{Bel}^{\widetilde{l}}(\cdot|\cdot,\cdot)$ суть переходные правдоподобие и доверие для пространств $((0,\infty)^2, \mathcal{P}((0,\infty)^2))$, $(\mathrm{L}, \mathcal{P}(\mathrm{L}))$, и указать множество $V(\lambda) \in \mathcal{P}(\mathrm{L})$ *важных* для него потерь, $\lambda \in (0,\infty)$. Тогда выражения

$$\mathrm{Pl}^{\widetilde{l}}(V(\lambda)|m,m') = \sup_{l \in V(\lambda)} t^{\widetilde{l}}(l|m,m') \stackrel{\mathrm{def}}{=} \mathrm{pl}l^{\widetilde{\mathsf{E}}\xi_\lambda}_{m,m'},$$

$$\mathrm{Bel}^{\widetilde{l}}(V(\lambda)|m,m') = \inf_{l \in \mathrm{L}\setminus V(\lambda)} s^{\widetilde{l}}(l|m,m') \stackrel{\mathrm{def}}{=} \mathrm{bel}l^{\widetilde{\mathsf{E}}\xi_\lambda}_{m,m'},$$

определят *правдоподобие* $\mathrm{pl}l^{\widetilde{\mathsf{E}}\xi_\lambda}_{m,m'}$ и *доверие* $\mathrm{bel}l^{\widetilde{\mathsf{E}}\xi_\lambda}_{m,m'}$ *потерь, сопутствующих решению* $\widetilde{\mathsf{E}}\xi_\lambda = m'$, *в то время как на самом деле* $\widetilde{\mathsf{E}}\xi_\lambda = m$, $m,m' \in (0,\infty)$. Это позволяет определить оптимальные значения m^* и m_* как минимизирующие правдоподобие и доверие потерь:

$$\sup_{m\in(0,\infty)} \min(\mathrm{pl}l^{\widetilde{\mathsf{E}}\xi_\lambda}_{m,m^*}, t^{\widetilde{\mathsf{E}}\xi_\lambda}(m)) = \\ = \min_{m'\in(0,\infty)} \sup_{m\in(0,\infty)} \min(\mathrm{pl}l^{\widetilde{\mathsf{E}}\xi_\lambda}_{m,m'}, t^{\widetilde{\mathsf{E}}\xi_\lambda}(m)),$$

$$\inf_{m\in(0,\infty)} \max(\mathrm{bel}l^{\widetilde{\mathsf{E}}\xi_\lambda}_{m,m_*}, s^{\widetilde{\mathsf{E}}\xi_\lambda}(m)) = \\ = \min_{m'\in(0,\infty)} \inf_{m\in(0,\infty)} \max(\mathrm{bel}l^{\widetilde{\mathsf{E}}\xi_\lambda}_{m,m'}, s^{\widetilde{\mathsf{E}}\xi_\lambda}(m)).$$

Пример 2. Рассмотрим байесовскую задачу статистической идентификации, в которой модель стохастической системы задана зависящим от неизвестного параметра $x \in X$ распределением $\mathrm{pr}^{\zeta,\varkappa}(z,k;x)$, $z \in Z$, $k \in \{1,\ldots,q\}$, пары $(\zeta,\varkappa)$ случайных элементов (наблюдение, состояние), первый из которых наблюдаем, а второй — нет (см. гл. 6 в [2]). В рассматриваемой задаче идентификации требуется по наблюдению значения $\zeta = z$ принять решение $d(z) \in \{1,\ldots,q\}$ о состоянии системы, если известно, что вероятность «проблем», сопутствующих эксплуатации системы, находящейся в состоянии $k \in \{1,\ldots,q\}$, по правилам, определенным для состояния $d(z)$, равна $\mathrm{pr}l_{k,d(z)}$, $z \in Z$. Ожидаемая вероятность «проблем», сопутствующих правилу идентификации состояния системы, определенному «решающей функцией» $d(\cdot): Z \to \{1,\ldots,q\}$,

$$\mathrm{Pr}\,\mathrm{L}(d(\cdot);x) = \int_Z \mu(dz) \sum_{k=1}^{q} \mathrm{pr}l_{k,d(z)}\mathrm{pr}^{\zeta,\varkappa}(z,k;x)$$

зависит от неизвестного значения $x \in X$.

Оптимальное правило идентификации $d_{\mathrm{minmax}}(\cdot)$ можно определить условием

$$\sup_{x\in X}\Pr \mathrm{L}(d_{\mathrm{minmax}}(\cdot);x)=\min_{d(\cdot)}\sup_{x\in X}\Pr \mathrm{L}(d(\cdot);x),$$

но в данном случае нас интересуют решения, основанные на модели параметра как значения неопределенного элемента $\widetilde{x}$ со значениями в X, позволяющей учесть неформальные соображения субъекта, принимающего решения (С. П. Р.) о значениях $\widetilde{x}$.

Пусть $\widehat{d}(\cdot;x):Z\to\{1,\dots,q\}$ — решение задачи $\Pr_d(z;x)\stackrel{\mathrm{def}}{=}\sum\limits_{k=1}^{q}\mathrm{pr}l_{k,d}\mathrm{pr}^{\zeta,\varkappa}(z,k;x)\sim\min\limits_{d\in\{1,\dots,q\}}$, $z\in$ $\in Z$, $x\in X$, т. е. пусть $\widehat{d}(\cdot;x):Z\to\{1,\dots,q\}$ — некоторая функция, при любом $x\in X$ удовлетворяющая условию

$$\widehat{d}(z;x)\in D(z;x)\stackrel{\mathrm{def}}{=}\{d\in\{1,\dots,q\},\ \Pr_d(z;x)=\\ =\min_{d'\in\{1,\dots,q\}}\Pr_{d'}(z;x)\},\ z\in Z. \quad (8)$$

Тогда, как нетрудно увидеть, функция $\widehat{d}(z;x)$, $z\in Z$, является решением задачи

$$\min_{d(\cdot)}\Pr \mathrm{L}(d(\cdot);x)=\Pr \mathrm{L}(\widehat{d}(\cdot;x),x)=\\ =\int_Z\Pr{}_{\widehat{d}(z;x)}(z;x)\mu(\mathrm{d}z),$$

определяющей оптимальное правило идентификации $\widehat{d}(\cdot;x):Z\to\{1,\dots,q\}$ для каждого $x\in X$, [2].

Пусть $t^{\widetilde{x}}(x)$ и $s^{\widetilde{x}}(x)$, $x\in X$, — правдоподобие и доверие высказываний С. П. Р., согласно которым $\widetilde{x}=x$ и $\widetilde{x}\neq x$ соответственно. Тогда

$$t^{\widetilde{\mathrm{pr}}l}(\mathrm{pr})\stackrel{\mathrm{def}}{=}\mathrm{P}l^{\widetilde{x}}(\Pr \mathrm{L}(\widehat{d}(\cdot;\widetilde{x});\widetilde{x})=\mathrm{pr})=\\ =\sup\{t^{\widetilde{x}}(x)|x\in X,\ \Pr \mathrm{L}(\widehat{d}(\cdot;x);x)=\mathrm{pr}\}; \quad (9)$$

$$s^{\widetilde{\mathrm{pr}}l}(\mathrm{pr})\stackrel{\mathrm{def}}{=}\mathrm{Bel}^{\widetilde{x}}(\Pr \mathrm{L}(\widehat{d}(\cdot;\widetilde{x});\widetilde{x})\neq\mathrm{pr})=\\ =\inf\{s^{\widetilde{x}}(x)|x\in X,\ \Pr \mathrm{L}(\widehat{d}(\cdot;x);x)=\mathrm{pr}\} \quad (10)$$

суть правдоподобие и доверие высказываний С. П. Р., согласно которым $\widetilde{\mathrm{pr}}l=\mathrm{pr}$ и соответственно $\widetilde{\mathrm{pr}}l\neq\mathrm{pr}$, $\mathrm{pr}\in[0,1]$, где $\widetilde{\mathrm{pr}}l\stackrel{\mathrm{def}}{=}\Pr \mathrm{L}(\widehat{d}(\cdot;\widetilde{x});\widetilde{x})$ — *неопределенная вероятность «проблем» со значениями в* $[0,1]$.

Качество правила идентификации, заданного неопределенной функцией $\widehat{d}(\cdot;\widetilde{x}):Z\to\{1,\dots,q\}$, см. (8), характеризуется значениями ожидаемой вероятности «проблем» $\mathrm{pr}\in[0,1]$, при которых равенство $\widetilde{\mathrm{pr}}l=\mathrm{pr}$ имеет максимальное правдоподобие, а неравенство $\widetilde{\mathrm{pr}}l\neq\mathrm{pr}$ — минимальное доверие, т. е. чем меньше значения $\operatorname{argmax}\limits_{\mathrm{pr}\in[0,1]}t^{\widetilde{\mathrm{pr}}l}(\mathrm{pr})$ в (9) и $\operatorname{argmin}\limits_{\mathrm{pr}\in[0,1]}s^{\widetilde{\mathrm{pr}}l}(\mathrm{pr})$ в (10), тем лучше неопределенное правило решения $\widehat{d}(\cdot;\widetilde{z})$; назовем его *минимаксным*.

Для построения *максиминных правил* определим распределения неопределенной вероятности «проблем», отвечающие некоторой решающей функции $d(\cdot):Z\to\{1,\dots,q\}$:

$$t^{\widetilde{\mathrm{pr}}l}(\mathrm{pr};d(\cdot))\stackrel{\mathrm{def}}{=}\mathrm{P}l^{\widetilde{x}}(\Pr \mathrm{L}(d(\cdot);\widetilde{x})=\mathrm{pr})=\\ =\sup\{t^{\widetilde{x}}(x)|x\in X,\Pr \mathrm{L}(d(\cdot);x)=\mathrm{pr}\},$$

$$s^{\widetilde{\mathrm{pr}}l}(\mathrm{pr};d(\cdot))\stackrel{\mathrm{def}}{=}\mathrm{Bel}^{\widetilde{x}}(\Pr \mathrm{L}(d(\cdot);\widetilde{x})\neq\mathrm{pr})=\\ =\inf\{s^{\widetilde{x}}(x)|x\in X,\Pr \mathrm{L}(d(\cdot);z)=\mathrm{pr}\},\ \mathrm{pr}\in[0,1].$$

Оптимальные минимаксные правила идентификации $d^*(\cdot)$ и $d_*(\cdot)$ определим как решения задач

$$\left.\begin{aligned}\operatorname*{argmax}_{\mathrm{pr}\in[0,1]}t^{\widetilde{\mathrm{pr}}}(\mathrm{pr};d^*(\cdot))&=\min_{d(\cdot)}\operatorname*{argmax}_{\mathrm{pr}\in[0,1]}t^{\widetilde{\mathrm{pr}}}(\mathrm{pr};d(\cdot));\\ \operatorname*{argmin}_{\mathrm{pr}\in[0,1]}s^{\widetilde{\mathrm{pr}}}(\mathrm{pr};d_*(\cdot))&=\min_{d(\cdot)}\operatorname*{argmin}_{\mathrm{pr}\in[0,1]}s^{\widetilde{\mathrm{pr}}}(\mathrm{pr};d(\cdot)).\end{aligned}\right\} \quad (11)$$

Правила $d^*(\cdot)$ и $d_*(\cdot)$ тем лучше, чем меньше соответствующие левые части равенств в (11).

В заключение заметим, что в тех редких случаях, когда исследователь не может априори охарактеризовать свойства объекта в терминах правдоподобий и доверий значений параметра $x\in X$, его модель «абсолютного незнания» этих свойств определяется распределениями $t^{\widetilde{x}}(x)=1$, $s^{\widetilde{x}}(x)=$ $=0$, $x\in X$, (2), инвариантными относительно выбора шкал $\mathcal{L}$ и $\widetilde{\mathcal{L}}$ значений $\mathrm{P}l$ и Bel. В этих случаях такие же распределения в (3) будет иметь любая функция $\varphi(\widetilde{x})$.

Заключение

На примере известного с точностью до значения $x\in X$ вероятностного пространства $(\Omega,\mathcal{A},\Pr(\cdot;x))$, моделирующего объект исследования, показано, что если модельер-исследователь моделирует значения $x\in X$ как значения неопределенного элемента $\widetilde{x}$ и *может априори* предположить, насколько, по его мнению, *относительно* правдоподобны равенства $\widetilde{x}=x$, $x\in X$, и насколько следует *относительно* доверять неравенствам $\widetilde{x}\neq x$, $x\in X$, то, как следствие, он сможет оценить правдоподобия и доверия любых суждений о значениях неопределенных вероятностей $\Pr(A;\widetilde{x})$, $A\in\mathcal{A}$, математических ожиданий $\int\limits_\Omega\varphi(\omega,\widetilde{x})\Pr(\mathrm{d}\omega;\widetilde{x})$ и других, основанных на модели $(\Omega,\mathcal{A},\Pr(\cdot;x))$, $x\in X$, свойствах объекта исследования, что практически *невозможно оценить априори*.

Автор глубоко признателен Ю. М. Нагорному, рассчитавшему зависимости на рис. 2 и подготовившему электронный вариант рукописи.

Литература

[1] *Дюбуа Д., Прад А.* Теория возможностей. — М.: Радио и связь, 1990.

[2] *Пытьев Ю. П.* Возможность как альтернатива вероятности. — 2-е изд. — М.: Физматлит, 2011.

[3] *Пытьев Ю. П.* Неопределенные нечеткие модели и их применения. // Интеллектуальные системы. — 2004. — Т. 8. Вып. 1–4. — С. 147–310.

Вероятностные и возможностные измерительно-вычислительные преобразователи как средства измерений: сравнительный анализ качества*

Пытьев Ю. П., Фаломкина О. В., Макеев И. В., Артемов А. В.
yuri.pytyev@gmail.com, olesya.falomkina@gmail.com
Москва, МГУ им. М. В. Ломоносова, физический факультет

Работа посвящена компьютерному моделированию вероятностных и возможностных измерительно-вычислительных преобразователей (ИВП) и сравнительному анализу их качества. ИВП, измеряемый объект и среда охарактеризованы вероятностной и возможностной моделями, причем возможностная модель максимально согласована с вероятностной [1]. В работе рассмотрены возможностные модели ИВП первого порядка на основе двух вариантов теории возможностей, зависимости качества ИВП как средства измерений от качества измерительного преобразователя (ИП), математические основы интеллектуального диалога исследователя с ИВП.

Измерительно-вычислительная система (ИВС) в точном соответствии с ее названием состоит из двух компонент — *измерительной* и *вычислительной*. Обычно первая является преобразователем специфического для измерения воздействия: радиационного, теплового, механического или какого-либо другого — в электрический сигнал. Принципы действия измерительных преобразователей, называемых также датчиками, основаны на известных физических явлениях электромагнитной индукции, термо- и пьезоэлектричества и многих других, см., например, [2]. Далее в вычислительной компоненте электрический сигнал оцифровывается и подвергается математическому преобразованию, которое призвано, с одной стороны, извлечь из результатов измерения все то, что интересует исследователя, а с другой, — облечь это в форму, удобную как для восприятия, так и для «диалога» исследователя с ИВС.

При измерении в результате взаимодействия измеряемого объекта, среды и измерительной компоненты ИВС на ее входе формируется сигнал f, несущий информацию об измеряемом объекте и среде. Измерительная компонента преобразует f в (электрический) сигнал

$$\xi = Af + \nu, \tag{1}$$

где A — оператор, моделирующий физические процессы, определяющие преобразование внешнего воздействия f в электрический сигнал, ν — шум, погрешность преобразования.

Понятно, что на уровне измерительной компоненты все процессы контролируются физическими законами со свойственными им хорошо известными ограничениями и запретами — термодинамическими, дифракционными, квантовыми и т. п. На уровне ИВС все выглядит принципиально подругому, поскольку вычислительная компонента обычно позволяет, грубо говоря, математически смоделировать и вычислить то, что непосредственно ненаблюдаемо. При этом решающую роль играют как математические свойства физических моделей измеряемого объекта, среды, измерительной компоненты ИВС и их взаимодействия, так и используемые математические методы и алгоритмы решения задач интерпретации измерений, которые в конечном счете и определят *предельные возможности* ИВС *как средства измерений* (точность, чувствительность, разрешение и т. д.). По этой причине теория ИВС как средств измерений имеет мало общего с тем, что составляет основы классического приборостроения. В частности, требования к измерительной компоненте, обеспечивающие наивысшее качество ИВС как средства измерений, существенно отличаются от требований, гарантирующих наивысшее качество самой измерительной компоненты как измерительного прибора того же назначения.

Еще одно принципиальное отличие ИВС как средства измерений от обычного измерительного прибора обусловлено возможностью «интеллектуального диалога» исследователя и ИВС [3]. На этапах анализа и интерпретации измерений теория ИВС позволяет исследователю наиболее полно учесть свой научный опыт, оценить как точность, так и адекватность найденных значений параметров исследуемого объекта или явления, охарактеризовать адекватность математических моделей, используемых при интерпретации измерений.

Можно без преувеличения сказать, что ИВС образуют принципиально новый класс измерительных средств, позволяющих создавать идеальные измерительные приборы для научных исследований и промышленности.

Концепция ИВС *как средства измерений* может быть разъяснена на примере системы «измеряемый объект — среда — измерительный прибор — вычислитель», характерной для большинства экспериментальных исследований. Как известно, в процессе измерения объект, измерительный прибор и среда, в которой находятся объект и измеритель-

Работа выполнена при финансовой поддержке РФФИ, проекты № 11-07-00722-а.

ный прибор, взаимодействуют и оказываются в той или иной степени возмущенными. Поскольку исследователя обычно интересуют значения параметров объекта, свойственные невозмущенной системе «исследуемый объект — среда» или, иначе говоря, значения параметров объекта в его естественном состоянии, в практике измерений получил распространение принцип, согласно которому измерительный прибор должен как можно меньше возмущать объект и среду. Концепция ИВС как средства измерений основана на совершенно ином принципе измерений: в процессе измерения объект и среда могут претерпевать существенные возмущения, но на выходе ИВС должны быть *максимально точные значения параметров исследуемого объекта, не искаженные при измерении* (в теории ИВС понятия «исследуемый (изучаемый) объект» и «измеряемый (наблюдаемый) объект» не совпадают). Поэтому выходной сигнал ИВС следует интерпретировать как *максимально точную версию выходного сигнала идеального измерительного прибора*, позволяющего исследователю получать информацию о принципиально ненаблюдаемых характеристиках исследуемого объекта. Подчеркнем, что речь идет, во-первых, о таких характеристиках, которые, как правило, не могут быть измерены непосредственно и должны быть *вычислены* на основе воздействия объекта и среды на измерительную компоненту ИВС, во-вторых, должны быть вычислены характеристики не измеряемого объекта, искаженные при измерении, а исследуемого, свойственные его естественному состоянию.

Теория ИВС включает новое понятие качества измерительной компоненты, обеспечивающей наивысшее качество ИВС как средства измерений, существенно отличающееся от традиционного. Дело в том, что для наиболее качественных измерений измерительный прибор как таковой и как измерительная компонента ИВС должны обладать значительно различающимися характеристиками. Например, плохой в обычном понимании измерительный прибор (низкое разрешение, высокий уровень шума и т. д.) как измерительная компонента ИВС может обеспечить параметры ИВС как измерительного прибора того же (или другого) назначения, существенно более высокие, чем при прочих равных условиях обеспечивает хороший (в том же смысле) измерительный прибор, используемый как измерительная компонента ИВС. Короче говоря, вопрос о том, какими физическими характеристиками должен обладать измерительный прибор, решается существенно по-разному в зависимости от того, как будет использоваться прибор: как измерительная компонента ИВС или сам по себе.

Разумеется, для построения ИВС необходима математическая модель системы «измеряемый объект — среда — измерительный прибор», отражающая свойства объекта, среды и прибора с учетом их взаимодействия. Кроме этого, необходима математическая модель, связывающая входной сигнал f измерительной компоненты ИВС со значениями параметров исследуемого объекта, свойственными его естественному состоянию. Теория ИВС позволяет, используя эти модели и выходной сигнал ξ измерительной компоненты ИВС, «вычислить» с помощью ее вычислительной компоненты *наиболее точную* версию характеристик ненаблюдаемой системы «исследуемый объект — среда» и оценить сопутствующую погрешность.

Иначе можно сказать, что вычислительная компонента ИВС решает *задачу редукции измерения* $\xi = \xi(f)$ (1) *к значению* $u = Uf$ *параметров исследуемого объекта в системе «исследуемый объект — среда» и оценивает погрешность редукции.*

В работе под методами редукции измерений понимается математический формализм, позволяющий по результатам измерений в системе «измеряемый объект — среда — измерительный прибор» получать *наиболее точное* описание характеристик ненаблюдаемой (виртуальной) системы «исследуемый объект — среда», *определенных целью исследования.* Операторы A и U определяются как составные части моделей систем «измеряемый объект — среда — измерительный прибор» и «исследуемый объект — среда».

Редукцией измерения ξ, *выполненного на приборе* A, *к виду, свойственному измерению на приборе* U (*к выходному сигналу* Uf *прибора* U), *называется такое преобразование* $\xi \to R_*\xi$, *при котором* $R_*\xi$ *является (в известном смысле) наиболее точной (в заданном классе преобразований* R*) версией выходного сигнала* Uf *прибора* U.

Тот факт, что при фиксированном классе $\mathbb{R}$ максимальная точность интерпретации $R_*\xi$ как Uf определяется только перечисленными выше моделями, позволяет ставить и решать задачи *оптимального проектирования* измерительной компоненты ИВС, обеспечивающей наиболее точный синтез выходного сигнала идеального прибора[1] U (см. [4]), и для каждой конкретной измерительной компоненты ИВС позволяет вычислять *предельную точность* синтеза выходного сигнала U, *определяющую качество* ИВС *как идеального измерительного прибора* U.

Метод редукции измерений отличается от многих широко распространенных методов «обработки» измерений, таких, например, как методы наименьших квадратов и их регуляризованные варианты [5], метод максимальной энтропии [6] и др. [7],

[1] На практике обычно одна и та же измерительная компонента ИВС используется для изучения различных объектов, каждому из которых, как правило, соответствует специфический «идеальный» измерительный прибор U.

не имеющих прямого отношения к минимизации погрешности интерпретации измерения. Эти методы не могут служить основой теории ИВС как средств измерений, поскольку, не гарантируя максимальную точность интерпретации измерения, они, во-первых, не позволяют определить предельные возможности конкретной ИВС как средства измерений и, во-вторых, не позволяют сформулировать требования к ее измерительной компоненте, обеспечивающие наивысшее качество ИВС как этого средства измерений [4].

Согласно теории ИВС измерительный преобразователь (ИП) и вычислительный преобразователь (ВП) рассматриваются как единый прибор ИВП, выполняющий функции средства измерения. На практике широко используются ИП, математические модели которых описываются дифференциальными уравнениями [2]. Порядком ИП и соответствующего ИВП называется порядок дифференциального уравнения, описывающего ИП.

Работа посвящена сравнительному анализу качества вероятностных и возможностных ИВП. Вероятностные модели ИВП рассмотрены в монографии [4]. ИВП, измеряемый объект и среда охарактеризованы вероятностной и возможностной моделями, причем возможностная модель максимально согласована с вероятностной [1].

В работе рассмотрены теоретико-возможностные модели ИВП первого порядка на основе двух вариантов теории возможностей, далее называемых «первым» и «вторым» соответственно [1]. Основное отличие этих вариантов друг от друга состоит в определении шкалы значений возможности $\mathcal{L}$. В «первом» варианте $\mathcal{L} = \{[0,1], \leqslant, \max, \min\}$ — отрезок $[0,1]$ с естественной упорядоченностью и операциями сложения max и умножения min. Во «втором» варианте операция сложения понимается как max, умножение — как обычное умножение. Это отличие приводит к существенно разным свойствам моделей ИВП, в частности, в «первом» варианте эффект накопления информации при повторных измерениях отсутствует, в то время как во «втором» варианте имеются законы больших чисел, центральная предельная теорема и т. п. [1].

В работе исследована зависимость качества ИВП как средства измерений от характеристик ИП. В статистической теории ИВП вопрос о качестве ИП, гарантирующем наивысшее качество ИВП, решен (см. гл. 10 [4]). Задача Коши, моделирующая датчик первого порядка, имеет вид [4]:

$$\begin{cases} \alpha \dot{u}(t) + \beta u(t) = f(t), & 0 < t < T; \\ u(0) = 0, & \dot{u}(t) \triangleq du(t)/dt, \end{cases} \qquad (2)$$

$u(t)$ — выходной сигнал ИП первого порядка в любой момент времени $t \in [0,T]$, где $[0,T]$ — промежуток времени, в течение которого производится измерение, $f(t)$— воздействие измеряемого объекта на ИП (температура, влажность и т. п.) в момент времени t, α и β — материальные параметры датчика, в данном случае не зависящие от времени. В начальный момент $t = 0$ значение $u(0)$ считается равным нулю. В монографии [4] показано, что при любом фиксированном операторе U, допускающем редукцию, множество точек на плоскости (α, β), на котором качество ИВП, охарактеризованное среднеквадратичной ошибкой редукции, постоянно, имеет вид эллипса (см. рис. 1). Этот важный результат позволяет, не меняя качество ИВП, выбирать технологические параметры ИВП так, чтобы минимизировать стоимость датчика. Аналогичный результат получен для возможностных, и, в частности, для интервальных моделей ИВП.

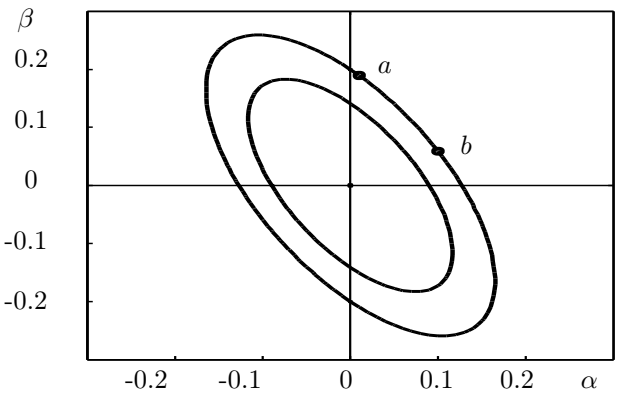


Рис. 1. Линии уровня с. к. погрешности синтеза выходного сигнала ИП, см. гл. 10 [4]. Технические параметры α и β определяют положение точки a на эллипсе.

В работе рассмотрены математические основы интеллектуального диалога исследователя с ИВП. Как известно, диалог исследователя с ИВП не может быть основан на вероятностной модели исследуемого объекта, поскольку исследователь не может формулировать свои мнения, гипотезы, заключения и т.п. в терминах формализма теории вероятностей и математической статистики (заключения и предложения исследователя о значениях вероятности, дисперсии и т.п. вероятностных характеристиках, как правило, не могут быть обоснованы его научным опытом и умозрительными построениями) [3]. Формализм мер правдоподобия и доверия, формально идентичных мерам возможности и необходимости, позволяет математически моделировать субъективные суждения исследователя об изучаемом объекте [3].

Выводы

Удалось сравнить качество вероятностных и возможностных моделей при решении задачи интерпретации данных, а также возможностные модели между собой. Основное отличие между воз-

можностными моделями, отмеченное в тексте, проиллюстрировано на примере решения задачи интерпретации данных. В задаче эмпирического восстановления модели ИП (в частности, материальных параметров α, β) «второму» варианту теории возможностей свойственны трудности, сопутствующие восстановлению вероятностных моделей. Что касается задачи построения множества значений α, β, отвечающих заданному качеству, то для теоретико-возможностных моделей эта задача находится в стадии исследования.

Литература

[1] *Ю. П. Пытьев.* Возможность как альтернатива вероятности. Математические и эмпирические основы, применения, издание 2-е, переработанное и дополненное, в печати.

[2] *Азизов А. М., Гордов А. Н.* Точность измерительных преобразователей. — Л.: Энергия, 1967. — 300 с.

[3] *Пытьев Ю. П.* Математическое моделирование неполноты знания модели объекта исследования // Доклады 15-й Всероссийской конференции «Математические методы распознавания образов», 2011.

[4] *Пытьев Ю. П.* Методы математического моделирования измерительно-вычислительных систем. — М.: Физматлит, 2004.

[5] *Тихонов А. Н., Арсенин В. Я.* Методы решения некорректных задач. — М.: Наука, 1979. — 286 с.

[6] *Теребиж В. Ю.* Восстановление изображений при минимальной априорной информации // Успехи физических наук. — 1995. — Т. 165, № 2. — С. 143–176.

[7] *Гончарский А. В., Романов С. Ю., Черепащук А. Н.* Конечно-параметрические обратные задачи астрофизики. — М.: Изд-во Моск. ун-та, 1991. — 192 с.

Теоретико-возможностные модели матричных игр двух субъектов*

Папилин С. С., Пытьев Ю. П.
papilin@physics.msu.ru, yuri.pytyev@gmail.com
Москва, Московский государственный университет имени М. В. Ломоносова

В докладе с позиций двух вариантов теории возможностей рассмотрены матричные игры двух субъектов. Найдены оптимальные стратегии игроков и цены игр, исследовано существование четких стратегий. Результаты сравниваются с качественными результатами в теории матричных игр.

В докладе рассмотрена матричная игра, в которой решения игроков влияют на возможность некоторого события W, желательного для одного игрока и нежелательного для другого. В биматричной игре учтены возможности этого события и его отрицания. В работе [1] анализ этих игр был проведен с позиций первого варианта теории возможностей, здесь он проводится с позиций двух вариантов, определенных в монографии [2].

Возможностная модель матричной игры двух субъектов

В игре участвуют два субъекта, «игрок A» и «игрок B». Игроки принимают нечеткие решения $\alpha \in \{1,\dots,m\}$ и $\beta \in \{1,\dots,n\}$ независимо друг от друга, где α и β суть независимые нечеткие элементы со значениями в $\{1,\dots,m\}$ и в $\{1,\dots,n\}$ соответственно,

$$\mathrm{p}_i^{\mathrm{A}} \stackrel{\mathrm{def}}{=} \mathrm{P}^{\mathrm{A}}(\alpha = i) \geqslant 0, \quad \max_{1\leqslant i\leqslant m} \mathrm{p}_i^{\mathrm{A}} = 1; \qquad (1)$$

$$\mathrm{p}_j^{\mathrm{B}} \stackrel{\mathrm{def}}{=} \mathrm{P}^{\mathrm{B}}(\beta = j) \geqslant 0, \quad \max_{1\leqslant j\leqslant n} \mathrm{p}_j^{\mathrm{B}} = 1, \qquad (2)$$

суть распределения возможностей значений α и β, в данном случае — возможностей решений игроков A и B. Наборы возможностей

$$\mathrm{p}^{\mathrm{A}} \stackrel{\mathrm{def}}{=} \{\mathrm{p}_1^{\mathrm{A}}, \mathrm{p}_2^{\mathrm{A}}, \dots, \mathrm{p}_m^{\mathrm{A}}\},\ \mathrm{p}^{\mathrm{B}} \stackrel{\mathrm{def}}{=} \{\mathrm{p}_1^{\mathrm{B}}, \mathrm{p}_2^{\mathrm{B}}, \dots, \mathrm{p}_n^{\mathrm{B}}\} \quad (3)$$

определяют нечеткие (фазифицированные) стратегии принятия решений игроками A и B соответственно, любые решения «i» игрока A и «j» игрока B определяют их четкие (нефазифицированные) стратегии принятия решений.

Определим матрицу переходных возможностей события W, матричные элементы которой

$$s_{ij} \stackrel{\mathrm{def}}{=} \mathrm{P}(W \,|\, \alpha = i, \beta = j) \qquad (4)$$

задают зависимость переходной возможности W от решений игроков.

Так как решения принимаются независимо, то

$$\mathrm{P}(\alpha = i, \beta = j) = \mathrm{P}^{\mathrm{A}}(\alpha = i) \bullet \mathrm{P}^{\mathrm{B}}(\beta = j) = \mathrm{p}_i^{\mathrm{A}} \bullet \mathrm{p}_j^{\mathrm{B}},$$

где «$\bullet$» есть символ операции минимума (min) в первом варианте теории возможностей и «обычного умножения» («$\times$») во втором.

Соответственно,

$$\mathrm{P}(W, \alpha = i, \beta = j) = \mathrm{P}(W \,|\, \alpha = i, \beta = j) \bullet \\ \bullet\, \mathrm{P}(\alpha = i, \beta = j) = s_{ij} \bullet \mathrm{p}_i^{\mathrm{A}} \bullet \mathrm{p}_j^{\mathrm{B}},$$

и *возможность события W как функция нечетких стратегий игроков*

$$\mathrm{P}(W \,|\, \mathrm{p}^{\mathrm{A}}, \mathrm{p}^{\mathrm{B}}) = \max_{\substack{1\leqslant i\leqslant m \\ 1\leqslant j\leqslant n}} \mathrm{P}(W, \alpha = i, \beta = j) = \\ = \max_{\substack{1\leqslant i\leqslant m \\ 1\leqslant j\leqslant n}} s_{ij} \bullet \mathrm{p}_i^{\mathrm{A}} \bullet \mathrm{p}_j^{\mathrm{B}} \stackrel{\mathrm{def}}{=} S(\mathrm{p}^{\mathrm{A}}, \mathrm{p}^{\mathrm{B}}) \quad (5)$$

определяет возможность «выигрыша» игрока A *и возможность «проигрыша» игрока* B, *сопутствующую используемым ими стратегиям* p^{A} *и* p^{B} *принятия решений.*

В рассматриваемой игре игрок A стремится возможность (5) максимизировать, игрок B — минимизировать, поскольку для игрока A событие W — «выигрыш», а для B — «проигрыш».

Определение 1. *Возможностной моделью фазифицированного акта игры назовем классы пространств с возможностями: пространств*

$$\big(\Omega_m \times \Omega_n, \mathcal{P}(\Omega_m \times \Omega_n), \mathrm{P}\big),$$

$\mathrm{P} \in \mathcal{P}_m \times \mathcal{P}_n$, *для каждого из которых нечеткий элемент* (α, β) *является каноническим, и пространств* $\big(\Omega, \mathcal{P}(\Omega), \mathrm{P}^{(i,j)}\big)$, $(i,j) \in \Omega_m \times \Omega_n$, *где* $\Omega_m = \{1,\dots,m\}$, $\Omega_n = \{1,\dots,n\}$, $\mathcal{P}_m$, $\mathcal{P}_n$ *суть классы всех возможностей* $\mathrm{P}^{\mathrm{A}}(\cdot)\colon \mathcal{P}(\Omega_m) \to [0,1]$, $\mathrm{P}^{\mathrm{B}}(\cdot)\colon \mathcal{P}(\Omega_n) \to [0,1]$ *в* (1), (2), Ω — *пространство «элементарных выигрышей» игрока A и «элементарных проигрышей» игрока B,*

$$\mathrm{P}(\alpha = i, \beta = j) = \mathrm{P}^{\mathrm{A}}(\{i\}) \bullet \mathrm{P}^{\mathrm{B}}(\{j\}) = \mathrm{p}_i^{\mathrm{A}} \bullet \mathrm{p}_j^{\mathrm{B}},$$

$i = 1,\dots,m,\ j = 1,\dots,n$; $\mathrm{P}^{(\cdot,\cdot)}(\cdot)$ — *переходная возможность для пространств* $\big(\Omega_m \times \Omega_n, \mathcal{P}(\Omega_m \times \Omega_n)\big)$ *и* $\big(\Omega, \mathcal{P}(\Omega)\big)$, *определяющая возможность* $s_{ij} = \mathrm{P}^{(i,j)}(W)$ *«выигрыша»* $W \in \mathcal{P}(\Omega)$ *игрока A, принявшего i–ое решение, и «проигрыша»* $W \in \mathcal{P}(\Omega)$

Работа выполнена при финансовой поддержке РФФИ, проект №11-07-00722-а.

игрока B, принявшего j–ое решение, $i = 1, \ldots, m$, $j = 1, \ldots, n$.

Результатом фазифицированного акта игры назовем возможность (5)

$$\mathrm{P}(W \,|\, \mathrm{p}^{\mathrm{A}}, \mathrm{p}^{\mathrm{B}}) = \max_{1 \leqslant i \leqslant m, 1 \leqslant j \leqslant n} \mathrm{P}^{(i,j)}(W) \bullet \mathrm{p}^{\mathrm{A}}_i \bullet \mathrm{p}^{\mathrm{B}}_j$$

«выигрыша» игрока A, использующего стратегию $\mathrm{p}^{\mathrm{A}} \in \mathcal{P}^{\mathrm{A}}$, *и «проигрыша» игрока B, использующего стратегию* $\mathrm{p}^{\mathrm{B}} \in \mathcal{P}^{\mathrm{B}}$.

Максиминная стратегия игрока A

Пусть игрок A выбирает стратегию p^{A}, а игрок B минимизирует возможность проигрыша, выбрав в (5) стратегию $\mathrm{p}^{\mathrm{B}} = \mathrm{p}^{\mathrm{B}}(\mathrm{p}^{\mathrm{A}})$; обозначим минимальную возможность проигрыша игрока B

$$S_{\mathrm{A}}(\mathrm{p}^{\mathrm{A}}) \stackrel{\mathrm{def}}{=} \min_{\mathrm{p}^{\mathrm{B}} \in \mathcal{P}^{\mathrm{B}}} S(\mathrm{p}^{\mathrm{A}}, \mathrm{p}^{\mathrm{B}}) = S\big(\mathrm{p}^{\mathrm{A}}, \mathrm{p}^{\mathrm{B}}(\mathrm{p}^{\mathrm{A}})\big).$$

Максиминную стратегию $\mathrm{p}^{*\mathrm{A}}$ *игрока* A определим как любое решение задачи

$$S_{\mathrm{A}}(\mathrm{p}^{*\mathrm{A}}) = \max_{\mathrm{p}^{\mathrm{A}} \in \mathcal{P}^{\mathrm{A}}} S_{\mathrm{A}}(\mathrm{p}^{\mathrm{A}}) \equiv$$

$$\equiv \max_{\mathrm{p}^{\mathrm{A}} \in \mathcal{P}^{\mathrm{A}}} \min_{\mathrm{p}^{\mathrm{B}} \in \mathcal{P}^{\mathrm{B}}} S(\mathrm{p}^{\mathrm{A}}, \mathrm{p}^{\mathrm{B}}) \stackrel{\mathrm{def}}{=} s_{\max\min}, \quad (6)$$

где $s_{\max\min}$ — *минимальная возможность* «выигрыша» игрока A, которую назовем *максиминной возможностью*; согласно (6)

$$\forall \mathrm{p}^{\mathrm{B}} \in \mathcal{P}^{\mathrm{B}} \quad S(\mathrm{p}^{*\mathrm{A}}, \mathrm{p}^{\mathrm{B}}) \geqslant s_{\max\min}.$$

Функцию $S_{\mathrm{A}}(\mathrm{p}^{\mathrm{A}})$ можно преобразовать к виду

$$S_{\mathrm{A}}(\mathrm{p}^{\mathrm{A}}) = \min_{1 \leqslant j \leqslant n} \max_{1 \leqslant i \leqslant m} s_{ij} \bullet \mathrm{p}^{\mathrm{A}}_i. \quad (7)$$

Максиминной стратегией игрока A будет, в частности, стратегия со всеми $\mathrm{p}^{\mathrm{A}}_i$, $1 \leqslant i \leqslant m$, равными единице, и *максиминная возможность*

$$s_{\max\min} = \max_{\mathrm{p}^{\mathrm{A}}} \min_{\mathrm{p}^{\mathrm{B}}} S(\mathrm{p}^{\mathrm{A}}, \mathrm{p}^{\mathrm{B}}) = \min_{1 \leqslant j \leqslant n} \max_{1 \leqslant i \leqslant m} s_{ij}. \quad (8)$$

Множество всех максиминных стратегий по (6) задается условием

$$\min_{1 \leqslant j \leqslant n} \max_{1 \leqslant i \leqslant m} s_{ij} \bullet \mathrm{p}^{*\mathrm{A}}_i = s_{\max\min}. \quad (9)$$

Что касается вопроса *о существовании четких максиминных стратегий* $\mathrm{p}^{*\mathrm{A}}$, в которых $\mathrm{p}^{*\mathrm{A}}_i = \begin{cases} 1, & i = i_0; \\ 0, & i \neq i_0, \end{cases}$ $i = 1, \ldots, m$, то ответ зависит от матрицы переходных возможностей (4). Для четких стратегий, согласно условию (9),

$$\min_{1 \leqslant j \leqslant n} s_{i_0 j} = s_{\max\min}. \quad (10)$$

Например, для матрицы $\{s_{ij}\} = \left(\begin{smallmatrix} 1/4 & 1/2 \\ 3/4 & 1 \end{smallmatrix}\right)$ ($m = 2$, $n = 2$, i — номер строки, j — номер столбца) по формуле (8) $s_{\max\min} = 3/4$. Согласно условию (10) чистая стратегия с $i_0 = 2$ является максиминной, а с $i_0 = 1$ — не является.

А для матрицы $\{s_{ij}\} = \left(\begin{smallmatrix} 3/4 & 1/2 \\ 1/4 & 1 \end{smallmatrix}\right)$ также $s_{\max\min} = 3/4$, но чистых максиминных стратегий нет (ни в одной строке минимум не равен $s_{\max\min}$).

Минимаксная стратегия игрока B

Пусть игрок B выбирает стратегию p^{B}, а игрок A максимизирует возможность выигрыша, выбрав в (5) стратегию $\mathrm{p}^{\mathrm{A}} = \mathrm{p}^{\mathrm{A}}(\mathrm{p}^{\mathrm{B}})$; обозначим эту максимальную возможность

$$S_{\mathrm{B}}(\mathrm{p}^{\mathrm{B}}) \stackrel{\mathrm{def}}{=} \max_{\mathrm{p}^{\mathrm{A}} \in \mathcal{P}^{\mathrm{A}}} S(\mathrm{p}^{\mathrm{A}}, \mathrm{p}^{\mathrm{B}}) = S\big(\mathrm{p}^{\mathrm{A}}(\mathrm{p}^{\mathrm{B}}), \mathrm{p}^{\mathrm{B}}\big). \quad (11)$$

Минимаксную стратегию $\mathrm{p}^{\mathrm{B}}_*$ *игрока* B определим как любое решение задачи

$$S_{\mathrm{B}}(\mathrm{p}^{\mathrm{B}}_*) = \min_{\mathrm{p}^{\mathrm{B}} \in \mathcal{P}^{\mathrm{B}}} S_{\mathrm{B}}(\mathrm{p}^{\mathrm{B}}) \equiv$$

$$\equiv \min_{\mathrm{p}^{\mathrm{B}} \in \mathcal{P}^{\mathrm{B}}} \max_{\mathrm{p}^{\mathrm{A}} \in \mathcal{P}^{\mathrm{A}}} S(\mathrm{p}^{\mathrm{A}}, \mathrm{p}^{\mathrm{B}}) \stackrel{\mathrm{def}}{=} s_{\min\max}, \quad (12)$$

в которой $s_{\min\max}$ — *максимальная возможность* «проигрыша» игрока B, которую назовем *минимаксной возможностью*; согласно (12) для всех $\mathrm{p}^{\mathrm{A}} \in \mathcal{P}^{\mathrm{A}}$ $S(\mathrm{p}^{\mathrm{A}}, \mathrm{p}^{\mathrm{B}}_*) \leqslant s_{\min\max}$.

Функцию $S_{\mathrm{B}}(\mathrm{p}^{\mathrm{B}})$ можно преобразовать к виду

$$S_{\mathrm{B}}(\mathrm{p}^{\mathrm{B}}) = \max_{1 \leqslant j \leqslant n} \left(\left(\max_{1 \leqslant i \leqslant m} s_{ij} \right) \bullet \mathrm{p}^{\mathrm{B}}_j \right). \quad (13)$$

Минимаксная возможность «проигрыша» игрока B

$$s_{\min\max} = \min_{\mathrm{p}^{\mathrm{B}}} \max_{\mathrm{p}^{\mathrm{A}}} S(\mathrm{p}^{\mathrm{A}}, \mathrm{p}^{\mathrm{B}}) =$$

$$= \min_{1 \leqslant j \leqslant n} \max_{1 \leqslant i \leqslant m} s_{ij} = S_{\mathrm{B}}(\mathrm{p}^{\mathrm{B}}_*). \quad (14)$$

Множество всех минимаксных стратегий $\mathrm{p}^{\mathrm{B}}_*$ в (12) определяется условием

$$\max_{1 \leqslant j \leqslant n} \left(\left(\max_{1 \leqslant i \leqslant m} s_{ij} \right) \bullet \mathrm{p}^{\mathrm{B}}_{*j} \right) = s_{\min\max} \stackrel{\mathrm{def}}{=} s. \quad (15)$$

Условие (15) можно записать в виде

$$\max_{1 \leqslant j \leqslant n} s_j \bullet \mathrm{p}^{\mathrm{B}}_{*j} = \min_{1 \leqslant j \leqslant n} s_j \stackrel{\mathrm{def}}{=} s, \quad (16)$$

где $s_j = \max\limits_{1 \leqslant i \leqslant m} s_{ij}$, $j = 1, \ldots, n$.

Заметим, что среди минимаксных стратегий игрока B всегда найдутся четкие. Точнее, если в (16) $s_{j_1} = \ldots = s_{j_k} = \min\limits_{1 \leqslant j \leqslant n} s_j = s$, то любая из k четких стратегий игрока B

$$\begin{cases} \mathrm{p}^{\mathrm{B}}_{*j_t} = 1, \\ \mathrm{p}^{\mathrm{B}}_{*j} = 0, \quad j \neq j_t, \end{cases} j = 1, \ldots, n, t = 1, \ldots, k,$$

является минимаксной.

Суммируем полученные результаты.

Теорема 1. *В любой одноматричной игре, возможностная модель которой охарактеризована в определении 1, существуют максиминные* $\mathrm{p}^{*\mathrm{A}}$ *и минимаксные* $\mathrm{p}^{\mathrm{B}}_{*}$ *стратегии, причем максиминная возможность выигрыша и минимаксная возможность проигрыша равны*

$$s_{\max\min} = s_{\min\max} \stackrel{\text{def}}{=} s = \min_{1\leqslant j\leqslant n}\max_{1\leqslant i\leqslant m} s_{ij}. \qquad (17)$$

Для любых максиминной $\mathrm{p}^{*\mathrm{A}}$ *и минимаксной* $\mathrm{p}^{\mathrm{B}}_{*}$ *стратегий*

$$S(\mathrm{p}^{*\mathrm{A}}, \mathrm{p}^{\mathrm{B}}_{*}) = s; \qquad (18)$$

тройку $\big(\mathrm{p}^{*\mathrm{A}}, \mathrm{p}^{\mathrm{B}}_{*}, s\big)$ *назовем решением одноматричной игры.*

Четкая стратегия игрока A, в которой $\mathrm{p}^{\mathrm{A}}_i = \begin{cases}1, & i = i_0;\\ 0, & i \neq i_0,\end{cases}$ $i = 1,\ldots,m$, *является максиминной, если* $\min\limits_{1\leqslant j\leqslant n} s_{i_0 j} = s$.

Четкие максиминные стратегии существуют не для любой матрицы переходных возможностей (4).

Четкая стратегия игрока B, в которой $\mathrm{p}^{\mathrm{B}}_j = \begin{cases}1, & j = j_0;\\ 0, & j \neq j_0,\end{cases}$ $j = 1,\ldots,n$, *является минимаксной, если* $\max\limits_{1\leqslant i\leqslant m} s_{ij_0} = s$.

Четкие минимаксные стратегии существуют для любой матрицы переходных возможностей (4).

В первом и втором вариантах теории возможностей формулировка теоремы выглядит одинаково, включая совпадение цен игры. Условия на множества всех максиминных и всех минимаксных стратегий отличаются операцией умножения возможностей (минимум или обычное умножение).

Возможностная модель биматричной игры

Как было отмечено, модель одноматричной игры характеризует ситуацию, в которой один игрок считает событие W «выигрышем», а другой — «проигрышем». Но поскольку в теории возможностей значения $P(W)$ и $P(\Omega\setminus W)$ не зависят друг от друга однозначно, а связаны лишь условием $\max\big(\mathrm{P}(W), \mathrm{P}(\Omega\setminus W)\big) = 1$, то модель одноматричной игры не может охарактеризовать ситуацию, в которой игрок A считает «выигрышем» событие W, а игрок B считает «выигрышем» событие $\Omega\setminus W$, или, наоборот, они оба считают соответствующие события «проигрышем». Для описания таких ситуаций следует ввести матрицы переходных возможностей W и $\Omega\setminus W$.

Определение 2. *Возможностной моделью фазифицированного акта игры назовем классы пространств с возможностями*

$$\big(\Omega_m\times\Omega_n, \mathcal{P}(\Omega_m\times\Omega_n), \mathrm{P}\big),\ \mathrm{P}\in\mathcal{P}_m\times\mathcal{P}_n$$
$$\text{и } \big(\Omega, \mathcal{P}(\Omega), \mathrm{P}_1^{(i,j)}, \mathrm{P}_2^{(i,j)}\big),\ (i,j)\in\Omega_m\times\Omega_n,$$

где $\Omega_m = \{1,\ldots,m\}$, $\Omega_n = \{1,\ldots,n\}$, $\mathcal{P}_m$, $\mathcal{P}_n$ *суть классы всех возможностей* $\mathrm{P}^{\mathrm{A}}(\cdot)\colon \mathcal{P}(\Omega_m)\to[0,1]$, $\mathrm{P}^{\mathrm{B}}(\cdot)\colon \mathcal{P}(\Omega_n)\to[0,1]$, Ω *— пространство «элементарных выигрышей» (или «элементарных проигрышей») игроков,* $\mathrm{P}_1^{(\cdot,\cdot)}(\cdot)$, $\mathrm{P}_2^{(\cdot,\cdot)}(\cdot)$ *— переходные возможности для пространств* $\big(\Omega_m\times\Omega_n, \mathcal{P}(\Omega_m\times\Omega_n)\big)$ *и* $\big(\Omega, \mathcal{P}(\Omega)\big)$, *определяющие* $s_{ij} = \mathrm{P}_1^{(i,j)}(W)$ *— возможность «выигрыша» («проигрыша») $W\in\mathcal{P}(\Omega)$ игрока A, и* $t_{ij} = \mathrm{P}_2^{(i,j)}(\Omega\setminus W)$ *— возможность «выигрыша» («проигрыша») $(\Omega\setminus W)\in\mathcal{P}(\Omega)$ игрока B, если игрок A принял i-ое решение, а игрок B — j-ое,* $i = 1,\ldots,m$, $j = 1,\ldots,n$.

В модели фазифицированного акта игры игроки A и B представлены независимыми нечеткими элементами α и β, принимающими значения в Ω_m и Ω_n, пара (α,β) которых является каноническим нечетким элементом для пространства с возможностью $\big(\Omega_m\times\Omega_n, \mathcal{P}(\Omega_m\times\Omega_n), \mathrm{P}\big)$, *определяющим возможность* $\mathrm{P}(\alpha = i, \beta = j) = (\mathrm{P}^{\mathrm{A}}\times\mathrm{P}^{\mathrm{B}})(\{i,j\}) = \mathrm{P}^{\mathrm{A}}(\{i\})\bullet\mathrm{P}^{\mathrm{B}}(\{j\}) = \mathrm{p}^{\mathrm{A}}_i\bullet\mathrm{p}^{\mathrm{B}}_j$, $i = 1,\ldots,m$, $j = 1,\ldots,n$.

Результатом фазифицированного акта игры назовем возможности

$$\mathrm{P}(W\,|\,\mathrm{p}^{\mathrm{A}}, \mathrm{p}^{\mathrm{B}}) = \max_{\substack{1\leqslant i\leqslant m\\ 1\leqslant j\leqslant n}} \mathrm{P}_1^{(i,j)}(W)\bullet\mathrm{p}^{\mathrm{A}}_i\bullet\mathrm{p}^{\mathrm{B}}_j$$

«выигрыша» («проигрыша») игрока A и

$$\mathrm{P}(\Omega\setminus W\,|\,\mathrm{p}^{\mathrm{A}}, \mathrm{p}^{\mathrm{B}}) = \max_{\substack{1\leqslant i\leqslant m\\ 1\leqslant j\leqslant n}} \mathrm{P}_2^{(i,j)}(\Omega\setminus W)\bullet\mathrm{p}^{\mathrm{A}}_i\bullet\mathrm{p}^{\mathrm{B}}_j$$

«выигрыша» («проигрыша») игрока B, если игрок A использует стратегию $\mathrm{p}^{\mathrm{A}}\in\mathcal{P}^{\mathrm{A}}$, *а игрок B — стратегию* $\mathrm{p}^{\mathrm{B}}\in\mathcal{P}^{\mathrm{B}}$.

Пусть заданы две матрицы переходных возможностей, матричные элементы которых

$$s_{ij} \stackrel{\text{def}}{=} \mathrm{P}(W\,|\,\alpha = i, \beta = j);$$
$$t_{ij} \stackrel{\text{def}}{=} \mathrm{P}(\Omega\setminus W\,|\,\alpha = i, \beta = j);$$
$$\max(s_{ij}, t_{ij}) = 1,\quad i = 1,\ldots,m,\ j = 1,\ldots,n.$$

Соответственно

$$\mathrm{P}(W\,|\,\mathrm{p}^{\mathrm{A}}, \mathrm{p}^{\mathrm{B}}) = \max_{\substack{1\leqslant i\leqslant m\\ 1\leqslant j\leqslant n}} s_{ij}\bullet\mathrm{p}^{\mathrm{A}}_i\bullet\mathrm{p}^{\mathrm{B}}_j \stackrel{\text{def}}{=} S(\mathrm{p}^{\mathrm{A}}, \mathrm{p}^{\mathrm{B}}); \qquad (19)$$

$$\mathrm{P}(\Omega\setminus W\,|\,\mathrm{p}^{\mathrm{A}}, \mathrm{p}^{\mathrm{B}}) = \max_{\substack{1\leqslant i\leqslant m\\ 1\leqslant j\leqslant n}} t_{ij}\bullet\mathrm{p}^{\mathrm{A}}_i\bullet\mathrm{p}^{\mathrm{B}}_j \stackrel{\text{def}}{=} T(\mathrm{p}^{\mathrm{A}}, \mathrm{p}^{\mathrm{B}}). \qquad (20)$$

Для рассматриваемой игры можно поставить две задачи: максимизации и минимизации.

Рассмотрим сначала задачу максимизации, в которой цель игрока A — максимизировать $\mathrm{P}(W \,|\, \mathrm{p}^{\mathrm{A}}, \mathrm{p}^{\mathrm{B}})$ в (19), а цель игрока B — максимизировать $\mathrm{P}(\Omega \setminus W \,|\, \mathrm{p}^{\mathrm{A}}, \mathrm{p}^{\mathrm{B}})$ в (20).

Назовем *точкой равновесия* в такой задаче пару стратегий $(\mathrm{p}^{*\mathrm{A}}, \mathrm{p}^{\mathrm{B}}_{*})$, для которых выполняется условие

$$\forall \mathrm{p}^{\mathrm{A}} \in \mathcal{P}^{\mathrm{A}} \quad S(\mathrm{p}^{\mathrm{A}}, \mathrm{p}^{\mathrm{B}}_{*}) \leqslant S(\mathrm{p}^{*\mathrm{A}}, \mathrm{p}^{\mathrm{B}}_{*});$$
$$\forall \mathrm{p}^{\mathrm{B}} \in \mathcal{P}^{\mathrm{B}} \quad T(\mathrm{p}^{*\mathrm{A}}, \mathrm{p}^{\mathrm{B}}) \leqslant T(\mathrm{p}^{*\mathrm{A}}, \mathrm{p}^{\mathrm{B}}_{*}).$$

Так как $S(\mathrm{p}^{\mathrm{A}}, \mathrm{p}^{\mathrm{B}})$ и $T(\mathrm{p}^{\mathrm{A}}, \mathrm{p}^{\mathrm{B}})$ — функции, неубывающие по каждому $\mathrm{p}^{\mathrm{A}}_i$ и $\mathrm{p}^{\mathrm{B}}_j$, то *пара тривиальных стратегий* $(\mathrm{p}^{\mathrm{A}}_{\mathrm{triv}}, \mathrm{p}^{\mathrm{B}}_{\mathrm{triv}})$ *(из всех единиц) является точкой равновесия, причем*

$$S(\mathrm{p}^{\mathrm{A}}_{\mathrm{triv}}, \mathrm{p}^{\mathrm{B}}_{\mathrm{triv}}) = \max_{\substack{1 \leqslant i \leqslant m \\ 1 \leqslant j \leqslant n}} s_{ij};$$
$$T(\mathrm{p}^{\mathrm{A}}_{\mathrm{triv}}, \mathrm{p}^{\mathrm{B}}_{\mathrm{triv}}) = \max_{\substack{1 \leqslant i \leqslant m \\ 1 \leqslant j \leqslant n}} t_{ij}.$$

Рассмотрим вопрос, существует ли точка равновесия из четких стратегий. Для наглядности будем обозначать i четкую стратегию первого игрока $(0, \ldots, 0, \overset{i}{1}, 0, \ldots, 0)$, а j — четкую стратегию второго игрока $(0, \ldots, 0, \overset{j}{1}, 0, \ldots, 0)$. *Пара* (i^*, j_*) *является точкой равновесия из четких стратегий в задаче максимизации тогда и только тогда, когда выполняется условие* $\max\limits_{1 \leqslant i \leqslant m} s_{ij_*} = s_{i^*j_*}$; $\max\limits_{1 \leqslant j \leqslant n} t_{i^*j} = t_{i^*j_*}$.

Например, для матриц $s = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, $t = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ это условие не выполняется ни для одной из возможных пар четких стратегий. Для матриц $s = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, $t = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$ точек равновесия из четких стратегий две, это пары $(i^* = 1, j_* = 1)$ и $(i^* = 2, j_* = 1)$.

Рассмотрим теперь задачу минимизации, в которой цель игрока A — минимизировать $\mathrm{P}(\Omega \setminus W)$, а цель игрока B — минимизировать $\mathrm{P}(W)$. Назовем *точкой равновесия* в такой задаче пару стратегий $(\mathrm{p}^{*\mathrm{A}}, \mathrm{p}^{\mathrm{B}}_{*})$, для которых выполняется условие

$$\forall \mathrm{p}^{\mathrm{A}} \in \mathcal{P}^{\mathrm{A}} \quad S(\mathrm{p}^{\mathrm{A}}, \mathrm{p}^{\mathrm{B}}_{*}) \geqslant S(\mathrm{p}^{*\mathrm{A}}, \mathrm{p}^{\mathrm{B}}_{*});$$
$$\forall \mathrm{p}^{\mathrm{B}} \in \mathcal{P}^{\mathrm{B}} \quad T(\mathrm{p}^{*\mathrm{A}}, \mathrm{p}^{\mathrm{B}}) \geqslant T(\mathrm{p}^{*\mathrm{A}}, \mathrm{p}^{\mathrm{B}}_{*}).$$

Пара (i^*, j_*) *является точкой равновесия (из четких стратегий) в задаче минимизации тогда и только тогда, когда выполняется условие* $\min\limits_{1 \leqslant j \leqslant n} s_{i^*j} = s_{i^*j_*}$; $\min\limits_{1 \leqslant i \leqslant m} t_{ij_*} = t_{i^*j_*}$.

Если таких пар нет, то точек равновесия в задаче минимизации нет (в том числе не из четких стратегий). Например, так получается для уже приведенной пары матриц $s = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, $t = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$. А для матриц $s = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, $t = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$ точек равновесия из четких стратегий две, это пары $(i^* = 1, j_* = 1)$ и $(i^* = 1, j_* = 2)$.

Суммируем полученные результаты.

Теорема 2. *В любой биматричной игре с задачей максимизации, возможностная модель которой охарактеризована в определении* 2*, существуют точки равновесия. Пара четких стратегий* (i^*, j_*) *есть точка равновесия тогда и только тогда, когда*

$$\max_{1 \leqslant i \leqslant m} s_{ij_*} = s_{i^*j_*}; \quad \max_{1 \leqslant j \leqslant n} t_{i^*j} = t_{i^*j_*}.$$

Точки равновесия из четких стратегий могут как существовать, так и не существовать в зависимости от матриц $\{s_{ij}\}$ *и* $\{t_{ij}\}$ *переходных возможностей.*

В такой же биматричной игре с задачей минимизации точки равновесия могут как существовать, так и не существовать в зависимости от матриц переходных возможностей. Если точки равновесия существуют, то среди них есть и точки равновесия из четких стратегий. Пара четких стратегий (i^*, j_*) *есть точка равновесия тогда и только тогда, когда*

$$\min_{1 \leqslant j \leqslant n} s_{i^*j} = s_{i^*j_*}; \quad \min_{1 \leqslant i \leqslant m} t_{ij_*} = t_{i^*j_*}.$$

Если таких пар нет, то в соответствии с вышесказанным точек равновесия, в том числе из фазифицированных стратегий, в задаче минимизации нет.

В первом и втором вариантах теории возможностей формулировка теоремы выглядит одинаково.

Выводы

Показано принципиальное совпадение результатов рассмотрения матричных игр с точки зрения первого и второго варианта теории возможностей. Доказано, что существуют матрицы, для которых не существует четких оптимальных стратегий.

Литература

[1] *Папилин С. С., Пытьев Ю. П.* Вероятностные и возможностные модели матричных игр двух субъектов // Математическое моделирование. — 2010. — Т. 22, № 12. — С. 144–160.

[2] *Пытьев Ю. П.* Возможность как альтернатива вероятности. Математические и эмпирические основы, применение. — Москва: Физматлит, 2007. — 464 с.

Методы интерпретации экспериментальных данных нечеткой модели измерений, восстановленной по тестам*

Копит Т. А., Чуличков А. И.
kopit_tanya@mail.ru, achulichkov@gmail.com
Москва, МГУ им. М. В. Ломоносова, физический факультет

Рассматривается метод редукции линейных измерений к виду, свойственному измерению параметров объекта идеальным измерительным прибором, в котором математическая модель измерительного прибора, связывающая результат измерения с его входным сигналом, неизвестна и извлекается из результатов тестовых измерений. Погрешность измерений описывается в терминах теории возможностей. Задача редукции измерений ставится как задача на максимум апостериорной возможности.

Широко распространен подход, в котором по тестовым экспериментам строится модель измерительного эксперимента, а затем эта модель используется для решения задач интерпретации данных. При этом на первом этапе стараются построить наиболее точную модель, а на втором — используют ее для получения наилучших оценок искомых параметров. Однако, как показано в [1], требования к точности математической модели, предъявляемые для решения прямой и обратной задач, существенно различаются. Поэтому актуальной является задача построения методов решения анализа и интерпретации данных, в которых тестовые измерения используются для извлечения информации о модели эксперимента так, чтобы решаемая на ее основе задача интерпретации данных обеспечивала максимальную точность.

Возможностная модель измерений

В данной работе рассматривается альтернатива теоретико-вероятностному подходу при анализе и интерпретации измерений, сопровождаемых нечеткостью, основанная на теории возможностей [2]. Фундаментальным понятием этой теории является понятие нечеткого элемента φ линейного пространства $\mathcal{R}$, построенного по аналогии со случайным элементом теории вероятностей. Нечеткий элемент $\varphi \in \mathcal{R}$ определяется распределением возможностей $\pi^{\varphi}(\cdot)\colon R \to [0,1]$, значение $\pi^{\varphi}(f)$ определяет возможность равенства $\varphi = f$, при этом если $\pi^{\varphi}(f) = 0$, то равенство $\varphi = f$ невозможно, а при $\pi^{\varphi}(f) = 1$ равенство $\varphi = f$ вполне возможно. В работе [2] значения возможности используется лишь для определения шансов того или иного события в сравнении с шансами другого, поэтому содержательный смысл имеет лишь высказывания «более возможно», «менее возможно», «равновозможно», в то время как конкретное значение возможности не имеет содержательного смысла. Поэтому все результаты теории возможностей инвариантны к изотонным изменениям шкалы возможностей, оставляющим неподвижными точки 0 и 1 интервала $[0,1]$.

Рассмотрим возможностную модель измерительного эксперимента, в котором на вход измерительного прибора A поступает сигнал f от измеряемого объекта. Измерение его выходного сигнала Af сопровождается аддитивной погрешностью z, и результатом измерения является вектор x. Считается, что сигналы x, f и z являются реализациями нечетких векторов $\xi \in \mathcal{R}_n$, $\varphi \in \mathcal{R}_N$, $\nu \in \mathcal{R}_n$, где $\mathcal{R}_N$ и $\mathcal{R}_n$ — линейные пространства. Моделью измерительного прибора является нечеткий элемент Λ пространства $(\mathcal{R}_N \to \mathcal{R}_n)$ линейных операторов. Его выходной сигнал является нечетким вектором $\Lambda\varphi$. Таким образом, схема измерительного эксперимента запишется в виде

$$\xi = \Lambda\varphi + \nu. \tag{1}$$

В теории измерительно-вычислительных систем [3, 4, 5] рассматриваются задачи интерпретации измерения (1), состоящие в наиболее точном оценивании параметров η изучаемого объекта, непосредственно не наблюдаемых, но связанных с сигналом φ известным соотношением

$$\eta = U\varphi. \tag{2}$$

Эта задача решается путем сведения (редукции) измерения (1) к виду, свойственному измерению сигнала φ с помощью «идеального измерительного прибора» U. Оценка сигнала $U\varphi$ строится на основании математической модели, связывающей результат ξ измерения (1) с состоянием объекта, характеризующегося сигналом φ, а также модели «идеального измерительного прибора» U.

В работе [2] разработаны методы редукции нечетких измерений, в которых модель измерительного прибора задана точно известным линейным оператором $A \in (\mathcal{R}_N \to \mathcal{R}_n)$. В настоящей работе модель измерительного прибора A неизвестна, и информация о ней извлекается из результатов тестовых экспериментов [6, 7] по измерению точно известных входных тестовых сигналов.

Работа выполнена при финансовой поддержке РФФИ, проекты № 11-07-00338-а, 09-01-96508 и 09-07-00505-а

Редукция измерения при известной нечеткой модели измерительного прибора Λ. Получим оценку нечеткого вектора η при условии, что в результате измерения (1) получено значение $\xi = x$, как оценку максимальной апостериорной возможности. Оптимальные свойства оценок максимальной возможности изучены в работе [2].

Лемма 1. *Пусть в (1) $\xi = x$ и задано совместное распределение возможностей значений следующих нечетких элементов: нечеткого вектора $\xi \in \mathcal{R}_n$, нечеткого оператора $\Lambda \in (\mathcal{R}_N \to \mathcal{R}_n)$, нечеткого вектора $\varphi \in \mathcal{R}_N$ и нечеткого вектора $\eta \in \mathcal{R}_M$ параметров исследуемого объекта:*

$$\pi^{\xi,\Lambda,\varphi,\eta}(x, A, f, u), \quad (x, A, f, u) \in \mathcal{R}_n \times (\mathcal{R}_N \to \mathcal{R}_n) \times \mathcal{R}_N \times \mathcal{R}_M; \quad (3)$$

тогда оценка максимальной возможности $\widehat{u}(x)$ нечеткого вектора η определится как решение вариационной задачи

$$\widehat{u}(x) = \sup_{u \in \mathcal{R}_M} \pi^{\xi,\eta}(x, u), \quad x \in \mathcal{R}_n,$$

где

$$\pi^{\xi,\eta}(x, u) = \sup_{\substack{f \in \mathcal{R}_N \\ A \in (\mathcal{R}_N \to \mathcal{R}_n)}} \pi^{\xi,\Lambda,\varphi,\eta}(x, A, f, u); \quad (x, u) \in \mathcal{R}_n \times \mathcal{R}_M. \quad (4)$$

Доказательство. Распределение (1) является маргинальным распределением нечетких векторов ξ и η. При известном ξ формула (1) равна условной возможности равенства $\eta = u$ при условии $\xi = x$, см. [2].

Замечание 1. Маргинальное распределение

$$\pi^{\xi}(x) = \sup_{u \in \mathcal{U}} \pi^{\xi,\eta}(x, u), \quad x \in \mathcal{R}_n, \quad (5)$$

позволяет получить оценку состоятельности модели эксперимента. Действительно, согласно определению, значение $\alpha(x) = \pi^{\xi}(x)$ есть возможность получения результата $\xi = x$, поэтому если, например, $\pi^{\xi}(x) = 0$, то модель (3) следует признать неадекватной.

В работе [2] задача редукции решена для случая, когда модель измерительного прибора точно известна. В этом случае считается, что $\Lambda = A$. Получим решение задачи (5) для ситуации, когда линейный оператор Λ задан своим распределением возможностей.

Теорема 2. *Пусть заданы распределения $\pi^{\varphi}(\cdot)$ и $\pi^{\nu}(\cdot)$ нечетких векторов $\varphi \in \mathcal{R}_N$ и $\nu \in \mathcal{R}_n$ и распределение $\pi^{\Lambda}(\cdot)$ нечеткого линейного оператора Λ, и φ, ν и Λ независимы, $\xi = x$ — результат измерения (1). Тогда оценка $\widehat{u}$ максимальной возможности вектора η равна $\widehat{u} = U\widehat{f}$, где $\widehat{f}$ — решение вариационной задачи*

$$(\widehat{A}, \widehat{f}) = \arg \max_{A,f} \min(\pi^{\nu}(x - Af), \pi^{\varphi}(f), \pi^{\Lambda}(A)). \quad (6)$$

Состоятельность модели измерения определяется априорным распределением возможностей

$$\pi^{\xi}(x) = \max_{A,f} \min\bigl(\pi^{\nu}(x - Af), \pi^{\varphi}(f), \pi^{\Lambda}(A)\bigr).$$

Доказательство. Запишем

$$\pi^{\xi,\varphi,\Lambda}(x, f, A) = \min\bigl(\pi^{\xi|\varphi,\Lambda}(x \,|\, f, A), \pi^{\varphi,\Lambda}(f, A)\bigr),$$

где переходная возможность $\pi^{\xi|\varphi,\Lambda}(x \,|\, f, A) = \pi^{\nu}(x - Af)$, а в силу независимости φ и Λ, $\pi^{\varphi,\Lambda}(f, A) = \pi^{\varphi}(f)\pi^{\Lambda}(A)$. Далее заметим, что распределение переходных возможностей $\pi^{\eta|\xi,\varphi,\Lambda}(u \,|\, x, f, A)$ при фиксированном φ не зависит от ξ, Λ, поэтому

$$\pi^{\xi,\varphi,\Lambda,\eta}(x, f, A, u) = \min\bigl(\pi^{\eta|\varphi}(u \,|\, f), \pi^{\xi,\varphi,\Lambda}(x, f, A)\bigr),$$

и поскольку

$$\pi^{\eta|\varphi}(u \,|\, f) = \begin{cases} 1, & \text{если } u = Uf; \\ 0, & \text{если } u \neq Uf, \end{cases} \quad f \in \mathcal{R}_N \text{ и } u \in \mathcal{R}_M, \quad (7)$$

то

$$\pi^{\xi,\varphi,\Lambda,\eta}(x, f, A, u) = \begin{cases} \min\bigl(\pi^{\nu}(x - Af), \pi^{\varphi}(f), \pi^{\Lambda}(A)\bigr), & \text{если } u = Uf; \\ 0, & \text{если } u \neq Uf. \end{cases} \quad (8)$$

Отсюда, согласно лемме 1, оценка $\widehat{u}$ максимальной возможности вектора η равна $\widehat{u} = U\widehat{f}$, где $\widehat{f}$ — решение задачи на максимин (6). Априорное распределение нечеткого вектора ξ получается из (7) и (8).

Редукция измерения при наличии тестов, уточняющих модель измерительного прибора Λ. Пусть теперь дополнительная информация о модели Λ измерительного прибора может быть извлечена из измерений известных тестовых сигналов $f_1, \ldots, f_m$, проведенных по схеме

$$\xi_j = \Lambda f_j + \nu_j, \quad j = 1, \ldots, m; \quad (9)$$

здесь $\nu_j \in R_n$ — нечеткий элемент, моделирующий погрешность j-го тестового измерения.

Для набора тестовых сигналов $f_1, \ldots, f_m$, набора результатов их регистрации $\xi_1, \ldots, \xi_m$ и набора погрешностей измерений $\nu_1, \ldots, \nu_m$ введем линейный оператор $F \in (\mathcal{R}_m \to \mathcal{R}_N)$ и нечеткие линейные операторы $\Xi \in (\mathcal{R}_m \to \mathcal{R}_n)$,

$N \in (\mathcal{R}_m \to \mathcal{R}_n)$, определенные для любого вектора $t = (t_1, \ldots, t_m) \in \mathcal{R}_m$ равенствами

$$Ft = \sum_{j=1}^{m} \widetilde{f}_j t_j, \quad \Xi t = \sum_{j=1}^{m} x_j t_j, \quad Nt = \sum_{j=1}^{m} \nu_j t_j. \quad (10)$$

Пользуясь обозначениями (10), запишем схему тестовых измерений (1) в виде

$$\Xi = \Lambda F + N. \quad (11)$$

Соотношение (11) позволяет записать переходную возможность $\pi^{\Xi|\Lambda}(\cdot \mid \cdot)$ в виде

$$\pi^{\Xi|\Lambda}(X \mid A) = \pi^{N}(X - AF).$$

Рассмотрим теперь схему (1) измерения нечеткого вектора $\varphi \in \mathcal{R}_n$, результат которого требуется редуцировать к виду, свойственному измерению f по схеме (2). Пусть заданы распределения $\pi^{\varphi}(\cdot)$ и $\pi^{\nu}(\cdot)$ нечетких векторов $\varphi \in \mathcal{R}_N$ и $\nu \in \mathcal{R}_n$. Тогда учитывая, что переходные возможности приобретают вид $\pi^{\xi|\varphi,\Lambda}(x \mid f, A) = \pi^{\nu}(x - Af)$, где $x \in \mathcal{R}_n$ — результат измерения нечеткого элемента ξ в (1), совместное распределение $\pi^{\xi,\varphi,\Lambda,\Xi}(x, f, A, X)$ получим из следующих соотношений:

$$\pi^{\xi,\varphi,\Lambda,\Xi}(x, f, A, X) =$$
$$= \min(\pi^{\xi|\varphi,\Lambda,\Xi}(x \mid f, A, X), \pi^{\varphi,\Lambda,\Xi}(f, A, X)) =$$
$$= \min(\pi^{\nu}(x - Af), \pi^{\Xi|\varphi,\Lambda}(X \mid f, A), \pi^{\varphi,\Lambda}(f, A) =$$
$$= \min(\pi^{\nu}(x - Af), \pi^{N}(X - AF), \pi^{\varphi}(f), \pi^{\Lambda}(A)).$$

Здесь учтено, что распределение ξ при фиксированных Λ и φ не зависит от Ξ, распределение Ξ при фиксированном Λ не зависит от φ.

Далее рассуждая так же, как при доказательстве теоремы 2, получим следующий результат.

Теорема 2*. *Пусть заданы распределения $\pi^{\varphi}(\cdot)$, $\pi^{\nu}(\cdot)$ нечетких векторов $\varphi \in \mathcal{R}_N$ и $\nu \in \mathcal{R}_n$ и распределения $\pi^{\Lambda}(\cdot)$ и $\pi^{N}(\cdot)$ нечетких линейных операторов Λ и N, и φ, ν, Λ и N независимы; $\xi = x$ — результат измерения* (1), *а $\Xi = X$ — результат тестового измерения* (11) . *Тогда оценка $\widehat{u}$ максимальной возможности вектора η равна $\widehat{u} = U\widehat{f}$, где $\widehat{f}$ — решение вариационной задачи*

$$(\widehat{A}, \widehat{f}) = \arg \max_{A,f} \min\Big(\pi^{\nu}(x - Af), \pi^{\mathcal{N}}(X - AF), \pi^{\varphi}(f), \pi^{\Lambda}(A)\Big). \quad (12)$$

Состоятельность модели измерения определяется априорным распределением возможностей

$$\pi^{\xi}(x) = \max_{A,f} \min\Big(\pi^{\nu}(x - Af), \pi^{\mathcal{N}}(X - AF), \pi^{\varphi}(f), \pi^{\Lambda}(A)\Big). \quad (13)$$

Редукция измерений при априорных нечетких ограничениях на координаты сигналов и матричные элементы оператора Λ. Пусть векторы пространств $\mathcal{R}_N$, $\mathcal{R}_n$ и $\mathcal{R}_M$ заданы своими координатами, операторы из $(\mathcal{R}_N \to \mathcal{R}_n)$ — своими матрицами и априорные распределение возможностей нечетких векторов и оператора Λ заданы в виде нечетких ограничений на координаты и матричные элементы следующими соотношениями: для векторов $\nu = (\nu_1, \ldots, \nu_n) \in \mathcal{R}_n$ и $\varphi = (\varphi_1, \ldots, \varphi_N) \in \mathcal{R}_N$ из (1):

$$\pi^{\nu}(x_1, \ldots, x_n) = \mu_0\left(\min_{i=1,\ldots,n} \frac{|x_i|}{\sigma_i}\right) \text{ и} \quad (14)$$

$$\pi^{\varphi}(f_1, \ldots, f_N) = \mu_0\left(\min_{i=1,\ldots,N}\left(\frac{|f_i - f_{0,i}|}{\sigma_i^{(\varphi)}}\right)\right) \quad (15)$$

соответственно, для векторов $\nu_j = (\nu_{j\,i})_{i=1}^{n} \in \mathcal{R}_n$, $j = 1, \ldots, m$ из (9), образующих матрицу $N_{ij} = (\nu_{j\,i})$,

$$\pi^{N}(x_{11}, \ldots, x_{mn}) = \mu_0\left(\min_{\substack{i=1,\ldots,n\\ j=1,\ldots,m}}\left(\frac{|x_{j_i}|}{\sigma_{ij}}\right)\right), \quad (16)$$

для матрицы (Λ_{ij}), $i = 1, \ldots, n$, $j = 1, \ldots, N$, линейного оператора Λ:

$$\pi^{\Lambda}(A_{11}, \ldots, A_{nN}) = \mu_0\left(\min_{\substack{i=1,\ldots,n\\ j=1,\ldots,N}}\left(\frac{|A_{ij} - A_{0,ij}|}{\sigma_{ij}^{(A)}}\right)\right). \quad (17)$$

Здесь $\mu_0(\cdot)\colon [0, \infty) \to [0, 1]$ — строго монотонно убывающая функция, $\mu_0(0) = 1$, $\lim\limits_{z\to\infty} \mu_0(z) = 0$, константы, стоящие в знаменателях формул (14)–(17) — заданные числа, определяющие величину «нечеткости» соответствующих величин, а константы $f_{0,i}$ и $A_{0,ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, N$ определяют наиболее возможные значения вектора $\varphi \in \mathcal{R}_N$ и матрицы оператора Λ.

Тогда задача (12) приводится к виду

$$(\widehat{A}, \widehat{f}) = \arg \inf_{A,f} \max\Bigg(\max_{i=1,\ldots,n}\Big(\tfrac{|x_i - \sum_{k=1}^{N} A_{ik} f_k|}{\sigma_i}\Big), \max_{\substack{s=1,\ldots,n\\ t=1,\ldots,m}}\Big(\tfrac{|\sum_{k=1}^{N} A_{sk} F_{kt} - X_{st}|}{\sigma_{st}}\Big), \max_{q=1,\ldots,N}\Big(\tfrac{|f_q - f_{0,q}|}{\sigma_q^{(\varphi)}}\Big), \max_{\substack{p=1,\ldots,n\\ l=1,\ldots,N}}\Big(\tfrac{|A_{pl} - A_{0,pl}|}{\sigma_{pl}^{(A)}}\Big)\Bigg). \quad (18)$$

Если оператор Λ и входной сигнал f редуцируемого измерения априори произвольны, то $\pi^{\Lambda}(A) = 1$ для любого $A \in (\mathcal{R}_N \to \mathcal{R}_n)$ и $\pi^{\varphi}(f) = 1$ для любого $f \in \mathcal{R}_N$, и задача (12) примет более

простой вид

$$(\widehat{A},\widehat{f}) = \arg\inf_{A,f}\max\left(\max_{i=1,\dots,n}\left(\frac{|x_i-\sum_{k=1}^{N}A_{ik}f_k|}{\sigma_i}\right),\right.$$
$$\left.\max_{\substack{s=1,\dots,n\\ t=1,\dots,m}}\left(\frac{|\sum_{k=1}^{N}A_{sk}F_{kt}-X_{st}|}{\sigma_{st}}\right)\right). \quad (19)$$

Суммируем полученные результаты.

Теорема 3. *Если* $(\widehat{A},\widehat{f})$ *— решение задачи* (18) *(или* (19) *при отсутствии априорных ограничений на возможные значения оператора* Λ *и вектора* f*), то редукция измерения* ξ (1) *к виду* (2)*, равна* $\widehat{u} = U\widehat{f}$*. Возможность согласия модели измерений* (11) *и* (1) *с результатами измерения* $\xi = x$ *и* $\Xi = X$ *равна* $\mu_0(z)$*, где* z *— значение минимакса, полученного при решении задачи* (18) *(или* (19)*).*

Метод вычисления оценки максимальной апостериорной возможности. Для получения решения задачи (18) (или (19) при отсутствии априорных знаний о возможных значениях оператора Λ) на первом этапе для каждого фиксированного вектора $f \in \mathcal{R}_N$ определим значение функции $q(f)$ как значение минимакса, полученного при решении задачи (18) (или (19)) относительно матричных элементов $\widehat{A}_{ij}$ матрицы оператора $\widehat{A}$. Эта минимаксная задача при фиксированном f сводится к задаче линейного программирования [8]. На втором этапе минимизация функции $q(\cdot)$ по $f \in \mathcal{R}_N$ проводится численно. Если минимум функции $q(\cdot)$ по $f \in \mathcal{R}_N$ достигается в точке $f = \widehat{f}$, то искомая оценка равна $\widehat{u} = U\widehat{f}$.

Редукция измерений при априорных нечетких ограничениях на евклидовы нормы сигналов и оператора Λ**.** Пусть пространства $\mathcal{R}_N$, $\mathcal{R}_n$ и $\mathcal{R}_M$ евклидовы, пространство линейных операторов $(\mathcal{R}_m \to \mathcal{R}_n)$ — евклидово со скалярным произведением

$$(A,B)_2 = \operatorname{tr} AB^* = \sum_{i=1}^{m}(Ae_i, Be_i),$$

$A, B \in (\mathcal{R}_m \to \mathcal{R}_n)$, где $\{e_i\}$ — любой ортонормированный базис евклидова пространства $\mathcal{R}_m$. (Если операторы $A, B \in (\mathcal{R}_m \to \mathcal{R}_n)$ заданы в некоторых ортонормированных базисах своими матрицами A_{ij}, B_{ij}, $i = 1,\dots,n$, $j = 1,\dots,m$, то их скалярное произведение $(A,B)_2 = \sum_{i,j} A_{ij}B_{ij}$, а квадрат нормы $\|B\|_2^2 = \sum_{i,j} B_{ij}^2$). Априорные распределение возможностей нечеткого вектора ν и оператора N заданы в виде нечетких ограничений на их нормы следующими соотношениями:

$$\pi^{\nu}(z) = \mu_0(\|z\|^2), \quad z \in \mathcal{R}_n,$$
$$\pi^{N}(Z) = \mu_0(\|Z\|_2^2), \quad Z \in (\mathcal{R}_m \to \mathcal{R}_n).$$

Здесь $\mu_0(\cdot)\colon [0,\infty) \to [0,1]$ — строго монотонно убывающая функция, $\mu_0(0) = 1$, $\lim\limits_{z\to\infty}\mu_0(z) = 0$. Линейный оператор $\Lambda \in (\mathcal{R}_N \to R_n)$ и входной сигнал φ редуцируемого измерения априори произвольны, так, что $\pi^{\Lambda}(A) = 1$ для любого $A \in (\mathcal{R}_N \to \mathcal{R}_n)$ и $\pi^{\varphi}(f) = 1$ для любого $f \in \mathcal{R}_N$.

Тогда задача (6) приводится к следующей задаче на минимакс:

$$\min_{A,f}\max\Big(\|x - Af\|^2, \|X - AF\|_2^2\Big). \quad (20)$$

В зависимости от того, какое из минимальных значений, $J_1\big(\widehat{A}_0, \widehat{f}(\widehat{A}_0)\big)$ или $J_2(\widehat{A}_0)$, меньше, выбираются различные методы решения задачи (20).

Выводы

В работе получен метод эмпирического восстановления нечеткой модели измерений, в которой математическая модель измерительного прибора, связывающая результат измерения с его входным сигналом, неизвестна и информация о ней извлекается из результатов тестовых измерений. Погрешность измерений описывается в терминах теории возможностей. Поставлена и решена задача редукции как задача на максимум апостериорной возможности. Рассмотрены примеры задач редукции в случае априорной информации об ограничениях на сигналы: для нескольких конкретных моделей распределений нечетких элементов φ, Λ, N и ν.

Литература

[1] *Пытьев Ю. П.* Надежность интерпретации эксперимента, основанной на приближенной модели // Мат.моделирование — 1989. — Т.1, № 2. — С. 49–64.

[2] *Пытьев Ю. П.* Возможность как альтернатива вероятности. — М.: Физматлит, 2007. — 464 с.

[3] *Пытьев Ю. П.* Методы математического моделирования измерительно-вычислительных систем. — М.: Физматлит, 2004. — 400 с.

[4] *Пытьев Ю. П.* Математические методы интерпретации эксперимента: Учеб. пособие для вузов. — М.: Высш. шк., 1989. — 352 с.

[5] *Чуличков А. И.* Основы теории измерительно-вычислительных систем сверхвысокого разрешения. — Тамбов: Изд-во Тамбовского гос. тех. ун-та., 2000. — 140 с.

[6] *Черемухин Е. А., Чуличков А. И.* О редукции к идеальному прибору по данным тестирующих измерений // Вестник Московского ун-та. Сер. 3. Физика. Астрон. — 2004. — № 3. — С. 15–18.

[7] *Голубцов П. В., Пытьев Ю. П., Чуличков А. И.* Построение оператора редукции по тестовым измерениям // Сб. «Дискретные системы обработки сигналов», Устинов, 1986. — С. 68–72.

[8] *Кириллов К. В., Чуличков А. И.* Редукция измерений в нечеткой модели эксперимента как решение задачи линейного программирования // Вестник Московского ун-та. Сер. 3. Физика. Астрон. — 1999. — Т.2. — С. 65–67.

Распознавание семантических и топологических свойств конфигураций пространств знаний*

Костенко К. И.
kostenko@kubsu.ru
г. Краснодар, Кубанский государственный университет

Рассматриваются вычислимые свойства абстрактных объектов, составляющих структурно-семантические представления целостных многообразий знаний, обеспечивающие возможность их сравнения и оценки различия. Свойства полезны для построения унифицированной содержательно полной системы классов операций моделирующих процессы обработки знаний.

Абстрактные пространства знаний — это математический формализм понятия пространства знаний. Его целью является проведение фундаментального исследования, связанного с созданием технологии построения и использования целостных систем знаний в различных областях деятельности. Теоретическая и практическая состоятельность данного формализма зависит от возможности уточнения и использования инвариантов современной математики в его основных структурных и функциональных конструктах. Абстрактное пространство знаний составляют модели нескольких типов, близкие к алгебраическим системам. Многообразие таких моделей составляет категорию, объектами которой являются классы однотипных систем, связанные вычислимыми операциями гомоморфного вложения и интеграции моделей [1, 2].

Структурное представление абстрактных знаний

Основным компонентом всякого пространства знаний являются пространства конфигураций. Пусть $\mathbf{M}$ — бесконечное вычислимое конфигураций (мгновенных представлений отдельных абстрактных знаний), содержащее пустую конфигурацию Λ, а $\mathbf{R}$ — вычислимое множество разрешимых бинарных отношений на $\mathbf{M}$, для которого разрешимо отношение вложения отношений ρ_1, включающее пустое отношение E, выполняющееся для любых пар конфигураций $z_1, z_2 \in \mathbf{M}$. Разложением конфигураций называется всюду определенное вычислимое отображение $\varepsilon : \mathbf{M} \mapsto \mathbf{M} \times \mathbf{M}$, для которого:

$$\varepsilon(\Lambda) = (\Lambda, \Lambda)$$

и

$$\forall\, z_1, z_2 \in \mathbf{M}\, \exists\, z \in \mathbf{M}(\varepsilon(z) = (z_1, z_2)).$$

Семантическим связыванием для разложения ε называется такое вычислимое отображение

$$\psi : \mathbf{M} \times \mathbf{M} \mapsto \mathbf{R},$$

что:

1. $\forall\, z_1, z_2 \in \mathbf{M}\, \exists\, z \in \mathbf{M}$

$$((\varepsilon(z) = (z_1, z_2)) \wedge (z_1 \neq \Lambda \vee z_2 \neq \Lambda) \rightarrow \rightarrow \psi(z) = (z_1, z_2));$$

2. $\forall\, z \in \mathbf{M}(\psi(z) \neq E \rightarrow \varepsilon(z) \in \psi(z))$;
3. ψ инъективно на множествах

$$\left\{ z \;\middle|\; z \in \mathbf{M} \wedge \varepsilon(z) = (z_1, z_2) \wedge (z_1 \neq \Lambda \vee z_2 \neq \Lambda) \right\}.$$

Пространством конфигураций называется пара $\mathbf{M} = (\mathbf{M}, d)$, где $\mathbf{M}$ — множество конфигураций, содержащее пустую конфигурацию, а $d = (\varepsilon, \psi)$ — каноническая декомпозиция, составленная отображениями разложения и связывания. Конфигурация $z \in \mathbf{M}$ называется элементарной, если $\varepsilon(z) = (\Lambda, \Lambda)$. Множество элементарных конфигураций упорядочено отношением ρ_0. Полные структурные представления (ПСП) конфигураций порождаются декомпозициями и имеют вид нагруженных бинарных деревьев, висячим вершинам которых сопоставляются элементарные конфигурации, а внутренние вершины размечены значениями отношений между конфигурациями, представляемыми деревьями, корнями которых являются непосредственные потомки таких вершин. Множество вершин деревьев ПСП конфигураций (обозначается как $\mathbf{I}$) составляют двоичные наборы с пустым набором, обозначаемым как λ. Множества вершин (висячих вершин) ПСП $z \in \mathbf{M}$ обозначается как $\mathbf{D}(z)(\mathbf{O}(z))$. Если $\alpha \in \mathbf{D}(z)$, то $[z]_\alpha$ — это разметка вершины α.

Моделирование процессов обработки знаний в абстрактных пространствах знаний позволяет обозначить систему классов отображений (морфизмов), исследование которых существенно для построения обоснованной технологии цифровых пространств знаний. Морфизмами указанных классов обеспечивается возможность исследования фундаментальных свойств пространств знаний, позволяющая получать описания программно реализуемых операций:

— декомпозиции и связывания представлений знаний, реализующих процессы создания регулярных многообразий знаний для отдельных предметных областей;

Работа выполнена при финансовой поддержке РФФИ, проект №11-07-96508-р_юг_ц.

— формирования топологической структуры пространства знаний, сравнения содержания отдельных знаний, оценки их различия и внутренней сложности, обеспечивающую формализацию схем обобщения и унификации знаний;
— распознавания достоверности и полноты систем знаний, отображаемых пространствами знаний;
— эквивалентных преобразований и трансформаций представлений знаний, связанных с решениям задач интеграции, дополнения и уменьшения избыточности содержания таких знаний.

Гомоморфные вложения и трассирования. Унифицированные эффективные сравнения на $\mathbf{M}$ связаны с возможностью такого сопоставления вершин ПСП конфигураций, при котором разметки соответствующих вершин оказываются сравнимыми.

Определение 1. *Конфигурация z_1 гомоморфно трассируется в конфигурацию $z_1 \leqslant z_2$, если существует такое монотонное относительно нестрого вложения двоичных наборов отображение ξ : $\mathbf{D}(z_1) \to \mathbf{D}(z_2)$, что*

1. $(\xi(\mathbf{D}(z_1)) \subseteq \xi(\mathbf{D}(z_2))) \wedge \forall \alpha \in \mathbf{D}(z_1)$
$(\alpha \in \mathbf{D}(z_1) \backslash \mathbf{O}(z_1) \leftrightarrow \xi(\alpha) \in \mathbf{D}(z_2) \backslash \mathbf{O}(z_2))$;
2. $\forall \alpha \in \mathbf{O}(z_1)([z_1]_\alpha \rho_0 [z_2]_{\xi(\alpha)})$;
3. $\forall \alpha \in \mathbf{D}(z_1) \backslash \mathbf{O}(z_1)([z_1]_\alpha \rho_1 [z_2]_{\xi(\alpha)}))$.

Частными случаями трассирований являются отображения, для которых первое условие приведённого определения заменено на

1. $\xi(\mathbf{D}(z_1)) \subseteq \xi(\mathbf{D}(z_2)) \wedge \forall \alpha \in \mathbf{D}(z_1)$
$(\alpha \in \mathbf{D}(z_1) \setminus \mathbf{O}(z_1) \leftrightarrow \xi(\alpha) \in \mathbf{D}(z_2) \setminus \mathbf{O}(z_2))$;
2. $\forall\, \alpha, \alpha\sigma \in \mathbf{D}(z_1), \sigma \in \{0,1\}\ \exists\, \beta, \gamma \in \mathbf{I}$
$((\xi(\alpha) \subset \xi(\alpha\sigma) \to \xi(\alpha\sigma) = \xi(\alpha)\beta\sigma\gamma)$.

Отображения трассирования конфигураций с дополнительными условиями на β и γ определяют классы o-трассирований (β — пустое) c — трассирований (β и γ — пустые) . Инъективные трассирования называются p-трассированиями. Будем называть последние два вида трассирований сжатиями и растяжениями.

Если $z \in \mathbf{M}$, то выражение $\Delta(z)$ обозначает множество конфигураций, получаемых из z с помощью опрации инвертирования [1].

Определение 2. *Конфигурация z_1 гомоморфно вкладывается в конфигурацию z_2 ($z_1 \subseteq z_2$), если существуют такие конфигурации $z^1 \in \Delta(z_1)$ и $z^2 \in \in \Delta(z_2)$, для которых $z^1 \leqslant z^2$.*

Различные виды вложения конфигураций порождают отношения на $\mathbf{M}$, обозначаемые как $\subseteq_o$, $\subseteq_c$ и $\subseteq_p$. Данные отношения рефлексивные и, в общем случае, не являются антисимметричными. Кроме того, $\subseteq_c$ и $\subseteq_p$ — транзитивны, а $\subseteq_o$ — может оказаться нетранзитивным.

Гомоморфные трассирования (вложения) обобщают трассирования (вложения) разных типов, рассматривавшиеся в [1].

Теорема 1. *Отношение гомоморфного вложения (трассирования) транзитивно на множестве конфигураций.*

Последнее позволяет рассматривать гомоморфные вложения в качестве унифицированного универсального уточнения понятия вложения для семантических структур знаний, вместо o-трассирований конфигураций, которые оказываются нетранзитивными и поэтому не всегда удобны при изучении свойств конфигураций.

Метрики и сходимость множеств конфигураций. Теоретический и прикладной интерес представляют топологические свойства пространств знаний, составляющие основу исследования и практического применения понятий расстояния (сходства), а также обобщения (сходимости) вычислимых множеств знаний. Включение указанных понятий в систему основных характеристик пространств знаний обеспечивает полноту унифицированного абстрактного моделирования существующих, преимущественно когнитивных и слабоформализованных, мер сложности и различия представлений знаний [2, 4].

Определим операции над конфигурациями, называемые изменением разметки, добавления и удаления вершин ПСП конфигураций из $\mathbf{M}$. Данные операции не замкнуты на $\mathbf{M}$, но для любых двух конфигураций $z_1, z_2 \in \mathbf{M}$ существует последовательность применения указанных операций, преобразующая z_1 в z_2. Функционал $p : \mathbf{M} \times \mathbf{M} \to \mathbf{M}$, определяемый соотношением $\forall\, z_1, z_2 \in \mathbf{M}(p(z_1, z_2)$ равно минимуму числа применений операций изменения разметки, добавления и удаления вершин для преобразования z_1 в z_2) является метрикой на $\mathbf{M}$.

Теорема 2. *Существуют пространства конфигураций, для которых метрика p является невычислимой.*

Особенностью p является слабая связь с трассированиями (вложениями), не позволяющая измерять различие конфигураций инвариантно относительно структуры и разметок вершин их ПСП. Этот недостаток преодолевается, если измерять расстояние (различие) пар конфигураций $z_1, z_2 \in \in M$, с помощью не являющегося метрикой функционала $P(z_1, z_2) = |\Gamma(z_1, z_2)| + |\Gamma(z_2, z_1)|$, где $\Gamma(z_1, z_2) = \mathbf{D}(z_1) \backslash \{\alpha | \alpha \in \mathbf{D}(z_1) \wedge \mathbf{D}(z_1) \wedge (z_1)_\alpha \subseteq z_2\}$.

Определение 3. *Вычислимое множество конфигураций $\Omega = \{z_i | z_i \in \mathbf{M} \wedge i \in \mathbf{N}\}$ сходится снизу (сверху) к конфигурации z, если*

$$\forall z_i \in \Omega(z_i \subseteq z) \wedge (\forall z^1 \in \mathbf{M}(\forall z_i \in \Omega(z_i \subseteq z) \rightarrow \rightarrow z \subseteq z^1);$$

$$\forall z_i \in \Omega(z \subseteq z_i) \wedge \forall z^1 \in \mathbf{M}(\forall z_i \in \Omega(z \subseteq z_i) \rightarrow \rightarrow z^1 \subseteq z)).$$

Для указанных видов сходимости будем называть z супремумом (инфинумом) множества Ω.

Теорема 3. *Если вычислимые множества Ω_1 и Ω_2 сходятся снизу, то множество $\Omega_1 \cup \Omega_2$ также сходится снизу.*

Перенос целостных систем знаний в пространства знаний. Формализация процессов построения пространств знаний связана с вычислимыми преобразованиями слабоформализованных и слабоструктурированных многообразий представлений знаний предметных областей (первичных интеллектуальных ресурсов) во множество конфигураций абстрактного пространства знаний. В общем случае такое преобразование осуществляется с помощью вычислимых отображений, моделирующих процессы извлечения знаний, основанных на распознавании семантической структуры ресурсов и называемых свободными декомпозициями. Для этого в унифицированной модели цифрового пространства знаний применяется специальная система атрибутов структурных представлений знаний, включающая классификаторы, роли и фильтры представлений знаний и их фрагментов [2]. В абстрактных пространствах знаний перечисленные атрибуты моделируются с помощью специальных классов вычислимых морфизмов конфигураций [5]. Автоматизация процессов извлечения знаний из первичных ресурсов достигается включением в структуру соответствующих морфизмов отображений, моделирующих элементарные процессы анализа контента. Такими отображениями моделируются схемы распознавания представлений элементарных знаний и определения их свойств, специфических для различных предметных областей.

Свободной декомпозицией называется пара $\mathbf{d} = (\varepsilon_d, \psi_d)$, где $\varepsilon_d : \mathbf{M} \rightarrow \mathbf{M} \times \mathbf{M}$ и $\psi_d : \mathbf{M} \rightarrow \mathbf{R}$ — вычислимые отображения, для которых выполняются условия:

1. $\forall z \in \mathbf{M}(\varepsilon_d(z) \in \psi_d(z))$;
2. $\forall z \in \mathbf{M}(\varepsilon_d(z) = (z_\mathbf{1}, z_\mathbf{2}) \rightarrow z_\mathbf{1} \oplus z_\mathbf{2} \subseteq z)$.

Здесь $\oplus$ — это операция прямой суммы конфигураций, представляющая простейшую схему интеграции знаний [1]. Свободные декомпозиции порождают структурные представления конфигураций, аналогичные ПСП, которые в общем случае могут не образовывать ПСП конфигураций, поскольку для них не требуются выполнимость соотношения $\varepsilon(z) \in \psi(z)$, где ε и ψ — отображение из канонической декомпозиции конфигураций. Многообразие свободных декомпозии реализует общий класс процессов извлечения и структуризации знаний, различающихся степенью полноты и детальности представления отдельных знаний, отражающих субъективные представления экспертов, а также учитывающих особенности предполагаемого применения пространства знаний. Построение развитой системы морфизмов декомпозиции может быть реализовано с использованием специальных сравнений таких морфизмов и операций их суммирования и умножесния декомпозиций.

Выводы

В работе изучаются теоретических аспекты процессов построения и использования систем формализованных знаний с использованием разнообразных математических инструментов. Реализуемый подход к моделированию связан с использованием алгебраических и алгоритмических конструктов при моделировании пространств знаний. Он основан на рассмотрении специальных множеств структурированных объектов и систем операций над ними, что отличает его от подходов к моделированию многообразий знаний в формализмах дескрипционных логик. Для формализма пространств знаний доказано существование невычислимых метрик на множествах конфигураций и возможность моделирования понятия расстояний с помощью разрешимых отношений вложения и трассирования конфигураций. Практическая применимость сравнений конфигураций, основанных на понятии трассирования, основывается на квадратичной сложности в сравнениях разметок конфигураций для таких важных видов трассирований, как растяжение и сжатие.

Литература

[1] *Костенко К. И.* Компоненты и операции абстрактных пространств знаний // Материалы Всероссийской конференции ЗОНТ09. — Новосибирск: Институт математики им. С. Л. Соболева, 2009. — Т. 2. — С. 36–40.

[2] *Костенко К. И., Левицкий Б. Е.* Унифицированные модели и автоматизированные технологии цифровых пространств знаний // Материалы конференции Новые информационные технологии и менеджмент качества. — ФГУ ГНИИ ИТТ Информика, 2011. — С. 118–120.

[3] *Гаврилова Т. А.* Субъективные метрики оценки онтологий // Материалы Всероссийской конференции ЗОНТ09. — Новосибирск: Институт математики им. С. Л. Соболева, 2009. — С. 178–186.

[4] *Анисимов А. В., Марченко А. А.* Система обработки текстов на естественном языке // Искусственный интеллект. — 2002. — № 4. — С. 157– 163.

[5] *Костенко К. И.* Классификация операций в пространствах знаний // XII национальная конференция по искусственному интеллекту с международным участием (труды конференции). — Тверь: ТГУ, 2010. — Т. 2. — С. 155–163.

Сходимость эмпирических случайных процессов, порождаемых процедурами обучения*

Хачай М. Ю.
mkhachay@imm.uran.ru
Екатеринбург, Институт математики и механики УрО РАН

Обосновывается равномерная (по классу функций) сходимость эмпирических средних к математическим ожиданиям и донскеровость соответствующих эмпирических случайных процессов при условии атомичности вероятностной меры

Решение многих задач теории статистического обучения: распознавания образов, восстановления регрессии и пр. — связано с обоснованием сходимости соответствующих случайных процессов, порождаемых обучающей выборкой. Подобные процессы принято [1] называть *эмпирическими*.

Возможно, самым ранним и в наибольшей степени исследованным является процесс, возникающий при оценке функции распределения $F\colon \mathbb{R} \to [0,1]$ «обыкновенной» случайной величины ξ по случайной независимой выборке

$$(t_1, \ldots, t_l) \tag{1}$$

при помощи эмпирической функции распределения

$$\mathbb{F}_l(t) = \frac{1}{l} \sum_{i=1}^{l} 1_{(-\infty,t)}(t_i),$$

где

$$1_{(-\infty,t)}(\tau) = \begin{cases} 1, & \text{если } \tau \in (-\infty, t); \\ 0, & \text{в противном случае.} \end{cases}$$

Из усиленного закона больших чисел следует, что при произвольном $t \in \mathbb{R}$

$$\mathbb{F}_l(t) \xrightarrow{\text{п.н.}} F(t).$$

Однако существенно более важными как для развития соответствующей теории, так и для статистических приложений явились два фундаментальных результата теории вероятностей: теорема Гливенко–Кантелли, обосновывающая наличие равномерной по t сходимости значений эмпирических функций распределения

$$\sup_{t\in\mathbb{R}} |\mathbb{F}_l(t) - F(t)| \xrightarrow{\text{п.н.}} 0,$$

и центральная предельная теорема (ЦПТ), утверждающая, что при каждом $t \in \mathbb{R}$ случайная величина

$$\mathbb{G}_l(t) = \sqrt{l}\big(\mathbb{F}_l(t) - F(t)\big) \tag{2}$$

слабо сходится к нормально распределенной случайной величине $\mathbb{G}(t)$ с нулевым средним и дисперсией $F(t)\big(1 - F(t)\big)$.

Существенным обобщением ЦПТ является теорема Донскера (см., напр. [2]), в которой аналогичный результат переносится на последовательности случайных процессов. Выборке (1) (в соответствии с соотношением (2)) сопоставляется случайный процесс $\mathbb{G}_l$, траектории которого являются элементами пространства ограниченных функций $l^\infty(\mathbb{R})$. В теореме обосновывается слабая сходимость $\mathbb{G}_l \rightsquigarrow \mathbb{G}$ к гауссовскому случайному процессу $\mathbb{G}$ с параметрами $E(\mathbb{G}(t)) = 0$ и

$$\operatorname{cov}(\mathbb{G}(s), \mathbb{G}(t)) = F(\min\{s,t\}) - F(s)F(t),$$

представимому в виде суперпозиции $\mathbb{G}(t) = \mathbb{B}\big(F(t)\big)$ стандартного броуновского моста $\mathbb{B}$ и восстанавливаемой функции распределения. Последнее разложение, ввиду изученности процесса $\mathbb{B}$, позволяет легко строить доверительные интервалы для $F(t)$ по эмпирическим данным.

Аналогичный подход может быть применен к построению эмпирических процессов более общей природы. Пусть заданы вероятностное пространство $(\mathcal{X}, \mathcal{A}, P)$ и семейство измеримых ограниченных функций $\mathcal{F} \subset [\mathcal{X} \to \mathbb{R}]$. Функции $f \in \mathcal{F}$ сопоставим математическое ожидание Ef. Задавшись случайной независимой выборкой $(X_1, \ldots, X_l) \in \mathcal{X}^l$ и определяемой ей эмпирической мерой

$$P_l = \frac{1}{l} \sum_{i=1}^{l} \delta_{X_i},$$

сопоставим элементам семейства $\mathcal{F}$ эмпирические средние

$$\mathbb{E}_l f = \int_{\mathcal{X}} f(x)\, dP_l(x) \equiv \frac{1}{l} \sum_{i=1}^{l} f(X_i),$$

задав, тем самым, случайные процессы

$$\{\mathbb{E}_l f : f \in \mathcal{F}\}; \tag{3}$$

$$\{\mathbb{G}_l f \equiv \sqrt{l}(\mathbb{E}_l f - Ef) : f \in \mathcal{F}\}. \tag{4}$$

В частности, если $f = 1_A$ — индикаторная функция события $A \in \mathcal{A}$, то значения Ef и $\mathbb{E}_l f$ совпадут, соответственно, с вероятностью $P(A)$ и частотой $\nu(A)$ этого события, вычисленной по заданной выборке.

Работа выполнена при финансовой поддержке Президиума УрО РАН, проекты № проекты 09-П-1-1001 и 09-С-1-1010, и РФФИ, проекты № 10-01-00273 и 10-07-00134

Соответствующим образом доопределив понятия равномерной и слабой сходимости, вводятся определения классов функций, удовлетворяющих свойству Гливенко–Кантелли и Донскера для заданной вероятностной меры P. Среди известных результатов следует отметить критерий В. Н. Вапника и А. Я. Червоненкиса [3] равномерной сходимости частот к вероятностям, результаты А. А. Боровкова [4] и Ван дер Ворта [2] о конечной аппроксимируемости классов событий и функций.

В статье исследуется частный случай вероятностного пространства, в котором вероятностная мера является атомической. Показано, что в этом случае произвольный не более чем счетный класс равномерно ограниченных измеримых функций $\mathcal{F}$ удовлетворяет условию Гливенко-Кантелли. В качестве следствия получено условие равномерной сходимости частот к вероятностям для произвольного не более чем счетного класса событий. Приведено условие на вероятностную меру, при котором произвольный не более чем счетный класс событий удовлетворяет условию Донскера.

Сходимость случайных процессов

Пусть фиксированы вероятностное пространство (Ω, Σ, Q) и множество $\mathcal{F}$.

Определение 1. *Случайным процессом, индексируемым множеством $\mathcal{F}$, называется семейство случайных величин $\{\xi_f : f \in \mathcal{F}\}$*

Удобно представлять себе случайный процесс как отображение $\Xi\colon \Omega \to \mathbb{D}$, принимающее значения в подходящем метрическом пространстве $(\mathbb{D}, d)$ функционалов, определенных на множестве $\mathcal{F}$. Каждому $\omega \in \Omega$ процесс Ξ сопоставляет случайную функцию $\Xi[\omega]$, называемую *траекторией процесса*. Всюду ниже в качестве $\mathbb{D}$ будет использоваться нормированное пространство ограниченных функционалов $l^\infty(\mathcal{F})$ с нормой $\|h\| = \sup\limits_{f\in\mathcal{F}} |h(f)|$.

Определение 2. *Случайный процесс Ξ называется эмпирическим, если*

1. $(\Omega, \Sigma, Q) = (\mathcal{X}^l, \mathcal{A}^l, P^l)$ *— пространство выборок длины l из базового пространства $(\mathcal{X}, \mathcal{A}, P)$,*
2. $\mathcal{F}$ *— множество измеримых функций, определенных на $\mathcal{X}$.*

Всюду ниже ограничимся рассмотрением процессов, индексируемых подмножествами $\mathcal{F} \subset l^\infty(\mathcal{X})$, и определяемых формулами (3)–(4). Введем определение свойств Гливенко–Кантелли и Донскера, которыми может обладать семейство функций $\mathcal{F}$, формулируемых в терминах сходимости этих процессов.

Определение 3. *Семейство $\mathcal{F}$ обладает свойством Гливенко–Кантелли относительно вероятностной меры P (Г-К(P)), если*

$$\sup_{f\in\mathcal{F}} |\mathbb{E}_l f - Ef| \xrightarrow{\text{п.н.}^*} 0, \tag{5}$$

т. е. если найдется последовательность случайных величин Δ_l такая, что

$$d(\mathbb{E}_l, E) \equiv \sup_{f\in\mathcal{F}} |\mathbb{E}_l f - Ef| \leqslant \Delta_l$$

и $P(\limsup\limits_{l\to\infty} \Delta_l = 0) = 1$.

Понятие «внешней сходимости почти наверное», используемое в определении, совпадает с привычным понятием «сходимости почти наверное» в случае измеримости величины $d(\mathbb{E}_l, E)$.

Свойством Г-К(P) для произвольной вероятностной меры P, как известно [3], обладает произвольный класс $\mathcal{F}$ конечной емкости ($\mathrm{VCD}(\mathcal{F}) < \infty$) равномерно ограниченных функций. Если $\mathcal{F}$ — семейство индикаторных функций событий, элементов $\mathcal{B} \subset \mathcal{A}$, то семейство $\mathcal{F}$ обладает свойством Г-К(P) тогда и только тогда, когда имеет место равномерная по классу событий $\mathcal{B}$ сходимость частот к вероятностям.

Вслед за монографией [2] введем понятие *интервального накрывающего числа*, естественным образом обобщающее понятие конечной аппроксимируемости класса событий. *Интервалом (порядковым, замкнутым)* $[l, u]$ называется [5] подмножество функций

$$\{f \in l^\infty(\mathcal{X}) :\ l(x) \leqslant f(x) \leqslant u(x),\ x \in \mathcal{X}\}.$$

Определение 4. *Для заданного $\varepsilon > 0$ ε-накрывающим числом $N(\varepsilon, \mathcal{F})$ семейства $\mathcal{F}$ называется мощность минимального конечного накрытия семейства $\mathcal{F}$ интервалами $[l_j, u_j]$, обладающими свойством*

$$\int_{\mathcal{X}} \big(u_j(x) - l_j(x)\big)\, dP(x) < \varepsilon.$$

Если любое накрытие бесконечно, то $N(\varepsilon, \mathcal{F}) = \infty$.

Справедлива

Теорема 1 ([2]). *Пусть для произвольного $\varepsilon > 0$ ε-накрывающее число $N(\varepsilon, \mathcal{F}) < \infty$. Тогда семейство $\mathcal{F}$ обладает свойством Гливенко–Кантелли для вероятностной меры P.*

Определение 5. *Семейство функций $\mathcal{F}$ называется Донскеровским, если последовательность $\mathbb{G}_l \rightsquigarrow \mathbb{G}$ слабо сходится к плотному гауссовскому [1] процессу $\mathbb{G}$.*

Естественным образом определение донскеровости переносится и на классы событий $\mathcal{B} \subset \mathcal{A}$.

Теорема 2 ([2]). *Пусть $\mathcal{B}$ — класс событий, обладающий свойством*

$$J(\mathcal{B}) = \int_{\varepsilon=0}^{\infty} \sqrt{\log_2 N(\sqrt{\varepsilon}, \mathcal{B})}\, d\varepsilon < \infty.$$

Класс $\mathcal{B}$ — Донскеровский.

Результаты

Все приведенные ниже утверждения (за исключением теор. 3 и 5) получены для вероятностного пространства $(\mathcal{X}, \mathcal{A}, P)$ с *атомической* мерой.

Определение 6. *Элемент $A \in \mathcal{A}$ называется атомом меры P, если $P(A) > 0$ и*

$$(\forall B \in \mathcal{A},\ B \subset A)\ \ (P(B) = P(A)) \vee (P(B) = 0).$$

Пусть $\mathcal{A}' \subset \mathcal{A}$ — семейство (всех) попарно не эквивалентных атомов. Очевидно, $\mathcal{A}'$ не более чем счетно.

Определение 7. *Мера P называется атомической, если $P(\mathcal{X} \setminus \bigcup_{\mathcal{A}'} A) = 0$.*

Элементарным примером вероятностного пространства с атомической мерой является дискретное пространство, мера в котором является смесью дираковских мер, сконцентрированных в одноточечных подмножествах (синглетонах) множества $\mathcal{X}$. Тем не менее, существуют и не дискретные пространства с атомической мерой.

Пример 1 [6]. Пусть, $\mathcal{X} = [0, 1]$, σ-алгебра $\mathcal{A}$ порождена синглетонами $\{x\}$, $x \in [0, 1]$. Тем самым, измеримыми являются только не более чем счетные подмножества единичного отрезка и их дополнения. Мера P задается равенством:

$$P(A) = \begin{cases} 0, & \text{если } A \text{ не более чем счетно,} \\ 1, & \text{в противном случае.} \end{cases}$$

Очевидно, заданная таким образом мера является атомической, $\mathcal{X} \in \mathcal{A}$ является ее единственным атомом (с точностью до эквивалентности).

В общем случае справедлива теорема.

Теорема 3 ([7]). *Пусть $(\mathcal{A}, \rho_p)$ — псевдометрическое пространство с полуметрикой $\rho_p(A, B) = P(B \setminus A) + P(A \setminus B)$. Следующие условия эквивалентны:*

1) *пространство $(\mathcal{A}, \rho_p)$ компактно;*
2) *пространство $(\mathcal{A}, \rho_p)$ σ-компактно;*
3) *мера P является атомической.*

Теорема 4. *Пусть $(\mathcal{X}, \mathcal{A}, P)$ — вероятностное пространство с атомической мерой, и $\mathcal{F} \subset l^\infty(\mathcal{X})$ — не более чем счетное семейство равномерно ограниченных измеримых функций. Семейство $\mathcal{F}$ обладает свойством Гливенко–Кантелли относительно меры P.*

Фактически, доказательство теоремы сводится к обоснованию конечности (для произвольного $\varepsilon > 0$) накрывающего числа $N(\varepsilon, \mathcal{F})$ и последующему применению результата теоремы 1.

Замечание 1. Если в условии теоремы 4 потребовать дополнительно эквивалентность каждого атома меры P некоторому синглетону $\{x\} \subset \mathcal{X}$, утверждение теоремы будет справедливо для произвольного (не обязательно счетного) семейства измеримых равномерно ограниченных функций $\mathcal{F}$.

Результат теоремы легко переносится на случай семейств индикаторных функций: $f(x) = 1_B(x)$, для $B \in \mathcal{B} \subset \mathcal{A}$.

Следствие 1. *Пусть $(\mathcal{X}, \mathcal{A}, P)$ — вероятностное пространство с атомической мерой, $\mathcal{B} \subset \mathcal{A}$ — произвольный не более чем счетный класс событий. Справедливо условие:*

$$\sup_{B \in \mathcal{B}} |\nu(B) - P(B)| \xrightarrow{\text{п.н.}} 0.$$

Замечание 2. Условие атомичности меры, в силу справедливости теоремы 3, эквивалентно условию σ-компактности пространства $(\mathcal{A}, \rho_p)$ и, очевидно, может быть ослаблено. Для наличия равномерной сходимости частот к вероятностям по классу $\mathcal{B}$ достаточно σ-компактности пространства $(\mathcal{A}(\mathcal{B}), \rho_{p'})$, где $\mathcal{A}(\mathcal{B}) \subset \mathcal{A}$ — σ-алгебра, индуцированная классом $\mathcal{B}$, а P' — сужение меры P на $\mathcal{A}(\mathcal{B})$.

Замечание 3. Ограничение не более чем счетности, накладываемое на мощность класса $\mathcal{B}$, наоборот, в общем случае не ослабляемо. В самом деле, рассмотрим вероятностное пространство, описанное в примере 1, и класс событий

$$\mathcal{B} = \{B \in \mathcal{A} : |B| < \infty\}.$$

По построению, $P(B) \equiv 0,\ (B \in \mathcal{B})$. С другой стороны, нетрудно убедиться, что для произвольной выборки найдется событие $B_0 \in \mathcal{B}$, для которого $\nu(B_0) = 1$.

Теорема 5. *Пусть в пространстве $(\mathcal{X}, \mathcal{A}, P)$ мера P не является атомической. Тогда найдется такой не более чем счетный класс $\mathcal{B} \subset \mathcal{A}$, компактный относительно полуметрики ρ_p, что для произвольной конечной выборки*

$$\sup_{B \in \mathcal{B}} |\nu(B) - P(B)| = 1.$$

Пример 2 Рассмотрим вероятностное пространство $(\mathcal{X}, \mathcal{A}, P)$, в котором $\mathcal{X} = [0, 1]$, $\mathcal{A}$ — борелевская σ-алгебра на отрезке $[0, 1]$, и P — мера Лебега. Мера P, очевидно, безатомная.

Построим последовательность множеств B_n (Рис. 1), обладающую свойством $\rho_p(B_n, \varnothing) \to 0$

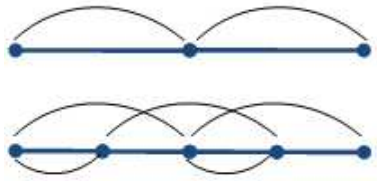


Рис. 1. Построение последовательности B_n

такую, что для произвольной конечной выборки $\varkappa = (X_1, \ldots, X_l)$ найдется элемент $B_n = B_{n(\xi)}$ сколь угодно малой меры, для которого $\nu(B_n) = 1$. Построение удобно провести по индукции. Для произвольного натурального k рассмотрим регулярное разбиение отрезка $[0, 1]$ с диаметром $\delta_k = 2^{-k}$ и добавим к строящейся последовательности всевозможные объединения из k элементов разбиения. Видно, что для каждого события B_p, добавленного на k-м шаге

$$P(B_p) = k2^{-k},$$

следовательно, $P(B_n) \to 0$. С другой стороны, очевидно, для произвольного N найдется $k \geqslant N$ и объединение k элементов регулярного разбиения с диаметром 2^{-k}, содержащее все элементы выборки $\varkappa$, то есть обладающее частотой, равной единице. Следовательно, построено такое компактное не более чем счетное подмножество

$$\mathcal{B} = \{B_n : n \in \mathbb{N}\} \cup \{\varnothing\} \subset \mathcal{A},$$

что

$$\sup_{B \in \mathcal{B}} |\nu(B) - P(B)| = 1.$$

Теорема 6. *Пусть $(\mathcal{X}, \mathcal{A}, P)$ вероятностное пространство с атомической мерой P, множество (всех) попарно неэквивалентных атомов $\mathcal{A}'$ упорядочено по убыванию меры P. Условие*

$$\sum_{N \geqslant 1} \left(\sum_{i > N} P(A_i) \right)^2 < \infty \qquad (6)$$

влечет донскеровость произвольного не более чем счетного класса $\mathcal{B} \subset \mathcal{A}$.

Замечание 4. Нетрудно привести пример последовательности, удовлетворяющей условию (6). Условию удовлетворяет, например, последовательность $\{P(A_i) = 2^{-i}\}$.

Замечание 5. Аналогично замечанию 1, отметим, что одноточечность (с точностью до эквивалентности) атомов меры P влечет справедливость утверждения теоремы 6 для произвольного (не обязательно счетного) класса $\mathcal{B} \subset \mathcal{A}$.

Выводы

В работе показано, что условие атомичности меры достаточно для того, чтобы произвольный не более чем счетный класс событий обладал свойствами равномерной сходимости частот к вероятностям и Донскера. Первый результат удалось распространить на семейства равномерно ограниченных измеримых функций. По-видимому, аналогичным образом может быть получено доказательство и второго факта. Открытым также остается вопрос построения доверительных оценок математических ожиданий для донскеровских классов функций.

Литература

[1] *Булинский А. А., Ширяев А. Н.* Теория случайных процессов. — Москва: Физматлит, 2005. — 408 с.

[2] *Van der Vaart A., Wellner J.* Weak convergence and empirical processes. — New York: Springer, 1996. — 268 p.

[3] *Vapnik V. N.* Statistical learning theory. — New York: Wiley, 1998. — 740 p.

[4] *Боровков А. А.* Математическая статистика. — Новосибирск: Наука; Изд-во Института математики СО РАН, 1997. — 772 с.

[5] *Биркгоф Г.* Теория решеток. — Москва: Наука, 1984. — 568 с.

[6] *Ченцов А. Г.* Элементы конечно аддитивной теории меры. I — Екатеринбург: УГТУ-УПИ. — 389 с.

[7] *Пыткеев Е. Г, Хачай М. Ю.* Сигма-компактность метрических булевых алгебр и равномерная сходимость частот к вероятностям // Труды Института математики и механики. — 2010. — Т. 16, № 1. — С. 127–139.

Эмпирические доверительные интервалы для условного риска в задаче классификации*

Неделько В. М.
nedelko@math.nsc.ru
Новосибирск, Институт математики СО РАН

Целью работы является исследование возможности улучшения оценок риска в задаче классификации за счёт построения оценок, дифференцированных по различным областям признакового пространства. Для получения дифференцированных оценок риска предлагается метод построения эмпирических доверительных интервалов. Приводятся примеры, демонстрирующие достоинства метода.

Как правило, оценки риска (вероятности ошибочной классификации) строятся для решающей функции в целом, независимо от значений переменных, описывающих объекты исследования. Вместе с тем, очевидна разумность идеи, заключающейся в построении оценок, дифференцированных по подобластям признакового пространства. Например, при построении дерева решений качество классификации обучающей выборки в разных конечных вершинах может очень сильно отличаться. Также при вычислении, например, оценки риска методом скользящего экзамена мы можем также подсчитать долю ошибочно классифицированных объектов для каждой конечной вершины в отдельности. Если полученные значения существенно различаются, то также напрашивается давать разные оценки риска для разных областей.

Для оценивания точности таких дифференцированных оценок риска будет использоваться метод эмпирических доверительных интервалов [3, 7].

Постановка задачи

Пусть X — пространство значений переменных, используемых для прогноза, а $Y = \{0, 1\}$ — пространство значений прогнозируемых переменных, и пусть C — множество всех вероятностных мер на заданной σ-алгебре подмножеств множества $D = X \times Y$. При каждом $c \in C$ имеем вероятностное пространство $\langle D, B, \mathsf{P}_c\rangle$, где B — σ-алгебра, P_c — вероятностная мера.

Решающей функцией называется соответствие $\lambda\colon X \to Y$.

Качество принятого решения оценивается заданной функцией потерь $\mathcal{L}\colon Y^2 \to [0, \infty)$. Под риском будем понимать средние потери:

$$R(c, \lambda) = \mathsf{E}^c\mathcal{L}\big(y, \lambda(x)\big) = \int_D \mathcal{L}\big(y, \lambda(x)\big)\ \mathsf{P}_c(dx, dy),$$

$x \in X$, $y \in Y$.

При $\mathcal{L}(y, y') = \mathcal{L}_I(y, y')$, где $\mathcal{L}_I(y, y') = \begin{cases} 0, & y{=}y'; \\ 1, & y{\neq}y', \end{cases}$ — индикаторная функция потерь, риск есть вероятность ошибочной классификации.

Работа выполнена при финансовой поддержке РФФИ, проекты № 10-01-00113-а и № 11-07-00346-а.

В общем случае функция потерь может зависеть от x. Почти все дальнейшие рассуждения при этом останутся без изменений.

Заметим, что значение риска зависит от c — распределения, которое неизвестно. Поэтому возникает задача оценивания риска по выборке.

Пусть $\nu = \{(x^i, y^i) \in D \,|\, i = 1, \ldots, N\}$, $\nu \in D^N$ — случайная независимая выборка из распределения P_c. Вероятностную меру на σ-алгебре пространства выборок D^N будем обозначать также P_c, поскольку это не приводит к неоднозначности.

Алгоритм (метод) построения решающих функций есть отображение $Q\colon \{\nu\} \to \Lambda$, где Λ — заданный класс решающих функций, а $\lambda_{Q,\nu}$ — функция, построенная по выборке ν методом Q.

Для риска можно строить как точечные, так и интервальные оценки. Мы будем использовать последние, поскольку в них по построению заложена оценка точности (степени достоверности).

Наиболее просто построить доверительный интервал для оценки риска по контрольной выборке:

$$R^*(\nu^*, \lambda) = \frac{1}{N^*}\sum_{i=1}^{N^*}\mathcal{L}\big(y_*^i, \lambda(x_*^i)\big),$$

где $\nu^* = \{(x_*^i, y_*^i) \in D \,|\, i = 1, \ldots, N^*\}$, $\nu^* \in D^{N^*}$ — случайная независимая выборка из распределения P_c. В этом случае доверительный интервал для риска есть классический односторонний доверительный интервал для параметра Биномиального распределения. Такой интервал не зависит от метода классификации Q.

Доверительные интервалы, строящиеся по обучающей выборке ν, зависят от Q. Доверительный интервал для R будем задавать [3] в виде $[0, \hat{R}(\nu)]$. При этом для всех $c \in C$ должно выполняться условие:

$$\mathsf{P}_c\big(R(c, \lambda_{Q,\nu}) \leqslant \hat{R}(\nu)\big) \geqslant \eta, \qquad (1)$$

где η — заданная доверительная вероятность. Здесь мы ограничиваемся односторонними оценками, поскольку на практике для риска важны именно оценки сверху. Таким образом, в данном случае построение доверительного интервала эквивалентно выбору функции $\hat{R}(\nu)$, которую будем называть оценочной функцией или просто оценкой (риска).

Обычно доверительные интервалы конструируются не как функция непосредственно выборки, а как функция от некоторой статистики. Для оценки риска в роли такой статистики естественно использовать эмпирические функционалы качества, то есть точечные оценки риска, такие как эмпирический риск, оценка скользящего экзамена, оценка bootstrap и т. п.

Эмпирический риск определяется как средние потери на выборке:

$$\widetilde{R}(\nu,\lambda)=\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}\big(y^i,\lambda(x^i)\big).$$

Функционал скользящего экзамена определяется как:

$$\breve{R}(\nu,Q)=\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}\big(y^i,f_{Q,\nu_i'}(x^i)\big),$$

где $\nu_i'=\nu\backslash\{(x^i,y^i)\}$ — выборка, получаемая из ν удалением i-го наблюдения,

Аналогичным образом можно строить доверительные интервалы и для среднего риска $R_Q(c)=\mathsf{E}^c_{D^N}R(c,\lambda_{Q,\nu})$, где $\mathsf{E}^c_{D^N}$ — усреднение по обучающим выборкам.

Величину $R_Q(c)$ обычно удаётся оценивать с большей точностью [2], хотя на практике, как правило, интересует именно $R(c,\lambda)$.

Варианты задания риска

Пусть выбрано некоторое разбиение $\mathcal{A}=\{A_1,...,A_k\}$ пространства X. Тогда области $D_i=A_i\times Y,\ i=1,...,k$, образуют разбиение пространства D. Предположим, что $\mathsf{P}_c(D_i)>0$. Условным риском будем называть величину

$$R_i(c,\lambda)=\mathsf{E}^c_{D_i}\mathcal{L}\big(y,\lambda(x)\big)=\\=\frac{1}{\mathsf{P}_c(D_i)}\int_{D_i}\mathcal{L}\big(y,\lambda(x)\big)\ \mathsf{P}_c(dx,dy).$$

Доверительным интервалом для условного риска будем называть такой интервал вида $[0,\hat{R}_i(\nu)]$, что для всех $c\in C$ выполняется условие

$$\mathsf{P}_c\big(R_i(c,\lambda_{Q,\nu})\leqslant\hat{R}_i(\nu)\big)\geqslant\eta,$$

Приведённое определение также имеет смысл, когда разбиение $\mathcal{A}$ зависит от Q и λ, которая в свою очередь зависит от ν.

Следует заметить, что различение оценок риска для разных областей пространства X позволяет рассчитывать на получение более адекватных оценок. Однако с ростом k объём выборки, «приходящийся» на каждую область, становится меньше, что делает оценку менее достоверной.

Желательно иметь способ учитывать различие в качестве классификации на разных подобластях, не увеличивая число оцениваемых величин, чтобы не уменьшать статистическую устойчивость оценок.

Один из таких способов вполне очевиден и применим, когда в зависимости от оцениваемого значения риска есть возможность предпринять какие-то действия по уменьшению потерь от ошибочной классификации.

Рассмотрим следующую модель управления потерями

$$\mathcal{L}_s(y,y')=\mathcal{L}(y,y')\cdot\big(1-\mathcal{S}(s)\big)+s,$$

где s — «страховая премия», т. е. расходы, которые мы понесли, чтобы уменьшить последствия ошибочной классификации, а $\mathcal{S}(s)$ — «страховое возмещение», а именно, доля «компенсируемых» потерь.

Терминология из области страхования здесь использована лишь для наглядности. Представленная модель применима во всех ситуациях, когда мы имеем возможность добровольно понести некоторые расходы, которые не «оправдаются» в случае правильной классификации, но существенно уменьшат потери в случае ошибочной классификации.

Если страховая премия составляет фиксированную ставку β, то $\mathcal{S}(s)=\left\{\begin{smallmatrix}s/\beta,\ s<\beta;\\ 1,\quad s\geqslant\beta\end{smallmatrix}\right.$.

Если мы знаем $R=\mathsf{E}\mathcal{L}$, то для минимизации $R_s=\mathsf{E}\mathcal{L}_s$ следует положить $s=\left\{\begin{smallmatrix}0,\ R<\beta;\\ \beta,\ R\geqslant\beta\end{smallmatrix}\right.$.

Такое решение можно интерпретировать следующим образом. Величина β — это потери, которые мы несём в случае отказа от классификации данного объекта. Если ожидаемые потери от ошибочной классификации маньше β, то мы принимаем решение классифицировать объект, иначе — отказываемся от классификации.

На практике вероятность ошибочной классификации неизвестна, поэтому решение об отказе от классификации может приниматься на основе некоторой оценки риска. В простейшем случае, когда используется оценка по контрольной выборке, функция коррекции потерь выглядит как $s=\left\{\begin{smallmatrix}0,\ R^*N^*<\mu;\\ \beta,\ R^*N^*\geqslant\mu\end{smallmatrix}\right.$, где μ — пороговое значение числа ошибок, которое может выбираться от 0 до N^*+1.

На рисунке 1 показаны зависимости ожидаемых потерь R_s от вероятности ошибочной классификации R при различных значениях μ порога отказа от решения для $\beta=0{,}2,\ N^*=20$. Номер кривой соответствует порогу μ.

Наиболее привлекательной выглядит кривая под номером 4, которая соответствует выбору $\mu=\beta N^*$, т. е. использованию вместо вероятности ошибки частоты ошибок на контроле.

Приведённый пример, несмотря на элементарность, интересен тем, что демонстрирует типичную для проблемы статистических решений ситу-

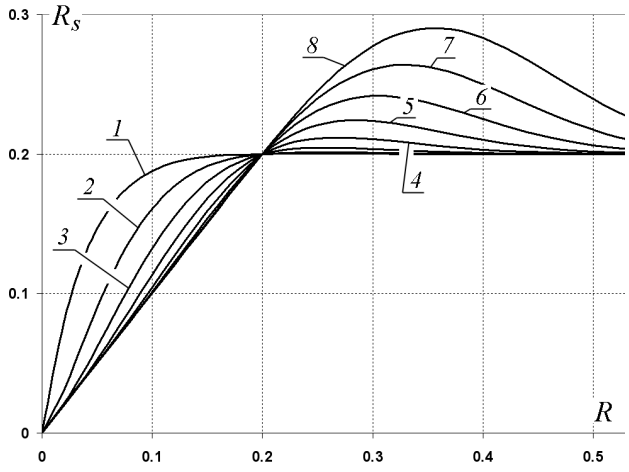


Рис. 1. Зависимости ожидаемых потерь от вероятности ошибочной классификации при различных порогах отказа от решения.

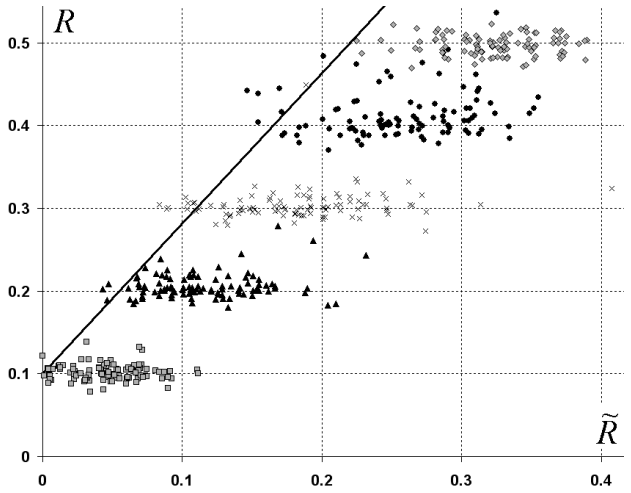


Рис. 2. Эмпирический доверительный интервал для риска классификации решающими деревьями.

ацию, когда различные стратегии (распознавателя) — в данном случае в роли такой стратегии выступает выбор μ — оказываются недоминируемыми, но при этом существует стратегия, «наилучшая» с субъективной точки зрения.

Численное моделирование

Идея построения эмпирических доверительных интервалов заключается в подборе некоторой функции $\hat{R}(\nu)$, такой чтобы свойство (1) выполнялось на всех распределениях из некоторого эмпирически выбранного набора. Множество распределений по возможности «представительным», т. е. таким, чтобы выполнение свойства (1) на этих распределениях позволяло бы практически рассчитывать на его выполнения для всех распределений, которые могут иметь место в реальных задачах.

Рассмотрим следующее семейство распределений.

Пусть $X = [0,1]^n$ — n-мерный гиперкуб, на котором задано равномерное распределение. Чтобы полностью определить вероятностную меру осталось задать $g(x) = \mathsf{P}(y = 1 \,|\, x)$ — условную вероятность первого класса при попадании в точку x. Будем задавать $g(x)$ в виде

$$g(x) = \begin{cases} g_1, & x_j < \theta^{\frac{1}{n}},\ j = 1, ..., n; \\ g_2, & \text{иначе.} \end{cases}$$

Иными словами, $g(x)$ является кусочно-постоянной, первая область постоянства есть гиперкуб объёма θ, вторая — дополнение внутреннего гиперкуба до единичного. Семейство распределений зададим следующим образом: положим $g_1 = 0{,}5$, $g_2 = 1$, а параметр параметр θ изменяется от 0 до 1.

Очевидно, что $\theta = 2R_0$, где R_0 — байесовский уровень ошибки.

Заметим, что данное семейство задано в некотором смысле «по подобию» распределений, доставляющих максимум смещения эмпирического риска для гистограммного классификатора [7].

Для иллюстрации приведём результаты моделирования для построения классификатора в виде дерева решений. Дерево строится методом направленного поиска (последовательного ветвления) при заданном числе конечных вершин M. Для примера выбраны значения параметров: $n = 2$, $N = 50$, $M = 3$, $\eta = 0{,}9$. Для R_0 использовались значения 0,1, 0,2, 0,3, 0,4, 0,5. При каждом R_0 генерировалось по 100 обучающих выборок, по каждой из которых строилось дерево решений, для которого вычислялись риск и эмпирический риск.

На рисунке 2 приведены результаты моделирования для случая индикаторной функции потерь. Разными видами маркеров отмечены точки $(R, \widetilde{R})$ для различных R_0. Сплошной линией показана функция $\hat{R}(\widetilde{R})$, подобранная так, чтобы для каждого R_0 выше этой кривой находилось не более 10 точек (порог получается как $100 \cdot (1 - \eta)$).

Заметим, что значительная часть точек на графике лежит ниже соответствующего Байесовского уровня ошибки. Это объясняется тем, что вероятность ошибки для построенного решающего правила вычислялась не точно, а по контрольной выборке (объёма 2000). Внесённая таким образом погрешность, однако, не является существенной в данном контексте.

Аналогичное моделирование проведено для условного риска на выбранной области. Каждый раз выбиралась область, соответствующая конечной вершине с максимальной разностью числа правильно и неправильно классифицированных объектов обучающей выборки. Результаты приведены на рисунке 3.

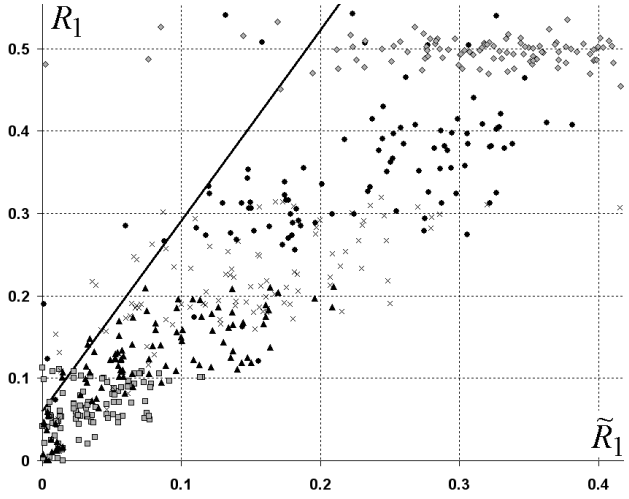


Рис. 3. Эмпирический доверительный интервал для условного риска.

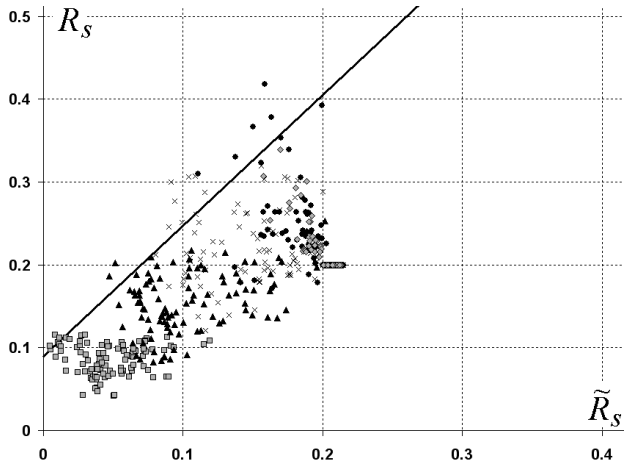


Рис. 4. Эмпирический доверительный интервал для риска при возможности отказа от классификации.

Как можно заметить, получаемые значения в среднем меньше, но имеют больший разброс, как и следовало ожидать.

Вместо условного эмпирического риска можно было использовать условную оценку скользящего экзамена: долю ошибочно классифицированных на скользящем контроле объектов из заданной области.

Следующее моделирование проведено для случая, когда есть возможность отказываться от классификации. Цена отказа $\beta = 0{,}2$. Отказ от классификации проводился для конечных вершин, в которых $\frac{N_j^{\neq}}{N_J} > \beta$, где N_j — число объектов обучающей выборки в j-й вершине, $N_J^{\neq}$ — из них неправильно классифицировано. Результаты приведены на рисунке 4.

Чтобы сравнить полученные результаты, вычислим средние значения оценочных функций для всех трёх случаев. Результаты приведены в таблице 1. Видим, что дифференцированное вычисление риска позволяет получать лучшие оценки риска, по сравнению с обычным подходом.

Таблица 1. Средние значения оценочных функций.

R_0	$\mathsf{E}\hat{R}$	$\mathsf{E}\hat{R}_1$	$\mathsf{E}\hat{R}_s$
0,1	0,17	0,12	0,15
0,2	0,30	0,22	0,24
0,3	0,42	0,38	0,32

Выводы

Полученные результаты могут быть полезны в двух аспектах. С одной стороны, как качественная оценка потенциальных возможностей, которые заложены в диференцированном оценивании риска в комбинации с использованием эмпирических доверительных интервалов. С другой стороны, приведённая схема моделирования может использоваться как образец использования метода для оценивания риска в реальных задачах.

Несмотря на большое число открытых вопросов, например, касающихся выбора семейств распределений, предложенный метод даже в существующем виде позволяет получать дополнительную по отношению к стандартным методам информацию, полезную для оценивания качества решающих функций в реальных задачах.

Литература

[1] *Лбов Г. С., Старцева Н. Г.* Логические решающие функции и вопросы статистической устойчивости решений. — Институт математики СО РАН, Новосибирск, 1999. — 211 с.

[2] *Efron B., Tibshirani R.* Improvements on Cross-Validation: The .632+ Bootstrap Method // Journal of the American Statistical Association. — 1997. — V. 92, N. 438. — Pp. 548–560.

[3] *Неделько В. М.* Об интервальном оценивании риска для решающей функции // Таврический вестник информатики и математики. Изд-во НАН Украины. — 2008. — № 2. — С. 97–103.

[4] *Vorontsov K. V.* Combinatorial probability and the tightness of generalization bounds // Pattern Recognition and Image Analysis. — 2008. — V. 18, N. 2. — Pp. 243–259.

[5] *Langford J.* Quantitatively tight sample complexity bounds. Carnegie Mellon Thesis. — 2002. — `http://citeseer.ist.psu.edu/langford02quantitatively.html`. — 130 p.

[6] *Неделько В. М.* О точности интервальных оценок вероятности ошибочной классификации, основанных на эмпирическом риске // Всеросс. конф. ММРО-14, М.: МАКС Пресс, 2009. — С. 56–59.

[7] *Неделько В. М.* Точные и эмпирические оценки вероятности ошибочной классификации // Научный вестник НГТУ. — 2011. — № 1 (42). — С. 3–16.

Системы достоверных эмпирических закономерностей в моделях оптимальных разбиений и методы их анализа*

Сенько О. В., Кузнецова А. В.
senkoov@mail.ru
Москва, Вычислительный центр им. А. А. Дородницына РАН, Институт биохимической физики им. Н. М. Эмануэля РАН

Представлен метод анализа данных, основанный на поиске систем статистически достоверных закономерностей, связывающих прогнозируемую величину с пространством объясняющих переменных, и исследовании различий между выявленными закономерностями. При этом под закономерностью понимается такое разбиение пространства объясняющих переменных, для которого существуют статистически достоверные отклонения между значениями прогнозируемой величины в областях разбиений. Поиск оптимальных разбиений ведётся в рамках нескольких моделей различных типов и различного уровня сложности. Для верификации закономерностей используются варианты перестановочного теста, заключающиеся в многократном повторении поиска на случайных выборках, полученных из исходной выборки данных с помощью перестановок значений прогнозируемой величины относительно фиксированных позиций объясняющих переменных. Рассматриваются методы анализа систем закономерностей в пространстве прогнозов, вычисляемых предикторами, которые строятся по закономерностям.

Ведение

Одним из способов исследования зависимости некоторой переменной Y от набора потенциальных объясняющих переменных $X_1, \ldots, X_n$ по выборкам данных является поиск областей пространства объясняющих переменных, в которых средние значения прогнозируемой переменной Y значимо отличаются от значений переменных во всей выборке или в соседних областях пространства. Предполагается, что обучающая выборка данных $\tilde{S}$ имеет вид: $(y_1, \mathbf{x}_1), \ldots, (y_m, \mathbf{x}_m)$, где y_j — значения прогнозируемой переменной Y, $\mathbf{x}_j$ — вектор объясняющих переменных X для j-го объекта. Для решения подобных задач может быть использован целый ряд подходов, включая методы поиска логических закономерностей [1, 2], модели регрессионных или классифицирующих деревьев. Одним из возможных способов поиска является метод Оптимальных достоверных разбиений (ОДР), основанный на построении оптимальных разбиений пространства объясняющих переменных в рамках нескольких моделей различного уровня сложности. При этом предусматриваются модели с заданной геометрической формой, включая одномерные модели с одной или двумя граничными точками, двумерные модели с границами, параллельными координатным осям, а также с границами, произвольно ориентированными относительно координатных осей. В том случае, если прогнозируемая величина Y является бинарной индикаторной функцией некоторого класса K, могут использоватьсчя также модели, в которых элементы разбиений, задаются как окрестности наборов опорных точек. В работе [5] поиказано, что задача поиска таких наборов может быть сведена к задаче поиска компонент связности специального графа смежности, заданного на объектах $\tilde{S} \cap K$ или $\tilde{S} \setminus K$. Метод оптимальных достоверных разбиений успешно использовался для решения ряда задач анализа биомедицинских данных в онкологии, психиатрии, неврологии, геронтологии, включая задачу оценки влияния разнообразных клинических, биохимических, инструментальных факторов на тяжесть течения дисциркуляторной энцефалопатии [6, 7, 8].

Построение и верификация закономерностей в методе ОДР

Среди всевозможных разбиений модели выбирается разбиение с максимальным значением специального функционала качества $F_q(R_{\text{opt}}, \tilde{S})$, возрастающего при увеличении степени разделения векторов $\mathbf{x}$. Для статистической верификации эмпирической закономерности, связанной с оптимальным разбиением R_{opt}, используется технология перестановочных тестов, основанная на сравнении качества оптимальных разбиений, полученных на исходной реальной обучающей выборке, с качеством разбиений, полученных на выборках, которые генерируются из исходной выборки с помощью датчиков случайных чисел. Первоначальный и простейший вариант перестановочного теста проверяет нулевую гипотезу о полной независимости Y от объясняющих переменных. В первоначальном варианте поиск оптимальных разбиений производится для набора выборок $\{\tilde{S}^r\}$, полученных из исходной выборки $\tilde{S}$ путём случайных перестановок величин $y_1, \ldots, y_m$ относительно фиксированных позиций векторов $\mathbf{x}_1, \ldots, \mathbf{x}_m$ с помощью полностью той же самой процедуры, что была использована для поиска оптимальных разбиений на реальной выборке. Для оценки статистической достоверности используются p-значения, определяемые как доли случайных выбо-

Работа выполнена при финансовой поддержке РФФИ, проект № 11-07-0715.

рок из множества $\{\tilde{S}^r\}$, для которых максимальное значение фукционала качества F_q превышает максимальное значение $F_q(R_{\text{opt}}, \tilde{S})$. Разбиение считается верифицированной закономерностью, если соответствующее p-значение не превышает некоторого заранее заданного уровня значимости. Для того, чтобы иметь возможность сравнить статистическую достоверность двух закономерностей, для которых p-значения равны 0, используется функционал P_q, равный отношению $F_q(R_{\text{opt}}, \tilde{S})$ к максимальному значению F_q, достигнутому на наборе $\{\tilde{S}^r\}$. Первоначальный вариант перестановочного теста может успешно использоваться при анализе с помощью наиболее простых моделей разбиений. Однако для случаев более сложных моделей первоначальный вариант не позволяет установить, является ли повышение сложноости оправданным, или для объяснения тенденций, существующих в данных, достаточно простых моделей. Кроме того первоначальный вариант не позволяет оценить вклад различных составляющих сложной закономерностей в описываемую зависмость. Так, при использовании двумерных моделей, данный тест не позволяет оценить вклад каждой из двух объясняющей переменной. Для решения упомянутых задач был подход, использующий методологический принцип «лезвие Оккама». Подход основан на модифицированной версии перестановочного теста, проверяющей возможность статистического объяснения различий, выявленных с помощью более сложных моделей, в рамках закономерностей, найденных ранее с использованием более простых моделей для тех же самых переменных. Проверяется нулевая гипотеза о независимости Y от объясняющих переменных внутри областей разбиений, соответствующих простым закономерностям. Предположим, что для пары переменных (X_i, X_j) в рамках двумерной модели найдено оптимальное разбиение R^2_{opt}. Однако ранее для переменной X_i в рамках простейшей одномерной модели было найдена статистически достоверная закономерность, соответствующая разбиению R^{1i}_{opt}. Тогда статистическая верификация закономерности, связанной с R^2_{opt} относительно простейшей одномерной закономерности для X_i, производится путём сравнения величины $F_q(R_{\text{opt}}, \tilde{S})$ с оптимальными значениями F_q, достигнутыми для выборок из множества $\tilde{S}^r_j$. Выборки из $\tilde{S}^r_i$ также формируются с помощью случайных перестановок величин $y_1, \ldots, y_m$ относительно фиксированных позиций векторов $\mathbf{x}_1, \ldots, \mathbf{x}_m$. Однако допустимыми считаются только перестановки внутри подмножеств объектов $\tilde{S}$ с x-векторами из одних и тех же подобластей пространства объясняющих переменных, принадлежащих разбиению R^{1i}_{opt}. Таким образом модифицированным вариантом перестановочного теста проверяется более мягкий вариант нулевой гипотеза о независимости Y от объясняющих переменных внутри подобластей из R^{1i}_{opt}. Для оценки статистической значимости относительно R^{1i}_{opt} используются: а) p_1-значения, определяемые как доли случайных выборок из множества $\{\tilde{S}^r_i\}$, для которых максимальное значение фукционала качества F_q превышает максимальное значение $F_q(R^2_{\text{opt}}, \tilde{S})$, б) функционал P^2_{q1}, равный отношению $F_q(R_{\text{opt}}, \tilde{S})$ к максимальному значению F_q, достигнутому на наборе $\{\tilde{S}^r_i\}$. Следует отметить, что величины p_1 фактически оценивают вклад в двумерную закономерность объясняющей переменной X_j. Совершенно аналогичным образом производится верификация закономерности R^2_{opt} относительно простейшей одномерной закономерности для X_j. При этом вычисляются величины p_2 и P^2_{q2}, рассчитанные относительно R^{1j}_{opt} по набору выборок $\{\tilde{S}^r_j\}$ с помощью тех же самых процедур, что p_1 и P^2_{q1} . В случае, если простейшая закономерности по какой либо из объясняющих переменных отсутствует, верификация производится путём проверки нулевой гипотезы о полной независимости Y относительно объясняющих переменных. Выходная система закономерностей формируется с использованием з рассчитанных р-значений. Двумерные разбиения считаются верифицированными закономерностями, если оба р-значения p_1 и p_2 не превышают заранее заданного уровня значимости.

Математические и алгоритмические средства анализа систем закономерностей

Общее число закономерностей в задачах высокой размерности очень сильно возрастает и может достигать нескольких сотен, что затрудняет их визуальный анализ и делает необходимой разработку математических и алгоритмических средств иследования структуры систем найденных закономерностей. Подобные средства позволили бы дать полную, подробную и статистически достоверную характеризацию зависимость Y от переменных $X_1, \ldots, X_n$.

Методы оценки информативности отдельных объясняющих переменных. Простейшими инструментами являются методы оценки степени участия в закономерностях отдельных объясняющих переменных. Оценка информативности переменных может производится по одномерным закономерностям. При этом для оценок информативности переменной X_i естественно использовать величину функционала P_q для одномерной закономерности R^{1i}_{opt}. Однако для многих задач информативность, связанная только с одномерными закономерностями, характеризует существующую зависимость до-

статочно поверхностно. Более точную характеристику может дать оценка информативности по системам закономерностей более высокой размерности. Пусть $\tilde{R}^2 = \{r^{ij} | i, j \in \{1, \ldots, n\}\}$, где r^{ij} — закономерность для пары переменных (X_i, X_j). Для оценки информативности переменной X_i может быть использован индекс

$$\gamma(X_i) = \sum_{r^{ij} \in \tilde{R}^2} P^2_{q2}(r^{ij}) + \sum_{r^{ji} \in \tilde{R}^2} P^2_{q1}(r^{ji}).$$

Эксперименты на реальных выборках данных продемонстрировали, что информативность объясняющих переменных, рассчитанная олько по одномерным закономерностям может существенно отличаться от информативности, рассчитанной по наборам двумерных закономерностей.

Методы основанные на сравнении закономерностей. Эффективным способом сравнения закономерностей, найденных для разных объясняющих переменных и описываемых с помощью разбиений из различных моделей, является сравнение соответствующих этим закономерностям предикторов. Пусть закономерность r описывается с помощью разбиения $\{q_1, \ldots, q_l\}$. Закономерности r можно поставить предиктор Z, вычисляющий для произвольного описания $\mathbf{x}$ в пространстве объясняющих переменных прогноз переменной Y. Пусть $\hat{y}_i = \frac{1}{m_i} \sum_{\mathbf{x}_j \in q_i} y_j$, где m_i — число объектов обучающей выборки, для которых вектор объясняющих переменных принадлежит элементу разбиения q_i, $i = 1, \ldots, l$. Прогноз в точке $\mathbf{x}$ вычисляется предиктором Z по формуле $Z(\mathbf{x}) = \sum_{i=1}^{l} \hat{y}_i I_i(\mathbf{x})$, где $I_i(\mathbf{x}) = 1$ при $\mathbf{x} \in q_i$ и $I_i(\mathbf{x}) = 0$ в противном случае, $i = 1, \ldots, l$. Выраженность закономерности, то есть степень разделения объектов с различными значениями Y может характеризоваться как с помощью p-значений и функционалов качества, так и с помощью величин ошибок соответствующих предикторов. Величина квадратичной ошибки предиктора Z далее будет обозначаться $\delta(Z)$. Пусть Z_1 и Z_2 — предикторы, соответствующие закономерностям r_1 и r_2 соответственно. Расстояние между предикторами $\rho(Z_1, Z_2)$ естественно задавать как $E_\Omega(Z_1 - Z_2)^2$, где Ω — просранство исследуемых объектов с заданной на нём вероятностной мерой. Каждая закономерность может характеризоваться вектором размерности m прогнозов, рассчитанных для объектов обучающей выборки $\tilde{S}$. Таким образом система выявленных закономерностей может быть описана с помощью матрицы $||P_{ij}||_{m \times N_r}$, где N_r — число найденных закономерностей, P_{ij} — прогноз, вычисленный предиктором Z_j для i-го объекта обучающей выборки. Анализ структуры матрицы $||P_{ij}||_{m \times N_r}$ может быть произведён с помощью алгоритмов кластерного анализа или метода главных компонент. Однако недостатком таких подходов является учёт в них только лишь параметров, характеризующих взаимное расстояние между предикторами или корреляцию предикторов. Ошибки предикторов при исследовании структуры явно не учитываются. Альтенативным подходом является выбор небольшой по размеру базисной системы достаточно «выраженных» и вместе с тем «удалённых» друг отдруга закономерностей. В качестве базисной системы может использоваться набор закономерностей $\widetilde{R}_B$, которому соответствует набор предикторов $\widetilde{Z}_B$. При этом достигается минимум квадратичной ошибки коллективного предиктора $\hat{Z}_{av}$, который вычисляет прогноз, усреднённый по $\widetilde{Z}_B$: $\hat{Z}_{av} = \sum_{i=1}^{l} Z_i(\mathbf{x})$, где $l = |\widetilde{Z}_B| = |\widetilde{R}_B|$. Усреднённый прогноз является частным случаем выпуклого коллективного корректора $\hat{Z}(\mathbf{x}) = \sum_{i=1}^{l} c_i Z_i(\mathbf{x})$, где $\sum_{i=1}^{l} c_i = 1$. Для ошибки выпуклого корректора справедливо [11, 12] разложение $\delta[\hat{Z}(\mathbf{c})] = \sum_{i=1}^{l} c_i \delta(Z_i) - 1/2 \sum_{i=1}^{l} \sum_{j=1}^{l} c_i c_j \rho(Z_i, Z_j)$. При равных весах предикторов разложение принимает вид $\delta[\hat{Z}(\mathbf{u})] = \sum_{i=1}^{l} c_i \delta(Z_i) - 1/(2l^2) \sum_{i=1}^{l} \rho(Z_i, Z_j)$, где $(u) = (1/l, \ldots, 1/l)$ — вектор размерности l. Из разложения видно, что величина квадратичной ошибки предиктора $\delta[\hat{Z}(\mathbf{u})]$ убывает по мере убывания квадратичных ошибок предикторов, составляющих ансамбль, а также по мере возрастания квадратичных отклонений между предикторами. Следовательно набор предикторов, минимизирующий такую квадратичную ошибку, должен соответствовать системе «выраженных» и «удалённых» друг от друга закономерностей. Для характеризации произвольной закономерности может быть использова набор расстояний до закономерностей базисной системы. Эффективность представленных методов для оценки информативности отдельных объясняющих переменных и поиска базисной системы закономерностей была продемонстрирована на примере задачи исследования влияния различных факторов на тяжесть ДЭП [9, 10]. Эксперименты показали, что индексы информативности, рассчитанные по системам одномерных закономерностей значительно отличаются от индексов, рассчитанных по системам более высокой размерности.

Ещё одним инструментом анализа систем закономерностей является изучение доминирования одних закономерностей над другими. Мы будем говорить, что закономерность R_1 доминирует над закономерностью R_2, если для соответствующих предикторов Z_1 и Z_2 справедливы условия:

а) ошибка прогнзирования для предиктора Z_1 меньше ошибки предиктора Z_2;

б) прогностическая способность Z_1 не может быть улучшена с помощью предиктора Z_2. Невозможность улучшения прогностической способности Z_1 с помощью Z_2 эквивалентна независимости Y

от Z_2 при заданном Z_1: $(Y \perp Z_2)|Z_1$. Оценка условия доминирования для всевозможных пар закономерностей позволяет отделить основные факторы, оказывающие влияние на величину Y от второстепенных индуцированных эффектов. Под индуцированным эффектом в данном случае понимается ситуация, при которой доминирующий фактор X_{dom} оказывает прямое воздействие одновременно на Y и на некоторую величину X'. При этом Y и X' оказываются связанными. Однако величина X' не несёт дополнительной информации, позволяющей увеличить точность прогноза, достигнутую с помощью X_{dom}. Для оценки доминирования могут быть использованы статистические методы, основанные на выявлении достоверных различий между значениями Y в элементах разбиений порождаемых R_2 подобластей R_1. Предварительная оценка отсутствия доминирования может производиться путём проверки выполнимости неравенства $|\delta(Z_1) - \delta(Z_2)| < \rho(Z_1, Z_2)$. Справедливость данного неравенства является необходимым и достаточным условием существования выпуклого корректора, имеющего ошибку прогнозирования менее $\delta(Z_1)$.

Заключение

Таким образом представленный метод анализа данных состоит из двух этапов. На первом этапе производится поиск и верификация закономерностей, описываемых в рамках моделей оптимальных разбиений. На втором этапе производится анализ множества найденных закономерностей, который включает:

а) оценку интегрального вклада в множество найденных закономерностей каждой из объясняющих переменных;

б) оценку различий между закономерностями в пространстве прогнозов, вычисляемых соответствующими закономерностям предикторами;

в) выделение подмножества «выраженных» и вместе с тем удалённых друг от друга закономерностей;

г) вычисление отношения доминирования между закономерностями.

Целью анализа множества закономерностей является выделение определяющих факторов, оказывающих непосредственное влияние на прогнозируемую величину Y, а также второстепенных факторов, чья взаимосвязь с Y индуцируется основными факторами. В ходе дальнейших исследований предполагается разработка средств анализа взаимосвязей существующих между различными показателями представленными в массивах данных. При этом в качестве прогнозируемых переменных может выступать целая совокупность величин, играющих существенную с точки зрения пользователя роль в исследуемом процессе. Конечным результатом подобного анализа будут компьютерные модели, сходные по смыслу с Байесовскими сетями.

Литература

[1] *Ryazanov V. V.* Logical Regularities in Pattern Recognition (parametric approach) // Computational mathematics and Mathematical Physics. — 2007. — V. 47. — Pp. 1793–1808.

[2] *Kovshov V. V., Moiseev V. L., Ryazanov V. V.* Algorithms for finding Logical Regularities in Pattern Recognition // Computational mathematics and Mathematical Physics. — 2008. — V. 48. — Pp. 314–328.

[3] *Sen'ko O. V., Kuznetsova A. V.* The use of partitions constructions for stochastic dependencies approximation // Proceedings of the International conference on systems and signals in intelligent technologies, Minsk (Belarus), 1998. — Pp. 291–297.

[4] *O. V. Senko and A. V Kuznetsova.* The Optimal Valid Partitioning Procedures // Statistics on the Internet. — 2006. `http://statjournals.net/`

[5] *Дедовец М. С. Сенько О. В.* Алгоритм распознавания, основанный на построении метрических закономерностейю // Доклады ММРО-13, Москва, 2007. — С. 117.

[6] *Kuznetsova A. V., Sen'ko O. V., Matchak G. N., Vakhotsky V. V., Zabotina T. N., Korotkova O. V.* The Prognosis of Survivance in Solid Tumor Patients Based on Optimal Partitions of Immunological Parameters Ranges // J. Theor. Med. — 2000. — V. 2. — Pp. 317–327.

[7] *А. С. Заковряшин, С. Е. Заковряшина, И. В. Доровских, О. В. Сенько, А. В. Кузнецова, А. А. Козлов.* Прогнозирование отдалённых последствий психогенных расстройств у военнослужащих в остром периоде боевой психической травмы (с использованием логико-статистических методов // Неврология и психиатрия им. С. С. Корсакова. — 2006. — Т. 106, № 3. — С. 31–38.

[8] *Водолагина Н. Н., Костомарова И. В., Кузнецова А. В., Малыгина Н. А., Сенько О. В.* Изучение связи тяжести дисциркуляторной энцефалопатии с полиморфизмом некоторых генетических факторов методами распознавания // Доклады ICMBB-III, Пущино, 2010. — С. 238–239.

[9] *Kostomarova I., Kuznetsova A., Malygina N., Senko O.* Methods for evaluating of regularities systems structure // New Trends in Classification and Data Mining, ITHEA, Sofia, Bulgaria, 2010. — Pp. 40–46.

[10] *Kostomarova I., Kuznetsova A., Malygina N., Senko O.* Methods for evaluating of discrepancy between regularities systems in different groups // CFDM 2011, Bulgaria, Varna (в печати).

[11] *A. Krogh, J. Vedelsby.* Neural network ensembles, cross validation, and active learning // NIPS. — 1995. — V. 7. — Pp. 231–238.

[12] *Senko O. V.* An Optimal Ensemble of Predictors in Convex Correcting Procedures // Pattern Recognition and Image Analysis. — 2009. — V. 19. — Pp. 465–468.

Комбинаторная теория переобучения: результаты, приложения и открытые проблемы*

Воронцов К. В.
vokov@forecsys.ru
Москва, Вычислительный центр им. А. А. Дородницына РАН

Статья содержит краткий обзор основных результатов комбинаторной теории переобучения. Рассматриваются применения комбинаторных оценок обобщающей способности для поиска логических закономерностей и отбора эталонов в методе ближайшего соседа. Приводится список открытых проблем.

Получение верхних оценок вероятности ошибки на основе доступной информации о семействе алгоритмов, обучающей выборке и методе обучения — одна из центральных проблем теории статистического обучения. Минимизация таких оценок позволяет строить алгоритмы с управляемой обобщающей способностью. Проблема в том, что большинство оценок сильно завышены, что может приводить к неоптимальным решениям. Например, метод структурной минимизации риска, основанный на теории Вапника-Червоненкиса [2], склонен чрезмерно упрощать алгоритмы [10]. Известны также оценки, обосновывающие более тонкие принципы обучения: максимизацию отступов, минимизацию аппроксимированного эмпирического риска с различными непрерывными функциями потерь, добавление различного вида штрафных слагаемых и регуляризаторов [7]. Однако эти оценки также завышены, и объясняют далеко не все важные особенности процесса обучения. Например, до сих пор не объяснены наблюдавшиеся в экспериментах эффекты переобучения бустинга [12].

В комбинаторным подходе вместо вероятности ошибки оценивается средняя ошибка на неизвестной контрольной выборке. Это не меняет сути оценок [3], но позволяет разобраться в причинах их завышенности экспериментально, измерив промежуточные оценки в цепочке неравенств по методике скользящего контроля [14], когда контрольная выборка всё же известна. Кроме того, завышенные неравенства концентрации меры заменяются функцией гипергеометрического распределения, которая может быть вычислена точно.

Основными причинами завышенности являются эффекты *расслоения* и *связности* [14, 15], которые снижают переобучение в реальных задачах классификации, однако современные вероятностные оценки учитывают в лучшем случае только один из них. Комбинаторный подход позволяет учесть оба эффекта одновременно и получить оценки вероятности переобучения, которые в некоторых случаях оказываются точными.

Работа поддержана РФФИ (проект №11-07-00480) и программой ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

Определения и обозначения

Пусть заданы конечные множества *объектов* $\mathbb{X} = \{x_1, \ldots, x_L\}$ и *алгоритмов* $A = \{a_1, \ldots, a_D\}$ и бинарная *функция потерь* $I\colon A \times \mathbb{X} \to \{0, 1\}$, $I(a, x) = 1$ тогда и только тогда, когда алгоритм a допускает ошибку на объекте x.

Вектором ошибок алгоритма a называется L-мерный бинарный вектор $\big(I(a, x_1), \ldots, I(a, x_L)\big)$.

Число ошибок алгоритма a на выборке $X \in \mathbb{X}$ определяется как $n(a, X) = \sum\limits_{x \in X} I(a, x)$.

Частота ошибок алгоритма a на выборке X определяется как $\nu(a, X) = n(a, X)/|X|$.

Методом обучения называется отображение $\mu\colon 2^{\mathbb{X}} \to A$, которое произвольной выборке $X \subset \mathbb{X}$ ставит в соответствие некоторый алгоритм $a \in A$.

Каноническим примером метода обучения является *минимизация эмпирического риска*:

$$\mu(X) = \arg\min_{a \in A} n(a, X), \quad X \subset \mathbb{X}.$$

Рассмотрим всевозможные разбиения множества объектов $\mathbb{X} = X \sqcup \bar{X}$ на две выборки — наблюдаемую обучающую X длины ℓ и скрытую контрольную $\bar{X}$ длины $k = L - \ell$. Допустим, что все C_L^ℓ разбиений реализуются с равными вероятностями.

Переобученностью алгоритма a называется величина отклонения частоты ошибок на контроле и обучении $\delta(a, X) = \nu(a, \bar{X}) - \nu(a, X)$.

Основной задачей является получение верхних оценок *вероятности переобучения*

$$Q_\varepsilon(\mu, \mathbb{X}) = \mathsf{P}\big[\delta(\mu(X), X) \geqslant \varepsilon\big], \quad \varepsilon \in (0, 1).$$

Для получения оценок, справедливых для любого метода обучения μ, вероятность переобучения Q_ε оценивается сверху вероятностью большого равномерного (по множеству алгоритмов A) отклонения частот в двух подвыборках:

$$Q_\varepsilon(\mu, \mathbb{X}) \leqslant \widetilde{Q}_\varepsilon(A, \mathbb{X}) = \mathsf{P}\Big[\max_{a \in A} \delta(a, X) \geqslant \varepsilon\Big].$$

Задача получения верхних оценок $\widetilde{Q}_\varepsilon$ часто постулируется как основная в теории статистического обучения [2, 11, 7]. Однако оценка $Q_\varepsilon \leqslant \widetilde{Q}_\varepsilon$ уже может оказаться сильно завышенной.

Значение $\widetilde{Q}_\varepsilon$ совпадает с вероятностью переобучения для *метода максимизации переобученности*

$$\mu(X) = \arg\max_{a\in A} \delta(a, X), \quad X \subset \mathbb{X},$$

который является искусственным «худшим случаем» и вряд ли имеет практические применения.

Оценки вероятности переобучения

Будем полагать, что все алгоритмы из A имеют попарно различные векторы ошибок. Введём на A естественное отношение порядка:

$$a \leqslant b \;\leftrightarrow\; I(a,x) \leqslant I(b,x),\ \forall x \in \mathbb{X};$$
$$a < b \;\leftrightarrow\; a \leqslant b \;\text{ и }\; a \neq b.$$

Обозначим через $\rho(a,b)$ хэммингово расстояние между векторами ошибок алгоритмов a и b.

Если $a < b$ и при этом $\rho(a,b) = 1$, то будем говорить, что a *предшествует* b и записывать $a \prec b$. Очевидно, что $n(a,\mathbb{X}) + 1 = n(b,\mathbb{X})$.

Графом расслоения–связности множества алгоритмов A будем называть направленный граф $\langle A, E\rangle$ с множеством рёбер $E = \big\{(a,b)\colon a \prec b\big\}$.

Граф расслоения–связности является многодольным, доли соответствуют *слоям алгоритмов* $A_m = \big\{a \in A\colon n(a,\mathbb{X}) = m\big\}$, рёбрами могут соединяться только алгоритмы соседних слоёв. Каждому ребру (a,b) соответствует единственный объект $x_{ab} \in \mathbb{X}$, такой, что $I(a,x_{ab}) = 0$ и $I(b,x_{ab}) = 1$.

Верхней (нижней) связностью алгоритма a будем называть число рёбер графа, исходящих из (входящих в) вершину a, соответственно:

$$q(a) = \#\big\{x_{ab} \in \mathbb{X} \;\big|\; a \prec b\big\};$$
$$d(a) = \#\big\{x_{ba} \in \mathbb{X} \;\big|\; b \prec a\big\}.$$

Связность есть реализуемое семейством A число способов изменить алгоритм a так, чтобы он стал делать на одну ошибку больше (или меньше). Связность можно интерпретировать как число степеней свободы семейства A в локальной окрестности алгоритма a. Для семейства линейных классификаторов значение связности концентрируется вокруг значения размерности пространства [4].

Неоптимальностью $r(a)$ алгоритма a будем называть число объектов $x \in \mathbb{X}$, на которых алгоритм a ошибается, при том, что существует алгоритм $b \in A,\ b < a$, не ошибающийся на x:

$$r(a) = \#\big\{x \in \mathbb{X} \;\big|\; \exists b \in A\colon b < a,\ I(b,x) < I(a,x)\big\}.$$

Другими словами, $r(a)$ есть число различных объектов x_{bc}, соответствующих всевозможным рёбрам (b,c) на путях, ведущих к вершине a.

В общем случае $d(a) \leqslant r(a) \leqslant n(a,\mathbb{X})$. Равенство $r(a) = d(a)$ достигается на всех алгоритмах двух самых нижних слоёв. Равенство $r(a) = n(a,\mathbb{X})$ достигается в случае, когда существует корректный алгоритм $a_0 \in A$: $n(a_0,\mathbb{X}) = 0$.

Оценки расслоения-связности. Определим для всех $m = 0,\ldots,L,\ z = 0,\ldots,\ell$ функцию гипергеометрического распределения

$$H_L^{\ell,\,m}(z) = \sum_{s=0}^{\lfloor z\rfloor} \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}.$$

Теорема 1 (оценка расслоения–связности). *Пусть μ — метод минимизации эмпирического риска, векторы ошибок всех алгоритмов из A попарно различны. Тогда для любого $\varepsilon \in (0,1)$*

$$Q_\varepsilon(\mu,\mathbb{X}) \leqslant \sum_{a\in A} \frac{C_{L-q-r}^{\ell-q}}{C_L^\ell} H_{L-q-r}^{\ell-q,\,m-r}\left(\tfrac{\ell}{L}(m-\varepsilon k)\right),$$

где q, r — верхняя связность и неоптимальность алгоритма a соответственно, $m = n(a,\mathbb{X})$.

Рассмотрим основные свойства этой оценки.

1. Комбинаторный множитель $\bar{P}_a = C_{L-q-r}^{\ell-q}/C_L^\ell$ есть верхняя оценка вероятности $P_a = \mathsf{P}\big[\mu(X) = a\big]$ получить алгоритм a в результате обучения. Величина $\bar{P}_a$ экспоненциально убывает с ростом неоптимальности r и связности q. Отсюда следуют два важных для практики вывода.

Во-первых, связные множества алгоритмов менее подвержены переобучению. Это относится к широкому классу параметрических семейств алгоритмов классификации, у которых разделяющая поверхность непрерывна по параметрам.

Во-вторых, лишь несколько нижних слоёв вносят существенный вклад в переобучение. Это позволяет вычислять приближённые (нижние) оценки $\bar{Q}_\varepsilon$, перебирая алгоритмы по слоям снизу вверх. Поскольку в нижних слоях находится обычно очень малая доля алгоритмов, такие оценки могут быть вычислены достаточно эффективно.

2. Если пренебречь расслоением и связностью, положив $r = q = 0$ для каждого $a \in A$, то получится оценка Вапника–Червоненкиса (при $\ell = k$):

$$Q_\varepsilon(\mu,\mathbb{X}) \leqslant \sum_{a\in A} H_L^{\ell,\,m}\left(\tfrac{\ell}{L}(m-\varepsilon k)\right) \leqslant |A| \cdot \tfrac{3}{2} e^{-\varepsilon^2 \ell}.$$

3. Оценка расслоения–связности является достижимой. Она обращается в равенство в случае специальных модельных семейств алгоритмов — монотонных цепей [16] и многомерных сетей [6].

Теорема 2 (оценка связности). *Пусть векторы ошибок всех алгоритмов из A попарно различны. Тогда для любого $\varepsilon \in (0,1)$*

$$\widetilde{Q}_\varepsilon(A,\mathbb{X}) \leqslant \sum_{a\in A} \frac{C_{L-q-d}^{\ell-q}}{C_L^\ell} H_{L-q-d}^{\ell-q,\;m-d}\left(\tfrac{\ell}{L}(m-\varepsilon k)\right),$$

где q, d, m — соответственно, верхняя связность, нижняя связность и число ошибок алгоритма a на генеральной выборке.

Теоремы 1 и 2 доказываются с помощью метода порождающих и запрещающих множеств [16].

Эта оценка гораздо слабее предыдущей и отличается от неё тем, что неоптимальность r заменяется на нижнюю связность d. Неоптимальность растёт с номером слоя m линейно, тогда как нижняя связность d не превышает r и примерно одинакова во всех слоях [4]. Поэтому алгоритмы всех слоёв вносят примерно равный вклад в оценку. Таким образом, оценка связности не учитывает эффект расслоения. На самом деле он игнорируется самим функционалом равномерного отклонения $\widetilde{Q}_\varepsilon$, а, значит, и любыми его верхними оценками, в частности, оценками рademахеровской сложности [7].

Модельные семейства алгоритмов являются «искусственными» и задаются непосредственно набором векторов ошибок, а не реальной выборкой данных. Каждое из них обладает некоторой «регулярностью» или симметрией, что и позволяет получать для них точные комбинаторные оценки.

К настоящему времени точные оценки получены для слоёв и интервалов булева куба, монотонных и унимодальных цепей [16] и многомерных сетей [6], хэмминговых шаров и некоторых их разреженных подмножеств [5]. Разработан теоретико-групповой подход [8], который позволяет получать точные оценки для семейств с произвольными симметриями. Монотонные сети, по всей видимости, наиболее близки к реальным семействам, поскольку обладают свойствами расслоения, связности и размерности. Экспериментально показано, что вероятность переобучения стандартных методов классификации (нейронных сетей, решающих деревьев, ближайшего соседа) на реальных задачах может быть аппроксимирована с помощью монотонных сетей подходящей размерности [1].

Логические классификаторы. Оценка расслоения–связности применена в [17] для уменьшения переобучения логических правил вида

$$r(x;\theta) = \prod_{j\in J} \bigl[x^j \leqslant \theta^j\bigr],$$

где $x = (x^1,\ldots,x^n) \in \mathbb{R}^n$ — вектор признакового описания объекта x, $J \subseteq \{1,\ldots,n\}$ — подмножество признаков, $\theta = (\theta^1,\ldots,\theta^n) \in \mathbb{R}^n$ — вектор порогов, θ^j — порог по j-му признаку. Если каждый признак принимает на объектах выборки разные значения, то данное семейство правил связно.

С помощью оценки расслоения–связности в [17] построен модифицированный критерий информативности логических правил, который учитывает переобучение, возникающее в результате оптимизации порогов θ по обучающей выборке. Эксперименты на реальных задачах классификации показали, что в результате модификации частота ошибок на независимых тестовых данных снижается в среднем на 1–2%.

Полный скользящий контроль

Обобщающая способность может быть охарактеризована не только вероятностью переобучения, но и функционалом полного скользящего контроля

$$C(\mu,\mathbb{X}) = \mathsf{E}\nu(\mu X, \bar{X}).$$

Теорема 3 (оценка расслоения–связности). *При условиях Теоремы 1 справедлива оценка*

$$C(\mu,\mathbb{X}) \leqslant \sum_{a\in A} \frac{C^{\ell-r}_{L-q-r}}{C^\ell_L} \left(\frac{m}{k} - \frac{(\ell-q)(m-r)}{(L-q-r)k}\right).$$

Известны точные оценки функционала C для двух частных случаев: метода ближайшего соседа и метода монотонной классификации.

Пусть $y(x)$ — истинная функция классификации, $I(a,x) = \bigl[a(x) \neq y(x)\bigr]$.

Метод ближайшего соседа запоминает обучающую выборку X и строит алгоритм $a = \mu(X)$, относящий произвольный объект $x \in \mathbb{X}$ к тому классу, которому принадлежит обучающий объект $x' \in X$, ближайший к x по заданной функции расстояния $\rho(x,x')$.

Профилем компактности называется функция $K(m)$, определяемая как доля объектов $x \in \mathbb{X}$, таких, что $y(x)$ не совпадает с истинной классификацией m-го соседа объекта x.

Теорема 4 (о профиле компактности [3]). *Для метода ближайшего соседа справедливо*

$$C(\mu,\mathbb{X}) = \sum_{m=1}^{k} \frac{C^{\ell-1}_{L-1-m}}{C^\ell_{L-1}} K(m).$$

В [9] предложен метод отбора эталонных объектов, основанный на минимизации данной оценки. Эксперименты показали, что он практически не переобучается, и оптимальное число эталонов надёжно определяется по обучающим данным.

Монотонные классификаторы. Пусть множество объектов $\mathbb{X}$ частично упорядочено. Монотонные функции вида $a\colon \mathbb{X} \to \{0,1\}$ будем называть *монотонными классификаторами*.

Нижняя область объекта x_i — это множество всех объектов x таких, что $x \leqslant x_i$.

Верхняя область объекта x_i — это множество всех объектов x таких, что $x_i \leqslant x$.

Определим для произвольной пары объектов $x,u \in \mathbb{X}$ расстояние $r(x,u)$ между нижней областью объекта x и верхней областью объекта u так, чтобы функция r обладала следующими свойствами:

1) $r(x,u) = 0$ тогда и только тогда, когда $x \geqslant u$;
2) $r(x,u)$ не возрастает по x не убывает по u.

Рассмотрим метод ближайшего соседа, определив функцию расстояния $\rho(x,x')$ от классифицируемого объекта x до обучающего объекта $x' \in X$ как расстояние до его нижней области, если $y(x') = 0$, и до его верхней области, если $y(x') = 1$.

Теорема 5. *При таком определении функции расстояния $\rho(x, x')$ метод ближайшего соседа строит монотонные классификаторы и справедлива теорема о профиле компактности.*

Оценки монотонных классификаторов по профилю компактности могут быть использованы при построении композиций алгоритмов классификации или ранжирования с монотонными корректирующими операциями [13].

Открытые проблемы

В заключение приведём (далеко не полный) перечень проблем, которые пока остаются открытыми в комбинаторной теории переобучения:

1) получение оценок, учитывающих связи между алгоритмами с хемминговым расстоянием, большим 1; их игнорирование является, по всей видимости, основной причиной завышенности оценки расслоения–связности;
2) получение оценок, не зависящих от характеристик полной выборки $\mathbb{X}$, а только от наблюдаемых величин, которые могут быть вычислены по конкретной обучающей подвыборке X;
3) расширение класса методов обучения, для которых комбинаторные оценки позволяют улучшать обобщающую способность;
4) получение эффективно вычислимых оценок для произвольных цепей алгоритмов;
5) получение оценок расслоения–связности для семейства линейных классификаторов;
6) обоснование и, возможно, уточнение принципов регуляризации, аппроксимации эмпирического риска, максимизации зазора;
7) обобщение комбинаторных оценок на случай небинарных функций потерь;
8) проверка гипотезы о том, что значения связности концентрируются вокруг значения локальной размерности пространства параметров;
9) проверка гипотезы о сепарабельности профиля расслоения-связности [16];
10) получение оценок обобщающей способности для задач динамического (online) обучения.

Литература

[1] *Ботов П. В.* Точные оценки вероятности переобучения для монотонных и унимодальных семейств алгоритмов // Всеросс. конф. Математические методы распознавания образов-14. — М.: МАКС Пресс, 2009. — С. 7–10.

[2] *Вапник В. Н., Червоненкис А. Я.* Теория распознавания образов. — М.: Наука, 1974.

[3] *Воронцов К. В.* Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики / Под ред. О. Б. Лупанова. — М.: Физматлит, 2004. — Т. 13. — С. 5–36.

[4] *Кочедыков Д. А.* Структуры сходства в семействах алгоритмов классификации и оценки обобщающей способности // Всеросс. конф. Математические методы распознавания образов-14. — М.: МАКС Пресс, 2009. — С. 45–48.

[5] *Толстихин И. О.* Вероятность переобучения плотных и разреженных семейств алгоритмов // Межд. конф. Интеллектуализация обработки информации ИОИ-8. — М.: МАКС Пресс, 2010. — С. 83–86.

[6] *Botov P. V.* Exact bounds on probability of overfitting for multidimensional model sets of classifiers // Pattern Recognition and Image Analysis. — 2011. — Vol. 21., no. 1. — Pp. 52–65.

[7] *Boucheron S., Bousquet O., Lugosi G.* Theory of classification: A survey of some recent advances // ESAIM: Probability and Statistics. — 2005. — no. 9. — Pp. 323–375.

[8] *Frei A. I.* Accurate estimates of the generalization ability for symmetric set of predictors and randomized learning algorithms // Pattern Recogn. and Image Analysis. — 2010. — Vol. 20, no. 3. — P. 241–250.

[9] *Ivanov M. N.* Prototype sample selection based on minimization of the complete cross validation functional // Pattern Recognition and Image Analysis. — 2010. — Vol. 20, no. 4. — Pp. 427–437.

[10] *Kearns M. J., Mansour Y., Ng A. Y., Ron D.* An experimental and theoretical comparison of model selection methods // 8th Conf. on Computational Learning Theory, Santa Cruz, USA. — 1995. — Pp. 21–30.

[11] *Koltchinskii V., Panchenko D.* Rademacher processes and bounding the risk of function learning // High Dimensional Probability, II / Ed. by D. E. Gine, J.Wellner. — Birkhauser, 1999. — Pp. 443–457.

[12] *Ratsch G., Onoda T., Muller K.-R.* Soft margins for AdaBoost // Machine Learning. — 2001. — Vol. 42, no. 3. — Pp. 287–320.

[13] *Spirin N. V., Vorontsov K. V.* Learning to rank with nonlinear monotonic ensemble // 10th International Workshop on Multiple Classidier Systems. Naples, Italy, June 15–17, 2011. — Lecture Notes in Computer Science. Springer-Verlag, 2011. — Pp. 16–25.

[14] *Vorontsov K. V.* Combinatorial probability and the tightness of generalization bounds // *Pattern Recognition and Image Analysis.* — 2008. — Vol. 18, no. 2. — Pp. 243–259.

[15] *Vorontsov K. V.* Splitting and similarity phenomena in the sets of classifiers and their effect on the probability of overfitting // *Pattern Recognition and Image Analysis.* — 2009. — Vol. 19, no. 3. — Pp. 412–420.

[16] *Vorontsov K. V.* Exact combinatorial bounds on the probability of overfitting for empirical risk minimization // Pattern Recognition and Image Analysis. — 2010. — Vol. 20, no. 3. — Pp. 269–285.

[17] *Vorontsov K. V., Ivahnenko A. A.* Tight combinatorial generalization bounds for threshold conjunction rules // 4th International Conference on Pattern Recognition and Machine Intelligence (PReMI'11). June 27 – July 1, 2011. — Lecture Notes in Computer Science. Springer-Verlag, 2011. — Pp. 66–73.

Уменьшение вероятности переобучения итерационных методов статистического обучения*

Ботов П. В.
pbotov@forecsys.ru
Московский физико-технический институт (государственный университет)

Для итерационных методов статистического обучения предложен метод минимизации предсказанного риска, основанный на аппроксимации семейства классификаторов унимодальной несимметричной сетью алгоритмов малой высоты и размерности, для которой известны точные комбинаторные формулы вероятности переобучения. Эксперименты на решающих деревьях и реальных задачах классификации показывают, что предложенный подход повышает обобщающую способность получаемых алгоритмов классификации.

Введение

Комбинаторная теория обобщающей способности в настоящее время позволяет получать точные (не асимптотические и не завышенные) оценки вероятности переобучения, главным образом, для модельных семейств алгоритмов. В [7] такие оценки были получены для монотонных и унимодальных цепей алгоритмов, в [4] — для связки монотонных цепей; в [1] — для многомерных монотонных и унимодальных сетей алгоритмов, в [2] и [3] — для многомерных несимметричных сетей алгоритмов.

Модельные семейства вряд ли могут порождаться практическими задачами. Тем не менее, они интересны тем, что обладают свойствами расслоения и связности, благодаря которым переобучение существенно понижается, и которыми с необходимостью должны обладать реальные семейства, чтобы их можно было применять на практике [6].

Некоторые модельные семейства (сети, связки) обладают также свойством «размерности» или «числа степеней свободы», что приближает их к реальным семействам. Вероятность переобучения нейронных сетей, решающих деревьев, байесовских классификаторов на реальных задачах классификации неплохо приближается вероятностью переобучения многомерной монотонной сети при соответствующем подборе её размерности [1].

Целью настоящей работы является применение комбинаторной теории переобучения для снижения переобучения в итерационных методах статистического обучения. Это широкий класс методов, в которых процесс обучения представляется в виде последовательности шагов. На каждом шаге вычисляется некоторая статистика (действительная функция выборки) и уточняется конструкция алгоритма. Примерами являются методы синтеза решающих списков, деревьев и лесов, индукции логических закономерностей, градиентного обучения нейронных сетей, и многие другие. В литературе такие методы называют также обучением по *статистическим запросам* [5]. В данной работе исследуются решающие деревья.

Примером статистического запроса является *эмпирический риск* — частота ошибок фиксированного классификатора на обучающей выборке объектов. В данной работе предлагается брать за основу *предсказанный риск* — оценку частоты ошибок на скрытой контрольной выборке, вычисленную по наблюдаемой обучающей выборке. Для этого семейство классификаторов аппроксимируется в окрестности оптимума модельным семейством — унимодальной несимметричной сетью (УНС) [3]. Для этого модельного семейства известна точная комбинаторная формула вероятности переобучения, по которой и оценивается предсказанный риск.

Для проверки предлагаемого метода обучения проводится эксперимент на четырёх реальных задачах классификации из репозитория UCI. Сравниваются два метода минимизации предсказанного риска — использующий УНС-аппроксимацию и использующий эмпирические оценки скользящего контроля (метод Монте-Карло). Оба метода дают выигрыш в качестве классификации в сравнении с методом минимизации эмпирического риска. При этом УНС-аппроксимация вычислительно гораздо эффективнее.

Основные понятия и обозначения

Задано множество объектов $\mathbb{X} = \{x_1, \ldots, x_L\}$, называемое *генеральной выборкой*, конечное множество A, элементы которого называются *алгоритмами*, и бинарная функция ошибки $I\colon A \times \mathbb{X} \to \to \{0, 1\}$. Если $I(a, x) = 1$, то алгоритм a ошибается на объекте x, иначе a не ошибается на x.

Для задачи классификации с конечным множеством классов $\mathbb{Y} = \{y_i\}$ вводится функция ошибки $I(a, x) = [a(x) \neq y(x)]$, где $y(x)$ — класс объекта x, выражение в квадратных скобках здесь и далее определяется как [истина] = 1 и [ложь] = 0.

Вектором ошибок алгоритма a называется бинарный вектор $\big(I(a, x_i)\big)_{i=1}^{L}$.

Числом ошибок алгоритма a на выборке $X \subseteq \mathbb{X}$ называется величина $n(a, X) = \sum\limits_{x \in X} I(a, x)$.

Работа поддержана РФФИ (проект №11-07-00480) и программой ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

Частотой ошибок алгоритма a на выборке X называется величина $\nu(a,X) = n(a,X)/|X|$.

Подмножество $A_t = \{a \in A\colon n(a,\mathbb{X}) = t + m\}$ называется *t-слоем* множества алгоритмов A относительно выборки $\mathbb{X}$, где m — число ошибок *лучшего алгоритма*, $m = \min\limits_{a\in A} n(a,\mathbb{X})$.

Методом обучения называется отображение $\mu\colon 2^{\mathbb{X}} \to A$, которое произвольной обучающей выборке X ставит в соответствие некоторый алгоритм $a = \mu X$ из A. Метод обучения μ называется методом минимизации эмпирического риска (МЭР), если для любой обучающей выборки X

$$\mu X = \arg\min_{a\in A} n(a,X).$$

Допустим, что все C_L^ℓ разбиений генеральной выборки $\mathbb{X} = X \sqcup \bar{X}$ на наблюдаемую обучающую выборку X длины ℓ и скрытую контрольную выборку $\bar{X}$ длины $k = L - \ell$ равновероятны.

Вероятность переобучения — это доля разбиений, при которых разность частоты ошибок на контроле и обучении превышает $\varepsilon \in (0,1)$:

$$Q_\varepsilon(\mu,\mathbb{X}) = \mathsf{P}\big[\nu(\mu X,\bar{X}) - \nu(\mu X,X) \geqslant \varepsilon\big],$$

Вероятность t-слоя — это доля разбиений, при которых результатом обучения является алгоритм из t-слоя:

$$P_t(\mu,\mathbb{X}) = P[\mu X \in A_t] = P[n(\mu X,\mathbb{X}) = t + m].$$

Унимодальная несимметричная сеть

Для получения оценок переобучения в настоящей работе используется унимодальная несимметричная сеть (УНС).

Вектор $\boldsymbol{w} = (w_j)_{j=1}^h \in \mathbb{Z}^h$ будем называть *вектором индексов*. Положим $|\boldsymbol{w}| = \sum_{j=1}^h |w_j|$. Введём на множестве целочисленных векторов частичный порядок: $\boldsymbol{w} \leqslant \boldsymbol{w}'$ тогда и только тогда, когда $w_j \leqslant w'_j$ для всех $j \in 1,\ldots,h$.

Определение 1. *Унимодальной несимметричной h-мерной сетью алгоритмов называется множество $A = \big\{a_{\boldsymbol{w}}\colon \boldsymbol{W}^- \leqslant \boldsymbol{w} \leqslant \boldsymbol{W}^+,\ |\boldsymbol{w}| \leqslant W_0\big\}$, где $\boldsymbol{W}^- = (W_j^-)_{j=1}^h$ и $\boldsymbol{W}^+ = (W_j^+)_{j=1}^h$ — векторы высот по каждой размерности, $\boldsymbol{W}^- \leqslant \boldsymbol{0} \leqslant \boldsymbol{W}^+$, W_0 — высота сети, если выполнены условия:*

1) *если $\boldsymbol{0} \leqslant \boldsymbol{w} < \boldsymbol{w}'$ или $\boldsymbol{w}' < \boldsymbol{w} \leqslant \boldsymbol{0}$, то для всех $x_i \in \mathbb{X}$ выполнено $I(a_{\boldsymbol{w}},x_i) \leqslant I(a_{\boldsymbol{w}'},x_i)$, причём ровно $|\boldsymbol{w} - \boldsymbol{w}'|$ из этих неравенств строгие;*
2) *$n(a_{\boldsymbol{w}},\mathbb{X}) = m + |\boldsymbol{w}|$ при некотором $m \geqslant 0$ для всех $a_{\boldsymbol{w}} \in A$; алгоритм $a_{\boldsymbol{0}}$ называется лучшим.*

Унимодальная сеть — это модель семейства алгоритмов с h непрерывными параметрами, в котором по мере увеличения j-го параметра ошибки возникают последовательно на объектах $x_1^j,\ldots,x_{W_j^+}^j$, а при уменьшении — на $x_{-1}^j,\ldots,x_{W_j^-}^j$.

Оценки вероятности переобучения и вероятности t-слоя для УНС получены в [2, 3]. В данной работе будет использоваться величина математического ожидания номера слоя

$$t^*(\mu,\mathbb{X}) = \sum_{t=0}^{W_0} tP_t(\mu,\mathbb{X}). \qquad (1)$$

Величина t^* предсказывает, насколько в среднем число ошибок алгоритма $a = \mu X$, получаемого в результате обучения, превышает число ошибок m лучшего алгоритма.

Обучение решающего дерева

Будем рассматривать бинарные решающие деревья с условиями ветвления во внутренних вершинах вида конъюнкций пороговых предикатов:

$$\beta(x;\theta) = \prod_{j\in J} \big[x^j \lessgtr_j \theta^j\big],$$

где $(x^1,\ldots,x^n) \in \mathbb{R}^n$ — вектор признакового описания объекта x, $\theta = (\theta^1,\ldots,\theta^n) \in \mathbb{R}^n$ — вектор порогов, θ^j — порог по j-му признаку, $J \subseteq \{1,\ldots,n\}$ — подмножество признаков, $\lessgtr_j$ — одна из операций сравнения $\{\leqslant,\geqslant\}$. Будем рассматривать только конъюнкции малого ранга, $|J| \leqslant 3$.

Обучение решающего дерева является итерационным. На каждом шаге один из листьев дерева заменяется внутренней вершиной с предикатом $\beta(x)$ указанного вида и двумя дочерними листьями, помеченными классами c_0 и c_1. Таким образом, за один шаг обучения дерево наращивается одной внутренней вершиной и одним листом. Допустим, выбранный лист относил множество попадающих в него объектов $U \subset \mathbb{X}$ к классу $c \in \mathbb{Y}$. Новое дерево будет относить объекты $x \in U$ к классу $c_{\beta(x)}$.

Рассмотрим два способа оценки качества добавленного предиката $\beta(x)$ с параметрами $(\theta_j,\lessgtr_j)_{j\in J}$ и листьями c_0 и c_1.

1. В методе *минимизации эмпирического риска* (МЭР) ищется предикат β, максимально уменьшающий число ошибок нового дерева по сравнению с предыдущим деревом:

$$n(\beta,U) = \sum_{x\in U} \big[c_{\beta(x)} \neq y(x)\big] - \big[c \neq y(x)\big] \to \min_\beta.$$

2. В методе *минимизации предсказанного риска* (МПР) ищется предикат β, максимально уменьшающий эмпирический риск с поправкой на ожидаемое число дополнительных ошибок, возникающих по причине переобучения:

$$n'(\beta,U) = n(\beta,U) + t^*(\mu,U) \to \min_\beta.$$

Первый метод, МЭР, рассматривается в качестве эталона для сравнения. Второй метод, МПР,

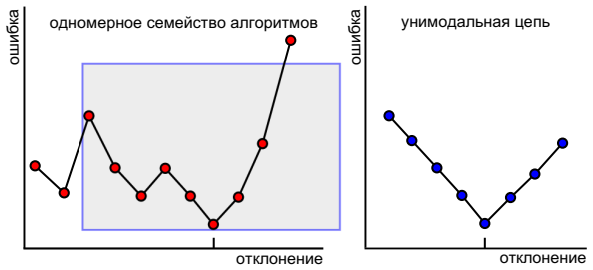


Рис. 1. Реальное одномерное семейство и аппроксимирующая его унимодальная цепь.

использует дополнительную информацию о структуре семейства, учитывающую эффекты расслоения и связности. Поэтому МПР должен приводить к классификаторам с меньшей разницей между частотами ошибок на обучении и контроле.

УНС-аппроксимация

Чтобы воспользоваться оценкой (1) для унимодальной несимметричной сети (УНС) необходимо подобрать адекватные параметры высот W_0, $\boldsymbol{W}^- = (W_j^-)_{j=1}^h$, $\boldsymbol{W}^+ = (W_j^+)_{j=1}^h$ и числа ошибок лучшего алгоритма m. Для этого предлагается следующая эвристическая процедура *УНС-аппроксимации*, которая по реальному семейству алгоритмов строит заменяющую его УНС.

1. Полным перебором параметров $(\theta_j, \lessgtr_j)_{j\in J}$, c_0, c_1 находится предикат $\beta(x)$, минимизирующий эмпирический риск на выборке U. Параметр УНС m полагается равным числу ошибок этого предиката на U.

2. Размерность УНС полагается равной числу предикатов в конъюнкции, $h = |J|$.

3. Путём варьирования каждого из порогов θ_j строятся одномерные семейства алгоритмов (в случае, когда значения вещественного признака x^j попарно различны на объектах выборки, одномерное семейство является цепью). Одномерное семейство ограничивается высотой в $W_0 = 5$ ошибок и расстоянием в 5 ошибочных объектов от лучшего положения порога. Высотам W_j^- и W_j^+ присваиваются значения максимального числа ошибок (минус m) в левой и правой ветвях оставшегося семейства. Данное правило схематично показано на рис. 1.

Значение параметра $W_0 = 5$ выбрано достаточно произвольно. Бóльшие значения приводят к увеличению времени вычисления оценки (1). Как показано в [3], число нижних слоёв УНС, дающих существенный вклад в переобучение, возрастает с ростом размерности h. Поэтому параметр высоты W_0 необходимо будет увеличить, если допустить использование конъюнкций большего ранга. Для ранга 3 пяти слоёв вполне достаточно.

Процедура УНС-аппрокимации имеет следующие недостатки. Во-первых, предполагается независимость объектов по каждой из размерностей. Если среди признаков из J есть высококоррелированные, оценка t^* будет завышенной. Во-вторых, предполагается, что одномерные семейства являются унимодальными цепями. Если это не так, оценка t^* может оказаться как выше, так и ниже действительной.

Эксперименты на реальных задачах классификации показали, что при УНС-аппроксимации оценка t^* завышается на 5–10% для одномерных семейств, на 10–30% для двумерных и до 50% для трёхмерных. Примерно в половине случаев одномерные семейства аппрокимируются вообще без ошибок, поскольку окрестность лучшего алгоритма радиуса 5 оказывается унимодальной цепью.

Метод Монте-Карло

Существует альтернативный способ получения оценки (1). Согласно определению, *вероятность t-слоя* P_t — это доля всех разбиений выборки $\mathbb{X} = X \sqcup \bar{X}$, при которых результатом обучения является алгоритм из t-слоя. Эмпирическое оценивание методом Монте-Карло предполагает, что берутся не все разбиения, а достаточно большое число N случайных разбиений (в наших экспериментах $N = 5000$), затем вероятность вычисляется буквально по определению.

Этот способ вычисления существенно более трудоёмкий с вычислительной точки зрения, по сравнению с УНС-аппроксимацией. При меньших N трудоёмкость пропорционально снижается, но оценка становится существенно менее точной.

В данной работе эмпирические оценки методом Монте-Карло использовались в качестве эталона для сравнения.

Эксперимент на реальных данных

Эксперимент проводился на четырёх реальных задачах классификации из репозитория UCI: Wdbc, Yeast, Sonar, German.

Сравнивалось качество классификации трёх методов: МЭР, МПР с УНС-аппроксимацией и МПР с использованием метода Монте-Карло.

Выборка 100 раз разбивалась на обучающую и тестовую в пропорции $9:1$. Для каждого разбиения по обучающей выборке строилось решающее дерево, которое итерационно усложнялось от 0 до 20 предикатов. При невозможности выбрать предикат для следующего шага в каком-либо из тестируемых методов обучение преждевременно останавливалось.

При вычислении оценок $t^*(\mu, U)$ выборка U разбивалась на «наблюдаемую обучающую» и «скрытую контрольную» равной или почти равной длины, $\ell = \lfloor |U|/2 \rfloor$, $k = |U| - \ell$.

Результаты приведены на рис. 2–5. На них изображены усреднённые по всем 100 разбиениям пары зависимостей частоты ошибок на обучении и тесте

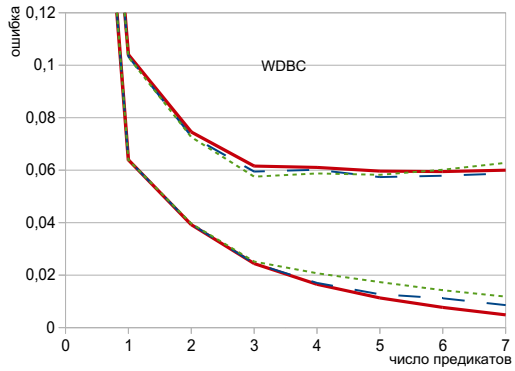


Рис. 2. Задача Wdbc, 659 объектов, 30 признаков, 2 класса.

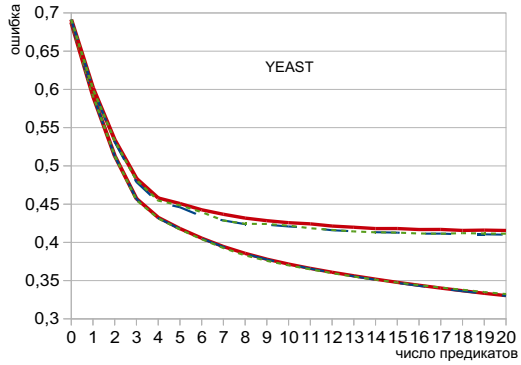


Рис. 3. Задача Yeast, 1484 объектов, 8 порядковых признаков, 10 классов.

в зависимости от числа внутренних узлов в решающем дереве.

Каждый график содержит три пары кривых (в каждой паре нижняя кривая — частота ошибок на обучающей выборке, верхняя — на тестовой):

1) сплошные линии — МЭР;

2) точки — МПР с УНС-аппроксимацией;

3) пунктир — МПР с методом Монте-Карло.

Во всех случаях при использовании МПР частота ошибок на тесте оказывается ниже, а на обучении — выше, чем при использовании МЭР.

МПР с методом Монте-Карло показывает немного меньший эффект, чем МПР с УНС-аппроксимацией, что может быть связано с завышенностью оценки t^* при аппроксимации реальных семейств модельными УНС.

Во всех случаях МПР уменьшает переобучение, но незначительно. Это объясняется тем, что переобучение возникает не только вследствие оптимизации порогов θ^j, но и вследствие оптимизации подмножеств признаков J в каждой внутренней вершине дерева. По всей видимости, вторая причина переобучения является более существенной. Её исследование в рамках комбинаторной теории переобучения пока остаётся открытой проблемой.

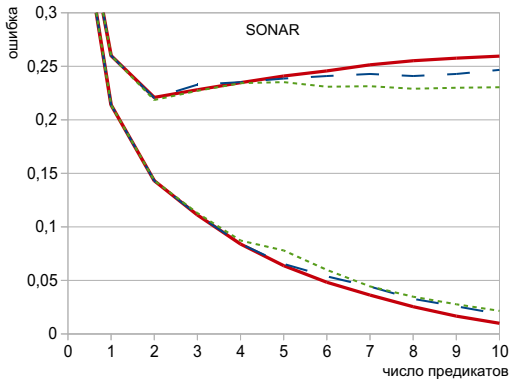


Рис. 4. Задача Sonar, 208 объектов, 60 порядковых признаков, 2 класса.

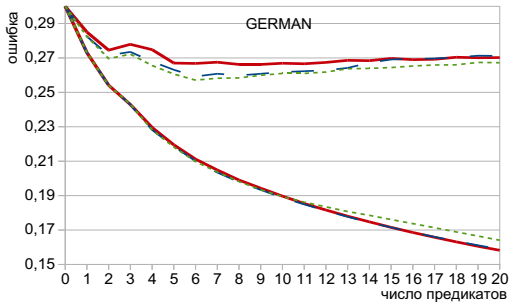


Рис. 5. Задача German, 1000 объектов, 12 порядковых признаков (из 20), 2 класса.

Литература

[1] *Ботов П. В.* Точные оценки вероятности переобучения для монотонных и унимодальных семейств алгоритмов // Всеросс. конф. ММРО-14. — М.: МАКС Пресс, 2009. — С. 7–10.

[2] *Ботов П. В.* Точные оценки вероятности переобучения для несимметричных многомерных семейств алгоритмов // Межд. конф. Интеллектуализация обработки информации ИОИ-8, 2010. — М.: МАКС Пресс, 2010. — С. 20–23.

[3] *Botov P. V.* Exact estimates of the probability of overfitting for multidimensional modeling families of algorithms // Patern Recognition and Image Analisys. — 2011. — Vol. 21, No. I. — Pp. 52–65.

[4] *Frey A. I.* Accurate estimates of the generalization ability for symmetric sets of predictors and randomized learning algorithms // Pattern Recognition and Image Analysis. — 2010. — Vol. 20, No. 3. — Pp. 241–250.

[5] *Kearns M.* Efficient noise-tolerant learning from statistical queries // Journal of the ACM, 1998. — Vol. 45, no. 6. — Pp. 983–1006.

[6] *Vorontsov K. V.* Splitting and similarity phenomena in the sets of classifiers and their effect on the probability of overfitting // Pattern Recognition and Image Analysis. — 2009. — Vol. 19, No. 3. — Pp. 412–420.

[7] *Vorontsov K. V.* Exact combinatorial bounds on the probability of overfitting for empirical risk minimization // Pattern Recognition and Image Analysis. — 2010. — Vol. 20, No. 3. — Pp. 269–285.

Критерии информативности пороговых логических правил с поправкой на переобучение порогов*

Ивахненко А. А., Воронцов К. В.
andrej_iv@mail.ru
Москва, Вычислительный центр им. А. А. Дородницына РАН

Комбинаторная оценка вероятности переобучения, учитывающая эффекты расслоения и связности, обобщена на широкий класс монотонных методов обучения. Предложена модификация критериев информативности с поправкой на переобучение для поиска логических закономерностей в зашумлённых данных.

Для поиска логических закономерностей в данных обычно используются критерии информативности, основанные на теории информации (энтропийный критерий), статистических тестах (точный тест Фишера, χ^2, ω^2), или чисто эвристических соображениях (индекс Джини, D-критерий Донского). Все эти критерии не учитывают переобучение, которое может возникать вследствие оптимизации параметров логического правила по критерию информативности на конечной обучающей выборке. Величина переобучения существенно зависит от свойств выборки, поэтому попытки «встроить» в критерий информативности оценки обобщающей способности, не зависящие от выборки (например, VC-оценки [1] или оценки мета-обучения [7]), не всегда приводят к успеху.

Комбинаторная теория переобучения [8, 2] даёт более точные оценки, учитывающие эффекты расслоения и связности в конкретной выборке данных. Методика встраивания этих оценок в критерии информативности развита в [4, 5, 9]. Эксперименты показали, что использование модифицированных критериев в стандартных процедурах поиска закономерностей понижает частоту ошибок на контроле на 1–2% в различных задачах.

В данной работе показано, что модифицированный критерий приводит к более правильному отбору закономерностей в тех случаях, когда в исходных данных присутствует шум.

Задача индукции логических правил

Пусть $\mathbb{X} = (x_1, \ldots, x_L)$ — выборка объектов, описанных n признаками, $x_i = (x_i^1, \ldots, x_i^n) \in \mathbb{R}^n$, и каждому объекту x_i соответствует ответ y_i из заданного конечного множества $\mathbb{Y}$.

Алгоритмы классификации $a\colon \mathbb{X} \to \mathbb{Y}$, основанные на взвешенном голосовании логических правил, имеют следующий вид:

$$a(x) = \arg\max_{y \in \mathbb{Y}} \sum_{r \in R_y} w_r r(x),$$

где w_r — вес правила r, обычно неотрицательный, R_y — множество правил класса y. В общем случае *правило* — это функция вида $r\colon \mathbb{X} \to \{0, 1\}$ из некоторого фиксированного параметрического семейства R. В данной работе рассматривается один из наиболее распространённых типов правил — семейство конъюнкций пороговых предикатов:

$$r(x;\theta) = \prod_{j \in J} \big[x^j \leqslant \theta^j\big], \qquad (1)$$

где $\theta = (\theta^1, \ldots, \theta^n) \in \mathbb{R}^n$ — вектор порогов, θ^j — порог по j-му признаку, $J \subseteq \{1, \ldots, n\}$ — подмножество признаков, $\leqslant_j$ — одна из операций сравнения $\{\leqslant, \geqslant\}$.

Говорят, что правило r *выделяет* объект x, если $r(x) = 1$. Предполагается, что правила класса y должны выделять как можно больше объектов класса y и как можно меньше объектов всех остальных классов. Поэтому для поиска (*индукции*) правил класса y по обучающей выборке $X \subset \mathbb{X}$ решается задача двухкритериальной оптимизации:

$$p(r, X) = \big|\{x_i\colon r(x_i) = 1,\ y_i = y\}\big| \to \max_r;$$
$$n(r, X) = \big|\{x_i\colon r(x_i) = 1,\ y_i \neq y\}\big| \to \min_r;$$

На практике оптимизируют некоторый критерий информативности $H(p, n)$. Известно множество эвристических критериев, однако ни один из них не является безусловно предпочтительным [7]. Большинство критериев оценивают степень неслучайности разбиения обучающей выборки X на два подмножества (положительные примеры $x\colon r(x) = 1$ и отрицательные $x\colon r(x) = 0$) относительно исходного разбиения выборки X на классы. Чем сильнее отличается отношение $p : n$ от исходной пропорции $P : N$ числа объектов в классах, $P(X) = \big|\{x_i \in X\colon y_i = y\}\big|$, $N(X) = \big|\{x_i \in X\colon y_i \neq y\}\big|$, тем выше информативность.

В результате оптимизации p и n по обучающей выборке X соответствующие величины $p' = p(r, \bar{X})$ и $n' = n(r, \bar{X})$ на контрольной выборке $\bar{X} = \mathbb{X} \backslash X$ уже не будут оптимальны. Поэтому предлагается оценивать (p', n') по обучающей выборке и подставлять эти оценки в критерий информативности $H(p', n')$, не меняя механизм поиска закономерностей. Поскольку модифицируется только критерий, данное решение легко встраивается в стандартные методы индукции правил.

Работа поддержана РФФИ (проекты №10-07-00422-а, №11-07-00480-а) и программой ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

Вероятность переобучения правил

Каждое правило $r \in R$ класса y индуцирует на выборке $\mathbb{X}$ бинарный вектор ошибок $(r_1, \ldots, r_L)$, где $r_i = \big[r(x_i) \neq [y_i{=}y]\big]$. Обозначим через R^* множество L-мерных бинарных векторов ошибок, индуцированных всевозможными правилами из R. Введём на множествах R^* и R отношение порядка $(r \leqslant r') \leftrightarrow (r_i \leqslant r'_i,\ i = 1, \ldots, L)$, хэммингово расстояние $\rho(r, r') = \sum\limits_{i=1}^{L} |r_i - r'_i|$ и отношение предшествования $(r \prec r') \leftrightarrow (r \leqslant r') \wedge \rho(r, r') = 1$.

Число ошибок правила r на выборке $X \subseteq \mathbb{X}$ есть

$$m(r, X) = \sum_{x_i \in X} r_i = n(r, X) + P(X) - p(r, X).$$

Частота ошибок правила r на выборке X есть

$$\nu(r, X) = m(r, X)/|X|.$$

Методом обучения называется отображение вида $\mu\colon 2^{\mathbb{X}} \to R$, которое произвольной обучающей выборке $X \subseteq \mathbb{X}$ ставит в соответствие некоторое правило $r = \mu(X)$ из R. Метод μ называется *минимизацией эмпирического риска* (МЭР), если

$$\mu(X) = \arg\min_{r \in R} m(r, X).$$

Метод МЭР часто используется для обучения алгоритмов классификации. Однако для обучения закономерностей на практике используется максимизация различных критериев информативности

$$\mu_H(X) = \arg\max_{r \in R} H\big(p(r, X), n(r, X)\big),$$

которая совпадает с МЭР лишь при $H(p, n) = p - n$.

Определение 1. *Метод обучения μ называется монотонным, если $\mu(X) = \arg\min\limits_{r \in R} K(r, X)$, где критерий $K(r, X)$ является строго монотонной функцией r, то есть для любых $X \subset \mathbb{X}$ и r, r', таких, что $r < r'$, справедливо $K(r, X) < K(r', X)$.*

Утверждение 1. *Если функция $H(p, n)$ строго монотонно возрастает по p и строго монотонно убывает по n, то $K(r, X) = -H\big(p(r, X), n(r, X)\big)$ является монотонным критерием, метод максимизации информативности также является монотонным.*

Предположим, что все C_L^ℓ разбиений исходной выборки $\mathbb{X} = X \sqcup \bar{X}$ на обучающую X длины ℓ и контрольную $\bar{X}$ длины $k = L - \ell$ равновероятны. Определим для любого $\varepsilon \in (0, 1)$ функционал *вероятности переобучения* метода μ на выборке $\mathbb{X}$:

$$Q_\varepsilon(\mu, \mathbb{X}) = \mathsf{P}\big[\nu(\mu(X), \bar{X}) - \nu(\mu(X), X) > \varepsilon\big], \quad (2)$$

где знак вероятности можно понимать как среднее по всем разбиениям: $\mathsf{P} \equiv \frac{1}{C_L^\ell} \sum\limits_X$.

Верхней связностью $q(r)$ правила r называется число объектов $x \in \mathbb{X}$, на которых правило r не ошибается, при том, что существует правило $r' \in R,\ r \prec r'$, ошибающееся на x.

Неоптимальностью $s(r)$ правила r называется число объектов $x \in \mathbb{X}$, на которых правило r ошибается, при том, что существует правило $r' \in R,\ r' < r$, не ошибающееся на x. Неоптимальность не превышает числа ошибок правила $s(r) \leqslant m(r, \mathbb{X})$, и в точности равна ему, когда в R содержится корректное правило (не ошибающееся ни на одном объекте).

Определим для всех $m = 0, \ldots, L,\ z = 0, \ldots, \ell$ функцию гипергеометрического распределения

$$H_L^{\ell,\, m}(z) = \sum_{s=0}^{\lfloor z \rfloor} \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}.$$

Теорема 1 (оценка расслоения–связности). *Для любого монотонного метода обучения μ и любого $\varepsilon \in (0, 1)$*

$$Q_\varepsilon(\mu, \mathbb{X}) \leqslant \sum_{r \in R^*} P_r H_{L-q-s}^{\ell-q,\, m-s}\left(\tfrac{\ell}{L}(m - \varepsilon k)\right), \quad (3)$$

где $m{=}m(r, \mathbb{X})$, $q{=}q(r)$, $s{=}s(r)$, $P_r = C_{L-q-s}^{\ell-q}/C_L^\ell$ есть верхняя оценка вероятности получить правило r в результате обучения.

Эта оценка была получена в [9] только для МЭР. Здесь мы показываем, что она справедлива для широкого класса *монотонных* методов обучения.

Эффективный алгоритм вычисления оценки (3) предложен в [9]. Он основан на переборе неэквивалентных правил по слоям снизу вверх. *Слоем* называется подмножество правил с фиксированным числом ошибок $m = m(r, \mathbb{X})$. Поскольку вклады слоёв в оценку экспоненциально убывают по s (следовательно, и по m), можно ограничиться перебором только небольшого числа нижних слоёв.

Критерий информативности с поправкой на переобучение

Оценка (3) остаётся в силе независимо от того, как определён индикатор ошибки. Рассмотрим два альтернативных способа определения:

$$r_i^p = \big[r(x_i) = 0\big]\big[y_i = y\big];$$
$$r_i^n = \big[r(x_i) = 1\big]\big[y_i \neq y\big].$$

Число ошибок, соответственно, будет равно

$$m_p(r, X) = \sum_{x_i \in X} r_i^p = P(X) - p(r, X);$$
$$m_n(r, X) = \sum_{x_i \in X} r_i^n = n(r, X).$$

Заметим, что $r_i = r_i^p + r_i^p$ и $m(r, X) = m_p(r, X) + m_n(r, X)$. Таким образом, вводится разделение ошибок на ошибки I и II рода.

Вычислим оценки (3) для обоих родов ошибок. Приравнивая их заданному уровню значимости $\eta \in (0,1)$ и выражая ε через η путём численного обращения функции, получим две оценки, справедливые для правила $r = \mu(X)$ с вероятностью $1-\eta$:

$$\tfrac{1}{k} p(r, \bar{X}) \geqslant \tfrac{1}{\ell} p(r, X) - \varepsilon_p(\eta),$$
$$\tfrac{1}{k} n(r, \bar{X}) \leqslant \tfrac{1}{\ell} n(r, X) + \varepsilon_n(\eta).$$

Подставляя эти оценки в критерий информативности H, получим модифицированный критерий

$$H'(p,n) = H\big(p - \ell\varepsilon_p(\eta), n + \ell\varepsilon_n(\eta)\big).$$

Поправочные слагаемые $\ell\varepsilon_p(\eta)$, $\ell\varepsilon_p(\eta)$ зависят от выборки, но не зависят от переменных p и n, поэтому максимизация модифицированного критерия не выводит за класс монотонных методов обучения. Фактически, мы выбираем в классе методов обучения, для которых верна оценка (3), такой метод, который максимизирует предсказанную информативность на контрольных данных. Это полностью соответствует общепринятой методике применения оценок обобщающей способности [1, 6].

Наряду с обращением оценки (3) в экспериментах проводилось также обращение эмпирической оценки функционала $Q_\varepsilon(\mu, \mathbb{X})$, вычисленной методом Монте-Карло по 200 случайным разбиениям.

Эксперимент на модельных данных

Возьмем две двумерных ($n = 2$) двухклассовых ($\mathbb{Y} = \{0,1\}$) сбалансированных ($P = N = L/2$) модельных выборки длины $L = 200$. Будем разбивать их на обучение и контроль поровну, $\ell = k = 100$. В семействе правил (1) фиксируем операцию сравнения $\leqslant$ и будем строить только правила класса 1. Для поиска правил используем гипергеометрический критерий информативности (точный тест Фишера) $H(p,n) = -\ln\big(C_P^p C_N^n / C_{P+N}^{p+n}\big)$.

Выборки сгенерируем следующим образом.

В первой выборке число ошибок лучшего правила равно 40. Граница между классами четкая, поскольку все 40 шумовых объектов расположены вдали от границы классов, рис. 1.

Во второй выборке число ошибок лучшего правила 23, т. е. почти вдвое меньше, шумовые объекты расположены вдоль границы классов, рис. 2.

Хотя во второй выборке существует правило, почти вдвое лучшее, чем в первой, найти его по обучающей подвыборке практически невозможно из-за переобучения. Значения критерия информативности на обучающей выборке почти одинаковы (незначимое предпочтение отдаётся второй выборке), однако на контроле найденное правило оказывается значимо лучше для первой выборки:

выборка:	1	2
обучение	40,29	40,47
контроль	39,13	34,46

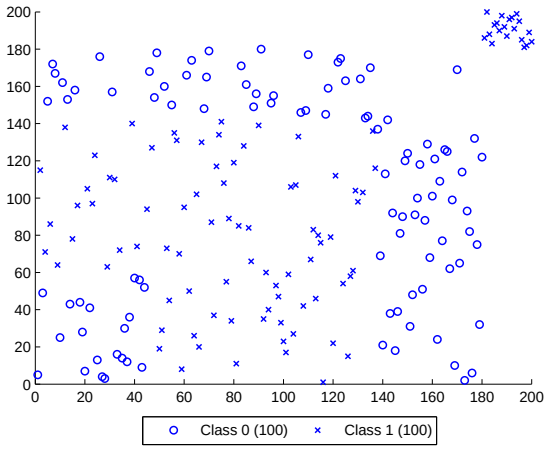


Рис. 1. Выборка 1 в осях двух признаков.

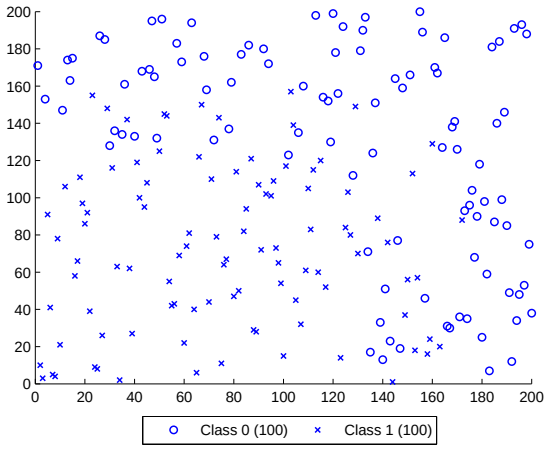


Рис. 2. Выборка 2 в осях двух признаков.

Причина в том, что у второй выборки существенно менее выражено расслоение, мощности нижних слоёв очень быстро возрастают, рис. 3.

Возникает вопрос — возможно ли с помощью комбинаторных оценок, использующих информацию о расслоении и связности на обучающей выборке, предсказать информативность на скрытой контрольной выборке? Следующая таблица даёт утвердительный ответ. Она содержит значения модифицированного критерия H' на обучении.

η	оценка H'	Монте-Карло		рассл.-связн.	
	выборка:	1	2	1	2
0,1	обучение	24,44	16,29	23,23	13,42
	контроль	39,13	34,06	39,13	34,06
0,2	обучение	28,07	22,13	27,11	20,58
	контроль	39,13	34,06	39,13	32,68
0,3	обучение	32,28	28,28	30,45	25,74
	контроль	39,13	34,43	39,13	34,40
0,4	обучение	34,35	32,98	33,39	27,59
	контроль	39,13	34,78	39,13	34,40
0,5	обучение	37,63	36,12	34,93	31,21
	контроль	39,13	34,79	39,13	34,40

Предсказанная информативность H' всегда меньше для второй выборки. Таким образом, нам действительно удаётся предсказать, что найти хорошее правило при сильно зашумлённой границе классов невозможно из-за переобучения.

Предсказанная информативность H', вычисляемая по оценке расслоения–связности, всегда пессимистично занижена, поскольку завышена оценка $p(r, \bar{X})$ и занижена оценка $n(r, \bar{X})$. Заниженность информативности составляет в данной задаче 1–2 для первой выборки и 3–5 для второй. Это согласуется с теоретическим выводом, что оценка расслоения–связности менее точна для выборок с менее выраженным расслоением. При поиске закономерностей такая смещённость оценки является благоприятной — чем слабее закономерность, тем сильнее занижена предсказанная информативность.

Заметим также, что оценка расслоения-связности вычисляется существенно быстрее, чем эмпирическая оценка по методу Монте-Карло.

Рассмотрим теперь вклад каждого слоя в оценку расслоения–связности (3). График зависимости накопленного значения P_r от номера слоя представлен на рисунках 4 и 5. Видно, что основной вклад в сумму (3) вносят слои с малым числом ошибок.

Если бы оценки вероятностей P_r были точными, их сумма равнялась бы единице. Накопленное значение P_r в крайней правой точке графика позволяет судить о степени завышенности оценок расслоения–связности. Для выборки 1 с «хорошим расслоением» оценка завышена всего в 2 раза; для выборки 2 с «плохим расслоением» — в 40 раз.

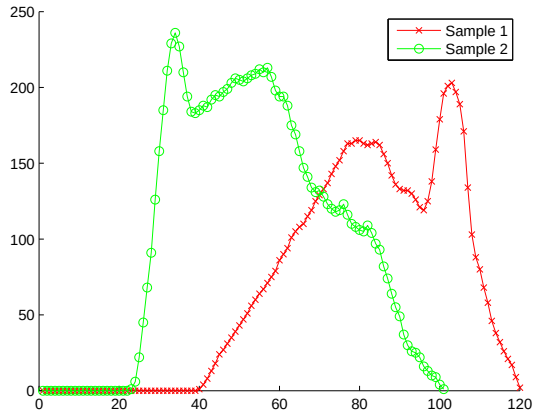


Рис. 3. Зависимость числа алгоритмов в слое от номера слоя $m(r, \mathbb{X})$ для двух модельных выборок.

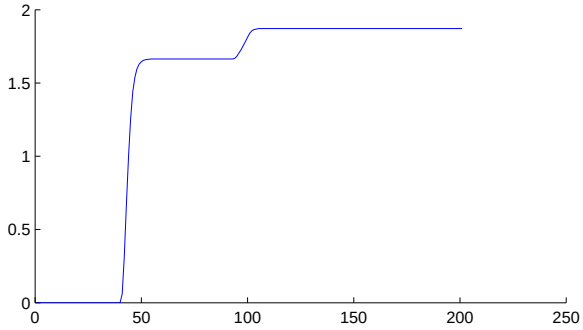


Рис. 4. Зависимость накопленного значения P_r от номера слоя $m(r, \mathbb{X})$ для выборки 1.

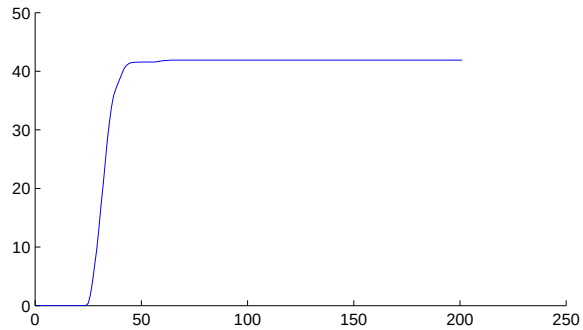


Рис. 5. Зависимость накопленного значения P_r от номера слоя $m(r, \mathbb{X})$ для выборки 2.

Литература

[1] *Вапник В Н., Червоненкис А. Я.* Теория распознавания образов. — М.: Наука, 1974.

[2] *Воронцов К. В.* Комбинаторная теория переобучения: результаты, приложения и открытые проблемы // ММРО-15, 2011. — С. 40–43.

[3] *Донской В. И., Башта А. И.* Дискретные модели принятия решений при неполной информации. — Симферополь: Таврия, 1992. — 166 с.

[4] *Ивахненко А. А.* О вероятности переобучения пороговых конъюнкций // Докл. межд. конф. Интеллектуализация обработки информации, ИОИ-8. — М.: МАКС Пресс, 2010. — С. 57–60.

[5] *Ивахненко А. А.* Точная верхняя оценка вероятности переобучения для корректных логических правил // Труды МФТИ. — Том 2, № 3 — М.: МФТИ., 2010. — С. 16–22.

[6] *Boucheron S., Bousquet O., Lugosi G.* Theory of classification: A survey of some recent advances // ESAIM: Probability and Statistics. — 2005. — no. 9. — Pp. 323–375.

[7] *Fürnkranz J., Flach P. A.* ROC 'n' rule learning — towards a better understanding of covering algorithms // Machine Learning, 2005. — Vol. 58, no. 1. — Pp. 39–77.

[8] *Vorontsov K. V.* Exact combinatorial bounds on the probability of overfitting for empirical risk minimization // Patt. Rec. and Image Anal., 2010. — Vol. 20, No. 3. — Pp. 269–285.

[9] *Vorontsov K. V., Ivahnenko A. A.* Tight combinatorial generalization bounds for threshold conjunction rules // 4-th Int'l Conf. on Pattern Recognition and Machine Intelligence (PReMI'11), June 27 – July 1, 2011. Lecture Notes in Computer Science. Springer-Verlag, 2011. — Pp. 66–73.

Комбинаторные оценки вероятности отклонения тестовой ошибки от ошибки скользящего контроля

Животовский Н. К.
nikita.zhivotovskiy@phystech.edu
Московский физико-технический институт (государственный университет)

В рамках комбинаторной теории переобучения получены оценки вероятности большого отклонения частоты ошибок на тестовой выборке от оценки скользящего контроля LOO. Они улучшают известные оценки «здравого смысла» (sanity-check bounds) благодаря учёту эффектов расслоения и связности.

На практике для оценивания обобщающей способности обучаемых алгоритмов часто используется оценка скользящего контроля с одним отделяемым объектом LOO (leave-one-out cross-validation). Для её вычисления каждый из объектов выборки длины ℓ по очереди отделяется в качестве контрольного, и обучение производится по оставшимся $\ell - 1$ объектам. Оценка LOO определяется как средняя ошибка на всех отделённых контрольных объектах. Известно, что LOO является несмещённой оценкой вероятности ошибки алгоритма. Эксперименты показывают, что эта оценка даёт более объективное представление о вероятности ошибки по сравнению со смещённой (оптимистично заниженной) частотой ошибок на обучении. Несмотря на широкое применение оценки LOO, её известные теоретические обоснования пока либо привязаны к конкретному методу обучения [6], либо представляют собой сильно завышенные оценки вероятности большого отклонения вероятности ошибки от оценки LOO [7, 4].

В статье [7] представлены так называемые *оценки «здравого смысла»* (sanity-check bounds). Термин пришёл из программирования, где он означает простую и быструю, но довольно грубую проверку общей работоспособности системы. Как правило, они показывают лишь то, что LOO даёт оценку вероятности ошибки не хуже, чем частота ошибок на обучении. Основной недостаток этих оценок — чрезмерная завышенность, вызванная, в частности, использованием VC-оценок из теории равномерной сходимости [2]. Кроме того, практически все известные оценки требуют сильного дополнительного предположения об *устойчивости* (stability) метода обучения. Неформально, оно означает, что вероятность ошибки алгоритма, выбираемого методом обучения, изменится не сильно, если из обучающей выборки исключить один объект.

Дисперсию оценки LOO, которая могла бы дать информацию о её точности, оказалось довольно трудно посчитать. Верхние оценки дисперсии выведены в [6, 5], а в [4] доказано, что в общем случае для LOO не существует несмещенной оценки дисперсии. В данной работе будет оцениваться не дисперсия, а функция распределения отклонения LOO от частоты ошибок на независимой тестовой выборке. Функция распределения дает более полную информацию об оценке LOO, чем дисперсия.

В работе используется комбинаторный подход [11], основанный на единственном вероятностном предположении, что все разбиения конечной генеральной выборки на две подвыборки — наблюдаемую обучающую и скрытую контрольную — равновероятны. Это предположение эквивалентно стандартному предположению о независимости элементов выборки. При этом оценка LOO вычисляется только по обучающей выборке.

Комбинаторный подход позволяет учесть эффекты *расслоения* и *связности*, отказавшись от гипотезы устойчивости метода обучения, роль которой фактически играет гипотеза *связности*.

Ещё одно преимущество комбинаторного подхода в том, что он позволяет сравнивать оценки эмпирически с помощью метода Монте-Карло как на модельных, так и на реальных данных [9, 10]. Так, в данной работе будут сравниваться четыре оценки вероятности большого отклонения LOO от частоты ошибок на тестовой выборке: точная эмпирическая оценка, комбинаторная оценка «здравого смысла», и две предлагаемые в данной работе оценки расслоения-связности.

Определения и обозначения

Будем использовать основные определения из комбинаторной теории переобучения. Пусть задано конечные множества *объектов* $\mathbb{X} = \{x_1, \dots, x_L\}$ и *алгоритмов* $A = \{a_1, \dots, a_D\}$ и существует бинарная *функция потерь* $I\colon A \times \mathbb{X} \to \{0, 1\}$, где $I(a, x) = 1$ тогда и только тогда, когда алгоритм a допускает ошибку на объекте x.

Вектором ошибок алгоритма a называется L-мерный бинарный вектор $\big(I(a, x_1), \dots, I(a, x_L)\big)$. Будем предполагать, что векторы ошибок всех алгоритмов из A попарно различны.

Число ошибок алгоритма a на выборке $X \in \mathbb{X}$ определяется как $n(a, X) = \sum\limits_{x \in X} I(a, x)$.

Частота ошибок алгоритма a на выборке X определяется как $\nu(a, X) = n(a, X)/|X|$.

Методом обучения называется отображение $\mu\colon 2^{\mathbb{X}} \to A$, которое произвольной выборке $X \subset \mathbb{X}$ ставит в соответствие некоторый алгоритм $\mu X \in A$.

Нас в первую очередь будет интересовать метод *минимизации эмпирического риска*:

$$\mu X = \arg\min_{a\in A} n(a, X), \quad X \subset \mathbb{X}.$$

Оценка *скользящего контроля* LOO определяется по формуле

$$\texttt{LOO}(\mu, X) = \frac{1}{|X|}\sum_{x\in X} I\big(\mu(X\backslash x), x\big).$$

Будем считать, что все C_L^ℓ разбиений генеральной выборки $\mathbb{X} = X \sqcup \bar{X}$ на две выборки — *наблюдаемую обучающую* X длины ℓ и *скрытую тестовую* $\bar{X}$ длины $k = L - \ell$ — равновероятны.

Отметим важное терминологическое отличие: *тестовой* называется выборка $\bar{X}$ длины k, *контрольными* называются одноэлементные подвыборки обучающей выборки, используемые для вычисления LOO.

Нашей целью является получение двусторонних оценок отклонения частоты ошибок на тестовой подвыборке от оценки LOO: для любого $\varepsilon \in (0, 1)$

$$\bar{Q}_\varepsilon^{\texttt{LOO}}(\mu, \mathbb{X}) = \mathsf{P}\big[\nu(\mu X, \bar{X}) - \texttt{LOO}(\mu, X) \geqslant \varepsilon\big];$$
$$\underline{Q}_\varepsilon^{\texttt{LOO}}(\mu, \mathbb{X}) = \mathsf{P}\big[\texttt{LOO}(\mu, X) - \nu(\mu X, \bar{X}) \geqslant \varepsilon\big];$$

Комбинаторные VC-оценки вероятности переобучения

В комбинаторной теории переобучения [11] одной из основных задач является получение точных либо верхних оценок *вероятности переобучения*

$$\bar{Q}_\varepsilon(\mu, \mathbb{X}) = \mathsf{P}\big[\delta(\mu X, X) \geqslant \varepsilon\big], \quad \varepsilon \in (0, 1),$$

где величина отклонения частоты ошибок на контроле и обучении $\delta(a, X) = \nu(a, \bar{X}) - \nu(a, X)$ называется *переобученностью* алгоритма a.

Таким образом, $\bar{Q}_\varepsilon$ есть вероятность того, что наблюдаемая частота ошибок на обучении $\nu(\mu X, X)$ *недооценивает* скрытую частоту ошибок на тесте $\nu(\mu X, \bar{X})$ на ε или более.

Нас также будет интересовать вероятность $\underline{Q}_\varepsilon$ того, что наблюдаемая частота $\nu(\mu X, X)$ *переоценивает* скрытую частоту $\nu(\mu X, \bar{X})$ на ε или более:

$$\underline{Q}_\varepsilon(\mu, \mathbb{X}) = \mathsf{P}\big[\nu(\mu X, X) - \nu(\mu X, \bar{X}) \geqslant \varepsilon\big].$$

Определим функцию гипергеометрического распределения

$$H_L^{\ell,\,m}(z) = \sum_{s=0}^{\lfloor z\rfloor} \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}.$$

Теорема 1 (VC-оценка). *Для любых μ, $\mathbb{X}$ и любого $\varepsilon \in (0, 1)$ справедливы оценки*

$$\bar{Q}_\varepsilon(\mu, \mathbb{X}) \leqslant D \max_{m=1,\ldots,L} H_L^{\ell,\,m}\left(\tfrac{\ell}{L}(m - \varepsilon k)\right);$$
$$\underline{Q}_\varepsilon(\mu, \mathbb{X}) \leqslant D \max_{m=1,\ldots,L} H_L^{k,\,m}\left(\tfrac{k}{L}(m - \varepsilon \ell)\right).$$

Правые части этих неравенств обозначим, соответственно, $\bar{\eta}(\varepsilon)$ и $\underline{\eta}(\varepsilon)$.

Оценки «здравого смысла»

Определение 1. *Если для метода обучения μ и семейства алгоритмов A выполнено*

$$\mathsf{P}_{X,x}[\nu(\mu X, \bar{X}) - \nu(\mu(X\backslash x), \bar{X}) \geqslant \beta_2] \leqslant \beta_1$$

для некоторых неотрицательных β_1, β_2, то говорят, что метод обучения (β_1, β_2)-устойчив.

Здесь и далее знак вероятности $\mathsf{P}_{X,x}$ означает, что равновероятными полагаются все разбиения генеральной выборки $\mathbb{X}$ на три подвыборки — обучающую $X\backslash x$ длины $\ell - 1$, контрольную, состоящую из одного объекта x, и тестовую $\bar{X}$ длины k.

Следующая оценка является комбинаторным аналогом оценки «здравого смысла» [7].

Теорема 2 (sanity-check bound). *Пусть метод минимизации эмпирического риска μ является (β_1, β_2)-устойчивым. Тогда для любого $\gamma \in [0, 1]$*

$$\mathsf{P}[\nu(\mu(X), \bar{X}) - \texttt{LOO}(\mu, X) > \varepsilon] \leqslant \frac{3\beta_1 + \beta_2 + \underline{\eta}(\gamma) + \gamma}{\varepsilon}.$$

Данная оценка может быть сильно завышенной (и в эксперименте мы покажем, что это действительно так) по следующим причинам:

— в её основе лежат VC-оценки, которые сильно завышены [3];
— наличие параметра ε, близкого к нулю, в знаменателе правой части неравенства;
— достаточно жесткое требование устойчивости, которое при $3\beta_1 + \beta_2 \geqslant 1$ делает оценку тривиальной.

Аналогично получается и вероятность того, что LOO переоценивает частоту ошибок на тестовой подвыборке.

Определение 2. *Если для метода обучения μ и семейства алгоритмов A выполнено*

$$\mathsf{P}_{X,x}[\nu(\mu(X\backslash x), \bar{X}) - \nu(\mu X, \bar{X}) \geqslant \beta_2] \leqslant \beta_1$$

для некоторых неотрицательных β_1, β_2, то говорят, что метод обучения $(\beta_1, \beta_2)^*$-устойчив.

Теорема 3 (sanity-check bound). *Пусть метод минимизации эмпирического риска μ является $(\beta_1, \beta_2)^*$-устойчивым. Тогда для любого $\gamma \in [0, 1]$*

$$\mathsf{P}[\texttt{LOO}(\mu, X) - \nu(\mu(X), \bar{X}) > \varepsilon] \leqslant \frac{3\beta_1 + \beta_2 + \bar{\eta}(\gamma) + \gamma}{\varepsilon}.$$

Полученные комбинаторные аналоги оценок «здравого смысла» будут использоваться в экспериментах в качестве эталона для сравнения с более точными оценками.

Оценка расслоения–связности для вероятности переобучения

Введём на множестве алгоритмов A естественное отношение порядка и метрику Хэмминга:

$$a \leqslant b \;\leftrightarrow\; I(a,x) \leqslant I(b,x),\ \forall x \in \mathbb{X};$$
$$a < b \;\leftrightarrow\; a \leqslant b \;\text{ и }\; a \neq b;$$
$$\rho(a,b) = \textstyle\sum_{i=1}^{L} \big[I(a,x_i) \neq I(b,x_i)\big].$$

Если $a < b$ и при этом $\rho(a,b) = 1$, то будем говорить, что a *предшествует* b и записывать $a \prec b$. Очевидно, что $n(a,\mathbb{X}) + 1 = n(b,\mathbb{X})$.

Графом расслоения–связности множества алгоритмов A будем называть направленный граф $\langle A, E\rangle$ с множеством рёбер $E = \big\{(a,b)\colon a \prec b\big\}$.

Граф расслоения–связности является многодольным, доли соответствуют *слоям алгоритмов* $A_m = \big\{a \in A\colon n(a,\mathbb{X}) = m\big\}$, рёбрами могут соединяться только алгоритмы соседних слоёв. Каждому ребру (a,b) соответствует единственный объект $x_{ab} \in \mathbb{X}$, такой, что $I(a,x_{ab}) = 0$ и $I(b,x_{ab}) = 1$.

Порождающим множеством X_a алгоритма a будем называть множество объектов, соответствующих исходящим из вершины a ребрам:

$$X_a = \big\{x \in \mathbb{X} \;\big|\; \exists b \in A\colon a \prec b,\ I(a,x) < I(b,x)\big\}.$$

Верхней связностью алгоритма a будем называть число рёбер графа, исходящих из вершины a:

$$q(a) = |X_a| = \#\big\{x_{ab} \in \mathbb{X} \;\big|\; a \prec b\big\}.$$

Неоптимальностью $r(a)$ алгоритма a будем называть число объектов $x \in \mathbb{X}$, на которых алгоритм a ошибается, при том, что существует алгоритм $b \in A$, $b < a$, не ошибающийся на x:

$$r(a) = \#\big\{x \in \mathbb{X} \;\big|\; \exists b \in A\colon b < a,\ I(b,x) < I(a,x)\big\}.$$

Для вероятности переобучения в [11] получена верхняя оценка, существенно зависящая от характеристик $q(a)$, $r(a)$ каждого алгоритма $a \in A$.

Теорема 4 (оценка расслоения–связности). *Пусть μ — метод минимизации эмпирического риска. Тогда для любого $\varepsilon \in (0,1)$*

$$\bar{Q}_\varepsilon(\mu,\mathbb{X}) \leqslant \sum_{a\in A} \frac{C_{L-q-r}^{\ell-q}}{C_L^\ell} H_{L-q-r}^{\ell-q,\,m-r}\left(\tfrac{\ell}{L}(m-\varepsilon k)\right),$$

где q — верхняя связность, r — неоптимальность алгоритма a, $m = n(a,\mathbb{X})$.

Чем выше связность $q(a)$ и неоптимальность $r(a)$, тем меньше вклад алгоритма a в вероятность переобучения. Поскольку отношение биномиальных коэффициентов убывает экспоненциально по r, существенный вклад в оценку дают только алгоритмы нескольких нижних слоёв.

Если в этой оценке пренебречь расслоением и связностью, положив $r = q = 0$ для каждого $a \in A$, то получится VC-оценка.

Оценка расслоения–связности для отклонения тестовой ошибки от LOO

Основным результатом данной работы являются оценки вероятности отклонения частоты ошибок на тестовой ошибке от оценки LOO, учитывающие свойства расслоения и связности.

Утверждение 1. *Для метода μ минимизации эмпирического риска и для любой обучающей выборки X имеет место неравенство*

$$\mathtt{LOO}(\mu,X) \geqslant \nu(\mu X, X).$$

Теорема 5. *Для метода μ минимизации эмпирического риска имеет место неравенство*

$$\bar{Q}_\varepsilon^{\mathtt{LOO}}(\mu,\mathbb{X}) \leqslant \bar{Q}_\varepsilon(\mu,\mathbb{X}).$$

Таким образом, верхняя оценка вероятности переобучения $\bar{Q}_\varepsilon$ справедлива также и для $\bar{Q}_\varepsilon^{\mathtt{LOO}}$.

Далее будет показано, что для $\bar{Q}_\varepsilon^{\mathtt{LOO}}$ справедлива более точная оценка расслоения-связности.

При минимизации эмпирического риска по выборке X может оказаться, что в семействе A существует много алгоритмов a с минимальным числом ошибок $n(a,X)$. Поэтому предлагается следующая гипотеза об устойчивости метода обучения по отношению к удалению одного обучающего объекта.

Гипотеза 1 (устойчивости). Для любого разбиения генеральной выборки $\mathbb{X} = X \sqcup \bar{X}$ и любого объекта $x \in X_{\mu X}$ выполняется $I\big(\mu(X\backslash x), x\big) = 1$.

Данная гипотеза выполнена, в частности, для *монотонных сетей* алгоритмов, если в качестве метода обучения выбрана *пессимистичная минимизация* эмпирического риска [1, 10, 11].

Теорема 6. *Пусть μ — метод минимизации эмпирического риска, удовлетворяющий гипотезе устойчивости. Тогда для любого $\varepsilon \in (0,1)$*

$$\bar{Q}_\varepsilon^{\mathtt{LOO}} \leqslant \sum_{a\in A} \frac{C_{L-q-r}^{\ell-q}}{C_L^\ell} H_{L-q-r}^{\ell-q,\,m-r}\left(\tfrac{\ell}{L}(m-\varepsilon k) - \tfrac{k}{L}q\right);$$

где q — верхняя связность, r — неоптимальность алгоритма a, $m = n(a,\mathbb{X})$.

Полученная оценка не превышает оценку расслоения-связности для вероятности переобучения (Теорема 4). Единственное отличие этих оценок — в аргументе функции гипергеометрического распределения. Чем выше связность, тем существеннее это отличие. Если пренебречь связностью ($q = 0$), то эти оценки совпадут.

Обе оценки являются достижимыми и обращаются в равенства в случае многомерных монотонных сетей алгоритмов, введённых в [1].

Вероятность переоценивания

Чтобы получить оценку величины $\underline{Q}_\varepsilon^{\text{LOO}}$, определим для данного семейства алгоритмов A подмножество *неулучшаемых алгоритмов*

$$A^* = \big\{a \in A \;\big|\; \nexists b \in A\colon b \prec a\big\}.$$

Теорема 7. *Пусть μ — метод минимизации эмпирического риска, удовлетворяющий гипотезе устойчивости. Тогда для любого $\varepsilon \in (0,1)$*

$$\underline{Q}_\varepsilon^{\text{LOO}} \leqslant \sum_{a\in A^*} \frac{C_{L-q}^k}{C_L^\ell} H_{L-q}^{k,\,m}\left(\tfrac{k}{L}(m+q-\varepsilon\ell)\right),$$

где q — верхняя связность, $m = n(a, \mathbb{X})$.

Эксперимент на модельных данных

Модельное семейство алгоритмов определяется тремя параметрами $m, H, h \in \{0,\ldots,L\}$ и параметризуется двумя индексами,

$$A = \big\{a_{sp}\colon s = 0,\ldots,H,\ p = 0,\ldots,\lfloor (H-s)/h\rfloor\big\};$$

векторы ошибок задаются следующим образом:

$$I(x_i, a_{sp}) = [i \leqslant s] \vee [i = s + hp] \vee [H < i \leqslant H + m],$$

для всех $i = 1,\ldots,L$. Таким образом, имеются m объектов, на которых ошибаются все алгоритмы, $L - m - H$ объектов, на которых не ошибается ни один алгоритм, на остальных H объектах возможны от 0 до H ошибок. Число непустых слоёв равно $H + 1$, верхняя связность практически всех алгоритмов равна $h - 1$, алгоритм с нулевой верхней связностью всего один.

В нашем эксперименте $H = 35$, $m = 5$, $h = 3$. Длина генеральной выборки $L = 150$, $k = \ell = 75$. Эмпирические «точные» оценки вычислялись методом Монте-Карло по 1000 случайных разбиений. Оценка «здравого смысла» (по Теореме 2) минимизировалась по параметрам γ и $3\beta_1 + \beta_2$.

Рис. 1 показывает, что оценка расслоения-связности завышена, но не так сильно, как оценка «здравого смысла», которая фактически вырождается в тривиальную. Аналогичный вывод справедлив и для вероятности переоценивания, см. рис. 2.

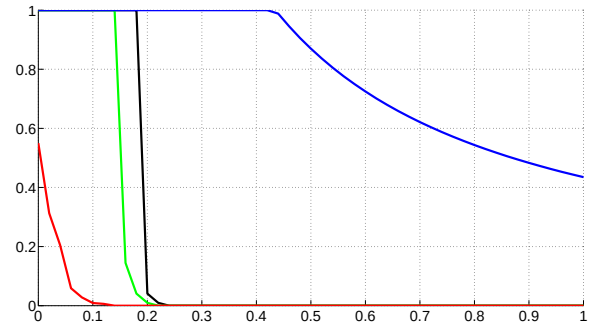


Рис. 1. Зависимость $\bar{Q}_\varepsilon^{\text{LOO}}$ от ε, слева направо: «точная» оценка; оценка расслоения-связности (по Теореме 6), оценка вероятности переобучения (по Теореме 4), оценка «здравого смысла» (по Теореме 2).

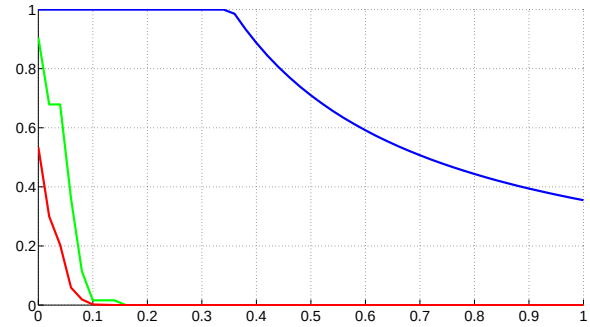


Рис. 2. Зависимость $\underline{Q}_\varepsilon^{\text{LOO}}$ от ε, слева направо: «точная» оценка; оценка расслоения-связности (по Теореме 7); оценка «здравого смысла» (по Теореме 3).

Литература

[1] *Ботов П. В.* Точные оценки вероятности переобучения для монотонных и унимодальных семейств алгоритмов // Всеросс. конф. Математические методы распознавания образов-14. — М.: МАКС Пресс, 2009. — С. 7–10.

[2] *Вапник В. Н., Червоненкис А. Я.* Теория распознавания образов. — М.: Наука, 1974.

[3] *Воронцов К. В.* Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики / Под ред. О. Б. Лупанова. — М.: Физматлит, 2004. — Т. 13. — С. 5–36.

[4] *Bengio Y., Grandvalet Y.* No unbiased estimator of the variance of k-fold cross-validation // J. of Mach. Learn. Res., 2003. — Vol. 25. — Pp. 1089–1105.

[5] *Bousquet O., Elisseeff A.* Stability and generalization // J. of Mach. Learn. Res., 2002. — No. 2. — Pp. 499–526.

[6] *Devroye L. P., Wagner T. J.* Distribution-free inequalities for the deleted and holdout error estimates // IEEE Transactions on Information Theory, 1979. — Vol 25, No. 2. — Pp. 202–207.

[7] *Kearns M., Ron D.* Algorithmic stability and sanity-check bounds for leave-one-out cross-validation // COLT, 1997. — Pp. 152–162.

[8] *Kale S., Kumar R., Vassilvitskii S.* Cross-validation and mean-square stability // Yahoo! Research, 2011.

[9] *Vorontsov K. V.* Combinatorial probability and the tightness of generalization bounds // Pattern Recognition and Image Analysis. — 2008. — Vol. 18, no. 2. — Pp. 243–259.

[10] *Vorontsov K. V.* Splitting and similarity phenomena in the sets of classifiers and their effect on the probability of overfitting // Pattern Recognition and Image Analysis. — 2009. — Vol. 19, no. 3. — Pp. 412–420.

[11] *Vorontsov K. V.* Exact combinatorial bounds on the probability of overfitting for empirical risk minimization // Pattern Recognition and Image Analysis. — 2010. — Vol. 20, no. 3. — Pp. 269–285.

Переобучение и комбинаторная рademахеровская сложность в задачах восстановления регрессии*

Каневский Д. Ю.
kanevskiy@forecsys.ru
Москва, ЗАО «Форексис»

Комбинаторная теория переобучения позволяет получать точные оценки обобщающей способности для задач классификации, но не допускает прямого обобщения на задачи восстановления регрессии с вещественной функцией потерь. Для таких задач известны оценки на основе рademахеровской сложности семейства обучаемых алгоритмов. В данной работе вводится понятие комбинаторной рademахеровской сложности и выводится оценка переобучения для задач восстановления регрессии в рамках комбинаторного подхода (слабой вероятностной аксиоматики). В модельном эксперименте полученная оценка сравнивается с «классической» и анализируются факторы завышенности обеих оценок.

Введение

Точные оценки вероятности переобучения, получаемые в рамках комбинаторной теории [1] существенно опираются на предположение о бинарности функции потерь (ошибка алгоритма может принимать только значения 0 или 1). Комбинаторные оценки не допускают прямого обобщения на случай функций потерь, принимающих значения из $\mathbb{R}$, используемых в задачах восстановления регрессии. Одним из наиболее перспективных подходов к оценке обобщающей способности, применимых к произвольным функциям потерь, являются оценки на основе *рademахеровской сложности (rademacher complexity)* [2, 5, 7]. Данная работа направлена на объединение сильных сторон обоих подходов путём переноса понятия рademахеровской сложности и основанных на нем оценок в слабую (перестановочную) вероятностную аксиоматику [1]. Для этого используется техника, предложенная в [3] для задачи трансдуктивного обучения.

Определения и обозначения

Пусть имеется выборка из L произвольных объектов $\mathbb{X} = \{x_1, \ldots, x_L\}$, которую мы будем называть *генеральной совокупностью*, и пусть каждому объекту x_i сопоставлен ответ $y_i \in \mathbb{R}$. В соответствии с комбинаторным подходом будем предполагать, что из генеральной совокупности случайно и равновероятно выбираются ℓ объектов (вместе со своими ответами), образуя наблюдаемую *обучающую выборку* X. Оставшиеся $k = L - \ell$ объектов составляют скрытую *контрольную выборку* $\bar{X}$. Таким образом, каждое из C_L^ℓ разбиений генеральной совокупности $\mathbb{X} = X \sqcup \bar{X}$ может реализоваться с равной вероятностью.

Задано множество *алгоритмов* $A = \{a\colon \mathbb{X} \to \mathbb{R}\}$, не обязательно конечное, и вещественная *функция потерь* $\mathscr{L}\colon \mathbb{R}^2 \to \mathbb{R}$, значение которой $z_i = \mathscr{L}\big(a(x_i), y_i\big)$ определяет величину *ошибки алгоритма a на объекте x_i*. Таким образом, каждому алгоритму a однозначно сопоставляется вектор ошибок $z = (z_1, \ldots, z_L)$, а множество алгоритмов A порождает множество векторов ошибок

$$Z = \big\{ z = \big(\mathscr{L}(a(x_i), y_i)\big)_{i=1}^L \;\big|\; a \in A \big\} \subseteq \mathbb{R}^L.$$

В дальнейшем вся работа будет выполняться именно с векторами ошибок, поэтому будем отождествлять алгоритмы с векторами ошибок, а множество A — с множеством Z.

Пусть фиксировано разбиение $(X, \bar{X})$. *Переобучением* алгоритма a с вектором ошибок z назовем разность средней ошибки на контроле и обучении:

$$\delta(z, X) = \frac{1}{k} \sum_{x_i \in \bar{X}} z_i - \frac{1}{\ell} \sum_{x_i \in X} z_i.$$

Методом обучения μ называется отображение, сопоставляющее произвольной обучающей выборке X некоторый алгоритм $a = \mu X$ из A.

Равномерной (верхней) оценкой переобучения будем называть величину

$$\delta(Z, X) = \sup_{z \in Z} \delta(z, X). \qquad (1)$$

Равномерная оценка переобучения априори завышена, но справедлива для всех методов обучения.

Рademахеровская сложность множества векторов ошибок Z является мерой разнообразия или «богатства» семейства алгоритмов A. Приведём классическое определение этого понятия согласно [4], затем несколько изменим его для применения в рамках комбинаторного подхода.

Определение 1. *Пусть $\sigma = (\sigma_1, \ldots, \sigma_n)$ — вектор независимых рademахеровских случайных величин: $P(\sigma_i = 1) = P(\sigma_i = -1) = \frac{1}{2}$. Классической рademахеровской сложностью множества векторов $V \subseteq \mathbb{R}^n$ называется величина*

$$R^n(V) = \frac{2}{n} \mathsf{E}_\sigma \sup_{v \in V} (\sigma^\intercal v).$$

Работа поддержана РФФИ (проект №11-07-00480) и программой ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

Смысл этой величины легко понять, если заметить, что $(\sigma^\intercal v)$ есть выборочная ковариация ошибок и случайного радемахеровского шума. Супремум выделяет вектор ошибок v, наиболее похожий на случайный шум, а математическое ожидание E_σ делает величину сложности R не зависящей от выбора вектора шума. Степень близости вектора ошибок и вектора шума, выраженная с помощью ковариации, вполне соответствует интуитивным представлениям о сложности: чем сложнее множество векторов, тем выше шансы найти в нём вектор, «похожий» на случайный шум. Благодаря линейности ковариации радемахеровская сложность обладает удобными алгебраическими свойствами, что упрощает её оценивание [2].

Как будет видно из дальнейшего, классическая радемахеровская сложность может применяться только в частном случае $\ell = k$, т. е. при равных по длине обучающей и контрольной выборках. В комбинаторном подходе нет такого ограничения, поэтому определение радемахеровской сложности приходится обобщать.

Определение 2. *Пусть* $n_l = \left\lfloor \frac{n\ell}{\ell+k} \right\rfloor$, $n_k = n - n_\ell$. *Пусть* $\Pi = \left\{\pi \subset \{1,\ldots,n\} \colon |\pi| = n_\ell\right\}$ *— множество всех* $C_n^{n_\ell}$ *разбиений множества из* n *элементов на два подмножества мощности* n_l *и* n_k *соответственно. Пусть все разбиения* $\pi \in \Pi$ *равновероятны. Комбинаторной радемахеровской сложностью множества* $V \subseteq \mathbb{R}^n$ *назовем величину*

$$R^n_{\ell,k}(V) = \mathsf{E}_\pi \sup_{v \in V} \big(\sigma(\pi)^\intercal v\big),$$

где $\sigma(\pi) = (\sigma_1,\ldots,\sigma_n)$, $\sigma_i = \begin{cases} -1/n_\ell, & i \in \pi; \\ +1/n_k, & i \notin \pi. \end{cases}$

Следующая лемма является комбинаторным аналогом неравенства МакДиармида [6], обычно используемого для получения оценок обобщающей способности на основе классической радемахеровской сложности [2, 7]. В дальнейшем эта лемма позволит связать понятия равномерного переобучения и комбинаторной радемахеровской сложности множества векторов ошибок Z. Её доказательство можно найти в [3] для случая, когда равномерное распределение вероятностей вводится на множестве перестановок выборки, а не разбиений. Однако если все используемые функции выборок X и $\bar{X}$ не зависят от порядка элементов в них, то лемма легко переформулируется в терминах разбиений.

Лемма 1. *Пусть* $(X, \bar{X})$ *— произвольное разбиение генеральной совокупности, разбиение* $(X', \bar{X}')$ *получено из него путём обмена пары объектов* $x \in X$, $x' \in \bar{X}$: $X' = X \setminus \{x\} \cup \{x'\}$, $\bar{X}' = \mathbb{X} \setminus X'$. *Пусть* $f(X, \bar{X})$ *— функция, не зависящая от порядка элементов в* X *и* $\bar{X}$, *удовлетворяет неравенству* $|f(X,\bar{X}) - f(X',\bar{X}')| < \beta$ *для всех* X, x, x'. *Тогда выполнено неравенство:*

$$\mathsf{P}\left[f(X,\bar{X}) - \mathsf{E}f(X,\bar{X}) \geqslant \varepsilon\right] \leqslant \\ \leqslant \exp\left\{\frac{-2\varepsilon^2\left(L - \frac{1}{2}\right)}{\ell k \beta^2}\left(1 - \frac{1}{2\max\{\ell,k\}}\right)\right\}.$$

Замечание 1. Поскольку разбиения являются равновероятными, $\mathsf{P}\,\varphi(X)$ обозначает долю разбиений выборки, при которых условие $\varphi(X)$ истинно, а $\mathsf{E}f(X)$ обозначает среднее по всем разбиениям значение функции $f(X)$.

Замечание 2. Легко видеть, что введённая в (1) величина $\delta(Z, X)$ не зависит от порядка элементов в подвыборках X и $\bar{X}$ и удовлетворяет лемме при $\beta = 2b(\frac{1}{\ell} + \frac{1}{k})$, где b — оценка сверху ошибки на одном объекте выборки.

Оценка обобщающей способности

Верхнюю оценку равномерного переобучения предлагается получать за два шага.

На первом шаге величина $\delta(Z, X)$ оценивается сверху с помощью комбинаторной радемахеровской сложности, определяемой по всей генеральной совокупности $\mathbb{X}$. Этот шаг следует логике работы [3].

На втором шаге доказываются два утверждения, связывающих радемахеровскую сложность на генеральной совокупности $\mathbb{X}$ с радемахеровской сложностью на случайной обучающей выборке X, которая может быть вычислена эмпирически.

Лемма 2. *Пусть* Z *— множество векторов ошибок семейства алгоритмов* A *на генеральной совокупности* $\mathbb{X}$; *ошибки ограничены сверху величиной* b; *разбиение* $(X, \bar{X})$ *выбрано случайно и равновероятно. Тогда для любого* η *с вероятностью не менее* $1 - \eta$ *выполнено неравенство*

$$\delta(Z, X) \leqslant R^L_{\ell,k}(Z) + b\sqrt{\frac{2SL}{\ell k}\log\frac{1}{\eta}}, \qquad (2)$$

где $S = \frac{L}{\left(L - \frac{1}{2}\right)\left(1 - \frac{1}{2\max(\ell,k)}\right)} \approx 1$.

Обозначим через $Z_X \subseteq \mathbb{R}^\ell$ множество векторов ошибок семейства A на обучающей выборке X.

Лемма 3. *При условиях леммы 2 радемахеровская сложность генеральной совокупности не превосходит математического ожидания радемахеровской сложности по обучающей выборке*[1]*:*

$$R^L_{\ell,k}(Z) \leqslant \mathsf{E}_X R^\ell_{\ell,k}(Z_X). \qquad (3)$$

[1] Доказательство леммы принадлежит Илье Толстихину.

Замечание 3. В действительности имеет место более сильное утверждение: $R^L_{\ell,k}(Z) \leqslant \mathsf{E}R^\ell_{\ell',k'}(Z_X)$ для любых ℓ' и k таких, что $\ell' + k' = \ell$. То есть утверждение леммы остаётся верным и в том случае, если обучающая выборка разбивается не в той же пропорции, что генеральная совокупность. Этим косвенно обосновывается использование для выбора модели скользящего контроля с разбиением на части, размер которых определяется соображениями скорости расчёта и устойчивости, а не предполагаемым размером тестовой выборки.

Лемма 4. *При условиях леммы 2*

$$\mathsf{E}_X R^\ell_{\ell,k}(Z_X) \leqslant R^\ell_{\ell,k}(Z_X) + b\sqrt{\frac{2SL}{\ell k}\log\frac{1}{\eta}}. \qquad (4)$$

Итак, неравенства (2)–(4), связывают оценку переобучения с радемахеровской сложностью, рассчитанной по обучающей выборке. Сформулируем теорему, обобщающую полученные результаты.

Теорема 5. *Пусть Z — множество векторов ошибок семейства алгоритмов A на генеральной совокупности $\mathbb{X}$; ошибки ограничены сверху величиной b; разбиение $(X, \bar{X})$ выбрано случайно и равновероятно. Тогда для любого η с вероятностью не менее $1-\eta$ выполнено неравенство*

$$\delta(Z, X) \leqslant R^\ell_{\ell,k}(Z_X) + 2b\sqrt{\frac{2SL}{\ell k}\log\frac{2}{\eta}}. \qquad (5)$$

Экспериментальный анализ верхней оценки равномерного переобучения

Для сравнения полученной оценки переобучения с известными ранее и анализа факторов её завышенности был поставлен эксперимент на модельных данных, генерируемых следующим образом.

Объекты $x_1, \ldots, x_L \sim N^d(0,1)$ — d-мерные нормальные случайные векторы. Ответы $y_i = \alpha^\intercal x_i + \varepsilon_i$, где $\alpha = \mathbf{1}^d$, шум $\varepsilon_i \sim N(0, \sigma_y)$ нормальный, некоррелированный, с дисперсией $\sigma_y = 0{,}1$. Семейство A — линейные функции регрессии. Размеры выборок $L = 200$, $l = k = 100$, для оценки переобучения рассматривались значения максимальной ошибки $b = 5\sigma_y$ и $\eta = 0{,}05$. Функция потерь квадратичная, но с дополнительным ограничением — не допускаются ошибки больше b: $\mathscr{L}(\hat{y}, y) = \min\big((\hat{y} - y)^2, b\big)$. Метод обучения μ — метод наименьших квадратов (МНК). Применение МНК для ограниченной функции потерь оправдано тем, что в рассматриваемой задаче ошибки на обучении не превышают b, и алгоритм, выбираемый по МНК, остаётся оптимальным и для введённой функции потерь.

Наряду с оценкой (5) вычислялась известная оценка обобщающей способности [4], основанная на

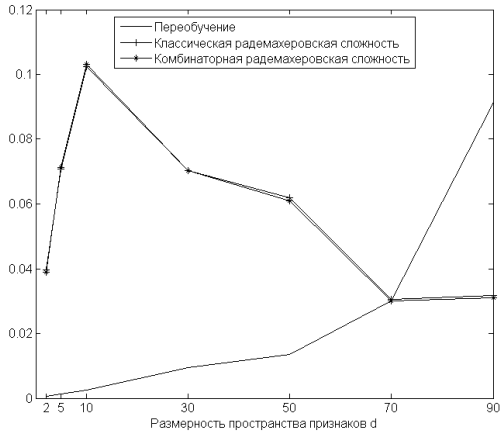


Рис. 1. Радемахеровская сложность и переобучение.

классической радемахеровской сложности (определение (1)): с вероятностью не менее $1 - 2\eta$

$$\psi_\tau(Z) \leqslant R^\ell(Z(\tau)) + 3b\sqrt{\frac{2}{\ell}\log\frac{2}{\eta}}. \qquad (6)$$

Данная оценка формулируется при тех же условиях, что и оценка (5), и получается с помощью неравенства МакДиармида, что позволяет проводить прямые аналогии между двумя оценками.

В эксперименте генерировались выборки различных размерностей $d \in \{2, 5, 10, 30, 50, 70, 90\}$. Для каждого фиксированного d было произведено $N = 100$ разбиений генеральной совокупности на обучающую и контрольную подвыборки. При каждом разбиении $(X, \bar{X})$ выполнялась настройка регрессионной функции по МНК и рассчитывалась величина переобучения $\delta(z, X)$, где z — вектор ошибок полученной функции. Также рассчитывались оценки классической и комбинаторной радемахеровской сложности $R^\ell(Z_X)$ и $R^\ell_{\ell,k}(Z_X)$ методом Монте-Карло по $M = 100$ генерациям векторов радемахеровских величин σ. На их основе вычислялись оценки переобучения (6) и (5), которые сравнивались с истинным значением переобучения $\delta(z, X)$. Окончательные оценки радемахеровской сложности, переобучения и его оценок получались взятием медианы по разбиениям.

Зависимость переобучения и радемахеровской сложности от d представлена на рис. 1. Завышенность оценок, представленная на рис. 2 и 3, вычислялась делением медианных значений оценок на медианное значение переобучения. Также взятием медианы по разбиениям оценивались средняя и максимальная ошибки на одном объекте, ограниченные по условиям эксперимента оценкой $b = 0{,}5$. Их зависимость от d представлена на рис. 4.

На основе этих результатов можно сделать следующие выводы.

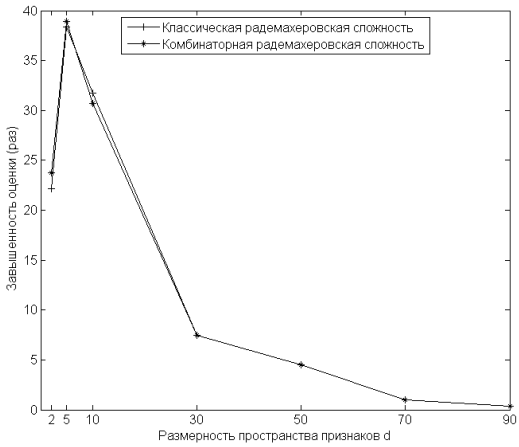


Рис. 2. Завышенность рademахеровской сложности.

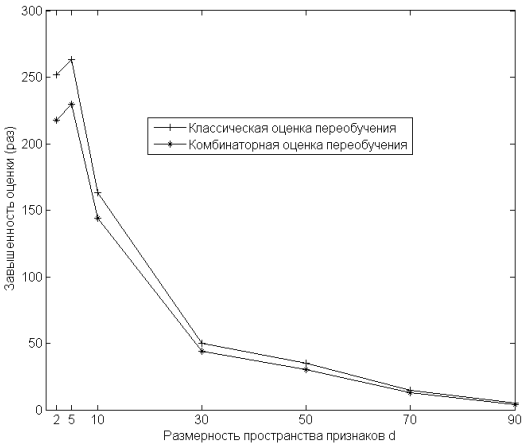


Рис. 3. Завышенность оценок переобучения.

1. Комбинаторная оценка немного лучше классической, но в целом обе оценки завышены на два порядка (для малых размерностей).

2. При увеличении размерности задачи точность оценки повышается. Это связано с тем, что при $d \sim \ell$ обучение по МНК даёт «плохие» алгоритмы, для которых $\delta(z, X)$ близка к $\delta(Z, X)$ и к рademахеровской сложности семейства.

3. Завышенность оценки, получаемая за счёт взятия супремума по всем алгоритмам семейства (заложенная в определении рademахеровской сложности), составляет для рассмотренной задачи лишь один порядок и быстро уменьшается с ростом d. При этом способ вычисления рademахеровской сложности не играет роли.

4. Основной фактор завышенности в данной задаче — применение неравенства МакДиармида: второе слагаемое оценки даёт завышение на два порядка. В этом завышении существенна равномерная мажоранта единичной ошибки: в эксперименте $b = 0{,}5$, в то время как среднее значение ошибок для малых размерностей едва превышает 0,01.

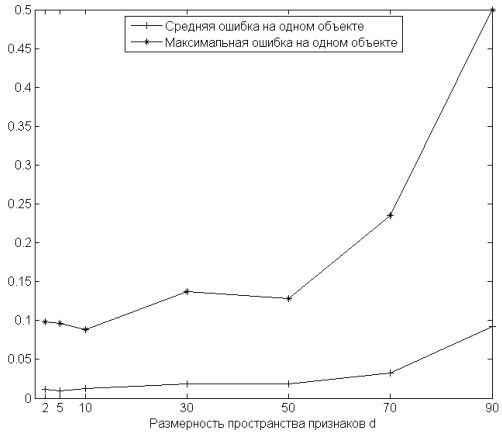


Рис. 4. Максимальная ошибка на одном объекте.

Заключение

Получена оценка переобучения для задач восстановления регрессии в слабой вероятностной аксиоматике. Введено понятие комбинаторной рademахеровской сложности и использован комбинаторный аналог неравенства МакДиармида. Переход к слабой вероятностной аксиоматике позволил упростить технику доказательств и немного улучшить оценку. Эксперимент на модельных данных показал, что обе оценки, и классическая, и комбинаторная, завышены на два порядка. Для повышения качества оценок необходимо не только сужать множество рассматриваемых алгоритмов и использовать более точные неравенства концентрации вероятностной меры, но и отказываться от априорной равномерной мажоранты функции потерь.

Литература

[1] *Vorontsov K. V.* Exact combinatorial bounds on the probability of overfitting for empirical risk minimization // *Pattern Recognition and Image Analysis.* — 2010. — Vol. 20, no. 3. — Pp. 269–285.

[2] *Boucheron S., Bousquet O., Lugosi G.* Theory of classification: A survey of some recent advances. — 2005.

[3] *El-Yaniv R., Pechyony D.* Transductive rademacher complexity and its applications. // J. Artif. Intell. Res. (JAIR). — 2009. — Pp. 193–234.

[4] *Koltchinskii V.* Rademacher penalties and structural risk minimization // IEEE Trans. on Information Theory. — 2001. — Vol. 47, no. 5. — Pp. 1902–1914.

[5] *Koltchinskii V., Panchenko D.* Rademacher processes and bounding the risk of function learning // High Dimensional Probability II. — Birkhäuser, 1999. — Pp. 443–459.

[6] *McDiarmid C.* On the method of bounded differences // In Surveys in Combinatorics, London Math. Soc. Lecture Notes Series. — 1989. — Vol. 141. — Pp. 148–188.

[7] *Philips P.* Data-Dependent Analysis of Learning Algorithms: Phd thesis / The Australian National University. — Canberra, Australia, 2005.

Метод порождающих и запрещающих множеств для рандомизированной минимизации эмпирического риска*

Фрей А. И.
frey@forecsys.ru
Москва, Московский физико-технический институт (государственный университет)

В комбинаторной теории переобучения основным инструментом вывода оценок обобщающей способности является метод порождающих и запрещающих множеств. В данной работе этот метод обобщается на случай рандомизированной минимизации эмпирического риска.

При решении задач машинного обучения требуется из заданного множества алгоритмов выбрать алгоритм, который ошибался бы как можно реже не только на объектах наблюдаемой обучающей выборки, но и на объектах скрытой контрольной выборки, которая в момент выбора алгоритма ещё неизвестна. Если частота ошибок на контрольной выборке оказывается значительно выше, чем на обучающей, то говорят, что произошло переобучение алгоритма — он слишком хорошо описывает конкретные данные, но не обладает способностью к обобщению этих данных, не восстанавливает порождающую их зависимость и не пригоден для построения прогнозов.

На практике переобучение оценивается количественно с помощью процедуры скользящего контроля (кросс-валидации). Фиксируется некоторое множество разбиений исходной выборки на две подвыборки — обучающую и контрольную. Для каждого разбиения выполняется настройка алгоритма по обучающей подвыборке, затем оценивается его средняя ошибка на объектах контрольной подвыборки. *Оценкой скользящего контроля* называется средняя по всем разбиениям величина ошибки на контрольных подвыборках. Аналогичным образом определяется и *оценка вероятности переобучения* — это доля разбиений, при которых средняя ошибка на контрольной подвыборке превышает среднюю ошибку на обучающей подвыборке более чем на заданную величину ε. В комбинаторной теории переобучения рассматривается множество всех возможных разбиений и ставится задача получения вычислительно эффективных формул для оценок скользящего контроля и вероятности переобучения [2, 7].

Теоретико-групповой подход [4, 5, 6] позволяет получать такие формулы для случаев, когда семейство алгоритмов обладает некоторой симметрией, а методом обучения является *рандомизированная минимизация эмпирического риска*. Рандомизация означает, что если в семействе существует несколько алгоритмов, допускающих одинаковое минимальное число ошибок на обучающей выборке, то из них равновероятно выбирается любой.

В данной работе техника порождающих и запрещающих множеств, разработанная в [2, 7] исключительно для детерминированных методов обучения, обобщается на случай рандомизированной минимизации эмпирического риска. С её помощью выводятся точные оценки вероятности переобучения для монотонных сетей алгоритмов — нетривиального модельного семейства, обладающего всеми ключевыми свойствами реальных семейств — расслоением, связностью и размерностью.

Основные обозначения

Пусть задано конечное множество объектов $\mathbb{X} = \{x_1, \ldots, x_L\}$, называемое *генеральной выборкой*. Произвольный алгоритм классификации a порождает бинарный *вектор ошибок* $\big(I(a, x_i)\big)_{i=1}^{L}$, где $I(a, x_i) \in \{0, 1\}$ — индикатор ошибки алгоритма a на объекте x_i. В дальнейшем будем отождествлять алгоритмы с их векторами ошибок.

Обозначим через $\mathbb{A} = \{0, 1\}^L$ множество всех возможных векторов ошибок длины L. Через $[\mathbb{X}]^\ell$ обозначим множество всех подвыборок $X \subset \mathbb{X}$ длины ℓ. Будем говорить, что генеральная выборка $\mathbb{X}$ разбивается на обучающую выборку X и контрольную выборку $\bar{X} = \mathbb{X} \setminus X$ длины $k = L - \ell$. *Число ошибок* алгоритма a на выборке $U \subseteq \mathbb{X}$ обозначим через $n(a, U) = \sum_{x \in U} I(a, x)$. Величину $\nu(a, U) = n(a, U)/|U|$ будем называть *частотой ошибок* алгоритма a на выборке U. *Уклонение частот* алгоритма a на разбиении $\mathbb{X} = X \sqcup \bar{X}$ определим как разность частот ошибок на контроле и на обучении: $\delta(a, X) = \nu(a, \bar{X}) - \nu(a, X)$.

Для произвольного подмножества алгоритмов $A \subset \mathbb{A}$ с попарно различными векторами ошибок введём множество алгоритмов $A(X)$ с минимальным числом ошибок на обучающей выборке X:

$$A(X) = \operatorname*{Argmin}_{a \in A} n(a, X). \qquad (1)$$

Детерминированным методом обучения называется отображение $\mu\colon [\mathbb{X}]^\ell \to \mathbb{A}$, которое произвольной обучающей выборке X ставит в соответствие некоторый алгоритм $a = \mu X$. Частоту ошибок на обучающей выборке $\nu(\mu X, X)$ называют *эм-*

Работа поддержана РФФИ (проект №11-07-00480) и программой ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

пирическим риском. Минимизация эмпирического риска — это метод обучения μ, для которого $\mu X \in A(X)$ для любой выборки $X \in [\mathbb{X}]^\ell$.

Говорят, что метод μ переобучается на разбиении $X \sqcup \bar{X}$, если $\delta(\mu X, X) \geqslant \varepsilon$. Переобучение может быть следствием «неудачного» разбиения генеральной выборки на обучение и контроль. Поэтому вводится функционал *вероятности переобучения*, равный доле разбиений выборки, при которых возникает переобучение [2, 3]:

$$Q_\varepsilon(A) = \mathsf{E}[\delta(\mu X, X) \geqslant \varepsilon], \text{ где } \mathsf{E} \equiv \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell} .$$

Здесь и далее квадратные скобки (нотация Айверсона) переводят логическое выражение в число 0 или 1 по правилам [истина] = 1, [ложь] = 0.

Функционал $Q_\varepsilon(A)$ уже не зависит от выбора разбиения и характеризует качество метода обучения μ на данной генеральной выборке $\mathbb{X}$.

Порождающие и запрещающие множества объектов

Первый подход, позволивший получать точные оценки вероятности переобучения в рамках слабой вероятностной аксиоматики, основан на выделении порождающих и запрещающих объектов [7].

Гипотеза 1. Пусть множество A, выборка $\mathbb{X}$ и детерминированный метод обучения μ таковы, что для каждого алгоритма $a \in A$ можно указать пару непересекающихся подмножеств $X_a \subset \mathbb{X}$ и $X'_a \subset \mathbb{X}$, удовлетворяющую условию

$$\big[\mu X{=}a\big] = \big[X_a \subseteq X\big]\big[X'_a \subseteq \bar{X}\big], \quad \forall X \in [\mathbb{X}]^\ell. \quad (2)$$

Множество X_a называется *порождающим*, X'_a — *запрещающим* для алгоритма a. Гипотеза 1 означает, что метод μ выбирает алгоритм a тогда и только тогда, когда в обучающей выборке X находятся все порождающие объекты и ни одного запрещающего. Все остальные объекты $\mathbb{X}\backslash X_a \backslash X'_a$ называются *нейтральными* для алгоритма a.

Для произвольного алгоритма $a \in A$ введём следующие обозначения:

$L_a = L - |X_a| - |X'_a|$ — число нейтральных объектов в генеральной выборке;

$\ell_a = \ell - |X_a|$ — число нейтральных объектов в обучающей выборке;

$m_a = n(a, \mathbb{X}\backslash X_a \backslash X'_a)$ — число ошибок алгоритма a на нейтральных объектах;

$s_a(\varepsilon) = \frac{\ell}{L}\big(n(a,\mathbb{X}) - \varepsilon k\big) - n(a, X_a)$ — наибольшее число ошибок алгоритма a на нейтральных обучающих объектах $X \setminus X_a$, при котором имеет место большое уклонение частот ошибок, $\delta(a, X) \geqslant \varepsilon$.

Введём функцию гипергеометрического распределения:

$$H_L^{\ell,m}(z) = \sum_{s=0}^{\lfloor z \rfloor} \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}.$$

Теорема 1. *Если справедлива гипотеза 1, то вероятность получить в результате обучения алгоритм a равна $P_a(A) = \mathsf{P}[\mu X{=}a] = C_{L_a}^{\ell_a}/C_L^\ell$, вероятность переобучения равна*

$$Q_\varepsilon(A) = \sum_{a \in A} P_a H_{L_a}^{\ell_a, m_a}\left(s_a(\varepsilon)\right).$$

Данный результат позволил получить формулы вероятности переобучения для широкого класса модельных семейств алгоритмов, в частности для монотонных и унимодальных сетей. В следующем параграфе аналогичная теорема будет получена для рандомизированных методов обучения.

Рандомизированный метод минимизации эмпирического риска

Рандомизированный метод минимизации эмпирического риска выбирает произвольный алгоритм из множества $A(X)$ случайно и равновероятно [4, 6]. Поскольку в задаче статистического обучения появляется второй независимый источник случайности, определение вероятности переобучения $Q_\varepsilon(A)$ приходится модифицировать. Наиболее естественный вариант модификации — усреднение по множеству $A(X)$:

$$Q_\varepsilon(A) = \mathsf{E}\frac{1}{|A(X)|} \sum_{a \in A(X)} \big[\delta(a, X) \geqslant \varepsilon\big]. \quad (3)$$

Для детерминированного метода обучения результатом обучения являлся алгоритм $a \in A$. В случае рандомизированного метода результатом обучения является подмножество $A(X) \subseteq A$. Таким образом, множество алгоритмов A порождает множество подмножеств алгоритмов, получающихся в результате обучения

$$\mathfrak{A} = \big\{A(X) \colon X \in [\mathbb{X}]^\ell\big\}.$$

Гипотеза 2. Пусть множество A и выборка $\mathbb{X}$ таковы, что для каждого подмножества $\alpha \in \mathfrak{A}$ можно указать пару непересекающихся подмножеств $X_\alpha \subset \mathbb{X}$ и $X'_\alpha \subset \mathbb{X}$, удовлетворяющую условию

$$\big[A(X){=}\alpha\big] = \big[X_\alpha \subseteq X\big]\big[X'_\alpha \subseteq \bar{X}\big], \ \ \forall X \in [\mathbb{X}]^\ell. \quad (4)$$

Следующая теорема является непосредственным обобщением теоремы 1 для рандомизированного метода минимизации эмпирического риска.

Теорема 2. *Если справедлива гипотеза 2, то вероятность переобучения рандомизированного метода минимизации эмпирического риска есть*

$$Q_\varepsilon(A) = \sum_{a \in A} \sum_{\alpha \in \mathfrak{A}} \frac{[a \in \alpha]}{|\alpha|} \frac{C_{L_\alpha}^{\ell_\alpha}}{C_L^\ell} H_{L_\alpha}^{\ell_\alpha, m_\alpha^a}\left(s_\alpha^a(\varepsilon)\right),$$

где введены следующие обозначения:

$$L_\alpha = L - |X_\alpha| - |\bar{X}_\alpha|; \quad \ell_\alpha = \ell - |X_\alpha|;$$
$$m_\alpha^a = n(a, \mathbb{X}\backslash X_\alpha \backslash X'_\alpha);$$
$$s_\alpha^a(\varepsilon) = \tfrac{\ell}{L}(n(a,\mathbb{X}) - \varepsilon k) - n(a, X_\alpha).$$

Лемма 3. *Пусть во множестве A найдётся алгоритм a_0, такой, что для любого $a \in A$ вектор ошибок алгоритма a_0 содержится в векторе ошибок алгоритма a. Обозначим через X_0 множество объектов, на которых ошибается алгоритм a_0. Пусть система порождающих и запрещающих множеств такова, что для всех $\alpha \in \mathfrak{A}$ выполнено $X_0 \cap X_\alpha = \varnothing$ и $X_0 \cap X'_\alpha = \varnothing$. Тогда*

$$m^a_\alpha = n(a_0, \mathbb{X}), \quad s^a_\alpha(\varepsilon) = \tfrac{\ell}{L}(n(a, \mathbb{X}) - \varepsilon k).$$

Теорема о порождающих и запрещающих объектах легко объединяется с теоремой о разбиении множества алгоритмов на орбиты действия группы симметрии [4].

Рассмотрим симметрическую группу S_L, элементы которой действуют на алгоритм a как на бинарный вектор ошибок длины L, и на произвольное множество алгоритмов $A \subset \mathbb{A}$ (поэлементно). *Группой симметрий* $\operatorname{Sym} A$ будем называть стационарную подгруппу $\operatorname{Sym} A = \{\pi \in S_L \colon \pi A = A\}$.

Теорема 4. *Пусть $G \subset \operatorname{Sym} A$ — подгруппа группы симметрии множества алгоритмов A, $\Omega(A)$ — множество орбит действия G на A, алгоритм $a_\omega \in \omega$ — представитель орбиты $\omega \in \Omega$. Тогда вероятность переобучения $Q_\varepsilon(A)$ можно записать в виде:*

$$Q_\varepsilon(A) = \sum_{\omega \in \Omega(A)} \sum_{\alpha \in \mathfrak{A}} \big[a_\omega \in \alpha\big] \frac{|\omega|}{|\alpha|} \frac{C^{\ell_\alpha}_{L_\alpha}}{C^\ell_L} H^{\ell_\alpha, m^{a_\omega}_\alpha}_{L_\alpha}\left(s^{a_\omega}_\alpha(\varepsilon)\right).$$

Монотонные и унимодальные сети алгоритмов

Монотонная сеть алгоритмов [1] — это модель параметрического семейства алгоритмов классификации с разделяющей поверхностью, непрерывной по вектору параметров. Предполагается, что в семействе имеется наилучший алгоритм a_0, и что при непрерывном изменении каждого его параметра число ошибок на полной выборке только увеличивается.

Пример 1. Монотонная двумерная сеть алгоритмов при $L = 4$ и $n(a_0, \mathbb{X}) = 0$:

$$\begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \end{array} \left(\begin{array}{ccccccccc} 0 & \mathbf{1} & \mathbf{1} & 0 & \mathbf{1} & \mathbf{1} & 0 & \mathbf{1} & \mathbf{1} \\ 0 & 0 & \mathbf{1} & 0 & 0 & \mathbf{1} & 0 & 0 & \mathbf{1} \\ \hline 0 & 0 & 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} \end{array}\right)$$

Унимодальная сеть алгоритмов [1] является более реалистичной моделью связного параметрического семейства, по сравнению с монотонной сетью. Если мы имеем лучший алгоритм a_0 с оптимальным значением вектора вещественных параметров, то отклонение значений компонент этого вектора как в бóльшую, так и в меньшую, сторону приводит к увеличению числа ошибок.

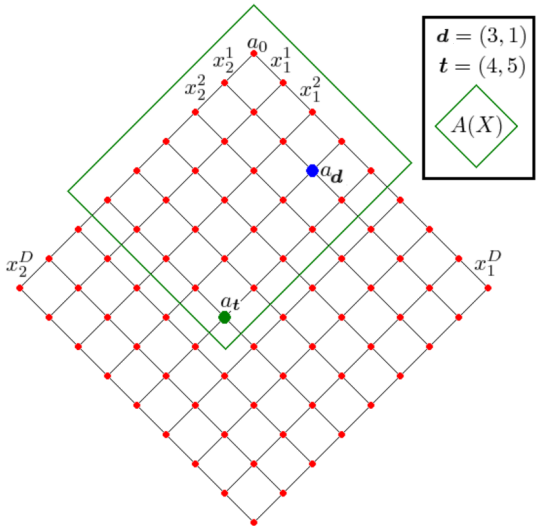


Рис. 1. Строение множества $A(X)$ для двумерной монотонной сети; $h = 2$, $D = 8$.

Формулы вероятности переобучения рандомизированного метода минимизации эмпирического риска для многомерных монотонной и унимодальной сетей получены в [5]. Их доказательства можно существенно упростить, если заметить, что монотонные и унимодальные сети удовлетворяют гипотезе 2.

Введём целочисленный вектор индексов $\boldsymbol{d} = (d_1, \ldots, d_h) \in \mathbb{Z}^h$. Обозначим $\|\boldsymbol{d}\| = \max\limits_{j=1,\ldots,h} |d_j|$, $|\boldsymbol{d}| = |d_1| + \ldots + |d_h|$. На множестве векторов индексов введём покомпонентное отношение сравнения: $\boldsymbol{d} < \boldsymbol{d}'$, если $d_j \leqslant d'_j$, $j = 1, \ldots, h$, и хотя бы одно из неравенств строгое.

Определение 1. *Множество алгоритмов $A = \{a_{\boldsymbol{d}}\}$, где $\boldsymbol{d} \geqslant 0$ и $\|\boldsymbol{d}\| \leqslant D$ называется монотонной h-мерной сетью алгоритмов размера D, если существуют упорядоченные наборы объектов $X_j = \{x_j^1, \ldots, x_j^D\} \subset \mathbb{X}$, для всех $j = 1, \ldots, h$, а также множества $U_1 \subset \mathbb{X}$ и $U_0 \subset \mathbb{X}$, такие, что:*

1) $\mathbb{X} = U_0 \sqcup U_1 \sqcup X_1 \sqcup \cdots \sqcup X_h$;
2) $a_d(x_j^i) = [i \leqslant d_j]$, где $x_j^i \in X_j$;
3) $a_d(x) = 0$ при всех $x \in U_0$;
4) $a_d(x) = 1$ при всех $x \in U_1$.

Отметим, что для произвольной обучающей выборки X множество $A(X)$ устроено специфическим образом. На рис. 1 показано, что в $A(X)$ всегда найдется такой алгоритм $a_{\boldsymbol{t}}$, что $A(X) = \{a_{\boldsymbol{d}} \,|\, \boldsymbol{d} \leqslant \boldsymbol{t}\}$.

Рассмотрим произвольное подмножество алгоритмов $\alpha \in \mathfrak{A}$. Согласно определению $\mathfrak{A}$ это значит, что найдется такое разбиение $X \in \mathbb{X}$, для которого $\alpha = A(X)$.

Обозначим через $i'(j)$ наименьший номер $i \in \{1, \ldots, D\}$, такой, что объект x_j^i попадает в обучающую выборку X. Возможно, что для некоторых $j \in \{1, \ldots, h\}$ выражение $i'(j)$ не определено,

поскольку все объекты x_j^i оказались в контроле $\bar{X}$. Пусть $J \subset \{1,\ldots,h\}$ — множество индексов j, для которых $i'(j)$ определено.

Лемма 5. *Положим*

$$X_\alpha = \bigcup_{j\in J} x_j^{i'(j)}, \quad X'_\alpha = \bigcup_{j=1}^{h} \bigcup_{i=1}^{i'(j)-1} x_j^i.$$

Тогда:

1) *X_α и X'_α зависят только от исходного множества α, но не от выбора представителя $X \in \mathbb{X}$;*
2) *X_α и X'_α удовлетворяют условию* (4).

Данное утверждение, в сочетании с теоремой 2 и леммой 3, позволяет выписать формулу вероятности переобучения рандомизированного метода обучения для монотонной сети алгоритмов.

Теорема 6. *Вероятность переобучения рандомизированного метода минимизации эмпирического риска, примененного к монотонной сети $A = \{a_{\boldsymbol{d}}\}$ размерности h, $\|\boldsymbol{d}\| \leqslant D$, дается выражением:*

$$Q_\varepsilon(A) = \sum_{\substack{\boldsymbol{d}\geqslant\boldsymbol{0},\\ \|\boldsymbol{d}\|\leqslant D}} \sum_{\substack{\boldsymbol{t}\geqslant\boldsymbol{0},\\ \|\boldsymbol{t}\|\leqslant D}} \frac{[\boldsymbol{t}\geqslant\boldsymbol{d}]}{V(\boldsymbol{t})} \frac{C_{L'}^{\ell'}}{C_L^\ell} H_{L'}^{\ell',m}\left(s(\varepsilon)\right),$$

где $V(\boldsymbol{t}) = \prod_{j=1}^{h}(t_j+1)$, $\ell' = \ell - \sum_{j=1}^{h}[t_j \neq D]$, $k' = k - |\boldsymbol{t}|$, $L' = \ell' + k'$, $s(\varepsilon) = \frac{\ell}{L}\big[m + |\boldsymbol{d}| - \varepsilon k\big]$.

Вычисление вероятности переобучения по данной формуле требует $O(h \cdot D^{2h})$ операций. Теорема 4 позволяет сократить количество вычислений за счет учета симметрий монотонной сети. Отметим, что группа симметрии монотонной сети размерности h изучена в [5], и содержит в качестве подгруппы группу S_h всевозможных перестановок множеств $X_1,\ldots,X_h$.

Теорема 7. *С учетом симметрий монотонной сети вероятность переобучения записывается в виде*

$$Q_\varepsilon(A) = \sum_{\boldsymbol{d}\in Y_h^D} \sum_{\substack{\boldsymbol{t}\geqslant\boldsymbol{d},\\ \|\boldsymbol{t}\|\leqslant D}} \frac{|S_h\boldsymbol{d}|}{V(\boldsymbol{t})} \frac{C_{L'}^{\ell'}}{C_L^\ell} H_{L'}^{\ell',m}\left(s(\varepsilon)\right),$$

где Y_h^D — множество целочисленных неотрицательных невозрастающих последовательностей длины h и не превосходящих D, $|S_h\boldsymbol{d}|$ — число различных слов, состоящих из символов $d_1,\ldots,d_h$.

Расчет новой формулы требует $O(h \cdot D^h \cdot C_{D+h}^h)$ операций.

Рассмотрим отношение D^h/C_{h+D}^h, показывающее во сколько раз сокращается объем вычислений благодаря учету симметрии. Данная величина максимальна при $D \gg h$. Это соответствует случаю сеток большой длины, на которых группа симметрии действует наиболее эффективно. В этом случае число операций сокращается в $h!$ раз, что в точности соответствует соответствует количеству элементов в группе симметрий. В остальных случаях (сетки больших размерностей и малой длины) выигрыш оказывается меньше.

Выводы

В данной работе метод получения комбинаторных оценок вероятности переобучения, основанный на выделении порождающих и запрещающих множеств объектов, обобщён на случай рандомизированной минимизации эмпирического риска. Полученный результат по-прежнему позволяет учитывать структуру симметрии множества алгоритмов (разбиение множества алгоритмов на орбиты действия группы симметрии).

С помощью предложенного подхода получены две оценки вероятности переобучения для монотонной сети: с учетом и без учета симметрий. Показано, что при определенных сочетаниях параметров учет симметрий позволяет сократить объем вычислений в $h!$ раз, где h — размерность сети.

Литература

[1] *Ботов П. В.* Точные оценки вероятности переобучения для монотонных и унимодальных семейств алгоритмов // Всеросс. конф. ММРО-14 — М.: МАКС Пресс, 2009. — С. 7–10.

[2] *Воронцов К. В.* Точные оценки вероятности переобучения // Доклады РАН, 2009. — Т. 429, № 1. — С. 15–18.

[3] *Воронцов К. В.* Комбинаторный подход к проблеме переобучения // Всеросс. конф. ММРО-14 — М.: МАКС Пресс, 2009. — С. 18–21.

[4] *Фрей А. И.* Точные оценки вероятности переобучения для симметричных семейств алгоритмов // Всеросс. конф. ММРО-14 — М.: МАКС Пресс, 2009. — С. 66–69.

[5] *Фрей А. И.* Вероятность переобучения плотных и разреженных многомерных сеток алгоритмов // Интеллектуализация обработки информации — М.: МАКС Пресс, 2010. — С. 87–90.

[6] *Frei A. I.* Accurate estimates of the generalization ability for symmetric set of predictors and randomized learning algorithms // Pattern Recognition and Image Analysis. — 2010. — Vol. 20, no. 3. — Pp. 241–250.

[7] *Vorontsov K. V.* Exact combinatorial bounds on the probability of overfitting for empirical risk minimization // Pattern Recognition and Image Analysis. — 2010. — Vol. 20, no. 3. — Pp. 269–285.

Принцип максимизации зазора для монотонного классификатора ближайшего соседа*

Воронцов К. В., Махина Г. А.
vokov@forecsys.ru, gmakhina@yandex.ru
Москва, Вычислительный центр им. А. А. Дородницына РАН; Симферополь, ТНУ им. В. И. Вернадского

Получены точные оценки полного скользящего контроля для монотонных классификаторов, основанных на принципе ближайшего соседа. Показано, что наилучшей обобщающей способностью обладает монотонный классификатор, в котором разделяющая поверхность проходит посередине зазора между классами. Показана связь данной задачи с задачей доопределения частично заданной монотонной булевой функции.

Монотонные классификаторы

Монотонным классификатором будем называть монотонное отображение $a\colon \mathbb{X} \to \mathbb{Y}$, действующее из множества объектов $\mathbb{X}$ в конечное множество ответов $\mathbb{Y}$. Предполагается, что множества $\mathbb{X}$ и $\mathbb{Y}$ частично упорядочены. Для простоты будем рассматривать только двухклассовые задачи, $\mathbb{Y} = \{0, 1\}$.

Ограничение монотонности довольно часто возникает при решении задач классификации.

1. Для некоторого подмножества признаков может быть априори известно, что чем больше значение признака, тем выше уверенность, что объект принадлежит классу 1. Это означает, что классификатор $a(x)$ должен быть монотонной функцией при соответствующем определении частичного порядка на $\mathbb{X}$.

2. Частным случаем монотонного классификатора является *пороговое решающее правило* $a(x) = \big[f(x) \geqslant \theta\big]$, где x — объект, $f(x)$ — некоторая вещественная функция, например, признак объекта, $\theta \in \mathbb{R}$ — порог. Пороговые правила используются как элементарные предикаты в конъюнктивных закономерностях, решающих деревьях [5] и решающих списках [10], как базовые классификаторы в композициях типа бустинга и бэггинга [13], как правила принятия окончательного решения во многих методах классификации на два класса.

3. *Композицией* T базовых классификаторов вида $a_t(x) = \big[b_t(x) \geqslant 0\big]$, где $b_t\colon \mathbb{X} \to \mathbb{R},\ t = 1, \ldots, T$, называется функция $a(x) = F\big(b_1(x), \ldots, b_T(x)\big)$, где $F\colon \mathbb{R}^T \to \mathbb{Y}$ — *корректирующая операция*. Значения $b_t(x)$ и $b(x)$ интерпретируются как степень уверенности классификаторов $a_t(x)$ и $a(x)$, соответственно, в том, что объект x принадлежит классу 1. Поэтому вполне естественно потребовать, чтобы функция F монотонно не убывала по всем своим T аргументам [1].

Частным случаем монотонной корректирующей операции является линейная с неотрицательными весами (взвешенное голосование) [6, 11]. Монотонные корректирующие операции позволяют строить короткие композиции (T от 2 до 6) из базовых классификаторов высокого качества [4], тогда как взвешенное голосование (в частности, алгоритмы бустинга и бэггинга) лучше подходит для построения сложных композиций (T порядка сотен) из простых базовых классификаторов низкого качества. Монотонные композиции более подвержены переобучению, поэтому для их построения хотелось бы иметь точные оценки обобщающей способности.

В [3] получены верхние оценки полного скользящего контроля, справедливые для произвольного множества монотонных классификаторов. В данной работе рассматривается семейство монотонных классификаторов ближайшего соседа, обобщающее конструкцию монотонной корректирующей операции из [2]. Для него получена точная оценка, в случае, когда исходная выборка монотонна.

Полный скользящий контроль

Рассмотрим задачу классификации на два класса, $\mathbb{Y} = \{0, 1\}$. Пусть $\mathbb{X} = \{x_1, \ldots, x_L\}$ — конечное множество объектов, $y_i = y(x_i)$ — истинная классификация объекта x_i.

Частота ошибок классификатора a на выборке X есть

$$\nu(a, X) = \frac{1}{|X|} \sum_{x \in X} \big[a(x) \neq y(x)\big].$$

Методом обучения называется отображение $\mu\colon 2^{\mathbb{X}} \to A$, которое произвольной выборке $X \subset \mathbb{X}$ ставит в соответствие некоторый классификатор $a = \mu(X)$ из заданного множества A.

Каноническим примером метода обучения является *минимизация эмпирического риска*:

$$\mu(X) = \arg\min_{a \in A} \nu(a, X), \quad X \subset \mathbb{X}.$$

Рассмотрим все C_L^ℓ разбиений множества объектов $\mathbb{X} = X \sqcup \bar{X}$ на две выборки — обучающую X длины ℓ и контрольную $\bar{X}$ длины $k = L - \ell$.

Функционал *полного скользящего контроля* CCV (complete cross-validation) характеризует обобщающую способность метода μ на конечном множестве объектов $\mathbb{X}$ и определяется как средняя по всем разбиениям частота ошибок на контроле [12]:

$$C(\mu, \mathbb{X}) = \frac{1}{C_L^\ell} \sum_X \nu\big(\mu(X), \bar{X}\big).$$

Работа поддержана РФФИ (проект №11-07-00480) и программой ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

Профиль монотонности

Выборка $X \subseteq \mathbb{X}$ называется монотонной, если из $x_i \leqslant x_j$ следует $y_i \leqslant y_j$ для всех $x_i, x_j \in X$.

Верхним и нижним клином объекта $x_i \in \mathbb{X}$ называются, соответственно, множества (рис. 1)

$$W_0(x_i) = \{x \in \mathbb{X}\colon x_i < x \text{ и } y(x) = 0\};$$
$$W_1(x_i) = \{x \in \mathbb{X}\colon x < x_i \text{ и } y(x) = 1\}.$$

Множество $W_i = W_{y_i}(x_i)$ будем называть просто *клином* объекта x_i. Это все объекты того же класса, что и x_i, лежащие между x_i и границей классов. Мощность клина $w_i = |W_i|$ характеризует глубину погружения объекта x_i в свой класс. Чем меньше w_i, тем ближе объект к границе класса.

Профилем монотонности выборки $\mathbb{X}$ называется функция $M(m)$, равная доле объектов с клином мощности m:

$$M(m) = \frac{1}{L}\sum_{i=1}^{L}[w_i = m]; \quad m = 0, \ldots, L-1.$$

Теорема 1. *Если выборка $\mathbb{X}$ монотонна и метод μ минимизирует эмпирический риск в классе всех монотонных функций A, то*

$$C(\mu, \mathbb{X}) \leqslant \sum_{m=0}^{k-1} M(m)\frac{C^{\ell}_{L-1-m}}{C^{\ell}_{L-1}}. \qquad (1)$$

Доказательство основано на том, что если монотонный классификатор ошибается на объекте x_i, то он ошибается и на всех объектах из клина W_i.

Отношение биномиальных коэффициентов в (1) быстро убывает по m. Поэтому для минимизации этой оценки необходимо, чтобы функция $M(m)$ принимала малые значения при малых m, то есть чтобы как можно меньше объектов имели клинья малой мощности. Для этого отношение порядка на множестве объектов X должно быть близко к линейному вблизи границы классов.

Профиль компактности

Метод ближайшего соседа μ запоминает выборку X и строит классификатор $a = \mu(X)$, относящий объект $x \in \mathbb{X}$ к тому классу, которому принадлежит обучающий объект $x' \in X$, ближайший к x по заданной функции расстояния $\rho(x, x')$:

$$a(x) = y\big(\arg\min_{x' \in X} \rho(x, x')\big). \qquad (2)$$

Предположим, что все расстояния $\rho(x, x')$ попарно различны. Для каждого объекта $x_i \in \mathbb{X}$ пронумеруем все остальные объекты $x_{i1}, \ldots, x_{i,L-1}$ по возрастанию расстояний $\rho(x_i, x_{im})$.

Профилем компактности называется функция $K(m)$, равная доле объектов $x_i \in \mathbb{X}$, у которых m-й сосед x_{im} находится в другом классе:

$$K(m) = \frac{1}{L}\sum_{i=1}^{L}\big[y_i \neq y(x_{im})\big], \quad m = 1, \ldots, L-1.$$

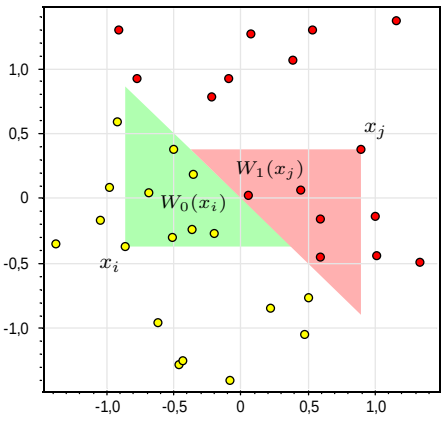


Рис. 1. Двумерная задача классификации с естественным отношением порядка на множестве $\mathbb{X} \subset \mathbb{R}^2$. Верхний клин объекта x_i, нижний клин объекта x_j.

Теорема 2. *Для метода ближайшего соседа μ*

$$C(\mu, \mathbb{X}) = \sum_{m=1}^{k} K(m)\frac{C^{\ell-1}_{L-1-m}}{C^{\ell}_{L-1}}. \qquad (3)$$

Обратим внимание на большое структурное сходство оценок (1) и (3), полученных для таких различных, на первый взгляд, методов, как метрический и монотонный классификатор.

Точные оценки CCV

Верхняя оценка (1) справедлива для любого множества монотонных классификаторов A. Именно в силу её чрезмерной общности она не является точной. Однако если ограничить множество A, то возможно получить точные оценки.

Введём следующие определения. *Нижняя область* объекта x_i — это множество всех объектов x таких, что $x \leqslant x_i$. *Верхняя область* объекта x_i — это множество всех объектов x таких, что $x_i \leqslant x$.

Если выборка $\mathbb{X}$ монотонна, то любой монотонный классификатор a правильно классифицирует все объекты в нижних областях обучающих объектов класса 0 и в верхних областях обучающих объектов класса 1, рис. 2. Классификация остальных объектов, *лежащих в зазоре* между нижними и верхними областями обучающих объектов, зависит от конструкции классификатора $a = \mu(X)$. На рис. 2 показана одна из возможных реализаций.

Граничные монотонные классификаторы. Рассмотрим два крайних случая — когда монотонный классификатор относит все объекты, лежащие в зазоре, либо к классу 0, либо к классу 1.

Назовём *профилем монотонности* класса $y \in \mathbb{Y}$ функцию $M_y(m)$, равную доле объектов класса y с клином мощности m:

$$M_y(m) = \frac{1}{L}\sum_{i=1}^{L}\big[y_i = y\big]\big[w_i = m\big]; \quad m = 0, \ldots, L-1.$$

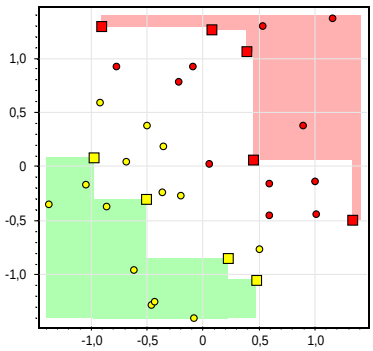


Рис. 2. Одна из допустимых монотонных классификаций двумерной выборки $\mathbb{X}$. Объекты обучающей выборки показаны квадратными точками □, ■.

Теорема 3. *Если выборка $\mathbb{X}$ монотонна и классификатор $a = \mu X$ относит все объекты, лежащие в зазоре, к классу y, то справедлива точная оценка*

$$C(\mu, \mathbb{X}) = \sum_{m=0}^{k-1} M_{1-y}(m) \frac{C^{\ell}_{L-1-m}}{C^{\ell}_{L-1}}. \qquad (4)$$

Очевидно, сумма профилей по классам совпадает с общим профилем: $M(m) = M_0(m) + M_1(m)$, $m = 0, \ldots, L-1$. Если предположить, что профиль монотонности делится между классами примерно поровну, то из теоремы 3 следует, что оценка (1) завышена приблизительно вдвое. На самом деле она завышена сильнее, поскольку такая классификация объектов из зазора явно не оптимальна с точки зрения обобщающей способности. Аналогия с машинами опорных векторов (SVM) и другими методами максимизации зазора [9] подсказывает, что оптимальное разделение должно проходить посередине зазора. Чтобы проверить эту гипотезу, сравним методы обучения, которые проводят разделяющую поверхность по границам зазора, по середине зазора или (для большей общности) по заданному отношению расстояний до нижней и верхней границ. Точная оценка CCV для случая разделения по середине зазора получена в [7].

Монотонный классификатор ближайшего соседа. Определим для произвольной пары объектов $x, u \in \mathbb{X}$ расстояние $r(x, u)$ между нижней областью объекта x и верхней областью объекта u так, чтобы оно обладало следующими свойствами:

1) $r(x, u) = 0$ тогда и только тогда, когда $x \geqslant u$;
2) $r(x, u)$ не возрастает по x не убывает по u.

Возьмём в методе ближайшего соседа (2) функцию расстояния $\rho(x, x')$ от классифицируемого объекта x до обучающего объекта $x' \in X$, равную расстоянию до его нижней области, если $y(x') = 0$, и до его верхней области, если $y(x') = 1$:

$$\rho_\lambda(x, x') = \begin{cases} (1-\lambda) r(x', x), & y(x') = 0; \\ (1+\lambda) r(x, x'), & y(x') = 1; \end{cases}$$

где параметр $\lambda \in (-1, 1)$ определяет положение разделяющей поверхности внутри зазора: при $\lambda \to -1$ все объекты из зазора относятся к классу 1; при $\lambda \to +1$ — к классу 0; при $\lambda = 0$ разделяющая поверхность проходит посередине зазора.

Теорема 4 ([2]). *Если выборка $X \subseteq \mathbb{X}$ монотонна и функция $r(x, u)$ удовлетворяет свойствам 1), 2), то функция $a(x)$ монотонно не убывает по x и проходит через точки обучающей выборки: $a(x_i) = y_i$ для всех $x_i \in X$.*

Следствие 1. *Классификатор ближайшего соседа является монотонным, и для него верна точная оценка полного скользящего контроля* (3).

В случае $\mathbb{X} = \mathbb{R}^n$ функцию $r(x, u)$ можно задать следующим образом. Положим для произвольных $x = (x^1, \ldots, x^n)$, $u = (u^1, \ldots, u^n)$

$$r(x, u) = \varphi\left((u^1 - x^1)_+, \ldots, (u^n - x^n)_+\right),$$

где индекс «+» обозначает операцию срезки: $z_+ = z \cdot [z \geqslant 0]$; функция $\varphi(z^1, \ldots, z^n)$ не убывает на всей области определения $[0, +\infty)^n$ и принимает нулевое значение $\varphi(z^1, \ldots, z^n) = 0$ тогда и только тогда, когда $z^1 = \cdots = z^n = 0$. В качестве функции φ подходят: максимум, сумма, p-норма, число ненулевых аргументов. Произведение и минимум не подходят, так как они принимают нулевые значения не только в точке $(0, \ldots, 0)$.

Взаимосвязь между профилями компактности и монотонности. Верхняя оценка CCV по профилю монотонности (1) может быть получена в результате ослабления точной оценки CCV по профилю компактности (3). Для доказательства достаточно заметить, что все объекты x из клина W_i являются первыми $m = |W_i|$ ближайшими соседями объекта x_i, поскольку для них и только для них расстояния $\rho_\lambda(x_i, x)$ равны нулю. Начиная с $(m+1)$-го соседа, могут возникать ошибки. Оценка худшего случая (при всех $m = 1, \ldots, L-1$)

$$\big[y_i \neq y(x_{im})\big] \leqslant [w_i < m] = \sum_{s=0}^{m-1} [w_i = s]$$

при подстановке в (3) даёт оценку (1).

Восстановление частично заданных монотонных булевых функций

Пусть объекты описываются n бинарными признаками, $\mathbb{X} = \{0, 1\}^n$. Тогда задача построения монотонного классификатора совпадает с задачей восстановления частично заданной монотонной булевой функции. Известны алгоритмы, решающие данную задачу при дополнительном требовании минимизации сложности получаемой булевой функции $a(x)$ [8]. Критерий обобщающей способности CCV является новым в данной задаче.

Дополнительное затруднение связано с тем, что естественное в данном случае расстояние Хэмминга $\rho(x, x')$ принимает дискретный набор значений $0, \ldots, n$, что не позволяет однозначно ранжировать соседей и определять $y(x_{im})$ — класс m-го соседа объекта x_i. В результате монотонные классификаторы вида (2) могут отличаться на объектах, равноудалённых от границ классов 0 и 1. Для получения верхней оценки CCV рассмотрим пессимистичный метод обучения μ_p, который строит классификатор $\mu_p(X)$, ошибающийся на всех таких объектах. Нижняя оценка получается для оптимистичного метода μ_o, при котором $\mu_o(X)$ правильно классифицирует все такие объекты.

Для каждого $x_i \in \mathbb{X}$ определим три функции целочисленного аргумента $v \in \{0, \ldots, n\}$:

$$t_i(v) = \sum_{x_j \in \mathbb{X} \setminus x_i} \big[\rho_\lambda(x_i, x_j) < (1 + \lambda - 2y_i\lambda)v\big];$$

$$s_i(v) = \sum_{x_j \in \mathbb{X} \setminus x_i} \big[\rho_\lambda(x_i, x_j) = (1 + \lambda - 2y_i\lambda)v\big]\big[y_i \neq y_j\big];$$

$$p_i(v) = \sum_{x_j \in \mathbb{X} \setminus x_i} \big[\rho_\lambda(x_i, x_j) = v\big]\big[y_i = y_j\big].$$

Теорема 5. *Пусть выборка $\mathbb{X}$ монотонна. Тогда $C_o \equiv C(\mu_o, \mathbb{X}) \leqslant C(\mu, \mathbb{X}) \leqslant C(\mu_p, \mathbb{X}) \equiv C_p$, где*

$$C_p = \sum_{i=1}^{L} \sum_{v=0}^{n} \frac{C^\ell_{L-t_i(v)-1} - C^\ell_{L-t_i(v)-s_i(v)-1}}{kC^\ell_L}.$$

$$C_o = \sum_{i=1}^{L} \sum_{v=0}^{n} \frac{C^\ell_{L-t_i(v)-p_i(v)-1} - C^\ell_{L-t_i(v)-s_i(v)-p_i(v)-1}}{kC^\ell_L}.$$

Численный эксперимент

Цель эксперимента — проверить гипотезу, что проведение разделяющей поверхности посередине зазора ($\lambda = 0$) минимизирует CCV.

Эксперимент проводился на бинарных модельных данных, при различных значениях n, L, ℓ, при различной сбалансированности классов, при различной форме зазора. Практически во всех экспериментах наблюдался чётко выраженный минимум CCV при $\lambda = 0$, при этом оценки C_o и C_p были очень близки, рис. 3. В задачах с малыми выборками или существенно несбалансированными классами положение минимума иногда незначительно отличалось от $\lambda = 0$.

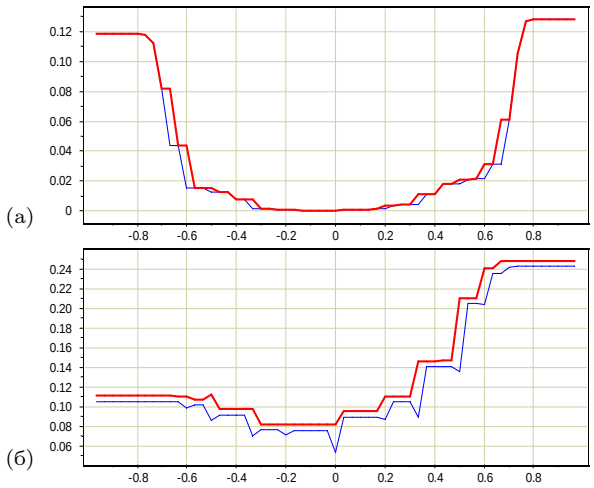


Рис. 3. Зависимость C_o и C_p от λ при $\ell = k = 100$, $n = 15$: (а) выборка с широким зазором и числом объектов в классах $(100, 100)$; (б) выборка с зазором сложной формы и несбалансированными классами $(150, 50)$.

Литература

[1] *Воронцов К. В.* О проблемно-ориентированной оптимизации базисов задач распознавания // ЖВМ и МФ. — 1998. — Т. 38, № 5. — С. 870–880.

[2] *Воронцов К. В.* Оптимизационные методы линейной и монотонной коррекции в алгебраическом подходе к проблеме распознавания // ЖВМ и МФ. — 2000. — Т. 40, № 1. — С. 166–176.

[3] *Воронцов К. В.* Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики / под ред. О. Б. Лупанова. — М.: Физматлит, 2004. — Т. 13. — С. 5–36.

[4] *Гуз И. С.* Нелинейные монотонные композиции классификаторов // всеросс. конф. Математические методы распознавания образов, ММРО-13. — М.: МАКС Пресс, 2007. — С. 111–114.

[5] *Донской В. И.* Алгоритмы обучения, основанные на построении решающих деревьев // ЖВМ и МФ. — 1982. — Т. 22, № 4. — С. 963–974.

[6] *Журавлёв Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Пробл. киберн. — 1978. — Т. 33. — С. 5–68.

[7] *Махина Г. А.* Оценка обобщающей способности для монотонных алгоритмов классификации // 16-я межд. конф. «Проблемы теоретической кибернетики». Нижний Новгород, 20–25 июня 2011.

[8] *Akers S. B.* A truth table method for synthesis of combinational logic // IRE Trans.. — 1961. — Vol. EC-10, no. 4. — Pp. 604–615.

[9] *Boucheron S., Bousquet O., Lugosi G.* Theory of classification: A survey of some recent advances // ESAIM: Probability and Statistics. — 2005. — no. 9. — Pp. 323–375.

[10] *Sokolova M., Marchand M., Japkowicz N., Shawe-Taylor J.* The decision list machine // Advances in Neural Information Processing Systems 15. — MIT-Press, Cambridge, MA, USA, 2003. — Pp. 921–928.

[11] *Kuncheva L.* Combining pattern classifiers. — John Wiley & Sons, Inc., 2004.

[12] *Mullin M., Sukthankar R.* Complete cross-validation for nearest neighbor classifiers // Proc. of Int'l Conf. on Machine Learning. — 2000. — Pp. 639–646.

[13] *Schapire R.* The boosting approach to machine learning: An overview // MSRI Workshop on Nonlinear Estimation and Classification, Berkeley, CA. — 2001.

Гибридные оценки полного скользящего контроля для монотонных классификаторов*

Гуз И. С.
ivan.guzz@gmail.com
Московский физико-технический институт (государственный университет)

Рассматривается задача обучения монотонного классификатора по выборке, которая не обязательно является монотонной. Цель работы — получение оценки полного скользящего контроля, которая могла бы быть использована для повышения обобщающей способности монотонных алгоритмических композиций.

Алгоритмические композиции позволяют строить надёжные классификаторы с высокой обобщающей способностью из ненадёжных базовых классификаторов [1, 3, 5, 6, 13, 14]. В данной работе рассматривается случай, когда классов два, -1 и $+1$; каждый базовый алгоритм определяет действительную оценку принадлежности объекта классу $+1$; эти оценки объединяются с помощью монотонной функции в пространстве оценок принадлежности, называемой монотонной корректирующей операцией [6]. Требование монотонности оправдано тем, что если для одного объекта оценки принадлежности не меньше, чем для другого, то и оценка принадлежности первого объекта, рассчитанная с помощью композиции, должна быть не меньше, чем для второго. Монотонные корректирующие операции образуют более широкое семейство по сравнению с выпуклыми (линейными с неотрицательными коэффициентами), используемыми в методах голосования, в частности, в бустинге [13]. Это позволяет точнее настраиваться на данные и использовать существенно меньшее число базовых алгоритмов, но, как было показано в [3], повышает риск переобучения.

Цель работы состоит в получении как можно более точной верхней оценки полного скользящего контроля для монотонных классификаторов. Ранее такие оценки были получены для семейства всех монотонных классификаторов [2], однако они не достаточно точны именно в силу своей общности. Точные оценки пока получены только для случая, когда исходная выборка является монотонной [7, 8]. Однако при построении монотонных корректирующих операций выборка, как правило, является не монотонной, но «почти монотонной», в том смысле, что путём удаления небольшой доли объектов её можно сделать монотонной.

В данной работе снимается требование монотонности выборки, и точность оценки повышается за счёт сужения семейства до множества монотонных классификаторов ближайшего соседа.

Работа поддержана РФФИ (проект №11-07-00480) и программой ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

Монотонные классификаторы

Задано конечное множество $\mathbb{X} = \{x_1, \ldots, x_L\}$, состоящее из L объектов, в котором каждый объект x_i описывается вектором из n вещественных признаков $x_i^1, \ldots, x_i^n \in \mathbb{R}^n$. Каждому объекту x_i соответствует метка класса $y_i \equiv y(x_i) \in \{-1, +1\}$.

Введем на множестве объектов отношение порядка: $x_i \leqslant x_j$, если $x_i^k \leqslant x_j^k$ для всех $k = 1, \ldots, n$. Объекты x_i и x_j несравнимы, если найдутся такие k, t: $x_i^k < x_j^k$, $x_i^t > x_j^t$. Назовем множество $\mathbb{X}$ *генеральной выборкой*, и будем считать, что среди объектов нет двух одинаковых.

Выборка $X \subseteq \mathbb{X}$ называется монотонной, если для любых $x_i, x_j \in X$ таких, что $x_i \leqslant x_j$, выполняется $y_i \leqslant y_j$.

Назовем пару объектов (x_i, x_j) генеральной выборки *дефектной парой*, если $x_i < x_j$ и $y_i > y_j$. Назовем объект x_i *дефектным*, если он входит хотя бы в одну дефектную пару, и *бездефектным*, если он не входит ни в одну дефектную пару. Обозначим через D_0 множество всех дефектных объектов генеральной выборки.

Назовем *клином* объекта $x_i \in \mathbb{X}$ множество, определяемое следующим образом [2]:

$$W(x_i) = \begin{cases} x_k \in \mathbb{X}, & x_i < x_k,\ y_i = y_k = -1; \\ x_k \in \mathbb{X}, & x_k < x_i,\ y_i = y_k = +1. \end{cases}$$

Бездефектным клином объекта $x_i \in \mathbb{X}$ назовем множество $\bar{W}(x_i) = W(x_i)/D_0$. Обозначим через w_i мощность клина объекта x_i, через $\bar{w}_i$ — мощность бездефектного клина объекта x_i.

Рассмотрим семейство A всех *монотонных классификаторов* вида $a\colon \mathbb{R}^n \to \{-1, +1\}$:

$$a \in A \Leftrightarrow \big(\forall x, x' \in \mathbb{R}^n\ x \leqslant x' \Rightarrow a(x) \leqslant a(x')\big).$$

Задана бинарная функция $I\colon A \times \mathbb{X} \to \{0, 1\}$, называемая *индикатором ошибки*. Если $I(a, x) = 1$, то классификатор $a \in A$ допускает ошибку на объекте x.

Рассмотрим подсемейство *монотонных классификаторов ближайшего соседа*, имеющих вид

$$a(x) = y\big(\arg\min_{x' \in U} \rho(x, x')\big), \qquad (1)$$

где U — некоторая монотонная подвыборка $\mathbb{X}$, функция расстояния от классифицируемого объекта x до объекта $x' \in U$ зависит от класса объекта x' и определяется следующим образом:

$$\rho(x,x') = \begin{cases} r_1(x,x'), & y(x') = +1; \\ r_0(x,x'), & y(x') = -1; \end{cases}$$

функции r_1 и r_0 определяются как

$$r_1(x_i,x_j) = \max(x_j^1 - x_i^1, \ldots, x_j^n - x_i^n, 0);$$
$$r_0(x_i,x_j) = \max(x_i^1 - x_j^1, \ldots, x_i^n - x_j^n, 0).$$

В [1] доказано, что построенный таким образом классификатор $a(x)$ является монотонной функций на $\mathbb{R}^n$.

Методом обучения называется отображение $\mu\colon 2^{\mathbb{X}} \to A$, которое произвольной выборке $X \subseteq \mathbb{X}$ ставит в соответствие некоторый алгоритм $a \in A$.

Рассмотрим метод обучения μ, минимизирующий эмпирический риск, то есть выбирающий из семейства A такой монотонный алгоритм, для которого количество ошибок классификации на выборке X минимально:

$$\mu(X) = \arg\min_{a \in A} \sum_{x \in X} I(a,x).$$

Назовем множество объектов генеральной выборки, на которых алгоритм $\mu(\mathbb{X})$, минимизирующий эмпирический риск, ошибается, множеством *немонотонности*, и обозначим его D. *Степень немонотонности* δ генеральной выборки $\mathbb{X}$ определим как $|D|/L$. В [5] предложен способ построения монотонного алгоритма, минимизирующего эмпирический риск, а также способ построения множества D.

Назовем *обучающей выборкой* X подмножество генеральной выборки $\mathbb{X}$ мощности ℓ, а *контрольной выборкой* $\bar{X}$ — множество $\mathbb{X} \setminus X$ мощности $k = L - \ell$. Функционал *полного скользящего контроля* CCV (complete cross-validation) определяется для генеральной выборки $\mathbb{X}$, семейства алгоритмов A и метода обучения μ как средняя частота ошибок на контроле алгоритмов, построенных методом μ по всем C_L^ℓ обучающим подвыборкам:

$$\mathrm{CCV} = \frac{1}{C_L^\ell} \sum_{X \sqcup \bar{X} = \mathbb{X}} \frac{1}{k} \sum_{x \in \bar{X}} I\big(\mu(X), x\big).$$

Этот функционал характеризует обобщающую способность метода μ на выборке $\mathbb{X}$. Целью данной работы является получение верхних оценок этого функционала для метода минимизации эмпирического риска и семейства монотонных классификаторов ближайшего соседа.

Гибридная верхняя оценка CCV

Заметим, что выборка $\mathbb{X} \setminus D$ является монотонной, то есть для всех $x, x' \in \mathbb{X} \setminus D$ из $x \leqslant x'$ следует $y(x) \leqslant y(x')$.

Для каждого объекта $x_i \in \mathbb{X}$ удалим из генеральной выборки множество $D \setminus \{x_i\}$. Обозначим через D_i мощность множества $D \setminus \{x_i\}$, тогда есть $D_i = |D| - [x_i \in D]$. Оставшиеся объекты упорядочим по возрастанию расстояния от объекта x_i, пронумеровав их двойными индексами: $x_{i,1}, \ldots, x_{i,L-1-D_i}$. Таким образом,

$$\rho(x_i, x_{i,1}) \leqslant \rho(x_i, x_{i,2}) \leqslant \ldots \leqslant \rho(x_i, x_{i,L-1-D_i}).$$

Обозначим через $I_m(x_i)$ ошибку, возникающую, если правильный ответ y_i для объекта x_i заменить ответом на его m-ом соседе:

$$I_m(x_i) = [y_i \neq y_{i,m}],\ m = 1, \ldots, L - 1 - D_i.$$

Используя введенные обозначения, доказывается верхняя оценка CCV.

Теорема 1. *Если μ минимизирует эмпирический риск в классе всех монотонных функций, то:*

$$\mathrm{CCV} \leqslant \sum_{i=1}^{L} \frac{1}{L} \Bigg(\sum_{d=q}^{D_i} \frac{C_{D_i}^d C_{L-1-D_i-\bar{w}_i}^{\ell-d}}{C_{L-1}^\ell} + \\ + [D_i + \bar{w}_i < k] \sum_{m=1}^{k-D_i} \frac{I_m(x_i) C_{L-1-D_i-m}^{\ell-1}}{C_{L-1}^\ell} \Bigg), \quad (2)$$

где $q = \max\{1, D_i + \bar{w}_i + 1 - k\}$.

Оценка (2) состоит из двух слагаемых.

Первое слагаемое объясняет зависимость CCV от степени немонотонности генеральной выборки. Чем больше мощность множества немонотонности D, тем больший вклад дает это слагаемое в оценку CCV. Поэтому назовем первое слагаемое *немонотонной частью* оценки CCV (2).

Второе слагаемое объясняет зависимость CCV от структуры монотонного алгоритма, то есть от выбора функций r_1 и r_0. Поэтому назовем второе слагаемое *структурной частью* оценки CCV (2).

Поскольку оценка CCV (2) учитывает как свойства выборки, так и структуру алгоритмов, назовем ее *гибридной*.

Основное преимущество гибридной оценки в том, что чем более монотонная выборка, тем ближе эта оценка к точному значению CCV. Если генеральная выборка является монотонной, то есть $|D| = 0$, то гибридная оценка совпадает с оценкой [8] и является точной.

Рассмотрим далее точность гибридной оценки CCV на модельных задачах и сравним ее с комбинаторной оценкой CCV, доказанной в [2]:

$$\mathrm{CCV} \leqslant \sum_{m=0}^{\delta L + k - 1} M(m, \mathbb{X}) \sum_{s=s_1}^{s_2} \frac{C_m^s C_{L-1-m}^{\ell-s}}{C_{L-1}^\ell},$$

где $M(m,\mathbb{X})$ — *профиль монотонности*, определяемый как доля объектов выборки $\mathbb{X}$ с клином мощности m; $s_1 = \max\{0, m-k+1\}$, $s_2 = \min\{\delta L, \ell, m\}$.

Экспериментальное исследование точности гибридной оценки CCV

Расчет оценок CCV проводился на 400 различных генеральных выборках, где каждая выборка состоит из 100 объектов класса -1 и 100 объектов класса $+1$, $x_i \in \mathbb{R}^3$. Значения признаков x_i^j, $j = 1,2,3$ генерировались независимо друг от друга для каждого объекта x_i случайным образом из нормальных распределений. Математическое ожидание бралось равным 5 для объектов класса -1, и 10 для объектов класса $+1$. Дисперсия нормального распределения при каждой генерации выбиралась случайным образом из отрезка $[0;7{,}5]$. Данный способ генерации позволил создать генеральные выборки с различной степенью монотонности, от строго монотонных выборок до существенно немонотонных.

Расчет точного значения CCV производился с помощью скользящего контроля по 100 случайным разбиениям генеральной выборки на обучающую и контрольную выборки. Длина обучающей выборки бралась равной 140 (70% от числа всех объектов).

На рис. 1–3 приводятся зависимости комбинаторной оценки CCV (белые точки) и гибридной оценки CCV (красные точки) от точного значения CCV для 400 выборок. Линией на всех рисунках показано точное значение CCV. На рис. 1 рассматривается одномерный случай, где каждый объект x_i описывается только первым признаком, на рис. 2 — первыми двумя признаками, на рис. 3 — всеми тремя признаками. Чем большим количеством признаков описывается каждый объект генеральной выборки, тем больше в ней несравнимых пар объектов. Эксперименты показывают, что точность оценок CCV существенно зависит от количества пар несравнимых объектов.

Таким образом, в одномерном случае, когда все объекты одной генеральной выборки сравнимы, использовать гибридную оценку CCV осмысленно только для достаточно монотонных выборок. При увеличении степени немонотонности гибридная оценка CCV начинает существенно уступать по точности комбинаторной оценке CCV, которая, в свою очередь, в разы превышает точное значение CCV. При появлении в выборках пар несравнимых объектов гибридная оценка CCV становится точнее комбинаторной оценки CCV (рис. 2 и 3). Причем в трехмерном случае на генеральных выборках, являющихся монотонными, гибридная оценка CCV совпадает с точной оценкой CCV (рис. 3).

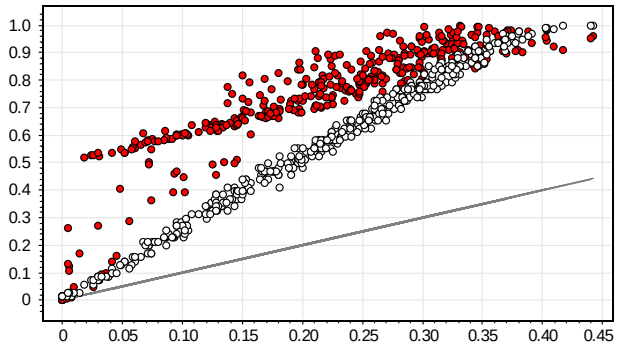


Рис. 1. Одномерные генеральные выборки, все объекты сравнимы. Комбинаторная оценка на большинстве выборок точнее гибридной оценки CCV.

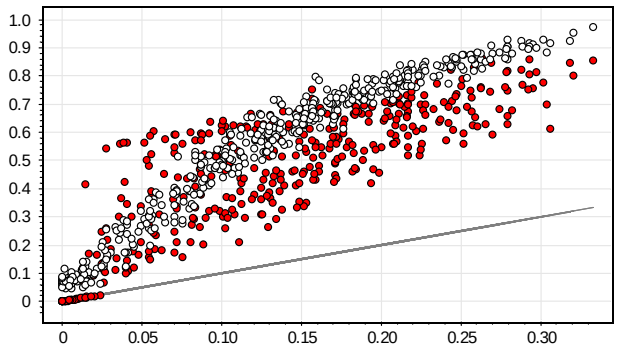


Рис. 2. Двумерные генеральные выборки содержат пары несравнимых объектов. Гибридная оценка на большинстве выборок точнее комбинаторной оценки CCV.

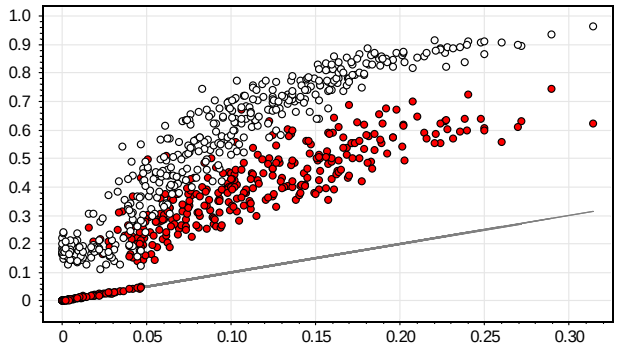


Рис. 3. Трёхмерные генеральные выборки содержат большое число пар несравнимых объектов. Гибридная оценка почти на всех выборках существенно точнее комбинаторной оценки CCV.

На рис. 4 показана зависимость структурной части гибридной оценки CCV от значения гибридной оценки CCV для трехмерного случая. Начальный прямой участок графика соответствует монотонным выборкам.

Рис. 5 показывает, что при увеличении степени немонотонности δ генеральной выборки вклад структурной части резко уменьшается и начинает преобладать немонотонная часть.

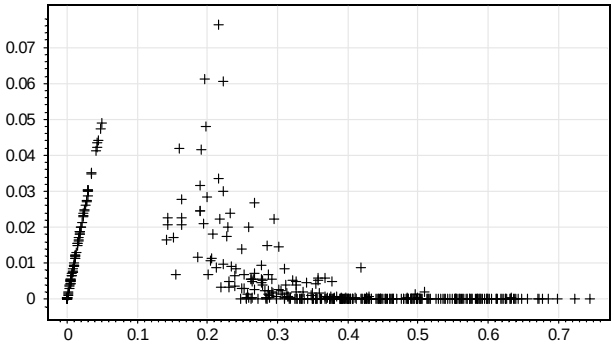


Рис. 4. Зависимость структурной части гибридной оценки CCV от значения гибридной оценки CCV для трехмерного случая.

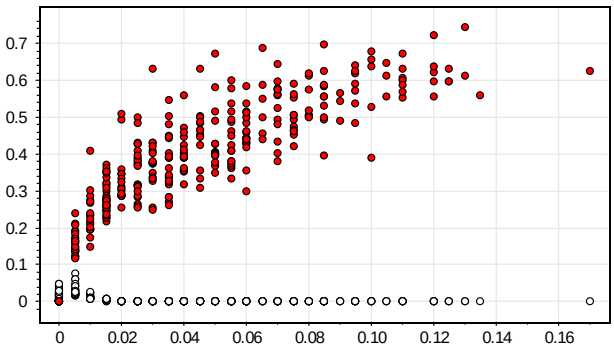


Рис. 5. Зависимость структурной части гибридной оценки CCV (белые точки) и немонотонной части гибридной оценки CCV (красные точки) от степени немонотонности δ генеральной выборки.

Выводы

Полученная гибридная оценка CCV хорошо аппроксимирует точное значение CCV для выборок большой размерности и малой степени немонотонности. Она может быть использована при построении монотонных композиций алгоритмов. Действительно, в [3] было показано, что уже после добавления второго алгоритма в композицию, генеральная выборка в пространстве оценок алгоритмов становится достаточно монотонной и содержит большое количество пар несравнимых объектов.

Открытым остается вопрос реализации алгоритма построения монотонных композиций, где выбор очередного алгоритма для добавления в композицию был бы основан на минимизации гибридной оценки CCV всей композиции.

Литература

[1] *Воронцов К. В.* О проблемно-ориентированной оптимизации базисов задач распознавания // ЖВМ и МФ. — 1998. — Т. 38, № 5. — С. 870–880.

[2] *Воронцов К. В.* Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики. — 2004. — № 13. — С. 5–36.

[3] *Гуз И. С.* Нелинейные монотонные композиции классификаторов // Математические методы распознавания образов-13, М.: МАКС Пресс, 2007. — С. 111–114.

[4] *Гуз И. С.* Исследование обобщающей способности семейства монотонных функций // Сборник трудов МФТИ. Моделирование и обработка информации, 2008.

[5] *Гуз И. С.* Минимизация эмпирического риска при построении монотонных композиций классификаторов // Сборник трудов МФТИ, 2011. — С 89–97.

[6] *Журавлёв Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. — 1978. — Т. 33. — С. 5–68.

[7] *Махина Г. А.* Оценка обобщающей способности для монотонных алгоритмов классификации // 16-я международная конференция. Проблемы теоретической кибернетики. Нижний Новгород, 20-25 июня 2011.

[8] *Махина Г. А.* Принцип максимизации зазора для монотонного классификатора ближайшего соседа // сборник докладов 15-й всероссийской конференции Математические методы распознавания образов, 2011. — С. 64–68.

[9] *Рудаков К. В., Воронцов К. В.* О методах оптимизации и монотонной коррекции в алгебраическом подходе к проблеме распознавания // Доклады РАН. — 1999. — Т. 367, № 3. — С. 314–317.

[10] *Agresti A.* Building and applying logistic regression models // An Introduction to Categorical Data Analysis. — Hoboken, New Jersey: Wiley. — p 138.

[11] *Breiman L., Friedman J. H., Olshen R. A., Stone C. J.* Classification and regression trees. — Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984.

[12] *Cortes C., Vapnik V.* Support-Vector Networks // Machine Learning. — 1995. — V. 20.

[13] *Freund Y., Schapire R. E.* A decision-theoretic generalization of on-line learning and an application to boosting // European Conference on Computational Learning Theory. — 1995. — Pp. 23–37.

[14] *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning. — Springer, 2001.

[15] *Kass. G. V.* An Exploratory Technique for Investigating Large Quantities of Categorical Data // Applied Statistics. — 1980. — V. 29, N. 2. — Pp. 119–127.

[16] *Loh W. Y. and Shih Y. S.* Split selection methods for classification trees // Statistica Sinica. — 1997. — V. 7. — Pp. 815–840.

Задачи построения линейных и нелинейных классификаторов в случае многих классов*

Журавлев Ю. И., Лаптин Ю. П., Виноградов А. П.
zhuravlev@ccas.ru, laptin_yu_p@mail.ru, vngrccas@mail.ru
Москва, Вычислительный Центр им. А. А. Дородницына РАН
Киев, Институт кибернетики им. В. М. Глушкова НАН Украины

Формулируются специальные задачи выпуклого программирования для построения линейных классификаторов в случае многих классов, как для линейно разделимых, так и для линейно неразделимых обучающих выборок. Производится сравнение с методом опорных векторов. Результаты обобщаются для задач построения нелинейных классификаторов. Рассматриваются методы решения сформулированных задач.

Проблемы построения классификаторов, линейных и нелинейных, обычно рассматриваются для двух непересекающихся множеств. Для линейно разделимых множеств естественным образом формулируются и эффективно решаются задачи построения оптимальных классификаторов. Для линейно неразделимых множеств как формулировка, так и решение задач построения классификаторов наталкиваются на определенные трудности. При построении линейных классификаторов в этом случае часто используются эвристические алгоритмы. Широко используется метод опорных векторов. Альтернативой для линейно неразделимых множеств является переход к нелинейным классификаторам - квадратичным, полиномиальным. Повидимому, увеличивая сложность классификатора, всегда можно добиться правильного разделения исходных множеств. Такая цель не всегда оправданна, поскольку исходные множества (обучающие выборки) могут формироваться в условиях возможного возникновения тех, или других ошибок. Естественным в этой ситуации является построение классификатора заданного типа (заданной сложности), минимизирующего число точек обучающей выборки, разделяемых неправильно (минимизация эмпирического риска).

Постановка задачи

Для заданной совокупности конечных непересекающихся множеств $\Omega_i \subset R^n$, $i = 1, \ldots, m$ (обучающей выборки из m непересекающихся классов) рассматривается задача построения алгоритма классификации следующего вида

$$a(x, W) = \arg\max_i \Big\{ f_i\left(x, W^i\right) : i = 1, \ldots, m \Big\}, \quad x \in R^n,\ W^i \in R^{L_i+1}, \quad (1)$$

где W^i — совокупность параметров функции $f_i\left(x, W^i\right)$, $i = 1, \ldots, m$, $W = \left(W^1, \ldots, W^m\right)$. Функции $f_i\left(x, W^i\right)$ называются (приоритетными) весовыми функциями.

Работа поддержана совместным грантом НАН Украины и РФФИ № 10-01-90419.

Говорят, что классификатор $a(x, W)$ правильно разделяет точки из Ω_i, $i = 1, \ldots, m$, если $a(x, W) = i$, для всех $x \in \Omega_i$, $i = 1, \ldots, m$. Будем предполагать, что для каждого класса i задана совокупность базисных функций $\varphi_0^i(x) \equiv 1$, $\varphi_j^i(x), j = 1, \ldots, L_i$, на основе которых формируются весовые функции $f_i\left(x, W^i\right)$, $W^i = \left(w_0^i, w_1^i, \ldots, w_{L_i}^i\right)$, $i = 1, \ldots, m$,

$$f_i\left(x, W^i\right) = \sum_{j=1}^{L_i} w_j^i \varphi_j^i(x) + w_0^i. \quad (2)$$

Совокупности базисных функций могут совпадать для разных классов, например, $\varphi_j^i(x) = x_j$, $j = 1, \ldots, n$, $i = 1, \ldots, m$ для линейных классификаторов. В общем случае совокупность базисных функций для каждого класса может быть своя.

Множества Ω_i, $i = 1, \ldots, m$ называются разделимыми в заданной совокупности базисных функций, если существует классификатор, сформированный на основе этой совокупности базисных функций, правильно разделяющий точки из этих множеств.

Обозначим $\Omega = \bigcup_{i=1}^m \Omega_i$. Пусть точки множества Ω перенумерованы, T — совокупность индексов, $\Omega = \{x^t : t \in T\}$, T_i — совокупность индексов множества Ω_i, $\Omega_i = \{x^t : t \in T_i\}$, $T = \bigcup_{i=1}^m T_i$. Положим $i(t)$ — номер множества Ω_i, которому принадлежит точка x^t, $t \in T$. Величина

$$g^t(W) = \min\left\{ f_i\left(x^t, W^i\right) - f_j\left(x^t, W^j\right) : j \in \{1, \ldots, m\} \setminus i,\ i = i(t) \right\} \quad (3)$$

называется отступом или зазором (gap) классификатора $a(x, W)$ на точке x^t, $t \in T$.

Величина $g(W) = \min\{g^t(W) : t \in T\}$ называется зазором классификатора $a(x, W)$ на совокупности множеств $\Omega_i, i = 1, \ldots, m$. Классификатор $a(x, W)$ правильно разделяет точки из $\Omega_i, i = 1, \ldots, m$, если $g(W) > 0$.

Замечание 1. Классификатор $a(x, W)$ инвариантен относительно умножения всех функций f_i на положительное число, зазор $g(W)$ линеен относительно такой операции умножения.

Величину $g(W)$ можно использовать как критерий качества классификатора $a(x, W)$ (чем боль-

ше значение $g(W)$, тем надёжнее разделяются точки из Ω_i, $i = 1, \ldots, m$), однако, при этом должна учитываться некоторая нормировка совокупности векторов W, которую обозначим $\eta(W)$ и будем называть нормой классификатора $a(x, W)$. Вопрос выбора нормировки имеет самостоятельное значение и должен рассматриваться отдельно. В дальнейшем будем использовать функцию:

$$\eta(W) = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{L_i} \left(w_j^i\right)^2}. \qquad (4)$$

С учетом введенных обозначений задачу построения оптимального классификатора (определения значений параметров W) запишем в следующем виде: найти

$$g^* = \max_W \Big\{ g(W) : \eta(W) \leqslant 1, W \in R^L \Big\}, \qquad (5)$$

где L — размерность вектора W. Поскольку вектор $W = 0$ является допустимым, то задача (5) имеет решение всегда и $g^* \geqslant g(0) = 0$. Заметим, что $g^* > 0$, если множества Ω_i, $i = 1, \ldots, m$ разделимы в заданной совокупности базисных функций, т.е. существует классификатор, правильно разделяющий эти множества. Рассмотрим также задачу: найти

$$\eta^* = \min_V \Big\{ \eta(V) : g(V) \geqslant 1, \quad V \in R^L \Big\}. \qquad (6)$$

Лемма 1 [1]. Пусть W^* — оптимальное решение задачи (5). Тогда если $g^* > 0$, то задача (6) также имеет оптимальное решение V^* и $V^* = W^*/g^*$, $\eta^* = 1/g^*$. Если $g^* = 0$, то задача (6) не имеет допустимых решений.

Замечание 2. Задачи (5) и (6) позволяют находить оптимальный классификатор только в случае, когда множества Ω_i, $i = 1, \ldots, m$ разделимы в заданной совокупности базисных функций.

Линейные классификаторы и минимизация эмпирического риска

Если функции $f_i\left(x, W^i\right)$ линейны,

$$f_i\left(x, W^i\right) = \left(w^i, x\right) + w_0^i, \quad i = 1, \ldots, m,$$

то классификатор $a(x, W)$ называется линейным. Задачи построения таких классификаторов рассматривались в [1–3], условия линейной разделимости множеств для произвольного m — в [1, 4].

В случае линейных классификаторов задача (5) есть задача линейного программирования с одним дополнительным квадратичным ограничением, (6) — задача квадратичного программирования. В случае $m = 2$ задача (6) эквивалентна задаче, которая решается с целью построения полосы максимальной ширины, разделяющей линейно разделимые множества Ω_1, Ω_2.

В случае линейно неразделимой выборки задача (6) не имеет решений, а решение задачи (5) не несет полезной информации. Естественным критерием выбора классификатора в этом случае является минимизация эмпирического риска.

Пусть задано $\bar{\delta} > 0$. Будем говорить, что точка $x^t, t \in T$ обрабатывается классификатором $a(x, W)$ ненадежно, если $g^t(W) < \bar{\delta}$. Определим эмпирический риск как число точек обучающей выборки, которые классификатор обрабатывает неправильно или ненадежно.

Каждой точке x^t, $t \in T$, поставим в соответствие переменную $y_t = 0 \vee 1$ так, что $y_t = 0$, если точка x^t учитывается при формировании задачи (5), $y_t = 1$ — в противном случае.

Пусть задано большое положительное число B. Задача минимизации эмпирического риска с учетом надежности, определяемой параметром $\bar{\delta}$, имеет вид [3]: найти

$$Q^* = \min_{w,y} \Big\{ \sum_{t \in T} y_t \Big\}, \qquad (7)$$

при ограничениях

$$\left(w^i - w^j, x^t\right) + w_0^i - w_0^j \geqslant \bar{\delta} - B \cdot y_t, \quad j \in \{1, \ldots, m\} \setminus i, \ t \in T_i, \ i = 1, \ldots, m, \qquad (8)$$

$$\eta(W) \leqslant 1, \qquad (9)$$

$$\sum_{t \in T_i} y_t \leqslant |T_i| - 1, \quad i = 1, \ldots, m, \qquad (10)$$

$$0 \leqslant y_t \leqslant 1, \quad t \in T, \qquad (11)$$

$$y_t = 0 \vee 1, \quad t \in T. \qquad (12)$$

Задача (7)–(12) — NP-трудная. Для вычисления оценки снизу q^* для минимального эмпирического риска Q^* рассматривается непрерывная релаксация сформулированной задачи — задача (7)-(11).

Положим $d^t(W) = \max\left(0, \frac{1}{B}\left(\bar{\delta} - g^t(W)\right)\right)$, где $g^t(W)$ определено в соответствии с (3). Релаксированная задача определения q^* и оптимальных значений переменных W приводится [3] к виду

$$q^* = \min \sum_{t \in T} d^t(W) \qquad (13)$$

при ограничениях

$$\eta(W) \leqslant 1, \qquad (14)$$

$$\sum_{t \in T_i} d^t(W) \leqslant |T_i| - 1, \quad i = 1, \ldots, m, \qquad (15)$$

$$d^t(W) \leqslant 1, \quad t \in T. \qquad (16)$$

Функции $d^t(W)$ — выпуклые кусочно-линейные, $\eta(W)$ — квадратичная положительно определенная. Задача (13)–(16) для случая линейных классификаторов рассматривалась в [3]. Было показано,

что лагранжева релаксация этой задачи при специальном подборе значений множителей Лагранжа эквивалентна задаче, которая решается в методе опорных векторов.

Задача (13)–(16) является специальной задачей выпуклого программирования и для ее решения целесообразно применять эффективные методы негладкой оптимизации [6]. Один из распространенных подходов к решению таких задач заключается в применении метода негладких штрафных функций, позволяющего сформировать эквивалентную задачу безусловной оптимизации. После этого для решения задачи безусловной оптимизации используются ускоренные методы субградиентного спуска с растяжением пространства. Некоторые сложности при таком подходе связаны с выбором штрафных коэффициентов.

В настоящее время предложены новые методы построения эквивалентной задачи безусловной оптимизации, основанные на выпуклом продолжении функций и конической регуляризации исходной задачи [7]. Их эффективность подтверждена вычислительным экспериментом. Особенности задачи (13)–(16) позволяют строить быстрые алгоритмы решения вспомогательных задач.

Нелинейные классификаторы

В случае, когда обучающая выборка является линейно неразделимой совокупность базисных функций можно расширить, т. е. от линейных классификаторов можно перейти к квадратичным или к полиномиальным более высокого порядка. Такие переходы к полиномиальным классификаторам должны быть обоснованы содержательно в результате анализа рассматриваемой прикладной задачи, поскольку неразделимость обучающей выборки в заданой совокупности базисных функций может быть связана не со сложностью конфигурации классов, а с ошибками определения координат точек множеств Ω_i, $i = 1, \ldots, m$.

Пусть задана совокупность базисных функций $\varphi_j^i(x), j = 0, \ldots, L_i$, на основе которых формируются весовые функции $f_i\left(x, W^i\right)$, $W^i = \left(w_0^i, w_1^i, \ldots, w_{L_i}^i\right)$, $i = 1, \ldots, m$

Для каждой точки x^t обучающей выборки обозначим $z_j^{it} = \varphi_j^i\left(x^t\right)$, тогда, учитывая (2), зазор $g^t(W)$ можно представить в виде

$$g^t(W) = \min\left\{\sum_{k=1}^{L_i} w_k^i z_k^{it} - \sum_{k=1}^{L_j} w_k^j z_k^{jt} + w_0^i - w_0^j : \right.$$
$$\left. j \in \{1, \ldots, m\} \setminus i, \quad i = i(t)\right\}. \quad (17)$$

Откуда следует [5], что в случае нелинейных классификаторов задача (5) также есть задача линейного программирования с одним дополнительным квадратичным ограничением, а (6) — задача квадратичного программирования.

В случае неразделимости обучающей выборки в заданой совокупности базисных функций необходимо рассматривать задачу минимизации эмпирического риска. Для вычисления оценки снизу для минимального эмпирического риска должна решаться задача (13)–(16), в которой зазор $g^t(W)$ определяется в соответствии с (17).

Выводы

В работе для случая многих классов рассмотрены задачи построения классификаторов, весовые функции которых представимы в виде линейной комбинации заданной совокупности базисных (в общем случае нелинейных) функций. В случае неразделимости обучающей выборки в заданной совокупности классификаторов формулируется задача построения классификатора, обеспечивающего минимизацию эмпирического риска. Такая задача является частично целочисленной и NP-трудной. Для практического использования рассмотрена непрерывная релаксация задачи минимизации эмпирического риска, которая сводится к задаче выпуклого программирования с кусочно-линейными функциями.

Литература

[1] *Laptin Yu. P., Likhovid A. P., Vinogradov A. P.* Approaches to Construction of Linear Classifiers in the Case of Many Classes // Pattern Recognition and Image Analysis. — 2010. — Vol. 20, № 2. — Pp. 137–145.

[2] *Zhuravlev Yu., Laptin Yu., Vinogradov A.* Minimization of empirical risk in linear classifier problem // New Trends in Classification and Data Mining, ITHEA, Sofia, Bulgaria, 2010. — Pp. 9–15.

[3] *Лаптин Ю. П., Журавлев Ю. И., Виноградов А. П.* Минимизация эмпирического риска и задачи построения линейныхклассификаторов // Кибернетика и системный анализ. — 2011. № 4. — С. 155–164.

[4] *Петунин Ю. И., Шульдешов Г. А.* Проблемы распознавания образов с помощью линейных дискриминантных функций Фишера // Кибернетика . — 1979. № 6. — С. 134–137.

[5] *Лаптин Ю. П., Журавлев Ю. И., Виноградов А. П.* Построение нелинейных классификаторов в случае многих классов // Applicable Information Models, ITHEA, Sofia, Bulgaria, 2011. — Pp. 7–13.

[6] *Shor N. Z.* Nondifferentiable Optimization and Polynomial Problems. — Dordrecht, Kluwer, 1998. — 394 p.

[7] *Лаптин Ю. П., Бардадым Т. А.* Некоторые подходы к регуляризации нелинейных задач оптимизации // Проблемы управления и информатики . — 2011. № 3. — С. 57–68.

Эффективное построение ДНФ функций с малым числом нулей*

Романов М. Ю.
mromanov@ccas.ru
Москва, ЗАО «Связной логистика»

В работе показан линейный алгоритм построения тупиковой ДНФ для специального класса булевых функций с малым числом нулей.

В работе рассматриваются булевые функции с малым числом нулей, задаваемые перечислением нулевых точек $\tilde{m}^i = \left(\alpha_1^i, \alpha_2^i, \ldots, \alpha_n^i\right)$, $i = 1, 2, \ldots, k$:

$$F\left(x_1,\ x_2,\ \ldots,\ x_n\right) = \bigwedge_{i=1}^{k} \left(x_1^{\alpha_i^1} \vee x_2^{\alpha_i^2} \vee \ldots \vee x_n^{\alpha_i^n}\right).$$

Согласно [1], класс функций, зависящих от n переменных и имеющих k нулей, будем обозначать P_k^n.

Функции этого класса применяются в большом числе оптимизационных задач, алгоритмов распознавания, задачи синтеза управляющих систем и т. п. В этих задачах необходимо эффективно строить минимальные или хотя бы достаточно простые дизъюнктивные нормальные формы (ДНФ) для рассматриваемого класса функций. В некоторых случаях достаточно построить произвольную ДНФ, а потом проводить процедуры упрощения такие, как методы Квайна или Блейка. В настоящей работе приведён эффективный алгоритм для некоторых функций этого класса.

Для каждой функции F рассмотрим матрицу $M_F = \left\|\alpha_j^i\right\|_{k\times n}$, строки которой являются нулевыми точками для F. Тогда, согласно [1], каждой функции $F \in P_k^n$ можно сопоставить функцию $\Phi \in P_k^l$ такую, что матрица M_Φ обладает следующими свойствами:

1) в M_Φ отсутствуют нулевые и единичные столбцы;
2) одинаковые столбцы в M_Φ расположены последовательно;
3) из любых двух столбцов, один из которых является отрицанием другого, в M_Φ присутствует не более одного.

Получение ДНФ функции F из ДНФ функции Φ далее будет рассмотрено подробнее.

Следуя [2], будем называть функцию F *полной*, если матрица M_F её нулевых точек состоит из всех $2^{k-1} - 1$ различных столбцов, исключая нулевой и единичный, причём из каждых двух двойственных столбцов в M_F содержится ровно один.

В работе [3] исследовался вопрос о количестве максимальных граней.

Работа выполнена при финансовой поддержке РФФИ, проект № 10-07-00726.

В работе [4] продолжено исследование максимальных граней булевых функций с малым числом нулей и получен ряд оценок степеней точек, входящих в максимальные грани.

Построение канонической тупиковой ДНФ

Введём $n_k = 2^{k-1} - 1$ — число переменных полной функции с k нулями. В классе полных функций с k нулями выделим одну функцию, обладающую следующей свойством. В матрице нулей этой функции i-й столбец (где $i = 1, \ldots, n_k$) представляет собой двоичную запись числа i, в которой старшие разряды дополнены спереди нулями для получения k-значной записи. Эту функцию $\mathfrak{F}_k$ назовём *базовой функцией* с k нулями.

Матрица нулей $M_{\mathfrak{F}_k}$ функции $\mathfrak{F}_k$ обладает следующими свойствами.

1) матрица содержит нулевую строку, потому что для любого i от 1 до n_k число значащих двоичных разрядов не превосходит $k - 1$.
2) если из матрицы исключить строку нулей, то последним столбцом будет единичный столбец, так как он соответствует числу $2^{k-1} - 1 = n_k$; кроме того, столбцы с индексами i и $n_k - i$ будут являться отрицанием друг друга при $i = = 1, \ldots, n_{k-1}$.
3) Матрица $M_{\mathfrak{F}_k}$ содержит матрицу $M_{\mathfrak{F}_{k-1}}$ в качестве подматрицы по построению, т. е. если $M_{\mathfrak{F}_k} = \|a_{ij}\|_{k\times n_k}$, то $M_{\mathfrak{F}_{k-1}} = \|a_{ij}\|_{k-1\times n_{k-1}}$.

Определение 1. *Характеристикой столбца матрицы нулей является число, для которого столбец является двоичным представлением.*

Тогда матрицу нулей $M_{k\times n}$ можно записать в виде вектора характеристик длины n. Матрица нулей базовой функции $\mathfrak{F}_k$ представляется в виде вектора чисел от 1 до n_k.

Пример 1. Пример матрицы нулей базовой функции для $k = 4$:

$$\begin{array}{ccccccc} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}$$

Рекурсивно опишем некоторую ДНФ $\mathfrak{D}_k$.

1. Для $k = 3$ определим ДНФ $\mathfrak{D}_3$, как

$$x_1x_2 \vee \overline{x_1}\,\overline{x_2}x_3 \vee x_1\overline{x_3} \vee x_2\overline{x_3}.$$

2. Пусть уже построена ДНФ $\mathfrak{D}_{k-1}$. Для получения ДНФ $\mathfrak{D}_k$ надо добавить следующие конъюнкции:

$$x_i x_{n_k-i},\ i=1,\dots,n_{k-1};$$
$$\overline{x_i}\,\overline{x_{n_k-i}}\,x_{n_k},\ i=1,\dots,n_{k-1};$$
$$x_i\overline{x_{n_k}},\ i=n_{k-1},\dots,n_k-1.$$

Теорема 1. *Рекурсивно описанная ДНФ $\mathfrak{D}_k$ является тупиковой ДНФ базовой функции $\mathfrak{F}_k$.*

Построенную ДНФ $\mathfrak{D}_k$ будем называть *канонической* тупиковой ДНФ для базовой функции $\mathfrak{F}_k$ с k нулями. Эта ДНФ состоит из конъюнкций 3-х видов:

1) $x_i x_{n_p-i}$ — эти конъюнкции соответствуют граням размерности n_k-2 и их $2^{k-1}-k$ штук;
2) $x_i\overline{x_{n_p}}$ — эти конъюнкции соответствуют граням размерности n_k-2 и их $2^{k-1}-2$ штук;
3) $\overline{x_i}\,\overline{x_{n_p-i}}\,x_{n_p}$ — эти конъюнкции соответствуют граням размерности n_k-3 и их $2^{k-1}-k$ штук.

Всего $\mathfrak{D}_k$ содержит $3\cdot 2^{k-1}-2k-2$ конъюнкций. Эта величина асимптотически равна $3n_k$ при $k\to\infty$. По теореме 4 из статьи [1] для базовой функции существует ДНФ с асимптотической оценкой числа конъюнкций равной $2n_k$. Построенная нами ДНФ $\mathfrak{D}_k$ близка к этой теоретической оценке.

Итак, нами предложена ДНФ полной функции, близкая к минимальной, и построение которой по сформулированному рекуррентному описанию имеет линейную сложность по n_k.

Построение тупиковой ДНФ произвольной функции

Согласно статьи [1] матрица нулей любой функции $F\in P_k^n$ может быть получена из матрицы нулей базовой функции $\mathfrak{F}_k$ следующими преобразованиями:

1) изменение порядка столбцов;
2) дублирование столбцов;
3) добавление столбца нулей;
4) инвертирование (отрицание) столбцов;
5) удаление столбцов.

Если имеется некоторая ДНФ базовой функции, то соответствующими преобразованиями можно получить ДНФ функции F.

В этой работе рассмотрим построение ДНФ для функций, которые можно получить из базовой функции любыми из перечисленных операций, за исключением операции «удаление столбцов». При каждом из необходимых преобразований происходит соответствующее изменение ДНФ:

1) изменение порядка столбцов с $(1,\dots,n)$ на $\big(a(1),\dots,a(n)\big)$ приводит к соответствующему изменению переменных в конъюнкциях;
2) дублирование i-го столбца столбцами $n+1,\dots,n+p$ приводит к добавлению конъюнкций

$$x_i\,\overline{x_{n+1}},\ x_{n+t}\,\overline{x_{n+t+1}},\ x_{n+p}\,\overline{x_i},$$

где $t=\overline{1,p-1}$;
3) добавление i-го столбца нулей приводит к добавлению конъюнкции $\overline{x_i}$;
4) инвертирование i-го столбца приводит к замене переменной x_i во всех конъюнкциях на $\overline{x_i}$.

Согласно результатам статьи [1] преобразованная таким образом ДНФ будет тупиковой ДНФ функции F. Все операции, рассмотренные в этом параграфе выполняются за линейное число шагов от числа конъюнкций. При этом результирующее число конъюнкций не превышает величины $n+2^k$.

Выводы

Для специального класса булевых функций получен линейный алгоритм построения тупиковой ДНФ. При этом полученная ДНФ имеет длину асимптотически в $\frac{3}{2}$ раза большую, чем минимальная ДНФ.

Эффективное построение тупиковых ДНФ для более широкого класса функций (в частности функций, получающихся в результате удаления некоторых столбцов матрицы нулей базовой функции) является предметом дальнейших исследований.

Литература

[1] *Журавлёв Ю. И., Коган А. Ю.* Реализация булевых функций с малым числом нулей дизъюнктивными нормальными формами и смежные задачи // Докл. АН СССР — 1985. — Т. 285, № 4. — С. 795–799.

[2] *Журавлёв Ю. И., Коган А. Ю.* Алгоритм построения дизъюнктивной нормальной формы, эквивалентной произведению левых частей булевых уравнений Нельсоновского типа // Ж. вычисл. матем. и матем. физ. — 1986. — Т. 26, № 8. — С. 1243–1249.

[3] *Коган А. Ю.* О дизъюнктивных нормальных формах булевых функций с малым числом нулей // Ж. вычисл. матем. и матем. физ. — 1987. — Т. 27, № 6. — С. 924–931.

[4] *Romanov M. Yu.* Maximal faces of Boolean functions with a small number of zeroes // Pattern Recognition and Image Analysis. — 2010. — V. 20, № 4. — Pp. 474–478.

Эффективная реализация логических алгоритмов в задачах классификации с малым числом эталонов

Максимов Ю. В.
yurymaximov@gmail.com
Москва, МФТИ

Важной частью многих современных алгоритмов распознавания в задачах с бинарной информацией является построение дизъюнктивных нормальных форм (ДНФ) характеристических функций классов. В работе [1] показано, что при достаточно малом числе нулей задача почти всегда может быть сведена к построению ДНФ некоторой канонической функции. Для данной функции в настоящей работе доказана нижняя оценка на число литералов, входящих в произвольную ДНФ, а также предложен алгоритм построения ДНФ, асимптотически оптимальной по числу литералов и по числу конъюнкций.

Существенной частью многих современных алгоритмов распознавания является построение ДНФ характеристических функций классов. Характеристическая функция класса K представляет собой частично определенную булеву функцию, которая равна единице на эталонах из K и нулю на эталонах из других классов. Обычно для ее построения сперва строится ДНФ полностью определенной булевой функции, которая принимает значение ноль лишь на эталонных объектах, не лежащих в классе K, а затем из нее удаляются лишние конъюнкции.

Таким образом, задача построения характеристической функции класса сводится к построению ДНФ булевой функции, заданной матрицей нулей. Очевидно, что построение сокращенной ДНФ путем перемножения скобок совершенной конъюнктивной нормальной формы осуществимо лишь при крайне малом числе переменных. Однако, как впервые указал С. В. Яблонский на примере функций, имеющих ровно два нуля, существуют эффективные методы построения ДНФ, позволяющие избежать прямого перемножения большого числа скобок.

В 80-х годах Ю. И. Журавлевым и А. Ю. Коганом было показано, что сложность проблемы построения сокращенной ДНФ по матрице нулей не является препятствием для решения конкретных практических задач распознавания. Ими был предложен ряд алгоритмов построения достаточно простых ДНФ функций, заданных перечнем своих нулей [1, 2]. Однако практическая реализация этих алгоритмов для построения ДНФ характеристических функций классов в задачах распознавания большой размерности может оказаться достаточно трудоемкой.

В начале 2000-х для функций с малым числом нулей А. Г. Дьяконовым был предложен ряд алгоритмов, позволяющих строить, возможно, более сложную ДНФ, но с меньшими временными затратами, чем алгоритмами Ю. И. Журавлева и А. Ю. Когана [3, 4, 5].

Базовым элементом значительной части предложенных алгоритмов было построение простой ДНФ для так называемой полной функции, от сложности реализации которой существенным образом зависит сложность ряда указанных выше алгоритмов.

В настоящей работе автором предложен достаточно эффективный алгоритм, позволяющий для указанной полной функции строить ДНФ с асимптотически минимальным числом букв и конъюнкций. Кроме того, впервые получены нетривиальные нижние оценки минимального числа букв в произвольной ДНФ рассматриваемой функции.

Построение ДНФ полной функции

Определения и обозначения. При исследовании сложности ДНФ реализации булевой функции обычно оцениваются две основные характеристики сложности: *длина*, равная числу входящих в нее конъюнкций, и *ранг*, равный суммарному числу входящих в ДНФ литералов. На неэквивалентность этих мер сложности указал Ю. И. Журавлёв в работе [6]. ДНФ с оптимальным значением ранга называется минимальной, а ДНФ с минимальным числом конъюнкций — кратчайшей. Обозначим через P_n^k класс булевых функций от n переменных, имеющих ровно k нулей.

Определение 1. *Булеву функцию f назовем приведенной, если в ее матрице нулей не содержится пары столбцов, один из которых равен дополнению другого, а также не содержатся пары равных столбцов.*

Определение 2. *Полной булевой функцией назовем приведенную булеву функцию при* $n = 2^{k-1} - 1$.

В работе [2] была доказана следующая теорема.

Теорема 1. *Построение ДНФ почти всех функций от n переменных, имеющих не более чем $k < \log_2 n - \log_2 \log_2 n + 1$ нулей, сводится к построению ДНФ полной функции с линейной по числу переменных сложностью.*

Оценки сложности ДНФ. К настоящему времени лучшими теоретической оценкой ранга

ДНФ полной функции обладал редукционный алгоритм, предложенный в работах Ю. И. Журавлева и А. Ю. Когана, позволяющий строить ДНФ длины, асимптотически равной $2n(1+o(1))$, и ранга $4n(1+o(1))$.

Алгоритм с оптимальной в асимптотике оценкой длины ДНФ полной функции был предложен А. Г. Дьяконовым в работе [3]. Данный алгоритм позволяет строить ДНФ длины $n(1+o(1))$, однако ранг подобной ДНФ может быть существенно нелинейным. Им же доказано, что длина ДНФ полной функции не может быть меньше n.

В настоящей работе автором доказывается следующая теорема, устанавливающая асимптотически точную оценку сложности ДНФ полной функции

Теорема 2. *Всякая ДНФ, реализующая полную функцию от n переменных, имеет ранг не менее $3n(1+o(1))$. Более того, для всякой полной функции f от n переменных существует ДНФ длины $n(1+o(1))$ и ранга $3n(1+o(1))$ при $n\to\infty$.*

Техника получения оценок. Назовем литерал $x_i^{\sigma_i}$ *собственным* для некоторой ДНФ D функции f, если он входит в единственную конъюнкцию из D. Конъюнкцию, содержащую хотя бы один собственный литерал, назовем *собственной*. Будем считать, что матрица нулей рассматриваемой полной функции преобразована так, что каждый ее столбец имеет число единиц не большее, чем число нулей.

Нормой литерала $x_i^{\sigma_i}$ назовем $\min(k_i, k-k_i)$, где k_i — число единиц в i-ом столбце матрицы нулей. Основой для получения нижней оценки служит

Лемма 3. *Пусть D — произвольная ДНФ, а K — конъюнкция приведенной функции f от n переменных, имеющая k нулей. Пусть, более того, матрица нулей f не содержит столбцов, у которых число единиц больше числа столбцов. Если K содержит хотя бы один собственный литерал нормы больше $\frac{k}{3}$, то справедливо:*

1) $\operatorname{rank} K \geqslant 3$;
2) *K не может содержать более двух собственных литералов нормы больше чем $\frac{k}{3}$;*
3) *если $K=\bar{x}K'$, а литерал $\bar{x}$ собственный, то K' не содержит отрицаний переменных.*

Указанная лемма позволяет выписать оценку для ранга полной функции в виде

$$\operatorname{rank} D \geqslant \max(2(2m-2k_1-k_2, 2k_2)) + k_2 + 3k_1;$$
$$m > (1-\varepsilon)n,\ \forall \varepsilon > 0;$$

где k_1 — число конъюнкций имеющих в точности два собственных литерала с нормой больше $\frac{k}{2}$, k_2 — число конъюнкций, которые имеют ровно одну собственную точку нормы больше $\frac{k}{3}$, m — число столбцов матрицы полной функции размера $k\times n$, содержащих более чем $\frac{k}{3}$ единиц. Указанная оценка дает требуемый результат.

Идею алгоритма получения верхних оценок сформулируем, используя теорию гиперграфов. Обозначим $M[i_1,\dots,i_t]$ подматрицу матрицы M, выделяемую столбцами с номерами $i_1,\dots,i_t$. Каждому литералу функции f сопоставим вершину гиперграфа. Набор различных литералов $x_{i_1}^{\sigma_{i_1}},\dots,x_{i_t}t^{\sigma_{i_t}}$ объединим ребром в том и только в том случае, если всякая строка матрицы $M[\alpha_{i_1},\dots,\alpha_{i_t}]$ отличается от строки $(\sigma_{i_1},\dots,\sigma_{i_t})$ не более чем в одном разряде. Кроме того, соединим ребрами все пары литералов $x_i,\bar{x}_i$. Построенный гиперграф назовем гиперграфом связей функции f.

Представим матрицу M нулей полной функции f в виде объединения двух непересекающихся подматриц T и M', так, что матрица T является тестом (не обязательно тупиковым). Скажем, что ребро e корневое, если все его вершины кроме одной относятся к T. Назовем вершину внешней, если соответствующий ей столбец не входит в матрицу T.

Определение 3. *Подгиперграф $H'=\langle V,E\rangle$ графа связей H назовем T-циклом, если*

1. *H' — связен, то есть из любой вершины H' можно добраться до любой другой, двигаясь только по ребрам гиперграфа.*
2. *Если в гиперграфе H' присутствует внешняя вершина z^σ, то в нем также необходимо присутствует 2-ребро $\{z^\sigma, z^{1-\sigma}\}$.*
3. *H' не имеет циклов по внешним вершинам, то есть в любом реберном цикле T-цикла как гиперграфа существует пересечение последовательных ребер, свободное от внешних переменных.*
4. *От каждой внешней вершины, проходя по ребрам T-цикла, можно добраться до корневого ребра, не проходя через 2-ребро, связанное с рассматриваемой вершиной.*

Основная идея конструкции состоит в том, что указанные T-циклы обладают достаточно простой ДНФ реализацией. Формализует указанную идею следующая лемма

Лемма 4. *Пусть $M_{k\times n} = T_{k\times(n-m)}M'_{k\times m}$, где M — матрица приведенной функции, T — тест. Пусть, более того, существуют непересекающиеся по 2-ребрам T-циклы $H_1,\dots,H_l$ мощности $m_1,\dots,m_l$, $\sum\limits_{i=1}^{l} m_i = m$. Пусть D_T — ДНФ, реали-*

зующая тест T, тогда

$$D = D_T \vee \bigvee_{H_i, i=1,\ldots,l} \bigvee_{\substack{e \in E \\ e=\{z_e^1, z_e^2, \ldots, z_e^{s_e}\} \\ s_e \geqslant 3}} \bar{z}_e^1 \bar{z}_e^2 \cdots \bar{z}_e^{s_e}$$

реализует функцию, заданную матрицей M. Более того, если M является матрицей полной функцией, а T — тестом, содержащим все столбцы, имеющие не более чем $\frac{k}{3}$ единиц, то

1) *указанное разбиение существует, при этом*
2) *каждое ребро T-цикла содержит не более трех вершин;*
3) *общее число T-циклов не превосходит $n(1 + o(1))$.*

Реализовать тестовую подматрицу можно, например, редукционным алгоритмом. Так как тестовая подматрица содержит небольшое число столбцов, ее реализация требует достаточно малого числа конъюнкций и литералов.

Выводы

В работе получена нижняя оценка на ранг полной булевой функции и предложен эффективный алгоритм построения ее асимптотически минимальной и кратчайшей ДНФ. Простая реализация данной функции помогает уменьшить сложность многих современных логических алгоритмов распознавания с малым числом эталонов.

Литература

[1] *Журавлев Ю. И., Коган А. Ю.* Реализация булевых функций с малым числом нулей дизъюнктивными нормальными формами и смежные задачи // Доклады АН СССР. — 2011. — Т. 285, № 4. — С. 795–799.

[2] *Журавлев Ю. И., Коган А. Ю.* Алгоритм построения дизъюнктивной нормальной формы, эквивалентной произведению левых частей булевых уравнений нельсоновского типа // Ж. вычисл. матем. и матем. физ. — 1986. — Т. 26. № 8. — С. 1243–1249.

[3] *Дьяконов А. Г.* Реализация одного класса булевых функций с малым числом нулей тупиковыми дизъюнктивными нормальными формами // Ж. вычисл. матем. и матем. физ. — 2001. — Т. 41. № 5. — С. 828–835.

[4] *Дьяконов А. Г.* Тестовый подход к реализации булевых функций с малым числом нулей дизъюнктивными нормальными формами // Ж. вычисл. матем. и матем. физ. — 2001. — Т. 41. №12. — С. 924–928

[5] *Дьяконов А. Г.* Построение дизъюнктивных нормальных форм в задачах распознавания образов с бинарной информацией // Доклады РАН. — 2002. — Т. 383. №6. — С. 747–749

[6] *Журавлев Ю. И.* О различных понятиях минимальности дизъюнктивных нормальных форм // Сибирский математический журнал. — 1960. — Т. 1, № 4. — С. 608–609.

О корректном понижении значности данных в задачах распознавания*

Дюкова Е. В.[1]***, Сизов А. В.***[2]***, Сотнезов Р. М.***[3]
edjukova@mail.ru, box.sizov@gmail.com, rom.sot@gmail.com
Москва, [1]ВЦ РАН, [2,3]МГУ им. М.В. Ломоносова

Исследуются вопросы применения логических процедур распознавания по прецедентам в случае вещественнозначной информации и целочисленной информации высокой значности. Рассмотрена задача корректного понижения значности данных. Разработаны генетические алгоритмы поиска оптимальной корректной перекодировки исходной информации. Проведено тестирование алгоритмов на реальных данных.

Введение

Один из подходов к решению задачи распознавания по прецедентам сводится к комбинаторному (логическому) анализу исходных признаковых описаний объектов. В данном случае для каждого признака определяется бинарная функция близости между его значениями, позволяющая различать объекты и их подописания. Особенно эффективен комбинаторный подход в случае целочисленной информации низкой значности, например бинарной. Поэтому актуальной является задача корректного понижения значности исходных данных.

Пусть $\{x_1, \ldots, x_n\}$ — система признаков и $\varepsilon_j, \varepsilon_j \geqslant 0$, — точность измерения признака x_j, $j \in \in \{1, \ldots, n\}$. Пусть далее $S_{i_1} = (a_{i_1 1}, \ldots, a_{i_1 n})$ и $S_{i_2} = (a_{i_2 1}, \ldots, a_{i_2 n})$ — обучающие объекты (здесь $a_{i_t j}$ — значение признака x_j для объекта S_{i_t}, $t \in \{1, 2\}$, $j = 1, \ldots, n$). Положим

$$\delta_j(S_{i_1}, S_{i_2}) = \begin{cases} 1, & |a_{i_1 j} - a_{i_2 j}| \leqslant \varepsilon_j; \\ 0, & \text{иначе.} \end{cases}$$

Функция $\delta_j(S_{i_1}, S_{i_2})$ называется функцией близости объектов S_{i_1} и S_{i_2} по признаку x_j. Предполагается, что обучающие объекты из разных классов имеют разные описания, т. е. для любых S_{i_1} и S_{i_2}, принадлежащих разным классам, существует хотя бы один признак x_j такой, что $\delta_j(S_{i_1}, S_{i_2}) = 1$. В случае бинарной информации полагают $\varepsilon_j = 0$, $j \in \{1, \ldots, n\}$.

Один из способов понижения значности данных состоит в преобразовании исходной выборки путем разбиения множества значений каждого признака на интервалы порогами. Значения признаков, попавшие в один интервал, считаются близкими и кодируются одним числом. Однако при произвольном выборе порогов обучающие объекты, принадлежащие разным классам, могут стать неразличимыми. При данном способе преобразования информации важным является понятие корректной перекодировки данных, т. е. такого преобразования обучающей информации, при котором объекты из разных классов остаются различимыми.

Ю. И. Журавлевым предложена методика корректного перекодирования исходных данных. Показано, что задача построения корректной перекодировки может быть сведена к построению специального вида покрытия булевой матрицы, которая строится по обучающей выборке. В [1, 2] предложен подход, позволяющий выбирать наилучшую в смысле качества распознавания корректную перекодировку. Недостаток подхода — его большая вычислительная сложность.

Целью данной работы является развитие методов корректного перекодирования данных и снижение вычислительной сложности этих методов. В работе предложены более эффективные способы оценки качества перекодировок. Для сокращения перебора при поиске оптимальной корректной перекодировки использован генетический подход. Приведены результаты тестирования генетических алгоритмов поиска оптимальной корректной перекодировки на реальных прикладных задачах из репозитория программной системы «Распознавание», описанной в [5].

Основные обозначения

Рассмотрим задачу распознавания по прецедентам с двумя непересекающимися классами K_1 и K_2 [4]. Пусть $T = (a_{ij})_{m \times n}$ — обучающая таблица, $a_{ij} \in \mathbb{R}$, $\mathbb{R}$ — множество действительных чисел. Столбцам таблицы T соответствуют признаки $x_1, x_2, \ldots, x_n$, а каждая строка является набором значений признаков, описывающим один из обучающих объектов. Предполагается, что в таблице T нет столбцов, состоящих из одинаковых чисел.

Пусть S_{i_1} и S_{i_2} — обучающие объекты, принадлежащие разным классам, $j \in \{1, \ldots, n\}$.

Определение 1. *Число $(a_{i_2 j} + a_{i_1 j})/2$ назовем порогом для признака x_j, если в T не существует элемента a_{ij} такого, что $a_{ij} \in (a_{i_1 j}, a_{i_2 j})$.*

Через $D^{(j)}$ обозначим множество всех порогов для признака x_j, $j \in \{1, \ldots, n\}$. Суммой двух элементов $a_{i_1 j}$ и $a_{i_2 j}$ таблицы T по порогу $d \in D^{(j)}$ $j \in \{1, \ldots, n\}$ назовем число $(a_{i_1 j} \oplus a_{i_2 j}|_d)$ равное 1,

Работа выполнена при финансовой поддержке РФФИ, проект №10-01-00770, гранта Президента РФ по поддержке ведущих научных школ НШ – 7950.2010.1.

если $a_{i_1 j}$ и $a_{i_2 j}$ лежат по разные стороны от порога d и равное 0, в противном случае. Пусть $D^{(j)} = \{d_1^{(j)}, \ldots, d_{u_j}^{(j)}\}$.

Через $\prod$ будем обозначать последовательность всех порогов

$$d_1^{(1)}, \ldots, d_{u_1}^{(1)}, d_1^{(2)}, \ldots, d_{u_2}^{(2)}, \ldots, d_1^{(n)}, \ldots, d_{u_n}^{(n)},$$

где $u_j = |D^{(j)}|$, при $j = 1, \ldots, n$. Суммой двух строк таблицы T с номерами i_1 и i_2 по последовательности порогов $\prod$ назовем строку

$$\Big(a_{i_1 1} \oplus a_{i_2 1}|_{d_1^{(1)}}, \ldots, a_{i_1 1} \oplus a_{i_2 1}|_{d_{u_1}^{(1)}},$$
$$a_{i_1 2} \oplus a_{i_2 2}|_{d_1^{(2)}}, \ldots, a_{i_1 2} \oplus a_{i_2 2}|_{d_{u_2}^{(2)}},$$
$$\ldots$$
$$a_{i_1 n} \oplus a_{i_2 n}|_{d_1^{(n)}}, \ldots, a_{i_1 n} \oplus a_{i_2 n}|_{d_{u_n}^{(n)}}\Big).$$

Пусть m_1 и m_2 — число обучающих объектов из классов K_1 и K_2 соответственно. Построим булеву матрицу L. Матрица L имеет размеры $h \times N$, где $h = m_1 m_2$, $N = |D^{(1)}| + \cdots + |D^{(n)}|$. Каждая ее строка образуется в результате попарного сложения строк таблицы T, описывающих объекты из разных классов, по последовательности порогов $\prod$. Порядок выбора пар может быть задан произвольным образом. Множеству порогов $D^{(j)}$, $j \in \{1, \ldots, n\}$ по построению соответствует группа из u_j столбцов матрицы L, обозначаемая через G_j.

Определение 2. *Набор столбцов H матрицы L назовем кодирующим покрытием, если выполнены следующие два условия: 1) H является покрытием L, т. е. для любой строки матрицы L в наборе H можно указать хотя бы один столбец, имеющий 1 на пересечении с этой строкой; 2) $H \cap G_j \neq \varnothing$ при $j = 1, \ldots, n$.*

Определение 3. *Кодирующее покрытие назовем неприводимым, если никакое его собственное подмножество кодирующим покрытием не является.*

Определение 4. *Число* $\max_{j \in \{1,\ldots,n\}} |H \cap G_j| + 1$ *назовем значностью кодирующего покрытия H.*

Кодирующее покрытие H задает очевидным образом преобразование таблицы T в таблицу T^H на основе замены элементов T числами из $\{0, \ldots, k-1\}$, где k — значность H. Действительно, пусть a_{pj} — произвольный элемент таблицы T, и пусть $\{d_1, \ldots, d_v\}$ — пороги, соответствующие столбцам из $H \cap G_j$, причем $d_1 < \cdots < d_v$ и $v < k$. Возможны три случая:

1) $a_{pj} \leqslant d_1$;
2) $d_t < a_{pj} \leqslant d_{t+1}$, $t \in \{1, \ldots, v-1\}$;
3) $d_v < a_{pj}$.

В случае 1) элемент a_{pj} кодируется числом 0, в случае 2) — числом t , в случае 3) — числом v. Легко видеть, что в таблице T^H описания объектов из разных классов различны. В дальнейшем T^H будем называть корректной перекодировкой таблицы T.

Таким образом, каждому кодирующему покрытию матрицы L соответствует корректная перекодировка. Мощность множества кодирующих покрытий матрицы L экспоненциально растет с ростом размеров задачи. Поэтому сложной в вычислительном плане является задача выбора наилучшей по качеству распознавания корректной перекодировки. Данная задача рассмотрена в [1]. В указанной работе построен алгоритм КОД1 поиска оптимальной корректной перекодировки.

Алгоритм поиска оптимальной корректной перекодировки КОД1

Введем понятие типичного элемента b_{ij} в таблице $T^H = (b_{ij})_{m \times n}$. Пусть $b_{ij} = a$ и q_t, $t \in \{1, 2\}$, — число строк в T^H, имеющих a в пересечении со столбцом с номером j и описывающих объекты из класса K_t. Элемент b_{ij} назовем типичным в T^H для класса K_1, если

$$\frac{q_1}{|K_1|} - \frac{q_2}{|K_2|} > \mu_j,$$

где $\mu_j, \mu_j \geqslant 0$, — заданный порог типичности значений признака x_j. Аналогично вводится понятие типичного элемента для класса K_2.

Положим

$$I_{ij}(a) = \begin{cases} 1, & \text{если } b_{ij} = a; \\ 0, & \text{иначе.} \end{cases}$$

Пусть $D \subseteq D^{(j)}$, Q_j — множество всех типичных элементов j-го столбца T^H.

Для каждого признака x_j зададим целое число k_j, $0 < k_j \leqslant u_j$, и для каждого $p, p \in \{1, \ldots, k_j\}$, поставим задачу максимизации функционала

$$F(D) = \frac{1}{m} \sum_{i=1}^{m} \sum_{a \in Q_j} I_{ij}(a), \ |D| = p, \ D \subseteq D^{(j)}.$$

Таким образом, для каждого признака x_j и для каждого $p, p \in \{1, \ldots, k_j\}$, выбираем подмножество порогов D_{pj}^* такое, что $F(D_{pj}^*) = \max F(D)$, $|D| = p$, $D \subseteq D^{(j)}$. Множество перекодировок признака x_j упорядочиваем по убыванию значений $F(D_{pj}^*)$. Будем считать, что перекодировка $H_1 = \{D_{p_1 1}^*, \ldots, D_{p_n n}^*\}$ следует за перекодировкой $H_2 = \{D_{q_1 1}^*, D_{q_2 2}^*, \ldots, D_{q_n n}^*\}$, если

$$\sum_{j=1}^{n} F(D_{q_j j}^*) \geqslant \sum_{j=1}^{n} F(D_{p_j j}^*).$$

В заданном порядке последовательно просматриваем всевозможные перекодировки таблицы T. Первая по порядку корректная перекодировка считается оптимальной.

Сложность алгоритма КОД1 быстро растет с ростом размеров задачи. В следующем разделе построены алгоритмы, в которых используются другие критерии оптимальности кодирующего покрытия. Для сокращения вычислительной сложности рассмотрен генетический подход.

Однокритериальные генетические алгоритмы поиска оптимальной корректной перекодировки КОД2, КОД3, КОД4

Пусть c_j,$j \in \{1,\dots,n\}$, — число единиц в j-ом столбце матрицы $L_{h\times N}$, $R_1(H)$ — множество номеров столбцов L, входящих в кодирующее покрытие H, $R_2(H)$ — множество номеров столбцов матрицы L, не входящих в кодирующее покрытие H. Тогда

$$f_1(H) = \sum_{j\in R_2(H)} c_j$$

$$f_2(H) = \frac{1}{|H|}\sum_{j\in R_1(H)} \frac{1}{c_j}.$$

Построены алгоритмы КОД2,КОД3 и КОД4. В генетическом алгоритме КОД4 особям соответствуют кодирующие покрытия, а в алгоритмах КОД2 и КОД3, особям соответствуют неприводимые кодирующие покрытия. Роль функции приспособленности играет один из двух функционалов: $f_1(H)$ для КОД2 и $f_2(H)$ для КОД3 и КОД4. Функция приспособленности особи является критерием качества кодирующего покрытия в перечисленных алгоритмах.

Для реализации генетических алгоритмов взята схема генетического алгоритма из [3], которая адаптирована к условиям задачи. Основной особенностью задачи поиска кодирующего покрытия является условие включения в покрытие хотя бы одного столбца из набора столбцов $G_j, j = 1,\dots,n$, матрицы L. Для выполнения этого условия внесены соответствующие изменения в процедуру восстановления допустимости решения.

В работе использован оригинальный оператор мутации с переменным числом мутирующих генов, что позволяет минимизировать влияние данного оператора на особь на ранних этапах работы и усиливать его влияние с увеличением числа итераций.

Двухкритериальный генетический алгоритм поиска оптимальной корректной перекодировки КОД5

Разработан двухкритериальный генетический алгоритм, основанный на схеме однокритериального генетического подхода. Отличия заключаются в процедуре вычисления функции приспособленности и процедуре добавления особи в популяцию.

На каждом этапе для каждой особи вычисляются значения функционалов $f_1(H)$ и $f_3(H)$, где $f_3(H)$ — значность перекодировки. На основе полученных векторов значений $v\big(f_1(H), f_3(H)\big)$ производится вычисление рангов. Пусть V — множество векторов v, вычисленных для каждой особи популяции. Рангом вектора $\mathrm{rg}(v)$, $v \in V$, называется величина, равная числу векторов, строго доминирующих данный вектор. В случае если таких векторов нет, ранг вектора v принимается равным 1. В качестве функции приспособленности используется $f(v) = \max\limits_{w\in V}\big(\mathrm{rg}(w)-\mathrm{rg}(v)\big)+1$. В процедуре обновления популяции заменяется особь, имеющая самый низкий ранг. Алгоритм рассматривает только неприводимые покрытия.

Результаты экспериментов

Построенные в настоящей работе генетические алгоритмы были протестированы на реальных данных и сравнивались с алгоритмами КОД1 и градиентным алгоритмом. В качестве распознающего алгоритма использовалась процедура голосования по представительным наборам с ограничением по длине набора. Сравнение проводилось на реальных задачах из репозитория программной системы «Распознавание», описанной в [5].

Результаты счета представлены в таблицах 1–4. В этих таблицах введены следующие обозначения:

N — *номер задачи;*

m_1, m_2 — *число объектов в классах;*

n — *число признаков;*

A_0 — *алгоритм голосования по представительным наборам, построенным по исходной обучающей выборке;*

A_0^* — *распознающий алгоритм, примененный к данным, перекодированным градиентный алгоритм поиска неприводимых кодирующих покрытий;*

A_1 — *распознающий алгоритм, примененный к перекодированным алгоритмом КОД1 данным;*

A_2 — *распознающий алгоритм, примененный к перекодированным алгоритмом КОД2 данным;*

A_3 — *распознающий алгоритм, примененный к перекодированным алгоритмом КОД3 данным;*

A_4 — *распознающий алгоритм, примененный к перекодированным алгоритмом КОД4 данным;*

A_5 — *распознающий алгоритм, примененный к перекодированным алгоритмом КОД5 данным;*

Z — *значность исходных данных.*

В таблице 1 приведены характеристики задач: число обучающих объектов в классах и число признаков. В случае вещественнозначной информации значность определяется по числу различных значений признаков в обучающей выборке.

Таблица 2 содержит значность полученных алгоритмами перекодировок. Прочерк означает, что алгоритм был исключен из эксперимента из-за слишком большого времени выполнения (более одного часа). Нетрудно заметить, что наименьшая

Таблица 4.

N	m_1, m_2, n	A_0^*	A_1	A_3
1	90,42, 9	6	59	10
2	60,71,13	8	244	17
3	30,30,33	13	415	29
4	50,50,19	9	194	17
5	70,80,24	24	-	67

Таблица 1.

N	m_1, m_2, n	Z
1	90,42, 9	112
2	60,71,13	97
3	30,30,33	51
4	50,50,19	78
5	70,80,24	133

Таблица 2.

N	КОД1	КОД2	КОД3	КОД4	КОД5
1	39	31	27	30	24
2	26	24	12	22	11
3	24	20	11	18	10
4	29	23	17	23	15
5	-	21	15	20	13

Таблица 3.

N	A_0	A_0^*	A_1	A_2	A_3	A_4	A_5
1	54	66	67	67	**76**	72	73
2	63	83	88	92	93	92	**94**
3	58	70	72	72	**82**	80	81
4	53	68	71	73	**79**	72	75
5	57	66	-	67	76	74	**78**

значность достигнута алгоритмом КОД5. Данный результат обусловлен использованием функционала $f_3(H)$ в этом алгоритме.

В таблице 3 приведено качество распознавания в процентах, полученное алгоритмами на скользящем контроле.

Из приведенных в таблице 3 результатов следует, что все предложенные способы перекодирования данных улучшают качество распознавания. Из тех же результатов следует, что решения алгоритмов A_2 и A_4 не хуже по качеству распознавания, чем решения алгоритмов A_0, A_0^* и A_1, а алгоритмы A_3 и A_5 превосходят другие алгоритмы по качеству распознавания.

Нетрудно заметить, что условие неприводимости кодирующего покрытия оказывает существенное влияние как на значность перекодировки, так и на качество распознавания. Среди рассмотренных функционалов оценки качества перекодировки лучшими очевидно являются $f_2(H)$ и пара функционалов $f_1(H)$ и $f_3(H)$, так как результаты алгоритмов A_3 и A_5 являются лучшими на всех рассмотренных задачах.

Алгоритм A_3 превзошел алгоритм A_5 на задачах 1,3,4, поэтому можно сделать вывод, что меньшая значность перекодировки (функционал $f_3(H)$) не всегда означает, что перекодировка лучше по качеству распознавания.

В таблице 4 приведено время счета алгоритмов в секундах. Прочерк означает, что алгоритм был исключен из эксперимента из-за слишком большого времени выполнения (более одного часа). В таблице не представлены алгоритмы A_2, A_4 и A_5 по той причине, что вычислительная сложность этих алгоритмов аналогична вычислительной сложности алгоритма A_3, а следовательно время счета практически не отличается. Из результатов следует, что распознающие алгоритмы, полученные на основе алгоритмов КОД2, КОД3, КОД4 и КОД5 по сравнению с распознающим алгоритмом, основанным на алгоритме КОД1, выигрывают не только по качеству решения, но и по скорости выполнения. Кроме того, качество решения полученных алгоритмов существенно лучше, чем качество решения алгоритма голосования по представительным наборам, примененного к неперекодированным данным.

Литература

[1] *Дюкова Е. В., Журавлев Ю. И., Песков Н. В., Сахаров А. А.* Обработка вещественнозначной информации логическими процедурами распознавания // Искусственный интеллект. НАН Украины. — 2004. — № 2. — С. 80–85.

[2] *Djukova E., Inyakin A., Peskov N., Sakharov A.* Combinatorial (Logical) Data Analysis in Pattern Recognition Problems // Pattern Recognition and Image Analysis. — 2005. — V. 15, N. 1. — Pp. 46–48

[3] *Sotnezov R. M.* Genetic Algorithms for Problems of Logical Data Analysis in Discrete Optimization and Image Recognition // Pattern Recognition and Image Analysis. — 2009. — V. 19, N. 3. — Pp. 469—477.

[4] *Журавлев Ю. И.* Об алгебраическом подходе к решению задач распознавания и классификации. // Проблемы кибернетики. М.: Наука. — 1978. — Вып. 33. — С. 5–68

[5] *Журавлев Ю. И., Рязанов В. В., Сенько О. В.* «Распознавание». Математические методы. Программная Система. Практические применения // М.: Фазис, 2006. — С. 176.

[6] *Дюкова Е. В., Карнеева И. Л.* Модели распознающих алгоритмов, основанные на различных способах перекодировки исходной информации // Математические методы в распознавании образов и дискретной оптимизации. М.: ВЦ АН СССР, 1990. — С. 43–56

Полные решающие деревья в задачах классификации по прецедентам*

Генрихов И. Е., Дюкова Е. В.
ingvar1485@rambler.ru, edjukova@mail.ru
Москва, Учреждение Российской академии наук Вычислительный центр им. А. А. Дородницина Российской академии наук

В докладе представлены результаты, полученные авторами, разработки алгоритмов классификации на основе полных решающих деревьев. Рассмотрены вопросы практического применения полных решающих деревьев с энтропийный критерием ветвления для задач распознавания по прецедентам в случае вещественнозначной информации. Построенны модели распознающих процедур, нацеленые на решение задач с неполными данными (с пропусками в признаковых описаниях объектов) и с неравномерным распределением обучающих объектов по классам. Дан обзор основных результатов, полученных авторами ранее в данной области.

Рассматривается задача распознавания по прецедентам с системой признаков $\{x_1, \ldots, x_n\}$, с непересекающимися классами $K_1, \ldots, K_l$ и множеством обучающих объектов $T = \{S_1, \ldots, S_m\}$, где $S_r = (a_{r1}, \ldots, a_{rn})$, $a_{rj} \in \{\mathbb{R}, «*»\}$, $r \in \{1, \ldots, m\}$, $j = 1, \ldots, n$. Если $a_{rj} =$«*», то значение признака x_j для объекта S_r не определено. Пусть далее $S = (b_1, \ldots, b_n)$ — распознаваемый объект, и $b_j \in \{\mathbb{R}, «*»\}$, $j = 1, \ldots, n$.

Одним из известных инструментов для решения рассматриваемой задачи являются деревья решений. Процедура построения классического дерева решений представляет собой итерационный процесс. Как правило, для построения очередной вершины дерева выбирается признак, наилучшим образом удовлетворяющий некоторому критерию ветвления. По значениям этого признака и осуществляется ветвление, далее указанная процедура повторяется для каждой из ветвей. Однако, если при построении дерева несколько признаков удовлетворяют критерию ветвления в равной или почти равной мере, то выбор одного из них происходит случайным образом. При этом в зависимости от выбранного признака построенные деревья могут существенно отличаться как по составу используемых признаков, так и по своим распознающим качествам. В [6] рассмотрен следующий подход к потроению решающих деревьев для задач распознавания. При возникновении ситуации, когда два или более признака удовлетворяют критерию ветвления в равной мере или почти равное мере, предложено проводить ветвление по каждому из этих признаков независимо. Полученная в результате конструкция названа полным решающим деревом (ПРД). Таким образом, в отличие от классического дерева решений, в ПРД на каждой итерации строится так называемая полная вершина, которой соответствует набор признаков $\{x_{j_1}, \ldots, x_{j_q}\}$. Затем для каждого признака x_{j_i}, $i = 1, \ldots, q$, строится внутренняя вершина, из которой осуществляется ветвление. Для случая целочисленной информации в [1, 2, 6] данный подход успешно продемонстрирован на примерах усовершенствования алгоритма построения допустимых разбиений [5] и алгоритма C4.5 [9].

Для случая вещественнозначной информации и наличия пропусков в признаковых описаниях объектов в [3, 4] построены распознающие процедуры на основе ПРД, в которых был модифицирован энтропийный критерий ветвления, используемый в алгоритме C4.5. Каждая висячая вершина построенного ПРД содержит вектор оценок за классы, которые учитываются в процедуре голосования. Алгоритмы из [3, 4] показали лучшие результаты на большинстве из рассматриваемых прикладных задачах по сравнению с алгоритмом C5.0 [10] (улучшенной коммерческой версии алгоритма C4.5) и методом построения бинарного решающего дерева представленного в работе [7].

Более подробный обзор результатов, полученных в [1–4], дан ниже. Кроме того в данной работе рассмотрен случай вещественнозначной информации, наличия пропусков в признаковых описании объектов и неравномерного распределения объектов по классам (в этом случае можно указать пару классов таких, что число обучающих объектов в одном из них существенно больше числа обучающих объектов в другом). Такая ситуация нередко встречается в прикладных задачах и ей недостаточно уделено внимания при построении классификаторов, основанных на решающих деревьях. Для данного случая предложен новый алгоритм AGI.Bias.

Основные понятия и полученные ранее результаты

Опишем структуру ПРД. Пусть $\hat{T}$ — подмножество обучающих объектов и $X(\hat{T})$ — подмножество признаков, рассматриваемые на текущем шаге построения ПРД. На первом шаге $\hat{T} = T$, $X(T) = \{x_1, \ldots, x_n\}$.

Работа выполнена при частичной финансовой поддержке РФФИ, проект № 10-01-00770, гранта Президента РФ по поддержке ведущих научных школ НШ № 7950.2010.1

На каждом шаге формируется набор из различных признаков X, $X \subseteq X(\hat{T})$, порождающий полную вершину. Из полной вершины, обозначаемой через X, выходит ровно q дуг. Каждая из этих дуг входит в простую внутреннюю вершину x, $x \in X$. При ветвлении из вершины x, происходит удаление признака x из $X(\hat{T})$ и удаление некоторых объектов из $\hat{T}$.

Рассмотрим два способа ветвления из вершины x.

Способ 1. Информация бинарная или целочисленная невысокой значности.

Пусть $\{0,\ldots,k-1\}$, $K \geqslant 2$, — множество различных значений признака x, встречающихся в описании объектов из $\hat{T}$. В этом случае при построении дерева решений из вершины x выходят k помеченных дуг. Пусть σ — метка одной из этих дуг, $\sigma \in \{0,\ldots,k-1\}$. При спуске из вершины x по дуге с меткой σ удаляются те объекты из $\hat{T}$, для которых значение признака x равно σ. Пусть v — вершина, порожденная ветвью дерева с простыми внутренними вершинами $x_{j_1},\ldots,x_{j_r}$, и пусть дуга, выходящая из вершины x_{j_i}, имеет метку σ_i, $i \in \{1,\ldots,r\}$. Набор $N_v = (\alpha_1,\ldots,\alpha_n)$, где $\alpha_{j_i} = \sigma_i$ при $i = 1,\ldots,r$, и $\alpha_j =$«*» при $j \notin \{j_1,\ldots,j_r\}$, назовем порождающим для вершины v. Если вершина v не является висячей, то поставим ей в соответствие набор N_v. Если вершина v — висячая, то поставим ей в соответствие пару $\big(N_v, \{\omega_v^1,\ldots,\omega_v^l\}\big)$, где ω_v^i, $i \in \{1,\ldots,l\}$, — оценка вершины v за класс K_i, специальным образом вычисленная по обучающей выборке. Описанием объекта $\bar{S} = (b_1,\ldots,b_n)$ в вершине v будем называть вектор $\bar{S}(v) = (\beta_1,\ldots,\beta_n)$, в котором $\beta_{j_i} = b_{j_i}$ при $i = 1,\ldots,r$, и $\beta_j =$«*» при $j \notin \{j_1,\ldots,j_r\}$.

Способ 2. Информация вещественнозначная или целочисленная высокой значности.

В этом случае для ветвления из вершины x осуществляется бинарная перекодировка текущих значений признака x. Рассматриваемая вершина помечается парой $\big(x, d(x)\big)$, где $d(x)$ — «оптимальный» порог перекодировки. Спуск из вершины $\big(x, d(x)\big)$ происходит по двум ветвям, при этом левая ветвь помечается 0, а правая — 1. При спуске из вершины $\big(x, d(x)\big)$ по левой ветви удаляются те объекты из $\hat{T}$, для которых значение признака x больше $d(x)$. При спуске из вершины $\big(x, d(x)\big)$ по правой ветви удаляются те объекты из $\hat{T}$, для которых значение признака x не больше $d(x)$. Пусть v — вершина, порожденная ветвью дерева с простыми внутренними вершинами $x_{j_1},\ldots,x_{j_r}$, и пусть дуга, выходящая из вершины x_{j_i}, имеет метку σ_i, $i \in \{1,\ldots,r\}$. Набор $N_v = (\alpha_1,\ldots,\alpha_n)$, где $\alpha_{j_i} = \sigma_i$ при $i = 1,\ldots,r$, и $\alpha_j =$«*» при $j \notin \{j_1,\ldots,j_r\}$, назовем порождающим для вершины v. Если v не является висячей вершиной, то ей ставится в соответствие набор N_v. Если v — висячая вершина, то поставим ей в соответствие пару $\big(N_v, \{\omega_v^1,\ldots,\omega_v^l\}\big)$, в которой N_v — порождающий набор для вершины v, $\{\omega_v^1,\ldots,\omega_v^l\}$ — вектор оценок за классы. Описанием объекта $\bar{S} = (b_1,\ldots,b_n)$ в вершине v будем называть вектор $\bar{S}(v) = (\beta_1,\ldots,\beta_n)$, в котором $\beta_{j_i} = 1$, если $b_{j_i} > d(x_{j_i})$, иначе $\beta_{j_i} = 0$ при $i = 1,\ldots,r$, и $\beta_j =$«*» при $j \notin \{j_1,\ldots,j_r\}$.

Висячая вершина v называется голосующей для S, если $S(v) = N_v$.

Заметим, что возможен случай, когда $S(v_1) = N_{v_1}$ и $S(v_2) = N_{v_2}$ при $v_1 \neq v_2$.

Пусть $Q(S)$ — множество всех голосующих висячих вершин ПРД для распознаваемого объекта S. Для каждого $i \in \{1,\ldots,l\}$ вычисляется оценка принадлежности объекта S классу K_i, имеющая вид $\Gamma(S,K_i) = \sum_{v \in Q(S)} \omega_v^i$, $i = 1,\ldots,l$.

Объект S зачисляется в класс K_i, если

$\Gamma(S,K_i) = \max\limits_{j=1,\ldots,l} \Gamma(S,K_j)$, $i = 1,\ldots,l$;

$\Gamma(S,K_i) \neq \Gamma(S,K_j)$ при $i \neq j$, $j = 1,\ldots,l$.

Если классов с максимальной оценкой несколько, то среди них выбирается только один, а именно тот, который имеет наибольшее число объектов в обучающей выборке, иначе происходит отказ алгоритма от классификации объекта S.

Пусть $S_r(v)$, $r = 1,\ldots,m$, — описание обучающего объекта S_r в висячей вершине v. Висячая вершина v называется корректной, если не существует пары обучающих объектов S_{i_1} и S_{i_2}, принадлежащих разным классам, таких что $S_{i_1}(v) = N_v$ и $S_{i_2}(v) = N_v$, i_1, $i_2 \in \{1,\ldots,m\}$. Решающее дерево, у которого каждая висячая вершина корректная называется корректным.

Алгоритмы из [1, 2, 6] строят корректные ПРД. При этом, если висячая вершина v — голосующая для S, то только одна координата в векторе оценок $\{\omega_v^1,\ldots,\omega_v^l\}$ равна 1, остальные координаты равны 0. Координата вектора ω_v^i равна 1, если существует хотя бы один обучающий объект $S_r \in K_i$ такой, что $S_r(v) = N_v$, $r = 1,\ldots,m$, $i \in \{1,\ldots,r\}$.

В случае вещественнозначной информации, важной является задача нахождения такого порога $d(x)$, который наилучшим образом разделяет объекты из $\hat{T}$ по признаку x, принадлежащие разным классам. Рассмотрим правила выбора порога $d(x)$ в алгоритме C5.0 и в алгоритмах из [3, 4].

В алгоритме C5.0 выбор признака x и порога $d(x)$ для ветвления на текущем шаге происходит следующим образом. Под порогом для признака x, $x \in X(\hat{T})$, понимается полусумма двух соседних значений из упорядоченного множества текущих значений признака x. Для каждого найденного порога определяется информативность признака x по энтропийному критерию, и в качестве оптимального порога $d(x)$ берется тот порог, для которого

эта информативность максимальна. Данная процедура повторяется для всех признаков из $X(\hat{T})$. После этого выбирается только один наиболее информативный признак x_{opt} с оптимальным порогом $d(x_{\text{opt}})$. Далее происходит ветвление из внутренней вершины с меткой $(x_{\text{opt}}, d(x_{\text{opt}}))$.

В алгоритмах из [3, 4] полусумма двух соседних значений из упорядоченного множества текущих значений признака x, $x \in X(\hat{T})$, является порогом только тогда, когда в $\hat{T}$ эти значения принадлежат объектам из разных классов. Для каждого найденного порога определяется информативность признака x по модифицированному энтропийному критерию (см. [3, 4]), и в качестве оптимального порога $d(x)$ берется тот порог, для которого эта информативность максимальна. Данная процедура повторяется для каждого признака из $X(\hat{T})$, после чего вызывается процедура выбора набора признаков $X = \{x_{j_1}, \ldots, x_{j_q}\}$, $X \subseteq X(\hat{T})$, для ветвления (см. [3, 4]). Далее осуществляется ветвление из полной вершины с меткой $(\{x_{j_1}, \ldots, x_{j_q}\}, \{d(x_{j_1}), \ldots, d(x_{j_q})\})$.

Рассмотрим теперь ситуацию, когда информация вещественнозначная и в признаковых описаниях объектов имеются пропуски.

Пусть на текущем шаге построения дерева построена простая вершина с меткой (x_t, d), где $d = d(x_t)$, и пусть значение признака x_t для обучающего объекта S_r неопределенно, т. е. равно «*». При спуске из вершины (x_t, d) по левой (правой) ветви образуется подмножество T_d^1 (T_d^2) множества $\hat{T}$ путем удаления объектов из $\hat{T}$, для которых значение признака x_t больше (не больше) d.

В алгоритме C5.0 предполагается, что пропущенные значения признака x_t вероятностно распределены пропорционально частоте появления встречающихся значений. До начала построения дерева решений все обучающие объекты имеют вес (вероятность) равный 1. Пусть w_0 — сумма весов объектов из $\hat{T}$, для которых значение признака x не больше d, w_1 — сумма весов объектов из $\hat{T}$, для которых значение признака x_t больше d, $w(S_r)$ — текущий вес объекта S_r. При ветвлении из вершины (x_t, d) объект S_r попадает в T_d^1 с весом $w_0 w(S_r)/(w_0 + w_1)$ и в T_d^2 с весом $w_1 w(S_r)/(w_0 + w_1)$. В висячей вершине v неявно, с учетом весов обучающих объектов для вершины v, вычисляется вектор оценок за классы.

Пусть значение признака x_t для распознаваемого объекта S неопределенно, т. е. равно «*» и пусть N_{x_t} — порождающий набор для вершины (x_t, d) вида $(\alpha_1, \ldots, \alpha_{t-1}, «*», \alpha_{t+1}, \ldots, \alpha_n)$, где $\alpha_i \in \{0, 1, «*»\}$, $i \in \{1, \ldots, n\} \backslash t$. При спуске из (x_t, d) по левой ветви образуется набор $N_{x_t}^1 = (\alpha_1, \ldots, \alpha_{t-1}, 0, \alpha_{t+1}, \ldots, \alpha_n)$. При спуске из (x_t, d) по правой ветви образуется набор $N_{x_t}^2 = (\alpha_1, \ldots, \alpha_{t-1}, 1, \alpha_{t+1}, \ldots, \alpha_n)$. До начала спуска по дереву объект S имеет вес равный 1. Пусть $\hat{w_0}$ — сумма весов объектов из T_d^1, $\hat{w_1}$ — сумма весов объектов из T_d^2, $w(S)$ — текущий вес объекта S. В этом случае, в алгоритме C5.0, $S(x_t) = N_{x_t}^1$ и $S(x_t) = N_{x_t}^2$ с определенными весами (вероятностями). Вероятность того, что $S(x_t) = N_{x_t}^1$ вычисляется по формуле $\hat{w_0} w(S)/(\hat{w_0} + \hat{w_1})$ и вероятность того, что $S(x_t) = N_{x_t}^2$ равна $\hat{w_1} w(S)/(\hat{w_0} + \hat{w_1})$. В голосующей висячей вершине v неявно, с учетом веса распознаваемого объекта S для вершины v, вычисляется вектор оценок за классы.

Следует заметить, что приписывание определенного веса (вероятности) обучающим объектам из $\hat{T}$, в описании которых значение признака x_t пропущено, вносит шум в обучающие данные. Если бы на месте пропущенного значения признака x_t находилось бы какое-либо реальное число, полученное в процессе сбора данных, то решающее дерево возможно имело бы другую структуру.

В алгоритмах из [3, 4] при ветвлении из вершины (x_t, d) пропущенные значения признака x_t для объектов из $\hat{T}$ не принимаются во внимание. Объект S_r не попадает ни в T_d^1, ни в T_d^2.

В случае, если на текущей шаге построения ПРД значение признака x_t неопределенно для распознаваемого объекта S, то в алгоритмах из [3, 4] признак x_t исключается из $X(\hat{T})$. Таким образом, при построении ПРД для классификации объекта S, учитываются только те признаки, для которых значения в S определены.

Стоит отметить, что при построении ПРД происходит лавинообразный рост вершин и ветвей, в связи с этим увеличивается время классификации объекта S. Для того, чтобы сократить это время в алгоритмах из [1–4] строятся только голосующие ветви ПРД. Эта методика позволяет сократить время классификации более, чем в 2 раза.

Описание популярных методик, используемых при решении задачи классификации с пропусками, представлено в работе [8]. Большинство из них основано на замене пропущенного значения одним из допустимых. Это значение может быть вычислено различными способами: как среднее по существующим значениям признака; случайно выбрано из существующих значений; получено с помощью методов k-ближайших соседей, регрессионного или кластерного анализа. Также существует методика основанная на удалении объектов с пропусками из обучающей выборки. При таком подходе теряется полезная информация.

Алгоритм AGI.Bias

В данной работе на базе алгоритма AGI.La.sum из [3, 4] построен алгоритм AGI.Bias предназначенный для случая вещественнозначной информации, наличия пропусков в признаковых описании объ-

Таблица 1. Эффективность алгоритмов.

№ задачи $(\lvert K_1\rvert, \ldots, \lvert K_l\rvert, n)$	C5.0	AGI. Bias	AGI. La. sum	ГМ	НС	ГТТ
№ 1 (48, 12, 69)	50,0	59,38	54,17	56,25	59,4	**60**,**42**
№ 2 (23, 173, 17)	46,53	**55**,**13**	50,01	49,71	49,15	49,75
№ 3 (32, 123, 19)	65,03	**79**,**37**	72,09	– – –	75,56	71,28
№ 4 (38, 107, 35)	**82**,**17**	81,70	80,94	78,39	77,75	66,78
№ 5 (38, 35, 35)	64,32	82,67	**83**,**76**	77,18	82,33	62,26
№ 6 (30, 102, 24)	60,29	53,63	59,80	51,67	61,37	**69**,**31**
№ 7 (60, 15, 39, 5)	72,65	70,17	64,19	71,03	**79**,**40**	76,67
№ 8 (47, 30, 7)	85,75	**89**,**68**	**89**,**68**	85,28	83,01	84,54
№ 9 (40, 40, 18)	70,0	70,00	**71**,**25**	53,75	62,50	70,0
№ 10 (89, 42, 9)	70,87	77,22	77,02	– – –	77,15	**77**,**45**
№ 11 (120, 150, 13)	76,33	82,58	79,92	81,50	81,17	**82**,**75**
№ 12 (39, 22, 18)	74,13	78,67	**81**,**24**	75,70	**81**,**24**	70,86
Среднее значение	68,17	**73**,**35**	72,01	– – –	72,5	70,17

ектов и неравномерного распределения объектов по классам. Отличие алгоритма AGI.Bias от алгоритма AGI.La.sum состоит в том, что в нем по другому осуществляется вычисление вектора оценок в висячих вершинах ПРД. Пусть v — висячая вершина, которой приписана пара $\bigl(N_v, \{\omega_v^1, \ldots, \omega_v^l\}\bigr)$. Обозначим m_v^i — число обучающих объектов класса K_i, описания которых равны N_v, m^i — число всех обучающих объектов класса K_i и пусть $m_v = m_v^1 + \cdots + m_v^l$. Тогда $\omega_v^i = (m_v^i + 1)/(m^i + l)$, $i = 1, \ldots, l$, в алгоритме AGI.Bias и $\omega_v^i = (m_v^i + 1)/(m_v + l)$, $i = 1, \ldots, l$, в алгоритме AGI.La.sum. Таким образом, в алгоритме AGI.Bias оценка за класс K_i для висячей вершины v нормируется с учетом числа всех объектов класса K_i в обучающей выборке.

Численный эксперимент

Тестирование алгоритмов осуществлялось на наборе из 12 реальных задач, собранных в отделе Математических проблем распознавания и методов комбинаторного анализа ВЦ РАН. Качество алгоритмов оценивалось методом скользящего контроля («leave-one-out»). Вычислялась величина $\Theta = \sum_{i=1}^{l} q_i/l$, где q_i — процент правильно классифицируемых объектов класса K_i, l — число классов. Для сравнения использовались алгоритмы: голосования по тупиковым тестам (ГТТ), генетический метод (ГМ), нейронная сеть (НС). Данные алгоритмы входят в систему интеллектуального анализа данных, распознавания и прогнозирования (версия 2.0, ООО «Центр технологий анализа и прогнозирования «Решения»») [7]. Для сравнения так же использовались алгоритм C5.0 [10], который запускался без применения методов «бустинга» и отсечения, и алгоритм AGI.La.sum [3, 4]. В таблице 1 представлены результаты счета.

Лучшими алгоритмами по показателю Θ среди представленных алгоритмов по всем задачам стали алгоритм голосования по тупиковым тестам, алгоритм AGI.La.sum и AGI.Bias .

По сравнению с алгоритмами, в которых строится дерево решений (алгоритм C5.0 и AGI.La.sum), алгоритм AGI.Bias по показателю Θ оказался лучше на большинстве задач. Так же по среднему значению показателя Θ алгоритм AGI.Bias показал неплохие результаты по сравнению со всеми тестируемыми алгоритмами.

Исследование и построение распознающих процедур на основе ПРД показало возможность использования данного подхода, для случая вещественнозначной информации, наличия пропусков в признаковых описании объектов и неравномерного распределения объектов по классам.

Литература

[1] *Генрихов И. Е.* Построение полного решающего дерева на базе алгоритма C4.5 // Сообщение по прикладной математике. Москва: ВЦ РАН — 2009.— 24 с.

[2] *Генрихов И. Е., Дюкова Е. В.* Усовершенствование алгоритма C4.5 на основе использования полных решающих деревьев // Математические методы распознавания образов // Доклады 14-й Всероссийской конференции. Москва: МАКС Пресс. — 2009. — С. 104–107.

[3] *Генрихов И. Е., Дюкова Е. В.* Построение и исследование распознающих процедур на основе полных решающих деревьях // Информационная обработка информации // Доклады 8-й Международной конференции. Москва: МАКС Пресс. — 2010. — С. 117–121.

[4] *Genrikhov I. E.* Synthesis and analysis recognition procedure on based of complete decision trees // J. Pattern Recognition and Image Analysis. — 2011. — V. 21, N. 1. — Pp. 45–51.

[5] *Донской В. И., Башта А. И.* Дискретные модели принятия решений при неполной информации. — Симферополь: Таврия, 1992. — С. 33–74.

[6] *Djukova E. V., Peskov N. V.* A classification algorithm based on the complete decision tree // J. Pattern Recognition and Image Analysis. — 2007. — V. 17, N. 3. — Pp. 363–367.

[7] *Журавлев Ю. И., Рязанов В. В., Сенько О. В.* Распознавание Математические методы Программная система Практические применения. — Москва: ФАЗИС. — 2006.

[8] *Peng L., Lei L.* A review of missing data treatment methods // An International Journal: Intelligent Information Management Systems and Technologies. — 2005. — V. 1, N. 3. — Pp. 412–419.

[9] *Quinlan J. R.* C4.5: Programs for Machine Learning. Morgan Kaufmann. San Mateo. CA. — 1993.

[10] *Quinlan J. R.* See5/C5.0 DEMO. RULEQUEST RESEARCH. — 2008. — `http://www.rulequest.com/see5-info.html`

Исследование комбинаторных свойств и сложности построения полных решающих деревьев*

Генрихов И. Е., Дюкова Е. В.
ingvar1485@rambler.ru, edjukova@mail.ru
Москва, Учреждение Российской академии наук Вычислительный центр им. А. А. Дородницина Российской академии наук

В докладе изучаются комбинаторные свойства полных решающих деревьев. Получены оценки максимального числа всех вершин, полных вершин, висячих вершин и число ребер дерева. В случае бинарной информации получена оценка емкости (VCD) класса полных решающих деревьев. Рассмотрены вопросы оценки времени построения полных решающих деревьев с энтропийный критерием ветвления для задач распознавания по прецедентам.

Одним из известных инструментов для решения задачи распознавания по прецедентам являются деревья решений. В работах [1–5] предложен новый подход к построению деревьев решений. Данный подход позволяет учитывать все признаки, удовлетворяющие критерию ветвления в равной мере или почти равное мере. В результате на каждой итерации (шаге) строится специальная вершина, называемая полной, которой соответствует набор признаков $X = \{x_{j_1}, \ldots, x_{j_q}\},\ q \geqslant 1$. Для каждого признака $x \in X$ строится обычная внутренняя вершина, из которой осуществляется ветвление. Полученная в результате конструкция названа полным решающим деревом (ПРД). Показано, что ПРД позволяет повысить точность распознавания на прикладных задачах по сравнению с классическими методами построения деревьев решений.

Одной из важных задач является изучение комбинаторных свойств деревьев решений, т. е. получение оценок количественных характеристик дерева. Знание комбинаторных свойств бинарных решающих деревьев (БРД) позволяет получить статистические оценки надежности БРД на контрольной выборке [6]. Эти свойства БРД используются для определения числа всех неизоморфных бинарных деревьев, для перечислении всех различных БРД, для перечислении всех решающих правил, получаемых при помощи БРД. Комбинаторные свойства БРД применяются при получении оценки емкости (VCD — «Vapnik-Chervonenkis Dimension») класса БРД, а также при получении оценки емкости этого класса с помощью метода программирования оценки VCD (pVCD) [7, 8].

В данной работе представлены результаты, касающиеся изучения комбинаторных свойств ПРД. Получены оценки максимального числа листьев дерева, максимального числа обычных вершин, максимального числа ребер, максимального числа полных вершин и максимальной глубины дерева. С использованием метода pVCD получена оценка VCD класса ПРД с μ листьями и n признаками.

Для задач распознавания с n признаками, m обучающими объектами и l классами, приведены временные оценки сложности построения ПРД. Указанные оценки получены для класса ПРД без ограничения на число признаков, образующих полную вершину, а также для некоторых подклассов с ограничениями на число признаков, образующих полную вершину.

Комбинаторные свойства и оценка емкости ПРД

Рассматривается задача распознавания по прецедентам с системой признаков $\{x_1, \ldots, x_n\}$, с непересекающимися классами $K_1, \ldots, K_l$, $l \geqslant 2$, и множеством обучающих объектов $T = \{S_1, \ldots, S_m\}$, где $S_r = (a_{r1}, \ldots, a_{rn})$, $a_{rj} \in \{\mathbb{R}, «*»\}$, $r \in \in \{1, \ldots, m\}$, $j = 1, \ldots, n$. Если $a_{rj} =$«*», то значение признака x_j для объекта S_r не определено. Пусть далее $S = (b_1, \ldots, b_n)$ — распознаваемый объект, и $b_j \in \{\mathbb{R}, «*»\}$, $j = 1, \ldots, n$.

Опишем структуру дерева. Пусть $\hat{T}$ — подмножество обучающих объектов и $X(\hat{T})$ — подмножество признаков, рассматриваемые на текущем шаге построения ПРД. На первом шаге $\hat{T} = T$, $X(\hat{T}) = \{x_1, \ldots, x_n\}$.

При построении ПРД могут встречаться два типа внутренних вершин: полные и обычные вершины. Обычная вершина в ПРД эта вершина дерева, которая соответствует некоторому признаку x и из которой выходит две дуги. При спуске по каждой из этих дуг происходит удаление некоторых объектов из $\hat{T}$ и удаление x из $X(\hat{T})$. Полная вершина — это вершина дерева, которой соответствует набор признаков $X \subseteq X(\hat{T})$, каждый из которых удовлетворяет критерию ветвления в равной или почти равной мере. На каждом шаге построения ПРД формируется в точности одна полная вершина. Из полной вершины X выходит ровно q дуг, где q — число признаков в наборе X. Каждая из этих дуг входит в обычную внутреннюю вершину x, $x \in X$.

Работа выполнена при частичной финансовой поддержке РФФИ, проект № 10-01-00770, гранта Президента РФ по поддержке ведущих научных школ НШ № 7950.2010.1

Определение 1. *Глубиной ветви в ПРД называется число обычных вершин, которые содержит эта ветвь.*

Определение 2. *Ребром в ПРД называется дуга выходящая из обычной вершины и входящая в полную или висячую вершину дерева, а также дуга выходящая из полной и входящая в обычную вершину.*

Определение 3. *Глубиной ПРД называется максимальная глубина среди всех построенных ветвей дерева.*

В случае вещественнозначной информации обычная вершина, соответствующая признаку x, помечается парой $\big(x, d(x)\big)$, где $d(x)$ — «оптимальный» порог перекодировки [3, 4]. Спуск из вершины $\big(x, d(x)\big)$ происходит по двум ветвям, при этом левая ветвь помечается 0, а правая — 1. При спуске из вершины $\big(x, d(x)\big)$ по левой (правой) ветви удаляются те объекты из $\hat{T}$, для которых значение признака x больше (не больше) $d(x)$. Обозначим через $\hat{T}_l$ ($\hat{T}_r$) — подмножество обучающих объектов для левой (правой) ветви. Пусть v — вершина, порожденная ветвью дерева с обычными вершинами $x_{j_1}, \ldots, x_{j_r}$, и пусть дуга, выходящая из вершины x_{j_i}, имеет метку σ_i, $i \in \{1, \ldots, r\}$. Набор $N_v = (\alpha_1, \ldots, \alpha_n)$, где $\alpha_{j_i} = \sigma_i$ при $i = 1, \ldots, r$, и $\alpha_j =$«$*$» при $j \notin \{j_1, \ldots, j_r\}$, назовем порождающим для вершины v. Если v не является висячей вершиной, то ей ставится в соответствие набор N_v. Если v — висячая вершина, то поставим ей в соответствие пару $\big(N_v, \{\omega_v^1, \ldots, \omega_v^l\}\big)$, в которой N_v — порождающий набор для вершины v, $\{\omega_v^1, \ldots, \omega_v^l\}$ — вектор оценок за классы. Описанием объекта $\bar{S} = (b_1, \ldots, b_n)$ в вершине v будем называть вектор $\bar{S}(v) = (\beta_1, \ldots, \beta_n)$, в котором $\beta_{j_i} = 1$, если $b_{j_i} > d(x_{j_i})$, иначе $\beta_{j_i} = 0$ при $i = 1, \ldots, r$, и $\beta_j =$«$*$» при $j \notin \{j_1, \ldots, j_r\}$. В случае бинарной информации ветвление из обычной вершины, помеченной x, осуществляется стандартным способом.

Основной числовой характеристикой ПРД является максимальная глубина дерева k. Покажем, что $k = \min\{n, m-1\}$. Очевидно, $k \leqslant n$, т. к. по построению ПРД при спуске из обычной вершины текущее число признаков уменьшается на единицу.

Теорема 1. *Если $n \geqslant m-1$, то $k = m-1$.*

Доказательство теоремы 1 проводится индукцией по m.

В приведеном ниже следствии 1 приведены условия, при которых $k = m-1$.

Следствие 1. *Пусть $n \geqslant m-1$, $\big|\hat{T}\big| > 2$ и пусть при спуске из обычной вершины строится в точности одна висячая вершина и в эту висячую вершину попадает описание только одного обучающего объекта $s \in \hat{T}$, такого что объекты из $\hat{T} \setminus \{s\}$ не принадлежат одному классу. Тогда $k = m-1$.*

Оценим максимальное число полных вершин FV, максимальное число обычных вершин SV, максимальное число висячих вершин (листьев, терминальных вершин) L, максимальное число ребер R в ПРД. Следует отметить, что данные числовые характеристики существенно зависят от максимальной глубины дерева k. В [6, 9] показано, что для БРД имеем: $R = 2SV$, $L = 2^k$ и $SV = L - 1$.

Обозначим через A_n^k — число размещений из n по k.

Теорема 2. *Если k — максимальная глубина ПРД, то $L = 2^k A_n^k$, $SV = \sum_{i=1}^{k} 2^{i-1} A_n^i$, $R = 3SV$, $FV = \sum_{i=1}^{k} 2^{i-1} A_n^{i-1}$.*

Теорема 3. *Для ПРД при выполнении условий следствия 1 имеет место $FV = \sum_{i=1}^{k} A_n^{i-1}$, $SV = \sum_{i=1}^{k} A_n^i$, $L = \sum_{i=1}^{k} A_n^i + A_n^k$, $R = 3SV$, где $k = m-1$ — максимальная глубина дерева.*

Доказательства теорем 2 и 3 основаны на оценке числовых характеристик SV, L, FV и R на каждом ярусе дерева. При этом под ярусом i – ого уровня ПРД понимается совокупность полных и обычных вершин с глубиной $i-1$ и листьев дерева с глубиной i. При доказательстве теоремы 3 используется оценка для k, приведенная в теореме 1.

Важной задачей является приближенное вычисление значений рассматриваемых характеристик при больших значениях n.

Приведем асимптотические оценки для SV при $n \to \infty$ Заметим, что для всех n, $i \in \mathbb{N}$, $n \geqslant 1$ и $i \leqslant n$, $A_n^i \leqslant n^i$. Поэтому при выполнении условий теоремы 2 получаем

$$SV = \sum_{i=1}^{k} 2^{i-1} A_n^i \leqslant \frac{1}{2} \sum_{i=1}^{k} 2^i n^i = \frac{2^k n^{k+1} - n}{2n-1} \lesssim 2^{k-1} n^k.$$

В условиях теоремы 3 имеем

$$SV = \sum_{i=1}^{k} A_n^i \leqslant \sum_{i=1}^{k} n^i = \frac{n(n^k - 1)}{n-1} \lesssim n^k;$$

$$SV = \sum_{i=1}^{k} A_n^i = A_n^k \left(1 + \frac{1}{(n-k+1)} + o\left(\frac{1}{n}\right)\right).$$

Если $n = m-1$, то в условиях теоремы 3 получаем

$$SV = \sum_{i=1}^{n} A_n^i = A_n^n \left(\sum_{i=0}^{n-1} \frac{1}{i!}\right) \approx n!e.$$

Для случая бинарной информации, с помощью метода программирования оценки VCD (pVCD) [7, 8], получим оценку емкости $FD_{\mu,n,l,k}$ —

семейство ПРД с не более μ листьями и глубиной не больше k, для задачи с n признаками и l классами.

Висячую вершину v в ПРД можно задать парой $\big(B_v, \{\omega_v^1, \ldots, \omega_v^l\}\big)$, где B_v — элементарная конъюнкция (э. к.) над переменными $x_1, \ldots, x_n$ вида $x_{j_1}^{\sigma_1}, \ldots, x_{j_r}^{\sigma_r}$, в которой $x^\sigma = x$, если $\sigma = 1$, иначе $x^\sigma = \overline{x}$; ω_v^i — оценка за класс K_i, вносимая вершиной v [1–4].

Пусть w — сумма рангов всех э.к., которые описывают все висячие вершины ПРД. Очевидно $w \leqslant k\mu$, где k — глубина дерева.

Закодируем ПРД из $FD_{\mu,n,l,k}$ в виде двоичного слова p. Слово p представляется в виде последовательности двоичных слов сформированных из блоков. Каждый блок содержит описание одной из висячих вершин и имеет следующий вид: первая часть блока — описание э. к. B_v или число 0 — разделитель блоков; вторая часть блока — вектор оценок за классы $K_1, \ldots, K_l$, $l \geqslant 2$. Описание B_v представляет собой последовательность двоичных слов полученных из блоков вида: первая часть блока — номер переменной x_j, $j = 1, \ldots, n$, или число $n+1$ — разделитель для выделения вектора оценок за классы; вторая часть блока — двоичная цифра 1, если переменная x_j входит в B_v с инверсией, иначе двоичная цифра 0. Если ПРД содержит менее μ листьев, то последние блоки слова p заполняться нулями. При этом, если в слове p встретится подряд два числа 0, то это означает конец слова p.

Для того, чтобы представить в двоичном коде число из $\{0, 1, ..., n, n+1\}$, достаточно $q = \lceil log(n+2) \rceil$ двоичных разрядов (бит). Поскольку номера переменных начинаются с единицы, то число $n+1$ в слове p можно использовать как признак разделения между э. к. B_v и вектором оценок в описании висячей вершины v, а число 0 можно использовать как разделитель между описаниями различных висячих вершин в p. Чтобы указать значение σ переменной x^σ требуется один бит. Для того, чтобы записать вектор оценок достаточно $32\,l$ двоичных разрядов, т. к. оценка за класс — вещественнозначное число. При таком кодировании на описание висячей вершины v_j понадобится $r_j(q+1) + q + 32l$ бит, где r_j — ранг э. к. B_{v_j}, еще $(\mu - 1)q$ бит понадобится на разделители между описаниями различных висячих вершин. Из этого следует, что длина слова p ($\mathrm{len}(p)$) не превысит величины

$$(\mu-1)q + \sum_{j=1}^{\mu}\big(r_j(q+1) + q + 32l\big) = w + 32\mu l + \\ + (2\mu - 1 + w)q \leqslant \mu(k + 32l) + \big((k+2)\mu - 1\big)q.$$

Поскольку приведено точное определение структуры слова p, то очевиден алгоритм его расшифровки.

Использование метода pVCD позволяет записать следующее неравенство

$$VCD(FD_{\mu,n,l,k}) \leqslant pVCD(FD_{\mu,n,l,k}) = \mathrm{len}(p).$$

Из этого неравенства следует, что для задачи с l классами и n булевыми признаками емкость семейства ПРД с не более чем μ листьями и глубиной не более k не превышает $len(p)$. Для сравнения оценка емкости класса БРД с μ листьями и n признаками, приведенная в [8], равна

$$(\mu - 1)\big(\lceil log(n) \rceil + \lceil log(\mu + 3) \rceil\big).$$

В алгоритмах, описанных в работах [1, 2], вектор оценок $\{\omega_v^1, \ldots, \omega_v^l\}$ — бинарный, т. е. $\omega_v^i = 1$, если описание хотя бы одного объекта из класса K_i попадает в висячую вершину v, иначе $\omega_v^i = 0$. В этом случае, для того чтобы записать вектор оценок для висячей вершины v достаточно l двоичных разрядов. Поэтому оценка емкости класса $FD_{\mu,n,l,k}$ принимает вид

$$pVCD(FD_{\mu,n,l,k}) \leqslant \mu(k+l) + \big((k+2)\mu - 1\big)r.$$

Оценка времени синтеза ПРД

Для оценки времени построения ПРД используется рекуррентное соотношение

$$T(n,m,l) = n\Big(O\big(n(m-1)\big) + \\ + T(n-1, m-1, l)\Big) + O\big(nm(m-1)l\big). \quad (1)$$

Выведем это рекуррентное соотношение. Под элементарными операциями понимаются следующие: простое присваивание, просмотр элемента матрицы, арифметические операции, операции сравнения и логические операции. Заметим, что время построения полной вершины на первом шаге синтеза ПРД равно $O\big(nm(m-1)l\big)$ (необходимо просмотреть n признаков, при этом для выбора оптимального порога $d(x)$ признака x требуется $O\big((m-1)ml\big)$ элементарных операций). Оценка времени спуска из обычной вершины на первом шаге равна $O\big(n(m-1)\big)$, т. к. в худшем случае $m-1$ обучающих объектов попадет в одну из двух ветвей и для того, чтобы построить подмножество обучающих объектов из $\hat{T}$ для этой ветви понадобится не более $O\big(n(m-1)\big)$ элементарных операций. В худшем случае на первом шаге построения дерева число признаков, образующих полную вершину, будет ровно n, на следующем шаге число признаков не превосходит $n-1$, и т. д.. Из сказанного следует, что для времени $T(n,m,l)$ построения ПРД (в худшем случае) справедливо (1).

Из (1) индукцией по числу шагов синтеза ПРД доказывается теорема 4.

Теорема 4. *Имеет место*

$$T(n,m,l) = O\big(A_n^k(m-k)(m-k+1)l\big) + \\ + O\big(A_n^k(m-k)(n-k+1)\big),$$

где максимальная глубина ПРД $k = \min\{n, m-1\}$.

В частном случае, если на каждом шаге построения в полную вершину попадает $\lceil \hat{n}/2 \rceil$ признаков, где $\hat{n} = \big|X(\hat{T})\big|$, то получаем

$$T(n,m,l) = \lceil n/2 \rceil \Big(O\big(n(m-1)\big) + \\ + T(n-1,m-1,l)\Big) + O\big(nm(m-1)l\big) \quad (2)$$

(здесь $\lceil N \rceil$ — наименьшее целое число, не меньшее N)

Теорема 5. *Если на каждом шаге синтеза ПРД в полную вершину попадает* $\lceil \hat{n}/2 \rceil$ *признаков, где* $\hat{n} = \big|X(\hat{T})\big|$, *то*

$$T(n,m,l) = O\bigg((k+1)\frac{A_n^k}{2^k}(m-k)(m-k+1)\,l\bigg) + \\ + O\bigg((k+1)\frac{A_n^k}{2^k}(m-k)(n-k+1)\bigg),$$

где максимальная глубина ПРД $k = \min\{n, m-1\}$.

При доказательстве теоремы 5 используется (2) и приведенная ниже оценка

$$\left\lceil \frac{n}{2} \right\rceil \left\lceil \frac{n-1}{2} \right\rceil \cdot \ldots \cdot \left\lceil \frac{n-k+1}{2} \right\rceil \leqslant (k+1)\frac{A_n^k}{2^k},$$

для всех $n, k \in \mathbb{N}$ и $k \leqslant n$.

Если на каждом шаге построения ПРД в полную вершину может попасть не больше трех признаков из $X(\hat{T})$, то получаем следующее рекуррентное соотношение

$$T(n,m,l) = \min\{3,n\}\Big(O\big(n(m-1)\big) + \\ + T(n-1,m-1,l)\Big) + O\big(nm(m-1)\,l\big). \quad (3)$$

Теорема 6. *Если на каждом шаге синтеза ПРД в полную вершину попадает не больше трех признаков из* $X(\hat{T})$, *то*

$$T(n,m,l) = \begin{cases} O\big(3^{n-1}(m-n)(m-n+1)\,l\big), & n \leqslant m; \\ O\big(3^{m-1}(n-m+2)\big), & n > m. \end{cases}$$

При доказательстве теоремы 6 используется (3).

Следует отметить, что наличие пропусков в признаковых описаниях обучающих объектов снижает время синтеза ПРД. Во-первых, при построении полной вершины $X \subseteq X(\hat{T})$ в нее могут попасть только признаки, для которых существует оптимальный порог. Наличие пропусков уменьшает возможность найти оптимальный порог $d(x)$ для признака x. Во-вторых, при вычислении информативности по энтропийному критерию используются только те объекты из $\hat{T}$, для которых определены значения по признаку x. В-третьих, на этапе спуска из обычной вершины, соответствующей признаку x, $x \in X(\hat{T})$, в подмножество объектов для левой (правой) ветви попадут только те объекты из $\hat{T}$, для которых определены значения по признаку x.

Это обусловлено применяемой в [3, 4] методикой построения ПРД, суть которой заключается в использовании при синтезе дерева только имеющейся информации, при этом никаких «замен» вместо пропусков не происходит.

Литература

[1] *Генрихов И. Е.* Построение полного решающего дерева на базе алгоритма C4.5 // Сообщение по прикладной математике. Москва: ВЦ РАН — 2009.— 24 с.

[2] *Генрихов И. Е., Дюкова Е. В.* Усовершенствование алгоритма C4.5 на основе использования полных решающих деревьев // Математические методы распознавания образов // Доклады 14-й Всероссийской конференции. Москва: МАКС Пресс. — 2009. — С. 104–107.

[3] *Генрихов И. Е., Дюкова Е. В.* Построение и исследование распознающих процедур на основе полных решающих деревьях // Информационная обработка информации // Доклады 8-й Международной конференции. Москва: МАКС Пресс. — 2010. — С. 117–121.

[4] *Genrikhov I. E.* Synthesis and analysis recognition procedure on based of complete decision trees // J. Pattern Recognition and Image Analysis. — 2011. — V. 21, N. 1. — Pp. 45–51.

[5] *Djukova E. V., Peskov N. V.* A classification algorithm based on the complete decision tree // J. Pattern Recognition and Image Analysis. — 2007. — V. 17, N. 3. — Pp. 363–367.

[6] *Донской В. И., Башта А. И.* Дискретные модели принятия решений при неполной информации. — Симферополь: Таврия, 1992. — С. 33–74.

[7] *Донской В. И.* Колмогоровская сложность классов общерекурсивных функций с ограниченной емкостью // Таврический вестник информатики и математики. — 2005. — № 1. — С. 25–34.

[8] *Донской В. И.* Оценки емкости классов алгоритмов эмпирического обобщения, полученные pVCD методом // Ученые записки Таврического национального университета им. В. И. Вернадского, Том. 23(62). — 2010. — № 2. — С. 56–65.

[9] *Мошков М. Ю.* Деревья решений. Теория и приложения. Учебное пособие. — Н. Новгород: изд–во Нижегородского ун-та, — 1994. — 175 с.

Оптимизация распараллеливания алгоритма построения диагностических тестов*

Янковская А. Е.
ayyankov@gmail.com
Томск, Томский государственный архитектурно-строительный университет

Рассматривается одно из направлений сокращения временных затрат при реализации алгоритма построения диагностических тестов (ДТ) в задачах распознавания образов, а именно, оптимизация путей его распараллеливания. Алгоритм построения ДТ реализуется в два этапа: построение по матричному представлению данных и знаний, так называемой, безызбыточной матрицы импликаций ($\mathbf{U}'$), задающей различимость объектов из разных образов, и нахождение безызбыточных (h-кратных при построении отказоустойчивых ДТ) столбцовых покрытий матрицы $\mathbf{U}'$. На 1-м этапе оптимизация осуществляется путём иерархической декомпозиции матричного представления данных и знаний, обеспечивающей равномерность загрузки процессоров на каждом уровне иерархии. На 2-м этапе задача сводится к построению дерева поиска в виде иерархической системы подматриц матрицы $\mathbf{U}'$ с одновременным обходом вершин дерева поиска, являющаяся по сути параллельной задачей обработки поддеревьев поиска и сводимая к выбору поддеревьев для параллельной обработки. Даны оригинальные алгоритмы оптимизации распараллеливания построения ДТ.

В проблематике тестового распознавания образов в связи с полиномиальной сложностью построения диагностических тестов (ДТ) важнейшее место по-прежнему занимает разработка и исследование алгоритмов построения безызбыточных безусловных диагностических тестов (ББДТ). Одним из направлений сокращения временных затрат при реализации алгоритма построения ДТ в интеллектуальных системах является оптимизация путей распараллеливания алгоритмов построения ББДТ.

Впервые задача распараллеливания при построении ББДТ была поставлена в статье [1], где и был приведён первый алгоритм, развитый далее в публикации [2]. В основу этих алгоритмов положена идея иерархической декомпозиции матричного представления данных и знаний на подматрицы, применяемая для построения, так называемой, безызбыточной матрицы импликаций (БМИ), задающей различимость объектов из разных образов и служащей для нахождения кратчайших и безызбыточных столбцовых покрытий с целью построения минимальных и безызбыточных подмножеств (задающих описание объектов характеристических признаков) и являющихся, по сути, соответственно минимальными и безызбыточными ДТ.

Поскольку задача поиска кратчайших и безызбыточных столбцовых покрытий сводится к построению дерева поиска в виде иерархической системы подматриц БМИ с одновременным обходом вершин дерева поиска [3], являющаяся по своей сути параллельной задачей обработки поддеревьев дерева поиска, то она естественно сводится к выбору поддеревьев для параллельной обработки.

Особенно возрастает объём перебора при необходимости принятия отказоустойчивых решений [4,5,1], т. е. решений устойчивых к ошибкам измерения (занесения) значений признаков исследуемого (распознаваемого) объекта, что существенно повышает актуальность процесса распараллеливания построения отказоустойчивых ДТ, на основе которых принимаются решения.

Основные понятия и определения

Для представления данных и знаний, применяемых для построения ББДТ, используется матричная модель [6], включающая булеву или целочисленную матрицу описаний ($\mathbf{Q}$), задающую описание объектов в пространстве характеристических признаков $z_1, \ldots, z_m$, и целочисленную матрицу различений ($\mathbf{R}$), задающую разбиение объектов на классы эквивалентности по каждому механизму классификации. Если значение характеристического признака несущественно для объекта, то данный факт отмечается прочерком («–») в соответствующем элементе матрицы $\mathbf{Q}$.

Множество всех неповторяющихся строк матрицы $\mathbf{R}$ сопоставлено множеству выделенных образов, представленных одностолбцовой матрицей $\mathbf{R}'$, элементами которой являются номера образов.

Матричное представление данных и знаний приведено на рис. 1.

$$\boldsymbol{Q}=\begin{array}{c}\begin{matrix}z_1 & z_2 & z_3 & z_4 & z_5 & z_6 & z_7 & z_8 & z_9 & z_{10} & z_{11}\end{matrix}\\ \begin{bmatrix}4 & 4 & 5 & 2 & 3 & 2 & 7 & 8 & 3 & 4 & 1\\ 3 & 4 & 5 & 3 & 3 & 2 & 4 & 5 & 3 & 4 & 1\\ 3 & 4 & 4 & 1 & 4 & 4 & 2 & 3 & 1 & 5 & 1\\ 2 & 4 & 2 & 1 & 6 & - & 4 & 5 & 2 & 3 & 1\\ 1 & 4 & 3 & 2 & 5 & 2 & 1 & 2 & 3 & 4 & 1\\ 3 & 4 & 2 & 2 & 6 & 2 & 2 & 3 & 3 & 2 & 1\end{bmatrix}\end{array}\begin{matrix}1\\2\\3\\4\\5\\6\end{matrix}\quad \boldsymbol{R}=\begin{array}{c}\begin{matrix}k_1 & k_2\end{matrix}\\ \begin{bmatrix}1 & 2\\ 1 & 2\\ 2 & 1\\ 2 & 1\\ 1 & 3\\ 1 & 3\end{bmatrix}\end{array}\quad \boldsymbol{R'}=\begin{bmatrix}1\\1\\2\\2\\3\\3\end{bmatrix}$$

Рис. 1. Матричное представление данных и знаний.

Данная модель позволяет представлять не только данные, но и знания экспертов, поскольку одной строкой матрицы $\mathbf{Q}$ можно задавать в интервальной форме подмножество объектов, для которых характерны одни и те же итоговые решения, задаваемые соответствующими строками матрицы $\mathbf{R}$.

Работа выполнена при финансовой поддержке РФФИ, проект № 10-01-00462 и РГНФ, проект № 10-06-64604.

Диагностическим тестом называется совокупность признаков, различающих любые пары объектов, принадлежащих разным образам. ДТ называется безызбыточным (тупиковым [7]), если содержит безызбыточное количество признаков.

Закономерности в данных и знаниях

К закономерностям [6] будем относить следующие подмножества признаков: константные (принимающие одно и тоже значение для всех образов), устойчивые (константные внутри образа, но не являющиеся константными), неинформативные (не различающие ни одной пары объектов), альтернативные (в смысле включения в ББДТ), зависимые (в смысле включения подмножеств различимых пар объектов), несущественные (не входящие ни в один ББДТ), обязательные (входящие во все ББДТ), а также все минимальные и все (либо часть — при большом признаковом пространстве) безызбыточные различающие подмножества признаков, являющиеся, по сути, соответственно минимальными ДТ и ББДТ.

Для выявления закономерностей применяется процедура построения булевой (целочисленной [8]) матрицы импликаций $\mathbf{U}$ [6], столбцы которой сопоставлены столбцам матрицы $\mathbf{Q}$, а строки — всевозможным парам объектов v, l, соответственно из разных образов a, b (классов); $v \in \{1, 2, \ldots, \sigma(\mathbf{Q^a})\}$, $l \in \{1, 2, \ldots, \sigma(\mathbf{Q^b})\}$, где $\sigma(\mathbf{Q^a})$ ($\sigma(\mathbf{Q^b})$) — количество строк в подматрице $\mathbf{Q^a}$ ($\mathbf{Q^b}$) матрицы $\mathbf{Q}$. Строка $\mathbf{U}_i$ матрицы $\mathbf{U}$ представляет собой значение булевой (целочисленной) вектор-функции различения, j-я ($j \in \{1, 2, \ldots, m\}$) компонента $u_{i,j}$ которой вычисляется по формуле: $u_{i,j} = |q^a_{v,j} - q^b_{l,j}|$, где $q^a_{v,j}$ ($q^b_{l,j}$) — значение признака z_j для объекта v (l);

$$i = \prod_{r=1}^{a} \sigma(\mathbf{Q}^r) \sum_{s=r+1}^{b-1} \sigma(\mathbf{Q}^s) + \sum_{r=1}^{v-1} \sigma(\mathbf{Q}^b) + vl,$$

а $u_{i,j}$ вычисляется для каждой пары образов.

Будем говорить, что строка $\mathbf{U}_d$ поглощает строку $\mathbf{U}_l$ ($\mathbf{U}_d \succ \mathbf{U}_l$), если

$$(\mathbf{U}_d \succ \mathbf{U}_l) \leftrightarrow \forall i \in I(u_{di} \geqslant u_{li}),$$

где $i \in \{1, \ldots, s\}$ — множество строк матрицы $\mathbf{U}$.

БМИ назовем такую матрицу $\mathbf{U}'$, в которой отсутствуют поглощающие строки.

Построение $\mathbf{U}'$ для булевых матриц $\mathbf{Q}$ дано в [6].

Оптимизация распараллеливания построения диагностических тестов

Параллельный алгоритм построения ДТ [2] включает два этапа: параллельный алгоритм построения БМИ ($\mathbf{U}'$), задающий различимость объектов из разных образов, и параллельный алгоритм построения безызбыточных столбовых покрытий матрицы $\mathbf{U}'$.

Основная идея механизма распараллеливания при построении матрицы $\mathbf{U}'$ заключалась в многоуровневом разбиении матриц $\mathbf{Q}$ и $\mathbf{R}'$ на пары подматриц $\mathbf{Q}_{i,j}$ и $\mathbf{R}'_{i,j}$ (i — номер уровня разбиения матриц $\mathbf{Q}$ и $\mathbf{R}'$ ($i \in \{1, 2, \ldots, \omega\}$); j — номер подматрицы $\mathbf{Q}_{i,j}$ на i-ом уровне разбиения, j является функцией от i, т. е. $j(i)$ и $j \in \{1, 2, \ldots, r_i\}$), параллельном построении подматриц $\mathbf{U}'_{i,g}$ (g — номер подматрицы $\mathbf{U}'_{i,g}$ на i-ом уровне разбиения, $g \in \{1, 2, \ldots, r_i/2\}$) по подматрицам $\mathbf{Q}_{i,j}$ и $\mathbf{R}'_{i,j}$ и объединении подматриц $\mathbf{U}'_{i,g}$ в матрице $\mathbf{U}'$ с одновременным удалением поглощающих строк.

В результате многоуровневого разбиения матриц $\mathbf{Q}$ и $\mathbf{R}'$ на пары подматриц $\mathbf{Q}_{i,j}$ и $\mathbf{R}'_{i,j}$ и $\mathbf{Q}_{i,j+1}$ и $\mathbf{R}'_{i,j+1}$ получим на 1-м уровне две пары подматриц $\mathbf{Q}_{1,1}$ и $\mathbf{R}'_{1,1}$ и $\mathbf{Q}_{1,2}$ и $\mathbf{R}'_{1,2}$, а для каждой подматрицы i-го уровня получим разбиение на 2 пары подматриц $i+1$-го уровня ($\mathbf{Q}_{i+1,j}$, $\mathbf{R}'_{i+1,j}$ и $\mathbf{Q}_{i+1,j+1}$ и $\mathbf{R}'_{i+1,j+1}$) и т. д. до тех пор, пока разбиение возможно, т. е. пока все строки каждой подматрицы $\mathbf{Q}_{i,j}$ не будут принадлежать только одному образу.

Отметим, что при параллельном построении матрицы $\mathbf{U}'$ для вычисления весовых коэффициентов признаков применялся m-компонентный вектор $\boldsymbol{\rho}$, каждая компонента которого равна сумме значений элементов m-го столбца матрицы $\mathbf{U}$. Значения компонент вектора $\boldsymbol{\rho}$ вычислялись одновременно с построением очередной вектор-функции различения. Весовой коэффициент характеристического признака определяется по его разделяющей способности для двоичного признака по формуле, приведённой в [6], а для целочисленного — по формуле, приведённой в [1].

Применение целочисленной матрицы $\mathbf{U}'$ необходимо для подсчета весовых коэффициентов целочисленных признаков. Однако, для поиска безызбыточных (h-кратных) столбцовых покрытий матрицы целесообразно использовать двоичную матрицу импликаций $\mathbf{U}''$, заменив отличные от нуля элементы целочисленной матрицы $\mathbf{U}'$ на элементы, имеющие значение 1 и удалив в ней поглощающие строки. При этом формула для вычисления количества подматриц $\mathbf{Q}_{i,j}$ на i-ом уровне рабиения имеет следующий вид: $r_i = 2^i - 2\eta|f - 2^i|$, где f — количество образов в матрице $\mathbf{R}'$; $]\log_2 f[$ — наименьшее сверху целое к $\log_2 f$; ω — количество уровней разбиения, вычисляемое по формуле $\omega =]\log_2 f[$; η — коэффициент, η=0 при $i \neq \omega$ и η=1 при i=ω.

Отметим, что если решается задача распараллеливания при построении отказоустойчивых ДТ, то удаление поглощающих строк при построении матрицы $\mathbf{U}'$ осуществляется по приведённому в [2] модифицированному алгоритму.

Аналогично, как и в публикации [2], представим на рис. 2 дерево построения матрицы $\mathbf{U}'$ на основе многоуровневого разбиения матриц $\mathbf{Q}$ и $\mathbf{R}'$ на подматрицы. Корню дерева сопоставим матрицы $\mathbf{Q}$

и $\mathbf{R}'$. Вершинам i-го ($i \in \{1, 2, \ldots, \omega\}$) уровня дерева сопоставим подматрицы $\mathbf{Q}_{i,j}$ ($j \in \{1, 2, \ldots, r_i\}$), $\mathbf{R}'_{i,j}$ и $\mathbf{Q}_{i,j+1}$, $\mathbf{R}'_{i,j+1}$ и построенные по ним подматрицы $\mathbf{U}'_{i,g}$, где $g = j/2$.

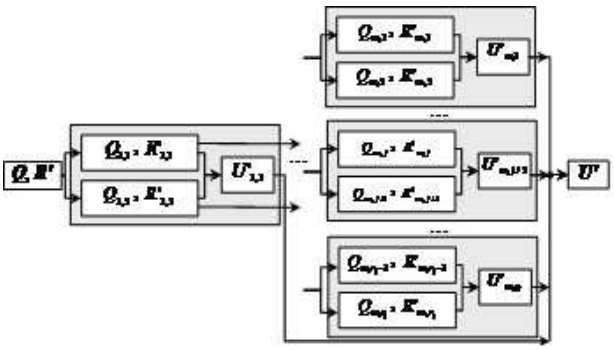


Рис. 2. Дерево построения матрицы $\mathbf{U}'$.

Однако такое разбиение приводит к неравномерной загрузки процессоров на каждом начиная со 2-го (на 1-м используется один процессор) уровне разбиения матриц на подматрицы, поскольку при разбиении на подматрицы не учитываются объем вычислений при построении БМИ.

В отличие от алгоритмов [1, 2] в данном докладе предлагается оптимизацию загрузки процессоров начинать как с последнего уровня разбиения, то есть с уровня ω, так и с 1-го уровня разбиения и выбрать лучший по загрузке процессоров вариант разбиения.

Объем вычислений в процессе обработки пар матриц ($\mathbf{Q}^a$, $\mathbf{R}'^a$) и ($\mathbf{Q}^b$, $\mathbf{R}'^b$) при построении БМИ прямо пропорционален величине $\mathbf{L}_{a,b} = \sigma(\mathbf{Q}^a) \cdot \sigma(\mathbf{Q}^b)$. Следовательно, если разбиение матриц на подматрицы на каждом i-м уровне разбиения произвести так, чтобы время обработки каждой пары подматрицы на i-м уровне было максимально близко друг к другу (точное равенство не всегда возможно в связи с решением реальных задач, в которых, как правило, количество объектов в каждом образе не совпадает), то время построения БМИ процессорами одного уровня разбиения будет максимально близко друг к другу, что обеспечит более равномерную загрузку процессоров.

В этих целях оптимизацию распараллеливания целесообразно производить с последнего уровня, то есть с уровня ω, когда в каждой паре подматриц $\mathbf{Q}^a$, $\mathbf{Q}^b$ будут содержаться подматрицы, состоящие из строк, входящих только в один образ.

Отметим, если бы нам надо было бы учитывать в данной задаче только количество строк в подматрицах (образах), то задача оптимизации загрузки процессоров сводилась бы к задаче разбиения чисел на классы с одинаковой суммой [9].

Однако, объём перебора при построении БМИ в процессе обработки каждой пары подматриц ($\mathbf{Q}^a$, $\mathbf{R}'^a$), ($\mathbf{Q}^b$, $\mathbf{R}'^b$) пропорционален не количеству объектов в образе, а велечине $L_{a,b}$.

При количестве образов, равном 2^k, где k — целое число, в исходных матрицах $\mathbf{Q}$, $\mathbf{R}'$, задача определения таких пар сводится к построению 2^{k-1} пар подматриц ($\mathbf{Q}^a$, $\mathbf{R}'^a$), ($\mathbf{Q}^b$, $\mathbf{R}'^b$), каждая их которых содержит только по 1-му образу, далее к построению 2^{k-2} пар подматриц, каждая их которых содержит по 2-а образа и т. д. до тех пор, пока каждая пара подматриц не будет содержать по $2^k - 1$ образов.

При количестве образов, не равном 2^k, процесс декомпозиции оканчивается для части подматриц на уровне $\omega - 1$, а для другой — на уровне ω.

Учитывая рамки доклада, приведём неформальное описание алгоритма иерархической декомпозиции матриц $\mathbf{Q}$, $\mathbf{R}'$ на подматрицы, при количестве образов, равном 2^k, блок–схема которого представлена на рис. 2 в виде дерева построения матрицы $\mathbf{U}'$.

1. Упорядочение образов, представленных матрицами $\mathbf{Q}$, $\mathbf{R}'$, по невозрастанию количества объектов в образах.
2. Разбиение пары матриц $\mathbf{Q}$, $\mathbf{R}'$ на пары подматриц ($\mathbf{Q}_{1,1}$, $\mathbf{R}'_{1,1}$) и ($\mathbf{Q}_{1,2}$, $\mathbf{R}'_{1,2}$) таким образом, чтобы $\sigma(\mathbf{Q}_{1,1})$ и $\sigma(\mathbf{Q}_{1,2})$ были максимально близки. $i := 1$. Построение подматрицы $\mathbf{U}'_{1,1}$. Вычисление величины $\boldsymbol{\rho}$.
3. $j := 1$.
4. $i := i + 1$.
5. Разбиение каждой пары подматриц $\mathbf{Q}_{i,j}$, $\mathbf{R}'_{i,j}$ на две пары подматриц ($\mathbf{Q}_{i+1,j}$, $\mathbf{R}'_{i+1,j}$) и ($\mathbf{Q}_{i+1,j+1}$, $\mathbf{R}'_{i+1,j+1}$) таким образом, чтобы $\sigma(\mathbf{Q}_{i,j})$ и $\sigma(\mathbf{Q}_{i,j+1})$ были максимально близки. Построение подматрицы $\mathbf{U}'_{i,j}$. Вычисление величины $\boldsymbol{\rho}^i$ ($\boldsymbol{\rho} := \boldsymbol{\rho} + \boldsymbol{\rho}^i$). Если $j > 2^k$, то перейти к п. 4.
6. $j := j + 2$. Если $i > k$, то произвести объединение всех матриц $\mathbf{U}'_{i,j}$ ($i \in \{1, 2, \ldots, k\}$, $j \in \{1, 2, \ldots, 2k\}$) с одновременным удалением поглощающих строк и перейти к п. 7, иначе перейти к п. 3.
7. Конец.

Приведём пример, иллюстрирующий работу алгоритма при количестве образов равном 9.

Пусть после упорядочения строк матриц $\mathbf{Q}$, $\mathbf{R}'$ по образам имеем количество объектов в образах $1, 2, 3, 4, 5, 6, 7, 8, 9$, равным соответственно $12, 10, 9, 8, 6, 5, 4, 3, 3$.

Разобьём множество образов на два подмножества $(2, 3, 6, 8, 9)$ и $(1, 4, 5, 7)$. При этом количество строк в каждом подмножестве равно 30, т. е. объём вычислений при построении матрицы $\mathbf{U}'$ прямо пропорционален величине 900. Выполнив 2-й уровень разбиения, получим следующую пару, каждую из 2-х подмножеств: $((2, 6), (3, 8, 9))$ и $((1, 7),$

$(4,5)$). Количество строк в 1-й паре будет равно $(15,15)$, а во второй — $(16,14)$, объём вычислений по 1-й паре прямо пропорционален величине 225, а по 2-й паре — 224, т. е. оба процессора будут загружены почти одинаково. На 3-м уровне разбиения получим следующие подмножества образов $((2),(6)),((3),(8,9)),\ ((1),(7)),((4),(5))$. Объём вычислений на каждой из пар прямо пропорционален величинам $50,54,48,48$ соответственно, т. е. максимальная загрузка процессора на 3-м уровне прямо пропорциональна величине 54. И наконец, на 4-м уровне разбиения получим подмножество образов $(8,9)$, объём вычислений для которого прямо пропорционален величине 9. Таким образом, общее время загрузки процессора прямо пропорционально величине 1198.

При разбиении множества образов на 2 подмножества $(1,2,3)$, $(4,5,6,7,8,9)$, количество строк, в которых равно 31 и 29 соответственно, и объём вычислений прямо пропорционален величине 899. Последующие разбиения приведут к неравномерной загрузке процессоров: на 2-м уровне максимальная загрузка процессора прямо пропорциональна величине 228, на 3-м уровне — величине 90. а на 4-м уровне — величине 30. Таким образом, общее время загрузки процессоров на всех уровнях прямо пропорционально величине 1347, что существенно выше, чем при 1-м варианте разбиения.

Что касается алгоритма построения ББДТ с использованием процедуры распараллеливания, то в зависимости от количества процессоров, могут быть выделены разные подходы. Так при большом количестве процессоров целесообразно разбиение дерева поиска безызбыточных h-кратных столбцовых покрытий БМИ, по которым строятся ББДТ, производить на максимально равные (по количеству вершин) поддеревья с целью равномерной загрузки процессоров. Построение ББДТ осуществляется с учётом выявленных закономерностей, что существенно сокращает объём перебора.

Выводы

Предложены пути оптимизации распараллеливания алгоритмов построения ДТ при реализации алгоритмов в интеллектуальных системах с матричным представлением данных и знаний.

Разработан оригинальный алгоритм оптимизации распараллеливания при построении БМИ, задающей различимость объектов из разных образов и служащей для нахождения безызбыточных (h-кратных при нахождении отказоустойчивых ДТ) столбцовых покрытий с целью построения ББДТ.

Несмотря на то, что рамки статьи позволили изложить лишь алгоритм только для числа образов, равном 2^k иллюстрирующий пример приведён для нечетного количества образов.

Также предложен новый алгоритм распараллеливания построенного по БМИ дерева поиска безызбыточных столбцовых покрытий БМИ, по которому осуществляется построение ББДТ, сводящийся к выбору поддеревьев для параллельной обработки.

В дальнейшем предполагается провести обоснование оптимальности предложенного алгоритма распараллеливания построения БМИ, выявить в целях оптимизации загрузки процессоров с какого уровня (1-го или последнего) реализовывать алгоритм, а также разработать различные подходы к оптимизации распараллеливания при построении матрицы $\mathbf{U}'$ и к разбиению деревьев поиска на поддеревья.

Литература

[1] *Янковская А. Е., Китлер С. В.* Принятие решений на основе параллельных алгоритмов тестового распознавания образов // Искусственный интеллект. Украина, Донецк: ІПШІ "Наука і освіта". — 2010. — № 3. — С. 151–159.

[2] *A. E. Yankovskaya and S. V. Kitler* Parallel Algorithm of Construction of k-Valued Fault-Tolerant Diagnostic Tests in Intelligent Systems // Pattern Recognition and Image Analysis. — 2011. — V. 21, N. 2, — Pp. 414–417.

[3] *Yankovskaya A. E., Gedike A. I.* Finding of All Shortest Column Coverings of Large Dimension Boolean Matrices // Proc. of the First International Workshop on Multi-Architecture Low Power Design (MALOPD), 1999. — Pp. 52–60.

[4] *Янковская А. Е.* Принятие решений, устойчивых к ошибкам измерения значений признаков в интеллектуальных системах // Искусственный интеллект. Интеллектуальные системы, Таганрог: Издво ТТИ ЮФУ, 2009. — С. 137–130.

[5] *Янковская А. Е., Китлер С. В.* Интеллектуальная система ускоренного построения k-значных отказоустойчивых диагностических тестов // Конференция по искусственному интеллекту (КИИ-2010): Тр. конф. Т. 4, М.: Физматлит, 2010. — С. 72–80.

[6] *Янковская А. Е.* Логические тесты и средства когнитивной графики в интеллектуальной системе // Новые информационные технологии в исследовании дискретных структур, Томск: СО РАН, 2000. — С. 163–168.

[7] *Журавлёв Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики: Вып. 33, Москва: Наука, 1978. — С. 5–68.

[8] *Янковская А. Е.* Построение k-значных диагностических тестов в интеллектуальной системе с матричным представлением знаний // Конференция по искусственному интеллекту (КИИ-2010): Тр. конф. Т. I, Пущино, 1998. — С. 264–269.

[9] *Агибалов Г. П., Беляев В. А., Янковская А. Е.* Алгоритм компоновки схем в модули ограниченной вместимости // УСиМ. — 1976. — № 1. — С. 84–89.

Коррекция информационной матрицы обучающей выборки и её применение к построению линейного решающего правила

Муравьева О. В.
muraveva@mail.ru
Москва, Московский педагогический государственный университет

Рассматривается задача построения линейного решающего правила для двух классов, заданных обучающей выборкой. Предлагается использовать для построения решающего правила методы оптимальной коррекции информационной матрицы. Критерием является наибольшее отклонение по всем признакам всех объектов обучающей выборки.

Имеем задачу классификации двух классов. Объекты определяются набором значений n числовых признаков. Исходной информацией является обучающая выборка. Класс K_1 представлен объектами с векторами признаков $x^1, \ldots, x^l$, класс K_2 — выборкой $x^{l+1}, \ldots, x^m$, $F(x) = a_1x_1 + a_2x_2 + \cdots + a_nx_n - b = (a, x) - b$ — неизвестная линейная разделяющая функция. Соответствующее решающее правило имеет вид: объект $x \in K_1$, если $F(x) \leqslant 0$; $x \in K_2$, если $F(x) > 0$.

Обозначим матрицу параметров модели (информационную матрицу обучающей выборки)

$$X = \left[x^1 \ldots x^l - x^{l+1} \cdots - x^m\right]^\intercal,$$

матрицу скорректированных (возмущенных) значений признаков

$$X' = \left[x'^1 \ldots x'^l - x'^{l+1} \cdots - x'^m\right]^\intercal,$$

$H = X' - X$ — матрица коррекции, вектор

$$(b, \ldots, b, -b, \ldots, -b)^\intercal = bp,$$

где $p = (1, \ldots, 1, -1, \ldots, -1)^\intercal \in \mathbb{R}^m$.

Имеем систему m линейных неравенств относительно $n + 1$ переменной a, b

$$\begin{cases} (a, x^i) \leqslant b, \ i = 1, \ldots, l; \\ (a, x^i) \geqslant b, \ i = l + 1, \ldots, m \end{cases} \quad \text{или} \quad Xa \leqslant bp. \quad (1)$$

Данная система может быть несовместной, иметь единственное решение или несколько решений. Рассмотрим методы выбора разделяющей функции $F(x) = (a, x) - b$ для каждого из этих случаев.

Неразделимая выборка

Если обучающие выборки невозможно безошибочно разделить гиперплоскостью, получим задачу коррекции (аппроксимации), т. е. минимального изменения параметров, при котором система становится совместной:

$$\inf_{H,a,b} \{\Phi(H) \colon (X + H)a \leqslant bp\}.$$

В качестве решающего правила исходной задачи принимается разделяющая гиперплоскость для ближайшей к заданной (в смысле критерия $\Phi(H) = \|H\|$) разделимой обучающей выборки. Минимаксному критерию соответствует l_∞-норма матрицы коррекции

$$\Phi(H) = \|H\|_\infty = \max_{i,j} |h_{ij}|.$$

Преобразуем задачу

$$\inf_{H,a,b} \{\|H\|_\infty \colon (X + H)a \leqslant bp\} \quad (2)$$

к виду

$$v = \inf_{H,a,b,z} \{\|H\|_\infty \colon (X + H)a - bp + z = 0, \ z \geqslant 0\}.$$

Лемма 1. *Система $Hx = d$ при заданных $x \in \mathbb{R}^n$, $x \neq 0$, $d \in \mathbb{R}^m$ имеет решение с минимальной матричной нормой*

$$\|H^*\|_\infty = \frac{\|d\|_\infty}{\|x\|_1} = \frac{\max\limits_i |d_i|}{\sum\limits_j |x_j|};$$

$$H^* = \frac{1}{\sum\limits_j |x_j|}\, d \cdot (\operatorname{sign} x_1, \ldots, \operatorname{sign} x_n).$$

Следовательно,

$$v = \inf_{z \geqslant 0, H,a,b} \{\|H\|_\infty \colon Ha = bp - Xa - z\} = \inf_{z \geqslant 0, a, b} \frac{\max\limits_i |(Xa - bp + z)_i|}{\|a\|_1} = \inf_{z \geqslant 0, a, b} \max_i \frac{|(p_i x^i, a) - bp_i + z_i|}{\sum\limits_j |a_j|},$$

где

$$p_i = \begin{cases} 1, & i = 1, \ldots, l, \\ -1, & i = l + 1, \ldots, m. \end{cases}$$

Последняя минимаксная задача стандартным образом сводится к задаче линейного программирования. Введем новые переменные: $c_j = |a_j|$, $j = 1, \ldots, n$ с соответствующими ограничениями $-c_j \leqslant a_j \leqslant c_j$. Равенство

$$w = \frac{1}{\sum_j c_j}$$

влечет $w\sum_j c_j = 1$, что после введения еще n переменных $r = (r_1, \ldots, r_n)$, $r = wc$ дает $\sum_j r_j = 1$. Обозначим $s = wa \in \mathbb{R}^n$, $t = wb \in \mathbb{R}$, $y = wz \in \mathbb{R}^m$, целевая функция

$$\frac{|(p_i x^i, a) - bp_i + z_i|}{\sum_j |a_j|}$$

примет вид $|(p_i x^i, s) - tp_i + y_i|$. Осталось ввести скалярную переменную $u = \max\limits_i \big|(p_i x^i, s) - tp_i + y_i\big|$. В результате получим задачу линейного программирования

$$\begin{gathered} u \to \min_{u,t\in R, r,s\in\mathbb{R}^n, y\in\mathbb{R}^m} \\ u \geqslant (p_i x^i, s) - tp_i + y_i, \quad i = 1, \ldots, m; \\ u \geqslant -(p_i x^i, s) + tp_i - y_i, \quad i = 1, \ldots, m; \\ -r_j \leqslant s_j \leqslant r_j, \quad j = 1, \ldots, n; \\ \sum_j r_j = 1, \quad y \geqslant 0. \end{gathered} \tag{3}$$

Коэффициенты системы линейных неравенств определяются с точностью до пропорциональности. Выберем для определенности правую часть равную по модулю 1 ($|b| = 1$). Таким образом, имеет место

Теорема 2. *Если u^*, t^*, r^*, s^*, y^* — решение задачи* (3)*, то решение задачи* (2) *определяется формулами*

$$\inf_{H,a,b} \{\|H\|_\infty \colon (X + H)a \leqslant bp\} = u^*;$$

$$a^* = \frac{1}{|t^*|} s^*, \quad b^* = \frac{t^*}{|t^*|}.$$

Построение разделяющей полосы

Полученная выше гиперплоскость проходит через некоторые точки обучающей выборки. Для устранения этого недостатка потребуем, чтобы точки обучающей выборки находились за пределами полосы, ограниченной параллельными гиперплоскостями $L_1 \colon (a, x) - b = 1$ и $L_2 \colon (a, x) - b = -1$.

Обозначим

$$F_1(x) = (a, x) - b - 1;$$
$$F_2(x) = (a, x) - b + 1,$$

решающее правило имеет вид: объект $x \in K_1$, если $F_1(x) \leqslant 0$; $x \in K_2$, если $F_2(x) \geqslant 0$; иначе класс не определен.

Определим ширину полосы, ограниченной гиперплоскостями L_1 и L_2, как минимальное расстояние между двумя точками, принадлежащими разным гиперплоскостям:

$$\min\{\|x - y\| \colon x \in L_1,\ y \in L_2\}.$$

Если рассматривать евклидово расстояние, то ширина полосы равна $\frac{2}{\|a\|_2}$. Этот показатель максимизируется в методе опорных векторов.

Используем расстояние между параллельными гиперплоскостями $L_1 \colon (a, x) - b = 1$ и $L_2 \colon (a, x) - b = -1$, определяемое ∞-нормой вектора.

Лемма 3.

$$\min\{\|x - y\|_\infty \colon (a, x) - b = 1, (a, y) - b = -1\} = \frac{2}{\|a\|_1}.$$

Рассмотрим систему неравенств

$$\begin{gathered} (a, x^i) \leqslant b - d\|a\|_1, \quad i = 1, \ldots, l; \\ (a, x^i) \geqslant b + d\|a\|_1, \quad i = l + 1, \ldots, m; \\ d > 0. \end{gathered}$$

Выполнение данных условий означает, что обучающая выборка разделена полосой ширины $2d$.

Если при заданных X и d решение не существует, выполним минимальную коррекцию таблицы обучения $X' = X + H$, в результате которой существует разделяющая полоса. Задача имеет вид

$$\begin{gathered} \|X' - X\|_\infty \to \min; \\ (a, x'^i) \leqslant b - d\|a\|_1, \quad i = 1, \ldots, l; \\ (a, x'^i) \geqslant b + d\|a\|_1, \quad i = l + 1, \ldots, m. \end{gathered}$$

Выполнив преобразования, аналогичные приведенным выше для задачи (2), получим задачу линейного программирования.

Теорема 4. *Если u^*, t^*, r^*, s^*, y^* — решение задачи линейного программирования*

$$\begin{gathered} u \to \min_{u,t\in R, r,s\in\mathbb{R}^n, y\in\mathbb{R}^m} \\ u \geqslant (x^i, s) - tp_i - d + y_i, \quad i = 1, \ldots, m; \\ u \geqslant -(x^i, s) + tp_i + d - y_i, \quad i = 1, \ldots, m; \\ -r_j \leqslant s_j \leqslant r_j, \quad j = 1, \ldots, n; \\ \sum_j r_j = 1, \quad y \geqslant 0, \end{gathered}$$

то разделяющая полоса ширина $2d$, соответствующая минимальной коррекции таблицы обучения, определяется гиперплоскостями

$$(a^*, x) \leqslant b^* - d\|a^*\|_1 \ \text{ и } \ (a^*, x) \geqslant b^* + d\|a^*\|_1;$$

$$\text{где } a^* = \frac{1}{|t^*|} s^*, \quad b^* = \frac{t^*}{|t^*|}.$$

Параметрическая устойчивость разделяющей гиперплоскости

Пусть теперь система (1) совместна и имеет несколько решений. Иначе говоря, представители классов в обучающей выборке линейно разделимы,

причем существует разделяющая гиперплоскость, не проходящая через точки обучающей выборки.

Рассмотрим, как изменяется свойство разделимости при измененении параметров задачи, т. е. информационной матрицы обучающей выборки. Предельное изменение любого параметра (элемента матрицы X), при котором функция остается разделяющей, назовем радиусом устойчивости данной гиперплоскости:

$$R(a,b) = \inf_H \{\|H\|_\infty \colon (X+H)a \not\leqslant bp\}.$$

Очевидно, что

$$R(a,b) = \min_i \left\{\|h_i\|_\infty \colon (p_i x^i + h_i)a = b\right\} = \\ = \min_i \frac{|b - (a, p_i x^i)|}{\sum_j |a_j|}.$$

где h_i — i-ая строка матрицы H.

Будем искать разделяющую гиперплоскость, для которой радиус устойчивости максимален

$$\sup_{\substack{a,b:\\ Xa \leqslant bp}} R(a,b) = \sup_{\substack{a,b:\\ Xa \leqslant bp}} \min_i \frac{b - (a, p_i x^i)}{\sum_j |a_j|}. \qquad (4)$$

В [1] показано, что задача определения наиболее устойчивого решения системы линейных неравенств по минимаксному критерию может быть сведена к задаче линейного программирования.

Теорема 5. *Если* $u^*, t^* \in \mathbb{R}, s^*, r^* \mathbb{R}^n$ *— решение задачи*

$$\max u;$$
$$0 \leqslant u \leqslant t p_i - (p_i x^i, s), \quad i = 1, \ldots, m;$$
$$-r_j \leqslant s_j \leqslant r_j, \quad j = 1, \ldots, n; \quad \sum_j r_j = 1,$$

то $a^* = s^*$, $b^* = t^*$ *— решение задачи* (4).

Выводы

Предложены методы построения линейного решающего правила в задаче классификации двух классов по прецедентам. Если представители классов в обучающей выборке не являются линейно разделимыми, выполняется минимальная коррекция информационной матрицы, в результате которой существует разделяющая гиперплоскость. Полученная гиперплоскость используется в качестве решающего правила в исходной задаче.

Дополнительно к минимизации наибольшего отклонения по значению признака задается минимальная ширина разделяющей полосы. Предлагается метод определения коэффициентов разделяющих функций $F_1(x)$, $F_2(x)$, которые определяют полосу требуемой ширины и правильно классифицируют объекты информационной матрицы обучающей выборки, ближайшей к заданной в смысле l_∞-метрики.

Если представители классов в обучающей выборке линейно разделимы, определяется наиболее устойчивая по параметрам разделяющая гиперплоскость, т. е. такая гиперплоскость, которая остается разделяющей при наибольшем возмущении информационной матрицы.

Во всех рассмотренных случаях задача определения параметров представляет собой задачу линейного программирования.

Предложенный подход можно распространить на задачи с дополнительными условиями: выделено подмножество точно измеренных признаков, не подлежащих коррекции (задача коррекции с фиксированными столбцами таблицы обучения); в обучающей выборке задано подмножество достоверных объектов (задача коррекции с фиксированными строками).

Литература

[1] *Муравьева О. В.* Возмущение и коррекция систем линейных неравенств // Управление большими системами, 2010. — Вып. 28. — С. 40–57.

Регуляризация обучения распознаванию образов по частично классифицированной обучающей совокупности*

Середин О. С.
oseredin@yandex.ru
Тула, Тульский государственный университет

Для регуляризации решающих правил распознавания образов в условиях недостаточного обучающего материала предлагается использовать неклассифицированные объекты, образующие базисную совокупность. Показано, что переход к пространству проекционных признаков обеспечивает регуляризацию непосредственно в процессе обучения.

Введение

Одной из актуальных проблем современной теории распознавания образов является необходимость обучения в условиях ограниченных обучающих выборок. Часто оказывается, что нехватка обучающего материала выражается не столько в наличии достаточного количества известных до этапа обучения объектов, сколько в отсутствии именно классифицированных объектов. Обычно это связано с тем, что технология регистрации характеристик объектов достаточно хорошо развита (сканирование отпечатков пальцев, секвенирование белков, фоторегистрация биологических объектов), а привлечение экспертов по тем или иным причинам затруднительно. Решающие правила распознавания, полученные опираясь на небольшое число классифицированных объектов распознавания, не обладают достаточными экстраполяционными свойствами. В такой ситуации оправдано желание использовать на этапе обучения так же и неклассифицированные объекты. Включение неклассифицированных объектов в этап обучения можно осуществить за счет эксплуатации идеи базисной совокупности [4, 5]. Особенно изящные результаты получаются при беспризнаковой постановке задачи обучения распознавания образов.

Нам кажется естественным предположить, что распределение объектов в признаковом пространстве имеет не одинаковую вытянутость в разных направлениях. Этот факт должна отражать базисная совокупность, и, по возможности, обучающая выборка. Нам представляется закономерной гипотеза, что множество объектов одного класса приблизительно одинаково протяженно во всех направлениях. Такая гипотеза имеет следующее математическое выражение: если для некоторого числа объектов одного класса собрать матрицу их попарных скалярных произведений, то все собственные числа такой матрицы будут примерно одинаковыми. Предположим теперь, что рассматривается совокупность, содержащая объекты как первого, так и второго класса. Допустим, что мы не знаем, какие именно объекты в этой совокупности принадлежат первому и второму классу, но знаем, что классы линейно разделимы. Заметим, что базисная совокупность является именно такой совокупностью объектов двух классов. Поскольку согласно нашей гипотезе каждый класс образует область, близкую к сфере, то оба класса вместе должны образовать область, вытянутую именно в том направлении, в котором классы разнесены друг от друга. Если построить матрицу попарных взаимных отношений составляющих ее объектов, то ее собственные числа будут существенно различны по величине. Собственный вектор, соответствующий максимальному собственному числу, как раз укажет это направление.

Получается, что при принятых предположениях естественно в процессе обучения отдавать предпочтение разделяющим гиперплоскостям, почти ортогональным оси инерции базисной совокупности объектов. Поэтому представляется разумным исключить из рассмотрения разделяющие гиперплоскости, ориентированные вдоль базисной совокупности, даже если максимальный зазор между объектами первого и второго классов имеет такую ориентацию, и предпочитать «поперечные» гиперплоскости. Предпочтение, отдаваемое решающим правилам, направляющий вектор которых близок к оси инерции базисной можно использовать как метод борьбы с проклятием малой выборки (метод стабилизации или регуляризации решающего правила распознавания), заключающийся в привлечении дополнительной информации, возможно априорной, не отраженной в обучающей выборке при построении решающего правила.

Алгоритмически учет главного направления вытянутости выборки можно выполнить, применив стандартный метод главных компонент и регуляризовав процедуру обучения например дополнительным штрафом на отклонение направляющего вектора разделяющей гиперплоскости от направления собственного вектора, соответствующего главному собственному числу. Такая методика описана, например, в [6]. Однако, оказывается, что нет необходимости специально искать направление близкое к оси инерции базисной совокупности (например

Работа выполнена при финансовой поддержке РФФИ, проект № 09-07-00394.

используя метод главных компонент. В основной части работы будет показано, что обучение необходимо проводить в пространстве проекционных признаков. Таким образом, для векторного представления объектов мы получаем прямую регуляризацию. Для задач беспризнакового распознавания по методу опорных элементов [4, 2] использование описанного подхода побочно решает и проблему положительной определенности матрицы отношений между объектами.

Понятие базисной совокупности объектов распознавания

Обозначим множество всех рассматриваемых объектов $\omega \in \Omega$ и классифицированных на два подмножества

$$\Omega_1 = \{\omega \in \Omega :\ g(\omega) = 1\} \text{ и}$$
$$\Omega_{-1} = \{\omega \in \Omega :\ g(\omega) = -1\},$$

$\Omega_1 \cup \Omega_{-1} = \Omega,\ \ \Omega_1 \cap \Omega_{-1} = \varnothing$ некоторой неизвестной индикаторной функцией $g(\omega) = \pm 1$. В общем случае даже не предполагается возможности измерения на объектах каких бы то ни было наблюдаемых признаков $\boldsymbol{x}(\omega) = \big(x_1(\omega), \ldots, x_n(\omega)\big)$, которые позволили бы применять методы обучения, разработанные для векторных признаковых пространств. Вместо этого будем предполагать, что для любых двух объектов $\omega' \in \Omega$ и $\omega'' \in \Omega$ может быть измерена числовая характеристика их сходства $\mu(\omega', \omega'')$. В [3, 4] показано, что если такая парная характеристика обладает свойствами потенциальной функции [1], то множество объектов распознавания фактически может быть рассмотрено как линейное (гильбертово) пространство. Предполагается, что даже если элемент $\omega \in \Omega$ гильбертова пространства действительно существует $\omega \in \tilde{\Omega} \subset \Omega$ он не может быть представлен иначе как через своё скалярное произведение (ω, ω') с каким либо другим, реально существующим элементом $\omega' \in \tilde{\Omega} \subset \Omega$.

Если $\vartheta \in \Omega$ — некоторый элемент гильбертова пространства, в общем случае воображаемый, то действительнозначная линейная дискриминантная функция $d(\omega|\vartheta, b) = (\omega, \vartheta) + b$, где $b \in R$ — некоторая константа, может быть использована как решающее правило $\hat{g}(\omega)\colon\ \Omega \to \{1,\ -1\}$, позволяющая судить о скрытой принадлежности некоторого рассматриваемого объекта $\omega \in \Omega$, к первому или второму классу, независимо от того существует этот объект в реальности, или нет

$$d(\omega|\vartheta, b) = (\omega, \vartheta) + b \begin{cases} > 0 \ \to\ \hat{g}(\omega) = 1; \\ < 0 \ \to\ \hat{g}(\omega) = -1. \end{cases}$$

Пусть наблюдатель выбрал конечную совокупность реально существующих объектов

$$\Omega^0 = \{\omega_1^0, \ldots, \omega_n^0\} \subset \tilde{\Omega} \subset \Omega,$$

называемую базисной совокупностью. В общем случае не предполагается, что элементы базисной совокупности классифицированы, то есть она не рассматривается как обучающая выборка. Базисная совокупность будет играть роль конечного базиса в гильбертовом пространстве, который определяет в нем n-мерное подпространство $\Omega_n(\omega_1^0, \ldots, \omega_n^0) = \big\{\omega \in \Omega\colon \omega = \sum_{k=1}^n a_k \omega_k^0\big\} \subset \Omega$. Мы ограничимся рассмотрением только тех разделяющих гиперплоскостей, направляющие элементы которых могут быть выражены в виде линейных комбинаций объектов базисной совокупности

$$\vartheta(\boldsymbol{a}) = \sum_{k=1}^{n} a_k \omega_k^0, \quad \boldsymbol{a} \in R^n. \tag{1}$$

Соответствующее параметрическое семейство разделяющих гиперплоскостей полностью определяется скалярным произведением их элементов с объектами составляющими базисную совокупность: $(\omega, \vartheta(\mathbf{a})) + b \ = \sum_{k=1}^n a_k(\omega, \omega_k^0) + b \ = \ 0$ Отметим, что если $(\omega, \vartheta(\mathbf{a})) + b = 0$, то $(\omega, \omega_k^0) = 0$ для всех $\omega_k^0 \in \Omega^0$. Это означает, что если выбраны направляющие векторы в соответствии с (1), то мы ограничиваем наше рассмотрение только теми разделяющими гиперплоскостями, которые ортогональны подпространству, определяемому базисной совокупностью объектов. Каждому элементу гильбертова пространства $\omega \in \Omega$ ставится в соответствие целый набор его скалярных произведений, которые мы будем рассматривать как действительнозначный «вектор признаков»

$$\boldsymbol{x}(\omega) = \big(x_1(\omega) \cdots x_k(\omega)\big)^{\mathsf{T}} \in R^k.$$

Не предполагается, что базис $\Omega^0 = \{\omega_1^0, \ldots, \omega_n^0\}$ является полным в гильбертовом пространстве Ω, поэтому произвольный элемент $\omega \in \Omega$ не может быть представлен линейной комбинацией элементов базисной совокупности, но вектор признаков $\mathbf{x}(\omega)$ полностью определяет проекцию элемента $\omega \in \Omega$ на это подпространство. Мы пришли к параметрическому семейству двухклассовых решающих правил распознавания образов в гильбертовом пространстве, опирающихся на проекционные признаки объектов:

$$d\big(\boldsymbol{x}(\omega)|\mathbf{a}, b\big) =$$
$$= \boldsymbol{a}^{\mathsf{T}}\boldsymbol{x}(\omega) + b \begin{cases} > 0 \ \to\ \hat{g}(\omega) = 1; \\ < 0 \ \to\ \hat{g}(\omega) = -1, \end{cases}$$
$$\mathbf{a} \in R^n,\ b \in R,\ \omega \in \Omega,\ g(\omega) = \pm 1.$$

По своей структуре это семейство решающих правил полностью соответствует семейству линейных решающих правил в методе опорных векторов [9]. Таким образом, идея проекционных при-

знаков сводит, по крайней мере внешне, задачу беспризнакового распознавания образов в гильбертовом пространстве к классической задаче в обычном линейном пространстве действительнозначных признаков. Действительно, пусть наблюдателю представлена классифицированная обучающая выборка из объектов $\Omega^* = \{\omega_1, \ldots, \omega_N\} \subset \Omega$, $g_1 = g(\omega_1), \ldots, g_N = g(\omega_N)$, которая, в общем случае, не совпадает с базисной совокупностью $\Omega^0 = \{\omega_1^0, \ldots, \omega_n^0\}$. У наблюдателя нет другого способа увидеть эти объекты, «почувствовать» их, иначе как через скалярные произведения их с объектами базисной совокупности, что как раз эквивалентно вычислению их проекционных признаков

$$\begin{aligned}\boldsymbol{x}(\omega_j) &= (x_1(\omega_j) \cdots x_n(\omega_j))^{\intercal} = \\ &= \big((\omega_j, \omega_1^0) \cdots (\omega_j, \omega_n^0)\big)^{\intercal} \in R^n. \end{aligned} \tag{2}$$

Поскольку направляющий элемент разделяющей гиперплоскости определён как конечномерный вектор параметров, то такая задача, рассмотренная в базисном подпространстве , полностью совпадает с классической постановкой задачи распознавания образов путём поиска оптимальной разделяющей гиперплоскости. Легко показать, что максимальный зазор (или минимальный дефицит зазора при пересекающихся выборках классов) между классами обеспечивается выбором направляющего элемента $\vartheta(\boldsymbol{a}) \in \Omega$ и порога $b \in R$ минимизирующего следующий критерий с ограничениями:

$$\begin{cases} \|\vartheta(\boldsymbol{a})\|^2 + C \sum\limits_{j=1}^{N} \delta_j \to \min; \\ g_j \left(\boldsymbol{a}^{\intercal} \boldsymbol{x}(\omega_j) + b\right) \geqslant 1 - \delta_j; \\ \delta_j \geqslant 0, \quad j = 1, \ldots, N. \end{cases} \tag{3}$$

Однако, в отличие от классической постановки задачи Вапника, норма направляющего элемента искомой гиперплоскости может трактоваться, по крайней мере, двумя способами, либо как норма элемента гильбертова пространства $\|\vartheta(\boldsymbol{a})\|^2 = \big(\vartheta(\boldsymbol{a}), \vartheta(\boldsymbol{a})\big)$, $\vartheta(\boldsymbol{a}) \in \Omega$, либо как норма направляющего вектора в пространстве проекционных признаков $\|\vartheta(\boldsymbol{a})\|^2 = ||\boldsymbol{a}||^2 = \boldsymbol{a}^{\intercal}\boldsymbol{a}$. Эти два варианта нормы приводят к алгоритмам обучения и распознавания, существенно различающимся по своей структуре и рассматриваются в следующих разделах.

Следует отметить, что идея вычисления новых признаков объектов через отношения с фиксированными фиксированных (эталонными, признакообразующими) объектами упоминается в литературе. В частности в работах научной школы профессора Р. Дьюина [7, 8] рассматривалось сравнение различных классификаторов в пространстве эталонов (dissimilarity space). В качестве порожденных признаков в основном рассматривались либо евклидовы расстояния от объектов до эталонов либо различные функции от расстояний, например квадраты. Также в упомянутых работах предлагаются эвристики — как отбирать подмножество эталонных объектов. Мы же показываем теоретически, что использование в качестве вторичных признаков значений скалярных произведений между объектами может быть способом регуляризации решающего правила распознавания при наличии априорной информации о взаимном расположении распознаваемых классов.

Отсутствие априорных предпочтений о направляющем элементе разделяющей гиперплоскости в гильбертовом пространстве

Рассмотрим сначала «естественную» норму направляющего элемента разделяющей гиперплоскости, которая измеряется непосредственно в гильбертовом пространстве $\|\vartheta(\boldsymbol{a})\|^2 = \big(\vartheta(\boldsymbol{a}), \vartheta(\boldsymbol{a})\big)$. В этом случае процедура обучения сводится к минимизации нормы направляющего элемента, и, следовательно, все его направления одинаково предпочтительны. В соответствии с (1)

$$\big(\vartheta(\boldsymbol{a}), \vartheta(\boldsymbol{a})\big)^2 = \sum_{j=1}^{n} \sum_{k=1}^{n} (\omega_j^0, \omega_k^0) a_j a_k = \boldsymbol{a}^{\intercal} \boldsymbol{M} \boldsymbol{a},$$

где $\boldsymbol{M}$ — матрица $(n \times n)$, построенная на скалярных произведениях элементов базисной совокупности $\omega_1^0, \ldots, \omega_n^0$:

$$\boldsymbol{M} = \begin{pmatrix} (\omega_1^0, \omega_1^0) & \cdots & (\omega_1^0, \omega_n^0) \\ \vdots & \ddots & \vdots \\ (\omega_n^0, \omega_1^0) & \cdots & (\omega_n^0, \omega_n^0) \end{pmatrix}.$$

Тогда мы приходим к следующему критерию обучения:

$$\begin{cases} \boldsymbol{a}^{\intercal} \boldsymbol{M} \boldsymbol{a} + C \sum\limits_{j=1}^{N} \delta_j \to \min; \\ g_j \big(\boldsymbol{a}^{\intercal} \boldsymbol{x}(\omega_j) + b\big) \geqslant 1 - \delta_j; \\ \delta_j \geqslant 0, \quad j = 1, \ldots, N. \end{cases} \tag{4}$$

Двойственная задача оптимизации имеет вид:

$$\begin{cases} W(\lambda_1, \ldots, \lambda_N) = \sum\limits_{j=1}^{N} \lambda_j - \\ - \frac{1}{2} \sum\limits_{j=1}^{N} \sum\limits_{k=1}^{N} \big(g_j g_k \boldsymbol{x}^{\intercal}(\omega_j) \boldsymbol{M}^{-1} \boldsymbol{x}(\omega_k)\big)\ \lambda_j \lambda_k \to \max; \\ \sum\limits_{j=1}^{N} \lambda_j g_j = 0, \quad 0 \leqslant \lambda_j \leqslant \frac{1}{2} C, \quad j = 1, \ldots, N. \end{cases}$$

Связь прямой и двойственной задач:

$$\boldsymbol{a} = \boldsymbol{M}^{-1} \sum_{j=1}^{N} \lambda_j g_j \boldsymbol{x}(\omega_j).$$

Множители Лагранжа $\lambda_1 \geqslant 0, \ldots, \lambda_N \geqslant 0$, или точнее те из них, которые отличны от нуля

$\lambda_j > 0$, формируют решающее правило распознавания применимое в терминах проекционных признаков (2) к любому объекту $\omega \in \Omega$ и определяют опорные элементы обучающей выборки $\boldsymbol{x}(\omega_j) = \big((\omega_j,\omega_1^0)\cdots(\omega_j,\omega_n^0)\big)^{\intercal} \in R^n$, $\omega_j \in \Omega^* = \{\omega_1,\ldots,\omega_N\}$ (напомним здесь, что $\Omega^0 = \{\omega_1^0,\ldots,\omega_n^0\}$ — есть базисная совокупность):

$$d\left(\boldsymbol{x}(\omega)\right) = \sum_{j:\lambda_j>0} g_j\lambda_j\boldsymbol{x}^{\intercal}(\omega)\boldsymbol{M}^{-1}\boldsymbol{x}(\omega_j) + b\begin{cases} > 0 \;\to\; \hat{g}(\omega) = 1; \\ < 0 \;\to\; \hat{g}(\omega) = -1. \end{cases} \tag{5}$$

Константа b может быть найдена так же, как и для случая линейных признаковых пространств:

$$b = -\frac{\sum_{j:\,0\leqslant\lambda_j\leqslant\frac{1}{2}C}\lambda_j\boldsymbol{a}^{\intercal}\boldsymbol{x}(\omega_j) + \frac{1}{2}C\sum_{j:\,\lambda_j=\frac{1}{2}C} g_j}{\sum_{j:\,0\leqslant\lambda_j<\frac{1}{2}C}\lambda_j}.$$

В частном случае, когда обучающая выборка совпадает с базисной совокупностью, т. е. $N = n$ и $\Omega^* = \Omega^0 = \{\omega_1,\ldots,\omega_N\}$, $g_1 = g(\omega_1),\ldots,g_N = g(\omega_N)$, матрица скалярных произведений построенная на объектах обучающей выборки и будет играть роль $\boldsymbol{M}$:

$$\boldsymbol{M} = \begin{pmatrix} (\omega_1,\omega_1) & \cdots & (\omega_1,\omega_N) \\ \vdots & \ddots & \vdots \\ (\omega_N,\omega_1) & \cdots & (\omega_N,\omega_N) \end{pmatrix}.$$

Двойственная задача приобретает особенно простую форму:

$$\begin{cases} W(\lambda_1,\ldots,\lambda_N) = \sum\limits_{j=1}^{N}\lambda_j - \\ \qquad -\frac{1}{2}\sum\limits_{j=1}^{N}\sum\limits_{k=1}^{N}\big(g_jg_k(\omega_j,\omega_k)\big)\lambda_j\lambda_k \to \max; \\ \sum\limits_{j=1}^{N}\lambda_jg_j = 0, \quad 0\leqslant\lambda_j\leqslant\frac{1}{2}C, \quad j = 1,\ldots,N. \end{cases}$$

Особенно простым будет также вид решающего правила распознавания, так как в (5) мы имеем $\boldsymbol{x}^{\intercal}(\omega)\boldsymbol{M}^{-1}\boldsymbol{x}(\omega_j) = (\omega,\omega_j)$ для любого объекта $\omega\in\Omega$, и следовательно

$$d(\omega) = \sum_{j:\lambda_j>0} g_j\lambda_j(\omega,\omega_j) + b\begin{cases} > 0 \;\to\; \hat{g}(\omega) = 1; \\ < 0 \;\to\; \hat{g}(\omega) = -1. \end{cases}$$

Предпочтительная ориентация направляющего элемента разделяющей гиперплоскости вдоль осей инерции базисной совокупности

Другой способ, которым может быть задана норма искомого направляющего элемента — это $\|\boldsymbol{a}\|^2 = \boldsymbol{a}^{\intercal}\boldsymbol{a}$, т. е. скалярное произведение вектора коэффициентов его представления в пространстве проекционных признаков R^n. В этом случае матрица $\boldsymbol{M}$ в (4) есть единичная матрица и задача обучения (3) примет вид:

$$\begin{cases} \boldsymbol{a}^{\intercal}\boldsymbol{a} + C\sum\limits_{j=1}^{N}\delta_j \to \min; \\ g_j\big(\boldsymbol{a}^{\intercal}\boldsymbol{x}(\omega_j) + b\big) \geqslant 1 - \delta_j; \\ \delta_j \geqslant 0, \quad j = 1,\ldots,N, \end{cases} \tag{6}$$

и в двойственной форме:

$$\begin{cases} W(\lambda_1,\ldots,\lambda_N) = \sum\limits_{j=1}^{N}\lambda_j - \\ -\frac{1}{2}\sum\limits_{j=1}^{N}\sum\limits_{k=1}^{N}\left(g_jg_k\boldsymbol{x}^{\intercal}(\omega_j)\boldsymbol{x}(\omega_k)\right)\lambda_j\lambda_k \to \max; \\ \sum\limits_{j=1}^{N}\lambda_jg_j = 0, \quad 0\leqslant\lambda_j\leqslant\frac{1}{2}C, \quad j = 1,\ldots,N. \end{cases}$$

В соответствии с (5), оптимальное решающее правило будет иметь структуру

$$d\big(\boldsymbol{x}(\omega)\big) = \sum_{j:\,\lambda_j>0} g_j\lambda_j\boldsymbol{x}^{\intercal}(\omega)\boldsymbol{x}(\omega_j) + b\begin{cases} > 0 \;\to\; \hat{g}(\omega) = 1; \\ < 0 \;\to\; \hat{g}(\omega) = -1, \end{cases} \tag{7}$$

в которое проекционные векторы признаков нового объекта $\boldsymbol{x}(\omega) = \big((\omega,\omega_1^0)\cdots(\omega,\omega_1^0)\big)^{\intercal}$ и опорных объектов обучающей выборки $\boldsymbol{x}(\omega_j) = \big((\omega_j,\omega_1^0)\cdots(\omega_j,\omega_1^0)\big)^{\intercal}$ входят таким же способом, как и в методе опорных векторов.

Нетрудно убедиться, что если $\vartheta\in\Omega$ и $\omega\in\Omega$ два произвольных объекта гильбертова пространства Ω то расстояние от ω до его проекции на луч, проходящий через ϑ равно $(\omega,\omega) - \frac{(\omega,\vartheta)^2}{(\vartheta,\vartheta)}$. Мы собираемся продемонстрировать, что критерий $\boldsymbol{a}^{\intercal}\boldsymbol{a}\to\min$ среди всех направляющих элементов с одинаковым квадратом нормы $\big(\vartheta(\boldsymbol{a}),\vartheta(\boldsymbol{a})\big) = \boldsymbol{a}^{\intercal}\boldsymbol{M}\boldsymbol{a} = \text{const}$ эквивалентен критерию $\sum_{j=1}^{n}\big(\omega_j,\vartheta(\boldsymbol{a})\big)^2 \to \max$. Затем мы покажем, что обучение по критерию $\boldsymbol{a}^{\intercal}\boldsymbol{a}\to\min$ предпочитает те разделяющие гиперплоскости, направляющие элементы которых близки к главной оси инерции базисной совокупности $\Omega^0 = \{\omega_1^0,\ldots,\omega_n^0\}$.

Во-первых, убедимся, что минимизация критерия $\boldsymbol{a}^{\intercal}\boldsymbol{a}\to\min$ при ограничениях $\big(\vartheta(\boldsymbol{a}),\vartheta(\boldsymbol{a})\big) = \boldsymbol{a}^{\intercal}\boldsymbol{M}\boldsymbol{a} = \text{const}$ обеспечивается главным собственным вектором матрицы $\boldsymbol{M}$. Действительно, функция Лагранжа для такой задачи минимизации имеет вид

$$L'(\boldsymbol{a},\beta') = \frac{1}{2}\boldsymbol{a}^{\intercal}\boldsymbol{a} - \frac{1}{2\beta'}(\boldsymbol{a}^{\intercal}\boldsymbol{M}\boldsymbol{a} - \text{const}),$$

и уравнение

$$\nabla_{\boldsymbol{a}} L(\boldsymbol{a}, \beta') = \boldsymbol{a} - \frac{1}{\beta'}\boldsymbol{M}\boldsymbol{a} = \boldsymbol{0}$$

приводит к условию $\boldsymbol{M}\boldsymbol{a} = \beta'\,\boldsymbol{a}$, которое в свою очередь приводит к равенству $\boldsymbol{a}^{\intercal}\boldsymbol{M}\,\boldsymbol{a} = \beta'\,\boldsymbol{a}^{\intercal}\boldsymbol{a} = \text{const}$. Так как мы минимизируем $\boldsymbol{a}^{\intercal}\boldsymbol{a} \to \min$, то, следовательно, неизбежно максимизируем $\beta' \to \max$, а это предполагает, что последнее равенство обеспечивается максимальным собственным числом матрицы $\boldsymbol{M}$. С другой стороны, главный собственный вектор матрицы $\boldsymbol{M}$ удовлетворяет условию $\boldsymbol{a}^{\intercal}\boldsymbol{M}\boldsymbol{M}\,\boldsymbol{a} \to \max$, при ограничениях $\boldsymbol{a}^{\intercal}\boldsymbol{M}\,\boldsymbol{a} = \text{const}$. Чтобы показать это продифференцируем функцию Лагранжа по $\boldsymbol{a}$:

$$L''(\boldsymbol{a}, \beta'') = \frac{1}{2}\boldsymbol{a}^{\intercal}\boldsymbol{M}\boldsymbol{M}\,\boldsymbol{a} - \frac{1}{2}\beta''(\boldsymbol{a}^{\intercal}\boldsymbol{M}\boldsymbol{a} - \text{const}),$$

что приведет к равенству

$$\nabla_{\boldsymbol{a}} L(\boldsymbol{a}, \beta'') = \\ = \boldsymbol{M}\boldsymbol{M}\,\boldsymbol{a} - \beta''\boldsymbol{M}\boldsymbol{a} = \boldsymbol{M}(\boldsymbol{M}\,\boldsymbol{a} - \beta''\boldsymbol{a}) = \boldsymbol{0},$$

если $\boldsymbol{M}\,\boldsymbol{a} = \beta''\boldsymbol{a}$.

Таким образом, требование $\boldsymbol{a}^{\intercal}\boldsymbol{M}\boldsymbol{M}\,\boldsymbol{a} \to \max$ эквивалентно требованию $\beta''\boldsymbol{a}^{\intercal}\boldsymbol{M}\,\boldsymbol{a} \to \max$, которое с учётом ограничения $\boldsymbol{a}^{\intercal}\boldsymbol{M}\,\boldsymbol{a} = \text{const}$ обеспечивается наибольшим собственным числом матрицы $\boldsymbol{M}$.

Итак, $\boldsymbol{a}^{\intercal}\boldsymbol{a} \to \min$ при ограничении $\big(\vartheta(\boldsymbol{a}), \vartheta(\boldsymbol{a})\big) = \boldsymbol{a}^{\intercal}\boldsymbol{M}\boldsymbol{a} = \text{const}$ эквивалентно $\boldsymbol{a}^{\intercal}\boldsymbol{M}\boldsymbol{M}\,\boldsymbol{a} \to \max$. Здесь $\boldsymbol{a}\,\boldsymbol{M} = \boldsymbol{a}^{\intercal}(\boldsymbol{x}_1 \cdots \boldsymbol{x}_N)$, где $\boldsymbol{x}_j$ — это столбец симметрической матрицы $\boldsymbol{M}$ и $\boldsymbol{a}^{\intercal}\boldsymbol{M} = (\boldsymbol{M}\,\boldsymbol{a})^{\intercal}$, так что $\boldsymbol{a}^{\intercal}\boldsymbol{M}\boldsymbol{M}\,\boldsymbol{a} = \sum_{j=1}^{n} \boldsymbol{a}^{\intercal}\boldsymbol{x}_j\boldsymbol{x}_j^{\intercal}\boldsymbol{a} = \sum_{j=1}^{n} (\boldsymbol{x}_j^{\intercal}\boldsymbol{a})^2$.

В свою очередь $\boldsymbol{x}_j^{\intercal}\boldsymbol{a} = \sum_{i=1}^{n} a_i(\omega_j, \omega_i) = (\omega_j,\ \sum_{i=1}^{n} a_i\omega_i) = \big(\omega_j, \vartheta(\boldsymbol{a})\big)$, и, таким образом, $\boldsymbol{a}^{\intercal}\boldsymbol{M}\boldsymbol{M}\,\boldsymbol{a} = \sum_{j=1}^{n} \big(\omega_j, \vartheta(\boldsymbol{a})\big)^2$.

Это означает, что если $\boldsymbol{a}^{\intercal}\boldsymbol{a} \to \min$ при ограничении $\big(\vartheta(\boldsymbol{a}), \vartheta(\boldsymbol{a})\big) = \boldsymbol{a}^{\intercal}\boldsymbol{M}\boldsymbol{a} = \text{const}$, то

$$\sum_{j=1}^{n} \big(\omega_j, \vartheta(\boldsymbol{a})\big)^2 \to \max,$$

и направляющий элемент разделяющей гиперплоскости $\vartheta(\boldsymbol{a})$ предпочтительно выбирать близким к главной оси инерции базисной выборки. Именно это мы и хотели показать.

Итак, обучение по критерию (6), то есть без всяких предпочтений в пространстве проекционных признаков, эквивалентно выражению предпочтений в исходном гильбертовом пространстве, связанных со склонностью направляющего элемента быть близким к главной оси инерции базовой совокупности объектов. Как результат, разделяющая гиперплоскость стремится быть ортогональной этой оси.

Выводы

Для регуляризации решающих правил распознавания в условиях недостаточного обучающего материала предлагается использовать неклассифицированные объекты, образующие базисную совокупность. Показано, что переход к пространству проекционных признаков обеспечивает регуляризацию непосредственно в процессе обучения. Такой подход можно использовать как в случае беспризнаковой концепции, так и оставаясь в рамках классического «признакового» распознавания. Если имеется априорное предположение о вытянутости генеральной совокупности вдоль распределения классов, использование проекционных признаков может повысить экстраполирующие свойства решающего правила.

Литература

[1] *Айзерман М., Браверман Э., Розоноэр Л.* Метод потенциальных функций в теории обучения машин. — М.: Наука, 1970. — 386 с.

[2] *Mottl V. V., Seredin O. S., Dvoenko S. D., Kulikowski C. A, Muchnik I. B.* Featureless pattern recognition in an imaginary Hilbert space // Proceedings of 16th International Conference Pattern Recognition, ICPR-2002, Quebec City, Canada. — 2002. — V. 2. — Pp. 88–91.

[3] *Моттль В. В.* Метрические пространства, допускающие введение линейных операций и скалярного произведения // Доклады академии наук. — 2003. — Т. 388, № 3. — С. 1–4.

[4] *Середин О. С.* Методы и алгоритмы беспризнакового распознавания образов // Дисс. на соискание звания канд. наук. -М., 2001.

[5] *Середин О. С.* Экспериментальное исследование априорных предпочтений о решающем правиле в гильбертовом пространстве объектов распознавания // Труды VI международной конференции «Распознавание образов и анализ изображений: новые информационные технологии». — 2002. — Т. 2. — С. 511–515.

[6] *Середин О. С., Моттль В. В.* Отбор информативных признаков при обучении распознаванию образов с упорядоченными признаками // Таврический вестник информатики и математики. — 2008. — № 2. — С. 180–185.

[7] *Pekalska E., Duin R.P.W.* Automatic pattern recognition by similarity representations // Electronic Letters. — 2001. — V. 37, N. 3. — Pp. 159–160.

[8] *Pekalska E., Duin R.P.W., Paclik P.* Prototype selection for dissimilarity-based classifiers // Pattern Recognition. — 2006. — V. 39. — Pp. 189–208.

[9] *Vapnik V.* Statistical Learning Theory. — NY.: J. Wiley, 1998. — 768 p.

Построение алгоритма обучения распознаванию образов в режиме реального времени на основе вероятностного подхода к методу опорных векторов

Турков П. А., Красоткина О. В.
pavel-turkov@gmail.com
Тула, Тульский государственный университет

В последнее время в области распознавания образов появилось достаточно много приложений, в которых обучающая совокупность объектов предъявляется для анализа не вся сразу, а постепенно в режиме реального времени. В данной статье с теоретико-вероятностных позиций изложен алгоритм обучения в режиме реального времени, который не требует хранения всего множества объектов для получения актуального решающего правила. Время обучения предложенного алгоритма остается постоянным вне зависимости от числа объектов поступивших на обучение, что позволяет применять его не только для построения решающего правила в режиме реального времени, но и для задач, характеризующихся огромным объемом обучающих данных. В статье приводится результаты экспериментального исследования предложенного алгоритма на реальных и тестовых данных.

Разработанный несколько десятилетий назад В. Н. Вапником метод опорных векторов (Support Vector Machine, SVM) [1] и в нынешнее время не теряет своей популярности в области анализа данных не только из-за простоты и понятности основной идеи, но и за счет скорости и точности нахождения параметров классификатора. Также следует отметить его практически неограниченную способность к модификации и расширяемости путем выбора подходящей потенциальной функции. Однако, существует ряд задач, условия в которых не позволяют эффективно применить классический метод опорных векторов. Поэтому, в связи со своими выдающимися свойствами часто этот метод используется для создания на его основе алгоритмов распознавания, более или менее сохраняющих преимущества базового метода в необычных для него условиях. К таким задачам относится задача обучения распознаванию образов в режиме реального времени, когда объекты обучения поступают на вход алгоритма в течение некоторого времени по одному или порциями.

Так, пусть после завершения построения решающего правила по заданному обучающему множеству в распоряжение разработчика поступают дополнительные объекты с известной для них скрытой характеристикой, которые было бы желательно использовать для дополнительного обучения, т. е. коррекции уже созданного классификатора. Использование существующих методов обучения распознавания образов, разработанных для анализа всей обучающей последовательности, потребует обучения на совокупности, состоящей из исходного и полученного обучающих наборов, что часто оказывается неприемлемым по отношению к вычислительным и временным ресурсам, имеющихся в распоряжении исследователя. В таком случае требуется метод обучения, который на каждой следующей итерации корректировал бы уже найденные параметры классификатора на основании только полученных объектов.

В работе [2] представлен алгоритм обучения, разработанный в предположении, что искомое решающее правило может изменяться с приходом новых объектов. Алгоритм основан на методе опорных векторов и процедуре динамического программирования. Степень изменения решающего правила контролируется специальным параметром нестационарности. Достоинством указанного алгоритма является то, что для построения решающего правила при поступлении новой порции объектов необходимо лишь знание параметров классификатора в данный момент и не требуется хранение обучающего множества. Недостаток метода состоит в том, что возможность численной реализации процедуры динамического программирования основана на предположении о существовании параметрического семейства, которому на каждом шаге процедуры принадлежат функции Беллмана. В данной задаче такого параметрического семейства не существует, и приходится аппроксимировать неквадратичные функции Беллмана их квадратичными аналогами, что неизбежно приводит к ухудшению качества распознавания.

На основании анализа литературы, можно отметить, что все существующие на сегодняшний момент методы обучения в режиме реального времени являются в той или иной степени эвристическими, причем, конкретный набор эвристик определяется спецификой решаемой задачи. В настоящее время в литературе известны три общих подхода к построению таких алгоритмов: алгоритмы, основанные на отборе объектов, алгоритмы основанные на взвешивании объектов и алгоритмы, основанные на слиянии и отборе классификаторов. Целью алгоритмов, использующих отбор объектов является селекция прототипов, релевантных для решающего правила в данный момент времени. Как правило, это реализуется с помощью техно-

логии скользящего окна, когда решающее правило в данный момент времени строится только на основании объектов, полученных в моменты времени, непосредственно предшествующих данному. Примерами таких алгоритмов может являться семейство алгоритмов FLORA [3]. Алгоритмы, основанные на взвешивании объектов [4], поступающих в разные моменты времени, эксплуатируют способность некоторых методов, таких как, например, метод опорных векторов, присваивать в процессе обучения веса различным объектам. Как правило, веса объектам присваиваются в зависимости от времени, прошедшего с момента поступления объектов. При обучении распознаванию образов в режиме реального времени с использованием технологии комбинирования классификаторов [6], [5] искомое решающее правило строится как голосование или взвешенное голосование классификаторов полученных для различных условий.

В данной работе предлагается алгоритм обучения распознаванию образов в реальном времени, основанный на байесовском подходе и являющийся в некоторой степени обобщением классического метода опорных векторов на случай, когда объекты поступают на вход алгоритма во времени. Алгоритм обучения, представленный в работе, обладает линейной вычислительной сложностью относительно длины обучающей последовательности. Такой метод обучения будем называть *методом обучения распознаванию в режиме реального времени*, соответственно, разработка такого метода и составляет тему настоящей статьи.

Вероятностная постановка задачи обучения распознаванию образов по методу опорных векторов

Рассмотрим задачу классификации объектов генеральной совокупности Ω. Каждый объект $\omega \in \Omega$ описывается набором действительнозначных признаков $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, т. е. представлен точкой в признаковом пространстве $X \subseteq R^n$. Ограничимся случаем бинарной классификации, тогда скрытая принадлежность объекта к одному из двух классов определяется меткой класса $y \in \{-1; +1\}$. Будем строить линейный пороговый классификатор $g(\mathbf{x}) = \operatorname{sign}(\mathbf{a}^\intercal \mathbf{x} + b)$, используя концепцию обучения с учителем: основываясь на ограниченном множестве объектов, для которых известна скрытая характеристика $(\mathbf{x}_i, y_i)_{i=1}^N$. Такое множество в дальнейшем будем называть обучающей выборкой. В [7] предложена вероятностная модель генеральной совокупности объектов, одним из ключевых моментов которой является несобственное параметрическое распределение

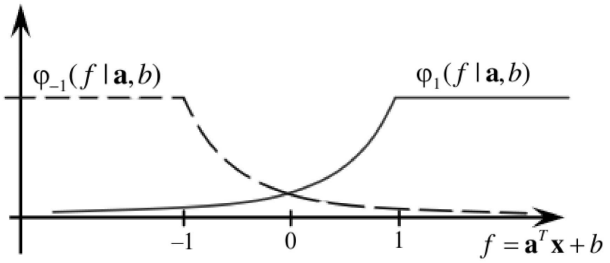


Рис. 1. Значения плотностей распределения (1) вдоль направляющего вектора гиперплоскости.

двух классов (рис. 1):

$$\varphi_y(\mathbf{x} \mid \mathbf{a}, b, y; c) \propto \propto \begin{cases} \text{const}, & yf(\mathbf{x}, \mathbf{a}, b) \geqslant 1; \\ e^{-c\left(1 - yf(\mathbf{x}, \mathbf{a}, b)\right)}, & yf(\mathbf{x}, \mathbf{a}, b) < 1, \end{cases} \quad (1)$$

определяемое объективно существующей гиперплоскостью $f(\mathbf{x}, \mathbf{a}, b) = \mathbf{a}^\intercal \mathbf{x} + b = 0$ с неизвестным направляющим вектором признаков $\mathbf{a} = (a_1, \ldots, a_n)$, имеющим априорное распределение $\Psi(\mathbf{a}, b|\sigma^2) \propto \propto \exp\left(-\frac{1}{2\sigma^2}\mathbf{a}^\intercal \mathbf{a}\right)$, и параметром положения b, который обладает несобственным равномерным распределением. Переменная штрафа c является параметром настройки модели. Применение принципа максимума апостериорной вероятности для оценивания параметров $(\mathbf{a}, b)$ гиперплоскости при таких предположениях приводит к широко известному критерию опорных векторов

$$(\hat{\mathbf{a}}, \hat{b}) = \arg\min_{\mathbf{a}, b} \left[\mathbf{a}^\intercal \mathbf{a} + C \sum_{i=1}^{N} \delta_i\right]; \quad \begin{cases} y_i(\mathbf{a}^\intercal \mathbf{x}_i + b) \geqslant 1 - \delta_i; \\ \delta_i \geqslant 0, \quad i = 1, \ldots, N, \end{cases} \quad (2)$$

который был сформулирован В. Н. Вапником с сугубо детерминистических позиций. При этом значение коэффициента штрафа C определяется выражением $2c\sigma^2$ [2].

Далее, для поставленной выше задачи классификации сформулируем следующее дополнительное условие: пусть объекты для обучения классификатора поступают группами или же объем обучающего множества столь велик, что обработать сразу все объекты технически не представляется возможным. Для определения параметров решающего правила классическим методом опорных векторов в условиях появления новых объектов обучения необходимо хранить информацию о всей имеющейся обучающей совокупности. Невозможность непосредственного применения метода опорных векторов в режиме реального времени обусловлена характером несобственных априорных распределений классов, так как именно их кусочность

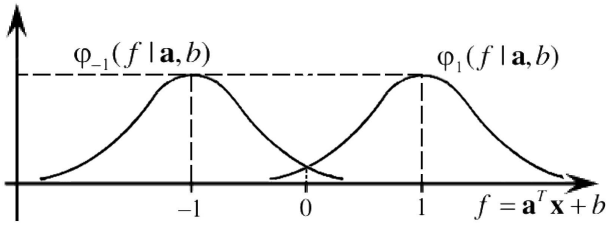


Рис. 2. Значения плотностей распределения (3) вдоль направляющего вектора гиперплоскости.

приводит к появлению в результирующем критерии ограничений и, собственно, к понятию опорных объектов. Добавление новых объектов в обучающую совокупность может приводить к изменению множества опорных векторов вследствие того, что часть старых элементов будет исключена из этого множества, в то время как другие объекты, считавшиеся до этого не опорными, станут ими на основании только что полученных данных. Таким образом, мы не можем «забыть» все зарегистрированные объекты, оставив лишь те, которые в текущий момент являются опорными и определяют решающее правило.

Вероятностная постановка задачи обучения распознаванию образов в режиме реального времени

Для построения задачи распознавания в режиме реального времени выберем априорные распределения для объектов классов следующего вида (см. рис. 2):

$$\varphi_y(\mathbf{x}\,|\,\mathbf{a},b,y,\sigma^2) \propto \exp\left(-\frac{\bigl(1-yf(\mathbf{x},\mathbf{a},b)\bigr)^2}{2\sigma_\varphi^2}\right), \quad (3)$$

которые также являются несобственными. Функция $f(\mathbf{x},\mathbf{a},b)$, как и ранее, описывает разделяющую гиперплоскость $\mathbf{a}^\intercal\mathbf{x}+b=0$. Отсюда следует, что объекты генеральной совокупности предполагаются нормально распределенными вдоль направляющего вектора разделяющей гиперплоскости и равномерно распределенными вдоль самой гиперплоскости. Направляющий вектор $\mathbf{a}$ будем как и ранее считать нормально распределенным $\Psi(\mathbf{a},b|\sigma^2) \propto \propto \exp\bigl(-\frac{1}{2\sigma^2}\mathbf{a}^\intercal\mathbf{a}\bigr)$, а параметр положения b - равномерно. Тогда принцип максимума апостериорной вероятности для нахождения оценки параметров разделяющей гиперплоскости приводит к критерию следующего вида:

$$(\hat{\mathbf{a}},\hat{b}) = \arg\min_{\mathbf{a},b}\left[\frac{1}{2}\mathbf{a}^\intercal\mathbf{a}+\frac{C}{2}\sum_{i=1}^{N}\delta_i^2\right]; \quad (4)$$
$$\delta_i = 1-y_i(\mathbf{a}^\intercal\mathbf{x}_i+b), \quad i=1,\ldots,N,$$

где $C=\sigma^2/\sigma_\varphi^2$. Задача оптимизации (4) в отличие от (27) не содержит ограничений, поэтому применение необходимого условия экстремума позволяет определить оптимальные значения параметров разделяющей гиперплоскости согласно выражениям:

$$\mathbf{a} = \bigl[\tfrac{I}{C}+N\mathsf{E}_N(\mathbf{x}\mathbf{x}^\intercal)+\tfrac{1}{n}N^2\mathsf{E}_N(\mathbf{x})\mathsf{E}_N(\mathbf{x}^\intercal)\bigr]^{-1}\times$$
$$\times\bigl(N\mathsf{E}_N(y\mathbf{x})+\tfrac{1}{n}N^2\mathsf{E}_N(y)\mathsf{E}_N(\mathbf{x})\bigr); \quad (5)$$
$$b = \tfrac{1}{n}N\mathsf{E}_N(y)-\tfrac{\mathbf{a}^\intercal}{n}N\mathsf{E}_N(\mathbf{x}),$$

где:

$$\mathsf{E}_N(\mathbf{x}\mathbf{x}^\intercal)=\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^\intercal, \quad \mathsf{E}_N(\mathbf{x})=\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i; \quad (6)$$
$$\mathsf{E}_N(y)=\frac{1}{N}\sum_{i=1}^{N}y_i, \quad \mathsf{E}_N(y\mathbf{x})=\frac{1}{N}\sum_{i=1}^{N}y_i\mathbf{x}_i.$$

Пусть в обучающую совокупность, состоящую из N объектов, добавлены еще M. Тогда, оценки параметров решающего правила принимают вид:

$$\mathbf{a} = \Bigl[\tfrac{I}{C}+N\mathsf{E}_N(\mathbf{x}\mathbf{x}^\intercal)+M\mathsf{E}_M(\mathbf{x}\mathbf{x}^\intercal)+$$
$$+\tfrac{1}{n}N^2\mathsf{E}_N(\mathbf{x})\mathsf{E}_N(\mathbf{x}^\intercal)+\tfrac{1}{n}M^2\mathsf{E}_M(\mathbf{x})\mathsf{E}_M(\mathbf{x}^\intercal)\Bigr]^{-1}\times$$
$$\times\Bigl(N\mathsf{E}_N(y\mathbf{x})+M\mathsf{E}_M(y\mathbf{x})+ \quad (7)$$
$$+\tfrac{1}{n}N^2\mathsf{E}_N(y)\mathsf{E}_N(\mathbf{x})+\tfrac{1}{n}M^2\mathsf{E}_M(y)\mathsf{E}_M(\mathbf{x})\Bigr);$$
$$b = \tfrac{1}{n}\bigl[N\mathsf{E}_N(y)+M\mathsf{E}_M(y)-$$
$$-\mathbf{a}^\intercal N\mathsf{E}_N(\mathbf{x})-\mathbf{a}^\intercal M\mathsf{E}_M(\mathbf{x})\bigr].$$

Так как равенства (27) зависят не от конкретных векторов признаков, а от средних по всей выборке (6), то хранение математических ожиданий $\mathsf{E}_N(\mathbf{x}\mathbf{x}^\intercal)$, $\mathsf{E}_N(\mathbf{x})$, $\mathsf{E}_N(y)$, $\mathsf{E}_N(y\mathbf{x})$, при добавлении в нее новой группы объектов (27) позволит скорректировать решающее правило, не обучаясь заново на всех когда-либо полученных объектах. Максимальный размер хранимой матрицы в этом случае будет составлять $n\times n$, вместо $n\times N$, соответственно, чем больше поступит объектов по сравнению с количеством измеренных на них признаков, тем больше будет выигрыш при использовании предложенного метода.

Экспериментальное исследование предложенного алгоритма

Первоначально для исследования предложенного алгоритма были применены искусственно сгенерированные тестовые выборки объектов двух равновероятных классов, признаковые описания которых представлены многомерным нормальным распределением с одинаковыми дисперсиями. Размер множества обучения составлял 8000 объектов, описываемых 30 признаками, контрольная выборка содержала 10 000 объектов. Для сравнения качества распознавания была использована реализация классического метода опорных векторов из статистического пакета R [9], в котором осуществлялось

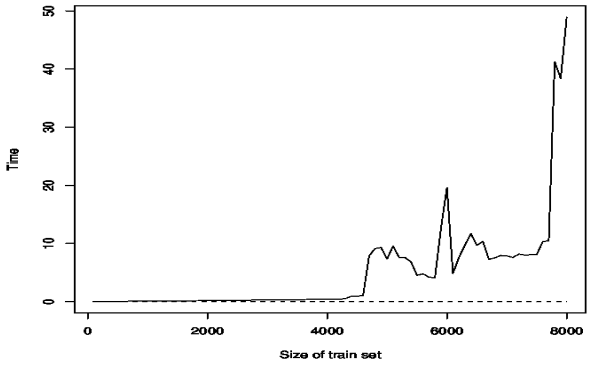


Рис. 3. Время обучения классификаторов. Сплошной линией обозначена зависимость в случае метода опорных векторов, пунктирной — для разработанного алгоритма обучения в режиме реального времени.

моделирование эксперимента. В ходе испытаний обучающая совокупность поступала на вход обоих алгоритмов блоками, по 100 объектов в каждом, по одному блоку данных за итерацию. При этом на каждой итерации производилось обучение одного классификатора на всех объектах, полученных к этому моменту, по методу SVM, в то время как другой классификатор корректировался в соответствии с алгоритмом обучения в режиме реального времени исключительно на данных, полученных на текущей итерации. На рис. 3 построен график зависимости времени, необходимого для обучения классификаторов, от количества объектов в обучающей совокупности.

Чтобы проверить качество работы разработанного метода применительно к объектам реального мира, использовались данные медицинских исследований BUPA в области печеночных заболеваний [8], содержащие информацию о 345 испытуемых по 6 признакам: первые 5 признаков соответствуют анализам крови, последний признак показывает количество регулярно принимаемого пациентом алкоголя. Принадлежность к одному из двух классов соответствует наличию или отсутствию заболеваний у подопытного. Для обучения использовались 145 объектов, проверка качества осуществлялась на остальных 200 объектах.

В таблице приведена доля ошибочно классифицированных объектов на контрольной выборке (для сгенерированных данных представлены усредненные результаты после проведения серии из 10 испытаний на момент окончания последней итерации). Как следует из этих результатов, разработанный алгоритм не уступает по качеству классическому методу опорных векторов, кроме того, время работы алгоритма обучения в режиме реального времени значительно меньше, чем аналогичный параметр для метода опорных векторов.

	Искусственные данные, %	Данные BUPA, %
SVM	0,405	32,5
Online-алритм	0,335	28,5

Выводы

В ходе работы предложен алгоритм обучения в режиме реального времени на основе вероятностного подхода к задаче распознаванию образов. По качеству работы созданный алгоритм сравним с классическим методом опорных векторов при меньших по сравнению с ним временных затратах на обучение.

Литература

[1] *Vapnik V. N.* The nature of statistical learning theory. — Springer, New York, 1995.

[2] *Красоткина О. В., Моттль В. В., Турков П. А.* Байесовский подход к задаче обучения распознаванию образов в нестационарной генеральной совокупности // Международная конференция ИОИ-8, 2010.

[3] *Widmer G., Kubat M.* Learning in the presence of concept drift and hidden contexts // Machine Learning. — 1996. — V. 23, N. 1. — Pp. 69–101.

[4] *Klinkenberg R.* Learning drifting concepts example selection vs. example weiting // Intellegent data analysis, Special Issue on Incremental Learning Systems Capable of Dealing with Concept Drift. — 2004. — V. 8, N. 3.

[5] *Kuncheva L.* Classifier Ensembles for Changing Environments // Proc. 5th Int. Workshop on Multiple Classfier Systems, Cagliari, Italy, Springer-Verlag, LNCS, 2004. — Pp. 1–15.

[6] *Policar R., Muhlbaier M.* An Ensemble Approach for Incremental Learning in Nonstationary Environments // MCS 2007, LNCS 4472, Springer-Verlag Berlin Heidelberg, 2007. — Pp. 490-–500

[7] *Татарчук А. И., Сулимова В. В., Моттль В. В., Уиндридж Д.* Метод релевантных потенциальных функций для селективного комбинирования разнородной информации при обучении распознаванию образов на основе байесовского подхода. — Всероссийская конференция ММРО-14.М.: МАКС Пресс, 2009. — С. 188—191.

[8] `https://archive.ics.uci.edu/ml/datasets/Liver\%20Disorders` — Liver Disorders Data Set. — 1990.

[9] *R Development Core Team* R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, 2009, URL http://www.R-project.org.

Методы интеллектуальной обработки информации на базе алгоритмов стохастической аппроксимации*

Дорофеюк А. А., Бауман Е. В., Дорофеюк Ю. А.
daa2@mail.ru
Москва, Институт проблем управления им. В. А. Трапезникова РАН

Рассмотрена задача размытого структурно-классификационного анализа, исследован вид оптимальной классификации, предложен общий итерационный алгоритм решения этой задачи. При конструировании и исследовании его сходимости использовались теория и процедуры стохастической аппроксимации.

Структурно-классификационный анализ данных часто приходится проводить для бесконечной выборки объектов. В таком случае актуальной является задача разработки рекуррентных (итерационных) алгоритмов типа стохастической аппроксимации, в которых решающие правила (разделяющие функции) последовательно пересчитываются при появлении очередного объекта выборки.

На базе вариационного подхода разработан и в достаточно общих предположениях исследован алгоритм решения задачи структурно-классификационного анализа данных. Исследование проводится для размытой постановки задачи структуризации, обобщающей многие частные постановки задач структурно-классификационного анализа. Для оценки качества размытой классификации используется широкий класс выпуклых функционалов, включающий значительную часть известных критериев качества классификации точек евклидова пространства, функционалы в неметрических шкалах и др. В том числе в него входят как частные случаи — функционал средневзвешенной дисперсии точек евклидова пространства; функционалы экстремальной группировки параметров; функционал диагонализации матрицы связи; функционал кусочно-линейной аппроксимации сложных зависимостей; функционалы классификации в бинарных, номинальных и ранговых шкалах.

В качестве наглядного примера использования процедур типа стохастической аппроксимации в задачах структурно-классификационного анализа в работе рассматривается задача классификации точек евклидова пространства, решаемая методом обобщённого среднего для критерия, зависящего от ненормированных моментов классов.

Постановка задачи

Пусть необходимо определить классификационную структуру произвольного множества X с заданной на нём вероятностной мерой (законом распределения вероятностей) $P(A)$, $A \in X$. Для случая конечного числа исследуемых объектов, т. е. когда $X = x_1, \ldots, x_n$, в качестве оценки $P(A)$ обычно используется $P(A) = |A|/n$, где $|A|$ — число точек множества A. В зависимости от конкретной постановки задачи классифицируемое множество может иметь различную природу (например, это может быть множество точек евклидова пространства, либо множество с заданной мерой близости между точками, либо некоторый набор параметров, описывающих исследуемое множество объектов, и др.).

Далее рассматривается размытая постановка задачи структурно-классификационного анализа. В дальнейшем для фильтрации грубых ошибок наблюдений и компенсации неправильного выбора числа классов (заниженном по отношению к «истинному») вводится в рассмотрение специальный класс, в пределах которого не учитывается близость объектов друг к другу. Такой класс называется «фоновым».

Структура (классификация) задаётся вектор-функцией $H(x) = \big(h_0(x), h_1(x), \ldots, h_r(x)\big)$, где $h_0(x)$ — функция принадлежности x к фоновому классу, а $h_i(x)$ — функция принадлежности x к i-му нефоновому классу, r — число классов. Функция $H(x)$ удовлетворяет следующим условиям: $H(*) \in L_2(X, P)$ и для любого x значение $H(x)$ принадлежит некоторому ограниченному множеству V пространства значений вектор-функции $H(x)$, т. е. $H(x) \in V \subseteq \mathbb{R}^k$. Множество V фактически определяет вид нормировки функций принадлежности $h_i(x)$ (тип размытости).

Задача размытого структурного анализа ставится так: необходимо найти такую структуру (классификацию) $H(x)$, которая обеспечила бы экстремальное (далее, для определённости — максимальное) значение функционала (H) — критерия качества структуризации, т. е. $H_{\text{opt}} = \arg\max\limits_{H \in V} \Phi(H)$. Исследования показали, что в качестве $\Phi(H)$ можно выбирать произвольный выпуклый функционал от вектор-функции $H(x)$. Выпуклость, как обычно, означает:

$$\Phi\big(\gamma H^{(1)}(x) + (\gamma - 1)H^{(2)}(x)\big) \geqslant \geqslant \gamma\Phi\big(H^{(1)}(x)\big) + (\gamma - 1)\Phi\big(H^{(2)}(x)\big). \quad (1)$$

Было показано, что условию (1) удовлетворяет большинство функционалов, используемых

Работа выполнена при частичной финансовой поддержке РФФИ, проекты № 08-07-00347, 08-07-00349, 10-07-00210.

как критерии качества в задачах структурно-классификационного анализа данных [1].

Алгоритм структурно-классификационного анализа

Вид оптимальной классификации. Для исследования вида оптимальной размытой классификации введём понятие опорной размытой классификации для линейного функционала $F(H)$

$$H_F(x) = \arg\max_{H\in V}\big(F(x), H\big). \qquad (2)$$

Показано, что если F — субградиент функционала Φ в точке H, то $\Phi(H_F) \geqslant (H)$. Доказана теорема о том, что оптимальная размытая классификация принадлежит классу опорных классификаций. Этот результат позволяет построить итерационный алгоритм максимизации функционала $\Phi(H)$, который состоит из двух последовательных шагов:

— на первом шаге по данному решающему правилу (функционалу) $F(H)$ находится его опорная классификация H_F;
— на втором шаге по этой опорной классификации находится функционал, являющийся субградиентом исходного функционала $\Phi(H)$ (если функционал дифференцируем, то это — градиент функционала).

Рассмотрим алгоритм более подробно. Пусть задана некоторая *начальная классификация* H_0. Далее алгоритм строится итерационно.

На q-ом шаге в соответствии с алгоритмом построена классификация H^q.

На $(q+1)$-ом шаге находится некоторый субградиент F_q функционала $\Phi(H)$ в точке H^q. По F_q строится классификация H^{q+1}, являющаяся опорной к функционалу F_q, т. е. $H^{q+1} = H_{F_q}$.

Итак, строятся две последовательности: последовательность классификаций $H^0, H^1, \ldots, H^q, \ldots$, а также соответствующая ей последовательность субградиентов $F_0, F_1, \ldots, F_q, \ldots$.

Доказана теорема о сходимости этого алгоритма.

В разделе, посвящённом методам стохастической аппроксимации сам алгоритм и смысл понятия «сходимость алгоритма» обсуждаются более подробно.

Типовые задачи классификационного анализа

Рассмотрим применение предложенного подхода к решению классического варианта задачи размытой автоматической классификации с фоновым классом, в которой требуется максимизировать функционал

$$I_1 = -\sum_{i=1}^{r}\int_X (x-\alpha_i)^2 h_i(x)dP(x) + \\ + B\int_X h_0(x)dP(x). \qquad (3)$$

Здесь α_i — эталон (центр) i-го класса, определяемый по формуле:

$$\alpha_i = \int_X x h_i(x)dP(x) \Big/ \int_X h_i(x)dP(x),$$

а константа B — «цена» отнесения объектов к фоновому классу (чем больше B, тем больше объектов попадает в фоновый класс).

Градиентом (в данном случае функционал $\Phi(H) = I_1$ дифференцируем) функционала (3), в точке H является такая вектор-функция $\widetilde{F}(x) = \big(f_0(x), \ldots, f_r(x)\big)$, что:

$$f_0(x) = B; \quad f_i(x) = -(x-\alpha_i)^2; \quad i = 1, \ldots, r, \qquad (4)$$

т. е. решающая функция $f_i(x)$ для оптимальной классификации должна соответствовать мере близости (расстоянию с обратным знаком) точки x к эталону (центру) i-го класса.

Заметим, что множество $D(I_1)$ всех допустимых градиентов функционала I_1 состоит из функционалов вида (4). Таким образом, при максимизации I_1 можно ограничиться функционалами, задаваемыми через эталоны α_i по формуле (4).

Функционал I_1 является частным случаем класса функционалов метода обобщённого среднего, которые также являются выпуклыми и, следовательно, удовлетворяют основному условию сходимости описанного выше алгоритма [2]. Согласно этому методу, в каждом классе строится эталон (модель) класса и максимизируется суммарная взвешенная мера близости объектов и эталонов (моделей) соответствующих классов.

Формально задача метода обобщённого среднего записывается следующим образом:

$$\Phi_1 = -\sum_{i=1}^{r}\int_X K(x,\alpha_i)^2 h_i(x)dP(x) + \\ + B\int_X h_0(x)dP(x), \qquad (5)$$

где α_i — эталон i-го класса, принадлежащая множеству эталонов Λ, $K(x,\alpha_i)$ — мера близости между объектом x и эталоном i-го класса α_i. В свою очередь, эталон i-го класса определяется как решение следующей оптимизационной задачи:

$$\alpha_i = \arg\max_{\alpha\in\Lambda}\int_X K(x,\alpha)h_i(x)dP(x);\ i = 1, \ldots, r. \qquad (6)$$

В этот класс критериев входят как частные случаи: функционал средневзвешенной дисперсии, описанный выше; функционалы экстремальной группировки параметров [3], в этом случае эталоны (модели) — факторы групп (обобщённые параметры); функционал диагонализации матрицы связи [4], в этом случае множества X и Λ совпадают с множеством строк матрицы, а элементы матрицы играют роль меры близости; и некоторые другие.

Процедуры стохастической аппроксимации в классификационном анализе

В настоящем разделе на примере одного из функционалов метода обобщённого среднего показано, как используются теория и процедуры стохастической аппроксимации для конструирования и исследования сходимости алгоритмов структурно-классификационного анализа.

Для простоты далее будем рассматривать размытую классификацию без фонового класса, которая определяется вектор-функцией $H(x) = \big(h_1(x),\dots,h_r(x)\big)$. Здесь, как и ранее, $h_i(x)$ — функция принадлежности точки x к i-му классу, удовлетворяющая условиям нормировки: $h_i(x) \geqslant 0$, $i = 1,\dots,r$; $\sum_{i=1}^{r} h_i(x) = 1$.

Далее предполагается, что критерий качества классификации $\Phi(H) = \Phi_2(H)$ зависит от вероятностей и моментов классов. Функционал $\Phi_2(H)$ является частным случаем функционала (5). Для того, чтобы рассматривать только первые ненормированные моменты классов, введём в рассмотрение вектор-функцию $z(x)$ ($z\colon X \to Z = \mathbb{R}^k$). Обычно пространство Z называют спрямляющим пространством [3], так как в нём все рассматриваемые моменты являются первыми. Предполагается, что выполняются следующие два очевидных ограничения:

1. существует $A > 0 : P\big(|z(x) > A|\big) = 0$ (множество X ограничено по мере P);
2. для всех $\in \mathbb{R}^k$ и $\forall d \in \mathbb{R}^1$, выполняется: $P\Big\{\big(c, z(x)\big) + d = 0\Big\} = 0$ (мера точек z, сосредоточенных на любой плоскости пространства $\mathbb{R}^k$ равно 0, т. е. отсутствие вырожденности).

Рассмотрим первые ненормированные моменты и вероятности классов (нулевые моменты):

$$M_i = \int_X z(x)\varphi\big(h_i(x)\big)dP(x),\ i = 1,\dots,r;$$
$$p_i = \int_X \varphi\big(h_i(x)\big)dP(x). \qquad (7)$$

Здесь $\varphi(h)$ — монотонно-возрастающая функция, отображающая отрезок [0,1] на себя, причем $\varphi(0) = 0$ и $\varphi(1) = 1$. Заметим, что выбор функции $\varphi(h)$ даёт возможность варьировать тип размытости оптимальной классификации.

Обозначим через $\mu(H) = (p_1, M_1, \dots, p_r, M_r)$ вектор, составленный из вероятностей (нулевых моментов) и ненормированных моментов классов, его размерность равна $r(k+1)$.

Далее используется критерий качества классификации следующего вида:

$$\Phi_2(H) = \zeta\big(\mu(H)\big), \qquad (8)$$

где $\zeta(h)$ — выпуклая функция от $r(k+1)$-мерного вектора $\mu(H)$.

Вид оптимальной классификации. Пусть задан $r(k+1)$-мерный вектор $\pi = (d_1, c_1, \dots, d_r, c_r)$, где $d_i \in \mathbb{R}_1$, $c_i \in \mathbb{R}^k$, $i = 1,\dots,r$. Назовём линейной с вектором коэффициентов π классификацию (функционал)

$$H_\pi(x) = \arg \max_{(h_1,\dots,h_r)} \sum_{i=1}^{r} \Big(\big(c_i, z(x)\big) + d_i\Big)\varphi(h_i). \quad (9)$$

Функционал $H_\pi(x)$ является опорной размытой классификацией (2) для частного вида линейного функционала $F(H)$.

Доказана следующая лемма о виде оптимальной классификации для функционала (9) (следует из общей теоремы о виде оптимальной классификации выпуклого функционала $\Phi(H)$).

Лемма 1. *Если π — субградиент функции ζ в точке $\mu(H)$, то $\zeta\big(\mu(H_\pi)\big) \geqslant \zeta\big(\mu(H)\big)$, то есть оптимальная классификация должна быть линейной.*

Рекуррентный алгоритм классификации. Необходимо построить алгоритм максимизации функционала (9) по бесконечной выборке объектов $S = x_1,\dots,x_n,\dots$, появляющихся независимо в соответствии с, вообще говоря, неизвестным законом распределения $P(x)$.

На каждом шаге алгоритма по конечной подвыборке объектов $S_n = x_1,\dots,x_n$ для классификации $H(x) = \big(h_1(x),\dots,h_r(x)\big)$, построенной к этому шагу, строятся оценки ненормированных моментов и вероятностей классов (7):

$$s_i = \frac{1}{n}\sum_{j=1}^{n} z(x_j)\varphi\big(h_i(x_j)\big), \quad v_i = \frac{1}{n}\sum_{j=1}^{n} \varphi\big(h_i(x_j)\big).$$

Обозначим через $\psi(H)$ вектор частот и оценок моментов классов, его размерность равна $r(k+1)$.

Лемма 2. *Для любых $\eta > 0$ и $\varepsilon > 0$ с вероятностью, большей $(1-\eta)$, одновременно для всех классификаций H выборки S_n выполняется неравенство*

$$\zeta\big(\mu(H_\pi)\big) \geqslant \zeta\big(\psi(H)\big) -$$
$$- |\pi| D\left[2\varepsilon + \sqrt{\frac{r(k+1)}{2}\ln\Big(\frac{4r(k+1)}{\varepsilon}\Big) + \ln\Big(\frac{2}{\eta}\Big)}\right],$$

где π — субградиент функции ζ в точке $\psi(H)$, а D — некоторая константа.

В разработанном алгоритме при появлении на n-ом шаге новой точки x_n пересчитываются оценки моментов и определяется текущая классификация множества X. Формально алгоритм записывается следующим образом:

$$\begin{aligned} v_i^n &= \frac{(n-1)v_i^{n-1} + h_i^{n-1}(x_n)}{n}; \\ s_i^n &= \frac{(n-1)s_i^{n-1} + h_i^{n-1}(x_n)z(x_n)}{n}; \\ \psi^n &= (v_1^n, s_1^n, \ldots, v_r^n, s_r^n). \end{aligned} \qquad (10)$$

Доказана следующая теорема о достаточных условиях сходимости алгоритма (10) к классификации, обеспечивающей стационарное значение критерия(8).

Теорема 3. *Если ζ — выпуклая функция и все её субградиенты ограничены, то в силу алгоритма (10) с вероятностью единица справедливо:*

1) $\underline{\lim}_{n\to\infty} \Phi_2(H^n) \geqslant \lim_{n\to\infty} \Phi_2(\psi^n) = C(S)$, *где $C(S)$ — константа, зависящая от выборки S, являющаяся односторонне-стационарным значением функции ζ;*
2) *если, кроме того, ζ — дважды непрерывно дифференцируемая функция, то $C(S)$ — стационарное значение функции ζ и любая предельная точка последовательности $\{\psi^n\}$ является стационарной.*

Выводы

В работе рассмотрена размытая постановка задачи структурно-классификационного анализа. Исследован вид оптимальной классификации, предложен общий итерационный алгоритм максимизации критерия качества структурно-классификационного анализа. На примере метода обобщённого среднего показано, как используются теория и процедуры стохастической аппроксимации для конструирования и исследования сходимости алгоритмов структурно-классификационного анализа.

Литература

[1] *Бауман Е. В., Дорофеюк А. А.* Классификационный анализ данных // «Избранные труды Международной конференции по проблемам управления. Том 1», Москва: СИНТЕГ, 1999.

[2] *Бауман Е. В., Блудян Н. О.,* Методы нахождения глобальных экстремумов функционалов в за-даче классификационного анализа данных. // Труды ИПУ РАН, Т.XIII, М: ИПУ РАН, 2001 — С. 129–136.

[3] *Браверман Э. М., Мучник И. Б.* Структурные методы обработки эмпирических данных. — М.: Наука, 1983.

[4] *Браверман Э. М., Дорофеюк А. А. и др.* Диагонализация матрицы связи и выделение скрытых факторов. // Проблемы расширения возможностей автоматов // Сб. трудов Института проблем управления, Вып. 1, М.: ИПУ РАН, 1971. — С. 42–79.

Оптимальные алгоритмы размытой кусочно-линейной аппроксимации сложных зависимостей*

Дорофеюк А. А., Бауман Е. В., Дорофеюк Ю. А.
daa2@mail.ru
Москва, Институт проблем управления им. В. А. Трапезникова РАН

Работа посвящена методам решения задачи кусочно-линейной аппроксимации, разработанных на базе общего подхода к задачам классификационного анализа данных. Основная идея кусочной аппроксимации сложной зависимости состоит в разбиении пространства входных параметров на такие области, в пределах каждой из которых сложную во всем пространстве зависимость можно аппроксимировать достаточно простой функцией, например, линейной. Введение достаточно естественного критерия качества аппроксимации в виде среднеквадратичного отклонения кусочно-линейной модели от выборки реальных входо-выходных данных, позволяет строить оптимальные алгоритмы его минимизации, а в некоторых случаях разрабатывать и глобально-оптимальные алгоритмы.

Постановка задачи кусочно-линейной аппроксимации

Случай чёткой классификации.

Рассматривается задача построения модели зависимости выходного показателя y от вектора входных показателей $x = (x^{(1)}, \ldots, x^{(k)}) \in X \subseteq \mathbb{R}^k$. Модель строится по выборке из n объектов, каждый из которых описывается вектором

$$(y_t, x_t) = (y_t, x_t^{(1)}, \ldots, x_t^{(k)}) \in \widetilde{X} \subseteq \mathbb{R}^{k+1}.$$

Классическая схема кусочно-линейной аппроксимации состоит в следующем [1]:

С помощью одного из алгоритмов автоматической классификации [2] строится классификация $H = (H_1, \ldots, H_r)$ имеющейся в пространстве $\mathbb{R}^k$ выборки (y_t, x_t), $t = 1, \ldots, n$ на r классов. Затем в каждом классе по методу наименьших квадратов строится линейная регрессия выходного показателя y от вектора входных показателей x. В i-м классе находятся такой вектор коэффициентов

$$c_i = (c_i^{(1)}, \ldots, c_i^{(k)})$$

и константа d_i линейной функции:

$$\big((c_i, x) + d_i\big) = d_i + \sum_{j=1}^{k} c_i^{(j)} x^{(j)},$$

которые минимизируют функционал

$$K_i = \sum_{x_t \in H_i} \big(y_t - \widetilde{F}_j(x_t)\big)^2.$$

Задача кусочно-линейной аппроксимации состоит в нахождении такого разбиения на классы, для которого сумма квадратов невязок по моделям всех классов была бы минимальна. Другими словами, необходимо найти такую классификацию H и такие векторы коэффициентов c_i и константы d_i, для которых функционал

$$I = \sum_{i=1}^{r} \sum_{x_t \in H_i} \Big(y_t - \big((c_i, x_t) + d_i\big)\Big)^2$$

принимал бы минимальное значение или функционал

$$J_{KA} = -I = -\sum_{i=1}^{r} \sum_{x_t \in H_i} \Big(y_t - \big((c_i, x_t) + d_i\big)\Big)^2$$

принимал бы максимальное значение. Последний функционал является частным случаем функционала классификационного анализа общего вида [2]

$$J(H, A) = \sum_{x \in X} \sum_{i=1}^{r} K\big(x, \alpha_i\big) \varphi\big(h_i(x)\big).$$

Классификацию H будем задавать через вектор-функцию принадлежностей

$$H\big(x\big) = \big(h_1(x), \ldots, h_r(x)\big).$$

Тогда функционал I можно переписать в следующих двух эквивалентных записях:

$$I_1 = \sum_{i=1}^{r} \sum_{j=1}^{n} \Big[y_t - \big((c_i, x_t) + d_i\big)\Big]^2 h_i(x); \qquad (1)$$

$$I_2 = \sum_{j=1}^{n} \Big[y_t - \sum_{i=1}^{r} \big((c_i, x_t) + d_i\big) h_i(x)\Big]^2. \qquad (2)$$

Когда классификация чёткая (т. е. каждый объект однозначно относится к одному из классов), функционалы (1) и (2) совпадают. Однако интерпретируются они по-разному. При минимизации функционала (1) линейные модели зависимостей строятся в каждом классе в отдельности, а затем суммируются квадраты отклонения ошибок. В функционале (2) выражение $\sum_{i=1}^{r} \big((c_i, x_j) + d_i\big) h_i(x)$

Работа выполнена при частичной финансовой поддержке РФФИ, проекты № 11-07-00178, 10-07-00210.

можно считать кусочно-линейной моделью выходного показателя y. Основной сложностью решения данной задачи является то, что при минимизации функционалов (1) или (2), как по классификации, так и по коэффициентам модели, решающие правила оптимальной классификации H записываются в терминах не только входных, но и выходного показателя. Это существенно уменьшает возможность использования построенной модели для прогноза. Приходится использовать различные варианты проекции классификации в расширенном пространстве $\tilde{X}$ на пространство X входных показателей. Подобное проектирование областей на пространство меньшей размерности приводит к появлению в пространстве входных показателей зон, в которых одновременно могут действовать модели как одного, так и другого класса. Именно исходя из этого, при построении кусочных моделей по существу возникает размытость между классами.

Случай размытой классификации. Рассмотрим размытую постановку задачи кусочной аппроксимации. Будем задавать классификацию через вектор-функцию принадлежностей

$$H\big(x\big) = \big(h_1(x), \ldots, h_r(x)\big),$$

удовлетворяющую ограничению:

$$\sum_{i=1}^{r} h_i(x) = 1; \quad h_i(x) \geqslant 0; \;\; x \in X; \;\; i = 1, \ldots, r. \quad (3)$$

В данном случае функционалы (1) и (2) не совпадают, и их минимизация приводит к разным результатам. Так как функционал

$$I_1 = I_1\big(c_i, d_i, i = 1, \ldots, r; H(x)\big)$$

линеен по $H(x)$, то функционал

$$\hat{I}_1 = -\min_{c_i, d_i} I_1\big(c_i, d_i, i = 1, \ldots, r; H(x)\big)$$

является выпуклым по $H(x)$. Отсюда следует, что оптимальная классификация $H(x)$ лежит на границе допустимой области, т. е. в ограничении все $h_i(x)$ равны либо 1, либо 0, что соответствует чёткой классификации. Следовательно, в оптимальной кусочной аппроксимации для функционала (1) при ограничениях (3) классификация — чёткая. Рассмотрим функционал (2). Достаточно легко показать, что, если на $h_i(x)$ не накладывать дополнительные ограничения, то за счёт большого количества степеней свободы можно всегда точно аппроксимировать все значения y на данной выборке.

Для того чтобы использовать функционал (2) в размытом варианте, необходимо ограничить класс разрешённых функций $h_i(x)$. В соответствии с общей методикой обобщённого среднего [2] для получения кусочной аппроксимации с размытой классификацией функционал (1) необходимо модифицировать следующим образом:

$$I_3 = \sum_{i=1}^{r} \sum_{j=1}^{n} \Big[y_j - \big((c_i, x_j) + d_i\big) \Big]^2 \varphi\big(h_i(x)\big). \quad (4)$$

Здесь $\varphi(h)$ — функция, определяющая тип размытости [2]. В частности, для $\varphi_1(h) = h$ имеет место чёткая кусочная аппроксимация (критерии (1) и (4) совпадают), а для $\varphi_2(h) = (h)^t$, $t > 1$ — размытая кусочная аппроксимация с функцией принадлежности $h_i(x)$. Поскольку функционал (4) является частным случаем функционала классификационного анализа, то для его оптимизации можно использовать общий алгоритм классификационного анализа [2], а именно, алгоритм — это последовательное применение двухэтапной процедуры: 1) фиксируется вектор-функция $H(x)$, для неё находятся оптимальные значения коэффициентов модели $c_i, d_i, i = 1, \ldots, r$; 2) фиксируются коэффициенты локальных регрессий $c_i, d_i, i = 1, \ldots, r$ и для них находится оптимальная классификация $H(x)$. Сходимость этого алгоритма следует из сходимости общего алгоритма [2].

Как и для чёткого случая, недостатком такого подхода является то, что в решающие правила аппроксимации входят не только входные показатели, но и выходной.

Случай с ограничениями на класс решающих правил. В ряде работ предлагается строить кусочную аппроксимацию так, чтобы классификация производилась по одному набору показателей, а аппроксимация в каждом классе — по другому (см., например, [3]). Будем считать, что кроме пространства X есть ещё пространство $Z = \mathbb{R}^s$, в котором и производится классификация объектов. Дальнейшие рассуждения не изменятся, если часть показателей в пространствах X и Z будут одни и те же. Считается, что каждый из n объектов исходной выборки, описывается $k + s + 1$ параметром, т. е. вектором

$$(y_t, x_t, z_t) = (y_t, x_t^{(1)}, \ldots, x_t^{(k)}, z_t^{(1)}, \ldots, z_t^{(s)}) \in \mathbb{R}^{k+s+1}.$$

Самый используемый критерий качества классификации — средневзвешенная дисперсия, который для пространства Z запишется как функционал, зависящий от эталонов классов $A = (\alpha_1, \ldots, \alpha_r)$ и вектор-функции принадлежностей $H(z)$:

$$J\big(A, H(z)\big) = \sum_{i=1}^{r} \sum_{t=1}^{n} (z_t - \alpha_i)^2 \varphi\big(h_i(z_t)\big).$$

Считается, что эталоны классов могут быть произвольными точками пространства Z, вектор-функция $H(z)$ удовлетворяет условиям (3), а функция φ определяет тип размытости. Минимизация

этого функционала производится как по классификации $H(z)$, так и по набору эталонов классов $A = (\alpha_1, \ldots, \alpha_r)$. В оптимальном случае эталон i-го класса записывается в виде

$$\alpha_i = \sum_{t=1}^{n} z_t \varphi\big(h_i(z_t)\big) / \sum_{t=1}^{n} \varphi\big(h_i(z_t)\big),$$

то есть совпадает с центром класса. Если $\varphi(t) = \varphi_1(t)$, то итерационный алгоритм минимизации этого функционала совпадает с известным алгоритмом ISODATA [2]. Если $\varphi(t) = \varphi_2(t)$, то — с размытым вариантом ISODATA. Если фиксирован набор эталонов классов A, то эталонная классификация $H^A(z) = \big(h_1^A(z), \ldots, h_r^A(z)\big)$ для каждой функции $\varphi(h)$ определяется однозначно. Так, например, для $\varphi_1(h)$: $h_i^A(x) = 1$, если $i = \arg \min\limits_{j=1,\ldots,r} (z - a_j)$, и равно 0 в противном случае, а для $\varphi_1(h)$:

$$h_i^A(x) = \frac{(z - a_i)^{\frac{2}{1-t}}}{\sum_{j=1}^{r} (z - a_j)^{\frac{2}{1-t}}}.$$

Другими словами для любого набора из r векторов пространства Z можно определить эталонную классификацию $H^A(z) = \big(h_1^A(z), \ldots, h_r^A(z)\big)$. Ограничимся при решении задачи аппроксимации множеством эталонных классификаций пространства Z.

Постановка задачи аппроксимации для случая: минимизировать функционал кусочной аппроксимации (1), (2) или (4) при условии, что классификация $H(x)$ является эталонной в пространстве Z. Свободными параметрами, по которым необходимо оптимизировать функционал качества аппроксимации, являются: во-первых, набор эталонов классов $A = (\alpha_1, \ldots, \alpha_r)$, задающих эталонную классификацию; а во-вторых, коэффициенты линейных моделей каждого из классов c_i, d_i, $i = 1, \ldots, r$. Всего $(l+k+1)r$ параметров. Функционалы (1), (2) или (4) легко переписываются как функционалы от эталонной классификации $H^A(z)$:

$$I_1' = \sum_{i=1}^{r} \sum_{j=1}^{n} \Big[y_t - \big((c_i, x_t) + d_i\big) \Big]^2 h_i^A(z_t); \qquad (5)$$

$$I_2' = \sum_{j=1}^{n} \Big[y_t - \sum_{i=1}^{r} \big((c_i, x_t) + d_i\big) h_i^A(z_t) \Big]^2; \qquad (6)$$

$$I_3' = \sum_{i=1}^{r} \sum_{j=1}^{n} \Big[y_j - \big((c_i, x_j) + d_i\big) \Big]^2 \varphi\big(h_i^A(z_t)\big). \qquad (7)$$

Функционалы (5)-(7) дифференцируемы по свободным параметрам и для нахождения их локальных экстремумов можно применять градиентные процедуры (для (7) это справедливо для подавляющего большинства используемых функций φ). Недостатком алгоритмов локальной оптимизации является сильная зависимость результата от начальных условий работы алгоритма. Поэтому актуальным является разработка методов глобальной оптимизации.

Алгоритм построения аппроксимации для конечного множества эталонов. Заметим, что если зафиксировать набор эталонов классов $A = (\alpha_1, \ldots, \alpha_r)$, то по методу наименьших квадратов однозначно находятся коэффициенты линейных моделей классов c_i, d_i, $i = 1, \ldots, r$, минимизирующие один из функционалов (5)–(7). Таким образом, если можно перебрать все возможные наборы эталонов классов, то можно найти глобальный минимум выбранного функционала. Пусть в пространстве Z выделено некоторое конечное множество точек $Z_p = \beta_1, \ldots, \beta_p$. Потребуем, чтобы эталоны классов выбирались только из множества Z_p. Число вариантов выбора различных эталонов будет равно p^r. Так как в прикладных задачах число классов в кусочной аппроксимации редко бывает больше пяти-шести, а число точек в Z_p для подавляющего числа прикладных задач порядка 100, то такой перебор вполне можно делать на современных ПЭВМ. В качестве множества Z_p можно взять, например, реализацию в пространстве Z исходной выборки объектов. В данном случае $Z_n = z_1, \ldots, z_n$. В качестве множества Z_p можно взять также достаточно разреженную решетку в пространстве Z. Такой вариант хорошо использовать в качестве начальных условий для градиентного алгоритма построения кусочной аппроксимации без ограничения на набор эталонов. Следует отдельно выделить случай одномерного пространства Z (один показатель).

Кусочная аппроксимация для одномерного классифицирующего пространства. Общая постановка задачи классификации не учитывает специфику одномерной классификации, а именно — упорядоченность точек числовой прямой. В результате в оптимальной классификации все точки (за исключением эталонов классов) принадлежат с некоторым весом всем классам одновременно. Более того, функция принадлежности i-го класса $h_i(x)$ не унимодальна и имеет достаточно сложную структуру. Однако в случае одномерной классификации естественно предполагать, что перекрываться могут лишь соседние классы. Тогда оказывается, что α_i является не только эталоном i-го класса, но и границей между $(i-1)$-м и $(i+1)$-м классами. В соответствии с этим естественно наложить на эталонную классификацию $H^A(z)$ следующие дополнительные ограничения (пусть $\alpha_1 < \alpha_2 < \cdots < \alpha_n$): если $z \leqslant \alpha_1$, то на этом луче все точки однозначно относятся к первому классу; если $\alpha_{i-1} \leqslant z \leqslant \alpha_i$, то на этом

отрезке ненулевые веса могут иметь лишь i-й и $(i-1)$-й классы; если $\alpha_r \leqslant z$, то на этом луче все точки однозначно относятся к последнему классу. Заметим, что функционал (5) — это частный случай (7), для которого $\varphi(h) = h$. Поэтому далее рассматривается алгоритм только для (7).

Эталоны классов разбивают числовую прямую на $(r+1)$ промежутков, в каждом из которых могут применяться не более двух локальных линейных моделей кусочной аппроксимации. В соответствии с этим перепишем функционал (7) в следующем виде:

$$I_3''(A; c_i, d_i, i = 1, \ldots, r) = \\ = \sum_{i=1}^{r} \big(\Delta(i, \alpha_{i-1}, \alpha_i, c_i, d_i) + \Delta(i, \alpha_i, \alpha_{i+1}, c_i, d_i)\big),$$

где

$$\Delta(i, \alpha, \beta, c, d) = \sum_{\alpha \leqslant z_t \leqslant \beta} \Big[\big(y_t - (cx_t + d)\big)^2 \varphi\big(h_i^A(z_t)\big)\Big].$$

Предполагается, что $\alpha_0 = -\infty$, а $\alpha_{r+1} = +\infty$. Если известны α_{i-1}, α_i и α_{i+1}, то однозначно известна функция $h_i(x)$, следовательно, по ней однозначно рассчитываются коэффициенты c_i и d_i. Обозначим

$$S_i(\alpha_{i-1}, \alpha_i, \alpha_{i+1}) = \\ = \big(\Delta(i, \alpha_{i-1}, \alpha_i, c_i, d_i) + \Delta(i, \alpha_i, \alpha_{i+1}, c_i, d_i)\big),$$

тогда

$$I_3''(A; c_i, d_i, i = 1, \ldots, r) = \sum_{i=1}^{r} S_i(\alpha_{i-1}, \alpha_i, \alpha_{i+1}).$$

Для $i = r-1, \ldots, 1$ строим функции

$$F_r(\alpha_{r-1}, \alpha_r) = S_r(\alpha_{r-1}, \alpha_r, \alpha_{r+1});$$

$$F_i(\alpha_{i-1}, \alpha_i) = \min_{\alpha_{i+1}} \big[S_i(\alpha_{i-1}, \alpha_i, \alpha_{i+1}) + F_{i+1}(\alpha_i, \alpha_{i+1})\big].$$

Последнее выражение — рекуррентное уравнение динамического программирования Беллмана [4]. Решая задачу

$$F_0(\alpha_0) = \min_{\alpha_i} \big[S_1(\alpha_0, \alpha_1, \alpha_2) + F_1(\alpha_0, \alpha_1)\big],$$

получим оптимальное значение α_1, по нему — α_2 и так далее до получения оптимального значения α_r.

Кусочная аппроксимация с классификацией по выходному параметру. В задаче кусочной аппроксимации (например, при идентификации сложного промышленного объекта) часто бывает полезно результирующую классификацию проектировать не на пространство входов X, а на выходной параметр y [1]. Тогда классы можно интерпретировать как режимы работы, которые обеспечивают соответствующий диапазон значений y. Для этого надо рассмотреть случай $z = y$ и глобальный экстремум соответствующих функционалов также находится с помощью процедуры Беллмана.

Выводы. В работе рассмотрены алгоритмы решения задачи кусочно-линейной аппроксимации сложной зависимости с использованием вариационного подхода к задачам классификационного анализа данных. Даются постановки этой задачи, как для чёткой, так и для размытой классификаций. Рассматривается два способа нахождения глобального экстремума функционала — для случая конечного множества эталонов и для одномерного классификационного пространства. Для последнего случая построена рекуррентная схема Беллмана, реализация которой и обеспечивает получение глобального экстремума.

Литература

[1] *Райбман Н. С., Дорофеюк А. А., Касавин А. Д.* Идентификация технологических объектов методами кусочной аппроксимации. — М.: ИПУ, 1977. — 70 с.

[2] *Бауман Е. В., Дорофеюк А. А.* Классификационный анализ данных // «Избранные труды Международной конференции по проблемам управления. Том 1», Москва: СИНТЕГ, 1999.

[3] *Бауман Е. В., Дорофеюк А. А., Чернявский А. Л.* Методы структурной обработки эмпирических данных // Измерение, контроль, автоматизация. — 1985. — № 3.

[4] *Bellman R.* Dynamic programming and Lagrange multipliers // Proc. USA National. Academy of Sciences. — 1956. — V. 42.— pp. 767–769.

Корректные расширения корректных ΣП-алгоритмов*

Шибзухов З. М.
szport@gmail.com
Москва, НИИ ПМА КБНЦ РАН

Исследуются корректные расширения семейств корректных ΣП-алгоритмов при помощи корректных операций, которые наборам корректных алгоритмов ставят в соответствие корректный алгоритм. Приводится обоснование того, что применение корректных операций в сочетании конструктивными методами обучения ΣП-алгоритмов позволяет эффективно строить определенным образом скорректированные ΣП-алгоритмы, которые могут выступать в качестве базовых в композиционных схемах обучения типа Boosting или Bagging.

Одной из важных проблем обучения алгоритмов является проблема построения *корректных алгоритмов* [1], которые не являются переобученными. Она часто возникает, когда метод обучения не позволяет строить алгоритмы, которые были бы одновременно корректными и требуемого качества. Существуют конструктивные методы, с помощью которых можно строить корректные алгоритмы [2, 3, 4], но нередко они оказываются переобученными. Важность свойства корректности алгоритмов, построенных по обучающему множеству показана при помощи комбинаторных оценок в рамках комбинаторного подхода к проблеме переобучения [5].

Существуют модели и методы обучения, в которых для любого непротиворечивого набора прецедентов $\mathcal{I}_0 \subset \mathcal{I}$ можно эффективно построить значительный набор алгоритмов, корректно функционирующих на $\mathcal{I}_0$, из которых затем можно выбирать конечные наборы алгоритмов, содержащие наилучшие по отношению к заданному внешнему критерию качества (например, с минимальным числом ошибок на $\mathcal{I} \setminus \mathcal{I}_0$).

Примером такой модели и метода обучения является модель ΣП-нейронов и последовательный конструктивный метод обучения ΣП-нейронов [4]. Особенностью данного метода является то, что построение алгоритма осуществляется за один проход предварительно упорядоченной последовательности прецедентов, так что в процессе обучения на каждом шаге построенный к этому моменту алгоритм является корректным на уже пройденной части набора прецедентов. А так как на очередном шаге, как правило, существует несколько вариантов «достраивания», то в результате можно получить значительное множество корректных алгоритмов. Поэтому из него отбираются те, которые являются наилучшими по отношению к заданному внешнему критерию. Для получения алгоритмов с хорошими оценками обобщающей способности приходится осуществлять построение либо по части полного набора прецедентов, либо, начав обучение по полному набору прецедентов, прерывать процесс обучения по достижении оптимального значения внешнего критерия на еще не пройденной части. В результате мы имеем наборы алгоритмов, корректные на некотором подмножестве прецедентов $\mathcal{I}_0 \subset \mathcal{I}$.

В связи этим возникает вопрос: как расширить набор алгоритмов, корректных на $\mathcal{I}_0$, до некоторого множества алгоритмов, так, чтобы

1) все алгоритмы сохраняли бы свойство корректности на $\mathcal{I}_0$;

2) эти множества в последствии расширялись так, чтобы можно было эффективно найти достаточный набор алгоритмов, корректных на бóльшем наборе прецедентов $\mathcal{I}_1 \supset \mathcal{I}_0$, и наилучших по отношению к заданному внешнему критерию.

В контексте этой проблемы изучаются *корректные операции* над алгоритмами, которые определяют специальный тип *корректирующих операций*. Корректные операции преобразуют наборы алгоритмов, корректных на $\mathcal{I}_0$, в алгоритмы также корректные на $\mathcal{I}_0$. Главная цель применения корректных операций состоит в построении более оптимальных алгоритмов на $\mathcal{I}$, чем корректные алгоритмы, построенные по $\mathcal{I}_0$. Это можно сравнить с поиском оптимального алгоритма из линейного замыкания набора алгоритмов с близкими моделями в рамках непараметрической схемы построения распознающих алгоритмов [6].

Показывается, что скорректированные таким образом алгоритмы являются достаточно хорошими кандидатами для использования в композиционных схемах типа Bagging или Boosting.

Корректные наборы алгоритмов. Пусть $\mathbf{X} \subset \mathbb{R}^n$, $\mathbf{Y} \subset \mathbb{R}$, $\mathcal{I} = \{\langle \boldsymbol{x}, y\rangle\}$ — конечный набор прецедентов. Каждый алгоритм $\mathsf{a}\colon \mathbf{X} \to \mathbf{Y}$ имеет вид: $\mathsf{a} = \mathsf{R} \circ \mathsf{A}$, где $\mathsf{A}\colon \mathbf{X} \to \mathbf{S}$ — алгоритмический (оценивающий) оператор, $\mathsf{R}\colon \mathbf{S} \to \mathbf{Y}$ — решающее правило ($\mathbf{S} \subseteq \mathbb{R}^p$, $p \geqslant 1$), $L(y_1, y_2)$ — функция потерь. Для любого $y \in \mathbf{Y}$ обозначим

$$\mathbf{S}_y = \big\{\boldsymbol{s} \in \mathbf{S}\colon L\big(\mathsf{R}(\boldsymbol{s}), y\big) = 0\big\},$$
$$\mathbf{Y}_y = \{z \in \mathbf{Y}\colon L(z, y) = 0\}.$$

Работа выполнена при поддержке Программы фундаментальных исследований ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики» и гранта РФФИ № 09-01-00166-а

Определение. Набор прецедентов $\mathcal{I}$ является *непротиворечивым*, если в нем нельзя найти пару прецедентов $\langle \boldsymbol{x}, y' \rangle$ и $\langle \boldsymbol{x}, y'' \rangle$, таких что $L(y', y'') \neq 0$.

Далее будем рассматривать только непротиворечивые наборы прецедентов. Понятие корректности алгоритмов будем рассматривать в следующем смысле.

Определение. Алгоритм a *корректен для прецедента* $\langle \boldsymbol{x}, y \rangle$, если $L\big(\mathsf{a}(\boldsymbol{x}), y\big) = 0$ и *корректен на* $\mathcal{I}$, если он корректен для каждого прецедента из $\mathcal{I}$. Набор алгоритмов $\mathbf{A} = \{\mathsf{a}\}$ *корректен на* $\mathcal{I}$, если всякий алгоритм $\mathsf{a} \in \mathbf{A}$ корректен на $\mathcal{I}$.

Конструктивные ΣП-алгоритмы.

Пусть $\mathbf{X} \subset \mathbb{R}^n$, $\mathcal{I}^0 = \{\langle \boldsymbol{x}_k, y_k \rangle \colon k = 1, \ldots, N\}$ — непротиворечивый набор прецедентов, $\mathbf{X}^0 = \{\boldsymbol{x}_k \colon k = 1, \ldots, N\}$, $\mathbf{Y}^0 = \{y_k \colon k = 1, \ldots, N\}$.

Рассмотрим ΣП-алгоритмы [7, 8] вида:

$$\mathrm{spa}(\boldsymbol{x}) = \mathrm{R} \circ \mathbf{sp}(\boldsymbol{x});$$
$$\mathbf{sp}(\boldsymbol{x}) = \mathbf{g}(\boldsymbol{x}) + \sum_k \mathbf{w}_k \mathrm{pn}_k(\boldsymbol{x}), \qquad (1)$$

где $\mathbf{sp} \colon \mathbf{X} \to \mathbb{R}^q$ — вектор ΣП-оценок, $\mathbf{g} \colon \mathbf{X} \to \mathbb{R}^q$ — произвольно заданная начальная функция, $\{\mathbf{w}_k \in \mathbb{R}^q\}$ — весовые векторы,

$$\mathrm{pn}_k(\boldsymbol{x}) = \eta_k\Big(\prod\nolimits_{j \in \boldsymbol{j}_k} \psi_j(\boldsymbol{x})\Big), \qquad (2)$$

где $\eta_k \colon \mathbb{R} \to \mathbb{R}$ — функция, т. ч. $\eta_k(s) = 0 \Leftrightarrow s = 0$, $\boldsymbol{j}_k \subseteq \{1, \ldots, M\}$, $k = 1, \ldots, M$.

Набор функций $\boldsymbol{\Psi} = \{\psi_j(\boldsymbol{x}) \colon j = 1, \ldots, M\}$ $(\psi_j \colon \mathbb{R}^n \to \mathbb{R})$, такой что:

1) каждая функция ψ_j имеет *конечный носитель*[1] $\mathbf{X}_j$;

2) набор $\{\mathbf{X}_j\}$ *образует покрытие* $\mathbf{X}^0$, которое *разделяет точки*[2] из $\mathbf{X}^0$.

Такие наборы функций и покрытия будем называть *допустимыми*.

Покрытие $\{\mathbf{X}_j\}$ можно построить разными способами, например путем оптимального выбора пороговых значений для каждой переменной $x_1, \ldots, x_n$, при помощи метода k-ближайших соседей и т. д.

Теорема 1 (о конструктивных ΣП-алгоритмах). *Для всякого допустимого конечного покрытия* $\mathbf{X}^0 \subset \bigcup\{\mathbf{X}_j\}$ *(соответственно, допустимого набора функций* $\boldsymbol{\Psi}$*) существует непустой набор* ΣП*-алгоритмов вида* (1)*, корректный на* $\mathcal{I}$*, который может быть эффективно построен [4, 8].*

Для минимизации сложности мультипликативных функций (2), входящих в ΣП-алгоритм (1), для каждого k строится множество минимальных конечных наборов индексов $\boldsymbol{J}_k = \{\boldsymbol{j}_k\}$, таких что 1) для всех $j \in \boldsymbol{j}_k$: $\boldsymbol{x}_k \in \mathbf{X}_j$; 2) для всякого $\ell \neq k$ найдется индекс j из $\boldsymbol{j}_k$, такой что $\boldsymbol{x}_\ell \notin \mathbf{X}_j$. Алгоритм поиска таких наборов индексов является незначительной модификацией соответствующего алгоритма, описанного в [4].

Алгоритм 1. Алгоритмическая схема построения корректных наборов ΣП-алгоритмов.

Вход: m
Выход: $\mathbf{A}$ — m корректных алгоритмов
// Строится набор $\mathbf{A}$*, состоящий*
// из m *корректных* ΣП*-алгоритмов на* $\mathcal{I}$*.*

```
k = 0; sp_0 = g; spa_0 = R ∘ sp_0; A_0 = {spa_0}
для k = 1 ... N
    A_k = ∅
    для всех spa_{k-1} в A_{k-1}
        если spa_{k-1}(x_k) = y_k то
            A_k = A_k ∪ {spa_{k-1}}
        иначе
            для всех j_k в J_k
                spa_k(x) = R ∘ sp_k(x), где
                sp_k(x) = sp_{k-1}(x) + w_k · pn_{j_k}(x),
                w_k = (s − sp_{k-1}(x_k)) / pn_{j_k}(x_k),
                s: L(R(s), y_k) = 0
                A_k ← A_k ∪ {spa_k}
    В A_k оставляется P_k ⩾ m наилучших алгоритмов, используя внешний критерий.
A ← A_N
В A оставляется m наилучших алгоритмов, используя внешний критерий.
```

Корректные операции и корректные замыкания. Введем понятие *корректной операции* — преобразования алгоритмов, которое набор корректных алгоритмов $\mathsf{a}_1, \ldots, \mathsf{a}_m$ на $\mathcal{I}_0 \subseteq \mathcal{I}$ переводит в алгоритм, корректный на $\mathcal{I}_0$. Определяемые ниже корректные операции представляют собой разновидность корректирующих операций.

Определение. Корректирующая операция Φ *по оценкам / ответам* является *корректной операцией* над алгоритмами, если для любого набора алгоритмов $\mathsf{a}_1, \ldots, \mathsf{a}_m$, корректных для прецедента $\langle \boldsymbol{x}, y \rangle$, алгоритм $\mathsf{a} = \mathsf{R} \circ \Phi(\mathsf{A}_1, \ldots, \mathsf{A}_m)$ / $\mathsf{a} = \Phi(\mathsf{a}_1, \ldots, \mathsf{a}_m)$ — также корректен для $\langle \boldsymbol{x}, y \rangle$.

Отметим два факта.

1) Пусть отображение $\Phi \colon \mathbf{S}^m \to \mathbf{S}$, такое что для любого $y \in \mathbf{Y}$ и любого набора $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_m \in \mathbf{S}_y$: $\Phi(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_m) \in \mathbf{S}_y$ (т. е. Φ не «смешивает» слои $\mathbf{S}_y$ в $\mathbf{S}$). Тогда Φ является корректной операцией по оценкам.

2) Пусть отображение $\Phi \colon \mathbf{Y}^m \to \mathbf{Y}$, такое что для любого $y \in \mathbf{Y}$ и любого набора $y_1, \ldots, y_m \in \mathbf{Y}_y$: $\Phi(y_1, \ldots, y_m) \in \mathbf{Y}_y$ (т. е. Φ не «смешивает» слои $\mathbf{Y}_y$

[1] т. е. $\varphi|_{\mathbf{X}_j} \neq 0$ и $\varphi|_{\bar{\mathbf{X}}_j} = 0$.

[2] т. е. для любой пары $\boldsymbol{x}' \neq \boldsymbol{x}''$ из $\mathbf{X}^0$ найдется носитель $\mathbf{X}_j$, такой что $\boldsymbol{x}' \in \mathbf{X}_j$, а $\boldsymbol{x}'' \notin \mathbf{X}_j$.

в $\mathbf{Y}$). Тогда Φ является корректной операцией по ответам.

По-существу, выбор корректных операций может определяться выбором решающего правила R и функции потерь L.

Пусть $\mathcal{F}$ — множество корректных операций. Введем понятие *корректного замыкания* набора алгоритмов относительно $\mathcal{F}$.

Определение. Множество алгоритмов, состоящих из всех алгоритмов, которые можно получить путем применения корректных операций из $\mathcal{F}$ к наборам алгоритмов из $\mathbf{A}_0$, называется *корректным замыканием* $\mathbf{A}_0$ *относительно* $\mathcal{F}$.

Примеры корректных операций.

I. Самые простые корректные операции — это *линейные*:

$$\Phi(\mathsf{A}_1,\dots,\mathsf{A}_m)=\sum\alpha_j(\boldsymbol{x})\mathsf{A}_j, \tag{3}$$

где $\alpha_j(\boldsymbol{x})\geqslant 0$ — вещественные веса, $\sum\alpha_j=1$.

Если для каждого $y\in\mathbf{Y}$ множества $\mathbf{S}_y$ — выпуклое, то линейные операции (3) являются корректными операциями.

Если набор $\{\Phi_\ell\}$ состоит из корректных операций, то операция $\sum w_\ell\Phi_\ell$, где $\sum w_\ell=1$, также является корректной.

II. Линейную операцию можно записать в более общей форме *обобщенно линейной* (по аналогии с обобщенными средними по Колмогорову):

$$\Phi(\mathsf{A}_1,\dots,\mathsf{A}_m)=\varphi^{-1}\Big(\sum\alpha_j(\boldsymbol{x})\varphi(\mathsf{A}_j)\Big), \tag{4}$$

где $\varphi\colon\mathbf{S}\to\mathbb{R}^q$ — обратимое преобразование ($q\geqslant 1$). Так если для каждого $y\in\mathbf{Y}$ множество $\varphi(\mathbf{S}_y)$ — выпуклое, то обобщенные линейные операции (4) также являются корректными операциями.

III. В общем случае если $\{\mathbf{S}_y\colon y\in\mathbf{Y}\}$ — набор выпуклых множеств, то всякая корректирующая операция Φ, индуцированная отображением $\mathbf{S}^m\to\mathbf{S}$, которая выпуклые подмножества $\mathbf{S}$ оставляет на месте, является корректной.

IV. В частном случае, когда $\mathbf{Y}=\mathbb{R}$, а решающее правило тривиальное ($\mathsf{R}(s)\equiv s$) можно предъявить общий способ порождения корректных операций. Так если функция $F\colon\mathbb{R}^m\to\mathbb{R}$, такая что функция $\varphi(s)=F(s,\dots,s)$ — обратимая, то $\varphi^{-1}\circ F$ — корректная операция. Например:

1. Если $F(s_1,\dots,s_m)=\exp\big(\sum w_js_j\big)$, то
   $\Phi(s_1,\dots,s_m)=\dfrac{1}{W}\sum w_js_j,\ W=\sum w_j.$
2. Если $F(s_1,\dots,s_m)=\prod s_j^{\lambda_j}$, то
   $\Phi(s_1,\dots,s_m)=\prod s_j^{\lambda_j/L},\ L=\sum\lambda_j.$
3. Если $F(s_1,\dots,s_m)=\sum w_{\boldsymbol{j}}\prod\limits_{j\in\boldsymbol{j}}s_j$, где $|\boldsymbol{j}|=r$, то
   $\Phi(s_1,\dots,s_m)=\big(\dfrac{1}{W}\sum w_{\boldsymbol{j}}\prod\limits_{j\in\boldsymbol{j}}s_j\big)^{1/r},\ W=\sum w_{\boldsymbol{j}}.$

V. Для задачи классификации на два класса с $\mathsf{R}(s)=\operatorname{sign}s$ если функция $\Phi(s_1,\dots,s_m)$ сохраняет знак: если $s_1,\dots,s_m<0$, то $\Phi(s_1,\dots,s_m)>0$, и если $s_1,\dots,s_m<0$, то $\Phi(s_1,\dots,s_m)>0$, то тогда она определяет корректную операцию.

VI. Для задачи многоклассовой классификации с $\mathsf{R}(s_1,\dots,s_p)=\arg\max\limits_{t=1,\dots,p}s_t$ преобразование

$$\Phi(\boldsymbol{s}_1,\dots,\boldsymbol{s}_m)=\big(\varphi(s_{11},\dots,s_{m1}),\dots,\varphi(s_{1p},\dots,s_{mp})\big),$$

где φ — монотонная функция[3] определяет корректную операцию.

VII. Мультипликативные операции (когда они имеют смысл) вида

$$\Phi(\mathrm{A}_1,\dots,\mathrm{A}_m)=\prod\mathrm{A}_j^{\beta_j(\boldsymbol{x})},$$

при условии что $\sum\beta_j(\boldsymbol{x})=1$, также являются корректными.

Композиция линейных и мультипликативных корректных операций вида

$$\Phi(\mathrm{A}_1,\dots,\mathrm{A}_m)=\sum\alpha_j(\boldsymbol{x})\prod_{t\in\boldsymbol{t}_j}\mathrm{A}_t^{\beta_{jt}(\boldsymbol{x})}$$

определяет корректную операцию, если $\sum\alpha_j(\boldsymbol{x})=1$ и $\sum\beta_j(\boldsymbol{x})=1$.

Семейства подходящих конструктивных ΣΠ-алгоритмов. Конструктивный алгоритм обучения ΣΠ-алгоритмов по заданному набору прецедентов $\mathcal{I}_0$ строит конечный набор алгоритмов, корректных на $\mathcal{I}_0$. Применяя корректные операции к этому набору (или поднабору) можно строить другие корректные алгоритмы. Можно решать задачу поиска среди них алгоритма, который дает большинство правильных ответов на $\mathcal{I}\setminus\mathcal{I}_0$.

В то же время алгоритм конструирования ΣΠ-алгоритмов легко адаптируется для построения семейств ΣΠ-алгоритмов, которые «подстраиваются» под заданную корректирующую операцию. В таких случаях корректирующая операция, по существу выступает в роли решающего правила и определяет определенный принцип принятия решения на основе «мнений экспертов».

Рассмотрим, например, корректирующие операции по ответам, индуцируемые функцией от m аргументов. Так, в задаче классификации, такой функцией может являться функция «ответа большинства»:

$$\Phi(y_1,\dots,y_m)=\arg\max_{y\in\{y_1,\dots,y_m\}}\sum_{i=1}^m[y=y_i].$$

В задаче регрессии это может быть функции нахождения «точки наибольшего сгущения» ответов:

$$\Phi(y_1,\dots,y_m)=\arg\max_y\sum_{i=1}^n\sigma(|y-y_i|),$$

[3] т. е. если $s'_1<s''_1,\dots,s'_m<s''_m$, то $\varphi(s'_1,\dots,s'_m)<\varphi(s''_1,\dots,s''_m)$.

где $\sigma(s)$ — быстро убывающая функция ($\sigma(0) = 1$, $\lim\limits_{s\to\infty} \sigma(s) = 0$).

В задаче классификации с функцией «большинства ответов» конструирование каждого ΣП-алгоритма, входящего в семейство, фактически осуществляется не более чем по половине обучающего набора прецедентов.

Условно корректные методы обучения и корректные операции по ответам. До сих пор речь шла о корректных алгоритмах на одном и том же наборе прецедентов $\mathcal{I}_0$. Теперь будем рассматривать корректные алгоритмы на разных подмножествах наборов прецедентов $\mathcal{I}_0 \subset \mathcal{I}$.

Определение. Метод обучения μ назовем *условно корректным* на $\mathcal{I}$, если для любого непротиворечивого набора прецедентов $\mathcal{I}_0 \subset \mathcal{I}$ можно построить алгоритм, корректный на нем.

Пусть $\{\mathcal{I}_j\}$ — некоторый набор обучающих наборов прецедентов, где $\mathcal{I}_j \subset \mathcal{I},\ j = 1, \ldots, m$.

Рассмотрим корректирующие операции, которые преобразуют наборы корректных алгоритмов $\{\mu(\mathcal{I}_j)\}$ в корректные алгоритмы на $\bigcup \mathcal{I}_j$. Это соответствует ситуации, когда некоторый набор алгоритмов уже построен и нужно как-то их «объединить».

Нетрудно привести простое достаточное условие, основанное на принципе голосования, для построения корректирующей операции по ответам алгоритмов из $\{\mu(\mathcal{I}_j)\}$. Для каждого прецедента $P \in \bigcup \mathcal{I}_j$ обозначим $\nu(P)$ — число элементов в множестве $\{j\colon P \in \mathcal{I}_j\}$.

Условие. Если метод μ — условно корректный и для каждого прецедента $\nu(P) > \alpha m$ ($\alpha > 0{,}5$, m — общее число алгоритмов), то число алгоритмов из $\{\mu(\mathcal{I}_j)\}$, которые дают правильный ответ для P, также $> \alpha m$.

Другими словами, ответ семейства алгоритмов соответствует выборочной *статистической моде* набора ответов алгоритмов, а соответствующая корректирующая операция по ответам алгоритмов соответствует функции $\Phi(y_1, \ldots, y_m)$, вычисляющей моду (точно или приближенно) набора значений $\{y_1, \ldots, y_m\}$.

Конструктивный метод обучения ΣП-алгоритмов является условно корректным и, поэтому, он позволяет по любому поднабору, содержащему более половины прецедентов, строить корректный, на этом поднаборе ΣП-алгоритм. Таким образом, использование ΣП-алгоритмов в качестве базовых гарантирует построение семейства корректных алгоритмов на базе методов типа Boosting или Bagging.

Заключение. В рамках данной работы рассмотрен класс конструктивных алгоритмов (ΣП-алгоритмов, в частности), которые можно корректно обучить по любому поднабору непротиворечивого обучающего набора прецедентов. При этом важно, что конструктивный метод обучения позволяет строить значительный набор корректных алгоритмов на обучающем материале. Поэтому, отбирая среди них наилучшие с точки зрения внешнего критерия качества, получаем набор корректных алгоритмов с наилучшими внешними показателями, который затем улучшается при помощи *корректных операций*. Если для целей обучения всегда использовать больше половины обучающего материала, то получаем корректные алгоритмы, которые *всегда* корректны не менее чем на половине общего набора прецедентов. Таким мы получаем гарантировано хорошие алгоритмы для использования их в качестве базовых в композиционных алгоритмах типа Boosting, Bagging и т. д.

За рамками данной работы остался эмпирический анализ корректных семейств ΣП-алгоритмов, оценки обобщающей способности и сложности ΣП-алгоритмов. Хотя успешность применения корректных конструктивных семейств ΣП-алгоритмов не вызывает сомнений с теоретической точки зрения, необходимо еще провести экспериментальные исследования на модельных и реальных задачах.

Литература

[1] *Журавлев Ю. И.* Избранные труды. — М.: Магистр, 1998.

[2] Корректные алгебры над множествами некорректных (эвристических) алгоритмов // Кибернетика. Ч. I. 1977. № 4. С. 5–17; Ч. II. 1977. № 6. С. 21–27; Ч. III. 1978. № 2. С. 35–43.

[3] *Матросов В. Л.* Синтез оптимальных алгоритмов в алгебраических замыканиях моделей алгоритмов распознавания // Распознавание, классификация, прогноз, Москва: Наука, 1989. — С. 149–176.

[4] *Шибзухов З. М.* Конструктивные методы обучения ΣП-нейронных сетей. — М: МАИК Наука, 2006.

[5] *Воронцов К. В.* Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики, М.: Физматлит, 2004. — Т. 13. — С. 5–36.

[6] *Журавлев Ю. И.* Непараметрические задачи распознавания образов // Кибернетика. — 1976. — № 6.

[7] *Шибзухов З. М.* Об одном конструктивном подходе к построению обобщенных алгебраических ΣП-нейронов в одном абстрактном классе алгебр // Математические методы распознавания образов (ММРО–14), М.: Макс-Пресс. 2009. — С. 70-76.

[8] *Шибзухов З. М.* О некоторых конструктивных и корректных классах алгебраических ΣП-алгоритмов // Доклады РАН, 2010. — Т. 432,№ 4.

Повышение качества комбинированного обучения нейронных сетей*

Цой Ю. Р.
yurytsoy@gmail.com
Томск, Томский политехнический университет
Томск, Томский университет систем управления и радиоэлектроники

В данной статье описывается применение комбинированного подхода к обучению искусственной нейронной сети (ИНС), объединяющего нейроэволюционное преобразование пространства признаков с последующим «градиентным» обучением ИНС, получающей на вход модифицированные описания объектов. Результаты показывают, что в качестве целевой функции для нейроэволюционного преобразования признаков возможно использование достаточно общего критерия, не зависящего от задачи и требующего минимизации недиагональных элементов матрицы Грамма для нормализованных векторов выходных сигналов. Использование такого критерия частично избавляет от проблемы поиска оптимальных параметров алгоритмов обучения. Предлагаются возможные направления будущих исследований, связанные с развитием предлагаемого подхода и реализацией инкрементного обучения.

При решении задач классификации для данного множества описаний $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\},\ \mathbf{x}_i \in R^n$ объектов часто априори неизвестно, каким образом их необходимо обработать, чтобы повысить качество классификации. В частности, неизвестно, нужно ли увеличивать или уменьшать размерность пространства поиска. Оба варианта имеют свои сильные и слабые стороны, и могут оказаться полезными в различных ситуациях. Например, широко распространены методы уменьшения размерности пространства признаков [1, 2]. С другой стороны, часто рассматривается увеличение размерности пространства признаков, направленное на такое преобразование описаний объектов, которое позволяет с большей вероятностью построить дискриминантную функцию, отделяющую описания объектов из разных классов [6].

Поскольку оптимальное преобразование пространства признаков заранее неизвестно, то разработка методов поиска такого преобразования является актуальной задачей. При этом критерии, которым должно подчиняться искомое преобразование также могут различаться. Например, при разложении исходных описаний по ортогональным функциям таким критерием может выступать минимизация отличий между исходным и полученным описаниям, а при использовании вероятностного подхода к классификации — максимизация вероятности порождения множества $\mathbf{X}$. В данной статье рассматривается применение нейроэволюционного подхода для преобразования пространства признаков с использованием достаточно общего критерия, требующего линейной независимости признаков в модифицированном описании, с последующим обучением искусственной нейронной сети (ИНС).

Исследование поддержано Российским фондом фундаментальных исследований (проекты № 09-08-00309-а, 11-07-00027-а).

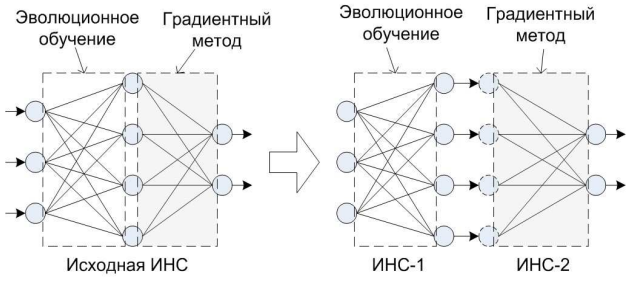


Рис. 1. Общая схема разделения ИНС для «комбинированного» обучения. Входные нейроны ИНС-2 имеют функцию активации $y(x) = x$.

Комбинированное обучение нейронной сети

В рамках комбинированного подхода к обучению ИНС предлагается разделение ИНС на две части (рис. 1) и обучение ИНС-1, отвечающей за преобразование описаний объектов, с использованием эволюционного подхода, а ИНС-2 — с применением традиционного градиентного метода. В данном исследовании рассматриваются ИНС с многослойной структурой, однако подобный подход применим и к ИНС с произвольной структурой связей.

Обученные ИНС-1 и ИНС-2 можно объединить в одну сеть таким образом, что выходные сигналы ИНС-1 будут являться входными для ИНС-2. Процедура комбинированного обучения ИНС представлена алгоритмом 1 (операция Net1 + Net2 обозначает объединение двух ИНС в одну).

Изменяя количество выходных нейронов ИНС-1 и их функцию активации можно изменять свойства преобразования признаков, осуществляемого ИНС-1. Будем обозначать количество выходных сигналов ИНС-1 как αn, где $\alpha \in R$ — некоторая константа, n — количество входных нейронов ИНС-1 (размерность исходных описаний объектов).

Алгоритм 1. Алгоритм комбинированного обучения ИНС.

Вход: Обучающее
$D_{\text{train}} = \{(\mathbf{X}_{t,0}, \mathbf{Y}_{t,0}), \ldots, (\mathbf{X}_{t,n1}, \mathbf{Y}_{t,n1})\}$
и проверочное
$D_{\text{validation}} = \{(\mathbf{X}_{v,0}, \mathbf{Y}_{v,0}), \ldots, (\mathbf{X}_{v,n2}, \mathbf{Y}_{v,n2})\}$ множества,
параметры ИНС-1 и ИНС-2,
параметры алгоритмов обучения ИНС-1 и ИНС-2,
количество K обучаемых ИНС-2;

1: Эволюционное обучение ИНС-1 $Net1$ с использованием D_{train};
2: Формирование обучающего множества для ИНС-2:

$$D'_{\text{train}} = \{(\mathbf{Y}'_{t,0}, \mathbf{Y}_{t,0}), \ldots, (\mathbf{Y}'_{t,n1}, \mathbf{Y}_{t,n1})\},$$

где $\mathbf{Y}'_{t,i}$ — выходной сигнал $Net1$ при подаче на вход вектора $(\mathbf{X}_{t,i})$;
3: Обучение K ИНС-2 с использованием D'_{train}: $\{\text{Net2}_1, \ldots, \text{Net2}_K\}$ и с применением градиентного алгоритма;
4: Выбор среди множества ИНС

$$\{\text{Net1} + \text{Net2}_i,\ i = 1, \ldots, K\}$$

лучшей на основании ошибки на проверочном множестве $D_{validation}$;
5: Вернуть лучшую найденную ИНС.

Критерий оценки нейроэволюционного алгоритма

Можно предложить различные критерии обучения ИНС-1. Например, в [3] рассматривается ошибка пробного обучения ИНС-2, а в [4] в качестве критерия выступает корреляция между выходными сигналами ИНС-1. Первый способ позволяет в явном виде оценить качество преобразования признаков, осуществляемое ИНС-1, однако он существенно сложнее с вычислительной точки зрения, так как на каждом поколении эволюционного поиска параметров ИНС-1 требуется $O(10^3 - 10^4)$ эпох обучения ИНС-2. В [4] было показано, что можно применять вычислительно более простой корреляционный критерий к оценке ИНС-1

$$f = \frac{2\sum_{i,j>i} |R_{Y_i,Y_j}|}{N(N-1)} \to \min, \qquad (1)$$

где R_{Y_i,Y_j} — коэффициент корреляции между i-м и j-м выходными сигналами. При этом в ряде случаев была получена более высокая точность распознавания на тестовом множестве. Однако использование критерия (1) требует тщательного подбора параметров алгоритмов обучения и структуры ИНС-1 и ИНС-2 [4], что затрудняет его использование. Это может быть вызвано недостатками преобразования, осуществляемого ИНС-1, в частности, тем, что модифицированные описания могут располагаться близко друг к другу.

Поскольку коэффициент корреляции равен 0, если равна нулю ковариация, то рассмотрим ковариацию двух векторов $\mathbf{Y}_1 = (y_{1,1}, y_{1,2}, \ldots, y_{1,N})^\intercal$ и $\mathbf{Y}_2 = (y_{1,1}, y_{1,2}, \ldots, y_{1,N})^\intercal$

$$Cov(\mathbf{Y}_1, \mathbf{Y}_2) = \frac{1}{N}(\mathbf{Y}_1 - \bar{Y}_1\mathbf{1})^\intercal(\mathbf{Y}_2 - \bar{Y}_2\mathbf{1}),$$

где $\bar{Y}_i$ — среднее значение компонент вектора $\mathbf{Y}_i$, $\mathbf{1}$ — вектор единиц. Раскрывая скобки, после упрощения получим:

$$NCov(\mathbf{Y}_1, \mathbf{Y}_2) = \mathbf{Y}_1^\intercal\mathbf{Y}_2 + r,$$

где $r = -N\bar{Y}_1\bar{Y}_2$. Отсюда следует, что ковариация двух ортогональных векторов обращается в нуль, только если среднее значение компонент хотя бы одного из векторов равно 0.

Учитывая, что многие распространенные функции активации (лог-сигмоидная, Гауссиан, единичного скачка) неотрицательны, критерий (1) может быть удовлетворен и в том случае, если выходные сигналы ИНС-1 не ортогональны. Это может обозначать, что модифицированные описания объектов занимают в пространстве признаков объем, меньший максимально возможного, что может уменьшить различие между описаниями объектов из разных классов. Для исправления этого недостатка предлагается использовать следующий критерий

$$f = \frac{2\sum_{i,j>i} G(i,j)}{\alpha n(\alpha n - 1)} + \pi \to \min, \qquad (2)$$

где $G(i,j)$ — элемент матрицы Грамма для i-го и j-го нормализованных выходных сигналов ИНС-1, $\hat{\mathbf{Y}}_i = \mathbf{Y}_i/|\mathbf{Y}_i|$, π — количество нулевых элементов на главной диагонали в матрице Грамма. Нормализация выходных сигналов нужна, чтобы избежать преимуществ у ИНС-1 с малыми значениями выходных сигналов.

Условия экспериментов

Для проверки комбинированного обучения с критерием (2) будем использовать задачи из набора Proben1 (cancer1, card1, diabetes1, glass1, heart1, horse1) [5]. Алгоритм для обучения ИНС-2 — RPROP[1]. Комбинированное обучение ИНС осуществляется с применением библиотеки Mental Alchemy[2].

Для всех запусков количество поколений обучения ИНС-1 равняется 50; размер популяции — 100

[1] Для реализации алгоритма RPROP использовалась библиотека Encog: `http://www.heatonresearch.com/encog`

[2] `http://code.google.com/p/mentalalchemy/`

Рис. 2. Результаты для задачи cancer1.

Рис. 3. Результаты для задачи card1.

особей; рассматривается турнирная селекция с турниром из 5 особей, одна элитная особь; вещественное кодирование. Количество эпох для обучения ИНС-2 — 100. Будем рассматривать 3 функции активации выходных нейронов ИНС-1:

- линейная $y = S$, где $S = \mathbf{w}^{\mathsf{T}}\mathbf{x}$ — взвешенная сумма входных сигналов $\mathbf{x}$, $\mathbf{w}$ — вектор весов входных сигналов нейрона;
- лог-сигмоидная $y = (1 + \exp(-aS))^{-1}$, $a \in R$;
- гауссова $y = \exp(-|\mathbf{w} - \mathbf{x}|^2)$.

Параметр α, определяющий количество выходных сигналов ИНС-1, изменяется в диапазоне $[0{,}5, \ldots, 3]$ с шагом 0,5. Для каждой ИНС-1 обучаются 10 ИНС-2. По 10 запускам комбинированного обучения выбирается ИНС дающая наименьшую ошибку классификации на тестовом множестве.

Результаты экспериментов

Средния значения ошибок классификации для различных алгоритмов обучения комбинированной ИНС и сравнение с лучшими результатами из [5] и [4] представлено на рис. 2–7.

Из приведенных данных видно, что в задаче card1 комбинированное обучение с применением критерия (2) уступает результатам из [4], однако необходимо отметить, что при использовании корреляционного криктерия (1) процесс поиска параметров алгоритма обучения занимает много времени. В случае использования критерия (2) удалось

Рис. 4. Результаты для задачи diabetes1.

Рис. 5. Результаты для задачи glass1.

Рис. 6. Результаты для задачи heart1.

Рис. 7. Результаты для задачи horse1.

получить сопоставимые результаты без трудоемкого подбора оптимального количества поколений и эпох обучения ИНС-1 и ИНС-2. Настраивались

только количество выходных сигналов ИНС-1 и их функции активации.

Заключение

В статье предлагается использование критерия, основанного на минимизации недиагональных элементов матрицы Грамма, для обучения ИНС-1 при комбинированном подходе к обучению. Результаты экспериментов показывают, что во многих случаях удается достичь сравнительно высокой точности распознавания без необходимости тщательного подбора параметров алгоритмов обучения ИНС-1 и ИНС-2. При этом получаемые результаты превосходят таковые для традиционного подхода с градиентным алгоритмом обучения ИНС.

Дальнейшие исследования могут проводиться по следующим направлениям:

— Адаптация эволюционного алгоритма обучения ИНС-1 для автоматического поиска количества выходных нейронов и их функций активации.

— Организация инкрементного обучения, при котором обучение ИНС-1 и ИНС-2 чередуется. Например, после обучения на протяжении t_1 поколений ИНС-1 производится обучение ИНС-2 в течение t_2 эпох, так же как описано в данной статье. После этого для имеющейся ИНС-2 снова осуществляется эволюционный поиск ИНС-1 (дообучение ИНС-1) на t_1 поколений, которая позволяет повысить точность классификации ИНС-2. Затем ИНС-2 снова дообучается. Данный процесс может повторяться до тех пор, пока не наступит условие останова (например, заданное количество таких итераций, либо проверка по ранней остановке (early stopping)). В дальнейшем процесс дообучения можно будет запустить снова. Благодаря подстройке преобразования входных признаков, путем адаптации ИНС-1 теоретически возможна адаптация к изменениям в обучающих данных, однако этот аспект требует отдельного изучения.

Литература

[1] *Gorban A., Kegl B., Wunsch D., Zinovyev A. (Eds.)* Principal Manifolds for Data Visualisation and Dimension Reduction, LNCSE 58, Berlin–Heidelberg–New York: Springer, 2007.

[2] *Vafaie H., De Jong K.* Genetic Algorithms as a Tool for Feature Selection in Machine Learning // Proc. of the Fourth Int. Conf. on Tools with Artificial Intelligence 1992 (TAI'92), IEEE Press, 1992. — Pp. 200–203.

[3] *Цой Ю. Р.* Об адаптивном увеличении размерности пространства признаков. // 12-я Национальная конференция по искусственному интеллекту с международным участием: Труды, Тверь, 20–24 сентября 2010, Москва: Физматлит, 2010. — Т. 4. — С. 134–140.

[4] *Цой Ю. Р.* Нейроэволюционное преобразование пространства признаков в задаче нейросетевой классификации // VI-я Международная научно-практическая конференция «Интегрированные модели и мягкие вычисления в искусственном интеллекте» (16-19 мая 2011 г., Коломна), 2011. — В печати.

[5] *Prechelt L.* PROBEN1 — a set of neural network benchmark problems and benchmarking rules. Technical Report 21/94. — Fakultat fur Informatik, Universitat Karlsruhe. — Karlsruhe, Germany, 1994.

[6] *Cover T. M.* Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition // IEEE Transactions on Electronic Computers. — 1965. — V. 14, N. 3. — Pp. 326–334.

Морфологический подход к синтезу метрических классификаторов и его реализация методом отыскания минимального разреза графа соседства для обучающей выборки*

Визильтер Ю. В., Горбацевич В. С.
viz@gosniias.ru, gvs@gosniias.ru

Предложен и обоснован морфологический подход к синтезу классификаторов в задаче обучения с учителем, предполагающий решение двух последовательных подзадач — переразметки обучающей выборки и оптимальной корректной интерполяции (расширения) решающего правила. Доказана проективность морфологических операторов машинного обучения. Показано, что классы морфологических классификаторов и алгоритмов обучения образуют систему вложенных классов возрастающей сложности в смысле Пытьева и могут быть использованы для структурной минимизации риска переобучения. Показано, что для случая двухклассовой классификации метод минимального разреза графов позволяет отыскивать глобальный оптимум введенного критерия качества классификации.

Введение

Непосредственным толчком к разработке предлагаемого подхода послужило изучение работы [1], в которой (как и в [2]) распознающие алгоритмы (классификаторы) анализируются, будучи представлены лишь векторами решений на объектах обучающей выборки. Для перехода от анализа к синтезу необходимо дополнительно опереться на *принцип компактности* [3, 4], заключающийся в том, что соседние объекты выборки должны с большей вероятностью принадлежать к одному классу. Основанные на этом предположении алгоритмы будем называть *метрическими классификаторами*. Предлагается рассматривать задачу синтеза метрического классификатора как задачу *оптимальной сегментации* (labeling) точек обучающей выборки, а «форму» и «сложность» классификаторов трактовать в терминах «формы» и «сложности» изображений (образованных метками классов на точках выборки), т. е. в терминах *математической морфологии*.

Основными источниками используемых далее морфологических конструкций и идей являются: *теория форм* М. Павель [5], *математическая морфология* Серра [6], *теория морфологического анализа* Пытьева [7], а также *критериальная проективная морфология* [8]. Для алгоритмической реализации процедур синтеза метрических классификаторов предлагается использовать технику построения минимальных разрезов графов [9–13], применяя ее к графам соседства элементов обучающей выборки.

Задача обучения с учителем

Пусть даны *пространство объектов* $\mathcal{A}$, конечное *множество классов* $C = \{c_1, \ldots, c_l\}$, и известно разбиение объектов по классам $c_{\mathcal{A}}(a)\colon a \in \mathcal{A} \mapsto c \in C$. Обозначение $c_{\mathcal{A}}$ указывает на то, что функция определена на $\mathcal{A}$.

Работа выполнена при финансовой поддержке РФФИ, проекты № 11-08-01114-а, 11-08-01039-а.

Производится описание объектов из $\mathcal{A}$ дескрипторами из *пространства описаний* (*признаков*) $\mathcal{X}$: $x_{\mathcal{A}}(a)\colon a \in \mathcal{A} \mapsto x \in \mathcal{X}$. Случайным образом формируется конечная *выборка объектов* $A \subseteq \mathcal{A}$, $\|A\| < +\infty$ и соответствующая *выборка описаний* $X \subseteq \mathcal{X}$, $\|X\| < +\infty$. Каждому значению x ставится в соответствие класс c породившего его объекта $c_X(x)\colon x_{\mathcal{A}}(a) \in X \mapsto c_{\mathcal{A}}(a) \in C$. По *обучающей выборке* c_X требуется построить такой *распознающий алгоритм* или *классификатор* $f_{\mathcal{X}}(x)\colon x \in \mathcal{X} \mapsto c \in C$, который обеспечивает *наилучшее разбиение* $\mathcal{X}$ на классы из C.

«Наилучшее разбиение» определим при помощи *тестовой выборки* $c'_Y(x)\colon x \in Y \mapsto c \in C$, $Y \subseteq \mathcal{X}$, $Y \cap X = \varnothing$, $\|X\| < +\infty$, и *критерия эмпирического риска* на выборке Y:

$$J_Y(f_{\mathcal{X}}) = d_H(f_Y, c'_Y)/\|Y\|\,,$$

$$d_H(f_Y, c'_Y) = \sum_{x \in Y} 1(f(x) \neq c'(x))\,,$$

где $1(true) = 1$, $1(false) = 0$, $\|Y\| = \sum\limits_{x \in Y} 1$.

Здесь *расстояние Хэмминга* d_H имеет смысл *числа ошибок классификации* на тестовой выборке Y. Отсюда *критерий среднего ожидаемого эмпирического риска* имеет вид

$$J_{\mathcal{X}}(f_{\mathcal{X}}) = \mathsf{E}_{Y \subseteq \mathcal{X}}\{J_Y(f_{\mathcal{X}})\},$$

где $\mathsf{E}_{Y \subseteq \mathcal{X}}\{.\}$ — математическое ожидание по всем возможным выборкам $Y \subseteq \mathcal{X}$.

Таким образом, задача построения *оператора оптимального синтеза* θ, заключается в минимизации критерия $J_{\mathcal{X}}(f_{\mathcal{X}}) = J_{\mathcal{X}}(\theta_{C_X})$:

$$\theta\colon c_X \in \Omega_X \mapsto f_{\mathcal{X}} \in \Omega_{\mathcal{X}},\ \theta\colon \arg\min_{\theta'}\{J_{\mathcal{X}}(\theta'_{C_X})\}. \quad (1)$$

Здесь Ω_X и $\Omega_{\mathcal{X}}$ — *множества всех возможных разбиений* X и $\mathcal{X}$ по классам из C.

Как правило, от *задачи синтеза* (1) сразу переходят к *задаче обучения* классификаторов задан-

ного класса при помощи *обучающего правила* известного типа:

$$\theta \in \Theta\colon c_X \in \Omega_X \mapsto f_{\mathcal{X}} \in F_{\mathcal{X}} \subseteq \Omega_{\mathcal{X}},$$

$$\theta = \arg\min_{\theta' \in \Theta}\{J_Y(\theta' C_X)\}, \qquad (2)$$

где $F_{\mathcal{X}}$ — класс классификаторов, Θ — класс алгоритмов обучения для $F_{\mathcal{X}}$ на выборках $X \subseteq \mathcal{X}$.

Кроме того, вместо недоступного критерия $J_{\mathcal{X}}(f_{\mathcal{X}})$, на практике используется критерий *наблюдаемого эмпирического риска* $J_X(\theta_{c_X})$, который имеет глобальный минимум при $f_X \equiv c_X$, непригодный для неизвестной тестовой выборки Y. Этой проблеме посвящена *теория оценки и контроля переобучения* [2]. Здесь риск оценивается по обучающей выборке, но сложность решающего правила искусственно ограничивается. Для этого вводится понятие *сложности классификатора* $Q(f_{\mathcal{X}})$, а точнее *сложности класса классификаторов* $Q(F_{\mathcal{X}})$. Соответственно вместо (2) решается *задача минимизации наблюдаемого риска с регуляризацией по сложности класса обучаемого классификатора*:

$$\theta \in \Theta\colon c_X \in \Omega_X \mapsto f_{\mathcal{X}} \in F_{\mathcal{X}} \subseteq \Omega_{\mathcal{X}},$$
$$\theta = \arg\min_{\theta' \in \Theta}\{J_X(\theta' C_X) + \alpha Q(F_{\mathcal{X}})\}, \qquad (3)$$

где $\alpha \geqslant 0$ — *параметр регуляризации.*

Решение задачи (3) сводится к следующему («*метод структурной минимизации риска*»). Пусть в рамках некоторого суперкласса $\mathbf{F}_{\mathcal{X}}$ определена последовательность *вложенных классов классификаторов нарастающей сложности*:

$$F^0_{\mathcal{X}} \subseteq F^1_{\mathcal{X}} \subseteq \ldots \subseteq F^j_{\mathcal{X}} \subseteq \mathbf{F}_{\mathcal{X}} \subseteq \Omega_{\mathcal{X}}\colon$$
$$Q(F^0_{\mathcal{X}}) \leqslant Q(F^1_{\mathcal{X}}) \leqslant \ldots \leqslant Q(F^j_{\mathcal{X}}) \leqslant \ldots. \qquad (4)$$

Тогда задача (3) последовательно решается для $F^j_{\mathcal{X}}$, $j = 0, 1, 2, \ldots$, пока значения критерия не перестанут улучшаться. Значение α подбирается *методом кросс-валидации с валидационной выборкой* $Z \subseteq \mathcal{X}$, $Y \cap X = \varnothing$, $\|Z\| < +\infty$.

Морфологический подход к синтезу классификаторов

Рассмотрим поход к машинному обучению, направленный непосредственно на решение задачи (1). Решение (1) предлагается отыскивать в виде композиции решений подзадач:

$$\theta_\alpha = \delta_\alpha \psi_\alpha\,, \qquad (5)$$

где ψ_α — оператор (процедура) *синтеза оптимального отклика классификатора на обучающей выборке X с учетом его сложности (локальной некомпактности)*

$$\psi_\alpha\colon c_X \in \Omega_X \mapsto f_X \in \Omega_X,$$
$$\psi_\alpha = \arg\min_{\psi'}\{J_X(\psi' c_X) + \alpha Q_X(\psi' c_X)\}; \qquad (6)$$

δ_α — оператор (процедура) *оптимальной корректной интерполяции (расширения)* классификатора f_X на $\mathcal{X}$ *с учетом сложности получаемого классификатора* $f_{\mathcal{X}}$:

$$\delta_\alpha\colon f_X \in \Omega_X \mapsto f_{\mathcal{X}} \in \Omega_{\mathcal{X}},$$
$$\delta_\alpha = \arg\min_{\delta'}\{J_{NN}(\delta' f_X) + \beta Q(\delta' f_X)\}; \qquad (7)$$

Здесь

$$J_{NN}(\delta f_X) = \begin{cases} +\infty, \text{если } \exists x \in X\colon \delta f_X(X) \neq f_X(X)\,; \\ d_H(\delta f_X(X)), \delta^{NN} f_X(X)) \text{ — иначе,} \end{cases}$$

где d_H — расстояние Хэмминга; δ^{NN} — оператор интерполяции, соответствующий *правилу ближайшего соседа* (Nearest Neighbor).

Как решается в рамках такого подхода проблема переобучения? Ответ связан с областью *критериальной морфологии* [8], где доказано следующее

Утверждение 1. *Если* ψ_α *— морфологический фильтр*

$$\psi_\alpha A = \arg\min_B\{J(A, B) + \alpha Q(B)\};$$

где A — исходный образ, B — выходной образ, $J(A, B)$ — критерий ошибки аппроксимации, $Q(B)$ — критерий сложности образа B, $\alpha \geqslant 0$ — параметр регуляризации, и критерий $J(A, B)$ обладает свойствами метрики, то

1) *оператор ψ_α является проектором: $\psi^2_\alpha = \psi_\alpha$;*
2) *проектор ψ_α определяет систему вложенных классов, монотонную по параметру α: $\alpha \geqslant \beta \Rightarrow \psi_\beta \psi_\alpha = \psi_\alpha$.*

Поскольку в задаче (1) функционал J имеет вид расстояния Хэмминга, из Утверждения 1 следует, что θ_α (5) является *алгебраическим проектором*:

$$\theta^2_\alpha = \theta_\alpha \Rightarrow \forall x \in X\colon \theta_\alpha f_X(x) = f_X(x). \qquad (8)$$

Кроме того, из Утверждения 1 следует, что на основе θ_α образуется система вложенных классов решающих правил, монотонная относительно α:

$$\forall \alpha \geqslant \beta \Rightarrow F^\alpha_{\mathcal{X}} \subseteq F^\beta_{\mathcal{X}}\colon Q(F^\alpha_{\mathcal{X}}) \leqslant Q(F^\beta_{\mathcal{X}}), \qquad (9)$$

где $F^\alpha_{\mathcal{X}} = \{f_X(x)\colon \theta_\alpha f_X(x) = f_X(x)\}$ — множество классификаторов (разбиений), стабильное относительно проектора θ_α. В морфологиях изображений такая система вложенных проективных классов рассматривается как множество Пытьевских «форм» нарастающей сложности. В задаче синтеза классификаторов последовательность «форм»

может быть использована для решения проблемы переобучения методом минимизации структурного риска. Необходимо лишь определить $Q_X(f_X)$ как критерий компактности.

Оценка компактности и минимизация сложности классификаторов

Для каждого $x \in X \subseteq \mathcal{X}$ определим систему вложенных окрестностей $O_k(x) \subseteq X$, $k = 1, \ldots, ||X|| - 1$, состоящих из k ближайших соседей. Принцип компактности предполагает, что близкие соседи должны с большей вероятностью принадлежать к одному классу. Введем локальную меру k-некомпактности f_X:

$$\left.\begin{aligned} Q_k(x, f_X) &= q_H(O_k(x))/||O_k(x)||; \\ q_H(O_k(x)) &= \Sigma_{y \in O_k(x)} 1(f_X(x) \neq f_X(y)) \end{aligned}\right\} \quad (10)$$

и глобальную меру k-некомпактности f_X:

$$\left.\begin{aligned} Q_X^k(f_X) &= Q_H(X, f_X)/||X||; \\ Q_H(X, f_X) &= \Sigma_{x \in X} Q_k(x, f_X). \end{aligned}\right\} \quad (11)$$

Значение $Q_X^k(f_X)$ (11) характеризует эмпирическую оценку вероятности того, что один из k ближайших соседей в разбиении $f_X(x)$ будет отнесен к другому классу. При любых фиксированных k и X усложнению классификатора f_X соответствует нарастание меры k-некомпактности $Q_X^k(f_X)$. Кроме того, как показано в [11], при увеличении параметра k в (11) преимущество получают более простые и «гладкие» разделяющие поверхности.

С учетом (11) задача (6) сводится к известной задаче оценивания параметров скрытой Марковской модели [14], для которой существует эффективное приближенное решение методом нахождения максимального потока / минимального разреза графа. Для случая двух классов метод разрезания графа может давать точное оптимальное решение. Алгоритм нахождения минимального разреза на графе с двумя терминальными вершинами позволяет минимизировать функционал энергии вида:

$$E(T) = E_0 + \sum_{i=1,\ldots,N} E_i(t_i) + \sum_{(i,j) \in V} E_{ij}(t_i, t_j), \quad (12)$$

где N — число нетерминальных вершин графа; $T = \langle t_1, \ldots, t_N \rangle, t_1, \ldots, t_N \in \{0, 1\}$ — метки ассоциирования нетерминальных вершин с терминальными; $E_i(0), E_i(1) \in \{0, 1\}$ — унарные потенциалы; $E_{ij}(t_i, t_j)$ — парные потенциалы $E_{ij}(0, 0)$, $E_{ij}(0, 1)$, $E_{ij}(1, 0)$, $E_{ij}(1, 1)$; V — подмножество пар индексов, задающее соседство на T.

Энергия (12) *субмодулярна* [13], если $\forall (i, j) \in V$:

$$E_{ij}(0, 0) + E_{ij}(1, 1) \leqslant E_{ij}(0, 1) + E_{ij}(1, 0). \quad (13)$$

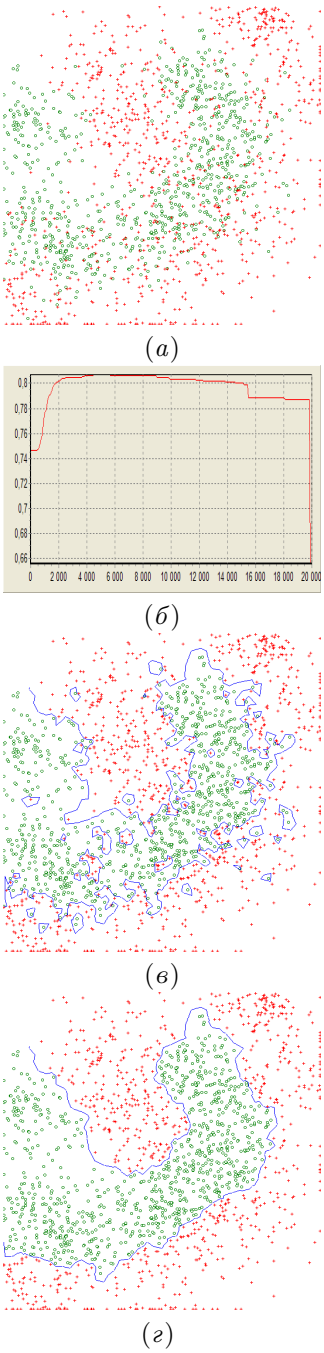
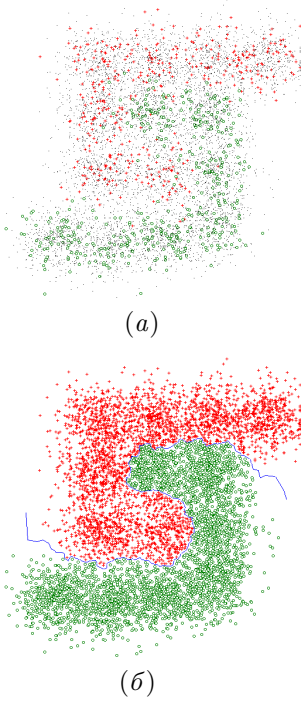


Рис. 1. Пример морфологического обучения: (*а*) обучающая выборка; (*б*) зависимость вероятности распознавания от параметра α; (*в*) переразметка выборки при $\alpha = 1000$ (переобучение); (*г*) $\alpha = 4500$ (оптимум)

Для субмодулярной энергии (12)–(13) метод минимального разрезания графа [9, 12] гарантирует нахождение точного минимума [10, 13].

Для задачи синтеза двухклассового классификатора (6), (11) примем: $C = \{0, 1\}$, $N = ||X||$, $T = \langle f_X(x_1), \ldots, f_X(x_N) \rangle$, $E_i(x) = 1(f_X(x) \neq c_X(x))$, $E_{ij}(t_i, t_j) = 1(f_X(x_i) \neq f_X(x_j))$, $V = (i, j) : j \in O_k(x_i)$. Легко убедиться, что соответствующая энергия (12) будет субмодулярной, а зна-



(*а*)

(*б*)

Рис. 2. Пример обучения на частично (на 10%) размеченной обучающей выборке: (*а*) обучающая выборка; (*б*) результат переразметки выборки

чит, метод действительно порождает α-семейства проекторов из Утверждения 1.

Моделирование процедур морфологического синтеза классификаторов

Для наглядности моделировались двумерные данные. Выборка формировалась на основе смеси гауссовых распределений. При построении графа соседства использовался алгоритм триангуляции Делоне с динамическим кэшированием [15]. Разрезы графов вычислялись при помощи библиотеки [16], что обеспечило время обучения не более 1 сек. На рис. 1. показан пример морфологического обучения с учителем. Обучающая выборка содержала 1000 объектов двух классов. Приведена полученная методом Монте–Карло зависимость вероятности правильного распознавания от параметра α. На рис. 2 показан пример обучения по частично размеченной выборке — т. н. semi-supervised learning (размечено 10% из 10 000 объектов).

Заключение

Предложенный подход к машинному обучению назван «*морфологическим*» в силу его алгоритмического и методического соответствия известному морфологическому подходу к анализу изображений. Дальнейшие работы в рамках предложенного подхода должны быть связаны с исследованием его поведения на многомерных модельных и реальных данных, связанных с задачами распознавания сложных образов.

Литература

[1] *Воронцов К. В.* Комбинаторная теория надёжности обучения по прецедентам. Диссертация на соискание ученой степени доктора физико-математических наук. ВЦ им. А. А. Дородницына РАН, Москва, 2010.

[2] *Вапник В. Н.* Восстановление зависимостей по эмпирическим данным. — М.: Наука, 1979.

[3] *Айзерман М. А., Браверман Э. М., Розоноэр Л. И.* Метод потенциальных функций в теории обучения машин. — М.: Наука, 1970 — 314 с.

[4] *Хачай М. Ю.* Топологический подход к выводу условий равномерной по классу событий сходимости частот к вероятностям // Интеллектуализация обработки информации: 8-я международная конференция (ИОИ-8), Кипр, г. Пафос, 2010 г.: Сборник докладов, Москва: МАКС Пресс, 2010, С. 91–94.

[5] *Pavel M.* Fundamentals of Pattern Recognition // Marcel Dekker. Inc., New York, 1989.

[6] *Serra J.* Image Analysis and Mathematical Morphology // Academic Press, London, 1982.

[7] *Пытьев Ю. П., Чуличков А. И.* Методы морфологического анализа изображений. М.: Физматлит, 2010 — 336 с.

[8] *Визильтер Ю. В.* Обобщенная проективная морфология. // Компьютерная оптика. — 2008. — Т. 32, № 4. — С. 384–399

[9] *Ford L., Fulkerson D.* Flows in Networks // Princeton University Press, 1962.

[10] *Greig D., Porteous B., Seheult A.* Exact maximum a posteriori estimation for binary images // Journal of the Royal Statistical Society 1989. — Vol. 51, No. 2. — Pp. 271–279.

[11] *Boykov Y., Kolmogorov V.* Computing geodesics and minimal surfaces via graph cuts // In Proc. IEEE International Conf. Computer Vision (ICCV) 2003. — Pp. 26–33.

[12] *Boykov Y., Kolmogorov V.* An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision // IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI) 2004. — Vol. 26, No. 9. — Pp. 1124–1137.

[13] *Kolmogorov V., Zabih R.* What energy functions can be minimized via graph cuts? // IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI) 2004. — Vol. 26, No. 2. — Pp. 147–159.

[14] *Geman S., Geman D.* Stochastic relaxation, Gibbs distributions, the Bayesian restoration of images // IEEE Trans. Pattern Analysis and Machine Intelligence 1984. — No. 6. — Pp. 721–741.

[15] *Скворцов А. В.* Обзор алгоритмов построения триангуляции Делоне // Вычислительные методы и программирование 2002. — Т. 3. — С. 14–39.

[16] *Boykov Y., Kolmogorov V.* MAXFLOW — software for computing mincut/maxflow in a graph. V. 3.01. — `http://www.cs.ucl.ac.uk/staff/V.Kolmogorov/software.html`.

Использование FRiS-функции при решении задачи распознования состояний объектов в функционально-топической диагностике*

Борисова И. А.
biamia@mail.ru
Новосибирск, Институт математики им. С. Л. Соболева СО РАН

В классической постановке задачи распознавания предполагается, что элементы обучающей выборки выбираются из генеральной совокупности независимо и случайно. Однако на практике встречаются ситуации, когда между элементами выборки существует связь и использование информации необходимо для успешного решения задачи распознавания. Примером такого рода задачи может выступать задача распознавания состояний объектов, которая особенно часто возникает при проведении различных медицинских, биологических и физиологических исследований. В данной работе демонстрируются возможности функции конкурентного сходства (FRiS-функции) для решения таких задач в процессе исследования постстимульной ритмической активности коры головного мозга при раздражении различных анализаторных систем.

При решении задачи распознавания предполагается, что элементы обучающей выборки выбираются из генеральной совокупности независимо и случайно. Классические алгоритмы классификации опираются на это предположение и обрабатывают элементы обучающей выборки единообразно, на первом этапе считая все признаки в исходной системе описания элементов равноправными. Исключение составляет лишь целевой признак. Однако, на практике встречаются ситуации, когда между элементами выборки существует связь и использование информации об этой связи является решающим фактором при решении задачи распознавания. Примером такого рода задачи может выступать задача распознавания состояний объектов. В этой задаче каждый анализируемый объект может находиться в нескольких различных состояниях. На основе обучающей выборки, содержащей описания ограниченного множества в этих состояниях, необходимо сформировать решающее правило, позволяющее различать исследуемые состояния, и выделить множество признаков, причинно-следственно связанных с переходом объектов в те или иные состояния.

Такая задача, в которой предположение о независимости выборки нарушается, может быть переформулированна в терминах марковских процессов, однако такое усложнение представляется нецелесообразным. С другой стороны эту задачу можно решать классическими методами распознавания образов, «забыв» о связи между элементами обучающей выборки. Но такое «огрубление» вероятно приведет к ухудшению качества получаемых решений. Именно поэтому актуальной становится задача создания технологии распознавания, учитывающей связи между элементами обучающей выборки и при наличии в выборке элементов, соответствующих разным состояниям одного объекта, позволяющей анализировать их с учетом особенностей этого объекта.

В качестве практического примера задачи распознавания состояний объектов, подтвержающего важность и актульность этой проблемы, может выступать задача исследования постстимульной ритмической активности коры головного мозга при раздражении различных анализаторных систем. Она возникает в процессе изучения функций центральной нервной системы с помощью программно-аппаратного комплекса для функционально-топической диагностики заболеваний внутренних органов [1]. В этих исследованиях ритмическая активность головного мозга измеряется в частотно-топической системе координат «Сегментарная матрица». При изучении специфичности постстимульной активности регистрируется реакция испытуемого на раздражение различных анализаторных систем и выявляются специфические участки спектра, активирующиеся этим раздражителем. Раздражители могут быть зрительными, вкусовыми, обонятельными, тактильными. Однако, из-за различной вегетатики, психотипа и прочих особенностей исходные матрицы испытуемых людей имеет очень разный спектральный рельеф, что затрудняет выделение общего спектрального рельефа, соответствующего раздражителю.

В данной статье предлагается несколько вариантов формальной постановки задачи распознавания состояний объектов, предлагаются простейшие подходы к решению таких задач, которые затем сравниваются на примере задачи распознавания многопериодического паттерна, характерного для визуального представления буквы «А». В качестве базового инструмента для решения задачи распознования состояний нами был выбран аппарат FRiS-функций [2], среди прочего успешно зарекомендовавший себя при решении задачи выбора информативных признаков и задачи построения решающего правила.

Работа выполнена при финансовой поддержке РФФИ, проект № 11-01-00156.

Формальная постановка задачи распознавания состояний объектов

Рассмотрим задачу распознавания K состояний L объектов в пространстве N характеристик в общем виде. В ней каждому элементу a обучающей выборки A соответствует некоторый объект i из множества $O = \{1, \ldots, L\}$ находящийся в некотором состояния j из множества $I = \{0, \ldots, K\}$, измеренный в пространстве характеристик $X = \{x_1, \ldots, x_N\}$. Объект i в состоянии j может встречаться в обучающей выборке несколько раз. При этом целесообразно бывает выделить 0-состояние, за которое принимается некоторое фоновое состояние, состояние без внешних раздражителей. В случае, если оно не задается отдельно, в качестве фонового состояния для объекта i можно использовать некоторое усреднение по всем активным состояниям для этого объекта:

$$a_i^0 = \frac{1}{|\{a : a \in A, O(a) = i\}|} \sum_{\substack{a \in A \\ O(a)=i}} a$$

Здесь $O(a)$ — имя объекта, соответствующего элементу a.

В наиболее сильной постановке задача состоит в выборе подпространства характеристик $Y \subseteq X$ и построении решающего правила $d(z)$, наиболее точно восстанавливающего зависимость между значениями признаков из множества Y у некоторого объекта z и номером состояния, в котором этот объект предположительно находится. При этом информация о том, что это за объект и присутствует ли он в каком-нибудь состоянии в обучающей выборке не дана.

В рамках данной работы нас будет интересовать частный случай задачи распознавания состояний, в которой необходимо научиться отличать одно выделенное состояние от фонового 0-состояния, то есть случай $K = 1$. При этом для распознавания решающему правилу d‘ предъявляется не только неизвестный объект z в неизвестном состоянии, но и этот же объект z^0 в фоновом состоянии, измеренный в пространстве признаков Y. Подобная постановка оправдана, когда первичным для нас является не распознавание состояния как такового, а выделение системы признаков Y связанных с возникновением этого состояния. В медицинских исследованиях это объясняется тем, что связь выделенной на качественном уровне информативной системы признаков, с исследуемым состоянием необходимо в дальнейшем подтверждать экспериментально и обосновывать теоретически. Решающее правило при этом заранее признается переобученным, так как обоснование точных количественные значения его параметров практически невозможно.

Для учета характерных особенностей объектов при распознавании состояний, будем рассматривать не сами состояния, а разницу между ними, некую нормализацию всех состояний для объекта относительно нулевого состояния (фонового, усредненного) состояния для этого объекта. Объекты, присутствующие в обучающей выборке только в одном состоянии при этом из рассмотрения исключаются. (Если таких объектов оказывается большинство, то-есть количество связанных между собой элементов выборки невелико, то эту задачу можно рассматривать как классическую задачу распознавания и решать существующими методами.) В результате мы переходим от исходной выборки A к нормализованной выборке ΔA, таблица данных для которой в строчке, соответствующей элементу a_i^j, где a_i^j — это описание объекта i в состоянии j содержит разность $\Delta a_i^j = a_i^j - a_i^0$, а в строчках, соответствующих нулевым состояниям объектов — нули. При этом, если известно фоновое состояние z^0 для объекта z, чье состояние необходимо распознать, то переходя к разности $\Delta z = z - z^0$, мы приходим к классической задаче распознавания K образов, соответствующих K ненулевым состояниям. Однако открытым остается вопрос, как отличать их в этом случае от фонового состояния. Об этом пойдет речь в следующем параграфе.

Простейший подход к решению задачи распознавания состояния объектов с использованием FRiS-функций

Вопрос о том, как отличать ненулевые состояния объектов от фоновых мы рассмотрим на примере более простой задачи распознавания состояний, где нужно различать фон и одно специфическое состояние, которая будет решаться с помощью функции конкурентного сходства.

Понятие функции конкурентного сходства возникло из идеи, что для оценки близости между объектами произвольной природы необходимо учитывать конкурентную ситуацию. Так, при распознавании двух образов, для оценки конкурентного сходства объекта z с первым образом необходимо учитывать не только расстояние r_1 от z до этого образа, но и расстояние r_2 до ближайшего конкурирующего образа (в случае двух образов это будет расстояние от z до второго образа).

Нормированная величина конкурентного сходства при этом вычисляется по следующей формуле:

$$F(z) = \frac{r_2 - r_1}{r_2 + r_1}.$$

Подробнее о свойствах функции конкурентного сходства можно узнать из [2]. Нас же интересует, как видоизменится функция конкурентного сходства в задаче распознавания состояний объектов и как она может быть записана при использовании нормализованной выборки ΔA. В качестве рас-

стояния от объекта до образа, а в данном случае от некоторого объекта z в некотором состоянии до образа, соответствующего выделенному состоянию j, будем рассматривать расстояние от этого объекта, до ближайшего к нему объекта в состоянии j:

$$r^j(z) = \min_{\substack{a \in A \\ O(a)=i}} r(z,a)$$

Тогда в случае идентификации единственного состояния для объекта a, находящегося в этом состоянии, в качестве ближайшего объекта конкурирующего образа выступает нулевое состояние этого объекта, то есть $r_2 = \|a - a^0\| = \|\Delta a\|$. При этом $r_1 = \min_{x \in A, O(x)=1} \|\Delta a - \Delta x\|$. Таким образом в терминах нормализованной выборки ΔA в качестве расстояний $r1$ и $r2$ используются нормы ненулевых строк и нормы разности ненулевых строк соответствующей таблицы.

В предыдущих работах [2] было показано, что среднее значение FRiS-функции вычисленное по анализируемой выборке в задаче распознавания является эффективным критерием выбора информативной системы признаков, потому именно его мы будем использовать в процедуре выбора признаков для задачи распознавания состояния объектов. При этом для направленного перебора различных подпространств признаков будет использоваться алгоритм AdDel [3].

В качестве решающего правила в этой задаче использовалось правило ближайшего соседа, которое при переходе к нормализованной выборке также видоизменялось. Для наблюдения Δz, наблюдаемый объект признавался находящимся в изучаемом состоянии, если $r_2 = \|\Delta z\|$ было больше, чем $r_1 = \min_{a \in A, O(a)=1} \|\Delta z - \Delta a\|$. Эта процедура также легко осуществляется с использованием нормализованной выборки ΔA.

Экспериментальная проверка предлагаемого подхода

Проверка эффективности предложенного подхода и его сравнение с другими возможными подходами проводилось на задаче распознавания многопериодического паттерна, характерного для визуального представления буквы «А». Испытуемому предъявлялась буква «А», размером 20 см, напечатанная на белом картоне и записывалась активность различных участков головного мозга испытуемого в различных частотных диапазонах за период 160 сек. Результатом измерений являлась частотная матрица 2435 отдельно для левого и правого полушарий головного мозга (всего 1680 признаков). Также для испытуемого записывался кадр «Фон», длительностью 160 сек, с закрытыми глазами в расслабленном состоянии. В разные дни и время исследование повторялось. Всего было 2 испытуемых. В результате анализируемая выборка состояла из 42 замеров, 24 из которых соответствовали первому испытуемому, а 18 — второму. Доля измерений фона равнялась 0,5.

Таблица 1. Сравнение эффективности алгоритмов на задаче распознавания состояния.

Type	FRiS-GRAD	CS+STOLP	CS+CNN
`CV`	0,95	0,91	0,89
`Test`	0,66	0,55	0,55
`CV`	1	0,88	0,74
`Test`	0,5	0,37	0,47
`CV`	0,85	0,66	0,75
`Test`	0,52	0,61	0,72
`CV`	0,73	0,65	0,71

На этой задаче сравнивалось три подхода. При первом информация о связи между измерениями, проведенными на одном и том же человеке, не использовалась. Эта задача интерпретировалась как классическая задача распознавания и для ее решения использовался алгоритм FRiS-GRAD [2]. При втором подходе информация о связи между объектами использовалась на этапе выбора признаков, а уже в выбранной системе решающее правило строилось алгоритмом FRiS-Stolp [3](Это сочетание при описании результатов мы будем обозначать «CS+Stolp».) Наконец результаты, полученные алгоритмом, учитывающим связь между объектами, как на этапе выбора системы признаков, так и непосредственно при распознавании, который был предложен в предыдущем параграфе, будут далее помечаться «CS+CNN». В качестве оценок эффективности использовался как скользящий экзамен (CV), так и разделение выборки на обучающую и контрольную (Test). Для увеличения объема контрольной выборки в нее включались все возможные пары соответствующие состоянию «ФОН» и «буква А».

В первых двух строках Таблицы 1 содержатся результаты эксперимента, в котором в качестве обучающей выборки использовались состояния только первого испытуемого. Состояния второго испытуемого при этом предъявлялись на контроль. В первой строке при этом приводятся результаты скользящего экзамена, а во второй — результаты на контроле. Результаты аналогичного эксперимент, когда измерения второго испытуемого составляют обучающую выборку приводятся в следующих двух строках. Усреднение по различным вариантам, когда половина замеров от первого и второго испытуемых используется при обучении, а другая половина отправляется на контроль — приводятся в третьем блоке таблицы. Последняя строчка соответствует скользящему контролю на всей выборке.

Анализ представленных результатов показывает, что на задачах, где связи между объектами обучающей выборки нет, так как они все относятся к одному объекту и потому равноправны алгоритм FRiS-Grad оказывается эффективнее. Это закономерно, так как правило ближайшего соседа — одно из самых слабых решающих правил. Однако на задачах, в которых в обучающей выборке присутствуют разные объекты и потому возникает связь между элементами выборки, учет этой связи уменьшает ошибку на тестовой выборке, которая является несмещенной оценкой ожидаемой ошибки.

Выводы

В данной статье предложен подход к решению задачи распознавания состояния объектов, который позволяет учитывать связь между элементами обучающей выборки. Экспериментально показано, что учет этой связи может быть важен при решении прикладных задач. Однако, алгоритм распознавания, легший в основу предложенного метода прост и потому имеет смысл провести подобные исследования для более сложных алгоритмов, например алгоритма FRiS-Stolp.

Литература

[1] *Лебедев Ю. А., Шабанов Г. А., Рыбченко А. А.* Магнитоэнцефалограф индукционный для регистрации и анализа ритмической активности биопотенциалов головного мозга // Информатика и системы управления. — 2008. — Т. 16, № 2. — С. 93–95.

[2] *Zagoruiko N. G., Borisova I. A., Dyubanov V. V., Kutnenko O. A.* Methods of Recognition Based on the Function of Rival Similarity // Pattern Recognition and Image Analysis. — 2008. — V. 18. — Pp. 1–6.

[3] *Загоруйко Н. Г.* Прикладные методы анализа данных и знаний. — Новосибирск: Издательство ИМ, 1999. — 270 с.

Построение взвешенных обучающих выборок w-объектов на основе сеточного подхода

Волченко Е. В.
Lm@mail.promtele.com
Украина, г.Донецк, Государственный университет информатики и искусственного интеллекта МОН Украины

В работе рассматривается проблема предобработки обучающих выборок в системах распознавания, решаемая путем построения взвешенной выборки w-объектов на основе сеточного подхода. Предложен метод формирования значений признаков w-объектов и их веса на основе анализа клеток сетки. Предложен способ сокращения взвешенной выборки w-объектов, выполняемого путем удаления w-объектов малого веса. Выполнен анализ особенностей предложенного метода и проведены экспериментальные исследования, подтвердившие его эффективность.

Проблема предобработки обучающих выборок в настоящее время увеличивает свою значимость за счет расширения круга прикладных задач, для которых выполняется построение систем распознавания, сложности и объема обрабатываемых входных данных [1, 2]. Обработка обучающих выборок, состоящих из десятков тысяч объектов, требует не только решения задач очистки и заполнения пробелов в данных, но и, в первую очередь, задач объединения и сжатия данных при условии сохранения информативности исходной выборки [1, 3]. Задача сжатия данных при условии неизменяющегося словаря признаков может быть, в том числе, решена двумя способами:

1. путем отбора некоторого множества объектов исходной обучающей выборки, каждый из которых отвечает предъявляемым требованиям;
2. путем построения множества новых объектов, каждый из которых строится по информации о некотором подмножестве объектов исходной обучающей выборки и обобщает его.

Первый способ, на на взгляд, является предпочтительным, когда объекты одного класса расположены в пространстве признаков достаточно близко друг к другу и классы пересекаются незначительно. Наиболее известными алгоритмами, реализующими такой способ, являются алгоритмы STOLP [4], FRiS-STOLP [5], NNDE (Nearest Neighbor Density Estimate) и MDCA (Multiscale Data Condensation Algorithm) [1]. Основными отличиями этих алгоритмов друг от друга являются способ отбора объектов, мера вычисления расстояния между ними и критерий оптимальности полученной подвыборки.

Второй способ более эффективен для обработки обучающих выборок, объекты которых неравномерно распределены в признаковом пространстве, классы имеют существенный разброс значений признаков и достаточно сильно пересекаются. Основой алгоритмов данного типа является дискретизация пространства признаков и анализ полученных частей пространства независимо друг от друга [6]. Одним из перспективных подходов данного направления является наложение на пространство признаков некоторой решетки, делящей все пространство на прямоугольные области, называемые в дальнейшем клетками. К алгоритмам, реализующим такой подход, относятся алгоритм LVQ (learning vector quantization) [6], алгоритм четкого разбиения пространства признаков [7], алгоритм GridDC [8]. Основным отличием алгоритма GridDC от двух других является формирование на выходе новой сокращенной обучающей выборки, а не множества классифицированных клеток, являющихся одновременно и решающим правилом классификации. Недостатком формирования классифицированных клеток является то, что получаемое разбиение в большинстве случаев очень грубо аппроксимирует границы классов и может оказаться как чрезмерно избыточным, так и крайне недостаточным по числу выделенных клеток [7]. Формирование выборки объектов позволяет использовать для построения решающих правил известные алгоритмы, дающие более эффективные решения классификации.

В предыдущих работах автора, например в [9], была предложена идея перехода к взвешенным сокращенным выборкам w-объектов, имеющим кроме значений признаков дополнительный параметр — вес. Вес содержит информацию о взаиморасположении, количестве или качестве заменяемых объектов и, исходя из результатов экспериментальных исследований, позволяет существенно повысить эффективность работы систем.

Целью данной работы является разработка метода построения взвешенной обучающей выборки w-объектов на основе сеточного алгоритма GridDC.

Постановка задачи

В качестве исходных данных дано некоторое множество объектов $X = \{X_1, \ldots, X_k\}$, представленное в виде объединения непересекающихся классов $X = \bigcup_{i=1}^{l} V_l$, и называемое обучающей выборкой. Каждый объект X_i из X описывается системой признаков, т. е. $X_i = \{x_{i1}, x_{i2}, \ldots, x_{in}\}$, и представляется точкой в линейном пространстве

признаков, т. е. $X_i \in R^n$. Для каждого объекта X_i известна его классификация $y_i \in \overline{1,l}$.

Необходимо сформировать классифицированную взвешенную обучающую выборку w-объектов $X^W = \{X_1^W, \ldots, X_m^W\}$, $y_i^W \in V$, где $X_i^W = \{x_{i1}, x_{i2}, \ldots, x_{in}, p_i\}$, p_i — вес i-ого w-объекта.

Построение взвешенной выборки w-объектов на основе алгоритма GridDC

Идеей метода GridDC [8] является наложение сетки на признаковое пространство для формирования множества клеток, определение объектов выборки, принадлежащих каждой из клеток и их замена на объекты новой сокращенной обучающей выборки. Формирование объектов новой выборки выполняется только в случае принадлежности всех объектов клетки одному классу. Вес w-объектов предлагается определять по количеству объектов исходной выборки, принадлежащих клетке. Далее приведем пошаговое описание метода.

Шаг 1. Формирование сетки. Рассчитывается шаг клетки s по формуле:

$$s = \left\lfloor 1 + \frac{\left(\sum_{i=1}^{n}\left(\max(x_i) - \min(x_i)\right)\right)^n \left(\lfloor \ln(k) \rfloor - 1\right)}{n \prod_{i=1}^{n}\left(\max(x_i) - \min(x_i)\right)} \right\rfloor,$$

где $\lfloor \ldots \rfloor$ — оператор округления до ближайшего целого значения;

$\max(x_i)$ — максимальное значение i-ого признака среди всех объектов выборки;

$\min(x_i)$ — минимальное значение i- ого признака среди всех объектов выборки.

Шаг 2. Формирование значений признаков множества новых обучающих объектов. Возможны следующие варианты обработки содержимого клеток.

1. Если все объекты клетки принадлежат к одному классу, то значения признаков объекта новой выборки рассчитываются как координаты центра масс объектов этой клетки:

$$x_{it} = \frac{1}{|W_i|} \sum_{X_j \in W_i} x_{jt},$$

где $t = 1, \ldots, n$, W_i — множество объектов исходной выборки, принадлежащих i-й клетке, $|W_i|$ — количество объектов исходной выборки, принадлежащих клетке.
2. Если клетка не содержит ни одного объекта, то объект новой выборки не формируется.
3. Если клетка содержит объекты нескольких классов, то она рекурсивно делится на две равные по размеру клетки (поочередно вертикально или горизонтально) до тех пор, пока любая из клеток внутри начальной клетки не будет содержать объекты только одного класса. Далее по каждой из полученных клеток формируются объекты новой выборки (согласно случаям 1 и 2).

Классификация w-объекта определяется по классификации объектов, по которым он сформирован.

Шаг 3. Определение веса w-объектов.

Вес w-объекта определяется по количеству объектов исходной выборки, принадлежащих клетке:

$$p_i = |W_i|.$$

В результате выполнения алгоритма будет получена новая сокращенная обучающая выборка w-объектов X^W.

Анализ предложенного алгоритма построения взвешенной выборки w-объектов.

Предложенный алгоритм и выборка w-объектов обладают следующими свойствами.

1. Выборка w-объектов формируется по всем объектам исходной выборки.
2. Никакие два и более w-объекта не содержат информацию об одном и том же объекте исходной выборки.
3. Вес w-объекта является целым числом и может принимать значения от 1 до количества объектов $|V_j|$ некоторого класса j.
4. Алгоритм требует выполнения не более чем

$$\prod_{i=1}^{n} \frac{\left(\max(x_i) - \min(x_i)\right)}{s}$$

шагов.
5. Временная сложность алгоритма равна $O(k)$.

Также отметим, что одним из наиболее существенных недостатков сеточного подхода в целом является сглаживание закона распределения признаков объектов, поскольку исходная выборка заменяется клетками равного размера вне зависимости от количества содержащихся в ней объектов. Введение веса для объектов формируемой выборки позволяет сохранить информацию о количестве заменяемых объектов в клетках, т. е. учитывать плотность распределения объектов в пространстве признаков.

Определение размера выборки w-объектов

Одним из преимуществ использования взвешенных обучающих выборок является возможность варьирования количеством объектов, включаемых в результирующую выборку [9]. Это выполняется

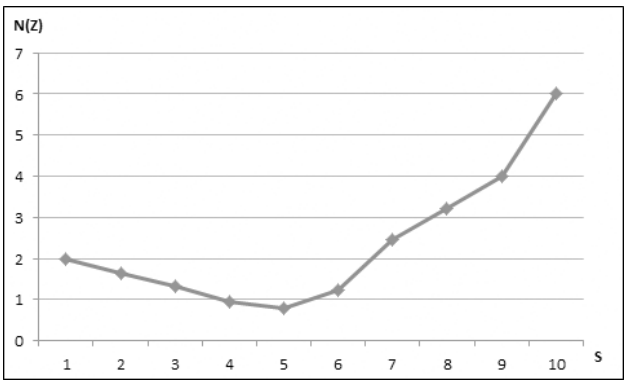


Рис. 1. Зависимость эффективности классификации от порога удаления объектов с малым весов.

путем установление порога на минимальный вес w-объектов. Очевидно, что w-объекты, имеющие единичный вес, не являются «типичными» объектами, могут быть «выбросами» и ухудшать эффективность работы систем распознавания.

Для анализа эффективности работы систем распознавания в зависимости от порога удаления объектов с малым весов был проведен ряд экспериментальных исследований. В экспериментах использовались выборки размером 1000–5000 объектов, площадь пересечения классов в пространстве признаков составляла 20-30%. В качестве решающего правила классификации использовался метод k-ближайших соседей, адаптированный к использованию на выборках w-объектов. Результаты исследований приведены на рисунке 1 и показывают, что наиболее эффективным является удаление w-объектов, вес которых меньше или равен 5. Отметим, что полученное значение может значительно варьироваться при изменении способа расчета размера клеток, площади пересечения классов в пространстве признаков и др.

Выводы

В работе предложен новый подход к построению взвешенных обучающих выборок w-объектов на основе сеточных алгоритмов. Описан метод формирования значений признаков w-объектов и их веса. Анализ предложенного метода показал его сходимость, низкую временную сложность, корректность обработки объектов исходной выборки. Предложен способ управления размером выборки путем удаления w-объектов, имеющих малый вес, показавший свою эффективность в тестовых испытаниях. Анализ эффективности использования выборок w-объектов, построенных по предложенному методу, для классификации объектов класса тестовых задач ADS 1 репозитория ISEC показал уменьшение неверных классификаций по сравнению с классическими методами — методом k-ближайших соседей и методом потенциальных функций — в среднем на 5,7% и 3,4% соответственно.

Литература

[1] *Pal S.,K.* Pattern Recognition Algorithms for Data Mining: Scalability, Knowledge Discovery and Soft Granular Computing. — Chapman and Hall/CRC, 2004. — 280 p.

[2] *Larose D. T.* Discovering knowledge in Data: An Introduction to Data Mining. — New Jersey, Wiley & Sons, 2005. — 224 p.

[3] *Basu M., Ho T. K.* Data Complexity in Pattern Recognition. — NSpringer-Verlag, London, 2006. — 310 p.

[4] *Загоруйко Н. Г.* Прикладные методы анализа знаний и данных. — Новосибирск: Издательство института математики, 1999. — 270 с.

[5] *N. G. Zagoruiko, I. A. Borisova, V. V. Dyubanov, and O. A. Kutnenko.* Methods of Recognition Based on the Function of Rival Similarity // Pattern Recognition and Image Analysis. — 2008. — V.frm[o]–8, N. 1. — Pp. 1–6.

[6] *Kohonen T.* Self-Organizing Maps. — Springer-Verlag, 1995. — 501 p.

[7] *Субботин С. А.* Метод обучения нейро-нечеткой сети распознаванию образов на основе прямоугольного разбиения пространства признаков // Складні системи і процеси. — 2009. — № 1. — С. 111–111.

[8] *Дрозд И. В., Волченко Е. В.* Метод сокращения обучающих выборок GridDC // Искусственный интеллект. — 2010. — № 4. — С. 185–190.

[9] *Volchenko E. V* Метод сокращения обучающих выборок GridDC Research of features in association of training sample objects to meta-objects // 9th International Conference on "Pattern recognition and image analysis: new information technologies": Conference Proceeding. Nizhny Novgorod, Russian Federation, 2008. — V. 2. — Pp. 291–294.

Построение нечетких характеристик классов образов по выборке прецедентов в задачах распознавания образов

Козловский В. А., Максимова А. Ю.
maximova.alexandra@mail.ru
Донецк, Институт прикладной математики и механики НАН Украины

В работе предлагается подход к решению задачи распознавания образов, в котором совмещен этап анализа данных с алгоритмом классификации. В предлагаемом методе анализа данных предлагается по проекциям признаков обучающей выборки строить интегральные характеристики — нечеткие портреты. Описаны предпосылки для выполнения внутриклассовой кластеризации с целью повышения качества алгоритма распознавания. Апробация подхода с использованием нечетких портретов выполнена на задаче контроля качества нефтепродуктов.

В работе рассматривается классическая задача распознавания образов. Акцент сделан на задачи, в которых классы образов обладают априорной неразделимостью. Подразумевается наличие в рамках одного класса нескольких кластеров. Предлагается недетерминированный адаптивный алгоритм решения рассматриваемой задачи. В качестве модельной характеристики классов строятся так называемые нечеткие портреты классов. Настройка алгоритма выполняется за счет подстройки параметров нечетких портретов с учетом внутриклассовой кластеризации.

В теории распознавания образов можно условно выделить два уровня алгоритмов распознавания образов. Алгоритмы первого уровня строят, например, разделяющие гиперповерхности для нескольких классов или логические решающие правила, и являются базисными для алгоритмов распознавания второго уровня, которые работают с обучающими выборками, обладающими большой мощностью и сложной структурой классов образов. К алгоритмам второго уровня относятся получившие большое развитие алгоритмы коллективного распознавания и др. использующие базисные алгоритмы как элементарные составляющие для принятия решений [1, 2].

Алгоритмы второго уровня являются инструментами для решения сложных прикладных задач, возникающих на производствах, в рамках управления технологическими процессами, в военной промышленности, в медицине. Важным этапом решения таких задач является интеллектуальный анализ данных с целью выявления структуры классов образов, которые могут пересекаться, иметь разные формы и размеры (с точки зрения геометрического подхода), обладать разной мощностью в обучающих выборках. Часто ситуация усложняется такими проявлениями нечеткости, как расплывчатость, неопределенность формы и взаимного расположения классов.

Постановка задачи

В работе рассматривается нечеткая модификация классической задачи распознавания образов. Имеется обучающая выборка

$$Y = \{(x^{(i)}, v^{(i)}) \mid x^{(i)} \in X,\ v^{(i)} \in V,\ i = 1, \ldots, n\},$$

где $x^{(i)} \in X \subset R^m$ — векторы в m-мерном признаковом пространстве, $v^{(i)} \in V,\ V = \{1, \ldots, k\}$ — номера классов образов, общее число которых равно k. Определим для каждого класса с номером j множество $V_j = \{x_j^{i_j} \mid i_j = 1, \ldots, k_j\}$, $j = 1, \ldots, k$, где k_j — количество элементов класса, представленного в обучающей выборке. Причем существуют такие классы образов V_i и V_j, что для них выполняется условие:

$$V_i \cap V_j \neq \varnothing, \tag{1}$$

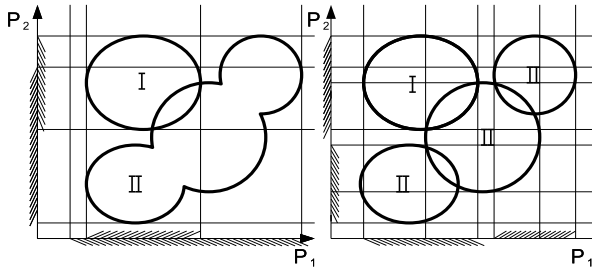
причем в случае, когда мощность таких пересечений сопоставима с мощностью самих классов, возникает необходимость получать результат работы алгоритма в виде нечеткого множества $\tilde{y} = \sum_{i=1}^{k} \mu_i / v_i$, где μ_i — степень принадлежности объекта классу образов v_i.

Особенностью данной задачи является тот факт, что в один класс образов могут объединяться объекты из нескольких однородных групп, т. е. класс может состоять из нескольких кластеров.

Анализ данных

Выбор алгоритма решения задачи распознавания образов обусловлен в первую очередь свойствами структуры классов образов.

Многие алгоритмы распознавания образов и кластеризации строятся на гипотезе компактности. Один из вариантов гипотезы компактности состоит в предположении, что множество X, компактное в m-мерном признаковом пространстве, обычно компактно и в его проекциях на координатные оси (гипотеза проективной компактности). Для построения классификатора это очевидное предположение не является достаточным. Необходимо, чтобы компактные и несовпадающие сгустки образов в m-мерном пространстве сохраняли это свойство и для своих одномерных проекций (гипотеза проективной локальной компактности) [3]. Однако в чистом виде обычно это требование не выполняется,



а) б)

Рис. 1. Пример ситуации, в которой невозможно построить эффективные решающие правила (а.) Иллюстрация успешного решения данной проблемы при использовании разбиения на кластеры (б.).

т. к. проекции точек разных образов на координатные оси образуют перекрывающиеся области.

Существует ряд алгоритмов, которые основаны на предположении, что найдутся такие комбинации несовпадающих перекрытий на нескольких осях, которые позволят построить эффективные решающие правила. Как видно из рис. 1 а), качество распознавания алгоритмов, строящих решающие правила или правила нечетких продукций, рассматриваемых в работе [4], не дадут верного решения.

Однако данную проблему можно решить применив алгоритмы кластеризации для поиска кластеров внутри класса образов, как это показано на рис. 1 б), где класс образов B разбивается на три кластера B_1, B_2, B_3.

Существование множества различных алгоритмов кластеризации (или таксономии) обусловлено субъективностью целей кластеризации. Отдельную нишу среди всех этих алгоритмов занимают алгоритмы нечеткой кластеризации [5]. Выбор подходящего алгоритма кластеризации в первую очередь обусловлен прикладной задачей. Следует отметить, что для внутриклассовой кластеризации необходимо выбирать алгоритмы, которые предполагают поиск кластеров разного относительного размера (в геометрической интерпретации это кластеры разных диаметров).

Внутриклассовую кластеризацию в данном контексте можно рассматривать как кластеризацию с «суперцелью». Это значит, что на более высоком уровне в данном случае стоит алгоритм распознавания образов. Качество алгоритма распознавания зависит от получаемых кластеров, которые фактически расширяют множество классов образов и оценивается с учетом разбиения исходных классов на кластеры. Именно по такому принципу работают алгоритмы классификации с «суперцелью».

В качестве алгоритма внутриклассовой кластеризации в работе предлагается использовать алгоритм нечеткой кластеризации, который является модификацией предложенного в [6] D-AFC(c) алгоритма, что формально можно записать как $R_j = F_{cl}(V_j)$, где F_{cl} — алгоритм вычисления нечеткого покрытия R_j по выборке V_j. Нечеткое покрытие $R = \{A^l \mid l = 1, \ldots, c\}$ — семейство нечетких множеств $A^l = \sum_{i=1}^{n} \mu_{A^l}^{(i)} / x^{(i)}$, где c — количество классов образов. Для A^l выполняется условие: $\sum_{j=1}^{k} \mu_{A^l}(x^{(i)}) > 0$ для всех $i = 1, \ldots, n$.

Необходимо найти такое $R(x)$, чтобы количество точек пересечения для каждой пары было не больше параметра *cross*, который определяет максимальное возможное число точек пересечения. Также накладывается ограничение на количество точек в каждом классе образов.

Для определения наличия пересечений между классами исходной выборки алгоритм запускается со значением параметром *cross*, сопоставимым с мощностями классов образов, представленных в обучающей выборке.

Построение нечетких портретов

Задачи распознавания образов могут рассматриваться в рамках решения задачи интеллектуального анализа данных и знаний. Исходя из этого в данной работе на основе предварительного анализа данных строятся интегральные характеристики классов образов в виде так называемых нечетких портретов классов образов.

Нечетким портретом класса образов v_t называется набор нечетких векторов $s_t = \{\overline{l_j} \mid j = 1, \ldots, c\}$, $\overline{l_j} = (l_j^1, \ldots, l_j^m)$, c — количество кластеров, на которые разбит исходный класс образов, m — размер признакового пространства, а $l_j^i = (P_i, U_i, \mu_{ij})$ — нечеткая переменная, где P_i — имя переменной, соответствующее имени признака, который она описывает, U_i — область допустимых значений признака для класса образов t, а μ_{ij} — нечеткое ограничения переменной .

Галереей нечетких портретов S назовем семейство нечетких портретов $\{s_i \mid i = 1, \ldots, k\}$, где k — количество классов образов в задаче. Алгоритм F_{fp} строит галерею портретов $S = F_{fp}(Y, \alpha, \beta, R)$, где α и β — параметры формирования нечетких портретов. Алгоритм формирования функций принадлежности нечетких портретов μ_{ij} основан на концепции скользящего окна. Ширина скользящего окна и шаг скольжения регулируются параметрами α и β соответственно. В работе [4] рассматривается данный алгоритм и определены диапазоны допустимых значений $\alpha \in (0, 1]$ и $\beta \in [1, 5]$.

Алгоритм F_{fp} формирует галерею нечетких портретов. Исходное обучающее множество расслаивается на группы образов, принадлежащие одному из классов $X = \{V_1, \ldots, V_k\}$. Для каждого множества V_j, $j = 1, \ldots, k$, строится нечеткий портрет s_j. $S = F_{fp}(Y, \alpha, \beta, R_z)$, где $R_z = \{R_i, i = 1, \ldots, k\}$ — семейство нечетких разбиений для всех классов образов, которое фор-

Алгоритм 1. Настройка нечетких портретов с учетом параметров α, β и внутриклассовой кластеризации.

Вход: Y, Y_{cv};
Выход: $S^{res}, w := 0$;
1: $R_z^{(0)} = \varnothing$;
2: **для всех** $\alpha \in A$
3: **для всех** $\beta \in B$
4: **повторять**
5: $S^{(w)} = F_{fp}(Y, \alpha, \beta, R_z^{(0)})$;
6: $Q_{cv}^{(w)} = F_{fi}(Y_{cv}, S^{(w)})$; $w := w + 1$;
7: **пока** не перебрали все α и β;
8: $S^{(w\prime)} = \arg\max_{S^w}(F_{fi}(Y_{cv}, S^{(w)}))$;
9: $R_z^{(1)} = F_{cl}(Y, S^{w\prime})$;
10: $S^{(w\prime\prime)} = F_{fp}(S^{(w\prime)}, \alpha, \beta, R_z^{(1)})$;
11: $Q_{cv}^{(w\prime\prime)} = F_{fi}(Y_{cv}, S^{(w\prime\prime)})$;
12: **если** $Q_{cv}^{(w\prime)} < Q_{cv}^{(w\prime\prime)}$ **то**
13: $S_{res} = S^{(w\prime)}$;
14: **иначе**
15: $S_{res} = S^{(w\prime\prime)}$;

мируется алгоритмом нечеткой кластеризации $F_{cl}(Y, S)$. Алгоритм кластеризации может использовать нечеткую галерею без учета кластеризации в качестве дополнительного параметра.

На практике для оптимизации небольшого числа параметров используют функционалы скользящего контроля [7]. Фактически методами скользящего контроля измеряется обобщающая способность метода обучения на заданной конечной выборке. В работе использованы функционал k-кратного скользящего, который вычисляется по результатам работы алгоритма распознавания F_{fi}: $Q_{cv} = F_{fi}(Y_{cv}, S)$. Y_{cv} — комбинации выборки прецедентов для вычисления качества алгоритма.

Приведем алгоритм настройки нечетких портретов с учетом параметров α, β и внутриклассовой кластеризации.

Апробация результатов

Данные подход был апробирован на задаче контроля качества нефтепродуктов. Данная задача является задачей обучения с учителем, однако технологические особенности производства приводят к возникновению кластеров в пределах одного класса. Для выявления этих кластеров были применены методы нечеткой кластеризации, т. к. классы достаточно сильно пересекаются. Данная задача решалась алгоритом, предложенным авторами в статье [4], однако в связи с ситуациями, которые аналогичны описанной в первой части статьи, качество не удовлетворило разработчиков и была предложена описанная в данной работе модификация. Благодаря модификации качество работы алгоритма на данных лаборатории контроля качества за 2008 год было улучшено с 92% верно распознанных объектов до 95%. Обучающая выборка состояла из 900 образцов и содержала описания 6 классов образов. Разделение на кластеры было выявлено по трем классам образов.

Выводы

В работе предлагается метод анализа данных, в результате применения которого строятся интегральные характеристики классов образов в виде нечетких портретов. Опираясь на гипотеза локальной проективной компактности нечеткие портреты строятся по проекциям. На их основе строить алгоритм распознавания образов на базе нечеткого вывода, используемый авторами, что не исключается других способов их применения. Данный подход используется для решений прикладных задач с обучающими выборками, обладающими различными аспектами неопределенности и был апробирован на задачи контроля качества нефтепродуктов, для которой было достигнуто качество распознавания до 95% верных ответов.

Использование проекций по признакам значительно упрощает алгоритм и позволяет строить характеристики, хорошо интерпретируемые экспертом. Алгоритм является эффективным, если для групп точек, описывающих класс, можно построить контур в виде эллипса, оси которого параллельны осям координат признакового пространства. Если данное условие не выполняется, то с помощью преобразований признакового пространства можно улучшить результат работы алгоритма.

Данной подход апробирован на задаче контроля качества нефтепродуктов.

Литература

[1] *Журавлев Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики, 1978. — Вып. 33. — С. 5–68.

[2] *Городецкий В. И., Серебряков С. В.* Методы и алгоритмы коллективного распознавания // Автоматика и телемеханика. — 2008. — № 11. — С. 3-40.

[3] *Загоруйко Н. Г.* Прикладные методы анализа данных и знаний. — Н.: ИМ СО РАН, 1999. — 27 с.

[4] *Козловский В. А., Максимова А. Ю.* Решение задачи распознавания образов по нечетким портретам классов // Искусственный интеллект. — 2010. — № 4. — С. 221–228.

[5] *Вятченин Д. А.* Нечеткие методы автоматической классификации. — Мн.: УП Технопринт, 2004. — 219 с.

[6] *Viattchenin D. A.* A new heuristic algorithm of fuzzy clustering // Control and cybernetics, 2004. — V. 33, N. 2. — Pp. 323-340.

[7] *Воронцов К. В.* Комбинаторный подход к оценке качества обучаемых алгоритмов. // Математические вопросы кибернетики. — М.: Физматлит, 2004. — Вып. 13. — С. 5-36.

Селективное комбинирование потенциальных функций при многомодальном восстановлении регрессионной зависимости*

Красоткина О. В., Нгуен Т. Ч., Ежова Е. О., Моттль В. В.
krasotkina@tsu.tula.ru, nguyentrongtinh75@yahoo.com.vn, lena-ezhova@rambler.ru, vmottl@yandex.ru
Тула, Тульский государственный университет; Москва,
Московский физико-технический институт; Москва, Вычислительный центр РАН

Мы рассматриваем проблему многомодального оценивания регрессионной зависимости, являющуюся обобщением классической задачи оценивания регрессии на случай, когда каждая модальность является фактически результатом сравнения объектов обучающей выборки между собой и представлена своей потенциальной функцией. Комбинирование моделей осуществляется на уровне сенсоров с помощью использования в качестве объединенной модели декартова произведения линейных пространств, в которые соответствующие потенциальные функции погружают выходные шкалы отдельных сенсоров или признаков. Для преодоления проблемы переобучения, неизбежно возникающей в данной задаче в случае большого количества построенных на данных потенциальных функций, в модель вводится структурный параметр, неявно управляющий числом потенциальных функций, присутствующих в окончательной модели.

Введение

Задача оценивания зависимостей по эмпирическим данным является одной из наиболее трудных в современной информатике. Пусть $\omega \in \Omega$ — множество объектов произвольной природы, которые характеризуются некоторой скрытой характеристикой $y \in \mathbb{Y}$. Как правило, функция $y(\omega) : \Omega \to \mathbb{Y}$ — известна нам только для некоторого ограниченного набора объектов, называемого обучающей совокупностью

$$\Omega^* \Rightarrow \{\omega_j, y(\omega_j)\}_{j=1}^N . \qquad (1)$$

Требуется продлить функцию на все множество $\hat{y}(\omega) : \Omega \to \mathbb{Y}$, чтобы иметь возможность оценивать значения целевой характеристики для других объектов $\omega \in \Omega \setminus \Omega^*$. Предметом рассмотрения данной работы является случай, когда выходная переменная является действительнозначной $\hat{y}(\omega) \colon \Omega \to \mathbb{R}$, и исходная задача представляет собой задачу оценивания регрессионной зависимости.

Компьютер не может непосредственно воспринимать физические объекты. Поэтому всегда необходима некоторая формальная переменная, выступающая как посредник между компьютером и природой. Практически, все способы представления объектов в задачах восстановления зависимостей делятся на две категории, одна из которых основана на признаковых представлениях объектов, а другая на понятии сходства и несходства объектов.

При признаковом представлении каждый объект связан с некоторой переменной $x(\omega) \colon \Omega \to \mathbb{X}$, которая называется его представлением в компьютере или признаком. Неизвестная регрессионная зависимость оценивается на основе обучающей совокупности, имеющей более специальный вид, чем (1):

$$\Omega^* \Rightarrow \{x(\omega_j), y(\omega_j)\}_{j=1}^N . \qquad (2)$$

Если признаковые описания объектов представляют собой последовательности действительных чисел, то регрессионная модель является простой как по форме, так и по способу оценивания параметров по обучающей совокупности (2) в виде

$$\hat{y}\big(\boldsymbol{x}(\omega)\big) = \hat{\boldsymbol{c}}^T \boldsymbol{x}(\omega) + \hat{b} = \sum_{i=1}^N \hat{c}_i x_i(\omega) + \hat{b}. \qquad (3)$$

Заметим, что сами отдельные признаки $x_i(\omega)$ в (3) являются простейшими модальностями представления объектов, а линейная комбинация в (3) — простейшим способом их слияния.

Мы рассматриваем более общий принцип представления объектов, основанный на понятии сходства или несходства объектов, и использующийся в том случае, когда единственным способом восприятия объектов является их попарное сравнение (ω', ω'') с помощью некоторой действительной двухместной функции $K(\omega', \omega'') \colon \Omega \times \Omega \to \mathbb{R}$. В большинстве практических ситуаций сходство или несходство измеряется на основании нескольких различных свойств или модальностей представления объектов, каждая из которых представлена своей двухместной функцией $(K_i(\omega', \omega''))_{i=1}^n$. При этом в обучающей совокупности каждый объект вместо индивидуальных признаков будет представлен n матрицами своих попарных сравнений с другими объектами обучающей выборки $y(\omega)$

$$\Omega^* \Rightarrow \{K_i(\omega_j, \omega_l)_{i=1}^n,\ y(\omega_j)\}_{j,l=1}^N . \qquad (4)$$

Мы покажем, что результат оценивания регрессионной модели на основе метода потенциальных функций имеет вид:

$$\hat{y}(\omega \,|\, \Omega^*) = \sum_{j=1}^N \sum_{i=1}^n \hat{a}_{ij} K_i(\omega_j, \omega) + \hat{b}, \qquad (5)$$

Работа выполнена при финансовой поддержке РФФИ, проекты № 11-07-00409, 11-07-00634, 09-07-00394.

где параметры $\big(\hat{a}_{ij}\big)_{i=1}^{n}{}_{j=1}^{N}$ и $\hat{b}$ оцениваются по обучающей совокупности (4).

Чем больше существует разных модальностей, тем шире круг свойств объектов, учитываемых в регрессионной модели (5). Но если число потенциальных функций n велико, такая сложная модель теряет свою обобщающую способность. Задача сокращения признакового описания (3) рассматривалась в литературе [2], но существующая методология оказывается неприменимой в задаче оценивания регрессии (5) по методу потенциальных функций. В этой статье мы попытаемся устранить этот недостаток.

Для селективного комбинирования потенциальных функций мы будем использовать принцип релевантных потенциальных функций (relevance kernel machine, RKM), который изначально был предложен для распознавания образов. Данная техника позволит определить «лишние» признаки и присвоить им малые регрессионные коэффициенты $((\hat{a}_{ij})_{i=1}^{n})_{j=1}^{N}$.

Желаемый уровень селективности модели определяется мета-параметром который регулирует степень элиминации «лишных» потенциальных функций. Подходящий уровень селективности определяется с помощью процедуры скользящего контроля.

Семейство потенциальных функций на объектах произвольного типа в задаче оценивания регрессионной зависимости

Пусть на множестве объектов $\omega \in \Omega$ определены потенциальные функции $K_i(\omega', \omega'')\colon \Omega \times \Omega \to \mathbb{R}$, $i = 1, \ldots, n$, которые выражают альтернативные способы количественного сравнения между всеми парами объектов. Всякая потенциальная функция $K(\omega', \omega'')$ погружает множество объектов $\mathbf{\Omega}$ в некоторое действительное линейное пространство со скалярным произведением $\Omega \subset \widetilde{\Omega}_i$, роль которого играет сама симметрическая исходная потенциальная функция. Важной особенностью метода потенциальных функций, часто эксплуатируемой в задачах анализа данных, является то, что он дает возможность работать с объектами произвольной природы в единых терминах линейной действительной функции $f_i(\omega)\colon \Omega \to \mathbb{R}$, для оценивания которой достаточно найти направляющий элемент (вектор, в терминах линейных пространств) $c_i \subset \widetilde{\Omega}_i$, тогда функция будет выражена в виде скалярного произведения $f_i(\omega \,|\, c_i) = K_i(c_i, \omega)$

Рассмотрим $\widetilde{\Omega}_i \times \ldots \times \widetilde{\Omega}_n \supset \Omega \times \ldots \times \Omega = \Omega^n$ — декартово произведение линейных пространств $\widetilde{\Omega}_i \supset \Omega$, определенных соответствующими потенциальными функциями, и определим подходящую комбинированную потенциальную функцию (скалярное произведение) в нем $(\widetilde{\Omega}_i \times \ldots \times \widetilde{\Omega}_n) \times (\widetilde{\Omega}_i \times \ldots \times \widetilde{\Omega}_n) \to \mathbb{R}$. Для обозначения комбинированной потенциальной функции мы будем использовать символ $K(\omega', \omega'')$. С этой точки зрения, любой выбор точки $\boldsymbol{c} = (c_i \in \widetilde{\Omega}_i)_{i=1}^{n} \in (\widetilde{\Omega}_i \times \ldots \times \widetilde{\Omega}_n)$ и действительного числа $b \in \mathbb{R}$ определяет регрессионную зависимость на исходном множестве объектов

$$\hat{y}(\omega) = K(\boldsymbol{c}, \omega) + b = \sum_{i}^{n} K_i(c_i, \omega) + b, \qquad (6)$$

давая, таким образом, потенциальную функцию.

Линейная гамма-нормальная модель скрытой регрессионной зависимости и ее байесовская оценка по обучающей совокупности

Пусть на множестве пар $\big(\omega, y(\omega)\big) \in \Omega \times \mathbb{R}$ введено вероятностное пространство. При этом любой наблюдаемый объект и его действительная целевая характеристика рассматриваются как пара случайных величин в пространстве $\Omega \times \mathbb{R}$.

Пусть наблюдения связаны со скрытыми признаковыми переменными линейной регрессионной зависимости $E(y \,|\, \omega; c_1, \ldots, c_n, b)$ с неизвестными коэффициентами $c_i \in \widetilde{\Omega}_i$. (6), тогда соответствующее условное параметрическое семейство плотностей распределений будет нормальным с неизвестной дисперсией шума наблюдения $\xi > 0$

$$\varphi(y \,|\, \omega; c_1, \ldots, c_n, \xi) = \big(1/\xi^{1/2}(2\pi)^{1/2}\big) \times \\ \times \exp\Big(- \tfrac{1}{2\xi}\big(y - \sum_{i=1}^{n} K_i(c_i, \omega) - b\big)^2\Big).$$

В этом случае совместное распределение целевых переменных $y_j = y(\omega_j)$ в обучающей совокупности (4) имеет вид

$$\Phi(y_1, \ldots y_N \,|\, \omega_1, \ldots, \omega_N; c_1, \ldots, c_n, \xi) = \\ = \prod_{j=1}^{N} \varphi(y_j \,|\, \omega_j; c_1, \ldots, c_n, \xi) = \big(1/\xi^{1/2}(2\pi)^{1/2}\big) \times \\ \times \exp\Big(- \tfrac{1}{2\xi}\big(y_j - \sum_{i=1}^{n} K_i(c_i, \omega) - b\big)^2\Big). \quad (7)$$

В свою очередь, будем рассматривать неизвестные коэффициенты регрессии $(c_i)_{i=1}^{n}$ как независимые скрытые случайные величины, каждая из которых распределена в своем линейном пространстве $c_i \in \widetilde{\Omega}_i$, определенном соответствующей потенциальной функцией, в соответствии с нормальными законами с нулевыми математическими ожиданиями.

Пусть дисперсии $r_i\xi$ коэффициентов регрессии различны для каждой модальности и пропорциональны дисперсии шума наблюдения ξ в (7) $(r_1 > 0, \ldots, r_n > 0)$, и, таким образом, априорное распределение коэффициентов регрессии будет

иметь вид

$$\psi(c_i \,|\, r_i, \xi) \propto \Big(\prod_{i=1}^{n} r_i \xi \Big)^{-1/2} \exp\big(-(1/2r_i\xi) K(c_i, c_i) \big).$$

Что касается константы b, то нет никакой априорной информации о ее распределении, поэтому совместная априорная плотность распределения будет несобственной

$$\psi(c_i, \ldots, c_n, b \,|\, r_1, \ldots, r_n, \xi) \propto \Big(\prod_{i=1}^{n} r_i \xi \Big)^{-1/2} \times \\ \times \exp\Big(-(1/2) \sum_{i=1}^{n} (1/r_i\xi) K(c_i, c_i) \Big). \quad (8)$$

Если дисперсии регрессионных коэффициентов $(r_1\xi, \ldots, r_n\xi)$ в априорном распределении (8) фиксированы, то точка максимума совместной априорной плотности $\mathsf{P}(c_i, \ldots, c_n, b \,|\, \Omega^*, r_1, \ldots, r_n, \xi)$ определяется в соответствии с байесовским правилом

$$(\hat{c}_1, \ldots, \hat{c}_n, \hat{b}) = \\ \arg\max \mathsf{P}\big(c_1 \in \widetilde{\Omega}_1, \ldots, c_n \in \widetilde{\Omega}_n, b \,|\, \Omega^*, r_1, \ldots, r_n, \xi\big) = \\ \arg\max \big[\ln \Phi(y_1, \ldots, y_N \,|\, \omega_1, \ldots, \omega_N; c_1, \ldots, c_n, \xi) + \\ + \ln \Psi(c_i, \ldots, c_n, b \,|\, r_1, \ldots, r_n, \xi) \big]. \quad (9)$$

Как видим из (7) и (8) данная оценка не зависит от предполагаемой дисперсии шума наблюдения ξ.

Теорема 1. *Регрессионная модель* (6) *полученная по обучающей совокупности* Ω^* *при фиксированных априорных дисперсиях регрессионных коэффициентов в соответствии с байесовским принципом имеет вид* (5) *с параметрами* $(\hat{a}_{ij}, \hat{b})$, *являющимися точкой минимума квадратичного критерия*

$$J(a_{ij},\, i = 1,...,n,\, b,\, \delta_j,\, j = 1,...,N \,|\, r_i,\, i = 1,...,n) = \\ = \sum_{i=1}^{n} (1/r_i) \sum_{j=1}^{N} \sum_{l=1}^{N} K_i(\omega_j, \omega_l) a_{ij} a_{il} + \\ + \sum_{j=1}^{N} \Big(y_j - \sum_{i=1}^{n} \sum_{l=1}^{N} K_i(\omega_j, \omega_l) a_{il} - b \Big)^2 \to \min.$$

Точка минимума критерия $(\hat{a}_{ij} = r_i \hat{\delta}_j, \hat{b})$ определяется решением системы $N + 1$ линейных уравнений относительно фиктивных переменных $(\hat{\delta}_1, \ldots, \hat{\delta}_N, \hat{b})$

$$\begin{cases} \Big(\sum_{i=1}^{n} r_i K_i(\omega_j, \omega_j) + 1 \Big) \delta_j + \\ \quad + \sum_{l=1, l\neq j}^{N} \Big(\sum_{i=1}^{n} r_i K_i(\omega_j, \omega_l) \Big) \delta_l + b = y_j, \\ j = 1, \ldots, N; \\ \sum_{j=1}^{N} \delta_j = 0. \end{cases} \quad (10)$$

Байесовская оценка регрессионной модели получается по формуле (5)

$$\hat{y}(\omega \,|\, \Omega^*, r_1, \ldots, r_n) = \sum_{j=1}^{N} \hat{\delta}_j \sum_{i=1}^{n} r_i K_i(\omega_j, \omega) + \hat{b}. \quad (11)$$

Нетрудно видеть из (11) что положительные коэффициенты $(r_1, \ldots, r_n)$ при оценке регрессионной модели выступают в роли весов потенциальных функций. Если все коэффициенты равны единице $(r_1 = 1, \ldots, r_n = 1)$, все потенциальные функции одинаково участвуют в модели, если же некоторые из дисперсий $r_i\xi$ стремятся к 0, то это означает, что соответствующая потенциальная функция практически исключается из модели.

Предположим теперь, что положительные обратные коэффициенты дисперсий $(1/r_1, \ldots, 1/r_n)$ имеют независимое неизвестное априорное гамма-распределение

$$\gamma\big((1/r_i) \,|\, \alpha, \beta\big) = \big(\beta^\alpha / \Gamma(\alpha)\big)(1/r_i^{\alpha-1}) \exp\big(-\beta(1/r_i) \big)$$

с двумя параметрами $\alpha > 1$ и $\beta > 0$, математическими ожиданием $\mathsf{E}(1/r_i) = (1 + \xi)\mu + 1$ и дисперсией $\mathsf{D}(1/r_i) = \alpha/\beta^2$. Выберем $\alpha = (1/2)\big[(1/\xi)(1 + \\ + 1/\mu) + 1\big]$, $\beta = 1/2\xi\mu$, и получим параметрическое семейство распределений относительно параметра $\mu \geqslant 0$, такое, что $\mathsf{E}(1/r_i) = (1 + \xi)\mu + 1$, $\mathsf{D}(1/r_i) = 2\xi\mu\big[(1 + \xi)\mu + 1\big]$.

Если $\mu \to 0$, то априорные случайные значения дисперсий $1/r_i$ будут одинаковыми $1/r_1 \cong \ldots 1/r_n \cong 1$, а при увеличении μ, дисперсии могут существенно различаться, так как $\mathsf{D}(1/r_i)$ увеличивается быстрее, чем $\mathsf{E}(1/r_i)$

Совместное априорное распределение независимых обратных дисперсий имеет вид

$$G(1/r_1, \ldots, 1/r_n \,|\, \mu, \xi) \propto \\ \propto \Big(\prod_{i=1}^{n} \frac{1}{r_i} \Big)^{(1/2)\left[(1/\xi)(1+1/\mu)-1\right]} \exp\Big(-\frac{\xi\mu}{2} \sum_{i=1}^{n} \frac{1}{r_i} \Big).$$

Итак, байесовское обучение (9) с добавлением априорных распределений дисперсий имеет вид

$$(\hat{c}_1, \ldots, \hat{c}_n,\, \hat{b},\, \hat{r}_1, \ldots, \hat{r}_n) = \\ = \arg\max \mathsf{P}(c_1 \in \widetilde{\Omega}_1, \ldots, c_n \in \widetilde{\Omega}_n, b, \\ r_1 \geqslant \varepsilon, \ldots, r_n \geqslant \varepsilon \,|\, \Omega^*, \mu, \xi) = \\ = \arg\max \big[\ln \Phi(y_1,...,y_N \,|\, \omega_1,...,\omega_N; a_1,...,a_n, \xi) + \\ + \ln \Psi(c_i, \ldots, c_n, b \,|\, r_1, \ldots, r_n, \xi) + \\ + \ln G\big(\tfrac{1}{r_1}, \ldots, \tfrac{1}{r_n} \,|\, \mu, \xi\big) \big]. \quad (12)$$

Как и в (9), байесовская оценка не зависит от ξ.

Теорема 2. *Регрессионная модель*

$$\hat{y}(\omega) = K(\boldsymbol{c}, \omega) + b = \sum_{i=1}^{n} K_i(c_i, \omega)$$

с коэффициентами, полученными с помощью метода максимального правдоподобия для апостериорного распределения, построенного по обучающей выборке Ω^* *в соответствии с байесовским критерием* (12) *имеет вид* $\hat{y}(\omega \mid \Omega^*) = \sum_{j=1}^{N} \sum_{i=1}^{n} \hat{a}_{ij} K_i(\omega_j, \omega) + \hat{b}$, *где параметры* $(\hat{a}_{ij}, \hat{b})$ *могут быть получены как минимум квадратичного критерия*

$$J(a_{ij}, r_i, i = 1, \ldots, n, b, j = 1, \ldots, N \mid \mu) =$$
$$= \sum_{i=1}^{n} \Big(\frac{1}{r_i} \Big(\sum_{i=1}^{N} \sum_{l=1}^{N} K_i(\omega_j \omega_l) a_{ij} a_{il} + \tfrac{1}{\mu} \Big) + \big(1 + \tfrac{1}{\mu}\big) \ln r_i \Big) +$$
$$+ \sum_{j=1}^{N} \Big(y_j - \sum_{i=1}^{n} \sum_{l=1}^{N} K_i(\omega_j \omega_l) a_{il} - b \Big)^2 \to \min .$$

Минимум критерия может быть найден с помощью покоординатного спуска по двум группам переменных $(r_i,\ i = 1, \ldots, n)$ и $(a_{ij},\ j = 1, \ldots, N)$: $\big((r_i)^0 = 1,\ i = 1, \ldots, n\big)$, $(a_{ij})^k = (r_i)^k (\delta_j)^k$. Вспомогательные переменные $\delta_j,\ j = 1, \ldots, N$ могут быть получены как решение системы линейных уравнений (10) а оценки дисперсий $(r_i,\ i = 1, \ldots, n)$ на очередном шаге получаются по формуле

$$(r_i)^{k+1} = \frac{(r_i)^k (r_i)^k \sum_{i=1}^{n} \sum_{l=1}^{N} K_i(\omega_j, \omega_l) (\delta_i)^k (\delta_l)^k + \frac{1}{\mu}}{1 + \frac{1}{\mu}}.$$

Обычно процесс сходится за 10-15 итераций, заглушающих «лишиние» потенциальные функции соответствующими малыми (но всегда ненулевыми) $\hat{r}_i$ в регрессионной модели (12).

Уровень селективности потенциальных функций, определяется параметром $\mu : 0 < \mu < \infty$, значение которого может быть подобрано с помощью процедуры скользящего контроля. Предложенный в данной работе алгоритм будем называть алгоритмом восстановления регрессионной зависимости с управляемым селективным комбинированием потенциальных функций (SSRKR — Relevance Kernel Regression with Supervised Selectivity).

Экспериментальное исследование предложенного алгоритма

Исследование качества работы алгоритма проводилось на тестовых данных, полученных в соответствии с моделью линейной регрессии (7)(3). Все эксперименты выполнялись на выборке из 1000 объектов, из которых только 20 было отведено на обучение, а остальные 980 использовались для контроля качества построенной модели. В ходе экспериментов варьировалось число потенциальных функций (признаков), измеренных на объектах, от 25 до 500. Причем, в скрытой модели только две потенциальных функции являлись релевантными. Это фактически означает, что только 2 коэффициента регрессии в зависимости (3) отличны от 0, а остальные являются нулевыми, что исключает соответствующие потенциальные функции из модели. Дисперсия шума в модели (3) была установлена на уровне 10% от дисперсии наблюдаемой переменной. В ходе экспериментов сравнивались между собой 3 алгоритма восстановления регрессионной модели: оценивание регрессии по методу наименьших квадратов (LS), алгоритм оценивания регрессионной модели с селективным комбинирования потенциальных функций (SSKR) и алгоритм восстановления регрессии с отбором признаков Lasso, разработанный Тибширани в 1996 [6]. В эксперименте использовалась версия алгоритма Lasso, предложенная в работе [7], основанная на градиентном походе к оптимизации критерия оценивания регрессии со штрафным членом на сумму модулей коэффициентов регрессии. В ходе экспериментов на контрольной выборке подсчитывался относительный средний квадрат ошибки восстановления выходной переменной. Для каждого числа потенциальных функций n генерировалось 100 вариантов входных данных. В таблице 1 приведены усредненные по 100 экспериментам значения ошибки.

n	LS	$Lasso$	$SSRKR$
25	1.12	0.79	0.062
100	1.25	0.88	0.075
500	1.41	0.98	0.076

Таблица 1. Средний относительный квадрат ошибки восстановления наблюдаемой переменной в задаче оценивания регрессионной зависимости.

Литература

[1] *Vapnik V.* Estimation of Dependencies Based on Empirical Data. Springer, 1982.

[2] *Fan J, Lv J.* A selective overview of variable selection in high dimensional feature space. // Statistica Sinica, 2010. — Vol. 20. — Pp. 111–122.

[3] *Tatarchuk A., Sulimova V., Windridge D., Mottl V., Lange M.* Supervised selective combining pattern recognition modalities and its application to signature verification by fusing on-line and off-line kernels // Proc. of the 8th Int'l Workshop on Multiple Classifier Systems, Reykjavik, Iceland, June 10–12, 2009. Lecture Notes In Computer Science, Vol. 5519. Springer-Verlag, 2009, Pp. 324–334.

[4] *Vapnik V.* Statistical Learning Theory. John-Wiley and Sons, Inc., 1998.

[5] *Mottl V.* Metric spaces admitting linear operations and inner product // Doklady Mathematics, 2003. — Vol. 67, No. 1. — Pp. 140–143.

[6] *Tibshirani R.* Regression shrinkage and selection via the lasso // Journal of the Royal Statistical Society, No. 58, —Pp. 267–288.

[7] *Kim J., Kim Y., Kim Y.* A gradient-based optimization algorithm for lasso // J. of Computational and Graphical Statistics, 2008, — Vol. 17, — Pp. 994–1009.

О способах введения байесовской регуляризации в регрессии на основе гауссовских процессов

Панов М. Е., Бурнаев Е. В., Зайцев А. А.
{maxim.panov, evgeny.burnaev, alexey.zaytsev}@datadvance.net
Москва, Институт Проблем Передачи Информации РАН,
Москва, DATADVANCE,
Долгопрудный, МФТИ

В работе предложен метод введения байесовской регуляризации на параметры ковариационной функции гауссовского процесса. В качестве априорных распределений параметров рассмотрены нормальное и гамма распределения. Применение разработанного метода к задаче восстановления неизвестной зависимости позволило заметно повысить качество аппроксимации, а также увеличить обобщающую способность и надежность алгоритма

Введение

Одной из основных задач, которые приходится решать при построении метамоделей (моделей на основе данных) является задача аппроксимации неизвестной зависимости по данным [1, 2]. Наиболее популярная модель для построения аппроксиматоров, основанная на гауссовских процессах [3, 4, 5], используется в большом количестве разнообразных прикладных задач, включая многокритериальную оптимизацию при проектировании [6], конструирование в аэрокосмической [7] и автомобильной отраслях [8], а также многие другие инженерные задачи.

В своих классических реализациях, таких как DACE [9], аппроксиматоры на основе гауссовских процессов склонны в некоторых случаях давать вырожденные аппроксимации, интерпретация которых с точки зрения исходной задачи является затруднительной. Для того, чтобы избежать подобных случаев, в данной работе предлагается ввести априорное распределение параметров ковариационной функции гауссовского процесса и оптимизировать совместное правдоподобие данных и модели как по параметрам ковариационной функции, так и по гиперпараметрам априорного распределения.

Постановка задачи

В наиболее общем виде задача аппроксимации может быть сформулирована следующим образом. Пусть $y = f(x)$ некоторая неизвестная функция со входом $x \in \mathbb{X} \subset \mathbb{R}^n$ и выходом $y \in \mathbb{R}^m$. В данной статье мы ограничимся случаем $y \in \mathbb{R}^1$. Пусть $D_{\text{learn}} = \big(X, Y\big) = \big\{\big(x_i, y_i = f(x_i)\big), i = 1, \ldots, N\big\}$ — обучающая выборка. Задача состоит в построении аппроксимации $\hat{y} = \hat{f}(x) = \hat{f}(x|D_{\text{learn}})$ для исходной зависимости $y = f(x)$ по обучающей выборке D_{learn}.

Если для всех $x \in \mathbb{X}$ (не только для $x \in D_{\text{learn}}$) имеет место примерное равенство $\hat{f}(x) \approx f(x)$, то считается, что аппроксиматор хорошо воспроизводит исходную зависимость. Это факт проверяется на независимой тестовой выборке $D_{\text{test}} = \big(X_*, Y_*\big) = \big\{\big(x_j, y_j = f(x_j)\big), j = 1, \ldots, N_*\big\}$. Мерой качества аппроксимации является средняя абсолютная ошибка на контрольной выборке: $\varepsilon(\hat{f}|D_{test}) = \frac{1}{N_*}\sum_{j=1}^{N_*}\big|y_j - \hat{f}(x_j)\big|$, а также 95 и 99 процентные квантили абсолютной ошибки аппроксимации.

Гауссовские процессы

Гауссовский процесс является одним из возможных способов задания распределения на пространстве функций. Гауссовский процесс $f(x)$ полностью определяется своей функцией среднего $m(x) = \mathsf{E}[f(x)]$ и ковариационной функцией

$$\operatorname{cov}(y, y') = k(x, x') = \mathsf{E}[\big(f(x) - m(x)\big)\big(f(x') - m(x')\big)].$$

Если положить функцию среднего нулевой: $m(x) = \mathsf{E}[f(x)] = 0$, а ковариационную функцию считать известной, то функция апостериорного (для заданной обучающей выборки) среднего значения гауссовского процесса в точках контрольной выборки X_* выглядит следующим образом [10]:

$$\hat{f}(X_*) = K_* K^{-1} Y, \tag{1}$$

где $K_* = K(X_*, X) = \big(k(x_i, x_j)\big)_{i=1\,j=1}^{N_*\;N}$, $K = K(X, X) = \big(k(x_i, x_j)\big)_{i=1\,j=1}^{N\;N}$.

В типичных, более реалистичных ситуациях при моделировании мы не имеем доступа непосредственно к значениям функции, а наблюдаем их только в зашумленном виде:

$$y(x) = f(x) + \varepsilon(x), \tag{2}$$

где шум $\varepsilon(x)$ моделируется независимыми одинаково распределенными нормальными случайными величинами с нулевым средним и дисперсией $\tilde{\sigma}^2$. В таком случае наблюдения $y(x)$ будут гауссовским процессом с нулевым средним и ковариационной функцией $\operatorname{cov}\big(y(x), y(x')\big) = k(x, x') + \tilde{\sigma}^2$. Таким образом, функция апостериорного (для заданной обучающей выборки) среднего значения гауссовского процесса $f(x)$ в точках контрольной выборки X_* принимает вид [10]:

$$\hat{f}(X_*) = K_*\big(K + \tilde{\sigma}^2 I\big)^{-1} Y, \tag{3}$$

где I — единичная матрица размера $N \times N$.

Заметим, что наличие в формуле (3) дисперсии шума $\tilde{\sigma}^2$ фактически приводит к регуляризации, что позволяет улучшить обобщаю способность аппроксиматора. При этом апостериорная ковариационная функция гауссовского процесса в точках контрольной выборки имеет вид:

$$\mathbb{V}\big[\hat{f}(X_*)\big] = K(X_*, X_*) + \tilde{\sigma}^2 I_* - \\ - K_*\big(K + \tilde{\sigma}^2 I\big)^{-1} K_*^T, \quad (4)$$

где $K(X_*, X_*) = \big(k(x_i, x_j)\big)_{i=1 j=1}^{N_* \; N_*}$, I_* — единичная матрица размера $N_* \times N_*$.

Дисперсии гауссовского процесса в точках контрольной выборки могут быть использованы как оценки ожидаемой ошибки аппроксимации в этих точках. Заметим, что для этого нет необходимости вычислять по формуле (4) всю матрицу $\mathbb{V}\big[\hat{f}(X_*)\big]$, а достаточно вычислить только элементы ее главной диагонали, которые и являются искомыми дисперсиями.

При работе с реальными данными ковариационная функция породившего их гауссовского процесса как правило не известна, поэтому необходимо уметь ее идентифицировать по данным.

Нахождение параметров гауссовского процесса

Предположим, что ковариационная функция гауссовского процесса является членом некоторого параметрического семейства $k(x, x') = k(x, x'|a)$, где $a \in \mathbb{R}^K$ — вектор параметров ковариационной функции. Семейство $k(x, x'|a)$ обычно берется из класса так называемых стационарных ковариационных функций, т. е. функций значение которых зависит только от разности значений аргументов $k(x, x'|a) = k(x - x'|a)$. Значение параметра a предлагается восстанавливать по обучающей выборке D_{learn}, исходя из принципа максимума правдоподобия. Для этого выпишем логарифм правдоподобия гауссовского процесса в точках обучающей выборки [10]:

$$\log p(Y|X, a, \tilde{\sigma}) = -\frac{1}{2} Y^T (K + \tilde{\sigma}^2 I_N)^{-1} Y - \\ - \frac{1}{2}\log|K + \tilde{\sigma}^2 I| - \frac{n}{2}\log 2\pi, \quad (5)$$

где $|K + \tilde{\sigma}^2 I|$ — детерминант матрицы $K + \tilde{\sigma}^2 I$.

Кроме параметров a ковариационной функции параметром функционала (5) является также значение дисперсии шума наблюдений $\tilde{\sigma}^2$, которое также можно настраивать по обучающей выборке. Таким образом, нахождение оптимальных значений параметров сводится к отысканию максимума правдоподобия по параметрам:

$$\log p(Y|X, a, \tilde{\sigma}) \to \max_{a, \tilde{\sigma}}. \quad (6)$$

Выбор конкретного семейства ковариационных функций $k(x, x'|a)$ обычно продиктован соображениями удобства, а также априорными представлениями о свойствах аппроксимируемой зависимости. В данной работе мы используем ковариационные функции вида

$$k(x - \tilde{x}|a) = \hat{\sigma}^2 \exp\left(-\left\{\sum_{i=1}^{n} \theta_i^2 |x_i - \tilde{x}_i|^p\right\}^q\right),$$

где параметры $p \in (0, 2]$, $q \in (0, 1]$ задаются априори, $a = \big\{\theta_i, i = 1, \ldots, n; \hat{\sigma}\big\}$ настраиваются по обучающей выборке при решении задачи (6).

Решение оптимизационной задачи (6) представляет собой отдельную задачу, которая осложняется тем, что функционал (5) в общем случае многоэкстремален. В данной работе для решения (6) мы будем пользоваться алгоритмом на основе метода сопряженных градиентов.

Вырожденные случаи

Заметим, что минимум правдоподобия может достигаться при значениях параметров, которые не имеют хорошей физической интерпретации. Примером такого поведения может служить подбор очень больших значений для параметров ковариационной функции $\boldsymbol{\theta} = \{\theta_i, i = 1, \ldots, n\}$. Рассмотрим аппроксимацию функции Растригина [14] с помощью метода (3)&(6) по 20 случайным точкам (см. рис. 1).

Заметим, что аппроксимация выродилась в отдельные очень узкие пики, которые не могут быть интерпретированы. Также плохо интерпретируемые аппроксимации получаются, когда параметры $\boldsymbol{\theta}$ оказываются сильно отличающимися друг от друга по величине. В следующем разделе будет предложен метод введения байесовской регуляризации, которая позволяет бороться с такими, как приведенные выше, вырожденными случаями.

Байесовская регуляризация

Предположим, что параметры $\boldsymbol{\theta}$ распределены с некоторой априорной плотностью $\bar{p}(\boldsymbol{\theta}|\gamma)$, где γ — это гиперпараметры априорного распределения. Тогда логарифм совместной плотности распределения данных и параметров модели принимает вид:

$$\log P(Y, \boldsymbol{\theta} \,|\, X, \hat{\sigma}, \tilde{\sigma}, \gamma) = \\ = \log p(Y \,|\, X, \boldsymbol{\theta}, \hat{\sigma}, \tilde{\sigma}) + \log \bar{p}(\boldsymbol{\theta}|\gamma). \quad (7)$$

Правдоподобие (7) можно максимизировать по параметрам $\hat{\sigma}, \tilde{\sigma}, \gamma, \boldsymbol{\theta}$ с целью нахождения их оптимальных значений. Слагаемое $\log \bar{p}(\boldsymbol{\theta} \,|\, \gamma)$ выполняет роль дополнительной регуляризации на значения параметров ковариационной функции $\boldsymbol{\theta}$, причем выбор конкретного распределения будет определять характер этой регуляризации. В данной работе предлагается рассмотреть в качестве распределения $\bar{p}(\boldsymbol{\theta} \,|\, \gamma)$ нормальное и гамма распределения.

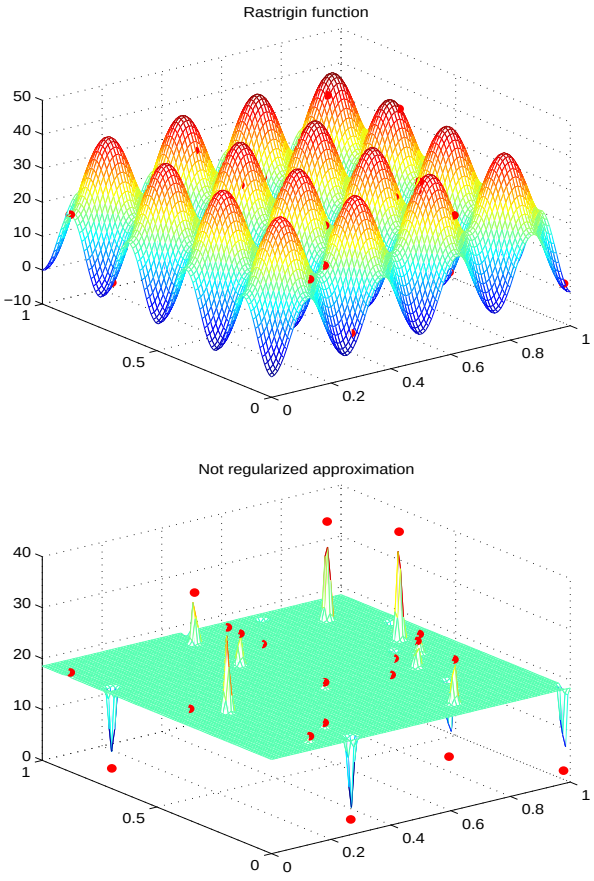


Рис. 1. Функция Растригина и ее аппроксимация.

В случае нормального распределения рассмотрим следующее преобразование вектора $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} = \boldsymbol{\theta}_{\|} + \boldsymbol{\theta}_{\perp},$$

где $\boldsymbol{\theta}_{\|} = |\boldsymbol{\theta}_{\|}| * \boldsymbol{e}$, $\boldsymbol{e} = \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)$, а вектор $\boldsymbol{\theta}_{\perp}$ ортогонален вектору $\boldsymbol{e}$. Предположим, что величина $|\boldsymbol{\theta}_{\|}|$ распределена нормально со средним $\mu_{\|} > 0$ и дисперсией $d_{\|}$, а вектор $\boldsymbol{\theta}_{\perp}$ не зависит от $|\boldsymbol{\theta}_{\|}|$ и имеет нормальное распределение с нулевым средним и диагональной ковариационной матрицей $d_{\perp} I_{\perp}$, где $I_{\perp}$ — единичная матрица размера $n-1$, $d_{\perp}$ — дисперсия компонент вектора $\boldsymbol{\theta}_{\perp}$. Такое предположение основано на нашем желании, чтобы все компоненты вектора $\boldsymbol{\theta}$ бы ли не слишком велики и не слишком малы по величине, а также не слишком сильно отличались по величине друг от друга. Таким образом в данном случае $\gamma = \{\mu_{\|}, d_{\|}, d_{\perp}\}$, а $\log \bar{p}(\boldsymbol{\theta}|\gamma) = \log N(\mu_{\|}, d_{\|}) + {} + \log N(\mathbf{0}, d_{\perp} I_{\perp})$, где $N(\mu, \Sigma)$ — функция плотности нормального распределения со средним μ и ковариационной матрицей Σ.

Теперь рассмотрим случай гамма распределения. Предположим, что все компоненты вектора $\boldsymbol{\theta}$ независимы и распределены с одним и тем же гамма распределением с параметрами k и λ. В таком

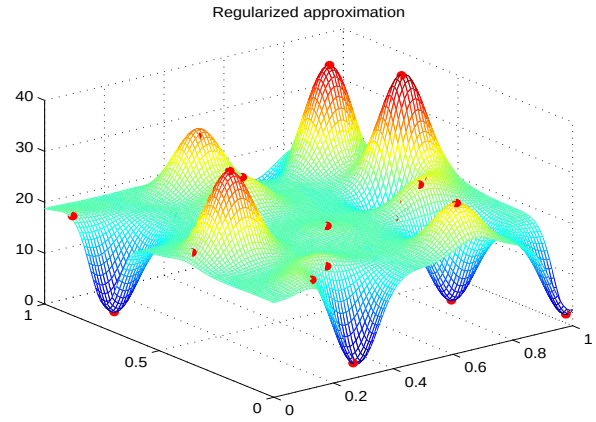


Рис. 2. Аппроксимация функции Растригина в случае введения байесовской регуляризации.

случае

$$\log \bar{p}(\boldsymbol{\theta} \,|\, \gamma) = \log \bar{p}(\boldsymbol{\theta} \,|\, k, \lambda) = \sum_{i=1}^{n} \big[k \log \lambda + (k-1) \log \theta_i - \lambda \theta_i - \log \Gamma(k)\big], \quad (8)$$

где $k > 1$, $\lambda > 0$, $\Gamma(k)$ — гамма-функция от аргумента k. Такой выбор априорного распределения параметров также штрафует слишком большие и слишком маленькие значения компонент вектора θ_i.

Если применить к вышеописанному примеру вырожденной аппроксимации каждый из этих методов, то получаются одинаковые аппроксимации, вид которых изображен на рисунке 2. Хорошо видно, что новая аппроксимация гораздо ближе к истинному виду функции Растригина, хотя и довольно сильно отличается от нее в силу малого объема обучающей выборки.

В следующем разделе мы проведем подробное экспериментальное сравнение рассматриваемых методов аппроксимации.

Экспериментальные результаты

В ходе экспериментов сравнивалась работа трех алгоритмов, основанных на гауссовских процессах, а именно:

- алгоритм без регуляризации (MGP на графике);
- алгоритм с байесовской регуляризацией на основе гамма распределения (MGP-Gamma на графике);
- алгоритм с байесовской регуляризацией на основе нормального распределения (MGP-Normal на графике);

Для демонстрации экспериментальных результатов был использован большой набор тестовых функций, которые применяются для тестирования задач оптимизации [11, 12]. Всего тестирование проводилось на 30 различных функциях, для

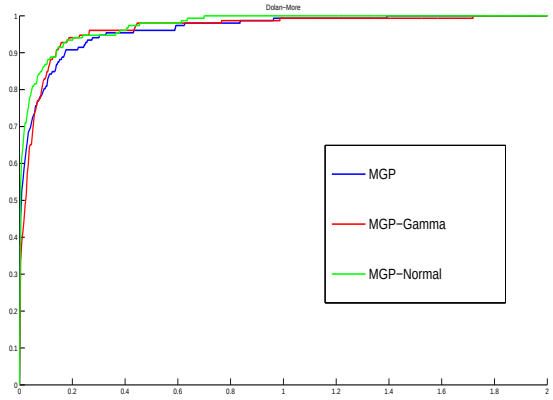


Рис. 3. Кривые Долан-Мора для тестовых функций.

каждой из которых генерировалось по 2 случайные обучающие выборки каждого из размеров 10, 20, 40, 80 и 160 точек. Результаты сравнивались по 95 процентному квантилю абсолютной ошибки на больших контрольных выборках из 10 000 точек. Для удобства результаты представлены в виде кривых Долан-Мора [13] на рисунке 3. Чем выше кривая находится на графике, тем выше качество работы соответствующего алгоритма. Мерой качества работы алгоритма также является площадь под кривой Долан-Море, значения которой для трех рассматриваемых алгоритмов приведены в таблице 1.

MGP	MGP-Gamma	MGP-Normal
1,926	1,930	1,951

Таблица 1. Площади под кривой Долан–Мора.

Заметим, что оба алгоритма с байесовской регуляризацией показали результаты лучше, чем алгоритм без регуляризации, причем улучшение, достигнутое с помощью алгоритма с регуляризацией на основе нормального распределения, является весьма значительным.

Выводы

В работе предложен метод введение байесовской регуляризации на параметры ковариационной функции гауссовского процесса. В качестве априорных распределений параметров рассмотрены нормальное и гамма распределения. Оба метода позволяют избежать получения плохо интерпретируемых аппроксимаций. Также результаты экспериментов показывают, что оба подхода позволяют увеличить обобщающую способность и надежность алгоритма. В качестве дальнейших направлений исследования предполагается введение аналогичных априорных предположений для других параметров алгоритма, таких как дисперсия шума наблюдений $\tilde{\sigma}$ и коэффициент перед ковариационной функцией $\hat{\sigma}$.

Литература

[1] *Бернштейн А. В., Бурнаев Е. В., Кулешов А. П.* Интеллектуальный анализ данных в метамоделировании. // Труды 17 Всероссийского Семинара «Нейроинформатика и ее приложения к Анализу Данных», Красноярск, 2009. — С. 23–28.

[2] *Forrester A., Sobester A., Keane A.* Engineering Design via Surrogate Modelling. A Practical Guide. — Wiley, 2008. — 238 p.

[3] *Giunta A., Watson L. T. A* Comparison of Approximation Modeling Technique: Polynomial Versus Interpolating Models // 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization. — 1998. — V. 1. — Pp. 392—404.

[4] *Simpson T. W., Booker A. J., Ghosh S., Giunta A., Koch P. N., Yang R. J.* Approximation Methods in Multidisciplinary Analysis and Optimization: A Panel Discussion // Structural and Multidisciplinary Optimization. — 2004. — V. 27, N. 5. — Pp. 302—313.

[5] *Batill S. M., Renaud J. E., Gu X.* Modeling and Simulation Uncertainty in Multidisciplinary Design Optimization // AIAA. — 2000.

[6] *Koch P. N., Wujek B. A., Golovidov O., Simpson T. W.* Facilitating Probabilistic Multidisciplinary Design Optimization Using Kriging Approximation Models // AIAA. — 2002.

[7] *Simpson T. W., Maurey T. M., Korte J. J., and Mistree F.* Kriging Metamodels for Global Approximation in Simulation-Based Multidisciplinary Design Optimization // AIAA. — 2001. — V. 39, N. 12. — Pp. 2233—2241.

[8] *Yang R. J., Wang N., Tho C. H., Bobineau J. P., and Wang B. P.* Metamodeling Development for Vehicle Frontal Impact Simulation // American Society of Mechanical Engineers, ASME Design Engineering Technical Conf.—Design Automation Conf., DETC2001/DAC-21012, Sept. 2001.

[9] *Nielsen H. B.* DACE. A Matlab Kriging Toolbox. `http://www2.imm.dtu.dk/~hbn/dace/`

[10] *Rasmussen C. E., Williams C. K. I.* Gaussian Processes for Machine Learning. — The MIT Press, 2006.

[11] Lappeenranta University of Technology: evoluationary computation pages — the function testbed. `http://www.it.lut.fi/ip/evo/functions/functions.html`

[12] Swiss International Institute of Technology (ETH): system optimization — testproblems. `http://www.tik.ee.ethz.ch/sop/download/supplementary/testproblems/`

[13] *Dolan E. D., More J. J.* Benchmarking optimization software with performance profiles // Mathematical Programmming, Ser. A 91. — 2002 — Pp. 201—213.

[14] *A. Torn, A. Zilinskas* “Global Optimization”. Lecture Notes in Computer Science, N. 350. — Springer-Verlag, Berlin, 1989.

Методика формирования функционального словаря в задаче аппроксимации многомерной зависимости

Беляев М. Г., Бурнаев Е. В., Любин А. Д.
mikhail.belyaev@datadvance.net
Москва, Институт проблем передачи информации им. А. А. Харкевича
Москва, Datadvance
Долгопрудный, Московский физико-технический институт (государственный университет)

В работе рассматривается задача построения аппроксимации многомерной зависимости на основе линейного разложения по словарю параметрических функций. Словарь формируется из функций нескольких типов: линейных, сигмоидальных и гауссовских. В работе предложен эффективный метод объединения функций разного типа в единый словарь и выбора из него подмножества, которое позволяет наиболее точно воспроизвести искомую зависимость.

Задача построения аппроксимации неизвестной зависимости получила достаточно широкое распространение. В частности, она активно используется в метамоделировании при создании сложных инженерных объектов [1]. Наиболее типичные методы, которые используются для решения этой задачи, — это линейная регрессия, однослойная нейронная сеть, радиальные базисные функции [2]. В данной работе рассматривается построение аппроксимации с помощью разложения по словарю параметрических функций разного типа. Модель такого рода позволяет моделировать достаточно широкий класс функций, в который, в частности, входят модели, используемые в упомянутых выше подходах.

Очевидно, что состав оптимального набора функций разного типа значительно зависит от свойств решаемой задачи. Поэтому цель данной работы — предложить метод выбора оптимального подмножества из словаря нелинейных параметрических функций разного типа, которое позволил бы получить более высокое качество приближения, чем аппроксиматоры на основе функций только одного типа.

Алгоритм аппроксимации

В рамках этого раздела кратко рассматриваются некоторые составляющие алгоритма аппроксимации. Более подробно он описан в работах [3, 4].

Постановка задачи. Сформулируем постановку задачи аппроксимации. Пусть $f(X) \in \mathbb{R}^1$ — некоторая неизвестная функция, зависящая от вектора $X \in \mathbb{R}^n$. Задана выборка данных $S_{\text{learning}} = \{\boldsymbol{X}, \boldsymbol{Y}\}$. $\boldsymbol{X} = \{X_i\}$, $\boldsymbol{Y} = \{Y_i\}$, $i = 1, \ldots, N_{\text{learning}}$ — номер точки в S_{learning}, N_{learning} — объем этой выборки. S_{learning} порождена неизвестной функцией f: $Y_i = f(X_i)$. Задача состоит в построении функции $\hat{f}(X) = \hat{f}(X|S_{\text{learning}})$, которая должна быть близка к оригинальной функции: $\hat{f}(X) \approx f(X)$. В силу того, что функция $f(X)$ в общем случае неизвестна, в качестве меры близости подсчитывают среднюю ошибку по некоторой тестовой выборке:

$$Q\left(\hat{f}|S_{\text{learning}}\right) = \sqrt{\frac{1}{N_{test}} \sum_{\{X,Y\} \in S_{test}} \left(Y - \hat{f}(X)\right)^2}.$$

Модель разложения по словарю. Аппроксимация $\hat{f}(X)$ строится как разложение по словарю параметрических функций:

$$\hat{f}(X) = \sum_{j=1}^{p} \alpha_j \psi_j(X) + \alpha_0. \qquad (1)$$

Каждая из функций словаря $\psi_j(X)$ задается набором параметров Θ_j. Кроме того, в отличие от стандартных подходов к аппроксимации, в изучаемой модели словарь состоит из функций нескольких типов. Пусть $x_1, \ldots, x_n$ — это компоненты вектора X, тогда опишем используемые функции:

1) Сигомидальные функции, $\Theta_j = \{\Gamma_j, \gamma_j^0\}$:

$$\psi_j(X) = \sigma\left(\Gamma_j^{\intercal} X + \gamma_j^0\right), \quad \sigma(x) = \frac{e^x - 1}{e^x + 1}.$$

2) Гауссовские функции, $\Theta_j = \{C_j, \sigma_j\}$:

$$\psi_j(X) = \exp\left(-\|X - C_j\|_2^2 / \sigma_j^2\right).$$

3) Линейные функции, $\Theta_j = \{\beta_j\}$:

$$\psi_j(X) = x_{\beta_j}, \quad \beta_j = 1, 2, \ldots, n.$$

Построение аппроксимации выполняется в два этапа:

1) Инициализация. Для каждой функции, входящей в словарь, должны быть заданы ее тип и начальные значения параметров.
2) Итеративная подстройка параметров аппроксиматора (коэффициентов разложения по словарю и параметров функций).

Более детально эти этапы описываются в следующих параграфах.

Инициализация. Инициализация гауссиан основана на применении регрессионного дерева.

Оно используется для разбиения пространства дизайна на непересекающиеся области. В центре каждой области размещается гауссиана, затем из полученного набора с помощью жадного алгоритма выбирается оптимальный поднабор [5].

В отличии от гауссиан сигмоиды не обладают свойством локальности, поэтому описанный выше подход к ним неприменим. Широкое распространение получили методы случайной инициализации параметров в некотором диапазоне. Границы этого диапазона выбираются из некоторых эмпирических соображений, в основе которых лежит попытка обеспечить наиболее стабильный процесс итеративного обучения. В исследуемом алгоритме применяется модификация метода SCAWI на основе использования латинских гиперкубов в сферических координатах [6].

Помимо инициализации параметров необходимо выбрать, в каких пропорциях в словаре будут использованы различные типы функций. Этот вопрос будет исследован в следующем разделе.

Алгоритм обучения. Для борьбы с переобучением используется классический подход: выборка данных S_{learning} разбивается на части S_{train} и $S_{\text{validation}}$ [2]. Первая из этих выборок используется для настройки параметров, вторая — для контроля за переобучением.

Пусть словарь некоторым образом инициализирован, то есть определены типы входящих в него функций и заданы их параметры. После этого обучение заключается в итеративном изменении параметров, минимизирующее функцию ошибки на множестве S_{train}. В силу вида аппроксимации (1), нахождение оптимальных коэффициентов разложения по словарю α_j — это задача линейной регрессии, которая решается аналитически. Сформулируем алгоритм обучения:

1) инициализируем параметры аппроксиматора;
2) обновляем параметры функций словаря Θ_j с помощью метода RPROP [7], используя градиент функции ошибки на множестве S_{train};
3) подсчитываем оптимальные коэффициенты разложения α_j;
4) если не началось переобучение (ошибка на множестве $S_{\text{validation}}$ не начала расти), возвращаемся к пункту 2.

Опишем более подробно алгоритм подсчета α_j. Пусть $\psi_j(X)$ — это j-я функция словаря, а $\boldsymbol{\Psi}$ — матрица значений функций словаря в точках обучающей выборки S_{train}. Тогда:

$$\Psi_{i,j} = \psi_j(X_i),\ i = 1, \ldots, N_{\text{train}},\ j = 1, \ldots, p. \quad (2)$$

Для подсчета коэффициентов используется ridge-регрессия:

$$\boldsymbol{\alpha} = \left(\boldsymbol{\Psi}^{\intercal}\boldsymbol{\Psi} + \lambda \boldsymbol{I}_p\right)^{-1} \boldsymbol{\Psi}^{\intercal}\boldsymbol{Y}, \quad \boldsymbol{Y} \in S_{\text{train}}. \quad (3)$$

Параметр регуляризации λ на каждой итерации выбирается так, чтобы критерий GCV (некоторая оценка обобщающей способности, [2]) был минимален при условии, что обусловленность матрицы $(\boldsymbol{\Psi}^{\intercal}\boldsymbol{\Psi} + \lambda \boldsymbol{I}_p)$ не меньше заданного порога.

Методика объединения

Вернемся к проблеме определения состава словаря. В зависимости от класса функции $f(X)$ оптимальным может быть как использование только одного типа функций словаря, так и комбинации нескольких типов. Алгоритм объединения функций разных типов в одном аппроксиматоре должен обладать очевидным свойством: качество приближения не должно быть хуже по сравнению с аппроксиматорами, построенными на основе одного типа функций.

Можно предложить следующий алгоритм:

1) инициализируем словарь гауссиан;
2) инициализируем словарь сигмоидов;
3) объединяем оба словаря в один;
4) отбираем наиболее значимые функции;
5) обучаем полученный аппроксиматор.

Предложенный подход не учитывает тот факт, что в алгоритмах инициализации есть существенное отличие: сигмоиды располагаются случайно, в то время как гауссианы уже подстроены под данные. Это различие приводит к тому, что в оптимальном словаре будут преобладать гауссианы. Такой подход показывает достаточно низкое качество аппроксимации на тех задачах, которые хорошо приближаются словарем на основе сигмоидов.

От указанного недостатка можно избавиться, если подстроить случайно инициализированные сигмоиды под данные перед объединением. Для этого можно сформировать аппроксиматор на основе функций одного типа и затем провести предварительное, неточное обучение (например, выполнив небольшое количество итераций обучения). Чтобы избежать возникновения обратной ситуации, когда сигмоиды будут лучше подстроены под данные, обучим аналогичным образом гауссианы. Таким образом, модифицируем первые два пункта алгоритма:

1) инициализируем словарь гауссиан и проводим грубое обучение;
2) инициализируем словарь сигмоидов и проводим грубое обучение.

Алгоритм отбора значимых функций

Способ отбора функций из объединенного словаря существенно влияет на итоговое качество ап-

проксимации. Фактически, необходимо решить задачу выбора наиболее значимых признаков. В качестве стандартных методов можно предложить [2]:

1) регрессия методом включения-исключения;
2) lasso;
3) жадный выбор по корреляции с невязками.

Основной недостаток этих подходов — подсчет коэффициентов разложения α_j отличным от алгоритма обучения образом, что может привести к выбору неоптимальной для последующего обучения модели. Опишем алгоритм, тесно связанный с методом подсчета коэффициентов разложения по словарю и лишенный этого недостатка. Он основан на модификации формулы (3) подсчета коэффициентов для стандартной ridge-регрессии. Добавка регуляризационной поправки λI_p эквивалентно минимизации функции ошибки со штрафом на большие значения коэффициентов:

$$Q_{\text{ridge}} = Q_{\text{train}} + \lambda \sum_{j=0}^{p} \alpha_j^2,$$

где $Q_{\text{train}} = \sum_{i=1}^{N_{\text{train}}} \left(Y_i - \hat{f}(X_i)\right)^2$ — ошибка аппроксимации на выборке S_{train}.

Рассмотрим штраф другого вида, в котором для каждого регрессора используется свой регуляризационный параметр:

$$Q_{ridge} = Q_{\text{train}} + \sum_{j=0}^{p} \lambda_j \alpha_j^2.$$

Минимизация этой функции ошибки эквивалента подсчету коэффициентов разложения по формуле

$$\boldsymbol{\alpha} = \left(\boldsymbol{\Psi}^{\mathsf{T}}\boldsymbol{\Psi} + \boldsymbol{\Lambda}\right)^{-1} \boldsymbol{\Psi}^{\mathsf{T}}\boldsymbol{Y}, \tag{4}$$

где $\boldsymbol{\Lambda}$ — диагональная матрица, состоящая из λ_j. Этот подход можно использовать для отбора наиболее значимых признаков, поскольку $\lambda_j = \infty$ равносильно $\alpha_j = 0$, то есть удалению регрессора $\psi_j(X)$ из модели.

Опишем алгоритм выбора $\boldsymbol{\Lambda}$, который позволяет получить оптимальную в некотором смысле модель. Выпишем формулу для GCV — оценки обобщающей способности аппроксимации, которая используется для выбора единого параметра регуляризации λ во время обучения [2]:

$$\text{GCV}\left(\boldsymbol{\Lambda}\right) = \\ = \frac{\sum_{i=1}^{N_{\text{train}}} \left(Y_i - \boldsymbol{\Psi}\boldsymbol{\alpha}\left(\boldsymbol{\Lambda}\right)\right)^2}{N_{\text{train}}\left(1 - \text{trace}\left(\left(\boldsymbol{\Psi}^{\mathsf{T}}\boldsymbol{\Psi} + \Lambda\right)^{-1}\boldsymbol{\Psi}^{\mathsf{T}}\boldsymbol{\Psi}\right)\right)^2}. \tag{5}$$

Следует отметить, что коэффициенты разложения также являются функцией от $\boldsymbol{\Lambda}$, см. формулу (4). Таким образом, $\text{GCV}\left(\Lambda\right)$ — это достаточно сложная функция, минимум по $\boldsymbol{\Lambda}$ которой не находится аналитически.

Рассмотрим поведение функции $\text{GCV}\left(\boldsymbol{\Lambda}\right)$, выбрав некоторое $j_0 \in \{1, \ldots, p\}$ и положив все λ_j, $j = 1, \ldots, p$, $j \neq j_0$ зафиксированными. Можно доказать, что уравнение (5) может быть сведено к следующему виду:

$$\text{GCV}\left(\lambda_{j_0}\right) = \frac{a\lambda_{j_0}^2 - 2b\lambda_{j_0} + c}{\left(d\lambda_{j_0} - e\right)^2}. \tag{6}$$

В формуле $a = a\left(\lambda_j\right), \ldots, e = e\left(\lambda_j\right)$, $j \neq j_0$ — это некоторые функции, не зависящие от λ_{j_0}. Функцию $\text{GCV}\left(\lambda_{j_0}\right)$ можно минимизировать по λ_{j_0} явно. С учетом того, что рассматривается только неотрицательные значения параметра регуляризации, возможны три типа оптимальных значений:

1) $\lambda_{j_0}^{opt} = 0$ — функция $\text{GCV}\left(\lambda_{j_0}\right)$ монотонно возрастает при $\lambda_{j_0} > 0$;
2) $\lambda_{j_0}^{opt} = \infty$ — функция $\text{GCV}\left(\lambda_{j_0}\right)$ монотонно убывает при $\lambda_{j_0} > 0$;
3) $\lambda_{j_0}^{opt} = \tilde{\lambda}_{j_0}$ — функция $\text{GCV}\left(\lambda_{j_0}\right)$ имеет минимум в точке $0 < \tilde{\lambda}_{j_0} < \infty$.

Таким образом, для части $j_0 \in \{1, \ldots, p\}$ будет справедливо $\lambda_{j_0}^{opt} = \infty$, что эквивалентно удалению регрессора ψ_j из модели. Сформулируем алгоритм объединения регрессоров нескольких типов.

1) Введем индекс $b = \{s, g, l\}$, задающий тот тип функций, для которого после грубой подстройки ошибка $Q_{\text{validation}}$ была наименьшей среди всех типов регрессоров;
2) Подсчитываем значение параметра регуляризации λ^b стандартной ridge-регрессии (3) для матрицы регрессоров Ψ^b, которая состоит из значений регрессоров типа l в точках обучающей выборки, см. (2);
3) Объединяем матрицы регрессоров $\boldsymbol{\Psi} = \{\boldsymbol{\Psi}^s, \boldsymbol{\Psi}^g, \boldsymbol{\Psi}^l\}$. Пусть $\tilde{p}$ — это суммарное количество регрессоров (столбцов матрицы Ψ);
4) Инициализируем матрицу регуляризации:

$$\Lambda_{j,k} = \begin{cases} \lambda_j, & j = k; \\ 0, & j \neq k, \end{cases} \quad j,\ k = 1 \ldots \tilde{p};$$

$$\lambda_j = \begin{cases} \lambda^b, & \text{если тип регрессора } \psi_j \text{ есть } b; \\ \infty, & \text{в противном случае.} \end{cases}$$

5) Подсчитываем коэффициенты разложения с текущей регуляризацией $\boldsymbol{\Lambda}$ по формуле (4).
6) Находим начальное значение GCV.
7) Для каждого $j = 1, \ldots, p$ выполняем:
   (а) Минимизируем $\text{GCV}\left(\lambda_j\right)$ (5) и находим λ_j^{opt};
   (б) Сохраняем полученное оптимальное значение $\lambda_j = \lambda_j^{\text{opt}}$;

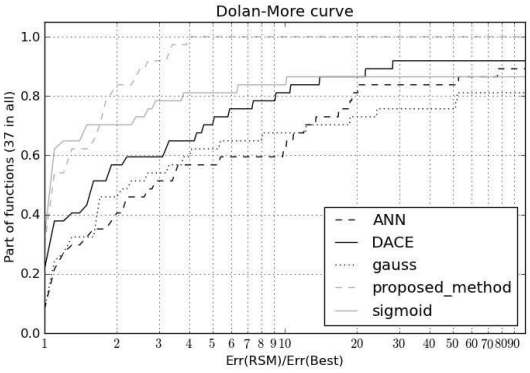


Рис. 1. Кривые качества алгоритмов аппроксимации.

(в) Находим новые коэффициенты разложения и обновленное значение GCV. При подсчете обращения $\left(\mathbf{\Psi}^{\mathsf{T}}\mathbf{\Psi}+\mathbf{\Lambda}\right)^{-1}$ учитываем, что в матрице $\mathbf{\Lambda}$ изменился только один элемент и применяем формулу Woodbury.

8) Составляем словарь объединенного аппроксиматора из тех ψ_j, для которых $\lambda_j \neq \infty$.

Экспериментальные результаты

Сравним описанный подход со следующими подходами к аппроксимации (в скобках будет дано *обозначение* на графике):

1) Разложение по словарю сигмоидов (*sigmoid*);
2) Разложение по словарю гауссиан (*gauss*);
3) Искусственная нейронная сеть с одним скрытым слоем, обучаемая алгоритмом Левенберга-Марквардта [9] (*ANN*);
4) Гауссовские процессы, реализованные в популярном пакете DACE [8] (*DACE*).

В качестве тестовых задач был использован набор многомерных функций [10]. Для визуального сравнения качества работы алгоритмов аппроксимации на наборе задач построим кривые Долана–Мора [11]. Эти кривые показывают, на какой доле задач ошибка метода не превосходит $x\,Q_{\min}$, где $Q_{\min}$ — минимальная среди методов ошибка для данной задачи, значение множителя x отложено по оси абсцисс. Таким образом, значение функции при $x = 1$ показывает долю задач, на которых алгоритм достиг наименьшей ошибки.

Можно отметить, что среди исследованных задач в наихудшем случае алгоритм получил ошибку, в четыре раза превышающую минимальную. В то же время для всех остальных методов аппроксимации есть выборки, на которых ошибка превышает минимальную на два порядка.

Выводы

В работе предложен алгоритм построения аппроксимации на основе разложения по словарю нескольких типов. Подход к объединению в словаре функций нескольких типов вычислительно эффективен и тесно связан с алгоритмом обучения, что позволяет получить более высокое качество аппроксимации по сравнению с аппроксиматорами на основе словаря из функций одного типа.

К недостатку метода можно отнести одну особенность процедуры оптимизации матрицы регуляризации. Во время итеративного процесса подстройки значений элементов этой матрицы регрессоры не упорядочены и перебор по функциям словаря идет в том порядке, в котором они были инициализированы. Однако более разумным подходом представляется сортировка функций по значимости перед началом этого процесса. Для реализации этой идеи необходимо предложить критерий значимости в рамках той же модели, которая используется при выборе оптимального набора.

Литература

[1] *Forrester A., Sobester A., Keane A.* Engineering Design via Surrogate Modelling. A Practical Guide // Wiley, 2008. — 238 p.

[2] *Hastie T., Tibshirani R., Friedman J.* The elements of statistical learning: data mining, inference, and prediction // Springer, 2008. — 763 p.

[3] *Burnaev E. V., Belyaev M. G., Prihodko P. V.* About hybrid algorithm for tuning of parameters in approximation based on linear expansions in parametric functions // Proceedings of the Intellectualization of information processing conference IIP-2010, Moscow, 2010.

[4] *Burnaev E. V., Belyaev M. G., Prihodko P. V., Chernova S. S.* Neural Approximation based on regression and boosting // Proceedings of the VI International Conference MMR-2009, Moscow, 2009. — Pp. 119-123.

[5] *Orr M., Hallam J., and others.* Combining Regression Trees and Radial Basis Function Networks // International Journal of Neural Systems. — 2000. — V. 10, N. 6. — Pp. 453–465.

[6] *Drago G., Ridella S.* Statistically controlled activation weight initialization // IEEE Transactions on Neural Networks. — 1992. — V. 3, N. 4. — Pp. 627–631.

[7] *Igel C., Husken M.* Improving the Rprop Learning Algorithm // Proceedings of the Second International Symposium on Neural Computation, NC'2000, ICSC Academic Press, 2000. — Pp. 115–121.

[8] `www2.imm.dtu.dk/~hbn/dace/` — DACE — 2002.

[9] *Nocedal J., Wright S.,* Numerical Optimization, 2nd Edition. — Springer, 2006. — 664 p.

[10] *Molga M., Smutnicki C.* Test functions for optimization needs // `www.zsd.ict.pwr.wroc.pl/files/docs/functions.pdf`

[11] *Dolan E., More J.* Benchmarking optimization software with performance profiles // Mathematical Programmming, Ser. A 91, 2002. — Pp. 201-–213.

Методы инициализации параметров нелинейной регрессионной модели

Беляев М. Г.[1,2,3]***, Бурнаев Е. В.***[1,2,3]***, Ерофеев П. Д.***[1,3]***, Приходько П. В.***[1,2,3]
belyaevmichael@gmail.com, burnaev@iitp.ru, erofeev.paul@gmail.com, prikhodkop@gmail.com
Москва, [1]Институт Проблем Передачи Информации РАН им. Харкевича, [2]DATADVANCE,
[3]Московский Физико-технический Институт (Государственный Университет)

При построении нелинейной регрессионной модели необходимо правильно инициализировать её параметры. В данной работе проводится сравнение некоторых широко распространенных методов и нескольких новых подходов к инициализации аппроксимационной модели, представляющей из себя разложение по словарю параметрических функций специального вида. Результаты численных экспериментов позволяют утверждать, что один из предложенных подходов к инициализации дает улучшение качества конечной аппроксимации на специфическом классе функций (двумерные гладкие и разрывные функции с множеством особенностей в области определения). Однако, в общем случае ни один из методов инициализации, в том числе и общепризнаных, не показал сколько-нибудь значимого улучшения качества аппроксимации или времени обучения.

При построении нелинейных регрессионных моделей возникает несколько типичных задач [1]: 1) первичная обработка исходных данных; 2) выбор и инициализация параметров используемой регрессионной модели; 3) обучение (подстройка параметров) модели и 4) оценка точности полученной аппроксимации. Обучение чаще всего является итеративным процессом, так как в общем (нелинейном) случае не существует явных формул, позволяющих точно оценить параметры регрессионной модели по данным. Известно, что начальный выбор архитектуры модели и значений параметров влияет не только на общее время обучения, но и на качество конечной аппроксимации [2, 3]. В данной работе исследуется влияние инициализации параметров на примере нелинейной регрессионной модели, представляющей из себя разложение по словарю параметрических функций.

Введение

Введем некоторые обозначения. Пусть задана выборка данных: $S = \{(\mathbf{X}_i, y_i); i = 1, \ldots, N\}$, $\mathbf{X}_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}^1$, порожденная неизвестной функцией $y = f(\mathbf{X})$. Необходимо построить функцию $\hat{f}(\mathbf{X})$, которая будет близка к исходной функции $f(\mathbf{X})$ в смысле некоторой нормы (обычно это среднеквадратичная ошибка). Мы будем рассматривать аппроксимирующие функции вида:

$$\hat{f}(\mathbf{X}) = \sum_{j=1}^{p} V_j \sigma(\mathbf{X} \times \mathbf{W}_j^{\mathsf{T}} + d_j) + V_0, \quad \mathbf{W}_j \in \mathbb{R}^n \quad (1)$$

— разложение по словарю параметрических функций [4]. Здесь в качестве функции $\sigma(\cdot)$ выступают функции специального вида — сигмоиды (гиперболический тангенс). Настраиваемые параметры модели: p, V_0, V_j, $\mathbf{W}_j$ и d_j, $(j = 1, \ldots, p)$. Веса V_j могут быть однозначно определены по остальным параметрам модели с помощью решения линейной регрессионной задачи. Для обучения модели используется алгоритм RProp (Resilient Propagation).

Таким образом, в контексте рассматриваемой проблемы инициализации модели необходимо решать сразу две задачи: 1) подбор количества p функций для словаря; 2) инициализация параметров этих функций. При этом порядок решения этих задач может быть разным.

В данной работе в рамках решения поставленных задач рассмотрены два принципиально разных подхода: **рандомизированные методы инициализации**, широко используемые в подобных задачах [2, 5, 9], предполагают случайные значения параметров модели; **детерминированные методы инициализации**, учитывающие характерные особенности аппроксимируемой выборки, являются более предпочтительными в смысле повторяемости результатов. Рандомизированные алгоритмы имеют следующие преимущества: простота реализации и незначительные вычислительные затраты. Это позволило рандомизированному подходу получить широкое распространение [2]. Однако, если качество полученной модели и время, затраченное на обучение, оказываются приемлемым, то в случае рандомизированной инициализации, как показали эксперименты, не представляется возможным добиться приемлимой повторяемости результатов даже на одних и тех же данных.

Статья организована следующим образом. Во втором и третьем разделах подробно описаны рассматриваемые алгоритмы в свете предложенной классификации. Четвертый раздел посвящен результатам численных экспериментов. В последнем, пятом разделе, подведены итоги работы.

Рандомизированная инициализация

Широкое распространение для инициализации моделей типа (1) получили алгоритмы рандомизированной инициализации. В этом разделе будут рассмотрены некоторые наиболее известные из них.

Инициализация Нгуена–Видроу. Самым распространенным способом инициализации нели-

нейных моделей типа (1) является рандомизированный алгоритм NW, предложенный Нгуеном и Видроу [9]. Весам $\mathbf{W}_j$ и b_j присваиваются начальные значения так, чтобы активные области соответствующих сигмоидов были распределены примерно равномерно в пространстве регрессоров. Таким образом? каждый элемент матрицы весов $\mathbf{W}$ инициализируется числом из равномерного распределения:[1]

$$\mathbf{W}_j \sim U[-I,\, I]^n, \tag{2}$$

где $I = p^{\frac{1}{N}}$. Компоненты вектора $\mathbf{b}$ также выбираются из равномерного распределения $U[-I, I]$.

Инициализация SCAWI. Подход к инициализации весов, используемый Драго и Риделла[2] [5], напоминает алгоритм NW, но с другой границей значений:

$$\mathbf{W}_j \sim U[-I,\, I]^n, \tag{3}$$

где $I = 1.3/\sqrt{1 + N\nu^2}$, $\nu^2 = \frac{1}{n\,N}\sum_{i=1}^{N}\sum_{j=1}^{n} x_{ij}^2$. Такая инициализация позволяет гарантировать, что значения аргументов сигмоидов будут находится в области ненасыщения сигмоида, при этом оказываясь значительно отличными от нуля. Компоненты вектора $\mathbf{b}$ выбираются аналогично предыдущему методу из равномерного распределения.

Сферическая инициализация. Для многомерных пространств покомпонентная случайная генерация векторов приводит к их кластеризации, порождая кластеризацию направлений сигмоидов. Существуют теоретические результаты [6], согласно которым наилучшая аппроксимация получается в случае равномерного распредления направлений по сфере. Представим веса модели (1) в виде $\mathbf{W}_j = R_j\mathbf{S}_j$, где $\mathbf{S}_j$ — случайный вектор, расположенный на единичной сфере, а R_j — некоторый радиус. Предлагается использовать следующую схему сферической инициализации весов (SWI). На первом этапе получаем значения углов φ_k $(k = 1, \ldots, n-1)$ из случайного распределения $U[-\pi, \pi]$, а затем переходим в декартовы координаты:

$$\begin{gathered} w_1 = R\cos(\varphi_1); \\ w_2 = R\sin(\varphi_1)\cos(\varphi_2); \\ \ldots \\ w_{n-1} = R\sin(\varphi_1)\ldots\sin(\varphi_{n-2})\cos(\varphi_{n-1}); \\ w_n = R\sin(\varphi_1)\ldots\sin(\varphi_{n-2})\sin(\varphi_{n-1}). \end{gathered}$$

Радиус по аналогии со SCAWI предлагается выбирать равным $R = \frac{1.3}{\sqrt{1+N\nu^2}}$. Компоненты вектора $\mathbf{b}$ выбираются аналогично методу Нгуена–Видроу из равномерного распределения.

Подбор числа сигмоидов. Приведенные алгоритмы не позволяют ответить на вопрос, сколько сигмоидов необходимо для построения аппроксимации. Предлагается два варианта решения этой проблемы: 1) подбор числа сигмоидов по сетке (по минимальной ошибке на валидационном множестве); 2) использование жадного набора сигмоидов [7] по критерию минимальной ошибки (до тех пор пока ошибка аппроксимации на валидационном множестве не начнет возрастать) или наибольшей корреляции. Последний подход предполагает решение задачи: $\min_{\mathbf{V}} \|\mathbf{V}\|_0$ при условии $\Xi\mathbf{V} = \mathbf{Y}$, где Ξ — матрица, состоящая из значений построенных сигмоидов в точках выборки: $\Xi_{ij} = \sigma(\mathbf{X}\mathbf{W}_j^{\intercal} + d_j)$, $i = 1, \ldots, N$, $j = 1, \ldots, p$; $\mathbf{Y} = [y_1, \ldots, y_N]^{\intercal}$ — вектор-столбец целевых переменных.

Выбор начального положения центров сигмоидов.

Во всех описанных ранее алгоритмах центры сигмоидов выбираются случайно (вследствие случайности выбора вектора $\mathbf{b}$). Однако, в связи с тем, что аппроксимируемая функция задана в ограниченном числе точек, разумно предположить, что мы сможем построить эффективное приближение функции только вблизи от этих точек. Поэтому правильно было бы располагать активные области сигмоидов рядом с точками выборки. Это легко сделать, если переписать функцию активации в следующем виде: $\sigma(\mathbf{X}\mathbf{W}_j^{\intercal} + b_j) = \sigma\big((\mathbf{X} - \mathbf{X}_0)\mathbf{W}_j^{\intercal}\big)$, где $b_j = \mathbf{X_0}\mathbf{W}_j^{\intercal}$, $\mathbf{X}_0$ — некоторая точка выборки.

Детерминированная инициализация

Рандомизированные алгоритмы инициализации имеют существенный недостаток: качество и время обучения могут существенно отличаться для двух разных запусков на одних и тех же данных. Также остается открытым вопрос подбора оптимального числа p функций в конечном словаре. В этом разделе приведено описание разработанных детерминированных алгоритмов инициализации, позволяющих решить эти проблемы.

Инициализация на основе локальных особенностей исходных данных.

В основе этого алгоритма лежит идея о том, что исходно сигмоиды необходимо располагать в тех областях, где аппроксимируемая функция имеет локальные особенности. Алгоритм может быть описан следующим образом.

Входные параметры алгоритма инициализации: выборка для обучения S и (опционально) количество требуемых сигмоидов: p[3].

[1] Здесь и далее будем считать, что в ходе предварительной обработки данных пространство регрессоров ограничено гиперкубом $[-1, 1]^n$.

[2] Алгоритм также известен под названием SCAWI (Statistically Controlled Activation Weight Initialization).

[3] Этот параметр задается, если подбор числа сигмоидов осуществляется по сетке.

Выходные параметры алгоритма: количество сигмоидов p, матрица весов $\mathbf{W} = [\mathbf{W}_j]_{j=1}^{p}$ и вектор весов $\mathbf{d} = [d_1, ..., d_p]$.

Локальная аппроксимация. На этом шаге для каждой точки $\mathbf{X}_i \in S$ исходной обучающей выборки строится локальная аппроксимация с помощью одного сигмоида. Для этого формируется веса

$$p_j = \frac{\exp\left(-\sum_{m=1}^{d} \frac{(x_{im}-x_{jm})^2}{h_m^2}\right)}{\sum_{l=1}^{N} \exp\left(-\sum_{m=1}^{d} \frac{(x_{im}-x_{lm})^2}{h_m^2}\right)},$$

где h_m — ширина ядра, которую можно оценить, например, по классической формуле Боумана–Аззалини [8]: $h_m = s_m \{4/(n+2)N\}^{\frac{1}{n+4}}$, где s_m — оценка стандартного отклонения по m-ой компоненте выходных векторов обучающей выборки, $j = 1, \ldots, N$. Затем решается задача линейной аппроксимации:

$$\min_{\mathbf{W}_i} \sum_{j=1}^{N} p_j^2 \left\| \sigma^{-1}\left(y_j/V_i\right) - (\mathbf{X}_j - \mathbf{X}_i)\mathbf{W}_i^{\intercal} \right\|_2^2,$$

где V_i подбирается по равномерной сетке. Таким образом, в каждую точку обучающей выборки ставится свой сигмоид, описывающий локальные особенности аппроксимируемой функции вблизи этой точки.

Отбор сигмоидов. После того как построены все сигмоиды из них необходимо выбрать наиболее коррелированные с заданными целевыми переменными. Если задан параметр p, то путем жадного набора, предложенного ранее формируется словарь из p сигмоидов. Если этот параметр не задан, то подбирается оптимальное (по ошибке на валидационном множестве) число сигмоидов для начальной аппроксимации.

Как показали опыты, качество конечной аппроксимации существенно зависит от ширины ядра h_m. И даже небольшие изменения ширины ядра h_m могут привести к значительному изменению качества аппроксимации.

Численные эксперименты

В этом разделе приведены описание и результаты численных экспериментов для разных вариантов инициализации: Нгуена–Видроу (NW), Драго–Риделла (SCAWI(1,2)), сферической инициализации (SWI) и инициализации на основе локальных особенностей исходных данных (DWI). Инициализация SCAWI(1) отличается от SCAWI(2) тем, что в первой сигмоиды располагаются случайно (оригинальный алгоритм), а во второй — в точках выборки. Для сравнения выбранных алгоритмов были использованы данные, полученные с помощью типичных в задачах нелинейной регрессии двумерных функций ($x_i \in [-1, 1], i = 1, 2$):

$$f_1(x_1, x_2) = \frac{\sin^2(\sum_{i=1}^{2}(x_i + 0{,}6)^2 - 0{,}3)}{\tanh[\sum_{i=1}^{2}\left((x_i + 0{,}6)^2 - 0{,}3\right)^2 + 0{,}4]};$$

$$f_2(x_1, x_2) = \frac{x_1 + x_2}{1 + 4(x_1^2 + x_2^2)};$$

$$f_3(x_1, x_2) = \sum_{i=1}^{2} x_i + 1 \cdot (\sum_{i=1}^{2} x_i^2 < 0{,}25) - 2 \cdot (\sum_{i=1}^{2}(x_i - 0{,}7)^2 < 1);$$

$$f_4(x_1, x_2) = \left((6x_1)^2 + 6x_2 - 11\right)^2 + \left(6x_1 + (6x_2)^2 - 7\right)^2;$$

Результаты численных экспериментов приведены в таблицах 1, 2 и 3. Эксперименты проводились на 20 случайных выборках мощностью 300 точек по 10 запусков на каждой — для рандомизированных алгоритмов и по одному запуску — для детерминированных. Все значения, приведенные в таблицах, являются десятичными логарифмами отношения соответствующих абсолютных значений к значениям, полученным при эталонной инициализации Нгуена–Видроу (NW)[4]. В таблице 1 приведены значения логарифмов отношения медиан среднеквадратичных ошибок конечной аппроксимации к соответствующим значениям при эталонной инициализации по всем запускам для каждой функции и для каждого метода инициализации. Аналогичные значения для 95 % квантилей приведены в таблице 2. Очевидно, что ни один из алгоритмов инициализации не дает существенного выигрыша в качестве конечной модели. Тем не менее расстановка центров сигмоидов в точках выборки дает небольшое стабильное улучшение. Если проводить сравнение по времени обучения, то несомненно лидирует детерминированный алгоритм инициализации, в то время как остальные алгоритмы по этой характеристике значимо не отличаются. Однако детерминированный алгоритм проигрывает по качеству аппроксимации конечной модели. На рассматриваемых двумерных функциях также не было выявлено существенного отличия сферической инициализации от других рандомизированных алгоритмов.

Заключение

В данной статье были предложены новые методы инициализации нелинейной регрессионной модели, представляющей из себя разложение по словарю параметрических функций специального ви-

[4]Таким образом, если некоторое значение близко к 0, то данная инициализация не отличается от эталонной, по данной характеристике; если значение имеет порядок 1, то данная инициализация имеет конечную ошибку аппроксимации (или время обучения), которая на порядок больше, чем при эталонной инициализации; если значение близко к -1, то инициализация имеет конечную ошибку аппроксимации (или время обучения), которая на порядок меньше, чем при эталонной инициализации.

Таблица 1. Качество конечной аппроксимации (ошибка среднеквадратичная) по отношению к эталонной.

Данные	SCAWI(1)	SCAWI(2)	SWI	DWI
f_1	0,08	0,07	0,20	0,57
f_2	-0,04	-0,08	-0,07	0,57
f_3	0,01	-0,00	-0,00	-0,03
f_4	0,14	-0,08	0,19	0,28

Таблица 2. Качество конечной аппроксимации (95 % квантиль абсолютной ошибки) по отношению к эталонной.

Данные	SCAWI(1)	SCAWI(2)	SWI	DWI
f_1	0,15	0,12	0,16	0,53
f_2	-0,03	-0,05	-0,10	0,55
f_3	0,00	-0,01	0,00	-0,01
f_4	0,06	-0,18	0,20	0,21

Таблица 3. Время обучения модели по отношению к эталонному.

Данные	SCAWI(1)	SCAWI(2)	SWI	DWI
f_1	-0,08	-0,14	-0,19	-0,61
f_2	0,00	-0,05	-0. 03	-0,77
f_3	0,12	0,15	0,07	-0,34
f_4	-0,06	-0,04	-0,14	-0,07

да. Также проведено сравнение предложенных алгоритмов с наиболее распространенными алгоритмами инициализации моделей подобного рода. Предложенный подход расстановки центров сигмоидов в точках выборки оказался эффективным на классе рассматриваемых функций. Однако, несмотря на разумный подход к проведению детерминированной инициализации, позволивший существенно сократить время последующего обучения, качество аппроксимации конечной модели при такой инициализации оказывалось хуже.

Следует отметить, что обучение модели проводилось с помощью алгоритма RProp. Использование других алгоритмов обучения может изменить характер зависимости конечной аппроксимации от начальной инициализации.

Литература

[1] *Bates D. M., Watts D. G.* Nonlinear Regression Analysis and Its Applications // Wiley Series in Probability and Statistics. — NY: Wiley, 1988. — V. 32. — p. 365.

[2] *Fernandez-Redondo M., Hernandez-Espinosa C.* Weight initialization methods for multilayer feedforward // Proc. of the 9th European Symposiumon Artificial Neural Networks. — 2001. — Pp. 25–27.

[3] *Thimm G., Fiesler E.* Optimal Setting of Weights, Learning Rate, and Gain // DIAP Research Rep. — 1997. — Pp. 97–04.

[4] *Burnaev E., Belyaev M., Prikhodko P.* About hybrid algorithm for tuning of parameters in approximation based on linear expansion in parametric functions // Intellectualization of inform. proc. conference. — V. 1. — 2010.

[5] *Drago G., Ridella S.* Possibility and Necessity Pattern Classification using an Interval Arithmetic Perceptron // Neural Computing & Applications. — 1999. — V. 8. — Pp. 40–52.

[6] *Maiorov V., Oskolkov K., Temlyakov V.* Gridge approximation and Radon compass // Approxim. Theory, Ed. B. Bojanov, DARBA. — 2002. — Pp. 284–309.

[7] *Bruckstein A. M., Donoho D. L., Elad M.* From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images // SIAM Review. — 2009. — V. 51, — P. 34.

[8] *Bowman A. W., Azzalini A.* Applied Smoothing Techniques for Data Analysis: The Kernel approach with S-Plus Illustrations // Oxford University Press, USA. — 1997.

[9] *Nguyen D., Widrow B.* Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights // IJCNN International Joint Conference on NN, 1990. — Pp. 21–26.

Задача выбора многоуровневых моделей с анализом ковариационной матрицы параметров*

Стрижов В. В.
strijov@ccas.ru
Вычислительный центр РАН, Москва, Россия

Обсуждается метод выбора активного набора признаков и фильтрации объектов выборки при восстановлении регрессии. Предполагается, что элементы рассматриваемой выборки естественным образом были разбиты на подмножества; для каждого из которых имеется своя, отличная от других, гипотеза порождения данных. Задача заключается в том, чтобы определить это разбиение и восстановить регрессионную модель для каждой подвыборки. При этом оценивается ковариационная матрица параметров каждой модели, и на основании анализа этой матрицы определяется вероятность принадлежности некоторого объекта данной подвыборке, а и некоторого признака — данной модели.

Введение

Работа опирается на следующие результаты. Предположим, что измеряемых свободных переменных недостаточно для восстановления адекватной регрессионной модели. Для пополнения их набора используем порождающие функции и вводим при этом меры их структурной сложности, аналогичные предложенным К. Владиславлевой [1].

В работе мы исходим из того, что процедура скользящего контроля недостаточно эффективна при решении прикладных задач. В случае, когда число измеряемых или порожденных признаков многократно превосходит объем выборки, однократное разбиение выборки не исключает переобучения модели и приводит к тому, что выборку приходится разбивать на несколько подвыборок: обучающую, тестовую, контрольную и так далее, как показано С. Ватанабе [2] и С. Арло [3].

Для выбора адекватной регрессионной модели используется функция правдоподобия модели, см. Д. МакКай [4]. Эта функция является составной частью связанного байесовского вывода, см. К. Бишоп [5]. Её использование согласуется с принципом минимальной длины описания, являющимся универсальным критерием выбора модели, см. П. Грюнвальд [6, 7]. Для оценки вероятности принадлежности признаков и объектов выборки к тем или иным моделям используются методы анализа ковариационных матриц, рассмотренные Дж. Нельдером [8]. Для оценки сходства двух и более моделей используется расстояние Дженсена-Шеннона, см. [9].

Предлагаемый метод заключается в следующем. Фиксируется класс моделей; порождается множество производных признаков. Индексы элементов выборки разбиваются на подмножества. Каждое из подмножеств соответствует модели. Число моделей выбирается таким, чтобы расстояние между моделями было статистически значимым [9]. Принадлежность элемента выборки к модели определяется по результатам анализа ковариационной матрицы зависимых переменных. Структура модели определяется по результатам анализа ковариационной матрицы параметров модели.

Результатом является многоуровневая модель оптимальной сложности — набор адекватных регрессионных моделей, описывающих выборку. В качестве иллюстрации приведена задача прогнозирования периодических временных рядов.

Постановка задачи

Задана выборка $D = \{(\mathbf{x}^i, y^i)\}$, проиндексированная $i \in \mathcal{I} = \{1, \ldots, m\}$. Элементы вектора $\mathbf{x}^i = [x_1^i, \ldots, x_j^i, \ldots, x_n^i] \in \mathbb{R}^n$ имеют индексы $j \in \mathcal{J}$. Само множество векторов представлено в виде матрицы плана $[\mathbf{x}^1, \ldots, \mathbf{x}^m]^T = X$, столбцы которой являются признаками и обозначаются нижним индексом: $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$. Соответственно, $D = (X, \mathbf{y})$, где матрица $X \in \mathbb{R}^{n \times m}$, а вектор $\mathbf{y} \in \mathbb{R}^m$.

Вектор зависимой переменной считается реализацией случайной величины; пусть ее распределение принадлежит семейству экспоненциальных распределений. Обозначим $\mathbf{f}$ — вектор восстановленных значений зависимой переменной посредством некоторой неизвестной функции регрессии, $\mathbf{f} = [f(\mathbf{w}_{\text{MP}}, \mathbf{x}^1), \ldots, f(\mathbf{w}_{\text{MP}}, \mathbf{x}^m]^T$, в которой $\mathbf{w}_{\text{MP}}$ — вектор наиболее вероятных параметров. Вектор свободных переменных $\mathbf{x}$, согласно классической постановке задачи восстановления регрессии, см. Г. Себер [10], будем считать неслучайной величиной.

Регрессионной моделью f будем называть поэлементное отображение $f : (\mathbf{w}, \mathbf{x}) \mapsto \mathrm{f}$. В терминах отображения соответствующих множеств, модель $f : \mathcal{W} \times \mathcal{X} \to \mathcal{Y}$. Будем считать, что $\mathcal{W}, \mathcal{X} \subset \mathbb{R}^n$, а $\mathcal{Y} \subset \mathbb{R}^m$.

Задача выбора модели ставится как задача нахождения такой модели f из класса допустимых моделей $\mathcal{F}$, которая имела бы максимальное правдоподобие при наиболее вероятных параметрах:

$$\hat{f}(\mathbf{w}_{\text{MP}}|\mathbf{x}) = \arg \max_{f \in \mathcal{F}, \mathbf{x} \in D} \mathcal{E}\left(f(\mathbf{w}_{\text{MP}}, \mathbf{x})\right).$$

Работа выполнена при поддержке РФФИ, грант: 10-07-00422

Наиболее вероятные параметры

$$\mathbf{w}_{\text{МР}} = \arg\max_{\mathbf{w}\in\mathcal{W}} p(\mathbf{w}|D, f),$$

модели f оцениваются с помощью формулы Байеса, в которой апостериорное распределение параметров

$$p(\mathbf{w}|D, f) = \frac{p(D|\mathbf{w}, f, B)p(\mathbf{w}|f, A)}{\int p(D|\mathbf{w}', f, B)p(\mathbf{w}'|f, A)d\mathbf{w}'},$$

функция правдоподобия параметров $p(D|\mathbf{w}, f, B)$ задана распределением зависимой переменной $\mathbf{y}$. Априорное распределение параметров задано классом моделей $\mathcal{F}$ и гипотезой порождения данных.

В качестве примера приведем распределение $\mathbf{y} \sim \mathcal{N}(\mathbf{f}, B)$ и класс линейных или линеаризованных существенно-нелинейных моделей. Параметрическая функция $\mathcal{N}$ переводится линейным отображением заданным матрицей плана X или линеаризованной матрицей плана

$$\underset{m\times n}{J} = \left[\frac{\partial f(\mathbf{w}, \mathbf{x}^i)}{\partial w_j}\right].$$

также в функцию $\mathcal{N}$ (так как линейное отображение, заданное матрицами X или J, представимо в виде произведения $U\Lambda V^T$ ортогональной, диагональной и ортогональной матриц). Параметры распределения при этом, в общем случае, изменяются. Следовательно, многомерная случайная величина $\mathbf{w}$ имеет ту же функцию распределения $\mathbf{w} \sim \mathcal{N}(\mathbf{w}_{\text{МР}}, A)$. Прочие примеры подробно рассмотрены Нельдером [8].

Правдоподобие модели $\mathcal{E}\left(f(\mathbf{w}, \mathbf{x})\right) \stackrel{\text{def}}{=} p(D|f)$ — сомножитель правой части формулы Байеса второго уровня связного вывода, см. [11]. Согласно этому выводу, наиболее вероятная модель отыскивается исходя из сравнения апостериоых вероятностей

$$p(f|D) \propto p(D|f)p(f),$$

или же из сравнения правдоподобий моделей $p(D|f)$, считая их априорные вероятности равными.

Правдоподобие модели при этом задается выражением

$$\mathcal{E}'\left(f(\mathbf{w}, \mathbf{x})\right) = \int p(D|\mathbf{w}, f, B)p(\mathbf{w}|f, A)d\mathbf{w}.$$

Предлагается вычислять правдоподобие модели в окрестности ее наиболее правдоподобных параметров $\mathbf{w}_{\text{МР}}$, используя только подынтегральное выражение

$$\mathcal{E}(f(\mathbf{w}_{\text{МР}}, \mathbf{x})) = p(D|\mathbf{w}_{\text{МР}}, f, B)p(\mathbf{w}_{\text{МР}}|f, A). \quad (1)$$

Ковариационные матрицы A и B при этом предполагаются уже оцененными и зафиксированными на этапе нахождения наиболее правдоподобных параметров.

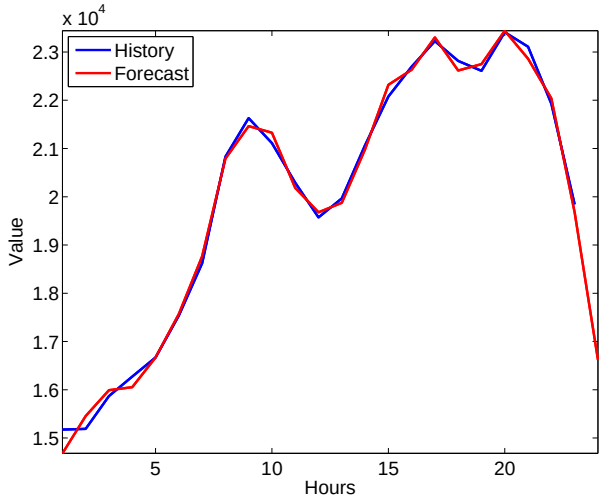


Рис. 1. Прогноз временного ряда $s(\tau)$ на 24 часа вперед

Выбор модели и фильтрация объектов

Линейная модель f однозначно задается активным множеством индексов признаков $\mathcal{A} \subseteq \mathcal{J}$. Предполагая частичную гомоскедаксичность выборки (например, среди объектов встречаются выбросы, которые должны быть исключены из рассмотрения), зададим «фильтрованную» выборку, иначе — активное множество объектов индексами $\mathcal{B} \subseteq \mathcal{I}$. Обозначим множество многомерных величин $\{\mathbf{x}^i|i \in \mathcal{B}\}$ как $\mathbf{x}^{\mathcal{B}}$. Задача выбора модели имеет вид

$$\mathcal{F} \ni \hat{f} = \arg\max_{\mathcal{A}\subseteq\mathcal{J},\mathcal{B}\subseteq\mathcal{I}} \mathcal{E}\left(f(\mathbf{w}_{\mathcal{A}}, \mathbf{x}^{\mathcal{B}})\right). \quad (2)$$

Способы решения этой задачи рассмотрены автором в [12]. Заметим, что для набору индексов признаков $\mathcal{J}$ мощности n соответствуют 2^n вершин двоичного куба. Каждая вершина задает некоторый активный набор признаков $\mathcal{A}$: считается, что j-й признак вошел в набор, если значение j-й координаты вершины единица. При решении задачи мы руководствуемся следующими предположениями.

1. Среди вершин куба существует по крайней мере одна, обозначим ее $\hat{\mathcal{A}}$, доставляющая матожидание правдоподобия модели.
2. От вершины $\mathcal{A} = \varnothing$ к вершине $\hat{\mathcal{A}}$ есть путь по ребрам куба (иначе — стратегия последовательного добавления-удаления признаков), который доставляет правдоподобию модели $\mathcal{E}\left(f(\mathbf{w}_{\mathcal{A}}, \mathbf{x})\right)$ сходимость по вероятности.

Множество индексов $\mathcal{B}$ задает выпуклую комбинацию $\{x_i|i \in \mathcal{B}\}$ — область $\mathcal{X}_{\mathcal{A}}$, «по крайней мере», в которой значения дисперсии $\{\beta_i|i \in \mathcal{B}\}$ за-

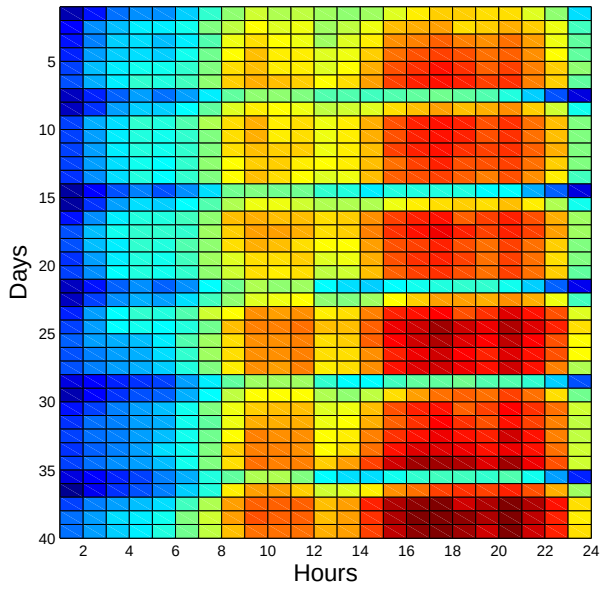


Рис. 2. Авторегрессионная матрица X, рабочие и выходные дни.

висимых переменных $\{\mathbf{y}_i | i \in \mathcal{B}\}$ меняются «незначительно». Другими словами, третий центральный момент, или коэффициент асимметрии случайной величины $\mathbf{y}$, соответствующей области $\mathcal{X}_\mathcal{A}$ равен нулю [13].

Выбор многоуровневых моделей

Многоуровневой моделью $\mathfrak{f}$ называется набор моделей $\mathfrak{f} = \{f_k | f \in \mathcal{F}\}$, $k = 1, \ldots, l$, такой, что

$$f_k : \mathcal{W}_k \times \mathcal{X}_{\mathcal{B}_k} \to \mathcal{Y}_{\mathcal{B}_k},$$

при разбиении $\mathcal{I} \supseteq \mathcal{B}^* = \sqcup \mathcal{B}_k$.

Введем функцию расстояния $\rho(f_k, f_l)$ между двумя моделями. Для этого используем дивергенцию Дженсена-Шеннона, в которой $\rho_{kl} \in [0, 1]$ является метрикой [9]:

$$\rho(p_k \| p_l) = 2^{-1} D_{\mathrm{KL}}\left(p_k \| p'\right) + 2^{-1} D_{\mathrm{KL}}\left(p' \| p_l\right),$$

где $p' = 2^{-1}(p_k + p_l)$ и здесь $p_k \stackrel{\text{def}}{=} (p(\mathbf{w}_\mathcal{A} | D, A, B, f_k)$. Несимметричная функция расстояния — дивергенция Кулльбака-Лейблера задана как

$$D_{\mathrm{KL}}\left(p \| p'\right) = \int\limits_{\mathbf{w} \in \mathcal{W}} p'(\mathbf{w}) \ln \frac{p(\mathbf{w})}{p'(\mathbf{w}) d\mathbf{w}}.$$

Отметим, что расстояние вводится только на моделях, имеющих одинаковый набор признаков $\mathcal{A}$.

Задача нахождения многоуровневых моделей ставится следующим образом:

$$\mathcal{F} \supset \hat{\mathfrak{f}} = \arg \max_{\mathcal{B}_1, \mathcal{B}_2 \subset \mathcal{B}} \rho(f_1, f_2) \tag{3}$$

при заданном множестве индексов признаков $\hat{\mathcal{A}}$, таком, что

$$\hat{\mathcal{A}} = \arg \max_{\mathcal{A} \subseteq \mathcal{J}} \mathcal{E}\left(f_1(\mathbf{w}_\mathcal{A}, \mathbf{x}^{\mathcal{B}_1})\right) \mathcal{E}\left(f_2(\mathbf{w}'_\mathcal{A}, \mathbf{x}^{\mathcal{B}_2})\right).$$

Иллюстрация: прогнозирование периодических временных рядов

Рассмотрим задачу авторегрессионного прогнозирования как одну из наиболее показательных при создании многоуровневой прогностической модели, в которых требуется одновременно выбрать объекты и признаки для каждой модели. Задан временной ряд $\{s(1), \ldots, s(\tau), \ldots, s(T-1)\}$, известен период $\varkappa$. Требуется спрогнозировать отсчет ряда в точке времени T. Для этого построим авторегрессионную матрицу X^* так, что ее строка i и столбец j отображались в номер отсчета как $(i-1)\varkappa \mapsto \tau$ при $\mathrm{mod}\frac{T}{\varkappa} = 0$. Представим X^* как матрицу, состоящую из присоединенных наборов векторов

$$X^* = \left[\begin{array}{c|c} \underset{m \times n}{X} & \underset{m \times 1}{\mathbf{y}} \\ \hline \underset{1 \times n}{\mathbf{x}^{m+1}} & \underset{1 \times 1}{s(T)} \end{array}\right].$$

Здесь X — матрица плана с числом столбцов $n = \varkappa - 1$ и $\mathbf{y}$ — последний столбец матрицы X^*. Принимая линейную модель зависимости $\mathbf{y} = X\mathbf{w}$, после оценки наиболее вероятного вектора параметров $\mathbf{w}$ получаем прогнозируемое значение $s(T) = \langle \mathbf{x}^{m+1}, \mathbf{w}_{\text{MP}} \rangle$.

Примем следующую гипотезу порождения данных: $\mathbf{y} \sim \mathcal{N}(\mathbf{f}, B)$ из которой следует $\mathbf{w} \sim \mathcal{N}(\mathbf{w}_{\text{MP}}, A)$. Тогда, при отсутствии гипотезы гомоскедаксичности регрессионных остатков и независимости элементов многомерной случайной величины $\mathbf{y}$, оптимизируемая функция S будет иметь вид

$$\begin{aligned} 2S(\mathbf{w} | D, f) =& (\mathbf{w} - \mathbf{w}_{\text{MP}})^T A (\mathbf{w} - \mathbf{w}_{\text{MP}}) + \\ &+ (\mathbf{f} - \mathbf{y})^T B (\mathbf{f} - \mathbf{y}). \end{aligned} \tag{4}$$

Учитываются также следующие предположения.

1. Существуют несколько типов периодов, каждый из которых должен быть спрогнозирован своей собственной моделью.
2. Не все фазы периода должны быть включены в модель.

Рисунки 1, 2 и 3 иллюстрирует результаты решения задач (2) и (3). На рис. 1 показан один период временного ряда и прогноз полученный этот период. На рис. 2 для каждого по следующего прогнозируемого значения показаны наиболее информативные признаки (имеющее наименьшие значения диагонали ковариационной матрицы A). Видно, что таковыми являются признаки, соответствующие столбцам авторегрессионной матрицы в окрестности периода данного прогнозируемого часа. На рис. 3 показана авторегрессионная матрица X. Ее строки (объекты выборки) можно условно разбить на два типа: соответствующие рабочим и выходным дням. Это дает основание для введения многоуровневой модели, состоящей из двух моделей одинаковой структуры с разными значениями параметров.

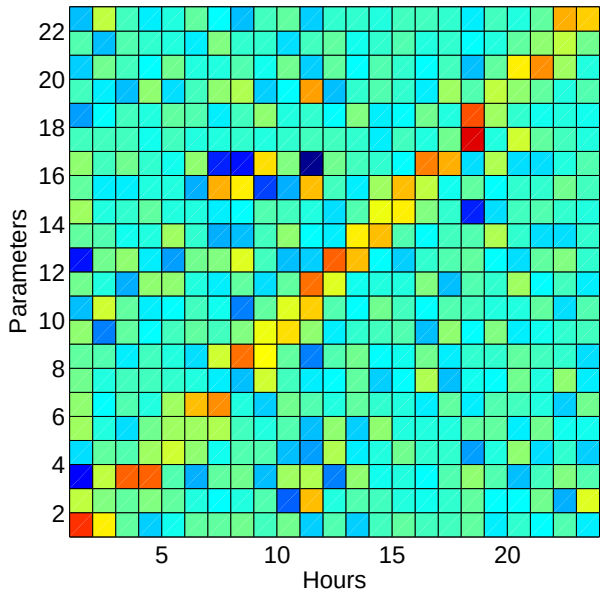


Рис. 3. Матрица информативности параметров $\mathbf{w}$ для различных значений времени τ.

Опишем алгоритм решения задачи (2).

1. Задаются единичные ковариационные матрицы A, B.
2. Для фиксированных значений матриц A, B оцениваются параметры $\mathbf{w}_{\text{MP}}$ модели f. При этом оптимизируется функция (4).
3. Оцениваются ковариационные матрицы A, B согласно гипотезе порождения данных.
4. Последние два шага повторяются до сходимости: пока изменение элементов матриц A, B не будут меньше заданных.
5. Выбираются те признаки $\mathcal{A}$ и объекты $\mathcal{B}$, которым соответствует наибольшие значения диагональных элементов матриц A, B.
6. Мощности множеств $\mathcal{A}, \mathcal{B}$ выбираются такими, чтобы они доставляли максимум функции правдоподобия (1).

Алгоритм решения задачи (3) состоит из двух основных шагов. Модели, включенные в $\mathfrak{f}$ заданы разбиением множества индексов $\mathcal{B}_1 \sqcup \mathcal{B}_2$, имеют различные ковариационные матрицы B_1, B_2 и общий набор признаков $\mathcal{A}$.

1. Решается задача максимизации правдоподобия $\mathfrak{f}$ на множестве $\mathcal{A}$ как в предыдущем алгоритме; разбиение $\mathcal{B}$ фиксировано.
2. Решается задача максимизации расстояния $\rho(f_1, f_2)$. Для этого значения диагональных элементов B_1, B_2 упорядочиваются по убыванию. Выполняется обмен b_1, b_2 индексами из разбиения $\mathcal{B}_1, \mathcal{B}_2$, соответствующими наименьшим значениям диагональных элементов. Числа b_1, b_2 выбираются такими, что расстояние $\rho(f_1, f_2)$ между двумя моделями было максимально.

Заключение

Рассмотренный метод позволяет решать задачу совместного выбора признаков и объектов как для одной регрессионной модели, так и для их набора. При этом особое внимание уделяется принятию статистических гипотез и, как следствие, корректности использования функций качества, с помощью которых отыскиваются оптимальные, в данном случае наиболее вероятные параметры моделей, а также их матрица их ковариаций.

Литература

[1] *Vladislavleva E., Smith G., Hertog D.* Order of nonlinearity as a complexity measure for models generated by symbolic regression via pareto genetic programming // *EEE Transactions on Evolutionary Computation.* — 2009. — Vol. 13(2). — Pp. 333–349.

[2] *Watanabe S.* Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory // *J. Machine Learning Research.* — 2010. — Vol. 11. — Pp. 3571–3594.

[3] *Arlot S., Blanchard G., Roquain E.* Some nonasymptotic results on resampling in high dimension // *Annals of Statistics.* — 2009. — Vol. 38 — Pp. 51–82.

[4] *MacKay D.* Information Theory, Inference, and Learning Algorithms. — Cambridge University Press, 2003.

[5] *Bishop C. M.* A new framework for machine learning // Computational Intelligence.— Springer, 2008. — Pp. 1–24.

[6] *Grunwald P. D.* The Minimum Description Length Principle. — MIT Press, 2007.

[7] *Grünwald P., Myung I. J., Pitt M.* Advances in Minimum Description Length. — MIT Press, 2005.

[8] *Lee Y., Nelder J. A., Pawitan Y.* Generalized linear models with random effects: unified analysis via h-likelihood. — Chapman & Hall/CRC, 2006.

[9] *Lin J.* Divergence measures based on the shannon entropy // *IEEE Transactions on Information Theory.* — 1991. — Vol. 37, no. 1. — P. 145.

[10] *Seber G. A. F., Wild C.* Nonlinear Regression. — Wiley-IEEE, 2003.

[11] *Strijov V., Weber G. W.* Nonlinear regression model generation using hyperparameter optimization // *Computers and Mathematics with Applications.* — 2010. — Vol. 60, no. 4. — Pp. 981–988.

[12] *Strijov V. V., Krymova E. A., Weber G. W.* Evidence optimization for consequently generated models // *Mathematical and Computer Modelling.* — 2011.

[13] *Rissanen J., Roos T., Myllymäki P.* Model selection by sequentially normalized least squares // *J. Multivariate Analysis.* — 2010. — Vol. 101, no. 4. — Pp. 839–849.

Выбор многоуровневых моделей в задачах банковскового кредитного скоринга*

Павлов К. В., Стрижов В. В.
kirill.pavlov@phystech.edu, strijov@ccas.ru
Московский физико-технический институт, Вычислительный центр Дородницина РАН, Москва, Россия.

Решается задача классификации с использованием логистической регрессии. Предлагается новый подход, заключающийся в совместной кластеризации объектов и выборе признаков моделей. Результатом подхода является многоуровневая модель — набор моделей оптимальной сложности. Для построения моделей предлагается использовать EM-алгоритм. На E-шаге происходит отнесение объектов к моделям. На M-шаге происходит выбор наиболее вероятных параметров модели. Алгоритм тестировался на данных кредитным займам наличными.

Введение

Данная работа посвящена проблеме выбора и настройки моделей логистической регрессии в задачах классификации. Авторы предлагают новый подход, заключающийся в совместной кластеризации объектов и выборе признаков многоуровневых моделей. Его результатом является набор моделей оптимальной сложности.

Известные подходы к выбору моделей заключаются в использовании шаговой регрессии с критерием Маллоуза [6], итеративного перевзвешивающего метода наименьших квадратов [2], порождения нелинейных регрессионных моделей [5].

Для построения моделей предлагается использовать EM-алгоритм [2]. На E-шаге происходит отнесение объектов к моделям на основе оценки правдоподобия многоуровневой модели. На M-шаге происходит выбор наиболее вероятных параметров модели по объектам, которые к ней отнесли.

Преимуществом данного подхода является его способность описывать принципиально многомодельные выборки и сегментировать объекты в соответствии с используемыми моделями. Алгоритм тестировался на модельных и реальных данных по кредитному займу наличными. Эксперименты показали преимущество использования многоуровневых моделей по сравнению с использованием одной модели.

Постановка задачи

Рассмотрим задачу восстановления регрессии

$$\mathsf{E}(y \,|\, \boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{w}), \tag{1}$$

в которой измеряемые данные представляют собой пары значений зависимой переменной y и независимой переменной $\boldsymbol{x}$. Зависимость f является функцией регрессии от независимой переменной $\boldsymbol{x}$, называемой регрессором, и вектора параметров $\boldsymbol{w}$. Задачей регрессионного анализа является нахождение функции f и параметров $\boldsymbol{w}$.

Определение 1. *Регрессионная выборка* $D = \{(\boldsymbol{x}^i, y^i)\}_{i=1}^m$ *— множество* m *пар, состоящих из векторов* $\boldsymbol{x}_i = (x_1, \ldots, x_n)^\intercal$ *и соответствующих этим векторам значений* y^i.

Далее предполагается, что переменные определены на подмножестве действительных чисел: $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^n,\ y \in \mathcal{Y} \subseteq \mathbb{R}$. Индексы элементов i и компонент вектора независимой переменной j являются элементами конечных множеств $i \in \mathcal{I} = \{1, \ldots, m\},\ j \in \mathcal{J} = \{1, \ldots, n\}$.

Определение 2. *Матрица плана* X *— матрица, строки которой есть компоненты независимой переменной* $\boldsymbol{x}$, $X = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^m)^\intercal$.

Регрессионную выборку, определенную в (1) будем обозначать $D = (X, \boldsymbol{y})$. Выборка может быть как функцией дискретного аргумента, так и отношением, при этом одному значению переменной $\boldsymbol{x}$ может соответствовать несколько значений переменной y. Для нахождения функции регрессии используется понятие регрессионной модели.

Определение 3. *Регрессионная модель — параметрическое семейство функций, отображающих декартово произведение областей определения объектов* $\mathcal{X}$ *и параметров* $\mathcal{W}$ *в область значений* $\mathcal{Y}$ *зависимой переменной*

$$f \colon \mathcal{X} \times \mathcal{W} \to \mathcal{Y}.$$

Определение 4. *Многоуровневой регрессионной моделью называется набор регрессионных моделей* f_k, $k = 1, \ldots, l$ *такой, что при разбиении множества индексов объектов* $\mathcal{I} = \sqcup \mathcal{I}_k$ *для всех объектов с индексами из* $\mathcal{I}_k$ *используется модель* f_k.

Ниже будут исследованы обобщенные линейные модели.

Обобщенные линейные модели

Впервые обобщенные линейные модели были введены Джоном Нельдером и Робертом Веддербурном в 1972г. [4]. В основе обобщенных линейных моделей лежат следующие предположения. Во-первых, считается, что зависимая переменная $\boldsymbol{y}$ имеет экспонентную плотность распределения [1] с вектором параметров $\boldsymbol{\theta}$,

$$p(\boldsymbol{y} \,|\, \boldsymbol{\theta}) = h(\boldsymbol{y}) g(\boldsymbol{\theta}) \exp\bigl(\boldsymbol{T}(\boldsymbol{y})^\intercal \boldsymbol{\eta}(\boldsymbol{\theta})\bigr), \tag{2}$$

Работа выполнена при поддержке РФФИ: 10-07-00422

где h, g, $\boldsymbol{T}$ и $\boldsymbol{\eta}$ — известные функции. Оказывается [1], что в случае экспонентного распределения и только в нем, $\boldsymbol{T}(\boldsymbol{y})$ является достаточной статистикой. Для удобства дальнейшей записи перепишем функцию плотности в следующем виде

$$p(\boldsymbol{y}\,|\,\boldsymbol{\theta}) = \exp\big(\boldsymbol{T}(\boldsymbol{y})^{\intercal}\boldsymbol{\eta}(\boldsymbol{\theta}) - b(\boldsymbol{\theta}) + c(\boldsymbol{y})\big). \quad (3)$$

Второе предположение заключается в том, что предиктор $\boldsymbol{\eta}$ линеен по координатам независимой переменной $\boldsymbol{x}$.

$$\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\theta}) = X\boldsymbol{w}. \quad (4)$$

Предполагается так же, что математическое ожидание $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$ зависимой переменной $\boldsymbol{y}$ есть монотонная функция вектора $\boldsymbol{\eta}$ [3]. При этом регрессионная модель имеет вид

$$\mathsf{E}(\boldsymbol{y}\,|\,\boldsymbol{\theta}) = \boldsymbol{\mu} = f(\boldsymbol{\eta}) = f(X\boldsymbol{w}). \quad (5)$$

Функция f называется функцией активации. В силу её монотонности существует обратная функция f^{-1}, которая называется функцией связи [2].

В частном случае экспонентного распределения $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$. При этом говорят [3], что распределение имеет каноническую форму. Функция плотности при этом

$$p(\boldsymbol{y}\,|\,\boldsymbol{\theta}) = \exp\big(\boldsymbol{T}(\boldsymbol{y})^{\intercal}\boldsymbol{\theta} - b(\boldsymbol{\theta}) + c(\boldsymbol{y})\big). \quad (6)$$

Для случая канонической формы можно выписать выражения математического ожидания и дисперсии достаточной статистики зависимой величины

$$\mathsf{E}\big(\boldsymbol{T}(\boldsymbol{y})\big) = \boldsymbol{\mu} = \nabla b(\boldsymbol{\theta}); \quad \mathsf{D}\big(\boldsymbol{T}(\boldsymbol{y})\big) = \nabla\nabla^{\intercal} b(\boldsymbol{\theta}). \quad (7)$$

В двухклассовой задаче классификации переменная y принимает два значения, $y \in \{0, 1\}$. Предположим, что зависимая величина принадлежит распределению Бернулли. Ниже рассмотрим этот случай.

Распределение Бернулли

Пусть случайная величина имеет распределение Бернулли с параметром p, $y \sim B(p)$, тогда

$$y = \begin{cases} 1, & p\,; \\ 0, & 1-p. \end{cases} \quad (8)$$

Покажем, что распределение Бернулли есть частный случай экспонентного распределения (6). Функция плотности $p(y\,|\,p)$ имеет вид

$$p(y\,|\,p) = p^y(1-p)^{1-y}. \quad (9)$$

Логарифмируя плотность получим функцию правдоподобия

$$l(y\,|\,p) = y\log p + (1-y)\log(1-p). \quad (10)$$

Сгруппируем члены

$$l(y\,|\,p) = y\log\frac{p}{1-p} + \log(1-p). \quad (11)$$

Полученное выражение имеет форму экспонентного семейства (6) для случая $\boldsymbol{T}(\boldsymbol{y}) = y$

$$\log p(y\,|\,p) = y\theta - b(\theta) + c(y) \quad (12)$$

со следующим соответствием: из вида первого слагаемового получим, что каноническй параметр соответствует логистической функции p:

$$\theta = \log\frac{p}{1-p}. \quad (13)$$

Решая данное уравнение получим, что

$$p = \frac{e^\theta}{1+e^\theta} = \sigma(\theta), \quad 1-p = \frac{1}{1+e^\theta} = \sigma(-\theta). \quad (14)$$

Во втором слагаемом

$$\log(1-p) = \log\Big(\frac{1}{1+e^\theta}\Big) = -\log(1+e^\theta),$$

откуда определяется функция $b(\theta)$:

$$b(\theta) = \log(1+e^\theta). \quad (15)$$

В случае распределения Бернулли $c(y) = 0$.

Проверим значения математического ожидания и дисперсии. Дифференцируя $b(\theta)$ получим

$$\mathsf{E}(y) = b^{'}(\theta) = \frac{e^\theta}{1+e^\theta} = p. \quad (16)$$

Вторая производная даст дисперсию

$$\mathsf{D}(y) = b^{''}(\theta) = \frac{e^\theta}{(1+e^\theta)^2} = p(1-p). \quad (17)$$

Для подбора параметров модели воспользуемся итеративным перевзвешивающим методом наименьших квадратов (IRLS).

Оценка правдоподобия модели

Рассмотрим распределение бернуллиевского случайного вектора $\boldsymbol{y}$ с независимыми компонентами $y_i \sim B(p_i)$. В рамках обобщенных линейных моделей натуральный параметр θ представляется как

$$\theta = \sum_{j=1}^{n} x_j w_j = \boldsymbol{x}^{\intercal}\boldsymbol{w}. \quad (18)$$

Функция плотности вектора $\boldsymbol{y}$ имеет вид

$$p(\boldsymbol{y}\,|\,\boldsymbol{w}) = \prod_{i=1}^{m} p_i^{y_i}(1-p_i)^{1-y_i}. \quad (19)$$

Определим функцию штрафа как минус логарифм правдоподобия

$$E(\boldsymbol{w}) = -\ln p(\boldsymbol{y} \mid \boldsymbol{w}) = \\ = -\sum_{i=1}^{m} y_i \ln p_i + (1-y_i)\ln(1-p_i). \tag{20}$$

Используя тождество

$$\frac{d\sigma(\theta)}{d\theta} = \sigma(1-\sigma) \tag{21}$$

и то, что $p = \sigma(\boldsymbol{x}^\intercal \boldsymbol{w})$, вычислим градиент функции штрафа.

$$\nabla E(\boldsymbol{w}) = -\sum_{i=1}^{m} \big(y_i(1-\sigma_i) - (1-y_i)\sigma_i\big)\boldsymbol{x}_i = \\ = \sum_{i=1}^{m} (\sigma_i - y_i)\boldsymbol{x}_i = X^\intercal(\boldsymbol{\sigma} - \boldsymbol{y}), \tag{22}$$

где $\sigma_i = \sigma(\boldsymbol{x}_i^\intercal \boldsymbol{w})$, $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_m)^\intercal$.

Для подбора параметров $\boldsymbol{w}$ модели воспользуемся методом Ньютона-Рафсона, который на каждой итерации вычисляет квадратичную аппроксимацию функции, используя её градиент и гессиан. Формула обновления весов

$$\boldsymbol{w}^{\text{new}} = \boldsymbol{w}^{\text{old}} - H^{-1}(\boldsymbol{w}^{\text{old}})\nabla E(\boldsymbol{w}^{\text{old}}). \tag{23}$$

Гессиан функции штрафа

$$H(\boldsymbol{w}) = \\ = \nabla\nabla^\intercal E(\boldsymbol{w}) = \sum_{i=1}^{m} \sigma_i(1-\sigma_i)\boldsymbol{x}_i\boldsymbol{x}_i^\intercal = X^T \Sigma X, \tag{24}$$

где введено обозначение Σ — диагональная матрица, $\Sigma_{ii} = \sigma_i(1-\sigma_i)$. Используя (17) заметим, что $\Sigma_{ii} = \mathsf{D}y_i$, а так как компоненты вектора $\boldsymbol{y}$ по предположению независимы, то Σ является корреляционной матрицей.

Из свойств сигмоидной функции $\Sigma_{ii} > 0$, ковариационная матрица положительно определена, а значит и гессиан положительно определён (он является матрицей Грама в пространстве весов) из чего следует, что функция $E(\boldsymbol{w})$ выпукла и имеет единственный минимум.

Формула Ньютона–Рафсона для обновления весов для модели логистической регрессии

$$\boldsymbol{w}^{\text{new}} = \boldsymbol{w}^{\text{old}} - (X^T\Sigma X)^{-1}X^\intercal(\boldsymbol{\sigma} - \boldsymbol{y}) = \\ = (X^T\Sigma X)^{-1}X^T\Sigma(X\boldsymbol{w}^{\text{old}} - \Sigma^{-1}(\boldsymbol{\sigma} - \boldsymbol{y})) = \\ = (X^T\Sigma X)^{-1}X^T\Sigma\boldsymbol{z}; \tag{25}$$

$$\boldsymbol{z} = X\boldsymbol{w}^{\text{old}} - \Sigma^{-1}(\boldsymbol{\sigma} - \boldsymbol{y}). \tag{26}$$

Процедура выбора модели

Для подбора многоуровневых моделей при решении задачи классификации для объекта нужно выбрать соответствующую ему модель. Это можно сделать на основе правдоподобия этой модели. Вероятность того, что объект $(\boldsymbol{x}_i, y_i)$ был порожден моделью f_k

$$p(f_k \mid \boldsymbol{x}_i, y_i) = \frac{p(f_k, \boldsymbol{x}_i, y_i)}{p(\boldsymbol{x}_i, y_i)} = \frac{p(y_i \mid f_k, \boldsymbol{x}_i)p(f_k, \boldsymbol{x}_i)}{p(\boldsymbol{x}_i, y_i)}. \tag{27}$$

Априорная вероятность объекта $p(\boldsymbol{x}_i, y_i)$ одинакова для всех моделей. Величина $p(f_k, \boldsymbol{x}_i)$ называется априорной вероятностью модели. Предположим, что заранее нет никаких предпочтений в выборе моделей и априорные вероятности их равны. Если мы предполагаем, что объект относится к наиболее вероятной модели, то принцип максимума правдоподобия модели можно представить в виде задачи оптимизации

$$k^* = \arg\max_{k \in \{1..l\}} p(y_i \mid f_k, \boldsymbol{x}_i). \tag{28}$$

Класс объекта неизвестен, в этом случае будем рассматривать наихудший вариант: объект имеет класс, доставляющий минимум $p(y_i \mid f_k, \boldsymbol{x}_i)$:

$$k^* = \arg\max_k \min_{y_i} p(y_i \mid f_k, \boldsymbol{x}_i). \tag{29}$$

Вероятности принадлежности объектов к классам выражаются через логистическую функцию (14), перепишем решающее правило для выбора модели

$$k^* = \arg\max_k \min\{\sigma(\boldsymbol{x}_i^\intercal \boldsymbol{w}_k), \sigma(-\boldsymbol{x}_i^\intercal \boldsymbol{w}_k)\}. \tag{30}$$

Преобразуем выражение

$$k^* = \arg\max_k \sigma(-|\boldsymbol{x}_i^\intercal \boldsymbol{w}_k|) = \\ = \arg\min_k \sigma(|\boldsymbol{x}_i^\intercal \boldsymbol{w}_k|) = \arg\min_k |\boldsymbol{x}_i^\intercal \boldsymbol{w}_k|. \tag{31}$$

Заметим, что $\frac{|\boldsymbol{x}_i^\intercal \boldsymbol{w}_k|}{|\boldsymbol{w}_k|}$ есть расстояние от $\boldsymbol{x}_i$ до гиперплоскости с нормальным вектором $\boldsymbol{w}_k$. Объекты относятся к той модели, расстояние до разделяющей гиперплоскости которой минимально с точностью до модуля нормального вектора. Ввиду того, что минимизация $|\boldsymbol{x}_i^\intercal \boldsymbol{w}_k|$ эквивалентна максимизации $\sigma(\boldsymbol{x}_i^\intercal \boldsymbol{w}_k)\big(1 - \sigma(\boldsymbol{x}_i^\intercal \boldsymbol{w}_k)\big) = \mathsf{D}(y_i \mid w_k)$, решающее правило (31) можно интерпретировать так: объекты относятся к модели, дисперсия относительно которой максимальна.

Перейдем к способу построения многоуровневых моделей.

Алгоритм выбора модели

Для построения моделей используется EM-алгоритм со следующими шагами:

M-step. На M-шаге настраиваются параметры моделей с помощью логистической регрессии и метода Ньютона–Рафсона (IRLS).

Алгоритм 1. EM-алгоритм для l моделей.

Вход: $X = \{\boldsymbol{x}_i^\intercal\}_{i=1}^m$ — матрица плана;
$\boldsymbol{y} = \{y_i\}_{i=1}^m$ — метки классов;
l — число моделей;
Выход: Набор моделей $(\text{model}_k)_{k=1}^l$;

```
EmIrls(X, Y, l):
Инициализировать модели объектов случайно;
повторять
    M-step:
    для k = 1..l
        Оценить параметры k-ой модели;
        X^k — объекты, отнесенные к k-ой модели;
        y^k — классы объектов k-й модели;
        w_k = IRLS(X^k, y^k);
    E-step:
    для всех i = 1..m
        model(x_i) = arg min_k |x_i^T w_k|;
пока модели не стабилизируются.
```

E-step. На E-шаге происходит отнесение объекта к моделям на основании их правдоподобия. Решающее правило имеет вид

$$k^* = \arg\min_k |\boldsymbol{x}_i^\intercal \boldsymbol{w}_k|. \tag{32}$$

Численный эксперимент

Алгоритм тестировался на модельных и реальных данных. Модельные данные представляли собой два кластера в каждом из которых объекты разных классов распределены нормально и линейно разделимы, однако сама выборка линейно разделимой не является. Алгоритм выявил наличие двух моделей и безошибочно классифицировал объекты гиперплоскостями, рис. 1.

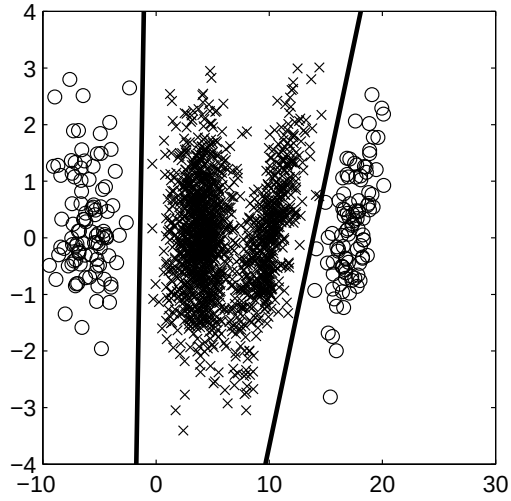


Рис. 1. Классификация модельной выборки.

Реальные данные представляли собой кредитные истории займа наличными. Выборка содержала данные о шести тысячах клиентах, каждый из которых описывался пятидесятью признаками. В качестве функции качества использовалась площадь под ROC-кривой. Метод сравнивался с логистической регрессией с градиентным методом настройки весов и итеративным перевзвешивающим методом наименьших квадратов, рис. 2.

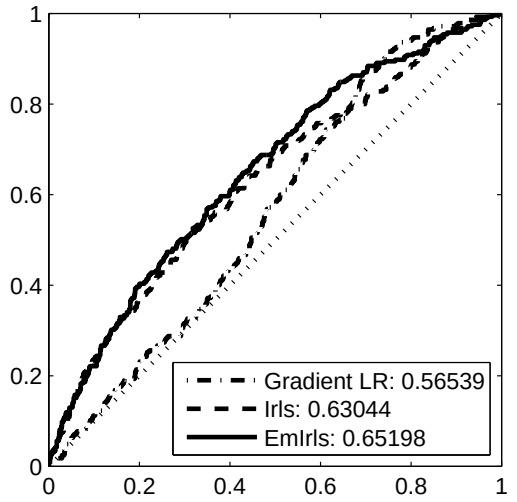


Рис. 2. ROC кривые и значения площади под кривой для различных моделей.

Заключение

В работе был предложен алгоритм выбора многоуровневых моделей; его работа проиллюстрирована на реальных и синтетических данных. Так же в работе показано преимущество использования многоуровневых моделей по сравнению с использованием одной модели на примере классификации линейно неразделимой выборки.

Литература

[1] *Ивченко Г., Медведев Ю.* Введение в математическую статистику. — Издательство ЛКИ, 2010. — P. 600.

[2] *Bishop C. M.* Pattern Recognition and Machine Learning. — Springer, Series: Information Science and Statistics, 2006. — 740 pp.

[3] *Lee Y., Nelder J. A., Pawitan Y.* Generalized Linear Models with Random Effects. — Taylor and Francis Group, LLC, 2006. — P. 396.

[4] *Nelder J., Wedderburn R.* Generalized linear models // Journal of the Royal Statistical Society. — 1972. — Pp. 370–384. — Series A (General) (Blackwell Publishing).

[5] *Strijov V., Weber G. W.* Nonlinear regression model generation using hyperparameter optimization // Computers and Mathematics with Applications. — 2010. — Vol. 60, no. 4. — Pp. 981–988.

[6] *Tibshirani R. J.* Regression shrinkage and selection via the lasso // Journal of the Royal Statistical Society. Series B (Methodological). — 1996. — Vol. 58, no. 1. — Pp. 267–288.

Уточнение ранговых экспертных оценок с использованием монотонной интерполяции*

Кузнецов М. П., Стрижов В. В.
mikhail.kuznecov@phystech.edu
Москва, ВЦ РАН

Описан способ построения интегральных индикаторов качества объектов с использованием экспертных оценок и измеряемых данных. Каждый объект описан набором признаков в линейных шкалах. Используются экспертные оценки качества объектов и важности признаков, которые корректируются в процессе вычисления. Предполагается, что оценки выставлены в ранговых шкалах. Рассматривается задача получения таких интегральных индикаторов, которые не противоречили бы экспертным оценкам. Предложено два подхода к уточнению экспертных оценок. При первом подходе вектор экспертных оценок рассматривается как выпуклый многогранный конус. Для уточнения экспертных оценок минимизируется расстояние между векторами в конусах. При втором подходе используется задача монотонной интерполяции с гиперпараметром. Проведен вычислительный эксперимент на следующих данных: экспертами оценивался фактор экологического воздействия на окружающую среду хорватских электростанций. Проведена процедура уточнения экспертных оценок.

При решении задач управления возникает необходимость дать каждому объекту оценку его качества. Интегральный индикатор — это число, поставленное в соответствие объекту, и рассматриваемое как оценка его качества. Интегральными индикаторами называется вектор оценок, поставленный в соответствие набору объектов.

При построении интегральных индикаторов выбирается критерий качества объектов. Формируется набор объектов, сравнимых в контексте выбранного критерия. Формируется набор показателей, которые эксперты считают необходимыми для описания этого критерия. Составляется матрица «объекты-признаки». Значения показателей приводятся к единой шкале и соответствуют принципу «чем больше, тем лучше»: большему значению показателя (при прочих равных) соответствует большее значение индикатора.

Данная работа посвящена уточнению экспертных оценок, выставленных в ранговых шкалах. Для построения интегральных индикаторов принимается линейная модель: строится линейная комбинация признаков с их весами. Вектор весов признаков и начальный интегральный индикатор выставляются экспертами в ранговой шкале. В общем случае, построенный по вектору весов интегральный индикатор не совпадает с индикатором, заданным экспертами, то есть экспертные данные противоречат друг другу. Данная работа посвящена устранению разногласия в оценках экспертов.

В работе будут рассмотрены два метода. Первый метод развивает идеи, описанные в [1]. Метод заключается в следующем: ранговые экспертные оценки весов показателей задают выпуклый многогранный конус. Матрица «объекты-признаки» задает линейное отображение этого конуса из пространства показателей в пространство интегральных индикаторов. Полученный в результате отображения конус может пересекаться с конусом, заданным ранговыми экспертными оценками интегрального индикатора. В этом случае, экспертные оценки показателей и объектов считаются непротиворечивыми, и отыскивается наиболее устойчивый интегральный индикатор. В противном случае, выполняется процедура рангового уточнения оценок.

Второй метод состоит в решении задачи монотонной интерполяции [2, 3, 4]. Метод заключается в том, что отыскивается вектор с монотонной последовательностью координат, наиболее близкий к заданному экспертами. Введенный в модель гиперпараметр отдает предпочтение экспертным оценкам индикаторов или оценкам весов признаков.

Предложенные алгоритмы используются для оценивания хорватских электростанций [5]. Данные являются матрицей «объекты-признаки» и заданными экспертами векторами оценок интегрального индикатора и весов признаков. Оценивается производительность электростанций.

Экспертные оценки, заданные в ранговых шкалах

Задана матрица описаний объектов $X = \{x_{ij}\}_{i=1\,j=1}^{m,n}$. Вектор $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ — описание $i-$го объекта. *Интегральный индикатор* — линейная комбинация вида

$$y_i = \sum_{j=1}^{n} w_j g_j(x_{ij}),$$

где g_j — функция приведения показателей в единую шкалу, например:

$$g_j : x_{ij} \mapsto (-1)^{\zeta_j} \frac{x_{ij} - \min_i x_{ij}}{\max\limits_i x_{ij} - \min\limits_i x_{ij}} + \zeta_j. \qquad (1)$$

Работа выполнена при финансовой поддержке РФФИ, проект № 10-07-00422.

Параметр ζ_j назначается равным 1, если оптимальное значение показателя минимально, и 0 иначе. Если знаменатель дроби 1 равен нулю для некоторых значений индекса j , то соответствующий признак исключается из дальнейшего рассмотрения. Будем обозначать теперь за X приведенную таким способом матрицу «объекты-признаки». Таким образом,

$$\mathbf{y} = X\mathbf{w}.$$

Заданы в ранговых шкалах экспертные оценки: $\mathbf{y}_0$, $\mathbf{w}_0$, допускающие произвольные монотонные преобразования. Пусть на наборах экспертных оценок введено отношение порядка такое, что

$$y_1 \geqslant y_2 \geqslant \cdots \geqslant y_m \geqslant 0;\ w_1 \geqslant w_2 \geqslant \cdots \geqslant w_n \geqslant 0.$$

Множество всех таких векторов задается системой линейных неравенств

$$J\mathbf{y} \geqslant 0,$$

где

$$\underset{m\times m}{J} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Если же порядок $y_{i_1} \geqslant y_{i_2} \geqslant y_{i_m} \geqslant 0$ произвольный, то матрица системы будет получается из J перестановкой соответствующих столбцов.

Таким образом, заданным $\mathbf{y}_0$ и $\mathbf{w}_0$ можно поставить в соответствие матрицы J_m и J_n размеров соответственно $m \times m$ и $n \times n$.

Решение задачи согласования экспертных оценок с использованием конусов

В этом параграфе опишем метод согласования экспертных оценок, предложенный в [1]. Дадим некоторые определения.

Определение 1. *Множество точек $\mathcal{Y}$ в $\mathbb{R}^m$ называется конусом, если для любой точки $y \in \mathcal{Y}$ точка λy также принадлежит $\mathcal{Y}$.*

Определение 2. *Выпуклый многогранный конус с вершиной в начале координат — это область решений системы однородных неравенств:*

$$\begin{cases} a_{11}w_1 + a_{12}w_2 + \ldots + a_{1n}w_n \geqslant 0; \\ a_{21}w_1 + a_{22}w_2 + \ldots + a_{2n}w_n \geqslant 0; \\ \ldots \quad\quad \ldots \quad\quad \ldots \quad\quad \ldots \\ a_{m1}w_1 + a_{m2}w_2 + \ldots + a_{mn}w_n \geqslant 0. \end{cases}$$

Эта система линейных неравенств задает в соответствующем пространстве выпуклый многогранный конус. Соответствуя данному определению, определим $\mathcal{Y}$ — конус, задаваемый матрицей J_m в пространстве интегральных индикаторов; $\mathcal{W}$ — конус, задаваемый матрицей J_n в пространстве весов признаков. Эти конусы характеризуются тем, что векторы внутри каждого из них имеют одинаковый ранговый порядок.

Поскольку A — линейное преобразование, оно переводит конус $\mathcal{W}$ в конус $A\mathcal{W}$, который лежит в пространстве интегральных индикаторов.

Задача 1. Требуется найти в конусах $\mathcal{W}$ и $\mathcal{Y}$ векторы $\mathbf{w}$ и $\mathbf{y}$, такие, что:

$$(\mathbf{y}_1, \mathbf{w}_1) = \min_{\mathbf{y}\in\mathcal{Y},\ \mathbf{w}\in\mathcal{W}} \|\mathbf{y} - A\mathbf{w}\|;$$
$$\text{при } \|A\mathbf{w}\| = 1, \|\mathbf{y}\| = 1,$$

где $\|.\|$ — евклидова метрика в пространстве $\mathbb{R}^m$.

Таким образом, отыскивается вектор весов $\mathbf{w}_1$, элементы которого имеют такой же ранговый порядок, что и $\mathbf{w}_0$. При этом приведенный в ранговую шкалу индикатор $A\mathbf{w}_1$ является ближайшим к $\mathbf{y}_0$.

В случае непустого пересечения конусов $\mathcal{Y}$ и $A\mathcal{W}$ решение задачи (1) дает вектор $\mathbf{y}$, который лежит в пересечении этих конусов. Если пересечение — пустое, предлагается найти ближайшие друг к другу лучи на ребрах или гранях конусов.

Отыскиваемая пара $(\mathbf{y}_1, \mathbf{w}_1)$ должна выполнять следующие условия:

$$\begin{array}{lll} \text{Минимизировать} & \|\mathbf{y} - A\mathbf{w}\| & \\ \text{при условиях} & \mathbf{y}^T\mathbf{y} = 1, & (A\mathbf{w})^T A\mathbf{w} = 1; \\ & J_n\mathbf{w} \geqslant \mathbf{0}, & J_m\mathbf{y} \geqslant \mathbf{0}. \end{array}$$

Постановка задачи согласования экспертных оценок с использованием монотонной интерполяции

В данном параграфе рассмотрим новый метод согласования экспертных оценок. Пусть $\mathbf{y}_0$ — заданное экспертами начальное приближение вектора $\mathbf{y}$. Вектор, наиболее близкий в пространстве весов признаков к $\mathbf{y}_0$, в смысле наименьших квадратов:

$$\widetilde{\mathbf{w}} = X^+\mathbf{y}_0, \text{ где}$$
$$X^+ = (X^TX)^{-1}X^T.$$

Задача 2. Требуется найти такую монотонную последовательность $w_1 \leqslant \ldots \leqslant w_n$, что она лучше всего приближает вектор $\widetilde{\mathbf{w}}$ в смысле среднего квадрата ошибки:

$$\begin{cases} \widehat{\mathbf{w}} = \arg\min\limits_{w\in\mathbb{R}^n} \sum\limits_{j=1}^{n} (\widetilde{w}_j - w_j)^2; \\ w_1 \leqslant \ldots \leqslant w_n. \end{cases}$$

Такую задачу можно решить, например, методом, описанным в [2]. Однако, чтобы получить согласованные экспертные оценки, введем в модель гиперпараметр. С его помощью мы сможем варьировать нашу «степень доверия» от экспертных оценок весов признаков (то есть, монотонной последовательности $w_1 \leqslant \cdots \leqslant w_n$) к экспертным оценкам интегральных индикаторов (вектору $\widehat{\mathbf{w}}$).

Задача 3. Требуется найти такой вектор $\widehat{\mathbf{w}}$, что:

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^n} \left(\frac{1}{2}\sum_{j=1}^{n}(\widetilde{w}_j - w_j)^2 + \right.$$
$$\left. + \lambda \sum_{j=1}^{n-1}(w_j - w_{j+1})_+ \right). \quad (2)$$

Решение задачи монотонной интерполяции с гиперпараметром

Для решения этой задачи воспользуемся идеями, описанными в [4].

Утверждение 1. *Пусть, для некоторого λ_0, совпадают две соседние координаты оценки: $\widehat{w}_j(\lambda_0) = \widehat{w}_{j+1}(\lambda_0)$. Тогда $\widehat{w}_j(\lambda) = \widehat{w}_{j+1}(\lambda)$ для всех $\lambda > \lambda_0$.*

Пусть при некотором λ совпадают некоторые соседние координаты вектора $\mathbf{w}$, и всего таких множеств совпадающих координат — K_λ. Обозначим за $A_1, \ldots, A_{K_\lambda}$ сами эти множества. Заметим, что $A_1 \cup \cdots \cup A_{K_\lambda} = \{1, \ldots, n\}$. Тогда функция потерь для задачи (2) перепишется в виде

$$\frac{1}{2}\sum_{k=1}^{K_\lambda}\sum_{l\in A_k}(\widetilde{w}_l - w_{A_k})^2 + \lambda\sum_{k=1}^{K_\lambda}(w_{A_k} - w_{A_{k+1}})_+.$$

Продифференцируем ее по всем w_{A_k}:

$$-\sum_{l\in A_k}\widetilde{w}_l + |A_k|\widehat{w}_{A_k}(\lambda) + \lambda(s_k - s_{k-1}) = 0$$
$$\text{для } k = 1, \ldots, K_\lambda,$$

где $s_k = 1$ при $\widehat{w}_{A_k}(\lambda) - \widehat{w}_{A_{k+1}}(\lambda) > 0$, и $s_k = 0$ иначе.

Пусть все $A_1, \ldots, A_{K_\lambda}$ не изменяются с увеличением λ. Тогда:

$$\frac{d\widehat{w}_{A_k}(\lambda)}{d\lambda} = \frac{s_{k-1} - s_k}{|A_k|}.$$

Когда λ увеличивается, множества A_k меняются. Однако, согласно утв. 1, они могут только объединяться, то есть, величины компонент $\widehat{w}_{A_k}(\lambda)$ внутри каждого множества A_k остаются равными. Можно посчитать величину следующего λ, при котором будут объединяться множества A_k, A_{k+1}. Обозначим это λ как $t_{k,k+1}$.

Алгоритм 1. Алгоритм решения задачи монотонной интерполяции.

Вход: $\lambda = 0,\ K_\lambda = n,\ A_k = \{k\},\ \widehat{w}_{A_k}(\lambda) = \widetilde{w}_k.$

1: **повторять**
2: $D_k := \frac{s_{k-1} - s_k}{|A_k|}$;
3: $t_{k,k+1} := \frac{\widehat{w}_{A_{k+1}}(\lambda) - \widehat{w}_{A_k}(\lambda)}{D_k - D_{k+1}} + \lambda$;
4: $\widehat{\lambda} := \min\limits_{k: t_{k,k+1} > \lambda} t_{k,k+1}$;
5: $\widehat{w}_{A_k}(\lambda) := \widehat{w}_{A_k}(\lambda) + D_k(\widehat{\lambda} - \lambda)$;
6: объединить $A_{k'}$ и $A_{k'+1}$, см. (3);
7: $\lambda := \widehat{\lambda}$;
8: **пока** существует k: $t_{k,k+1} \geqslant \lambda$.

Утверждение 2. *Множества A_k и A_{k+1} будут объединяться при*

$$t_{k,k+1} = \frac{\widehat{w}_{A_{k+1}}(\lambda) - \widehat{w}_{A_k}(\lambda)}{D_k - D_{k+1}} + \lambda,$$

для всех $k = 1, .., K_\lambda - 1$, где

$$D_k = \frac{d\widehat{w}_{A_k}(\lambda)}{d\lambda}.$$

Доказательство. Поскольку производные

$$\frac{d\widehat{w}_{A_k}(\lambda)}{d\lambda}$$

не являются функциями λ, можно записать следующую систему уравнений:

$$\begin{cases} \widehat{w}_{A_k}(\lambda) = \lambda D_k + C_k; \\ \widehat{w}_{A_{k+1}}(\lambda,) = \lambda D_{k+1} + C_{k+1}. \end{cases}$$

В точке $t_{k,k+1}$ происходит объединение множеств A_k и A_{k+1}, то есть:

$$\widehat{w}_{A_k}(t_{k,k+1}) =$$
$$= \widehat{w}_{A_{k+1}}(t_{k,k+1}) \ \Rightarrow\ t_{k,k+1} = \frac{C_{k+1} - C_k}{D_k - D_{k+1}} =$$
$$= \frac{(\widehat{w}_{A_{k+1}}(\lambda) - \lambda D_{k+1}) - (\widehat{w}_{A_k}(\lambda) - \lambda D_k)}{D_k - D_{k+1}} =$$
$$= \frac{\widehat{w}_{A_{k+1}}(\lambda) - \widehat{w}_{A_k}(\lambda)}{D_k - D_{k+1}} + \lambda.$$

Что и требовалось доказать.

Таким образом, на каждой итерации нужно вычислять величину

$$\widehat{\lambda} = \min_{k: t_{k,k+1} > \lambda} t_{k,k+1}$$

и объединять множества $A_{k'}$ и $A_{k'+1}$, где

$$k' = \arg\min_{k: t_{k,k+1} > \lambda} t_{k,k+1}. \quad (3)$$

Результат работы алгоритма монотонной интерполяции

Проиллюстрируем работу алгоритма решения задачи монотонной интерполяции на модельной выборке, порожденной с помощью функции $y_i = x_i + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, 20)$. Ломаная линия на рис. 1 — восстановленная зависимость, для различных значений регуляризатора λ.

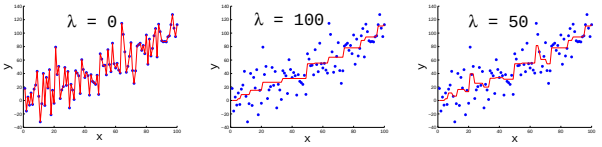


Рис. 1. Монотонная интерполяция.

Видно, что при $\lambda = 100$ и более функция, восстанавливающая зависимость, монотонная.

При $\lambda = 0$, наоборот, никакой монотонной коррекции нет.

Вычислительный эксперимент

Был проведен вычислительный эксперимент уточнения экспертных оценок экологического воздействия на окружающую среду хорватских электростанций. Для этого были собраны следующие данные: матрица «объекты-признаки», где объекты — это 7 электростанций, описываемых 11 признаками, экспертные оценки весов показателей и интегральных индикаторов электростанций. Приведем таблицу с данными (в ней показано только 6 первых признаков):

N	Power Plant	Available net capacity (MW)	Electricity (GWh)	Heat (TJ)	SO_2 (t)	NO_x (t)	Particles (t)
1	Plomin 1 TPP	98	452	0	1950	1378	140
2	Plomin 2 TPP	192	1576	0	581	1434	60
3	Rijeka TPP	303	825	0	6392	1240	171
4	Sisak TPP	396	741	0	3592	1049	255
5	TE-TO Zagreb CHP	337	1374	481	2829	705	25
6	EL-TO Zagreb CHP	90	333	332	1259	900	19
7	TE-TO Osijek CHP	42	114	115	1062	320	35
	Optimal value	max	max	max	min	min	min

Таблица 1. Электростанции

На рис. 2 показаны графики интегральных индикаторов, вычисленных различными алгоритмами:

а) начальный интегральный индикатор $\mathbf{q}_0$,

б) интегральный индикатор, построенный по $\mathbf{w}_0$,

в) алгоритм минимизации расстояния между векторами в конусах,

г) алгоритм максимизации корреляции между векторами в конусах,

д) алгоритм монотонной интерполяции со значением гиперпараметра $\lambda = 1$,

е) алгоритм монотонной интерполяции со значением гиперпараметра $\lambda = 0.5$.

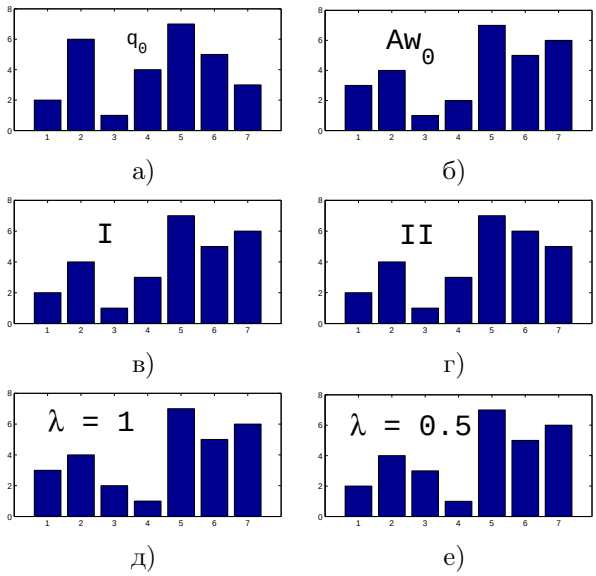


Рис. 2. Интегральные индикаторы электростанций.

Заключение

В работе рассматривалась задача получения согласованных оценок качества объектов и важности показателей. В результате выполнения работы обобщены ранее полученные результаты по согласованию экспертных оценок с использованием конусов. Предложено использовать алгоритм монотонной интерполяции для уточнения экспертных оценок. Исследованы свойства этого алгоритма при различном значении гиперпараметра, введенного в модель. Проведен вычислительный эксперимент уточнения экспертных оценок качества хорватских электростанций, составлен рейтинг электростанций, основанный на оценках экспертов и измеряемых данных.

Литература

[1] *V. Strijov, G. Granić et al.* Integral indicator of ecological impact of the Croatian thermal power plants // Energy. — 2011. — V. 4, N. 30.

[2] *J. de Leeuw, K. Hornik, P. Mair* Isotone Optimization in R: Pool-Adjacent-Violators Algorithm (PAVA) and Active Set Methods // Jounal of Statistic Software. — 2009. — V. 29.

[3] *R. E. Barlow, H. D. Brunk* The Isotonic Regression Problem and Its Dual // Journal of American Statistical Association. — 1972. — V. 67. — Pp. 140–147.

[4] *R. J. Tibshirani, H. Hoefling, R. Tibshirani* Nearly-isotonic Regression // Technometrics. — 2011. — V. 53.

[5] *R. Kos, Z. Krisic, T. Tarnik* Hrvatska elektoprivreda and the environment 2005–2006 // Zagreb, Hrvatska Elektroprivreda, 2008.

Метод многомерной регрессии, основанный на нерасширяемых и несократимых выпуклых комбинациях предикторов*

Сенько О. В., Докукин А. А.
senkoov@mail.ru
Москва, ВЦ РАН

Описывается новый метод многомерной регрессии, использующий выпуклые комбинации предикторов. Предполагается, что предикторы представляют собой достаточно простые регрессии малой размерности, построенные с использованием исходного набора независимых переменных. В основе разработанного метода лежит понятие нерасширяемого несократимого набора (ННН). Были найдены необходимые условия, позволяющие эффективно проверять несократимость набора предикторов, а также строить их оптимальные выпуклые комбинации. В статье приводятся результаты экспериментов, демонстрирующие высокую прогностическую способность метода а также высокую чувствительность отбора переменных.

В последние годы были разработаны статистические методы, позволяющие значительно улучшить прогностическую способность регрессионных моделей в задачах большой размерности, например, метод наименьших углов или LARS [1, 2]. Успешность этих методов связана с эффективным отбором прогностических переменных. Тем не менее, проблема низкой обобщающей способности эмпирических моделей в задачах высокой размерности не может считаться полностью решенной. Разработка новых альтернативных подходов может быть полезна для получения верхних оценок прогностической способности или оценки оптимального числа отбираемых переменных. В настоящей работе обсуждается подход, при котором оптимальные модели строятся на основе наборов обученных простых предикторов, например, одномерных или двумерных регрессий. Предположим, что имеется набор из L предикторов $z_1, \ldots, z_L$, прогнозирующих некоторую переменную Y. Пусть $\mathbf{c} = (c_1, \ldots, c_L)$ — вектор неотрицательных коэффициентов, удовлетворяющих условию $\sum_{i=1}^{l} c_i = 1$. Выпуклая корректирующая процедура (ВКП) вычисляет прогнозируемое значение как взвешенную сумму прогнозов отдельных предикторов $Z_{ccp}(\mathbf{c}) = \sum_{i=1}^{L} c_i z_i$.

Выпуклые комбинации широко используются в распознавании. В качестве примеров можно привести процедуры бэггинга и бустинга [3, 4], а также коллективные решения по наборам логических закономерностей [5, 6, 7]. Кроме того, ранее было показано, что ошибка выпуклой комбинации предикторов не может превышать ту же выпуклую комбинацию обобщенных ошибок отдельных предикторов [8]. Выпуклая коррекция используется и в задачах регрессии — в [9] была продемонстрирована эффективность выпуклых комбинаций пар регрессий, в [10, 11] был исследован метод оптимизации ВКП, основанный на минимизации оценок обобщенной ошибки. Эксперименты с искусственными данными показали, что оптимизации ошибки в ВКП приводила к эффективному отбору информативных прогностических переменных. Легко показать, что уменьшение дисперсии ВКП по сравнению с той же комбинацией отдельных дисперсий также является свойством выпуклых комбинаций. Такое уменьшение ухудшает прогностическую способность ВКП. Поэтому прогноз ВКП должен быть дополнительно скорректирован с помощью одномерной регрессии. Но прогностическая способность линейной регрессии монотонно зависит от коэффициента корреляции между Z_{cpp} и Y. В данной работе будет описан новый метод построения ВКП максимально коррелирующей с Y. Этот метод основан на той же концепции поиска несократимых наборов предикторов, которая была использована в [11].

Определение 1. *Предиктор z будет называться приведенным, если для любых значений α, β выполняется неравенство*

$$E_\Omega(Y - \alpha z - \beta)^2 \leqslant E_\Omega(Y - z)^2 \ .$$

Здесь $E_\Omega(X)$ — математическое ожидание X по пространству допустимых объектов с определенной σ-алгеброй и вероятностной мерой (для краткости обозначим $\hat{X}$). Далее предполагается, что все исходные предикторы — приведенные, в противном случае они приводятся с помощью одномерной регрессии. Известно, что для приведенных предикторов справедливы равенства: $\hat{z} = \hat{X}$, $cov(Y, Z) = E_\Omega\,(z - \hat{z})^2$.

Свойства выпуклых комбинаций

Известно, что квадратичная ошибка выпуклой комбинации предикторов всегда меньше той же комбинации ошибок отдельных предикторов [8]. Пусть $\delta_i = E_\Omega(Y - z_i)^2$ — квадратичная ошибка предиктора z_i, $\rho_{ij} = E_\Omega(z_i - z_j)^2$ — величина, характеризующая расхождение i-го и j-го предикторов. Ошибку выпуклой комбинации можно пред-

Работа выполнена при финансовой поддержке РФФИ, проекты № 10-01-90015-а, 10-01-90419.

ставить [11] в виде квадратичного функционала:

$$\delta(Z_{ccp}) = \sum_{i=1}^{L} c_i\delta_i - \frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{L}\rho_{ij} \ . \qquad (1)$$

Можно показать также, что представление, подобное (1), существует [8] для компонент обобщенной ошибки $\Delta(Z_{ccp}) = E_{\Omega_{tr}}\delta(Z_{ccp})$, где Ω_{tr} — пространство всех возможных обучающих информаций. При этом базовая и вариационная компоненты ошибки корректора также уменьшаются по сравнению с отдельными предикторами.

Структура дисперсии

По определению

$$V(Z_{ccp}) = \sum_{i=1}^{L}\sum_{j=1}^{L} c_i c_j E_\Omega\left[(z_i - \hat{z}_i)(z_j - \hat{z}_j)\right] \ ,$$

где $V(X) = (X - \hat{X})^2$. После преобразований получаем

$$V(Z_{ccp}) = \sum_{i=1}^{L} c_i V(z_i) - \frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{L} c_i c_j \rho_{ij}^v \ , \qquad (2)$$

где $\rho_{ij}^v = E_\Omega(z_i - \hat{z}_i - z_j + \hat{z}_j)^2$.

Таким образом, структура квадрата дисперсии ВКП также совпадает со структурой её ошибки. Отметим, что для пар приведенных предикторов $\rho_{ij} = \rho_{ij}^v$. Из (2) следует, что квадрат дисперсии ВКП всегда меньше, чем соответствующая комбинация квадратов дисперсий отдельных предикторов. Такое уменьшение часто приводит к ухудшению прогностической способности. Тем не менее, ухудшение может быть скомпенсировано с помощью дополнительного линейного преобразования, увеличивающего дисперсию. Очевидно, что наилучшая прогностическая способность после линейного преобразования будет достигнута на переменной, максимально коррелирующей с Y. Далее обсуждается подход, основанный на максимизации коэффициента корреляции Пирсона между Z_{ccp} и Y.

Наборы, несократимые относительно коэффициента корреляции

Для приведенных предикторов коэффициент корреляции Пирсона может быть записан в виде

$$K\left[Y, Z_{ccp}(\mathbf{c})\right] = \frac{\sum\limits_{i=1}^{L} c_i V(z_i)}{\sqrt{V(Y)}\sqrt{\sum\limits_{i=1}^{L} c_i V(z_i) - \frac{1}{2}\sum\limits_{i=1}^{L}\sum\limits_{j=1}^{L} c_i c_j \rho_{ij}^v}} \ .$$

Дальнейшие рассуждения основываются на понятии несократимости. Набор предикторов $\tilde{z}$ называется несократимым, если удаление из него любого элемента не позволяет построить ВКП с то же прогностической способностью, что и исходный набор. Запишем строгое определение.

Определение 2. *Множества $\bar{D}_{\mathrm{L}}$, D_{L} из $\mathbb{R}^L$ определяются как*

$$\bar{D}_{\mathrm{L}} = \Big\{\mathbf{c}\Big|\sum_{i=1}^{L} c_i = 1,\ c_i \geqslant 0,\ i = 1,\ldots,L\Big\};$$

$$D_{\mathrm{L}} = \Big\{\boldsymbol{c}\Big|\sum_{i=1}^{L} c_i = 1,\ c_i > 0,\ i = 1,\ldots,L\Big\}.$$

Определение 3. *Набор предикторов $z_1,\ldots,z_l$ называется несократимым относительно некоего функционала $F(\mathbf{c})$, характеризующего прогностическую способность, если существует такой вектор $\mathbf{c}^* \in D_L$, что для всех $\mathbf{c}' \in \bar{D}_{\mathrm{L}}$, $F(\mathbf{c}^*) > F(\mathbf{c}')$.*

Множество точек $\mathbb{R}^l$, одновременно удовлетворяющее ограничениям $\sum\limits_{i=1}^{L} c_i = 1$ и $\sum\limits_{i=1}^{L} c_i V(z_i) = \Theta$, далее будет обозначаться как $\mathbf{W}(\Theta)$.

Теорема 1. *Необходимым условием несократимости набора предикторов $z_1,\ldots,z_L$ относительно $K(Y, Z_{ccp})$ является существование такого вещественного θ, что квадратичный функционал*

$$\mathbf{P}_f(\mathbf{c}) = \sum_{i=1}^{L}\sum_{j=1}^{L} c_i c_j \rho_{ij}^v$$

достигает строгого максимума на $\mathbf{W}(\Theta)$ в точке $c_1^*,\ldots,c_L^*$, удовлетворяющей условиям: $c_i^* > 0$, $i = 1,2,\ldots,L$.

Необходимым условием максимума является существование положительного $\Theta > 0$, такого, что справедлива формула

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \rho(z_i, z_j) \to \max$$

при условиях: $\sum\limits_{i=1}^{n} c_i E(z_i^2) = \Theta$, $\sum\limits_{i=1}^{n} c_i = 1$, $c_i \geqslant 0$, $i = 1,\ldots,n$.

Введем обозначения: $D = \|\rho_{ij}\|_{L\times L}$, $C = \|c_i\|_{1\times L}$, $E = \|E(z_i^2)\|_{1\times L}$, $I = \|1\|_{1\times L}$. Записав уравнение Лагранжа и проведя дополнительные преобразования, установим зависимость между C и Θ:

$$c_k + \frac{\beta - \Theta\gamma}{\alpha\gamma - \beta^2}\sum_{i=1}^{n} d_{ki}E(z_i^2) + \frac{\alpha - \Theta\beta}{\beta^2 - \alpha\gamma}\sum_{i=1}^{n} d_{ki} = 0,$$

где d_{ij} — элемент матрицы D^{-1}, $\alpha = E^{\intercal}D^{-1}E$, $\beta = I^{\intercal}D^{-1}$, $E = E^{\intercal}D^{-1}I$, $\gamma = I^{\intercal}D^{-1}I$.

Итак, получаем условия для Θ $(k = 1, \ldots, n)$:

$$\frac{\Theta\gamma - \beta}{\alpha\gamma - \beta^2} \sum_{i=1}^{n} d_{ki} E(z_i^2) + \frac{\Theta\beta - \alpha}{\beta^2 - \alpha\gamma} \sum_{i=1}^{n} d_{ki} > 0 \ . \quad (3)$$

Необходимо отметить, что точка $\mathbf{c}^*$ может являться точкой строгого максимума $\mathbf{P}_f$ только при условии

$$\sum_{i=1}^{L} \sum_{j=1}^{L} \rho_{ij} \varepsilon_i \varepsilon_j > 0 \quad (4)$$

для любых $(\varepsilon_0, \ldots, \varepsilon_l)$, таких, что $\sum\limits_{i=1}^{l} \varepsilon_i = 0$.

Пусть Θ_{min} — минимальный, а Θ_{max} — максимальный из значений, обращающих одно из неравенств (3) в равенство. Введем обозначения:

$$R_k^v = \sum_{i=1}^{L} V(z_i)\rho_{ki}, \ \ P_k = \sum_{i=1}^{L} \rho_{ki} \ ;$$

$$\Gamma_i^1 = \frac{GR_i^v + BP_k}{AG - B^2}, \ \ \Gamma_i^0 = \frac{GR_i^v + BP_k}{AG - B^2} \ ;$$

$$\beta_0 = \sum_{i=1}^{l} \sum_{j=1}^{l} \Gamma_i^0 \Gamma_j^0 \rho_{ij} \ ;$$

$$\beta_1 = \sum_{i=1}^{l} \sum_{j=1}^{l} (\Gamma_i^0 \Gamma_j^1 + \Gamma_i^1 \Gamma_j^0) \rho_{ij} \ ;$$

$$\beta_2 = \sum_{i=1}^{l} \sum_{j=1}^{l} \Gamma_i^1 \Gamma_j^1 \rho_{ij} \ .$$

Тогда $\mathbf{P}_f = \beta_0 + \beta_1\Theta + \beta_2\Theta^2$, и легко показать, что

$$K(Y, Z_{ccp}) = \varkappa(\Theta) = \frac{\Theta}{\sqrt{\beta_1\Theta - \beta_2\Theta^2 - \beta_0}} \ .$$

Теорема 2. *Одновременное выполнение неравенств* $\Theta_{min} < \frac{2\beta_0}{\beta_1} < \Theta_{max}$, $\varkappa(\frac{2\beta_0}{\beta_1}) > \varkappa(\Theta_{min})$ *и отрицательность выражения (4) являются необходимым условием несократимости набора предикторов* $(z_1, \ldots, z_L)$.

Необходимые условия позволяют эффективно оценивать несократимость наборов предикторов. Достаточно вычислить Θ_{min} и Θ_{max}, чтобы оценить условия отрицательности (4) и проверить неравенства $\Theta_{min} < \frac{2\beta_0}{\beta_1} < \Theta_{max}$. Очевидно, что если при выполнении необходимых условий значение $\varkappa\left(\frac{2\beta_0}{\beta_1}\right)$ для рассматриваемого набора $\tilde{z}$ с l^* предикторов больше, чем для любого несократимого набора с менее чем l^* предикторов, рассматриваемый набор несократим.

Регрессии, основанные на нерасширяемых несократимых наборах предикторов

На первой стадии с помощью стандартного одномерного метода наименьших квадратов формируется исходное множество приведенных предикторов $\tilde{\mathbf{Z}} = \{z_1, \ldots, z_L\}$.

Несократимый набор $\hat{\mathbf{z}}'$, состоящий из l' элементов называется нерасширяемым (ННН), если не существует несократимых наборов в $\tilde{\mathbf{Z}}$ с числом предикторов, большим, чем l', содержащих все предикторы из $\tilde{\mathbf{z}}'$.

Было рассмотрено два способа построения регрессий по множеству ННН, основанных на переборе. Первый способ предполагает выбор единственного лучшего ННН с максимальным коэффициентом корреляции оптимального Z_{ccp} и Y. Этот оптимальный Z_{ccp} (Z_{ccp}^{max}) является окончательным. При втором способе выбирается множество ННН, в которых коэффициент корреляции оптимального Z_{ccp} с Y превышает $Tr \cdot K(Y, Z_{ccp}^{max})$, $Tr \in (0, 1)$ — пороговый параметр. При этом параметры окончательной регрессии вычисляются как среднее по найденным ННН. Перебор ННН в обоих случаях основан на индуктивном увеличении набора, с сохранением условия несократимости. Проведенные эксперименты показали, что второй способ более эффективен. Метод, основанный на ННН зависит от квадратов дисперсий отдельных предикторов V и расстояний между ними ρ. Эти параметры оценивались по обучающей выборке по стандартным формулам:

$$v(z) = \frac{1}{M} \sum_{j=1}^{M} [z(j) - \hat{z}]^2,$$

$$\rho(z^1, z^2) = \frac{1}{M} \sum_{j=1}^{M} [z^1(j) - z^2(j)]^2,$$

где M — размер выборки. Испытания показали, что такой тип оценок приводит к отбору слишком большого числа переменных, и, соответственно, к уменьшению прогностической способности. Тем не менее, эффективность может быть улучшена с помощью штрафного коэффициента для ρ, равного $\frac{1}{1+5/M}$. Этот эффект требует отдельного исследования и объяснения.

Реализация метода LARS

Эффективность регрессий, основанных на ННН, была оценена с использованием большого числа искусственных задач. При этом его прогностическая эффективность сравнивалась с эффективностью известного метода многомерной регрессии — метода наименьших углов (LARS). Алгоритм LARS был запрограммирован согласно схеме, описанной в [1]. Для оценок на шаге окончательного отбора предикторов были протестированы критерии C_p и Акаике. При этом было использовано простое приближение для степеней свободы [1]: $df = k$, где k — число отобранных заранее предикторов. По результатам экспериментов C_p продемонстрировал гораздо более высокую прогностическую способность.

	CCP				LARS			
M	K	N_c	N_f	R_f	K	N_c	N_f	R_f
20	0,729	15,13	6,03	0,132	0,668	1,84	0,16	0,0239
30	0,752	16,81	5,76	0,054	0,685	1,99	0,01	0,0007
40	0,772	17,36	7,35	0,066	0,704	1,99	0,01	0,00004
50	0,776	17,21	5,69	0,0296	0,705	2,0	0	0

Таблица 1. Результаты прогнозирования.

Эксперименты

Во всех задачах прогнозируемая переменная Y и переменные X являются функциями трех латентных переменных U_1, U_2, U_3. Значения переменных U независимы и принадлежат многомерному нормальному распределению со средним равным 0 и стандартным отклонением 1. Значение зависимой переменной Y в j-м случае определяется формулой $y_j = \sum_{k=1}^{3} u_{jk} + e^j_{\text{y}}$, где u_{jk} — значение латентной переменной U_k, e^{y}_j — случайная ошибка, распределенная нормально $N(0, d_y)$. Значения релевантных переменных X_i генерировались с помощью бинарного вектора $\boldsymbol{\beta^i} = \{\beta^i_1, \beta^i_2, \beta^i_3\}$. В j-м случае $x_{ij} = \sum_{k=1}^{3} u_{jk}\beta^i_k + e^j_{\text{x}_i}$, где u_{jk} — значение латентной переменной U_k, e^j_{xr} — случайная ошибка $(N(0, d_{xi}))$. Значения нерелевантных переменных X_i распределены нормально с параметрами $(0, d_{xi})$. В каждом эксперименте с помощью генератора случайных чисел синтезировались 100 пар выборок (обучающая и контрольная). Результаты экспериментов показаны в таблице 1. Для каждой пары выборок размера M приведены следующие характеристики для метода LARS и множественной ННН с $Tr = 0{,}95$: K — корреляция между Y вычисленным прогнозом, N_c — среднее число релевантных переменных, корректно отобранных в регрессионной модели, N_f — среднее количество ошибочно отобранных нерелевантных переменных, R_f — значение $|\beta|/|\beta_{max}|$ для нерелевантных переменных. Здесь $|\beta|$ — абсолютное значение коэффициента некоторой переменной в регрессии, $|\beta_{max}|$ — максимальное абсолютное значение коэффициента среди всех переменных модели.

Из таблицы видно, что эффективность ВКП регрессий, основанных на ННН выше, чем у LARS с простой аппроксимацией df. Важно отметить, что метод множественных ННН отбирает большую часть релевантных переменных. Число отобранных нерелевантных также относительно велико, однако реальных вклад этих переменных характеризуется параметрами R_f и является существенно меньшим. Видно также, что реализованный метод LARS в большинстве задач корректно отбирал пары релевантных переменных. Количество ошибочно отобранных нерелевантных в экспериментах с $M > 20$ пренебрежимо мало.

Заключение

Предложенный в работе метод продемонстрировал достаточно высокую прогностическую способность даже по сравнению с таким эффективным конкурентом, как LARS, а также высокую чувствительности и точность отбора переменных. Основным недостатком метода является быстрый рост вычислительной сложности поиска ННН. Ускорение метода до приемлемого уровня при числе переменных 60–70 является основным направлением дальнейших исследований.

Литература

[1] *Efron B., Hastie T., Jonnstone I. and Tibshirani R..* Least Angle Regression // Annals of Statistics. — 2004. — V. 32, N. 2. — Pp. 407–499.

[2] *Tibshirani R.* Regression shrinkage and selection via the lasso // J. Roy. Stat. Soc. — 1996. — V. 58. — Pp. 267–288.

[3] *Breiman L.* Random forests — random features // Technical report 567. Statistics department. University of California, Berkley. — 1999. — `www.boosting.org`.

[4] *Kuncheva L. I.* Combining Pattern Classifiers. Methods and Algorithms. — New Jersey: Wiley Interscience, 2004.

[5] *Zhuravlev Yu. I., Kuznetsova A. V., Ryazanov V. V., Senko O. V., Botvin M. A.* The Use of Pattern Recognition Methods in Tasks of Biomedical Diagnostics and Forecasting // Pattern Recognition and Image Analysis. — 2008. — V. 18, N. 2. — Pp. 195–200.

[6] *Журавлев Ю. И., Рязанов В. В., Сенько О. В.* РАСПОЗНАВАНИЕ. Математические методы. Программная система. Приложения. — М.: Фазис, 2006.

[7] *Kuznetsov V. A., Senko O. V. et all.* Recognition of fuzzy systems by method of statistically weighed syndromes and its using for immunological and hematological norm and chronic pathology // Chemical Physics. — 1996. — V. 15, N. 1. — Pp. 81–100.

[8] *Krogh A. and Vedelsby J.* Neural network ensembles, cross validation, and active learning // NIPS, 1995. — V. 7. — Pp. 231–238.

[9] *Senko O. V.* The Use of Collective Method for Improvement of Regression Modeling Stability // InterStat. Statistics on the Internet http://statjournals.net/, 2004.

[10] *Senko O. V.* An Optimal Ensemble of Predictors in Convex Correcting Procedures // Pattern Recognition and Image Analysis. — 2009. — V. 19, N. 3. — Pp. 465–468.

[11] *Senko O., Dokukin A.* Optimal Forecasting Based on Convex Correcting Procedures // New Trends in Classification and Data Mining — ITHEA, Sofia, Bulgaria, 2010. — Pp. 62–72.

Агрегирование адаптивных алгоритмов прогнозирования*

Романенко А. А.
alexromsput@gmail.com
Московский физико-технический институт (государственный университет)

Рассматриваются методы построения композиций адаптивных алгоритмов прогнозирования с помощью агрегирующего алгоритма В. Вовка. Получена оценка качества прогнозов в худшем случае. Эксперименты на реальных данных об объёмах продаж товаров в розничной сети показывают, что агрегирующий алгоритм обеспечивает в среднем более высокое качество прогнозов по сравнению с другими видами композиций.

В данной работе рассматривается задача прогнозирования временных рядов, и, как практическое приложение, задача прогнозирования потребительского спроса. В ней временные ряды представляют собой данные об объёмах ежедневных продаж товаров в магазинах розничной сети. Задача прогнозирования потребительского спроса обладает специфическими особенностями. Во-первых, число временных рядов имеет порядок 10^7, что накладывает определённые ограничения на используемые алгоритмы. Во-вторых, выборки, по которым требуется проводить прогнозирование, могут иметь различную длину, в том числе очень короткую. В-третьих, для данных характерно высокое отношение шум/сигнал и нестационарность.

В силу особенностей задачи для её решения используются быстрые адаптивные алгоритмы прогнозирования, такие, как экспоненциальное сглаживание, алгоритм Брауна, скользящее среднее и др. [1]. Большое разнообразие возможных типов временных рядов и методов прогнозирования приводит к необходимости автоматического подбора для каждого ряда наиболее адекватного метода или композиции нескольких наиболее адекватных методов. При этом часто привлекаются вероятностные предположения о характере самих данных или ошибок прогнозирования, например гипотезы стационарности или нормальности [7]. На практике эти предположения часто не выполняются, и лучшие результаты показывают простые эвристические алгоритмы и их композиции [4].

В данной работе рассматривается агрегирующий алгоритм В. Вовка [3], свободный от вероятностных допущений, но тем не менее имеющий строгие теоретические гарантии, что качество прогнозов композиции будет близко к качеству прогнозов лучшего из базовых алгоритмов.

Предлагается несколько обобщений и модификаций агрегирующего алгоритма, необходимых для его реализации и проверки на практических задачах прогнозирования временных рядов.

Работа поддержана РФФИ (проект №11-07-00480) и программой ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

Постановка задачи

Пусть заданы множество исходов Ω, множество допустимых предсказаний Γ и функция потерь

$$\lambda\colon \Omega \times \Gamma \to \mathbb{R}^+ \cup \{\infty\}.$$

Временным рядом будем называть упорядоченную по времени последовательность элементов $X_T = x_1, \ldots, x_T \in \Omega$. Алгоритмом прогнозирования называется функция $A\colon \Omega^T \to \Gamma$, которая по конечной последовательности $X_T = x_1, \ldots, x_T$ предсказывает значение $\hat{x}_{T+1}$ следующего её элемента: $\hat{x}_{T+1} = A(X_T)$.

Определение 1. *Алгоритм A прогнозирует последовательность $x_1, \ldots, x_T$, если он предсказывает поочередно каждый из её элементов и заключается в выполнении следующих шагов:*

Для $t = 0, \ldots, T-1$

1) *предсказать значение $\hat{x}_{t+1} \in \Gamma$;*
2) *получить исход $x_{t+1} \in \Omega$;*
3) *вычислить величину потерь $\lambda(x_{t+1}, \hat{x}_{t+1})$.*

Определение 2. *Процессом потерь при прогнозировании алгоритмом A последовательности $x_1, \ldots, x_T$ называется функция*

$$\mathcal{L}_A(x_1, \ldots, x_T) = \mathcal{L}_A(T) = \sum_{t=1}^{T} \lambda(x_t, \hat{x}_t).$$

Задача прогнозирования временного ряда $Z = \langle \Omega, \Gamma, \lambda \rangle$ сводится к минимизации процесса потерь в каждый момент времени.

Определение 3. *Композицией конечного числа алгоритмов прогнозирования $A^1, \ldots, A^N$ называется функция:*

$$C\colon \Omega^T \times \Gamma^N \to \Gamma,$$

которая по временному ряду $x_1, \ldots, x_T$ и прогнозам $\hat{x}^N_{T+1}, \ldots, \hat{x}^N_{T+1}$ следующего его значения, полученным от $A^1, \ldots, A^N$ соотвественно, делает свой прогноз $\hat{x}_{T+1}$ следующего отсчета временного ряда.

Агрегирующий алгоритм

Пусть задано некоторое (необязательно конечное) множество алгоритмов прогнозирования $\mathfrak{A}$, которых будем называть *экспертами*.

Определение 4. *Агрегирующий алгоритм*

$$M\colon \Omega^T \times \Gamma^{\mathfrak{A}} \to \Gamma$$

прогнозирует последовательность $x_1, \ldots, x_T$, опираясь на предсказания экспертов из $\mathfrak{A}$, если он поочерёдно прогнозирует каждый из её элементов согласно следующему алгоритму:

Для $t = 1, \ldots, T-1$

1) *получить прогнозы экспертов $\hat{x}_{t+1}^A$, $\forall A \in \mathfrak{A}$;*
2) *сделать предсказание $\hat{x}_{t+1} \in \Gamma$;*
3) *получить исход $x_{t+1} \in \Omega$;*
4) *вычислить величину ошибки $\lambda(x_{t+1}, \hat{x}_{t+1})$.*

Агрегирующий алгоритм — это обобщение понятия композиции, так как он может агрегировать как конечное, так и бесконечное число экспертов.

Задача 1. Найти агрегирующий алгоритм M такой, что для всех $x_1, \ldots, x_T$ и всех $A \in \mathfrak{A}$

$$\mathcal{L}_M(T) \leqslant f\left(\mathcal{L}_A(T)\right),$$

где $f(x)$ — функция, мало отличающаяся от x.

Агрегирующий алгоритм Вовка [2] отличается тем, что каждому эксперту A^j назначается вес p_t^j, полученный взвешиванием в экспоненциальном пространстве ошибок алгоритма A^j, см. Алгоритм 1.

При конечном множестве экспертов $|\mathfrak{A}| = N$ процесс потерь оценивается следующим образом:

$$\mathcal{L}_M(T) = c \cdot \mathcal{L}_A(T) + a \cdot \ln(N),\ \forall A \in \mathfrak{A}, \qquad (1)$$

где c и a — некоторые положительные константы. Описание агрегирующего алгоритма также приведено для случая конечного числа экспертов.

Функция $g_t(x)$, называемая *смешиванием*, каждому возможному исходу $x \in \Omega$ ставит в соответствие некоторую обобщенную потерю экспертов.

Подстановочный функционал $S\colon \Gamma^\Omega \to \mathbb{R}$ определяет для функции $g_T\colon \Omega \to \Gamma$ такое значение прогноза $\hat{x}_{T+1} = S(g_T)$, чтобы было выполнено неравенство:

$$\forall x \in \Omega \quad \lambda(x, \hat{x}_{T+1}) \leqslant c(\beta) \cdot g_T(x). \qquad (2)$$

Примеры подстановочных функционалов:

$$S(g_T) = \arg\min_{\hat{x}\in\Gamma} \max_{x\in\Omega} \frac{\lambda(x,\hat{x})}{g_T(x)};$$
$$S(g_T) = \arg\min_{x\in\Omega} g_T(x).$$

Важным является случай, когда $c(\beta) = 1$ для некоторого $\beta \in (0,1)$, тогда процесс потерь агрегирующего алгоритма будет расти не быстрее процесса потерь лучшего из экспертов.

Алгоритм 1. Агрегирующий Алгоритм В. Вовка.

Вход: $x_1, \ldots, x_T$ — временной ряд;
$A^1, \ldots, A^N$ — эксперты;
$\beta \in (0,1)$ — параметр настройки весов;
$p_0 \in \mathbb{R}^N$ — начальное распределение весов экспертов;
$S\colon \Gamma^\Omega \to \mathbb{R}$ — подстановочный функционал.
Выход: $\hat{x}_{T+1}$.

для $t = 0, \ldots, T$
  для $j = 1, \ldots, N$
    $\hat{x}_{t+1}^j = A^j(x_1, \ldots, x_t)$;
    $p_{t+1}^j = p_t^j \cdot \beta^{\lambda(x_{t+1}, \hat{x}_{t+1}^j)}$;
    $p_{t+1}^j = p_{t+1}^j \Big/ \sum_{k=1}^N p_{t+1}^k$;
  определить функцию
  $g_t(x) = \log_\beta \sum_{k=1}^N p_t^k \cdot \beta^{\lambda(x, \hat{x}_{t+1}^k)}$;
$\hat{x}_{T+1} = S\big(g_T(x)\big)$.

Определение 5. *Задача прогнозирования Z является β-смешиваемой, если существует $\beta \in (0,1)$ такой, что $c(\beta) = 1$. Задача Z является смешиваемой, если она β-смешиваемая при некотором $\beta \in (0,1)$.*

Построение композиций с помощью агрегирующего алгоритма

Будем рассматривать квадратичную задачу прогнозирования $Z = \langle [Y_1, Y_2], [Y_1, Y_2], \lambda \rangle$, где $Y_1, Y_2 \in \mathbb{R},\ Y_1 < Y_2,\ \lambda(x, \hat{x}) = (x - \hat{x})^2$.

Выбор параметров β, p_0, S агрегирующего алгоритма, при которых квадратичная задача смешиваема, основан на следующих результатах.

Лемма 1. *Квадратичная задача Z является β-смешиваемой тогда и только тогда, когда*

$$\beta \geqslant \exp\left(-\frac{2}{(Y_2 - Y_1)^2}\right).$$

Лемма 1 даёт условие, при котором будут существовать такие допустимые предсказания, для которых выполнено неравенство (2) при $c(\beta) = 1$. Однако для решения задачи следует еще выбрать начальное распределение экспертов и подстановочный функционал.

Лемма 2. *Квадратичная задача Z является смешиваемой, если*

1) начальное распределение экспертов $p(\hat{x})$ таково, что для всех $\beta \in (0,1)$ и всех $x \in [Y_1, Y_2]$ функция $\varphi(\hat{x}) = p_0(\hat{x}) \cdot \beta^{(x-\hat{x})^2}$ является с точностью до константы функцией плотности вероятности на $\mathbb{R}$;

2) подстановочный функционал $S(g)$ принимает такие значения, что

$$\big(Y_1 - S(g)\big)^2 \in [0, g(Y_1)]; \quad \big(Y_2 - S(g)\big)^2 \in [0, g(Y_2)].$$

Эта лемма есть обобщение леммы 3 из [3] и позволяет значительно расширить множество функций $p_0(x)$, допустимых в качестве начального распределения экспертов. Второе условие имеет ясную геометрическую интерпретацию. Если отложить на плоскости кривую $\big((Y_1-\hat{x})^2,\,(Y_2-\hat{x})^2\big)$, где $\hat{x}\in[Y_1,Y_2]$, то подстановочный функционал может принимать только такие значения из $[Y_1,Y_2]$, которым соответствуют точки кривой, лежащие в прямоугольнике $[0,g(Y_1)]\times[0,g(Y_2)]$. Ниже приведены варианты подстановочных функций и графически пояснено, какой точке на плоскости соответствует выбранное допустимое предсказание.

$$S(g_t)=\frac{Y_2\sqrt{g_t(Y_1)}+Y_1\sqrt{g_t(Y_2)}}{\sqrt{g_t(Y_1)}+\sqrt{g_t(Y_2)}} \tag{3}$$

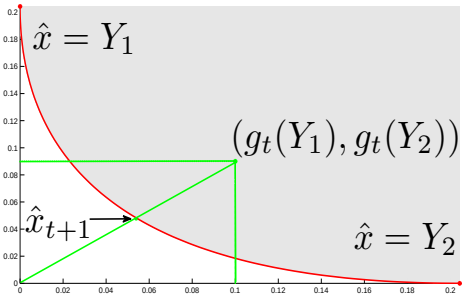


Рис. 1. Формирование прогноза $\hat{x}_{t+1}=S(g_t)$ при подстановочном функционале (3).

$$S\left(g_t\right)=\frac{g_t(Y_1)-g_t(Y_2)}{2(Y_2-Y_1)}+\frac{Y_1+Y_2}{2} \tag{4}$$

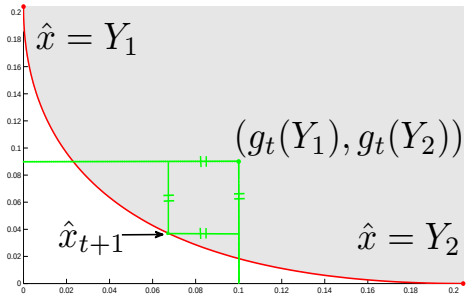


Рис. 2. Формирование прогноза $\hat{x}_{t+1}=S(g_t)$ при подстановочном функционале (4).

Построение композиций следует начать с определения множества экспертов на входе агрегирующего алгоритма. Мы рассмотрим два варианта задания множества экспертов.

В первом варианте экспертами будут базовые алгоритмы прогнозирования $\{B^1,\ldots,B^M\}$. Выбрав равномерное начальное распределение весов экспертов $p_0^j=1/M$, для $j=1,\ldots,M$, а в качестве подстановочного функционала (3), получим композицию АА.

Теорема 3. *Композиция АА на любой конечной последовательности* $(x_1,\ldots,x_T)\in[Y_1,Y_2]$, *на любом наборе экспертов* $\{B^1,\ldots,B^M\}$, *имеет следующую оценку процесса потерь:*

$$\mathcal{L}_{AA}(T)\leqslant\min_{j=1,\ldots,M}\mathcal{L}_{B^j}(T)+\frac{(Y_2-Y_1)^2}{2}\cdot\ln M. \tag{5}$$

Таким образом, процесс потерь композиции АА растёт не быстрее процесса потерь самого лучшего из экспертов. Поскольку процесс потерь определяется как суммарная потеря, накопленная к моменту времени T, относительный вклад второго слагаемого в оценку (5) убывает со временем.

Во втором варианте экспертами являются аффинные комбинации прогнозов базовых алгоритмов.

$$\hat{x}_{t+1}=\sum_{j=1}^{M}\hat{x}_{t+1}^j\cdot w^j,$$

где $\mathbf{w}=\left(w^1,\ldots,w^M\right)^{\intercal}\in\mathbb{R}^M$. В этом случае экспертом будет вектор весов $\mathbf{w}$, а множество экспертов континуально. Начальное распределение весов обозначим $p_0(\mathbf{w})$. В момент времени t обозначим $\mathbf{y}_t=(x_1,\ldots,x_t)$, $\hat{\mathbf{x}}_t=(\hat{x}_t^1,\ldots,\hat{x}_t^M)^{\intercal}$, $\mathbf{X}=$ $=(\hat{\mathbf{x}}_1,\ldots,\hat{\mathbf{x}}_t)$ — матрица размера $M\times t$.

Применим подстановочную функцию (4), при этом при различных начальных распределениях будут получаться разные композиции.

Зададим для агрегирующего алгоритма начальное распределение весов экспертов как M-мерное нормальное распределение с математическим ожиданием $\mathbf{w}_0$.

$$p_0(\mathbf{w})=\left(\frac{\alpha\ln(\frac{1}{\beta})}{\pi}\right)^{M/2}\cdot\exp\Big(-\ln\big(\tfrac{1}{\beta}\big)\alpha\|\mathbf{w}-\mathbf{w_0}\|^2\Big).$$

Таким образом, построим композицию AAR, прогноз которой записывается в аналитической форме:

$$\hat{x}_{t+1}^{AAR}=\left(\widetilde{\mathbf{y}}\widetilde{\mathbf{X}}'(\alpha\mathbf{I}+\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}')^{-1}+\mathbf{w}_0'\right)\hat{\mathbf{x}}_{t+1}, \tag{6}$$

где $\widetilde{\mathbf{X}}=(\hat{\mathbf{X}},\hat{\mathbf{x}}_{t+1})$, $\widetilde{\mathbf{y}}=(\mathbf{y}_t,0)$.

Формула (6) похожа на прогноз гребневой регрессии [5], поэтому данный алгоритм получил название Aggregating Algorithm Regression (AAR). Исследование работы AAR для решения задачи регрессии, сравнение его с гребневой регрессией проводится в работе [6].

Теперь пусть веса всех экспертов изначально будут равны $p_0(\mathbf{w})=1$, получим композицию AAR_ZERO, прогноз которой на шаге t будет иметь вид

$$\hat{x}_{t+1}^{AAR\_ZERO}=\widetilde{\mathbf{y}}\widetilde{\mathbf{X}}'(\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}')^{-1}\hat{\mathbf{x}}_{t+1}, \tag{7}$$

где $\widetilde{\mathbf{X}} = (\hat{\mathbf{X}}, \hat{\mathbf{x}}_{t+1})$, $\widetilde{\mathbf{y}} = (\mathbf{y}_t, 0)$.

Данный выбор начального распределения весов экспертов является естественным при отсутствии априорных предпочтений в пользу какой-либо комбинации прогнозов экспертов из-за широкого разнообразия временных рядов.

Теорема 4. *Для любого $\alpha > 0$, любого $M \in \mathbb{N}$, любого $\mathbf{w}_0 \in \mathbb{R}^M$ при условии $\|\mathbf{x}_t\|_\infty < C < \infty$ справедлива оценка процесса потерь*

$$\mathcal{L}_{AAR}(T) \leqslant \inf_{\mathbf{w}} \left(\mathcal{L}_{\mathbf{w}}(T) + \alpha\|\mathbf{w} - \mathbf{w_0}\|^2\right) + \\ + \frac{M(Y_2 - Y_1)^2}{4} \ln\left(\frac{TC^2}{\alpha} + 1\right), \quad (8)$$

$$\mathcal{L}_{AAR\_ZERO}(T) \leqslant \inf_{\mathbf{w}}(\mathcal{L}_{\mathbf{w}}(T)) + O\left(\ln(T)\right). \quad (9)$$

В неравенствах (8) и (9) вторые слагаемые в правой части растут медленнее первых, поэтому их вклад в оценку процессов потерь AAR и AAR_ZERO соответственно с течением времени уменьшается.

Временная сложность композиции AA линейно зависит от числа базовых алгоритмов и длины временного ряда $O(MT)$. Для AAR и AAR_ZERO временная сложность равна $O(M^3T)$.

Численные эксперименты проводились на реальных данных продаж в розничной сети магазинов Лама (Томск). Фактическая длина рядов варировалась от 50 до 1500 отсчетов, всего 2000 временных рядов.

В качестве базовых алгоритмов прогнозирования использовались экспоненциальное сглаживание (ES), адаптивное экспоненциальное сглаживание (AES), алгоритм Тейла-Вейджа (TW), скользящее среднее (MA). Настройка параметров базовых алгоритмов прогнозирования и параметров композиций проводилась на репрезентативной выборке из 200 рядов (в этих рядах наблюдались тренд, сезонность, выбросы, пропуски данных). На остальных рядах проводился анализ качества предложенных композиций (при уже настроенных параметрах) и сравнение их с ранее известными композициями.

Оценки (5), (8), (9) подтвердились в экспериментах. При этом вторые слагаемые в формулах (5), (9) дают небольшой вклад в финальную ошибку алгоритмов, поэтому соотвествующие алгоритмы прогнозируют не хуже лучшего из базовых. Реализация композиции AAR_ZERO потребовала обработку случая, когда матрица $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}'$ из (7) вырождена (тогда веса базовых алгоритмов берутся с предыдущего шага). Поэтому качество данного метода сильно зависит (особенно при малых t) от прогнозов базовых алгоритмов (процесс потерь AAR_ZERO может сильно возрасти, если один из базовых алгоритмов плохо прогнозирует).

Сравнение композиций AA, AAR, AAR_ZERO (ZERO) с уже известными композициями: ЛАВР [8], AFTER [7], IW [4] приведено в таблице 1.

Число базовых	AA	AAR	ZERO	AFTER	IW	ЛАВР
10	6,43	6,37	**6,35**	6,57	6,66	6,74
25	6,39	**6,31**	6,59	6,50	6,62	6,92
40	**6,35**	6,37	6,72	6,55	6,57	6,90

Таблица 1. Среднеквадратичная ошибка композиций при различном числе M базовых алгоритмов.

Основной вывод состоит в том, что предложенные алгоритмы в среднем работают лучше ранее известных методов: на 500 рядах алгоритмы AA, AAR провосходят по качеству все остальные композиции на 3-5%. При этом время работы предложенных композиций сравнимо со временем работы самых быстрых из ранее известных композиций (IW).

Итак, в данной работе показано, что композиции, основанные на агрегирующим алгоритме В. Вовка, вполне применимы для практического прогнозирования временных рядов.

Литература

[1] *Лукашин Ю. П.* Адаптивные методы краткосрочного прогнозирования временных рядов. — // Москва: Финансы и статистика, 2003. — 415 с.

[2] *Vovk V.* A Game of Prediction with Expert Advice // Journal of Computer and System Sciences. — 1997. — Т. 56. — С. 153–173.

[3] *Vovk V.* Competitive on-line statistics // International Statistical Review, 1999. — Vol. 69. — Pp. 2001.

[4] *Timmermann A. G.* Forecast Combinations — `http://ideas.repec.org/p/cpr/ceprdp/5361.html` — CEPR Discussion Papers — 2006.

[5] *Hoerl A. E., Kennard R. W.* Ridge regression: biased estimation for nonorthogonal problems // Technometrics, 2000. — Vol. 42, No. 1. — Pp. 80–86.

[6] *Busuttil S., Kalnishkan Y.*, Weighted Kernel Regression for Predicting Changing Dependencies Proceedings of the 18th European conference on Machine Learning. — Berlin: Springer-Verlag, 2007. — С. 535–542.

[7] *Yang Y.* Combining forecasting procedures: some theoretical results // Econometric Theory, 2004. — Vol. 20. — Pp. 176–222.

[8] *Воронцов К. В., Егорова Е. В.* Динамически адаптируемые композиции алгоритмов прогнозирования // Донецк: Искусственный Интеллект, №2, 2006.— С. 277–280.

Результаты исследования методов прогнозирования и моделей данных

Андреев А. В., Пытьев Ю. П.
alvlandr@gmail.com, yuri.pytyev@gmail.com
Москва, физический факультет МГУ им. М. В. Ломоносова

Рассматриваются математические методы прогнозирования на примере прогноза динамики курсов акций. Проводится сравнение использующихся методов. Делается попытка восстановить вероятностную и возможностную модель данных.

Известно, что самый ценный ресурс в мире — информация, способность же качественно прогнозировать — один из основных инструментов для получения надёжной информации. В работе рассмотрены математические методы, основанные на анализе временных рядов, на примере предсказания значения стоимостей акций пяти различных компаний. На данный момент известно много методов и моделей, используемых при анализе временных рядов и прогнозировании: модель авторегрессии (AR) [1, 2], модель авторегрессии-скользящего среднего (ARMA) [1, 2], подход Бокса–Дженкинса [1] и т. д.

Целью данной работы являлось:

1) создание методов, обеспечивающих высокую точность прогноза стоимостей акций на l шагов вперёд («на l следующих дней»);

2) создание методов, позволяющих с высокой точностью прогнозировать направление изменения стоимостей акций (повышение/понижение).

Прогнозирование стоимости акций на l «следующих дней».

Сформирован портфель из акций пяти различных компаний. Известна динамика изменения курса этих акций — стоимость акций за определённый период времени: n дней. Требуется оценить стоимость акции каждой компании в период с $(n+1)$-го по $(n+l)$-ый день. В работе представлены три подхода, позволяющие решить поставленную задачу.

Модель с предиктором. Данный метод является методом линейного прогнозирования, аналогичным методу наименьших квадратов (МНК). В его основе лежит решение следующей задачи на минимум:

$$||YX - Z|| \sim \min_{X \in \mathbb{R}^M}, \qquad (1)$$

где Y — матрица, каждая строка которой составлена из пяти временных рядов (по количеству типов акций) одинаковой длины, элементами (уровнями) которых являются стоимости акций, следующая строка получается сдвигом предыдущей на один шаг вперёд (таким образом, что в каждой следующей строке кроме первой используется пять новых значений относительно предыдущей), X — предсказывающая матрица или предиктор, Z — матрица, каждая строка которой состоит из $5l$ значений стоимостей акций, опережающих значения в соответствующей строке матрицы Y на число шагов от 1 до l. Проще говоря, при умножении одной строки матрицы Y на предиктор появятся $5l$ будущих значений стоимостей акций, составляющих матрицу Z.

Вероятностный аналог данного метода рассмотрен в [4].

Метод, минимизирующий среднеквадратичное отклонение. Пусть A_k — матрица размера $5\times m$, каждая из пяти строк которой (по числу типов акций) представляет собой временной ряд длины m, $\alpha_k \in \mathbb{R}^5$ — вектор, составленный из цен акций, опережающих соответствующие значения в A_k на один шаг (на один день). Введём векторы $x \in \mathbb{R}^m$ и y:$5\times l$ такие, что преобразование $A_k x + y$ было как можно ближе в среднеквадратичном к α_k. Таким образом, необходимо решить следующую задачу:

$$\frac{1}{s-1}\sum_{k=m+1}^{s+m} ||\alpha_k - A_k x - y||^2 \sim \min_{x,y}. \qquad (2)$$

Требуется решить задачу на минимум (2) и найти x, y, если известны вектор α_k и матрица A_k.

Возможностное прогнозирование. Рассмотрим следующую схему наблюдений:

$$\eta = AX + \nu, \qquad (3)$$

где $\eta = (\eta_1, \ldots, \eta_5)$, $j = 1, \ldots, 5$, η_j — нечёткий вектор, моделирующий стоимость акций, A и X являются полными аналогами матриц Y и X в (1) соответственно, $\nu = (\nu_1 \ldots \nu_5)$, $j = 1, \ldots, 5$, ν_j — нечёткий вектор «ошибок» (под ошибкой следует понимать отклонение стоимости от некоторого стабильного положения — линии тренда). Определённые предположения относительно нечётких векторов ошибок позволяют свести задачу (3) к задаче линейного программирования.

Прогнозирование направления изменения стоимости акций на следующий день.

Известна динамика изменения стоимости акций одной компании — стоимость акций за определённый период времени: n дней. Требуется определить

повысится или понизится цена за акцию в $(n+1)$-й день.

Метод с предиктором. Пусть $A_j \in \mathbb{R}^s$, $j = 1, 2, \ldots$ — временной ряд, элементами (уровнями) которого являются стоимости акций компании. Матрица парных сравнений $M_j = \{m_{lk}\}_j$, $l, k = 1, \ldots, s$ получается из временного ряда A_j следующим образом:

$$m_l k = \begin{cases} 1, & \text{если } a_l > a_k; \\ 0, & \text{если } a_l = a_k; \\ -1, & \text{если } a_l < a_k. \end{cases} \qquad (4)$$

Матрица M_j антисимметрична, и все диагональные элементы равны нулю. Пусть M_Y — матрица, состоящая из матриц $M_1, M_2, \ldots$; M_Z — также матрица, состоящая из матриц $M_1, M_2, \ldots$, но опережающих соответствующие значения в M_Y на один шаг, т.е. если $M_Y = (M_1, \ldots, M_i, \ldots, M_{n-s})^\intercal$, то $M_Z = (M_2, \ldots, M_{i+1}, \ldots, M_{n-s+1})^\intercal$. Решается следующая задача на минимум:

$$||M_Y X - M_Z|| \sim \min_x . \qquad (5)$$

Метод, вычисляющий условные вероятности. На заданном временном интервале согласно (4) строятся матрицы парных сравнений $M_i^{(3)}$ и $M_j^{(4)}$, $i = 1, \ldots, (n-3)$, $j = 1, \ldots, (n-4)$, $M_i^{(3)}$:3×3, $M_j^{(4)}$:4×4. Далее, на основе построенных матриц, находится вероятность появления матриц $M_i^{(3)}$ и$M_j^{(4)}$ при «следующем испытании» (на следующий день). Полученные значения вероятностей позволяют рассчитать условную вероятность появления матрицы $M_{(n-3)}^{(4)}$, если матрица размера 3×3 равна $M_{(n-3)}^{(3)}$, иначе говоря, вероятность того, как поведёт себя стоимость акции на следующий день при условии, что в течение трёх дней до этого цена изменялась определённым образом.

Построение вероятностной и возможностной моделей данных. Описанные выше методы подходят для любых временных рядов и не используют данные об их модели. Рассмотрим матрицы парных сравнений и, на основании данных, полученных за определённый фиксированный период времени, посчитаем частоту появления различных матриц. Далее исследуем эволюцию этих частот во времени. Полученная зависимость и учёт влияния «дрейфа», присутствующего в данных, позволяют судить о том, возможно ли эмпирически восстановить вероятностную модель. Возможностная же модель может быть достаточно простой, но для эмпирического восстановления возможностной модели данных требуется, чтобы при полученной эволюции вероятностей сохранялась их упорядоченность.

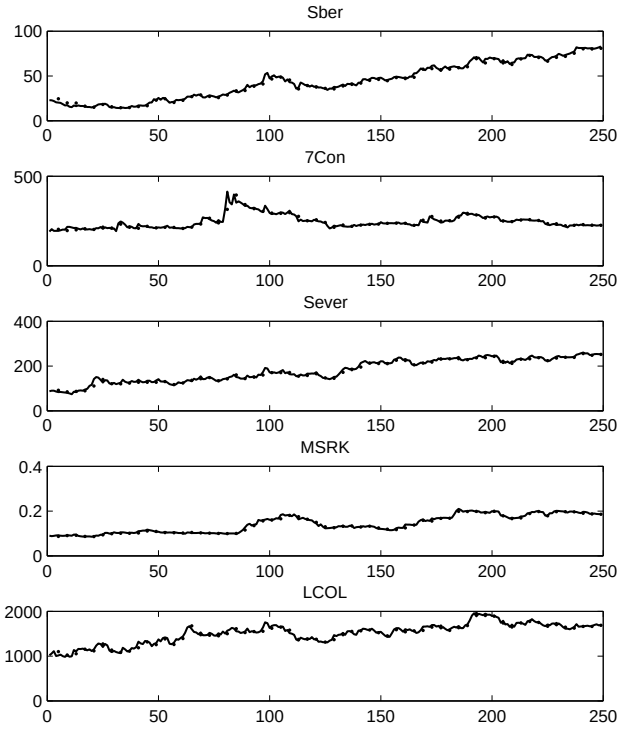


Рис. 1. Графики зависимости стоимости акций пяти компаний от времени. Сплошной линией показана реальная динамика курса акций, пунктирной — прогнозируемая, полученная методом, минимизирующим среднеквадратичное отклонение, при $l = 1$.

Выводы

Предложены методы прогнозирования изменения динамики курсов акций, обеспечивающие точность предсказания сравнимую с уже известными методами. В проведённом исследовании описанные методы модифицировались: учитывалось различное влияние разных типов акций друг на друга при помощи весовых коэффициентов, вычислялась эффективная размерность множества данных.

Компьютерное моделирование было проведено для акций пяти компаний, торгующихся на ММВБ: Сбербанк, 7 Континент, МСРК, Северсталь, ЛУКОЙЛ. Результаты моделирования при использовании модели, минимизирующей среднеквадратичное отклонение, показаны на рис. 1. На рис. 2 показан график изменения относительной ошибки прогноза в i-ый день, которая определялась следующим образом:

$$E_i = \frac{k_i - k_i^{pr}}{k_i}, \qquad (6)$$

где где k_i — стоимость акции в i-ый день, k_i^{pr} — полученное в результате прогноза значение стоимости акции в i-ый день. Полученные результаты показывают, что наилучшую в среднем точность обеспечивает использование модели, минимизирующей

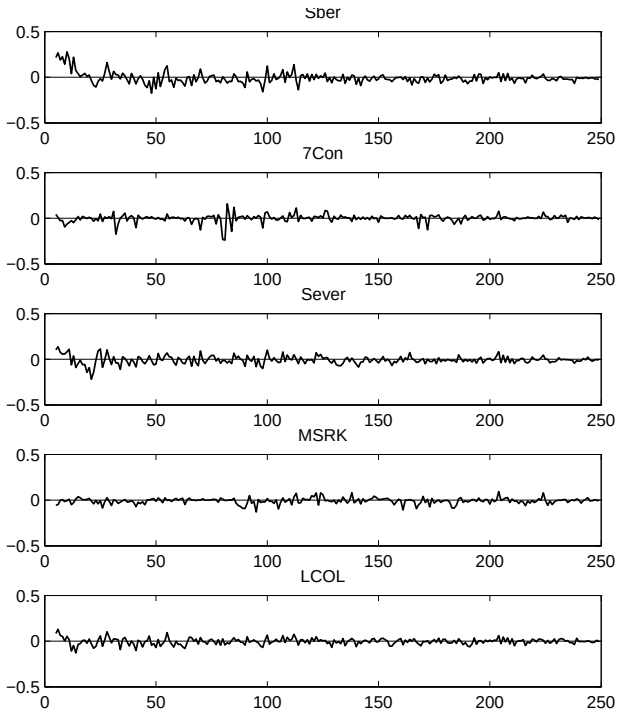


Рис. 2. Относительная ошибка прогнозирования E_i. Использовался метод, минимизирующий среднеквадратичное отклонение, $l = 1$.

среднеквадратичное отклонение. Также она является наименее затратной относительно компьютерной реализации. Возможностное прогнозирование является наименее точным и наиболее затратным. Относительно низкая точность метода этого связана с предположениями, сделанными при решении задачи, а затраты объясняются тем, что решается 5 задач линейного программирования с большим количеством условий.

Точность, получаемая при прогнозировании стоимостей акций, существующими методами, по-видимому, не может быть значительно повышена, поэтому наибольший интерес представляет прогнозирование не конкретных значений цен, а поведения этих значений (прогнозирование тренда). Актуальным остаётся вопрос о построении вероятностной и/или возможностной модели по данным значениям. На данный момент можно утверждать, что эмпирически построить стохастическую модель не удастся. Одним из следующих этапов исследования несомненно будет восстановление возможностной модели. Более подробно о достигнутых результатах, модификациях исходных методов будет рассказано на конференции.

Литература

[1] *Канторович Г. Г.* Лекции «Анализ временных рядов» // Экономический журнал ВШЭ. — 2002. — Т. 6, № 1–4.

[2] *Мельников А. В., Попова Н. В., Скорнякова В. С.* Математические методы финансового анализа. — Москва: Анкил, 2006. — 440 с.

[3] *Пытьев Ю. П.* Возможность как альтернатива вероятности — Москва: Физматлит, 2007. — 464 с.

[4] *Цыплаков А.* Введение в прогнозирование в классических моделях временных рядов // Квантиль, № 1. — С. 3–19.

Методика аппроксимации временного ряда разностью двух выпуклых функций одной переменной*

Финкельштейн Е. А., Горнов А. Ю.
Evgeniya.Finkelstein@gmail.com
Иркутск, Институт математики экономики и информатики ИГУ,
Институт динамики систем и теории управления СО РАН

Предлагается методика структурной идентификации временного ряда, основанная на его аппроксимации с помощью кубических сплайнов и разложении аппроксимирующих функций на разность двух выпуклых. В качестве инструмента разложения используется задача оптимального управления, формируемая на основе нескольких различных интегральных функционалов. Приводятся результаты численных экспериментов.

Введение

Проблема исследования структуры существенно невыпуклой функции естественно возникает в задачах анализа временных рядов. С применением простых методов интерполяции задача исследования свойств временного ряда может быть сведена к задаче анализа функции одной переменной. Представляется, для такой редукции целесообразно использовать хорошо разработанные и достаточно надежные алгоритмы сплайн-аппроксимации. Исследовав полученные в результате аппроксимационных процедур функции можно, с той или иной степенью достоверности, пытаться делать заключения и о свойствах изучаемого временного ряда.

Среди теоретических результатов, полученных для функций действительного переменного, особое, на наш взгляд, место занимают работы по разложениям функций на разность выпуклых («ПРВ-функций»). Вопрос о возможности такого точного разложения рассматривался в ряде работ (см., например, [1–3]) и решен положительно. Заметим, что термин, введенный для этого класса функций еще в 40-х годах прошлого века советским геометром А. Д. Александровым — ПРВ («поверхности, представимые разностью выпуклых» [1]) в последние годы, к сожалению, старательными усилиями зарубежных специалистов вытесняется в научной литературе другим — DC (difference of convex, см., например, [2]). Класс ПРВ-функций достаточно обширен: он содержит, очевидно, все выпуклые функции и также все дважды непрерывно дифференцируемые [3]. Потенциал этого контринтуитивного теоретического результата огромен, но пока явно не реализован, и только в последние годы изредка наблюдаются попытки создания на его основе вычислительных методов. По-видимому, дело в том, что поиск конструктивных алгоритмов построения такого разложения, удовлетворяющего, помимо того, дополнительным требованиям, пока остается проблемой.

В докладе рассматривается методика разложения невыпуклой дважды непрерывно дифференцируемой функции $f(t) = g(t) - h(t)$, где $g(t)$ и $h(t)$ — выпуклые, построенной с помощью кубических сплайнов [4] на основе исходного временного ряда, позволяющая синтезировать конкретные выпуклые составляющие функции. Обсуждаемая задача, очевидно, относится к классу некорректно поставленных, поэтому для обеспечения единственности решений предложено применять регуляризирующие функционалы, которые, с другой стороны, можно рассматривать как дополнительные критерии качества, позволяющие придавать разложениям те или иные новые свойства. В качестве основного инструмента численного анализа используются алгоритмы поиска оптимального управления, реализованные в комплексе программ OPTCON-I [5].

Построение методики

Предложенная методика заключается в формировании линейной задачи оптимального управления с четырьмя фазовыми переменными и двумя управляющими воздействиями.

$$\begin{cases} \dot{x}_1 = x_2(t); \\ \dot{x}_2 = u_1(t); \\ \dot{x}_3 = x_4(t); \\ \dot{x}_4 = u_2(t), \end{cases} \quad u_1 \geqslant 0;\ u_2 \geqslant 0.$$

Введя ограничения неотрицательности на управления, мы гарантируем и неотрицательность вторых производных нечетных траекторий $x_1(t)$ и $x_3(t)$, т. е. их выпуклость, разумеется, не строгую.

Сформировав динамическую систему, поставим задачу минимизации невыпуклого интегрального функционала, штрафующего за расхождение разности нечетных траекторий системы и исходной декомпозируемой функции.

$$J_0(u) = \int_{t_0}^{t_1} \Big(f(t) - \big(x_1(t) - x_3(t)\big) \Big)^2 dt \to \min .$$

Вычислительная технология

На первом этапе предложенной технологии исследуемый временной ряд аппроксимируется непрерывной функцией с применением алгоритмов

Работа выполнена при финансовой поддержке РФФИ, проект № 09-07-00267.

кубической сплайн-интерполяции, строится $S(t)$. Далее, для отыскания конкретных выпуклых составляющих временного ряда, необходимо решить вопрос начальных значений. Согласно постановке задачи разность искомых функций в начальной точке равна значению ряда в этой точке. То же мы можем сказать и о производных:

$$\begin{cases} x_1(t_0) - x_3(t_0) = S(t_0); \\ x_2(t_0) - x_4(t_0) = S'(t_0). \end{cases}$$

В случае искусственной тестовой задачи, когда мы знаем значение производной, мы можем его использовать, в случае реального ряда — только грубо оценить. В любом случае этого не достаточно, поэтому установим, что начальное значение одной из функций, например первой, равно первому значению ряда, а второй — нулю, что в терминых задачи выглядит так:

$$\begin{cases} x_1(t_0) = S(t_0); \\ x_3(t_0) = 0. \end{cases}$$

Начальные значения x_2, x_4 задаем на основе имеющихся условий, соображений здравого смысла и экспертных гипотез. Гипотезы относительно начал выпуклых составляющих, если таковые имеются, также легко учесть.

Определим набор дополнительных функционалов, записываемых в терминах траекторий управляемой системы:

- «Минимум энергии»

$$J_1(u) = \int_{t_0}^{t_1} \big(u_1^2(t) + u_2^2(t)\big)dt \to \min;$$

- «Минимум воздействия»

$$J_2(u) = \int_{t_0}^{t_1} \big(u_1(t) + u_2(t)\big)dt \to \min;$$

- «Минимум расхождения»

$$J_3(u) = \int_{t_0}^{t_1} \big(x_1(t) + x_3(t)\big)^2 dt \to \min;$$

- «Минимум амплитуды» составляющих функций

$$J_4(u) = \int_{t_0}^{t_1} \big(x_1^2(t) + x_3^2(t)\big)dt \to \min.$$

Вычислительный эксперимент

Предложенная методика тестирования включает выбор тестовой функции, задание сетки по времени с постоянным шагом, построение на основе

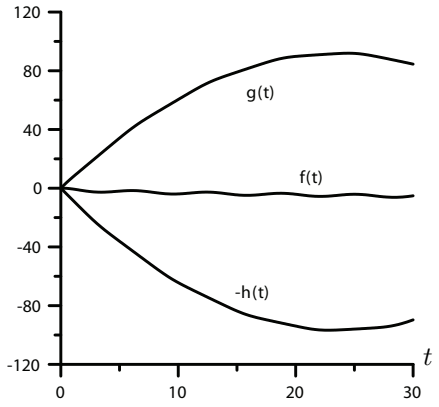


Рис. 1. Задача 1, $J^* = J_0 = 0{,}2063$.

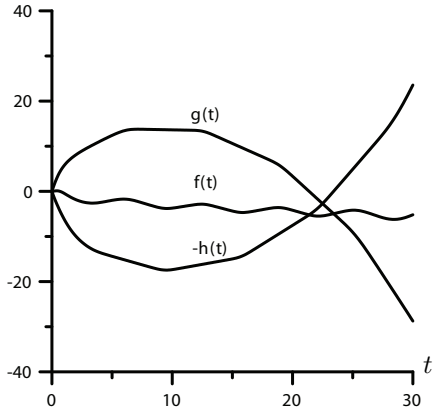


Рис. 2. Задача 1, $J^* = J_0 + 10^{-4}J_3 = 2{,}3277$.

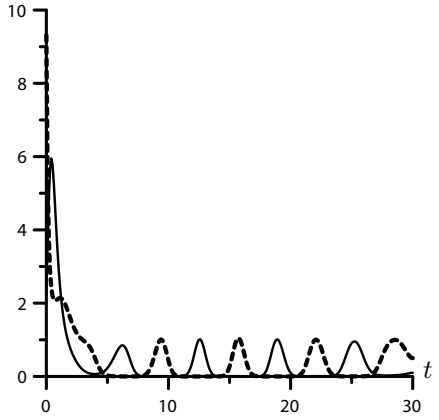


Рис. 3. Задача 1, $J^* = J_0 + 10^{-4}J_3$, u_1 — пунктиром, u_2 — сплошной линией.

выбранной функции временного ряда и применение алгоритма сплайн-интерполяции. В сформированной задаче оптимального управления необходимо учесть как требование близости разности траекторий к исследуемому временному ряду, так и дополнительные критерии качества, что достигается путем свертки функционалов с коэффициентами, подобранными экспериментальным путем.

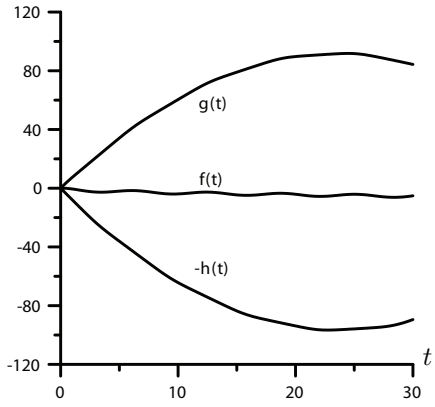


Рис. 4. Задача 1, $J^* = J_0 + J_1 = 0{,}2687$.

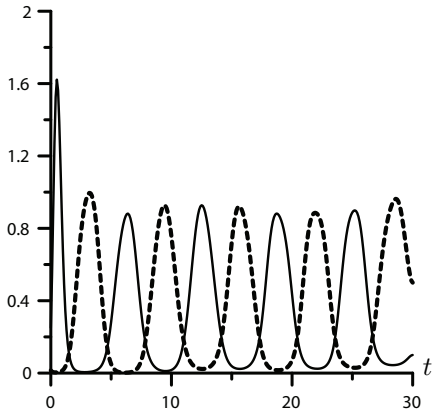


Рис. 5. Задача 1, $J^* = J_0 + J_1$, u_1 — пунктиром, u_2 — сплошной линией.

На графиках будут представлены ПРВ-разложения тестовой функции, полученные с применением предлагаемой методики, значения суммарного функционала J^* и полученные управления системы, когда это имеет смысл.

Тестовая задача 1.

$$f(t) = \sin\bigl(\cos(t)\bigr) - \sqrt{t}, \quad t \in [0, 30].$$

В данном случае целесообразно поставить задачу на «минимум расхождения». Результат решения приведен на рис. 2, на рис. 3 приведены полученные управления.

В эксперименте, показанном на рис. 2, дополнительный функционал J_3 входит в целевой функционал J^* с коэффициентом 10^{-4}, оптимальным для сохранения качества аппроксимации на приемлемом уровне и уменьшения расхождения.

При решении задачи с дополнительным функционалом «минимума энергии» также разумно привести график динамики управлений (рис. 5), и сравнить его с графиком, полученным при решении задачи с другим функционалом (рис. 3).

Тестовая задача 2.

$$f(t) = 0{,}1(t-15)^2 - \sin(t) - \cos(2t), \ t \in [0, 30].$$

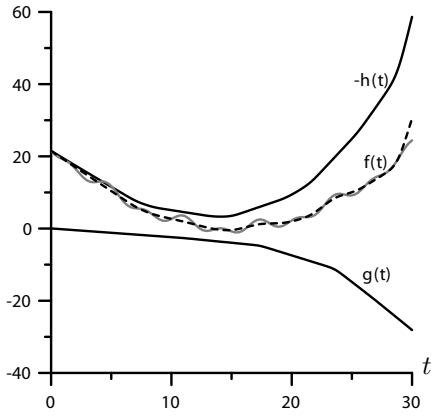


Рис. 6. Задача 2, $J^* = J_0 + J_1 + 10^{-4} J_3$.

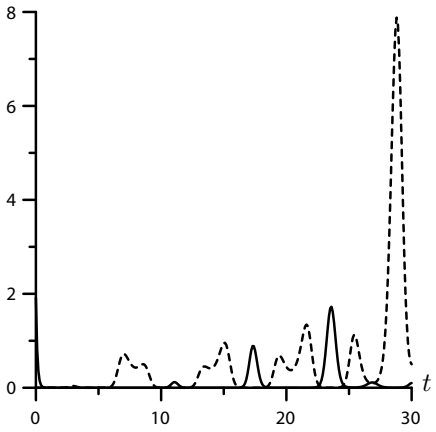


Рис. 7. Задача 2, u_1 — пунктиром, u_2 — сплошной.

Приводится один вариант решения задачи аппроксимации временного ряда разностью двух выпуклых, построенный на основе следующей свертки функционалов: $J^* = J_0 + J_1 + 10^{-4} J_3$. Поскольку выбор именно такого функционала исходит из предположения о наличии шумов, то дополнительно на графике покажем разность полученных траекторий.

На рис. 6 серым показана тестовая функция, а пунктиром — разность полученных разложений $g(t) - h(t)$. На рис. 7 показаны графики управлений системы.

Задача анализа уровня загрязнения воздуха. Предложенная методика использована при решении задачи оценки показателей критического загрязнения воздуха в г. Чита. На основе данных в виде временных рядов, предоставленных экспертом — заведующей лабораторией Института медицины труда и экологии человека СО РАМН д. м. н. Н. В. Ефимовой, был выполнен структурный анализ краткосрочного загрязнения в период март–май 2008 г., когда вблизи города были отмечены массовые лесные пожары. Выполненные расчеты (см. рис. 8) позволили оценить предельный уровень

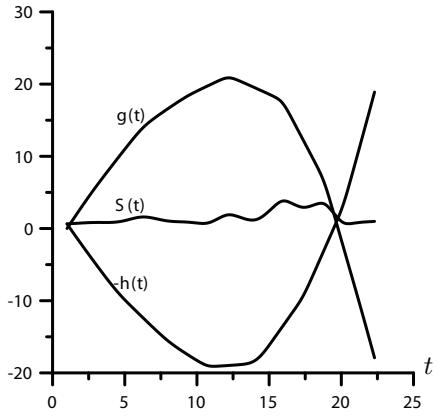


Рис. 8. Задача анализа уровня загрязнения воздуха. $J^* = J_0 + J_3 = 0{,}2442$.

краткосрочного загрязнения атмосферного воздуха примесями, выше которого наблюдается устойчивое повышение ежедневной обращаемости за скорой медицинской помощью и смертности населения. На основе полученных результатов экспертами были сформулированы рекомендации для органов муниципального самоуправления региона.

Выводы

Проведенные численные эксперименты, включающие несколько десятков модельных примеров, подтвердили принципиальную работоспособность предложенного подхода. Полученные результаты позволяют надеяться на возможность построения более общих вычислительных методик, основанных на новых динамических моделях.

Литература

[1] *Александров А. Д.* О поверхностях, представимых разностью выпуклых функций // Изв. АН КазССР. Сер. математика и механика. — 1949. — Вып. 3. — С. 3–20.

[2] *Tuy H.* Convex Analysis and Global Optimization // Kluwer Academic Publishers, Dordrecht, 1996.

[3] *Александров А. Д.* Поверхности, представимые разностями выпуклых функций // Докл. АН СССР. — 1950. — Т. 72, № 4. — С. 613–616.

[4] *Форсайт Дж., Малькольм М., Моулер К.* Машинные методы математических вычислений. — М.: Мир, 1980. — 279 с.

[5] *Горнов А. Ю, Дивакова А. О.* Комплекс программ для численного решения задач оптимального управления. Руководство пользователя. — Иркутск, 1990. — 36 с.

Структурная идентификация сложных объектов управления*

Дорофеюк Ю. А., Дорофеюк А. А.
dorofeyuk_julia@mail.ru
Москва, Институт проблем управления им. В. А. Трапезникова РАН

Рассмотрена задача структурной идентификации сложных объектов. Для её решения предложено использовать методологию классификационного анализа, в том числе алгоритмы кусочной аппроксимации сложной зависимости.

Работа посвящена решению задачи построения модели функционирования сложного объекта управления с помощью алгоритмов структурной идентификации. Основная идея структурной идентификации состоит в разбиении пространства аргументов (входных параметров) на такие локальные области, в пределах каждой из которых сложную во всём пространстве функцию (зависимость) можно аппроксимировать достаточно простыми функциями (например, линейными).

Задача структурной идентификации сложных объектов

Далее для простоты рассматривается статическая модель $y = F(x)$ функционирования идентифицируемого объекта — как модель зависимости выходного показателя y от вектора входных показателей $x = (x^{(1)}, \ldots, x^{(k)}) \in X \subseteq \mathbb{R}^k$. Такая модель строится по выборке

$$(y_t, x_t) = (y_t, x_t^{(1)}, \ldots, x_t^{(k)}) \in \widetilde{X} \subseteq \mathbb{R}^{k+1}$$

из n векторов размерности $(k+1)$, получаемых в режиме нормальной эксплуатации идентифицируемого объекта. Без особого труда можно показать, что предлагаемый далее подход может использоваться также для идентификации динамической модели достаточно общего вида

$$y(t) = F[x(t), \ldots, x(t-m)],$$

где m — «глубина памяти» динамической модели.

Было замечено, что многие сложные объекты могут работать в нескольких режимах, существенно различающихся своими моделями $y = F_j(x)$, где j — индекс режима [1]. При этом j-му режиму соответствует определённая область H_j в пространстве входных параметров X. В [2] для идентификации такого рода объектов впервые было предложено использовать методы структурной (кусочной) аппроксимации. А именно, предлагается модель $F(x)$ рассматривать в виде:

$$F(x) = \sum_{j=1}^{r} \varepsilon_j(x) F_j(x),$$

где $\varepsilon_j(x)$ — характеристические функции областей $H_j \in X$, $\bigcup\limits_{j=1}^{r} H_j = X$, r — число областей. По определению $\varepsilon_j(x) = 1$, если $x \in H_j$ и $\varepsilon_j(x) = 0$ в противоположном случае. Обычно в качестве оценок локальных моделей $\widetilde{F}_j(x)$ используются достаточно простые функции, например — линейные. По этой причине далее рассматривается только кусочно-линейная модель.

Классическая схема структурной идентификации состоит в следующем [3, 5]. Пространство разбивается на r классов $(H_1, \ldots, H_r)$. Затем в каждом классе H_i строится линейная оценка $\widetilde{F}_j(x)$ локальной модели идентифицируемого объекта (зависимости $y = F(x)$ выходного показателя y от вектора входных показателей x). Другими словами, для каждого класса H_i находится линейная оценка локальной модели

$$\tilde{F}_j(x) = \big((c_i, x) + d_i\big) = d_i + \sum_{j=1}^{k} c_i^{(j)} x^{(j)},$$

то есть находится такой k-мерный вектор коэффициентов $c_i = (c_i^{(1)}, \ldots, c_i^{(k)})$ и константа d_i, для которых минимизируется функционал остаточной дисперсии y относительно модели $\widetilde{F}_j(x_j)$:

$$K_i = \sum_{x_j \in H_i} \big(y_j - \widetilde{F}_j(x_j)\big)^2,$$

где $y_j = F_i(x_j)$.

Тогда задача структурной кусочно-линейной идентификации состоит в нахождении такого разбиения на классы (структуризации множества входных векторов), для которого остаточная дисперсия в среднем по всем классам была бы минимальна. Другими словами, необходимо найти такую классификацию $(H_1, \ldots, H_r)$ и такие векторы коэффициентов $c_i = (c_i^{(1)}, \ldots, c_i^{(k)})$ и константы d_i, для которых функционал

$$I = \sum_{i=1}^{r} K_i = \sum_{i=1}^{r} \sum_{x_j \in H_i} \Big(y_j - \big((c_i, x_j) + d_i\big)\Big)^2$$

принимал бы минимальное значение [5].

Работа выполнена при частичной финансовой поддержке РФФИ, проекты № 08-07-00347, 10-07-00210.

Удобнее записать функционал качества аппроксимации в форме:

$$J = -I = -\sum_{i=1}^{r}\sum_{x_j\in H_i}\Big(y_j - \big((c_i, x_j) + d_i\big)\Big)^2.$$

В этом случае задача состоит в максимизации функционала J. Последний функционал является частным случаем функционала классификационного анализа общего вида [4]:

$$J(H, A) = \sum_{x\in X}\sum_{i=1}^{r} K\big(x, \alpha_i\big)\varphi\big(h_i(x)\big). \qquad (1)$$

Классификация $H = (H_1, \ldots, H_r)$ задаётся через вектор-функцию принадлежностей

$$H\big(x\big) = \big(h_1(x), \ldots, h_r(x)\big).$$

Тогда функционал I можно переписать в следующих двух эквивалентных записях:

$$I_1 = \sum_{i=1}^{r}\sum_{j=1}^{n}\Big[y_j - \big((c_i, x_j) + d_i\big)\Big]^2 h_i(x); \qquad (2)$$

$$I_2 = \sum_{j=1}^{n}\Big[y_j - \sum_{i=1}^{r}\big((c_i, x_j) + d_i\big)h_i(x)\Big]^2. \qquad (3)$$

Если классификация чёткая (каждый объект однозначно относится к одному из классов, т. е. равно либо 1, либо 0), то функционалы (2) и (3) совпадают, однако интерпретируются они по-разному. При минимизации функционала (2) оценки локальных линейных моделей строятся отдельно для каждого класса, а затем суммируются квадраты отклонения ошибок по всем классам. В функционале (3) выражение $\sum\limits_{i=1}^{r}\big((c_i, x_j) + d_i\big)h_i(x)$ можно считать оценкой в целом кусочно-линейной модели выходного показателя y, т. е. для всего пространства X. Решению задачи построения кусочно-линейной модели в такой постановке посвящены, например, публикации [1, 3, 5].

Основной сложностью решения данной задачи является то, что при минимизации функционалов (2) и (3) по классификации и по коэффициентам модели решающие правила (границы) оптимальной классификации записываются в терминах не только входных, но и выходного показателя. Это исключает возможность использования такой модели для прогноза. Поэтому используются проекции классификации в пространстве $\tilde{X} = X \cup y$ на пространство X входных показателей. Подобное проектирование областей на пространство меньшей размерности приводит к появлению в пространстве входных показателей зон, в которых одновременно могут действовать модели как одного, так и другого класса. Именно исходя из этого, при построении структурно-классификационных моделей по существу возникает размытость между классами.

Использование размытой кусочно-линейной аппроксимации в задаче структурной идентификации

Рассмотрим случай размытой структурно-линейной идентификации. Классификацию будем задавать через вектор-функцию принадлежностей $H(x)$ с ограничениями:

$$\sum_{i=1}^{r} h_i(x) = 1; \quad h_i(x) \geqslant 0; \quad x \in X; \quad i = 1, \ldots, r.$$

В данном случае функционалы (2) и (3) не совпадают, и их минимизация приводит к разным результатам. Легко показать, что в оптимальной кусочно-линейной модели для функционала (2) классификация будет чёткой. В соответствии с общей методикой обобщённого среднего [4] для получения структурно-линейной модели с размытой классификацией функционал (2) необходимо модифицировать следующим образом:

$$I_3 = \sum_{i=1}^{r}\sum_{j=1}^{n}\Big[y_j - \big((c_i, x_j) + d_i\big)\Big]^2 \varphi\big(h_i(x)\big). \qquad (4)$$

В выражении (4) функция $\varphi(h)$ может принимать следующий вид:

1) $\varphi_1(h) = h$ приводит к случаю чёткой классификации;
2) $\varphi_2(h) = (h)^t$, $t > 1$ приводит к случаю структурно-линейной модели с реально размытой классификацией, ($h_i(x) > 0$ для нескольких классов);
3) $\varphi_3(h) = t - \sqrt{t^2 - (t-1)h}$, $t > 1$ приводит к случаю классификации с размытыми границами (размываются только границы классов).

С точностью до знака функционал (4) является частным случаем функционала (2). Следовательно, для его оптимизации можно использовать общий итерационный алгоритм классификационного анализа [4]. Алгоритм состоит в последовательном применении двухэтапной процедуры:

— на первом этапе фиксируется вектор-функция $H(x)$, и для неё находятся оптимальные значения оценок коэффициентов линейных моделей из (4);
— на втором этапе найденные оценки коэффициентов фиксируются, и для них находится оптимальная вектор-функция $H(x)$.

Сходимость этого алгоритма следует из сходимости общего алгоритма [4]. Также как и для случая чёткой классификации, тем же недостатком этого подхода является вхождение в уравнения границ оптимальной классификации как входных, так и выходных показателей.

Для ликвидации этого недостатка в работе предлагается строить структурно-линейную модель так, чтобы классификация производилась по одному набору показателей, а аппроксимация в каждом классе — по-другому [5]. В соответствии с этим будем считать, что кроме пространства входных показателей X есть ещё пространство $Z = \mathbb{R}^s$, в котором производится классификация объектов. Часть показателей в пространствах X и Z могут совпадать. Соответственно, считается, что каждый из n объектов исходной выборки описывается $k + s + 1$ параметром, т. е. вектором $(y_t, x_t, z_t) \in \mathbb{R}^{k+s+1}$. Задача состоит в построении структурно-линейной модели, для которой классификация в Z имела бы простой вид. При решении этой задачи ограничимся множеством эталонных классификаций в Z [4]. Тогда задача ставится следующим образом: необходимо минимизировать функционал (2), (3) или (4) при условии, что классификация $H(x)$ является эталонной в пространстве Z.

Если в эталонной классификации $\varphi(h)$ равно $\varphi_2(h)$ или $\varphi_3(h)$, то соответствующие функционалы дифференцируемы по своим свободным параметрам, и для нахождения их локальных экстремумов можно использовать стандартные градиентные процедуры.

Одним из недостатков такого типа алгоритмов является сильная зависимость результата от начальных условий. В этом случае необходимо либо использовать специальные процедуры выбора в определённом смысле «хороших» начальных условий [6], либо разрабатывать методы глобальной оптимизации. В работе разработаны алгоритмы глобальной оптимизации для случаев конечного множества эталонов и для одномерного классификационного пространства.

Литература

[1] *Дорофеюк А. А., Касавин А. Д., Торговицкий И. Ш.* Применение методов автоматической классификации для построения статической модели объекта // Автоматика и телемеханика. — 1970. — № 2.

[2] *Дорофеюк А. А., Торговицкий И. Ш.* Применение методов автоматической классификации данных в задаче контроля качества изделий // Стандарты и качество. — 1967. — № 4.

[3] *Райбман Н. С., Дорофеюк А. А., Касавин А. Д.* Идентификация технологических объектов методами кусочной аппроксимации. — М.: ИПУ, 1977. — 70 с.

[4] *Бауман Е. В., Дорофеюк А. А.* Классификационный анализ данных // Избранные труды Международной конференции по проблемам управления. Том 1. — Москва: СИНТЕГ, 1999.

[5] *Бауман Е. В., Дорофеюк А. А., Корнилов Г. В.* Алгоритмы оптимальной кусочно-линейной аппроксимации сложных зависимостей // Автоматика и телемеханика. — 2004. — № 10.— С. 163–171.

[6] *Дорофеюк Ю. А.* Комплексный алгоритм автоматической классификации и его использование в задачах анализа и принятия решений // Таврический вестник информатики и математики. — 2008. — № 1. — С. 171–177.

Метод структурного прогнозирования на базе адаптивного алгоритма кластер-анализа*

Дорофеюк Ю. А.
dorofeyuk_julia@mail.ru
Москва, Институт проблем управления им. В. А. Трапезникова РАН

Рассмотрен метод адаптивного структурного прогнозирования, позволяющий отслеживать существенные изменения структуры исследуемого множества объектов, например, числа классов. Такая адаптация достигается за счёт специально разработанного для этой цели адаптивного алгоритма кластер-анализа, оптимального по числу классов.

В работе [1] рассмотрена структурно-классификационная методология прогнозирования, используемая в задачах выработки управленческих решений для крупномасштабных систем управления. Основная идея структурно-классификационного подхода к прогнозированию состоит в том, что для каждого объекта требуется прогнозировать не точные значения параметров в последующие моменты времени, а лишь класс, к которому он будет принадлежать в эти моменты времени в рамках некоторой структуры множества объектов изучаемой системы. В [1] для выявления такой структуры использовался комплексный алгоритм кластер-анализа (автоматической классификации) [2].

Подобная схема структуризации хорошо работает в условиях стационарного функционирования исследуемой системы управления, то есть когда структура объектов в пространстве параметров X меняется незначительно. В условиях же существенной динамики структуры исследуемых объектов (например, когда изменяется число классов такой структуры) необходимо разрабатывать более адекватные схемы структуризации и прогнозирования.

Методы адаптивной структуризации

Пусть исследуемая система состоит из n объектов, каждый из которых характеризуется набором из k параметров. Вводится в рассмотрение k-мерное пространство параметров X, в котором каждый объект представляется точкой

$$x_j = \left(x_j^{(1)}, \ldots, x_j^{(k)}\right), \quad j = 1, \ldots, n.$$

Предполагается, что вектор значений параметров x_j достаточно полно характеризует состояние j-го объекта на момент сбора информации, то есть взаиморасположение множества точек $x_1, \ldots, x_n$ в пространстве X отражает реальную структуру исследуемых объектов на этот момент времени. Как уже говорилось выше, в рамках классификационного подхода к анализу данных [3] для выявления такой структуры используется комплекс алгоритмов структурно-классификационого анализа, специально разработанный для решения таких задач [2]. Комплекс включает алгоритмы: m-локальной оптимизации заданного критерия J, выбора информативных параметров, выбора начального разбиения, выбора «оптимального» числа классов. Так как часть алгоритмов комплекса используется и в адаптивном варианте структурно-классификационого анализа, приведём краткое описание некоторых из них.

Алгоритм m-локальной оптимизации. Вначале опишем работу алгоритма 1-локальной оптимизации (для простоты рассматривается случай двух классов $r = 2$). Пусть задано начальное разбиение R_0 всех точек классифицируемой выборки $x_1, \ldots, x_n$. Обозначим через $x_j \in A_1$ и $x_j \in A_2$ точки, относящиеся к первому и второму классам соответственно. Алгоритм итерационный, на каждом шаге рассматривается одна точка из «зацикленной» классифицируемой последовательности $x_1, \ldots, x_n, x_1, \ldots, x_n, \ldots$. Отнесение точки к одному из двух классов обозначается с помощью индекса $\rho(x_j)$, который равен 1, если x_j принадлежит первому классу, и -1, если x_j принадлежит второму классу. Тогда алгоритм 1-локальной оптимизации определяется следующим образом:

$$\rho(x_j) = \operatorname{sign}[J(x_j \in A_1) - J(x_j \in A_2)].$$

Алгоритм заканчивает работу, если на некотором цикле не будет сделано ни одной «переброски» точки из класса в класс.

Алгоритм m-локальной оптимизации — это поэтапное применение к исследуемой выборке алгоритма s-локальной оптимизации, где $s = 1, \ldots, m$. На s-ом этапе алгоритм работает по той же схеме, только на каждом его шаге происходит пробная «переброска» из класса в класс не одной, а s точек. Подсчитывается значение критерия J до и после «переброски». Принадлежность каждой из s точек к классу либо остаётся неизменной (J до «переброски» больше, чем после), либо меняется на другой класс — в противном случае. В данном случае цикл — это число шагов, равное числу всевозможных различных наборов, в каждый из которых входит s точек, выбранных из n точек исходной выборки. Доказана сходимость алгоритма за конечное число шагов к локальному максимуму критерия J.

Работа выполнена при частичной финансовой поддержке РФФИ, проект № 10-07-00210-а.

Разработан эвристический алгоритм сокращённого перебора, который на каждом шаге для пробной «переброски» использует s точек в определённом смысле ближайших к границе между классами.

Алгоритм выбора числа классов. Для выбора числа классов используется специальная экспертно-компьютерная процедура, принцип работы которой состоит в следующем [2]. Сначала эксперт-пользователь оценивает диапазон $(r_{\min}, r_{\max})$, в пределах которого заведомо находится искомое число классов. Далее, используя алгоритм m-локальной оптимизации, проводится разбиение анализируемого множества объектов на $r_{\min}, \ldots, r_{\max}$ классов. Качество каждой из полученных классификаций оценивалось с помощью критерия:

$$J_3(r) = J_1(r) - qJ_2(r), \qquad (1)$$

где J_1 — величина средней по классам меры близости точек в классе, а J_2 — величина средней меры близости между классами. Величина q в (1) является масштабирующим параметром, приводящим к одному масштабу средние значения функционалов J_1 и J_2.

В качестве «оптимального» можно выбрать такое число классов r_{opt}, которое соответствует максимальному значению критерия (1) для $r_j = r_{\min}, \ldots, r_{\max}$. Однако наличие существенной, но неиспользованной при классификации информации (например, ввиду отсутствия данных) может привести к тому, что полученное таким способом r_{opt} не будет «истинно оптимальным». Для компенсации этого недостатка используется процедура экспертной коррекции [2]. При классификации многомерных объектов такая процедура (в случае необходимости) используется также для коррекции классификации точек, расположенных вблизи границы между соответствующими классами.

При моделировании и в приложениях критерий J_1 определялся выражением

$$J_1 = \sum_{i=1}^{r} \frac{n_i}{n} K(A_i, A_i);$$

$$K(A_i, A_i) = \frac{2}{n_i(n_i - 1)} \sum_{i=1}^{n_i} \sum_{j>1} K(x_i, x_j),$$

где n_i — число точек в классе A_i, а $K(x_i, x_j)$ — потенциальная функция (мера близости точек x_i и y_j) использовалась в виде [2]:

$$K(x_i, x_j) = \frac{1}{1 + \alpha R^p(x_i, y_j)},$$

где α и p — настраиваемые параметры алгоритма. Критерий J_2 определялся выражением:

$$J_2 = \frac{1}{r-1} \sum_{i=1}^{r} \sum_{j>1} \frac{n_i + n_j}{n} K(A_i, A_j),$$

где $K(A_i, A_j)$ — мера близости классов A_i и A_j, которое определяется выражением

$$K(A_i, A_j) = \frac{1}{n_i n_j} \sum_{x_l \in A_i} \sum_{x_p \in A_j} K(x_l, x_p).$$

На практике величина q выбирается в диапазоне значений 2–7 (обычно во столько раз отличается средняя близость внутри классов от средней близости между самими классами). Более подробно процедуры выбора значений свободных параметров рассмотрены в [4].

Адаптивный алгоритм структуризации объектов

Описанный выше алгоритм выбора «оптимального» числа классов хорошо работает в стабильных условиях развития исследуемой системы. Для отслеживания изменений структуры объектов в условиях существенной динамики предлагается следующая схема.

В момент времени t_1 с помощью описанного выше алгоритма m-локальной оптимизации [2] производится структуризация n точек в пространстве X на r классов, каждый из которых и характеризует определённый тип объекта. Число классов $r = r_{\text{opt}}(t_1)$ определяется с помощью описанного в предыдущем подразделе алгоритма выбора «оптимального» числа классов. Вводится понятие модели (эталона) класса $a_i(t)$, $i = 1, \ldots, r$ (чаще всего — это центр класса) [3]. Для каждого объекта кроме принадлежности к классу вычисляются расстояния до эталонов всех классов $R_{ij}(t)$, $i = 1, \ldots, r$, $j = 1, \ldots, n$. Для отслеживания существенных изменений структуры исследуемой системы объектов во времени все результаты обработки данных на момент времени t_1, включая результаты классификации на $r \in (r_{\min}, \ldots, r_{\max})$ классов и расстояния до эталонов $R_{ij}(t_1)$, сохраняются.

Заметим, что на практике структуризация объектов чрезвычайно редко проводится в пространстве исходных признаков, обычно сначала производится выделения набора информативных параметров. В работе для этой цели используются алгоритмы экстремальной группировки параметров, входящие в комплексный алгоритм [2].

Данные об исследуемых объектах, собранные в момент времени t_2, распределяются по классам каждой классификации на r_l классов, $r_l \in (r_{\min}, r_{\max})$. Для этой цели используются алгоритмы распознавания образов с учителем метода потенциальных функций [5, 6]. После этого подсчитываются значения критерия $J_3(r_l, t_2)$ (1) и в качестве оптимального (на момент времени t_2) выбирается такое число классов $r_{\text{opt}}(t_2)$, которое соответствует максимальному значению этого критерия. В случае необходимости используется экспертная процедура коррекции $r_{\text{opt}}(t_2)$.

После того, как определена принадлежность всех точек к тому или иному классу в пределах каждой классификации $r_l \in (r_{\min}, r_{\max})$, производится пересчёт эталонов $a_{il}(t_2)$, $i = 1, \ldots, r_l$, $r_l \in (r_{\min}, r_{\max})$. Для каждой точки с предыдущего шага пересчитываются, а для каждой новой точки вычисляются расстояния до новых эталонов $R\big(a_i(t_2), x_j(t_2)\big)$, $i = 1, \ldots, r_l$, $j = 1, \ldots, n$, $r_l \in (r_{\min}, r_{\max})$. Такая процедура выполняется для всех m моментов времени в пределах диапазона T стационарности, выбираемого с помощью специальной экспертной процедуры [7]. В первый же момент времени вне этого диапазона система «перезагружается» — все предыдущие результаты отправляются в архив, структуризация проводится заново как для момента времени t_1. Разработан вариант итерационного адаптивного алгоритма, когда структуризация объектов производится для данных, собранных в моменты времени, находящихся в пределах скользящего окна ширины T.

Адаптивный алгоритм прогнозирования

Как и для обычной (не адаптивной) схемы структурного прогнозирования [1] в качестве прогнозной модели для каждого объекта используется марковская цепь с r состояниями (r — число классов), то есть на каждом шаге рассчитываются элементы матрицы переходных вероятностей $\boldsymbol{P} = \|p_{ji}\|$, $j = 1, \ldots, n$, $i = 1, \ldots, r$. Разработан алгоритм пересчёта на каждом шаге соответствующих переходных вероятностей с использованием информации о значениях расстояний до центров классов, а также условия нормировки $\sum_{i=1}^{r} p_{ji} = 1$ для всех $j = 1, \ldots, n$ [1].

Адаптивный вариант алгоритма отличается от обычного только тем, что на каждом шаге рассчитывается не одна, а v матриц P_l, где $v = r_{\max} - r_{\min}$, а r_l — число классов l-ой классификации, $r_l \in (r_{\min}, r_{\max})$. Пусть после первого шага для всех классификаций на r_l классов подсчитаны расстояния

$$R_{lji}^{(1)} = R\big(x_j(t_1), a_{li}(t_1)\big),$$

$i = 1, \ldots, r_l$, $j = 1, \ldots, n$, $r_l \in (r_{\min}, r_{\max})$ от точек $x_j(t_1)$ до эталонов $a_{il}(t_1)$. Индекс l в этой и последующих формулах обозначает, что соответствующая величина рассчитана для классификации на r_l классов. Тогда элементы матрицы переходных вероятностей $p_{lji}^{(1)} = p_{lji}(t_1)$ рассчитываются следующим образом:

$$p_{lji}^{(1)} = \frac{\alpha_{lj}^{(1)}}{R_{lji}^{(1)}}, \quad i = 1, \ldots, r_l, \quad j = 1, \ldots, n, \qquad (2)$$

где нормирующий множитель $\alpha_{lj}^{(1)}$ определяется выражением:

$$\alpha_{lj}^{(1)} = \frac{\prod_{i=1}^{r_l} R_{lji}^{(1)}}{\sum_{i=1}^{r_l} \frac{1}{R_{lji}^{(1)}} \prod_{i=1}^{r_l} R_{lji}^{(1)}}. \qquad (3)$$

На s-ом шаге для каждой классификации $r_l \in (r_{\min}, r_{\max})$ элементы матрицы переходных вероятностей (2) модифицируется при помощи следующей процедуры. Введём обозначения:

$$\Delta R_{lji}^{(s)} = R_{lji}^{(s-1)} - R_{lji}^{(s)}; \quad \Delta \widehat{R}_{lji}^{(s)} = \frac{R_{lji}^{(s-1)} - R_{lji}^{(s)}}{R_{lji}^{(s-1)} + R_{lji}^{(s)}}.$$

Если j-ая точка совпадает с эталоном i_0-го класса ($x_j(t_s) = a_{li_0}(t_s)$), то есть $R_{lji_0}^{(s)} = 0$, тогда $p_{lji}^{(s)}$ равно 1, если $i = i_0$ и 0 в противном случае.

Для случая, когда $R_{lji_0}^{(s)} \neq 0$, происходит модификация всех переходных вероятностей по другой схеме:

$$p_{lji}^{(s)} = \gamma\Big[p_{lji}^{(s-1)} + \big(S_1 - p_{lji}^{(s-1)} S_2\big)\Delta \widehat{R}_{lji}^{(s)}\Big], \qquad (4)$$

где $S_1 = \frac{1+\mathrm{sign}(\Delta R_{lji}^{(s)})}{2}$, $S_2 = \mathrm{sign}(\Delta R_{lji}^{(s)})$, а γ — нормирующий множитель аналогичный (3), который определяется условием нормировки переходных вероятностей $\sum_{i=1}^{r_l} p_{lji}^{(s)} = 1$ и имеет следующий вид:

$$\gamma = \frac{1}{1 + (S_1 - p_{lji}^{(s-1)} S_2)\Delta R_{lji}^{(s)}}. \qquad (5)$$

Введение в (4) и (5) величины $\mathrm{sign}(\Delta R_{lji}^{(s)})$ вызвано необходимостью производить различными способами модификацию переходных вероятностей для случаев увеличения и уменьшения расстояния от точки $x_j(t_s)$ до эталонов классов $a_{li}(t_s)$ на s-ом шаге. А именно: в случае уменьшения величины $R_{lji}^{(s)}$ по отношению к $R_{lji}^{(s-1)}$ (то есть $\Delta R_{lji}^{(s)} < 0$) изменение соответствующей переходной вероятности происходит за счёт её увеличения на некоторую долю от $(1 - p_{lji}^{(s-1)})$; а в случае увеличения величины $R_{lji}^{(s)}$ по отношению к $R_{lji}^{(s-1)}$ (то есть $\Delta R_{lji}^{(s)} > 0$) изменение соответствующей переходной вероятности происходит за счёт её уменьшения на некоторую долю от $p_{lji}^{(s-1)}$. Это необходимо для выполнения условий нормировки для переходных вероятностей $0 < p_{lji}^{(s)} < 1$, $i = 1, \ldots, r_l$.

Построенные при помощи описанного выше алгоритма матрицы переходных вероятностей $\boldsymbol{P}_l$ используются для прогнозирования принадлежности объекта тому или иному классу. При этом на каждом шаге используется только одна матрица переходных вероятностей, соответствующая классификации на «оптимальное» для данного шага число

классов, то есть на s-ом шаге для прогнозирования используется матрица $\boldsymbol{P}_{l(s)}$, где $r_{l(s)} = r_{\text{opt}}(t_s)$.

На практике обычно используется не рандомизированная, а байесовская схема, когда объект относится к тому классу i_0, для которого $p_{ji_0} = \max\limits_{i=1,\ldots,r_{\text{opt}}(t_s)} p_{ji}$. В случае равенства переходных вероятностей p_{ji} для прогнозируемого объекта для двух или нескольких классов, он относится к классу с наименьшим номером.

Выводы

Разработанная методология использовалась при анализе и совершенствовании процедур принятия решений для нескольких крупномасштабных социально-экономических систем управления, в основном регионального характера; в том числе — региональная система управления здравоохранением, пассажирскими автоперевозками, система анализа и прогнозирования социально-экономического развития субъектов РФ и др. Во всех приложениях, а также при машинном моделировании была подтверждена высокая эффективность разработанной методологии структурно-классификационного анализа и прогнозирования.

Оказалось, что для некоторых приложений (с достаточно высоким уровнем помех при измерении параметров) существенно более эффективным оказывается использование алгоритмов размытой классификации, в том числе с фоновым классом [3].

Литература

[1] *Дорофеюк Ю. А.* Структурно-классификационные методы анализа и прогнозирования в системах управления // Таврический вестник информатики и математики. — 2008. — № 1. — С. 166–170.

[2] *Дорофеюк Ю. А.* Комплексный алгоритм автоматической классификации и его использование в задачах анализа и принятия решений // Таврический вестник информатики и математики. — 2008. — № 1. — С. 171–177.

[3] *Бауман Е. В., Дорофеюк А. А.* Классификационный анализ данных // «Избранные труды Международной конференции по проблемам управления. Том 1», Москва: СИНТЕГ, 1999.

[4] *Дорофеюк Ю. А.* Моделирование и анализ эффективности комплексного алгоритма классификационного анализа сложно организованных данных // Управление развитием крупномасштабных систем (MLSD'2009): Труды Третьей международной конференции, М.: ИПУ РАН, 2009. — С. 299–308.

[5] *Айзерман М. А., Браверман Э. М., Розоноэр Л. И.* Метод потенциальных функций в теории обучения машин. — М.: Наука, 1970.

[6] *Браверман Э. М., Мучник И. Б.* Структурные методы обработки эмпирических данных. — М.: Наука, 1983.

[7] *Дорофеюк А. А., Покровская И. В., Черняв-ский А. Л.* Экспертные методы анализа и совершенствования систем управления // Автоматика и телемеханика. — 2004. — № 10.— С. 172–188.

Структурно-классификационный алгоритм коррекции квазистационарных временных рядов в задачах статистического и социально-экономического мониторинга*

Чернявский А. Л., Дорофеюк А. А., Дорофеюк Ю. А., Лайкам К. Э.
achern@ipu.ru
Москва, Институт проблем управления им. В. А. Трапезникова РАН
Москва, Федеральная служба государственной статистики РФ

В работе используются новые методы анализа временных рядов, значения которых в каждый момент времени получены на выборках малого объёма. Эти методы базируются на современной методологии интеллектуального анализа сложно организованных данных, в том числе на методах структурно-классификационного анализа и прогнозирования.

В настоящее время для многих крупномасштабных объектов управления исходные для анализа данные задаются в виде квазистационарных временных рядов, — это и данные о колебаниях цен на товарных рынках, биржевые котировки акций, данные о пассажиропотоках, изменяющихся во времени и др. Временные ряды характеристик солнечной активности и космического излучения широко используются для прогноза погоды и неблагоприятного влияния на здоровье метеозависимых людей. Но наибольший интерес и социальную значимость имеют исследования временных рядов основных социально-экономических характеристик объектов управления в народном хозяйстве (региональные объекты, отраслевые, экологические, социальные и пр.). Так, например, при исследовании социального развития регионов (субъектов РФ — далее СРФ) [1] необходимо было проанализировать динамические ряды таких важнейших параметров, как среднедушевой доход, доля оплаты труда в среднедушевом доходе, превышение доходов над расходами, число пенсионеров на 1000 чел. населения, уровень безработицы, общий объем финансовой помощи на душу населения. Основным поставщиком такого рода данных являются органы государственной статистики, объединённые в Федеральную службу государственной статистики России (Росстат). Причём ключевым инструментом наблюдения за состоянием экономики страны, анализа, моделирования и прогнозирования экономического развития является статистика макроэкономических показателей с месячной и квартальной периодичностью.

Основная проблема, с которой сталкиваются службы статистики и социально-экономической аналитики развитых стран — это проблема фильтрации и коррекции собираемых статистических данных, в первую очередь — сезонное и календарное сглаживание исходных временных рядов. Дело в том, что почти все временные ряды используемых социально-экономических показателей подвержены влиянию сезонных и календарных факторов, маскирующих краткосрочные и долгосрочные тенденции развития и изменения социально-экономических показателей и препятствующих чёткому пониманию происходящих экономических явлений. Сезонное сглаживание (или сезонная корректировка) заключается в оценке и вычленении из исходных рядов сезонных и календарных факторов.

Наиболее широко используемыми в настоящее время и детально разработанными являются модели типа *ARIMA* [2]. Так, например, процедура сезонного сглаживания *X12-ARIMA* (разработчик Бюро Цензов США) наряду с процедурой того же типа *TRAMO-SEATS* (разработчик Банк Испании) рекомендованы ОЭСР и Евростатом в качестве стандартных методов сезонного сглаживания и применяются на практике многими национальными статистическими органами. По заказу Евростата было также разработано специальное программное обеспечение *DEMETRA*, реализующее оба этих метода, а также процедуры автоматического выбора моделей и контроля качества результатов сезонного сглаживания. Однако эти методы можно использовать только для временных рядов, полученных на репрезентативных выборках.

Начиная с 2009 г. в России стала собираться на регулярной основе помесячная статистика (с 1999 по 2008 гг. имеется только поквартальная статистика) многих важнейших показателей (в том числе таких актуальных в кризисный и посткризисный периоды показателей, как безработица и занятость). Однако из-за недостаточного финансирования этих работ Росстат не может обеспечить необходимую достоверность помесячных данных по этим показателям в разрезе большинства субъектов Российской Федерации. Попытки повышения достоверности за счёт скользящего 2–3 месячного сглаживания оказались неприемлемыми, в основном из-за существенной сезонной динамики этих показателей.

В настоящей работе используются новые методы анализа временных рядов, значения которых в каждый момент времени получены на вы-

Работа выполнена при частичной финансовой поддержке РФФИ, проекты № 08-07-00349, 08-07-00427, 10-07-00210.

борках малого объёма. Эти методы базируются на современной методологии интеллектуального анализа сложно организованных данных, в том числе на методах структурно-классификационного анализа [3, 4] и прогнозирования [5]. Особенность этой методологии состоит в том, что за счёт привлечения дополнительной информации, получаемой на основе структурно-классификационного анализа как СРФ по основным социально-экономическим показателям, так и самих временных рядов (траекторий) изучаемых показателей, а также использования современных процедур структурно-классификационного моделирования динамических объектов, можно ожидать значимого повышения достоверности скорректированных помесячных значений изучаемых показателей для СРФ.

Постановка задачи

Пусть задан набор n исследуемых объектов (в нашем случае это набор СРФ). Каждый объект x_j характеризуется набором k параметров, то есть каждый объект может быть представлен в виде точки в k-мерном пространстве параметров X. Если рассматривать изменение положения точки x_j во времени, то есть последовательность $x_j(t_l)$, $l = 1, \ldots, N$, то получим k-мерный временной ряд длины N, или траекторию изменения во времени положения точки $x_j(t)$ в пространстве X. Пусть параметр $x(j)$ — это параметр, временные ряды которого являются предметом исследования (для определённости далее будем считать, что это — уровень безработицы).

Очевидно, что на динамику уровня безработицы в каждом конкретном регионе оказывают влияние огромное число факторов: работа властных структур региона по созданию новых рабочих мест, развитие и поддержка малого и среднего бизнеса, обучение и переобучение за счёт бюджета работников обанкротившихся и депрессивных предприятий и организаций, стимулирование работодателей на переориентацию неконкурентного бизнеса на новые направления (в том числе на общественно полезные работы, финансируемые из бюджета) и многие другие. Однако существуют разные типы (классы) регионов, в каждом из которых используются различные факторы и разные механизмы, реализующие такое влияние, — это определяется как уровнем компетентности и эффективности работы руководства соответствующего СРФ, так и фактическим уровнем его экономического и социального развития. Таким образом, если удастся выявить такую типологию (классификацию) объектов (в нашем случае — субъектов РФ), то появляется возможность анализа в каждом классе корреляционных зависимостей между параметром $x(j)$ (уровень безработицы) и другими макроэкономическими показателями, полученными на основе более достоверных выборок. Для выявления такой типологии в работе используется комплексный алгоритм автоматической классификации [6].

Кроме того, в пределах каждого класса можно проводить дальнейший классификационный анализ уже самих временных рядов параметра $x(j)$, то есть выявлять типологию моделей исследуемых временных рядов. Другими словами, для объектов каждого класса регионов (классификация первого уровня) строится классификация траекторий этого параметра (классификация второго уровня). Классификация второго уровня позволяет получить два типа результатов.

Первый тип — выделение СРФ, имеющих близкие (в статистическом смысле) траектории параметра $x(j)$. Это позволяет провести взвешенное объединение выборок для таких СРФ, что значимо повышает достоверность оценок значений соответствующих временных рядов.

Второй тип результатов связан с построением описаний (в том числе качественных) эталонов классов второго уровня (эталонных траекторий), что позволяет целенаправленно выбирать адекватные модели описания соответствующих временных рядов для каждого класса этого уровня. Для этой цели в работе используются методы структурно-классификационного анализа динамических объектов [4]. Следует подчеркнуть, что выбор адекватной модели временного ряда также приводит к значимому повышению достоверности оценок значений ряда. Это происходит за счёт того, что при оценке значения параметра $x(j)$ в каждый момент времени используется не только выборка данных на этот момент времени (как это делается сейчас), а выборки за все предыдущие моменты. Всё зависит от числа оцениваемых параметров модели. Так, например, если модель ряда в пределах исследуемого временного периода (например, года) является линейным трендом, то для оценки двух неизвестных значений параметров такой модели будут использоваться 12 выборочных оценок параметра $x(j)$ (при помесячном мониторинге), что увеличивает достоверность корректирующей оценки в 6 раз. Реальные модели обычно содержат большее число оцениваемых параметров, но даже для достаточно сложных моделей сезонного сглаживания оно не превышает 5–6, что даёт двукратное увеличение достоверности корректирующей оценки.

Схема структурно-классификационной коррекции временных рядов

Для решения задачи структурно-классификационной коррекции временных рядов на малых выборках была разработана специальная методология, которая содержит следующие основные этапы.

1. Выявление структуры ансамбля изучаемых объектов по набору показателей, являющихся адекватными характеристиками их функционирования. Для СРФ это означает их классификацию по основным информативным характеристикам социально-экономического развития (включая занятость и безработицу). Информативные характеристики выбираются из исходного множества показателей с помощью алгоритмов экстремальной группировки параметров [7]. С целью учёта динамики значений показателей в пределах классифицируемого периода классификация проводится по схеме двухуровнего структурного анализа [7]. На первом уровне производится классификация n_1 объектов в k_1-мерном пространстве показателей. Для этой цели используется комплексный алгоритм автоматической классификации [6]. Объектом классификации на первом этапе является «СРФ–квартал», при этом если используются только квартальные данные, то для каждого СРФ используются значения информативных показателей за каждый квартал для l лет имеющихся данных, таким образом для этого случая $n_1 = 4\,ln$, где n — число исследуемых СРФ, $k_1 = k_0$, где k_0 — число информативных показателей. Если же используются помесячные данные, то сначала производится их скользящее сглаживание с помощью 3-х позиционного окошка (псевдоквартальные данные), то есть значение каждого временного ряда для каждого месяца заменяется средним из трёх значений — его самого, соседнего слева (предыдущий месяц) и соседнего справа (последующий месяц). И, в отличие от поквартальных данных, каждый СРФ в каждом 3-х позиционном окошке («скользящем квартале») характеризуется тремя значениями каждого из информативных показателей. Таким образом для помесячных данных $n_1 = 12\,ln$, $k_1 = 3k_0$. Для вычисления усреднённых граничных («крайних») значений каждого временного ряда используются экстраполяционные процедуры, либо они не используются при классификации. В последнем случае число классифицируемых объектов равно $n_1 = (12\,l - 2)n$. Полученная в результате классификация для каждого СРФ порождает либо $4\,l$-позиционный код для квартальных данных, либо $12\,l$-позиционный код для сглаженных помесячных данных, в каждой позиции которого стоит номер класса, к которому был отнесён этот субъект для соответствующих значений информативных параметров.

На втором уровне на базе полученных кодов, каждая позиция которого трактуется как номинальный признак, производится окончательная классификация СРФ, учитывающая структурную динамику регионов в исследуемый временной период.

2. В каждом классе полученной на втором уровне классификации проводится классификация СРФ на небольшое число классов (2-4) как динамических объектов только по временным рядам исследуемого параметра (безработицы). Для этого используются помесячные данные, сглаженные с помощью 3-х позиционного окошка таким же образом, как и на предыдущем этапе. Для получения таких классификаций используется алгоритм структурно-классификационного анализа динамических объектов [4]. Для каждого класса множества полученных классификаций строятся эталонные траектории (временные ряды), которые в определённом смысле являются характеристиками «центров» таких классов. Кроме того, строится качественное описание как эталонных траекторий, так и траекторий СРФ, входящих в соответствующий класс.

3. На третьем этапе строятся модели временных рядов для каждого класса той классификации, которая получена на втором этапе.

— Если для этой классификации найдутся СРФ, для которых траектории близки со статистической точки зрения, то для них помесячные выборки объединяются, для таких СРФ модели не строятся.
— Для каждого класса строится модель усреднённых помесячных значений безработицы, которая используется как начальные условия для адаптивной модели каждого СРФ из этого класса.

4. Эти модели используются для вычисления скорректированных помесячных значений для тех СРФ, для которых исходная выборка недостоверна.

5. В некоторых случаях требуется провести согласование (балансировку) откорректированных оценок — сохранение (в пределах заданных доверительных интервалов) значений оценок параметра $x(j)$ для Федеральных округов и Российской Федерации в целом, вычисленных в те же моменты времени. Для этой цели разработаны специальные процедуры согласования, позволяющие минимизировать отклонение невязок за счёт целенаправленного изменения корректирующих оценок для СРФ, имеющих наименьшую достоверность значений оценок параметра $x(j)$.

Литература

[1] *Дорофеюк Ю. А., Дорофеюк А. А., Покровская И. В.* Методология экспертно-классификационного анализа в задаче оценки развития крупномасштабных региональных систем // Управление развитием крупномасштабных систем (MLSD'2007): Труды первой международной конференции, М.: ИПУ РАН, 2007. — С. 132–139.

[2] *Дж. Бокс, Г. Дженкинс* Анализ временных рядов. // Прогноз и управление.— М.: Изд. «Мир», 1974. — 406 с.

[3] *Бауман Е. В., Дорофеюк А. А.* Классификационный анализ данных // «Избранные труды Международной конференции по проблемам управления. Том 1», Москва: СИНТЕГ, 1999.

[4] *Бауман Е. В., Дорофеюк А. А., Дорофеюк Ю. А.* Методы динамического структурного анализа многомерных объектов // Четвертая международная конференция по проблемам управления (МКПУ-IV): Сборник трудов, М: ИПУ РАН, 2009. — С. 338–343.

[5] *Дорофеюк Ю. А.* Структурно-классификационые методы анализа и прогнозирования в крупномасштабных системах управления // Проблемы управления. — 2008. — № 4. — С. 78–83.

[6] *Дорофеюк Ю. А.* Комплекс алгоритмов экспертно-классификационного анализа для решения прикладных задач // Четвертая международная конференция по проблемам управления (МКПУ-IV): Сборник трудов, М: ИПУ РАН, 2009. — С. 373–379.

[7] *Браверман Э. М., Мучник И. Б.* Структурные методы обработки эмпирических данных. — М.: Наука, 1983.

Модель распознавания и оценивания состояний сложного объекта

Колесникова С. И.[12]***, Мертвецов А. Н.***[2]
skolesnikova@yandex.ru
[1]Томск, Томский университет систем управления и радиоэлектроники, [2]Национальный исследовательский Томский политехнический университет

Формализована задача распознавания и оценивания состояний сложных объектов как задача интеллектуального анализа данных в стохастических временных рядах. Создан подход к построению модели корректной алгоритмической композиции для решения задачи и апробирован на прикладной задаче обнаружения предвестников разрушения горных пород. Дан метод скользящей аппроксимации стохастического временного ряда для оценивания состояний сложных объектов в реальном времени.

Постановки задач, связанные со сложными (плохо формализуемыми, с неполным аналитическим описанием) объектами, как правило, допускают неоднозначность решения, поэтому ставится вопрос разработки алгоритмов (их комбинации) и их теоретическом обосновании, корректно решающих поставленные задачи в определённом классе [1–3].

Современные исследования распознавания состояний сложных объектов связаны с автоматизацией огрублённого численного исследования динамических систем на основе методов распознавания образов и статистического моделирования [4], распознавания последовательности состояний сложного источника как чередования и наложения характерных последовательностей сигналов [5] и многими другими. Однако существует ряд нерешенных проблем, порождаемых особенностями нелинейных и нестационарных СДО и приводящий к большому проценту ошибок при обнаружении предвестников зарождающихся «опасных» состояний (дефектов, разрушений), связанный с тем, что разброс величин измеряемых параметров превышает изменения, характерные для появления контролируемых состояний.

В данной работе формализована общая задача распознавания и оценивания состояний СДО, для решения которой использована идеология проблемно-ориентированной технологии, основы которой заложены Ю. И. Журавлёвым и К. В. Рудаковым и далее развитой в работах их учеников.

Постановка задачи распознавания состояний СДО

Рассматривается класс сложных динамических объектов с описанием:

$$Z = \{X,\, Y\} = \{X(t),\, Y(t),\, t_0 \leqslant t \leqslant T\},$$
$$Y(t) = f(X(t)) + \xi(t),\; t \geqslant t_0, \qquad (1)$$

где Z — стохастический временной ряд, сопровождающий функционирование СДО и характеризующий состояние СДО; $Z(t)$ принимает значения в произвольном измеримом (фазовом) пространстве $(Z,\ \mathfrak{F}_Z)$, где $\mathfrak{F}_Z$ — σ-алгебра подмножеств пространства Z; $X(t),\ Y(t)$ — векторы ненаблюдаемых и наблюдаемых переменных состояния объекта, соответственно; $\xi(t)$ — неизвестный шум с ограниченной дисперсией. Относительно динамики поведения процесса (1) на $[t_0, T]$ выдвинуто $I > 0$ альтернативных гипотез $(\Omega_1, \ldots, \Omega_I)$, составляющих полную группу событий и интерпретируемых как классы состояний СДО. Предполагается, что с вероятностью 1 за конечный промежуток времени происходит конечное число изменений состояний СДО, и процесс (1) допускает представление:

$$Z(t) = \sum_{i=1}^{I} Z_i(t)\chi\left\{t \in [t^{(i-1)},\, t^{(i)})\right\},$$

где $\{t^{(t)}\}$ — возрастающая последовательность случайных моментов времени $t^{(i-1)} < t^{(i)},\ i = 1, \ldots, I$, $Z_i(t)$ — случайный элемент (при фиксированном t) из $(Z,\ \mathfrak{F}_Z)$, $\chi(A)$ — индикатор события . Наблюдение $Y(t)$ осуществляется в соответствии с дискретным планом:

$$t \in \{t_0, t_1, \ldots, t_n\},\; t_j = t_0 + j\Delta,\; \Delta > 0,\; j = 0, \ldots, n.$$

Требуется: 1) выяснить условия разрешимости и регулярности задачи; 2) построить решающее правило (удовлетворяющее локальным (зависящим от обучающей выборки) и универсальным ограничениям (не зависящим от прецедентов)), относящее наблюдаемый фрагмент реализации случайного процесса $Y(t)$ (1) к одному из образов состояний $\Omega_i,\ i = 1, \ldots, I$; 3) оценить качество решающего правила в смысле заданных внешних и внутренних критериев; 4) апробировать модель распознавания и оценивания состояний СДО для решения прикладных задач: обнаружение нежелательных состояний СДО (мониторинг СДО), оценивание состояний (реконструкция временного ряда), с целью управления и наблюдения неизвестных возмущений и параметров.

Решение задачи распознавания состояний СДО

Первый шаг к решению поставленной задачи целесообразно связать с задачей выделения нелинейного тренда стохастического временного ряда

(СВР), впервые поставленной для конечных плоских конфигураций как задачи классификации в работе [3]. Для стохастического ряда на базе [3] задачу формализуем следующим образом.

Пусть L — линейное нормированное пространство всевозможных числовых последовательностей; дан временной ряд $y_1, \ldots, y_n$, y_j, $j = 1, \ldots, n$:

$$y_j = x_j + \xi_j,\ j \geqslant 0, \tag{2}$$

где x_j — детерминированная составляющая, ξ_j — белый (измерительный) шум, $M\xi_j = 0$, $M\xi_j^2 = \sigma_\xi^2 < \infty$. Предполагается, что для iго фрагмента ряда (2) (интерпретируемого как состояние СДО) процесс описывается моделью:

$$y_j^{(i)} = x_j^{(i)} + \xi_j^{(i)},\ j \geqslant 0,\ i = 1, \ldots, I, \tag{3}$$

где $x_j^{(i)} = f_j^{(i)} = f^{(i)}(j\Delta), \Delta > 0, f^{(i)}(t) \in \mathbb{R}$ — неизвестная функция.

Задача выделения тренда определяется пятеркой зафиксированных параметров:

$$\Theta_{TS}(\Phi,\ M,\ \boldsymbol{\mu},\ LS_\mu,\ \sigma_\xi^2),$$

где $\Phi = \{f_k(t),\ k = 1, \ldots, n_f\}$ — множество функциональных зависимостей, метки (символы) которых составляют алфавит $M = l_0, \ldots, l_m$ разметки СВР; $\boldsymbol{\mu}$ — система аксиом (правил разметки); LS_μ — обучающая выборка. Ставится задача классификации: каждой точке ряда $\boldsymbol{y}$ должен быть сопоставлен символ из алфавита M (l_0=«не размечено», l_k — метка функции, $l_k \in \{1, \ldots, n_f\}$).

Определен стохастический вектор-объект:

$$\boldsymbol{Y}^d = ((t_1, y_1), \ldots, (t_d, y_d)),\ d \geqslant 1,\ t_j, y_j \in \mathbb{R},$$
$$\boldsymbol{Y}_j = (t_j, y_j) \in \mathbb{R}^2,\ t_j = j\Delta,\ \Delta > 0,\ t_1 < \cdots < t_d,$$

где y_j удовлетворяет (2).

Объекты $\boldsymbol{Y}^d$, $\boldsymbol{Y}'^d$ названы статистически эквивалентными, если переменные y_d, y'_d удовлетворяют описанию (2) и имеют равные детерминированные составляющие: $x_j(t) = x'_j(t)$, $j = 1, \ldots, d$.

Под состоянием СДО будет пониматься набор статистически эквивалентных вектор-объектов, порождаемых динамикой объекта.

Разметки l_1^d, l_2^d определены как эквивалентные ($l_1^d \approx l_2^d$), если число несовпадающих (неразмеченных) позиций меньше порогового значения.

Условие корректности алгоритма A_μ. Задача Θ_{TS} выделения трендов заключается в синтезе такого алгоритма A_μ, что для всех статистически эквивалентных вектор-объектов из любого поднабора $LS'_\mu \subseteq LS_\mu$ выходом алгоритма A_μ являются эквивалентные разметки.

Для обеспечения разрешимости некорректной задачи выделения тренда СВР решаются вопросы:

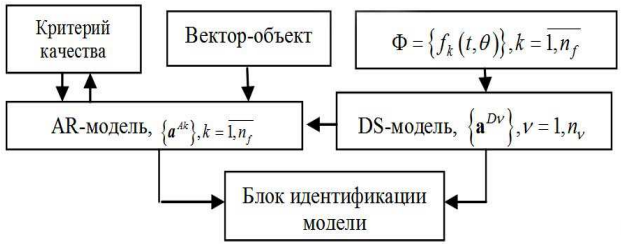


Рис. 1. Концептуальная схема разметки СВР в методе ARADS (AutoRegression, Adaptive algorithm, Difference Scheme).

вопрос локализации аксиом и алгоритмов разметки, поиска оптимальной системы окрестностей; вопрос регуляризации задачи разметки. Решение первого вопроса обеспечивается системой аксиом (правил разметки) $\boldsymbol{\mu} = \{\mu_u\}$, согласно которой каждой точке ряда $(t_j,\ y_j)$ (объекту) по её окрестности:

$$\boldsymbol{O}(Y_j,\ j-i+1) = ((t_i, y_i, \ldots, (t_j, y_j)),\ i \in [1,\ d],\ j \geqslant i,$$

сопоставляется номер функции (класса) $l_m \in M$. Основанием системы аксиом $\boldsymbol{\mu}$ является существование однозначного соответствия [6] между функциями с определёнными свойствами и значениями векторов разностных схем (рис. 1) и статистическая эквивалентность AR-объектов (авторегрессий) и DS-объектов.

Правила разметки СВР (ARADS)

Аксиома $\mu_u = \mu_u\left(\boldsymbol{O}(Y_j,\ d),\ l^d,\ \boldsymbol{a}^{Ak}\right)$ — бинарная функция, задаваемая на одной точке $Y_j = (t_j,\ y_j)$ с окрестностью:

$$\boldsymbol{O}(Y_j, d) = ((t_{d-j+1}, y_{d-j+1}), \ldots, (t_j, y_j)).$$

По правилу: точке (t_j, y_j) сопоставляется метка $l_j = f_u$, если в условиях модели (2) имеют место (4), (5):

$$\left(a_1^{Ak}, \ldots, a_{p_k}^{Ak}\right) = \underset{a^{Ai} \in \mathbb{R}^{p_k}}{\arg\min}\ Crt_\varphi\left(\boldsymbol{a}^{Ai}, j\right),$$
$$Crt_\varphi\left(\boldsymbol{a}^{Ai}, j\right) = \sum_{l \in j-d+1, \ldots, j} \varphi\left(\left| y_i - \sum_{o=1}^{p_k} a_o^{Ai} y_{l-o} \right|\right), \tag{4}$$

$$u = \begin{cases} \underset{k=1,\ldots,n_f}{\arg\min}\ \rho\left(\boldsymbol{a}^k, \boldsymbol{a}^A k\right), Crt_\varphi\left(\boldsymbol{a}^A k(p_k), j\right) < e_\varphi j; \\ l_0, Crt_\varphi\left(\boldsymbol{a}^A k(p_k), j\right) \geqslant e_\varphi j. \end{cases} \tag{5}$$

где φ — монотонно возрастающая на $(0, \infty)$ функция, $\varphi(0) = 0, \varphi(x) < 0, \forall x > 0$, $\sup\limits_{x \in \mathbb{R}^+} |x^2 \varphi(x)| < \infty$; $\rho\left(\boldsymbol{a}^k, \boldsymbol{a}^A k\right) = \|a^k - \boldsymbol{a}^A k\|$; $\boldsymbol{a}^A k = \boldsymbol{a}^A k[j, j+d-1]$, $\boldsymbol{a}^k$ — векторы AR- и DS-коэффициентов, сопоставленные отрезку $[t_j, t_j + d - 1]$ и функции $f_k(t)$

с дробно-рациональным Z-преобразованием соответствующих последовательностей $\{f_k(j\Delta)\}, j \geqslant 0$, соответственно.

Рассмотрено два варианта систем аксиом: $\boldsymbol{\mu}_1 = \{\mu_u^1\}$, $\boldsymbol{\mu}_2 = \{\mu_u^2\}$, где μ_u^1 — бинарная функция, задаваемая на $d = (i_2 - i_1 + 1)$ точках скользящего окна $[i_1,\ i_2]$ переменной длины с величиной сдвига, равной размеру предыдущего окна, $\forall j \in [i_1,\ i_2]$ объект (t_j, y_j) имеет метку $l_j = f_u$; μ_u^2 — функция, задаваемая на одной точке (t_j, y_j) по её окрестности $\boldsymbol{O}(Y_j, d)$.

Условия корректности системы аксиом $\boldsymbol{\mu}$ выражены в требованиях полноты: $\forall Y_j = (t_j, y_j)$, $\exists \mu' \in \boldsymbol{\mu}$, $l_i = m$ и однозначности $\forall Y_j$: $l_j = \mu', l_j = \mu'', l_j = \mu'' \Rightarrow \mu' = \mu''$ при фиксированной нетривиальной системе окрестностей $\{\boldsymbol{O}(Y_j, d_j,\ j \in [1, d])\}$.

Теорема 1. *Локальные системы аксиом* $\boldsymbol{\mu_1} = \{\mu_u^1\}$, $\boldsymbol{\mu_2} = \{\mu_u^2\}$ *на основе ARADS являются корректными.*

Для решения *вопроса регуляризации задачи разметки* наряду с алфавитом $M_\alpha = \{l_1^\alpha, \ldots, l_m^\alpha\}$ вводится алфавит M_β как основа регуляризирующей системы аксиом с приоритетами $\boldsymbol{\mu^\beta}$: аксиома l_B^b выполнима только тогда, когда не выполнимы аксиомы $\{l_j^\alpha, j \notin B\}$: $M_\beta = \{l_B^\beta | B \in \in (B_1, \ldots, B_I)\}$, $B_i \subseteq \{1, \ldots, m\}, i = 1, \ldots, I$. Соответствующие алфавитам $M_\alpha \subseteq M_\beta$ системы аксиом μ^α, μ^β и разметки названы $\alpha(\beta)$-системой и $\alpha(\beta)$-разметкой. Обоснование способа регуляризации опирается на результаты теории информации [7, 8]. Вводятся следующие условия.

Условие статистической согласованности $\alpha(\beta)$-разметки статистически согласованы при выполнении условий:

$$\mathsf{P}(l_j^\alpha, l_B^\beta) = \mathsf{P}(l_B^\beta)\,\mathsf{P}(l_j^\alpha / l_B^\beta), j \in B;$$
$$\mathsf{P}(l_j^\alpha) = \sum_{B \in (B_1, \ldots, B_I)} \mathsf{P}(l_B^\beta)\,\mathsf{P}(l_j^\alpha / l_B^\beta), \forall j = 1, \ldots, m. \tag{6}$$

Условие регуляризации. Алгоритм β-разметки при фиксированных алфавитах M_α, M_β назван корректным асимптотическим регуляризатором, если выполнены условия:

$$\mathsf{P}(l_B^\beta / l_j^\alpha) = 0, j \notin B; \mathsf{P}(l_B^\beta / l_j^\alpha) > 0, j \in B;$$
$$\mathsf{P}(l_B^\beta / l_j^\alpha) > 0, \text{ для } B = \{j\}. \tag{7}$$

Задача разметки Θ_{TS} названа β-разрешимой, если существует корректный (безошибочный на прецедентах) алгоритм β-разметки.

Поставлена задача оценивания распределения $\mathsf{P}_\alpha = (p_{\alpha 1}, \ldots, p_{\alpha m})$ α-разметки $\boldsymbol{l}^{\alpha n}$ по распределению $\mathsf{P}_\beta = (p_{\beta 1}, \ldots, p_{\beta I})$ β-разметки $\boldsymbol{l}^{\beta n}$.

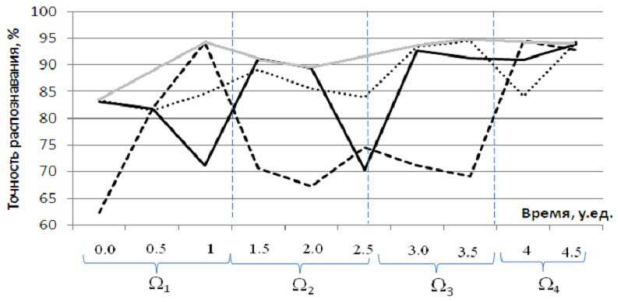


Рис. 2. Динамика усреднённой точности распознавания 4-х состояний ЭМО тремя базовыми алгоритмами $A_1(\varphi, \Phi, \rho)$ при фиксированных φ, ρ.

Условие асимптотической регулярности. β-разрешимая задача разметки Θ_{TS} является асимптотически регулярной тогда и только тогда, когда для любого допустимого набора возможных решений, порождённого алгоритмом β-разметки, существует α-разметка, вероятностное распределение которой $\mathsf{P}_\alpha = (p_{\alpha 1}, \ldots, p_{\alpha m})$ является единственным, максимизирующим вероятность β-разметки.

Теорема 2. *(Критерий асимптотической регулярности) Задача разметки* Θ_{TS} *асимптотически регулярна тогда и только тогда, когда для неё существует корректный алгоритм* β*-разметки* $A_{\mu\beta}$*, удовлетворяющий условиям* (6), (7).

Сконструированы алгоритмические композиции (рис. 2) на базе семейства алгоритмов — отображений $\{A_\mu(\varphi, \Phi, \rho)\}$, параметризованных множеством функций Φ (основой алфавита M), типом функции φ в критерии качества AR-модели (4), нормой (5) линейного пространства ρ (рис. 2) с весовыми коэффициентами, найденными на основе нелинейной модификации [9] метода анализа иерархий АНР (Т. Саати).

Дан *метод скользящей реконструкции временного ряда* (на базе разметки), для решения задачи получения оценки «скрытой» переменной x_n по измерениям $y_n, n \geqslant 1$ вида (2). Метод реализован следующими положениями: 1) используется идеология скользящего окна, размер и величина сдвига которого зависят от положения на временной оси анализируемого вектор-объекта; 2) даётся правило выбора вида функциональной зависимости для каждого окна; 3) модели аппроксимирующих функций могут быть неравными на разных сегментах, а длина временного ряда не ограничена; 4) траектории сшиваются в скользящем режиме: в j-м окне строится новая траектория и подгоняется к построенной в $(j-1)$-м окне, которая полагается окончательной к моменту начала j-го окна («начальное условие» каждой следующей модели задано) (рис. 3).

Отметим, что метод реконструкции Бока [10] обладает особенностями: отсутствие правила выбора числа сегментов и стартовых догадок относи-

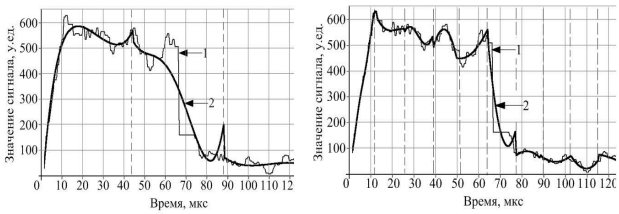


Рис. 3. Скользящая реконструкция реализации переменной состояния ЭМО: а) по трем отрезкам; б) по 10 отрезкам (1 — реальные измерения; 2 — кривая аппроксимации).

тельно начальных условий по каждой переменной; существенное ограничение на длину ряда; принадлежность модельных функций одному классу; аппроксимация требует знания данных всего ряда.

Решение прикладной задачи. Актуальны исследования по обнаружению предвестников разрушения горных пород по характеристикам их сигналов электромагнитной эмиссии (ЭМС), в основе которых лежит экспериментально доказанное явление увеличения электромагнитной активности на этапе предразрушения.

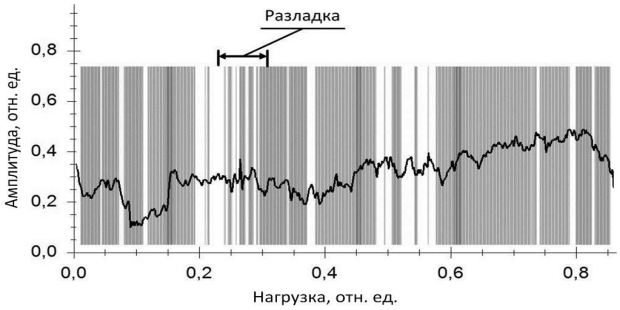


Рис. 4. Скользящая разметка СВР значений регистрируемой ЭМС в процессе одноосного сжатия образца горной породы (на рисунке отмечена зона разладки).

На рис. 4 — результат применения предложенного алгоритма разметки СВР (серии значений регистрируемой ЭМС) с целью анализа процессов деформирования лабораторного образца, в частности обнаружение момента разладки как предвестника разрушения образца.

Заключение

Рассмотрена модель распознавания и оценивания состояний сложных объектов на основе выявления закономерностей в стохастическом временном ряду, сопровождающем его функционирование. Апробация предложенного подхода и реализующих его метода и алгоритмов разметки осуществлена на решении актуальной прикладной задачи обнаружения предвестников разрушения горных пород по характеристикам сигналов электромагнитной эмиссии.

Целесообразно дальнейшее исследование данного подхода на предмет указания и уточнения границ его применимости в зависимости от вида шума и его числовых характеристик, а также правил выбора минимального числа признаков функциональных зависимостей, обладающих «хорошими» дискриминирующими свойствами, аналогичных гипотезе компактности.

Литература

[1] *Журавлев Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. 1979. — Т. 33.— С. 5–68.

[2] *Журавлев Ю. И., Рудаков К. В.* Об алгебраической коррекции процедур обработки (преобразования) информации // Проблемы прикладной математики и информатики. — 1987. — С. 187–198.

[3] *Рудаков К. В., Чехович Ю. В.* Алгебраический подход к проблеме синтеза обучаемых алгоритмов выделения трендов // Доклады РАН. — 2003. — Т. 388, № 1. — С. 33–36.

[4] *Неймарк Ю. И., Таранова Н. Н., Теклина Л. Г.* О возможностях изучения хаотических движений в конкретных динамических системах методами распознавания образов и математического моделирования // Математические методы распознавания образов (ММРО-14): доклады XIV Всеросс. конф. Сборник докладов, Москва: МАКС Пресс, 2009. — С. 422–425.

[5] *Грызлова Т. П.* Формализация задачи распознавания последовательности состояний сложного источника // Математические методы распознавания образов (ММРО-14): доклады XIV Всеросс. конф. Сборник докладов, Москва: МАКС Пресс, 2009. — С. 333–337.

[6] *Семенычев В. К.* Идентификация экономической динамики на основе моделей авторегрессии. — Самара: АНО «Изд. СНЦ РАН», 2004. — 243 с.

[7] *Галлагер Р.* Теория информации и надежная связь. — Москва: «Советское радио», 1974. — 720 с.

[8] *Шоломов Л. А.* Исследование одного класса динамических процедур коллективного выбора. — Нелинейная динамика и управление. Вып. 5. — Москва: Физматлит, 2007. — 400 с.

[9] *Колесникова С. И.* Свойства корректной модификации метода парных сравнений // Интеллектуальные системы. — 2010. — Т. 14, вып. 1–4. — С. 183–202.

[10] *Baake E., Baake M., Bock H. G., Briggs K. M.* Fitting ordinary differential equations to chaotic data // Phys. Rev. A. — 1992. — V. 45, № 8. — P. 5524–5529.

Алгоритм и автоматизированный метод построения алгоритмов распознавания участков фазовых траекторий

Коваленко Д. С., Щербинин В. В., Костенко В. А.
kovalenkods@gmail.com, victorshch@gmail.com, kost@cs.msu.su
Москва, МГУ им. Ломоносова

В данной работе рассматривается задача построения алгоритмов распознавания нештатного поведения систем по обучающей выборке, информация о поведении которых доступна с окружающих систему датчиков. Рассмотрены основные идеи построения алгоритмов распознавания, предложен алгоритм для автоматического построения распознавателей нештатного поведения и указаны подходы к улучшению качества его работы за счет вовлечения эксперта в промежуточные этапы работы алгоритма.

Задача распознавания нештатного поведения системы

Рассмотрим систему, которая окружена набором датчиков. Считаем, что датчики, окружающие систему, опрашиваются с одинаковой частотой. Фазовая траектория X представляет собой последовательные измерения всех датчиков системы: $X = (\bar{x}_1, \ldots, \bar{x}_n)$, где $\bar{x}_i = \bar{x}(t_0 + i \cdot \tau)$ — это точка в многомерном пространстве показаний датчиков; $1/\tau$ — частота опроса датчиков.

Все множество траекторий, которые могут быть получены с датчиков системы, назовем допустимыми траекториями и обозначим $V = \{X\}$.

Состояние системы характеризуется показаниями датчиков, окружающих ее. Со временем система может изменять свое состояние. Последовательные изменения состояния системы будем называть ее поведением. Поведение системы можно разделить на два типа:

- Нормальное поведение: система стабильно выполняет заложенные в нее функции.
- Нештатное поведение: система в скором времени гарантированно перестанет выполнять заложенные в нее функции.

Может существовать несколько классов нештатного поведения, приводящих систему к аварийному состоянию. Будем считать, что каждому классу нештатного поведения соответствует некоторая характерная траектория X_{Anom}, такие траектории будем называть эталонными.

Пусть число классов нештатного поведения системы равно L. Обозначим: $W = \{w\}_{w=1}^{w=L} \cup \{0\}$ — множество ответов, где 0 — соответствует нормальному поведению системы, w — нештатному поведению под номером w из L возможных.

Участки траекторий, соответствующие нештатному поведению, могут входить в анализируемую траекторию X в искаженном относительно эталонных траекторий виде. Искажения могут быть по амплитуде и времени. Под искажением траектории по амплитуде будем понимать изменение абсолютных значений точек траектории, без изменения числа отсчетов. Под искажением траектории по времени будем понимать изменение числа отсчетов, на которых определена траектория. Отдельно выделим подкласс искажений по амплитуде — стационарный шум, возникающий в датчиках, окружающих систему.

Искажения могут быть нелинейными, но траектории с искажениями, соответствующие различным классам нештатного поведения не должны пересекаться: $X^1_{\text{Anom}} \cap X^2_{\text{Anom}} = \varnothing$. Две траектории пересекаются, если одна из них целиком содержит другую, т.е. одна траектория содержит последовательность точек, равную всем точкам другой траектории, причем порядок следования совпадает.

Задача распознавания нештатного поведения заключается в следующем. Дано:

- Набор из L классов нештатного поведения системы и соответствующее множество эталонных траекторий $\{X^l_{\text{Anom}}\}_{l=1}^{L}$.
- Наблюдаемая многомерная траектория X.
- Ограничения на полноту и точность распознавания:
$$e_1 \leqslant \text{const}_1 \text{ и } e_2 \leqslant \text{const}_2,$$
где: e_1, e_2 – число ошибок распознавания первого и второго рода; const_1, const_2 – заданные числовые ограничения.
- Ограничение на время распознавания:
$$t(X \to \Psi) \leqslant \text{const}_3$$
где: $t(X \to \Psi)$ — время работы алгоритма распознавания; const_3 — заданное ограничение.

Требуется провести распознавание в наблюдаемой траектории участков, соответствующих эталонным траекториям, с учетом ограничений на время работы, полноту и точность распознавания.

Задача построения алгоритма распознавания по обучающей выборке

Определение 1. *Набор пар* $\Phi = \{(I^j, w^j)\}_{j=1}^{J}$, *определенный на фазовой траектории* X, *для которого выполняются следующие условия:*

- *Для всех* $j \in [1, J]:\ w^j \in W$,
- *Точка* I^j *траектории* X *соответствует наступлению аварии в наблюдаемой системе, возникшей в результате нештатного поведения класса* w^j, *для всех* j *от* 1 *до* J.

будем называть маркировкой точек аварий.

Пусть задана выборка TS в виде экземпляров траекторий X, полученных в различных условиях работы системы, с различными искажениями и шумами. При этом для каждой траектории задана маркировка точек аварии Ф. Считаем, что в траекториях из выборки TS участок нештатного поведения входит в последние r отсчетов, предшествующих точке аварии, где r является параметром и устанавливается до запуска алгоритма обучения для каждой конкретной системы. При таком способе задания обучающей выборки требуется, кроме построения алгоритма распознавания, определить эталонные траектории для различных классов нештатного поведения.

Всю заданную выборку разделим на обучающую $\widetilde{TS}$ и контрольную $\widehat{TS}$:

$$TS = \widetilde{TS} \cup \widehat{TS}, \quad \widetilde{TS} \cap \widehat{TS} = \varnothing.$$

Пусть так же определена целевая функция $\varphi(e_1, e_2)$, где e_1 и e_2 — число ошибок распознавания первого и второго рода.

Задача построения алгоритма распознавания заключается в следующем.

Требуется построить алгоритм распознавания Al и определить эталонные траектории с учетом следующих требований:

1. Алгоритм Al должен выдавать ограниченное число ошибок на обучающей выборке $\widetilde{TS}$:

$$e_1(Al, \widetilde{TS}) \leqslant \text{const}_1, \quad e_2(Al, \widetilde{TS}) \leqslant \text{const}_2,$$

   где: const_1 и const_2 — заданные ограничения.
2. Алгоритм Al должен минимизировать целевую функцию $\varphi(e_1, e_2)$ на контрольной выборке $\widehat{TS}$.
3. Вычислительная сложность работы алгоритма распознавания $\Theta_{Al}(m)$ на произвольной траектории, длины не больше m, должна быть ограничена наперед заданной функцией $\theta(m)$, которая определяется характеристиками используемого вычислителя и скоростью развития процессов в анализируемой системе:

$$\Theta_{Al}(m) \leqslant \theta(m) \ .$$

Описание аксиоматического подхода

Идея использования аксиоматического подхода для выделения трендов была предложена в работе [1], в работе [2] было предложено использование этого подхода для обнаружения нештатных режимов работы динамических систем. Основой аксиоматического подхода является переход от поиска вхождения траекторий к поиску подстрок.

Определение 2. *Элементарное условие $ec(x_t^*,t,P)$ — это функция, определенная на отчете t и некоторой его окрестности x_t^* на траектории X, зависящая от набора параметров P, которая принимает значения из множества* $\{\text{true}, \text{false}\}$.

Определение 3. *Аксиома $a = a(x_t^*, t)$ — это булева функция, заданная в виде формулы от элементарных условий, определенных на отсчете t и некоторой его окрестности x_t^* на траектории X:*

$$a(x_t^*, t) = \bigvee_i \bigwedge_j ec_{ij}(x_t^*, t, P_{ij})^{\delta_{ij}};$$

$$ec(x_t^*, t, P)^{\delta} = \begin{cases} ec(x_t^*, t, P), & \text{при } \delta = 0; \\ \overline{ec}(x_t^*, t, P), & \text{при } \delta = 1. \end{cases}$$

Точка траектории размечается аксиомой, если в данной точке аксиома принимает значение true.

Определение 4. *Конечное множество аксиом $as = \{a_1, \ldots, a_M\}$ будем называть системой аксиом, если оно удовлетворяет условию:*

$$\forall X \ \forall x_t \in X \ \exists! \ a_i \in as : \ a_i(x_t^*, t) = \text{true}.$$

Далее будем считать, что аксиомы в системе аксиом as пронумерованы последовательными натуральными числами: $as = \{a_1, \ldots, a_M\}$.

Любое множество аксиом возможно представить в виде системы аксиом выполнив следующее:

1. Введение порядка на множестве аксиом: если аксиома с индексом i выполняется в некоторой точке произвольной фазовой траектории, то считаем, что никакая аксиома с индексом j, $j > i$, не выполняется в данной точке.
2. Добавление тождественной аксиомы a_∞ с наименьшим приоритетом: $as = \{a_1, \ldots, a_M, a_\infty\}$ Тождественная аксиома a_∞ — это аксиома, которая выполняется в любой точке любой фазовой траектории.

Определение 5. *Разметкой траектории $X = (x_1, \ldots, x_n)$ относительно системы аксиом as будем называть последовательность $J = (j_1, \ldots, j_n)$, где j_t — индекс аксиомы в системе аксиом as, условия которой выполняются на отсчете t траектории X.*

Для поиска траекторий нештатного поведения размечаются эталонные траектории $\{X_{\text{Anom}}\}$ и наблюдаемая фазовая траектория X. Далее, в ряду разметки J ищутся последовательности аксиом соответствующие разметкам эталонных траекторий. Для поиска разметок используются методы нечеткого поиска подстрок, такие как алгоритм на основе метрики Минковского [3], DTW [5] и другие. Таким образом, определение нештатного поведения в работе наблюдаемой системы ведется путем поиска разметок эталонных траекторий в ряду разметки наблюдаемой фазовой траектории.

Для решения задачи построения алгоритма распознавания в рамках используемого подхода необходимо построить систему аксиом и определить разметки эталонных траекторий.

Алгоритм обучения

Преобразуем всю выборку TS. Каждую траекторию $X \in TS$, содержащую точку аварии, разделим на непересекающиеся части: траектория X' — это участок траектории X с отсчета 1 по отсчет $\max(1, I_X - r)$, траектория X'' — это участок X с отсчета $\max(1, I_X - r)$ по отсчет I_X, где I_X — это точка аварии на траектории X. После преобразований выборка TS представляет собой набор траекторий:

$$TS = \{X'\} \cup \{X''\}.$$

Часть из них является траекториями нормального поведения $\{X'\}$, другая часть $\{X''\}$ — представляет собой набор траекторий, каждая из которых содержит участок нештатного поведения, однако неизвестно на каких именно отсчетах.

Разделим преобразованную выборку TS на три равные части: выборка $\overline{TS}$ для определения разметок участков нештатного поведения, обучающая $\widetilde{TS}$ и контрольная $\widehat{TS}$ выборка.

$$TS = \overline{TS} \cup \widetilde{TS} \cup \widehat{TS};$$
$$\widetilde{TS} \cap \widehat{TS} = \varnothing, \quad \overline{TS} \cap \widehat{TS} = \varnothing, \quad \widetilde{TS} \cap \overline{TS} = \varnothing.$$

Обозначим: $\{ec\}$ — множество всех описанных элементарных условий.

Алгоритм обучения состоит в следующем:

1. Для каждого класса $l \in [1, L]$ нештатного поведения системы независимо формируется набор элементарных условий. Для каждого условия $\forall ec_i \in \{ec\}$ и его отрицания из $\overline{ec}_i$ строится сетка значений его параметров Π_i^{ec}. Из всех элементарных условий со всевозможными значениями параметров выбирается заданное число k условий с фиксированными параметрами, которые чаще выполняются на траекториях, содержащих участки нештатного поведения класса l, и реже на траекториях нормального поведения из обучающей выборки $\widetilde{TS}$.
2. Для каждого класса $l \in [1, L]$ нештатного поведения системы независимо формируется набор аксиом из выбранных элементарных условий.
   (а) Для каждого элементарного условия $ec_i^{\delta_i}$, выбранного ранее, создается аксиома: $a^i = ec_i^{\delta_i}$. Набор таких аксиом обозначим: Set_a.
   (б) Итеративное изменение набора аксиом Set_a. Для каждой пары различных аксиом a^i и a^j в наборе Set_a строится две новые аксиомы: $a^k = a^i \wedge a^j$ и $a^l = a^i \vee a^j$. В Set_a добавляются те из новых аксиом, которые чаще чем их родители выполняются на траекториях, содержащих участки нештатного поведения класса l, и реже на траекториях нормального поведения из $\widetilde{TS}$. Далее, множество Set_a сокращается до заданного числа аксиом $M_a^{\max}$ в соответствии с таким же критерием частоты выполнения на $\widetilde{TS}$. Процесс повторяется до тех пор, пока либо не было создано ни одной новой аксиомы, которая была добавлена в набор Set_a, либо число итераций превысило наперед заданный параметр.
3. Построение системы аксиом и разметок эталонных траекторий:
   (а) Для каждой аксиомы $a^i \in Set_a$ создается система аксиом: $as_i = \{a_1^i, a_\infty\}$. Множество всех таких систем обозначим: Set_{as}.
   (б) Для каждой системы $as_i \in Set_{as}$ формируются разметки эталонных траекторий. Эти разметки получаются из выборки $\overline{TS}$ путем выделения наибольшей общей подпоследовательности разметок траекторий системой as_i, содержащих участки нештатного поведения одного класса.
   (в) Проверяется критерий останова: если получена система аксиом с требуемым значением $\varphi(e_1, e_2)$ на контрольной выборке $\widehat{TS}$ или число итераций алгоритма превысило заданный параметр, то алгоритм останавливается.
   (г) Множество Set_{as} сокращается до заданного числа систем аксиом N_{as}^{max} с лучшими значениями $\varphi(e_1, e_2)$ на $\widetilde{TS}$.
   (д) Для каждой системы $as_i = \{a_1, \ldots, a_n, a_\infty\} \in Set_{as}$ и каждой аксиомы $a^j \in Set_a$, $a^j \notin as_i$, создается новая система аксиом, которая затем добавляется в Set_{as}: $as_u = \{a_1, \ldots, a_n, a_{n+1}^j, a_\infty\}$. Переход на шаг 3б.

   Выбор лучшей системы as в Set_{as} осуществляется по значению целевой функции $\varphi(e_1, e_2)$ на контрольной выборке $\widehat{TS}$.

Ключевой особенностью предложенного алгоритма является то, что он позволяет выделять разметки участков нештатного поведения, используемые как разметки эталонных траекторий, за счет выбора таких элементарных условий и аксиом, которые чаще выполняются на траекториях, содержащих участки нештатного поведения, и реже на траекториях нормального поведения.

Однако, данный алгоритм обладает и рядом недостатков, среди них:

— Алгоритм обладает высокой вычислительной сложностью, поскольку он использует направленный перебор.
— Алгоритм имеет большое число параметров, от которых зависит качество получаемой системы аксиом и скорость работы алгоритма. Параметры очередного этапа алгоритма могут зависеть от результатов работы предыдущего.
— Критерии отбора условий и аксиом не учитывают особенностей конкретной системы.
— Невозможно отследить, какой из этапов алгоритма в наибольшей степени влияет на качество получаемого решения и соответственно изменить параметры или перезапустить этап.

Автоматизированный метод обучения

Разработанная модификация алгоритма предполагает взаимодействие с пользователем. Метод состоит из четырех шагов:

1. Задание параметров и выполнение п. 1 основного алгоритма обучения.
2. Оценка результатов первого этапа, возможный его перезапуск, задание параметров и выполнение п. 2 основного алгоритма обучения.
3. Оценка результатов второго этапа, возможный его перезапуск, задание параметров и выполнение п. 3 основного алгоритма обучения.
4. Оценка результатов третьего этапа основного алгоритма и его возможный перезапуск.

Для каждого шага данного метода разработаны эвристические предположения, на основании которых пользователю предоставляется возможность:

— Сортировать и отсеивать часть полученных после 1-го и 2-го этапов алгоритма множеств элементарных условий и аксиом.
— Задавать параметры очередного этапа основного алгоритма.
— Возвращаться к одному из более ранних этапов основного алгоритма перезапускать его.

Действия пользователя на шаге 1:

— Ограничение множества различных типов элементарных условий $\{ec\}$. Исключение заведомо неэффективных условий приведет к снижению времени работы 1-го этапа алгоритма.
— Уточнение выбора участков траекторий, предшествующих точкам аварии. Возможно выделять не только последние r отсчетов, а ограничивать выделение с 2-х сторон и делать его специфичным для каждой траектории в TS.

Действия на шаге 2:

— Сокращение числа выбранных элементарных условий согласно следующим критериям:
1) Проведение автоматической кластеризации для каждого элементарного условия по отсчетам его выполнения на траекториях обучающей выборки $\widetilde{TS}$, при этом задаваемым параметром кластеризации является расстояние между отсчетами выполнения условия. Сокращение тех условий, размеры кластеров выполнения которых оказались слишком малыми или большими.
2) Удаление условий, которые выполняются на малом числе траекторий, содержащих участки нештатного поведения, и часто выполняются на траекториях нормального поведения.
— Перезапуск первого этапа основного алгоритма с новыми значениями параметров.

Действия на шаге 3:

— Сокращение числа выбранных аксиом согласно перечисленным выше критериям, как и для элементарных условий.
— Перезапуск второго этапа основного алгоритма с новыми значениями параметров.

Действия на шаге 4:

— Модификация построенных разметок эталонных траекторий. Ручная корректировка разметок может позволить уменьшить число ошибок распознавания.
— Перезапуск третьего этапа основного алгоритма с новыми значениями параметров.
— Выбор из построенных систем аксиом с наилучшим значением целевой функции $\varphi(e_1, e_2)$ той, которую считать решением задачи построения системы аксиом.

Для кластеризации выбран метод иерархической кластеризации ближайшего соседа [4], т. к. допускает задание минимального расстояния между кластерами и допускает эффективную реализацию.

Результаты исследования

Было проведено численное исследование, которое показало, что точность основного алгоритма и автоматизированного метода на его основе оказалась практически равной. Однако, при использовании автоматизированного метода удалось значительно сократить время обучения: выигрыш составил от 6-ти до 50-ти раз. Для многих реальных задач, например для задачи обучения алгоритма прогнозирования микроснов, время работы основного алгоритма может составлять до нескольких недель.

Разработанный метод позволяет пользователю применить экспертные знания о конкретной технической системе для увеличения точности и уменьшения времени построения системы аксиом.

Литература

[1] *Рудаков К. В., Чехович Ю. В.* О проблеме синтеза обучающих алгоритмов выделения трендов (алгебраический подход) // Прикладная математика и информатика. — 2001. — № 8. — С. 97–114.

[2] *Коваленко Д. С., Костенко В. А., Васин Е. А.* Исследование применимости алгебраического подхода к анализу временных рядов // Методы и средства обработки информации. Издательство фак. ВМиК МГУ, 2005. — С. 553–559.

[3] *Энслейн К., Рэлстон Э., Уилф Г.* Статистические методы для ЭВМ. — М: Наука, 1986.

[4] *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning (2nd ed.). — NY: Springer, 2009. — 745 p.

[5] *Keogh E. J., Michael J. Pazzani* Derivative Dynamic Time Warping // First SIAM International Conference on Data Mining (SDM'2001), Chicago, USA, 2001.

Постановка обобщенной задачи синтеза динамического объекта как задачи распознавания образов с активным экспериментом*

Неймарк Ю. И., Теклина Л. Г.
neymark@pmk.unn.ru
Нижний Новгород, Научно-исследовательский институт прикладной математики и кибернетики Нижегородского государственного университета им. Н. И. Лобачевского

В работе представлен новый подход к решению обобщенной задачи синтеза динамического объекта как задачи распознавания образов с активным экспериментом.

Введение

С нашей точки зрения, перспективы распознавания образов — в расширении сферы применения, в решении нестандартных для распознавания образов проблем. Ведь распознавание образов — это не только методы, это новый взгляд на решаемую проблему, когда из большого объема информации об исследуемом объекте вычленяется главное — то, что называется в распознавании «образами», а дополнительные детали, усложняющие и затуманивающие картину, отбрасываются. Такой взгляд на сложную проблему позволяет упростить задачу и найти для нее нетривиальное решение. Именно такой подход позволил разработать новую технологию численного исследования конкретных многомерных динамических систем с большим числом параметров методами распознавания образов [1, 2]. Результат такого исследования — построение огрубленных фазовых и параметрического портретов, включающих в себя определение всех видов устойчивых движений в системе, их областей притяжения в фазовом пространстве, а также областей в пространстве параметров, при которых существуют исследуемые устойчивые движения. С помощью новой методики исследован ряд математических моделей с большим числом параметров, причем исследование зависимости движений в системе от всех параметров моделей проведено впервые [3]. Новая методика реализована в виде комплекса программ, которые могут стать основой для автоматизации процесса исследования. Успешный опыт создания новой методики позволяет рассчитывать на успех в решении еще одной сложной прикладной для распознавания образов проблемы, а именно: обобщенной задачи синтеза динамического объекта.

Постановка обобщенной задачи синтеза динамического объекта

Рассматривается обобщенная задача синтеза динамического объекта с заданным качеством функционирования. Примерами таких объектов могут служить системы квазиинвариантного управления, которые должны не устранять возникающие ошибки, а предотвращать их, т. е. сделать объект управления невосприимчивым (инвариантным) к внешним воздействиям [4].

Синтезируемый объект описывается двумя группами переменных:

$$\mathbf{X} = \{\mathbf{x} = (x_1, x_2, \ldots, x_n)\};$$
$$\mathbf{A} = \{\mathbf{a} = (a_1, a_2, \ldots, a_m)\},$$

которые связаны между собой дифференциальными и алгебраическими уравнениями или неравенствами. Область изменения одной группы переменных $\mathbf{X}$ известна, это $\mathbf{X}^* \subset \mathbf{X}$. Обобщенная задача синтеза сводится к выбору множества переменных $\mathbf{A}$ и отысканию области их изменения $\mathbf{A}^* \subset \mathbf{A}$ такой, что для начальных значений переменных $x_1, x_2, \ldots, x_n$ из $\mathbf{X}^*$ и для значений переменных $a_1, a_2, \ldots, a_m$ из $\mathbf{A}^*$ решения систем уравнений $x_1(t), x_2(t), \ldots, x_n(t)$ удовлетворяли бы всем требованиям, предъявляемым к функционированию конструируемого объекта. Это означает, что известны технические характеристики синтезируемого объекта

$$\mathbf{Y} = \{\mathbf{y} = (y_1, y_2, \ldots, y_k)\},$$

и в этом множестве переменных задается область $\mathbf{Y}^* \subset \mathbf{Y}$, отвечающая определенным требованиям эксплуатации объекта.

Так, например, в задаче синтеза систем квазиинвариантного управления по известной математической модели, включающей в себя

- дифференциальные уравнения объекта управления;
- дифференциальные уравнения управления объектом с неизвестными параметрами;
- требования к техническим характеристикам управляемого объекта в виде систем уравнений или неравенств,

надо найти такой закон управления и такие значения неизвестных параметров, чтобы технические характеристики синтезируемой системы удовлетворяли всем предъявляемым к ней требованиям, а именно: обеспечение устойчивости системы управления, малость ошибки управления в установившемся режиме, соответствующие быстроту и каче-

Работа выполнена при финансовой поддержке РФФИ, проект № 11-01-00379.

ство переходного процесса, ограниченность значений управления [5, 6].

В математической постановке сказанное выше означает, что для заданных $\mathbf{X}$ и $\mathbf{A}$ существует преобразование Γ, в общем случае неоднозначное, которое объектам

$$\mathbf{x} \in \mathbf{X} \ \text{ и } \ \mathbf{a} \in \mathbf{A}$$

ставит в соответствие некоторый объект

$$\Gamma(\mathbf{x}, \mathbf{a}) = \mathbf{y} \in \mathbf{Y}.$$

Решение обобщенной задачи синтеза состоит в формировании множества $\mathbf{A}$ и поиске такой области $\mathbf{A}^* \subset \mathbf{A}$, что для всякого $\mathbf{x} \in \mathbf{X}^*$ и для всякого $\mathbf{a} \in \mathbf{A}^*$

$$\Gamma(\mathbf{x}, \mathbf{a}) = \mathbf{y} \in \mathbf{Y}^*.$$

Постановка задачи синтеза как задачи распознавания с активным экспериментом

Для постановки проблемы в виде задачи распознавания образов необходимо сформировать пространство признаков и выделить распознаваемые образы. В обобщенной задаче синтеза динамического объекта в качестве пространства признаков выбирается множество переменных $a_1, a_2, \ldots, a_m$. Именно эти данные определяют качество синтезируемого объекта. Когда признаки выбраны, за распознаваемый образ принимается область $\mathbf{A}^*$ в пространстве признаков, для всех точек которой соответствующие технические характеристики $\mathbf{y} = (y_1, y_2, \ldots, y_k)$ синтезируемого объекта отвечают всем предъявляемым к ним требованиям, т. е.

$$\mathbf{y} = \mathbf{\Gamma}(\mathbf{x}, \mathbf{a}) = \mathbf{\Psi}(\mathbf{a}) \in \mathbf{Y}^*,$$

где $\Psi(\mathbf{a})$ — преобразование, ставящее в соответствие любому описанию $\mathbf{a}$ (для всех $\mathbf{x} \in \mathbf{X}^*$) технические характеристики $\mathbf{y}$ синтезируемого объекта. Таким образом, при *заданном* описании $\mathbf{A}$ объектов распознавания по *известным* техническим характеристикам синтезируемого объекта $\mathbf{Y}^* \subset \mathbf{Y}$ надо по *известному* решающему правилу —

$$\mathbf{a} \in \mathbf{A}^*, \ \text{ если } \ \mathbf{\Psi}(\mathbf{a}) \in \mathbf{Y}^*,$$

— определить и описать распознаваемый класс $\mathbf{A}^* \subset \mathbf{A}$, т. е. решается проблема, обратная классической задаче распознавания, в которой по известной выборке данных о распознаваемом образе строится решающее правило.

Поставленная задача имеет две особенности:

— пространство признаков заранее неизвестно, поэтому возможна ситуация, когда множество $\mathbf{A}^*$ пусто;

— известное решающее правило достаточно сложно и требует больших вычислительных затрат при определении принадлежности $\mathbf{a}$ к $\mathbf{A}^*$.

С учетом этих особенностей задача определения и описания $\mathbf{A}^*$ в пространстве $\mathbf{A}$ решается двумя путями:

1. построение решающих правил на малых по объему обучающих выборках;
2. аппроксимация преобразования Γ другим более простым оператором $\tilde{\Gamma}$, требующим меньших вычислительных затрат.

Все обучающие выборки формируются в процессе решения задачи, т. е. решается задача распознавания с активным экспериментом. Для формирования таких выборок решается задача планирования эксперимента с целью достижения поставленных целей за минимальное число шагов. В задаче планирования эксперимента удобнее использовать решающую функцию вида

$$F(\mathbf{a}) = \rho\big(\mathbf{\Psi}(\mathbf{a}), \mathbf{Y}^*\big),$$

представляющую собой расстояние в пространстве $\mathbf{Y}$ от точки $\mathbf{y} = \mathbf{\Psi}(\mathbf{a})$ (технические характеристики объекта) до множества $\mathbf{Y}^*$. Эта функция не только указывает на принадлежность точки $\mathbf{a}$ к распознаваемому образу ($F(\mathbf{a}) = \mathbf{0}$ для $\mathbf{a} \in \mathbf{A}^*$), но и может служить мерой близости $\mathbf{a}$ к $\mathbf{A}^*$ при $F(\mathbf{a}) \neq \mathbf{0}$.

Что касается выбора описания $\mathbf{A}$, то оно требует конкретных знаний о конструируемом объекте и формируется на основе экспертных знаний. Предлагается начинать процесс решения с синтеза объекта минимальной сложности и увеличивать сложность объекта путем увеличения размерности пространства $\mathbf{A}$ или путем замены части переменных при условии, что для уже рассмотренной модели множество $\mathbf{A}^*$ пусто. Например, при синтезе систем квазиинвариантного управления можно начать с управления в виде кусочно-постоянной функции, а при необходимости перейти к линейным или кусочно-линейным функциям и т.д. Принятие решений об изменении пространства признаков производится с помощью решающих правил, построенных на базе экспертных оценок и с учетом результатов поиска $\mathbf{A}^*$ на предыдущем этапе.

Общая схема решения задачи синтеза в адаптивном режиме

Для решения обобщенной задачи синтеза динамического объекта методами распознавания предлагается простая схема, описываемая следующими шагами:

1. Формирование первичного пространства признаков $\mathbf{A}$.
2. Поиск $\mathbf{a}^* \in \mathbf{A}^*$ путем решения задачи планирования эксперимента с использованием функ-

ции $F(\mathbf{a})$. Если $\mathbf{a}^*$ существует, переходим к этапу 3. Если $\mathbf{a}^*$ не существует, переходим к этапу 4.

3. Построение $\bar{\mathbf{A}}^* \subseteq \mathbf{A}^*$ в виде области простой конфигурации по известному объекту $\mathbf{a}^*$ на основе гипотезы компактности.
4. Изменение пространства (соответственно и решающей функции $F(\mathbf{a})$) либо путем увеличения размерности, либо путем изменения части признаков. Переход к этапу 2.

Для поиска $\mathbf{a}^* \in \mathbf{A}^*$ на этапе 2 применяется процедура обучения. Для формирования первичного пространства признаков (этап 1), а при необходимости и изменения его (этап 4) используются знания экспертов и построенные на их основе правила принятия решений.

Заключение

В представленном докладе мы не пытаемся дать полное решение обобщенной задачи синтеза динамического объекта, а лишь намечаем подход к нестандартному решению поставленной проблемы. Новая задача — задача синтеза — рассматривается как дальнейшее развитие подхода, использованного для решения проблемы численного исследования динамических систем. При больших различиях в постановках этих двух актуальных задач их объединяют, во-первых, общность подхода, отличительной особенностью которого является переход от классических методов исследования к формированию обучающих выборок путем проведения экспериментов с последующей обработкой полученных результатов методами интеллектуального анализа данных, а во-вторых, возможность применения новой методики исследования устойчивости динамических систем в задаче синтеза. Предлагаемый подход универсален, он не опирается на конкретные особенности в постановке задачи. Его реализация позволит автоматизировать процесс решения сложной, но актуальной задачи синтеза динамического объекта.

Литература

[1] *Неймарк Ю. И., Котельников И. В., Теклина Л. Г.* Новый подход к численному исследованию конкретных динамических систем методами распознавания образов и статистического моделирования // Изв. вузов. Прикладная нелинейная динамика. — 2010. — Т. 18, № 2. — С. 3–15.

[2] *Неймарк Ю. И., Котельников И. В., Теклина Л. Г.* Новая технология численного исследования динамических систем методами распознавания образов // Математические методы распознавания образов. Доклады 14-й Всероссийской конференции. Москва: МАКС Пресс, 2009. — С. 418–421.

[3] *Neimark Yu. I., Teklina L. G.* About possibilities of application of the pattern recognition methods for research of the mathematical models // Pattern Recognition and Image Analysis: New Information Technologies. Proceedings of the 8-th International Conference (PRIA-10-2010). St. Petersburg: Politechnika, 2010. — Pp. 109–112.

[4] *Неймарк Ю. И.* О парадоксе и идеальном регуляторе Щипанова // Вестник ННГУ им. Н. И. Лобачевского. Математическое моделирование и оптимальное управление. — 2006. — Вып. 3(32). — С. 83–88.

[5] *Неймарк Ю. И.* Синтез и функциональные возможности простейшего квазиинвариантного управления // Вестник ННГУ им. Н. И. Лобачевского. — 2007. № 6. — С. 140–146.

[6] *Неймарк Ю. И.* Синтез и функциональные возможности квазиинвариантного управления // Автоматика и телемеханика. — 2008. — № 10. — С. 48–56.

Использование субмодулярного разложения в релаксационном подходе к обучению структурного метода опорных векторов*

Ветров Д. П., Осокин А. А., Шаповалов Р. В.
VetrovD@yandex.ru, Anton.Osokin@gmail.com, Shapovalov@graphics.cs.msu.su
Москва, Московский Государственный Университет им. М. В. Ломоносова

В работе рассматривается вопрос обучения структурного метода опорных векторов с применением т. н. релаксационного подхода для настройки параметров марковских случайных полей. Для поиска наиболее нарушаемого ограничения предлагается использовать метод субмодулярного разложения. Благодаря этому удается расширить область применимости релаксационного подхода.

Задачи байесовского вывода в марковских случайных полях (поиска наиболее вероятных значений скрытых переменных в вероятностной модели, представляющей марковскую сеть) получили большое распространение в течение последних 10 лет. Они возникают в таких прикладных областях как сегментация изображений, выделение объектов в видеопотоках, восстановление стерео, декодировка сообщений, анализ социальных сетей и многих других. Такие задачи сводятся к оптимизации логарифма совместного распределения, представляющего собой сумму функций от небольшого количества скрытых переменных.

Особый интерес представляют задачи, в которых скрытые переменные марковской сети принимают конечное число значений. Такая постановка является естественным обобщением стандартной задачи классификации на случай, когда объекты в выборке не являются независимыми, например, при анализе пикселей изображения или отсчетов сигнала. Можно показать, что если граф взаимозависимостей между объектами выборки имеет множественные циклы, то задача байесовского вывода становится NP-трудной [2]. В связи с этим в последние годы было предложено большое количество приближенных методов вывода в циклических марковских сетях, например, loopy belief propagation [1], α-расширение [2], max-sum diffusion [3], tree-reweighted message passing (TRW) [4, 6, 13], субмодулярное разложение (SMD) [5].

Задача автоматической настройки параметров марковских случайных полей долгое время оставалась открытой, т. к. даже вывод в рамках известной модели марковского поля оказался нетривиальной задачей. В последние годы, с разработкой эффективных методов вывода, интерес исследователей переключился на задачу обучения модели марковского поля по набору марковских полей с известными значениями скрытых переменных. Если задача байесовского вывода в марковском поле соответствует в стандартной задаче классификации принятию решения обученным классификатором, то задача настройки параметров поля соответствует обучению классификатора по прецедентам. Наиболее удачными оказались попытки использовать дискриминативный подход к настройке параметров, перенося на структурное обучение приемы, хорошо зарекомендовавшие себя в классических задачах машинного обучения: принцип максимизации зазора [8] и бустинг [9].

Для обучения дискриминативных марковских полей с помощью максимизации зазора (т. н. max-margin Markov networks) было предложено несколько обобщений известного метода опорных векторов, например, структурный SVM (SSVM) [8, 10]. Исследование нескольких подходов к обучению SSVM, проведенное в работе [11] выявило преимущества т. н. релаксационного (overgenerative) подхода над жадным (undergenerative). Тем не менее, хотя релаксационный подход и показал лучшее качество обучения, его практическое использование сопряжено со значительными вычислительными трудностями.

В настоящей работе использован новый подход к обучению релаксационных SSVM, основанный на использовании недавно предложенного субмодулярного разложения в качестве приближенного метода вывода [5]. К преимуществам этого подхода следует отнести возможность получения верхних оценок на максимум логарифма совместного распределения и нецелостных значений скрытых переменных. Благодаря использованию нецелостных значений скрытых переменных удается существенно упростить релаксационный подход.

Структурный метод опорных векторов

Рассмотрим некоторый граф $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, задающий структуру марковского поля. Здесь и далее будем считать, что логарифм совместного распределения переменных является парно-сепарабельной функцией вида

$$F(T) = \sum_{i \in \mathcal{V}} \varphi_i(t_i) + \sum_{(i,j) \in \mathcal{E}} \varphi_{ij}(t_i, t_j), \qquad (1)$$

где $t_i \in \mathcal{K} = \{1, \ldots, K\}$ — скрытые переменные поля. Предположим, что каждая вершина характеризуется вектором унарных признаков $\boldsymbol{x}_i \in \mathbb{R}^d$,

Работа выполнена при финансовой поддержке РФФИ (проект № 09-01-00409), гранта Президента РФ (МК3827.2010.9).

а каждое ребро — вектором парных признаков $\boldsymbol{x}_{ij} \in \mathbb{R}^e$. Будем считать, что потенциальные функции $\varphi_i(t_i)$, $\varphi_{ij}(t_i, t_j)$ задаются следующим образом:

$$\varphi_i(k) = \boldsymbol{w}_k^\intercal \boldsymbol{x}_i, \quad \varphi_{ij}(k,l) = \boldsymbol{w}_{kl}^\intercal \boldsymbol{x}_{ij},$$

где $\boldsymbol{w}_k \in \mathbb{R}^d$, $\boldsymbol{w}_{kl} \in \mathbb{R}^e$ — унарные и парные веса, соответственно. Рассматривается семейство решающих правил вида $T^* = \operatorname{argmax} F_{\boldsymbol{w}}(T)$.

Одним из способов настройки вектора весов $\boldsymbol{w} = \big((\boldsymbol{w}_k)_{k=1}^K, (\boldsymbol{w}_{kl})_{k,l=1}^K\big)$ по обучающей выборке[1] $(\tilde{X}, \tilde{T})$ является структурный метод опорных векторов, приближенно решающий следующую задачу:

$$\frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\xi \to \min_{\xi \geqslant 0, \boldsymbol{w}}; \tag{2}$$

$$F_{\boldsymbol{w}}(\tilde{T}) \geqslant F_{\boldsymbol{w}}(T) + \delta(\tilde{T}, T) - \xi, \quad \forall T \in \mathcal{K}^{|\mathcal{V}|}. \tag{3}$$

Здесь $C \geqslant 0$ — структурный параметр, определяющий величину штрафа за ошибку на обучении, а $\delta(\tilde{T}, T) \geqslant 0$ — функция, задающая разницу в значениях двух конфигураций. Здесь и далее в качестве δ будем рассматривать Хэммингово расстояние между разметками T и $\tilde{T}$.

Заметим, что при каждом T функция $G(\boldsymbol{w}, T) = F_{\boldsymbol{w}}(T) + \delta(\tilde{T}, T) - F_{\boldsymbol{w}}(\tilde{T})$ линейна по $\boldsymbol{w}$, а значит может быть записана в виде $\boldsymbol{a}_T^\intercal \boldsymbol{w} + b_T$, где $\boldsymbol{a}_T$ и b_T не зависят от $\boldsymbol{w}$ и ξ. Для каждого T условие (3) можно переписать в виде $\boldsymbol{a}_T^\intercal \boldsymbol{w} + b_T \leqslant \xi$. Все множество ограничений можно записать в виде $\tilde{A}^\intercal \boldsymbol{w} + \tilde{\boldsymbol{b}} \preceq \xi \mathbf{1}$[2], где матрица $\tilde{A}$ составлена из всех столбцов $\boldsymbol{a}_T$, а вектор $\tilde{\boldsymbol{b}}$ составлен из чисел b_T.

Количество ограничений (3) растет экспоненциально с ростом мощности $\mathcal{V}$, что делает невозможным решение задачи (2)–(3) при помощи стандартного алгоритма квадратичного программирования даже при размерах $\mathcal{V}$ порядка нескольких десятков. Стандартным способом решения задачи (2)–(3) является метод секущих плоскостей (алгоритм 1), на каждой итерации которого поддерживается текущие значения $\boldsymbol{w}_0$ и ξ_0, рабочее множество ограничений, определяемое матрицей A и вектором $\boldsymbol{b}$.

Метод секущих плоскостей подразумевает, что задачу (4) можно точно и эффективно решить. Во многих практически важных случаях это не так. В частности, в случае наличия циклов в графе $\mathcal{G}$ и количестве меток $K > 2$ задача (4) является NP-трудной. В этом случае используются приближенные методы. Выделяют два подхода к обобщению SSVM на случай NP-трудной задачи (4) [11]: жадный и релаксационный.

Алгоритм 1. Метод секущих плоскостей для обучения SSVM.

Вход: (X, T), C, ε;
Выход: $\boldsymbol{w}_0$, ξ_0;

1: инициализация:
   $\boldsymbol{w}_0 := \mathbf{0}$, $\xi_0 := 0$, $A := \varnothing$, $\boldsymbol{b} := \varnothing$;
2: **цикл**
3: решить оптимизационную задачу

$$T_0 = \operatorname*{argmax}_{T \in \mathcal{K}^{|\mathcal{V}|}} G(\boldsymbol{w}_0, T); \tag{4}$$

4: приведя подобные слагаемые в функции $G(\boldsymbol{w}, T_0)$, представить ее в виде $\boldsymbol{a}^T \boldsymbol{w} + b$;
5: **если** $\boldsymbol{a}^\intercal \boldsymbol{w}_0 + b > \xi_0 + \varepsilon$ **то**
6: $A := [A,\ \boldsymbol{a}]$, $\boldsymbol{b} := [\boldsymbol{b},\ b]$;
7: $(\boldsymbol{w}_0, \xi_0) := \operatorname*{argmin}_{\substack{\xi \geqslant 0 \\ A^\intercal \boldsymbol{w} + \boldsymbol{b} \preceq \xi \mathbf{1}}} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\xi$;
8: **иначе**
9: **выход**;

В рамках жадного подхода вместо точного решения задачи (4) используется *нижняя* оценка. В частности, в качестве приближения максимума можно выбрать приближенное решение задачи (4), полученное каким-либо методом дискретной оптимизации [1, 2, 4, 6, 13, 5].

В рамках релаксационного подхода вместо точного решения задачи (4) используется *верхняя* оценка. В частности, в качестве верхней оценки можно выбрать точное решение LP-релаксации задачи (4) [4]. Чтобы полученную таким образом оценку можно было представить в виде линейной функции от весов $\boldsymbol{w}$, необходимо вычислить, не только саму границу, но и значения релаксированных переменных. Эффективный метод решения LP-релаксированной задачи (TRW-S [6]) не позволяет получить значения релаксированных переменных, а значит неприменим в данном случае. Общие же методы решения задач линейного программирования слишком неэффективны и ограничивают размеры задач, для которых можно обучать SSVM десятками переменных. В реальных же задачах, возникающих, например в компьютерном зрении, количество переменных составляет тысячи.

И релаксационный, и жадный подходы решают задачу (2)–(3) приближенно. Жадный подход может получить решение, удовлетворяющее не всем ограничениям (3). Релаксационный подход может получить решение, при котором целевая функция принимает неоптимальное значение. Экспериментальное сравнение двух подходов, проведенное в работе [11], показывает, что релаксационный подход позволяет получать решающие правила, допускающие меньше ошибок на тестовой выборке, а зна-

[1] Для простоты дальнейшего изложения будем считать, что выборка состоит из одного объекта, представляющего собой марковское поле. Для каждой вершины и ребра графа $\mathcal{G}$ известны признаки $\tilde{X} = \big\{\{\tilde{\boldsymbol{x}}_i\}_{i \in \mathcal{V}}, \{\tilde{\boldsymbol{x}}_{ij}\}_{(i,j) \in \mathcal{E}}\big\}$. Также заданы правильные ответы $\tilde{T} = \big\{\tilde{t}_i\big\}_{i \in \mathcal{V}}$.

[2] Здесь и далее символом «$\preceq$» обозначается поэлементное неравенство векторов и матриц.

чит, является более предпочтительным. В то же время, из-за высокой вычислительной сложности решения возникающих задач линейного программирования, область применимости релаксационного подхода сильно ограничена.

Использование лагранжинана в релаксационном SSVM

В последние годы появились методы оптимизации правдоподобия марковских сетей (1), основанные на подходе двойственного разложения. К таким методам можно отнести DD-TRW [13] и SMD [5]. Подход двойственной декомпозиции позволяет получить верхнюю оценку точного решения задачи (4) в виде минимакса функции Лагранжа. В данной статье предлагается в рамках релаксационного подхода к обучению SSVM вместо LP-релаксации задачи (4) использовать верхнюю оценку, полученную с помощью двойственного разложения. Рассмотрим применение данного подхода с использованием метода SMD (все последующие выкладки можно провести и для DD-TRW).

Введем индикаторные переменные $y_{ik} = [t_i = k]$.[3] Тогда задачу (4) можно переписать:

$$D_{\boldsymbol{w}}(Y) = \sum_{i\in\mathcal{V}}\sum_{k=1}^{K} \boldsymbol{w}_k^\intercal \boldsymbol{x}_i y_{ik} + \sum_{(i,j)\in\mathcal{E}}\sum_{k,l=1}^{K} \boldsymbol{w}_{kl}^\intercal \boldsymbol{x}_{ij} y_{ik} y_{jl} + \\ + \sum_{i\in\mathcal{V}}\sum_{k=1}^{K} y_{ik}[\tilde{t}_i \neq k] - F_{\boldsymbol{w}}(\tilde{T}) \to \max_{\substack{y_{ik}\in\{0,1\} \\ \sum_{k\in\mathcal{K}} y_{ik}=1}}, \quad (5)$$

где условия $\sum_{k\in\mathcal{K}} y_{ik} = 1$ обеспечивают согласование индикаторных переменных y_{ik}.

Обозначим $\varphi_{\boldsymbol{w}}^{ik} = \boldsymbol{w}_k^\intercal \boldsymbol{x}_i + [\tilde{t}_i \neq k]$, $\varphi_{\boldsymbol{w}}^{ij,kl} = \boldsymbol{w}_{kl}^\intercal \boldsymbol{x}_{ij}$. Если бинарные потенциалы ассоциативны,[4] то задачу (5) можно решать при помощи метода SMD. В работе [5] показано, что лагранжиан задачи (5) можно записать в следующем виде:

$$L_{\boldsymbol{w}}(Y,\Lambda) = \sum_{k\in\mathcal{K}} \left(D_{\boldsymbol{w}}^k(Y_k) + \sum_{i\in\mathcal{V}} \lambda_i y_{ik} \right) - \sum_{i\in\mathcal{V}} \lambda_i, \quad (6)$$

где $Y_k = \{y_{ik}\}_{i\in\mathcal{V}}$, а $D_{\boldsymbol{w}}^k(Y_k)$ — выражения, линейные по $\boldsymbol{w}$. При этом максимум выражений $D_{\boldsymbol{w}}^k(Y_k) + \sum_{i\in\mathcal{V}} \lambda_i y_{ik}$ по Y_k можно эффективно вычислить при помощи алгоритма поиска минимального разреза графа. Получаем, что

$$\max_{\substack{y_{ik}\in\{0,1\} \\ \sum_{k\in\mathcal{K}} y_{ik}=1}} D_{\boldsymbol{w}}(Y) = \max_{y_{ik}\in\{0,1\}} \min_{\Lambda} L_{\boldsymbol{w}}(Y,\Lambda) \leqslant \\ \leqslant \min_{\Lambda} \max_{y_{ik}\in\{0,1\}} L_{\boldsymbol{w}}(Y,\Lambda).$$

Таким образом, минимакс лагранжиана (6) является верхней оценкой решения задачи (4). В работе [5] показано, что значение этого минимакса в точности равно значению условного максимума LP-релаксации задачи (4). Из этого факта следует, что при текущих весах $\boldsymbol{w}$ оценки, полученные при помощи функции Лагранжа и LP-релаксации, совпадают. Легко показать, что на модифицированный метод обучения SSVM распространяются все теоретические результаты по релаксационному подходу, полученные в работе [11].

Глобальные ограничения при обучении и выводе

Одним из достоинств субмодулярного разложения является возможность приближенного учета ограничений на статистики первого порядка от индикаторных переменных при поиске наиболее вероятных значений скрытых переменных. В самом деле, любые ограничения вида равенства или неравенства на функцию вида $S(Y) = \sum_{i\in\mathcal{V}}\sum_{k\in\mathcal{K}} v_{ik} y_{ik}$ приводят к появлению дополнительного множителя Лагранжа в выражении (6) и могут быть эффективно учтены. В частности, положив $v_{ik} = [k = k_0]$, мы можем получить ограничения на общее количество скрытых переменных, отнесенных к классу k_0.

Теорема 1. *Значение минимакса функции Лагранжа с линейными ограничениями на индикаторные переменные Y в точности равно значению условного максимума LP-релаксации* (6) *при тех же линейных ограничениях.*

Из теоремы 1 следует, что введение глобальных ограничений на Y не ухудшает качества субмодулярного вывода по отношению к LP-релаксации. Предположим теперь, что мы хотим настроить параметры марковского поля для определения значений скрытых переменных при глобальных линейных статистиках ограничениях. Примером такой задачи может являться задача сегментации изображения на области при известной суммарной площади областей, принадлежащих к одному классу. В этом случае, при поиске наиболее нарушаемого ограничения при обучении SSVM разумно проводить поиск не по всевозможным вариантам значений скрытых переменных, а только по тем, которые удовлетворяют глобальным ограничениям. Не ограничивая общности, рассмотрим ограничения вида равенства $S_m(Y) = \mu_m$, $m = 1,\ldots,M$. Тогда в качестве верхней оценки на решение задачи (4) можно использовать следующий минимакс:

$$\min_{\Lambda,\Pi} \max_{Y\in\{0,1\}^{K|\mathcal{V}|}} L_{\boldsymbol{w}}(Y,\Lambda) + \sum_{m=1}^{M} \pi_m \big(S_m(Y) - \mu_m\big).$$

[3] $[P] = 1$ тогда и только тогда, когда предикат P принимает значение «истина».

[4] Бинарные потенциалы $\varphi_{\boldsymbol{w}}^{ij,kl}$ называются ассоциативными, если $\varphi_{\boldsymbol{w}}^{ij,kl} = 0$ при $k \neq l$ и $\varphi_{\boldsymbol{w}}^{ij,kl} \geqslant 0$ при $k = l$. Обеспечить выполнение таких условий можно, положив $\boldsymbol{w}_{kl} = \boldsymbol{0}$ при $k \neq l$, $\boldsymbol{w}_{kl} \succeq \boldsymbol{0}$ при $k = l$, $\boldsymbol{x}_{ij} \succeq \boldsymbol{0}$.

Таблица 1. Время работы (в секундах) методов оптимизации энергии марковской сети со структурой, задаваемой графом вида «квадратная решетка» со стороной квадрата размера N.

Метод	$N = 10$	$N = 20$	$N = 30$	$N = 50$
TRW-S	0,01	0,04	0,09	0,25
SMD	0,04	0,08	0,12	0,5
LP	1,8	8,9	31	150

Таблица 2. Средние ошибки на объектах контрольной выборки для жадного и релаксационного подходов в SSVM в зависимости от структурного параметра C. Размер обучающей выборки был равен 10, размер тестовой выборки — 100. Для получения ответов на тестовой выборке использовался алгоритм TRW-S с весами, полученными алгоритмом SSVM.

Подход	$C = 10^2$	$C = 10^4$	$C = 10^6$
жадный	11%	10%	7.5%
релаксационный	7.5%	4.7%	5.7%

Результаты экспериментов

Алгоритмы вывода. В таблице 1 представлены результаты сравнения времени работы алгоритмов TRW-S [6], SMD [5], общего метода решения задач линейного программирования (прямодвойственный метод внутренней точки [12]) на модельных задачах. В рамках данного эксперимента рассматривались марковские сети над переменными, принимающими 10 значений ($K = 10$), со структурой, задаваемой графом вида «квадратная решетка». Унарные и бинарные потенциалы генерировались случайно: $\varphi_i(x_i) \sim \mathcal{N}(0,1)$, $\varphi_{ij}(t_i, t_j) = 10\,|c_{ij}|\,[t_i \neq t_j]$, $c_{ij} \sim \mathcal{N}(0,1)$. Таблица 1 содержит средние времена работы каждого из алгоритмов на задачах со стороной квадрата размером N.

Результаты эксперимента показывают, что алгоритм SMD лишь немного уступает по скорости работы алгоритму TRW-S и намного опережает общий метод решения задач линейного программирования, большое время работы которого объясняется большим количеством переменных, возникающих в задаче ($O(N^2K^2)$).

Алгоритмы обучения. В рамках данного эксперимента проводилось сравнение жадного и релаксационного (основанного на SMD) подходов к обучению структурного метода опорных векторов. Рассматривались графические модели вида «квадратная решетка» с ассоциативными парными потенциалами; размер стороны квадрата $N = 100$; количество меток переменных $K = 3$. Унарные признаки x_i^k генерировались случайно из равномерного распределения на отрезке $[0, 1]$; бинарные признаки строились по следующим правилам: $x_{ij}^1 = 1$, $x_{ij}^2 = |n_i - N/2|\,/N$, где n_i — позиция пикселя i по горизонтальному направлению решетки. Ответы T как контрольной, так и обучающей выборок генерировались при помощи оптимизации правдоподобия по разметкам методом TRW-S при «правильном» значении весов $\boldsymbol{w}$. В таблице 2 приведены результаты сравнения жадного и релаксационного (основанного на SMD) подходов к обучению структурного метода опорных векторов. Среднее время обучения в жадном подходе составляет 800 секунд, в релаксационном подходе — 5400 секунд.

Выводы. Проведенные эксперименты подтверждают вывод [11] о превосходстве релаксационного подхода над жадным и позволяют сделать вывод о целесообразности использования SMD для ускорения обучения SSVM в рамках релаксационного подхода.

Литература

[1] *Frey B., MacKay D.* A Revolution: Belief propagation in graphs with cycles // NIPS, 1998.

[2] *Boykov Y., Veksler O., Zabih R.* Fast Approximate Energy Minimization via Graph Cuts // TPAMI, 2001. — V. 23, N. 11. — Pp. 1222–1239.

[3] *Werner T.* A Linear Programming Approach to Max-sum Problem: A Review // TPAMI, 2007. — V. 29, N. 7. — Pp. 1165–1179.

[4] *Wainwright M., Jaakkola T., Willsky A.* MAP estimation via agreement on trees: message-passing and linear programming // Transactions on Information Theory. — 2005. — V. 51, N. 11. — Pp. 3697–3717.

[5] *Osokin A., Vetrov D., Kolmogorov V.* Submodular Decomposition Framework for Inference in Associative Markov Networks with Global Constraints // CVPR, 2011.

[6] *Kolmogorov V.* Convergent Tree-reweighted Message Passing for Energy Minimization // TPAMI. — 2006. — V. 28, N. 10. — Pp. 1568–1583.

[7] *Lafferty J., McCallum A., Pereira F.* Conditional Random Fields: Probabilitics models for segmenting and labeling sequence data // ICML, 2001.

[8] *Taskar B., Guestrin C., Koller D.* Max margin Markov networks // NIPS, 2003.

[9] *Munoz D., Bagnell A., Vandapel N., Hebert V.* Contextual classification with functional Max-Margin Markov Networks // CVPR, 2009.

[10] *Joachims T., Finley T., Yu C.* Cutting-Plane Training of Structural SVMs // Machine Learning. — 2009. — V. 77, N. 1. — Pp. 27–59.

[11] *Finley T., Joachims T.* Training structural SVMs when exact inference in intractable // ICML, 2008.

[12] *Mehrotra S.* On the Implementation of a Primal-Dual Interior Point Method // SIAM Journal on Optimization. — 1992. — V. 2. — Pp. 575–601.

[13] *Komodakis N., Paragios N., Tziritas G.* MRF Energy Minimization and Beyond via Dual Decomposition // TPAMI. — 2010. — V. 33, N. 3. — Pp. 531–552.

Решение задач оптимизации на марковских полях с помощью разложения, сохраняющего структуру графа*

Осокин А. А., Ветров Д. П.
Anton.Osokin@gmail.com, VetrovD@yandex.ru
Москва, Московский Государственный Университет им. М. В. Ломоносова

В работе предложен подход к приближенному решению сложных задач дискретной оптимизации на циклических графах, основанный на разложении K-значной задачи на множество бинарных подзадач, для точного, либо приближенного решения которых, существуют алгоритмы построения минимальных разрезов графов. Особое внимание в работе уделено сравнению предложенного подхода с существующим аналогом, основанным на разложении задачи на графе с циклами на подзадачи на ациклических подграфах. Получены условия, когда оба подхода дают эквивалентные результаты. Установлены частные случаи, в которых предложенный подход позволяет получить точное решение исходной задачи.

Задачи байесовского вывода в марковских случайных полях (поиска наиболее вероятных значений скрытых переменных в вероятностной модели, предсатвляющей марковскую сеть) получили большое распространение в течение последних 10 лет. Они возникают в таких прикладных областях как сегментация изображений, выделение объектов в видеопотоках, восстановление стерео, декодировка сообщений, анализ социальных сетей и мн. др. Такие задачи сводятся к оптимизации отрицательного логарифма правдоподобия, представляющего собой сумму функций от небольшого количества скрытых переменных и по аналогии со статистической физикой называемого энергией.

Особый интерес представляют задачи, в которых скрытые переменные марковской сети принимают конечное число значений. Такая постановка является естественным обобщением стандартной задачи классификации на случай, когда объекты в выборке не являются независимыми, например, при анализе пикселей изображения или отсчетов сигнала. Можно показать, что возникающая задача дискретной оптимизации является, в общем случае, NP-трудной. Для решения этой задачи в последние годы было предложено большое количество приближенных методов вывода на циклических марковских сетях, например, loopy belief propagation [1], α-расширение [2], max-sum diffusion [3], и др. Одним из наиболее интересных подходов является разбиение NP-трудной исходной задачи на полиномиально разрешимые подзадачи и последующая максимизация нижней границы энергии исходной задачи. В случае использования в качестве подзадач деревьев подход получил название tree-reweighted message passing (TRW) [4, 5, 6], а в случае использования субмодулярных функций — субмодулярного разложения (SMD) [7].

Особенностью субмодулярного разложения является тот факт, что разбиение исходной K-значной задачи проводится на субмодулярные бинарные подзадачи. При этом структура графа, на котором задана оптимизируемая функция, не меняется. Это позволяет ввести т. н. индикаторную параметризацию на метки классов, которая может быть использована для учета глобальных ограничений как функций от индикаторных переменных. В частности, в задачах сегментации изображений можно ввести глобальные ограничения на площадь класса, средний цвет класса, звездовидность формы и др. [7]. При этом на оптимизируемый функционал накладываются жесткие ограничения: ассоциативность марковского поля.

В настоящей работе предложено обобщение субмодулярного разложения на случай, когда разложение на бинарные подзадачи возможно, но получающиеся подзадачи больше не являются субмодулярными. Для их приближенного решения использован алгоритм квадратичной псевдобулевой оптимизации (QPBO) [8, 9]. Это обобщение получило название GPLD (Graph preserving label decomposition). Хотя, в общем случае, показано, что GPLD уступает по качеству нижней границы энергии методу TRW, установлен ряд случаев, когда методы эквивалентны. Преимуществом GPLD, аналогично SMD, является возможность учета глобальных ограничений за счет сохранения структуры графа. Получен ряд нетривиальных условий частичной оптимальности нецелостного решения, получаемого с помощью GPLD, и условий, когда по нецелостному решению удается построить оптимальное решение исходной задачи.

Марковские случайные поля

Рассмотрим произвольный граф $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ и сопоставим каждой его вершине K-значную переменную. Рассмотрим на графе функционал, представимый в виде суммы функций, определенных на вершинах и ребрах графа $\mathcal{G}$:

$$E(T) = \sum_{i \in \mathcal{V}} \varphi_i(t_i) + \sum_{(i,j) \in \mathcal{E}} \varphi_{ij}(t_i, t_j),$$

где $t_i \in \{1, \ldots, K\}$ — переменные, соответствующие вершинам графа. При использовании веро-

Работа выполнена при финансовой поддержке РФФИ (проект № 09-01-00409), гранта Президента РФ (МК3827.2010.9).

ятностного (байесовского) формализма часто вводят распределение на множестве вершин графа $p(T) \propto \exp(-E(T))$. Вероятностную модель такого вида называют марковским полем, а функцию $E(T)$ — энергией марковского поля. Задача байесовского вывода заключается в нахождении наиболее вероятной конфигурации переменных T, то есть в поиске

$$\hat{T} = \operatorname*{argmax}_T \mathsf{P}(T) = \operatorname*{argmin}_T E(T).$$

Данная задача является задачей дискретной оптимизации и, в общем случае, NP-трудна. Практический интерес представляют частные случаи, когда поиск минимума $E(T)$ может быть проведен эффективно: графа $\mathcal{G}$ является деревом; субмодулярные энергии при $K = 2$. Субмодулярность означает выполнение следующих условий:

$$\varphi_{ij}(1,1)+\varphi_{ij}(2,2) \leqslant \varphi_{ij}(1,2)+\varphi_{ij}(2,1), \quad \forall (i,j) \in \mathcal{E}.$$

Легко показать [10], что субмодулярные энергии могут быть эффективно минимизированы с помощью поиска минимального разреза в специальном графе.

В дальнейшем будем рассматривать т. н. индикаторную нотацию, при которой каждой переменной t_i будем сопоставлять K-мерный бинарный вектор $\boldsymbol{y}_i$, причем $y_{ik} = [t_i = k]$.[1] Обозначив $\varphi_{ik} = \varphi_i(k)$, $\varphi_{ij,kl} = \varphi_{ij}(k,l)$, энергию E можно переписать в виде квадратичной псевдобулевой функции

$$E(Y) = \sum_{i\in\mathcal{V}} \sum_{k=1}^{K} \varphi_{ik} y_{ik} + \sum_{(i,j)\in\mathcal{E}} \sum_{k,l=1}^{K} \varphi_{ij,kl}\, y_{ik}\, y_{jl}. \quad (1)$$

Минимизация этой функции по бинарным Y при ограничениях

$$\sum_{k=1}^{K} y_{ik} = 1, \quad \forall i \in \mathcal{V} \quad (2)$$

эквивалентна минимизации исходной энергии.

Разложение марковских полей с сохранением структуры графа

Предположим, что $\varphi_{ij,kl} = C_{ij,k}[k = l]$. Тогда задача (1) может быть переписана в виде

$$E(Y) = \sum_{k=1}^{K} \underbrace{\left[\sum_{i\in\mathcal{V}} \varphi_{ik} y_{ik} + \sum_{(i,j)\in\mathcal{E}} C_{ij,k} y_{ik} y_{jk} \right]}_{E_k(Y_k)} \to \min_{\sum_k y_{ik}=1}. \quad (3)$$

[1] $[P] = 1$ тогда и только тогда, когда предикат P принимает значение «истина».

Отказ от условия (2) приводит к разбиению задачи (3) на подзадачи относительно переменных $Y_k = \{y_{ik}\}_{i\in V}$. Очевидно, что

$$\min_{\sum_k y_{ik}=1} E(Y) \geqslant \sum_{k=1}^{K} \min_{Y_k} E_k(Y_k), \quad (4)$$

Если все $C_{ij,k} \leqslant 0$, все подзадачи нижней границы (4) могут быть эффективно решены при помощи алгоритма поиска минимального потока в графе. В общем случае, для приближенного решения подзадачи можно воспользоваться методом QPBO [9].

Теорема 1. *(Колмогоров и др., [9]) Метод QPBO получает нижнюю оценку на значение минимума функции $E_k(Y_k)$, причем значение этой нижней оценки совпадает со значением LP-релаксации задачи $\min\limits_{Y_k} E_k(Y_k)$:*

$$\min_{\mathcal{Y}_k} \sum_{i\in\mathcal{V}} \varphi_{ik} y_{ik} + \sum_{(i,j)\in\mathcal{E}} C_{ij,k} y_{ij,kk} \quad (5)$$

при условиях

$$y_{ij,kk} + y_{ij,k\bar{k}} = y_{ik}, \quad y_{ij,kk} + y_{ij,\bar{k}k} = y_{jk};$$
$$y_{ij,\bar{k}k} + y_{ij,\bar{k}\bar{k}} = 1 - y_{ik}, \quad y_{ij,k\bar{k}} + y_{ij,\bar{k}\bar{k}} = 1 - y_{jk};$$
$$y_{ik}, y_{ij,kk}, y_{ij,\bar{k}k}, y_{ij,\bar{k}\bar{k}} \in [0,1],$$

где множество переменных $\mathcal{Y}_k$ состоит из релаксированных унарных (y_{ik}) и бинарных ($y_{ij,kk}$, $y_{ij,\bar{k}k}$, $y_{ij,k\bar{k}}$, $y_{ij,\bar{k}\bar{k}}$) индикаторов.

Несмотря на то, что метод QPBO позволяет получить только целочисленное значения бинарной скрытой переменной, отвечающей вершине, либо отказ от определения ее значения, нетрудно конструктивно построить значения LP-релаксированных переменных, введенных в теореме 1. Для этого напомним известный факт из теории (0,1)-целочисленного линейного программирования.

Теорема 2. *(Хаммер и др., [11]) Существует оптимальное решение задачи (5), в котором все переменные y_{ik}, $y_{ij,kk}$, $y_{ij,\bar{k}k}$, $y_{ij,k\bar{k}}$, $y_{ij,\bar{k}\bar{k}}$ принимают значения из множества $\{0, 1/2, 1\}$.*

Основываясь на двух последних теоремах, можно показать, что вершинам, класс которых был неопределен в ходе работы QPBO будет соответствовать LP-релаксированная унарная индикаторная переменная, равная $1/2$. Парные индикаторные переменные однозначно определяются по значениям индикаторных переменных за исключением случая $y_{ik} = y_{jk} = 0.5$. Здесь возможно два варианта в зависимости от знака $C_{ij,k}$. При $C_{ij,k} < 0$ верно, что $y_{ij,kk} = y_{ij,\bar{k}\bar{k}} = 0{,}5$, $y_{ij,\bar{k}k} = y_{ij,k\bar{k}} = 0$, а при $C_{ij,k} \geqslant 0$ верно, что $y_{ij,\bar{k}k} = y_{ij,k\bar{k}} = 0{,}5$, $y_{ij,kk} = y_{ij,\bar{k}\bar{k}} = 0$.

Выпишем задачу нахождения наиболее точной нижней оценки на (3). Введя множители Лагранжа для ограничений (2) и приведя подобные слагаемые, получим

$$\min_{\sum\limits_k y_{ik}=1} E(Y) \geqslant$$
$$\sum_{k=1}^{K} \min_{Y_k\in\{0,1\}^{|\mathcal{V}|}} \left[E_k(Y_k) + \sum_{i\in V}\lambda_i y_{ik}\right] - \sum_{k=1}^{K}\lambda_i \geqslant$$
$$\sum_{k=1}^{K} \min_{Y_k\in\{0,1/2,1\}^{|\mathcal{V}|}} \left[E_k^{LP}(Y_k) + \sum_{i\in V}\lambda_i y_{ik}\right] - \sum_{k=1}^{K}\lambda_i = \Phi(\Lambda),$$

где $E_k^{LP}(Y_k)$ — целевая функция задачи (5).

Поскольку последнее неравенство верно для всех Λ, нижняя оценка, наиболее близкая к минимуму энергии будет достигаться при $\Lambda^* = \operatorname{argmax}\Phi(\Lambda)$. При максимизации функции $\Phi(\Lambda)$ для все большего числа вершин будет выполняться условие (2), а значит разметки вершин графа будут приближаться к целостным. Заметим, что функция $\Phi(\Lambda)$ является вогнутой и может быть эффективно максимизирована с помощью метода субградиентного подъема.

Сравнение с TRW

Одним из наиболее широко применяемых методов приближенного решения NP-трудных задач на графах является алгоритм TRW, основанный на декомпозиции исходной задачи на подзадачи, определенные на деревьях, в совокупности покрывающих весь граф $\mathcal{G}$. Можно показать [4], что TRW сходится к нижней границе энергии $E(Y)$, совпадающей с ее LP-релаксацией:

$$\min_Y \sum_{i\in\mathcal{V}}\sum_{k=1}^{K}\varphi_{ik}y_{ik} + \sum_{(i,j)\in\mathcal{E}}\sum_{k,l=1}^{K}\varphi_{ij,kl}y_{ij,kl} \qquad (6)$$

при условиях (2),

$$y_{ik}\geqslant 0,\quad \sum_{k=1}^{K} y_{ij,kl} = y_{jl},\quad \sum_{l=1}^{K} y_{ij,kl} = y_{ik}.$$

В общем случае, GPLD сходится к точке с меньшим значением нижней границы. Следующая теорема устанавливает условия, при которых методы дают эквивалентные результаты.

Теорема 3. *Если все элементы множества*

$$\operatorname*{Argmin}_{Y_k\in\{0,1/2,1\}^{|\mathcal{V}|}} E_k^{LP}(Y_k)$$

принадлежат $\{0,1\}$*, метод GPLD сходится к тому же значению нижней границы, что и TRW.*

Заметим, что условие теоремы не зависит от множителей Лагранжа Λ, поэтому его проверку можно провести один раз перед запуском собственно метода.

Основным достоинством GPLD является возможность учета т. н. глобальных ограничений на статистики первого порядка от индикаторных переменных y_{ik}, например, на общее число вершин, относящихся к одному классу. Учет ограничений

$$\sum_{i\in V}\sum_{k=1}^{K} v_{ik}^m y_{ik} = c^m,\quad m = 1,\ldots,M; \qquad (7)$$

$$\sum_{i\in V}\sum_{k=1}^{K} u_{ik}^l y_{ik} \leqslant d^l,\quad l = 1,\ldots,L \qquad (8)$$

приводит к появлению дополнительных множителей Лагранжа. В отличие от GPLD, метод TRW не позволяет напрямую учитывать глобальные ограничения подобного рода.

Теорема 4. *Пусть выполнены условия теоремы 3. Тогда GPLD, учитывающий глобальные ограничения вида* (7), (8)*, сходится по значению нижней границы к решению задачи* (6) *при дополнительных ограничениях* (7), (8).

Последняя теорема позволяет рассчитывать лишь на приближенное выполнение глобальных ограничений на решение, поскольку при переходе к целостным целочисленным значениям переменных y_{ik} условия (7), (8) могут нарушаться. Точное решение задачи линейного программирования в этом случае уже не удается восстановить.

Частичная оптимальность нецелостных решений

Пусть $\hat{Y} = \{\hat{y}_{ik}\}_{i\in\mathcal{V},k=1,\ldots,K}$ — точка сходимости GPLD.[2] В общем случае, по $\hat{Y}$ не удается восстановить разметку $\hat{T}$, т. к. не для всех вершин выполнены условия (2). Возникает естественный вопрос, в каких случаях это можно сделать и в каких случаях можно гарантировать, что вершины, для которых выполнено условие (2), будут иметь ту же принадлежность и в точке, соответствующей глобальному минимуму энергии $E(T)$. Ответ на этот вопрос дают следующие три теоремы

Теорема 5. *(Хаммер и др., [11]) Рассмотрим произвольное* $Y_k^* = \{y_{ik}^*\}_{i\in\mathcal{V}}$*, такое что*

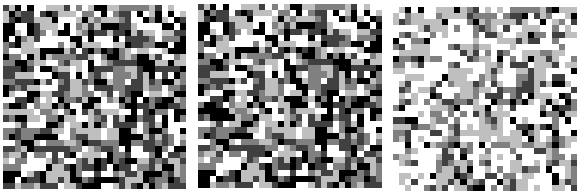
$$Y_k^* \in \operatorname*{Argmin}_{Y_k\in\{0,1/2,1\}^{|\mathcal{V}|}} E_k(Y_k).$$

Тогда найдется $Y_k^0 \in \operatorname*{Argmin}_{Y_k\in\{0,1\}^{|\mathcal{V}|}} E_k(Y_k)$*, такой что для всех* $i \in \left\{i \mid y_{ik}^* \in \{0,1\}\right\}$ *справедливо* $y_{ik}^* = y_{ik}^0$.

[2] Нетрудно показать, что $\hat{Y} = \arg_Y \max_\Lambda \min_Y L(Y,\Lambda)$, где $L(Y,\Lambda)$ функция Лагранжа, возникающая при релаксации ограничений (2).

Таблица 1. Результаты работы алгоритмов TRW-S, GPLD, GPLD-C. Для каждого метода приведены средние и стандартные отклонения от нижних границ (LB), энергий (E), суммарного нарушения всех ограничений (C).

Метод	LB	E	C
TRW-S	-13334 ± 383	-13305 ± 384	550 ± 32
GPLD	-13334 ± 383	-13240 ± 393	381 ± 38
GPLD-C	-857 ± 341	-10104 ± 523	110 ± 40



(а) (б) (в)

Рис. 1. Примеры результатов работы алгоритма TRW (а), GPLD (б), GPLD с глобальными ограничениями (в) Цвет пикселя обозначает номер метки в полученном решении.

Теорема 6. *Пусть для каждой вершины i графа $\mathcal{G}$ справедливо $|\{k \mid \hat{y}_{ik} \neq 0\}| \leqslant 2$. Тогда найдется $Y^0 = \{y^0_{ik}\}_{i\in\mathcal{V},k=1,\dots,K}$ — решение задачи* (3)*, такое что для всех $i \in \{i \mid \exists! k : \hat{y}_{ik} \neq 0\}$ справедливо $\hat{y}_{ik} = y^0_{ik}$.*

Теорема 7. *Пусть выполнено условие теоремы 6. Тогда существует алгоритм, позволяющий за линейное время восстановить решение задачи* (6) *и, в частности, определить, имеет ли LP-релаксация целочисленные решения.*

Алгоритм из последней теоремы выполняет проверку двудольности некоторого специального графа. Эксперименты показывают, что условия последних двух теорем часто бывают выполнены. В этом случае мы, по меньшей мере, можем частично восстановить оптимальное решение исходной задачи. Заметим, что все установленные результаты справедливы также для субмодулярного разложения, являющегося частным случаем GPLD.

Эксперименты

В рамках данного эксперимента рассматривались марковские сети вида «квадратная решетка» со стороной квадрата размера 30 и количеством классов 5. Унарные и бинарные потенциалы генерировались случайно: $\varphi_{ik} \sim \mathcal{N}(0,1)$, $\varphi_{ij,kl} = C_{ij}[k = l]$, $C_{ij} \sim \mathcal{N}(0,1)$. Проведено сравнение алгоритмов TRW-S [5], GPLD, GPLD с глобальными ограничениями (GPLD-C). В качестве глобальных ограничений для последнего алгоритма были выбраны жесткие ограничения на размеры сегментов: для каждого класса m в ограничении 7 параметры задавались следующим образом: $v^m_{ik} := [k = m]$, $c^m = 30^2 \frac{m}{\sum_{k=1}^K k}$. Такие глобальные ограничения сильно нарушаются при построении безусловного минимума.

Рис. 1 демонстрирует примеры работы алгоритмов TRW-S, GPLD, GPLD-C на одной из случайно сгенерированных модельных задач. Цвет пикселя кодирует значение переменной в решении, полученном соответствующим методов. Черный цвет соответствует метке 1, белый — метке 5.

Таблица 1 показывает усредненные по 50 экспериментам результаты работы трех алгоритмов. Приведены средние значения и стандартные отклонения нижних границ, энергий, и общего нарушения ограничений, вычисляемого по решению $\tilde{Y}$:

$$C(\tilde{Y}) = \sum_{m=1}^{K} \left| \sum_{i\in V} \sum_{k=1}^{K} v^m_{ik} \tilde{y}_{ik} - c^m \right|.$$

Литература

[1] *Frey B., MacKay D.* A Revolution: Belief propagation in graphs with cycles // NIPS, 1998.

[2] *Boykov Y., Veksler O., Zabih R.* Fast Approximate Energy Minimization via Graph Cuts // TPAMI. — 2001. — V. 23, N. 11. — Pp. 1222–1239.

[3] *Werner T.* A Linear Programming Approach to Max-sum Problem: A Review // TPAMI. — 2007. — V. 29, N. 7. — Pp. 1165–1179.

[4] *Wainwright M., Jaakkola T., Willsky A.* MAP estimation via agreement on trees: message-passing and linear programming // Transactions on Information Theory. — 2005. — V. 51, N. 11. — Pp. 3697–3717.

[5] *Kolmogorov V.* Convergent Tree-reweighted Message Passing for Energy Minimization // TPAMI. — 2006. — V. 28, N. 10. — Pp. 1568–1583.

[6] *Komodakis N., Paragios N., Tziritas G.* MRF Energy Minimization and Beyond via Dual Decomposition // TPAMI. — 2010. — V. 33, N. 3. — Pp. 531–552.

[7] *Osokin A., Vetrov D., Kolmogorov V.* Submodular Decomposition Framework for Inference in Associative Markov Networks with Global Constraints // CVPR, 2011.

[8] *Boros E., Hammer. P. L.* Pseudo-Boolean optimization // Discrete Applied Mathematics. — 2002. — V. 123, Nn. 1–3. — Pp. 155—225.

[9] *Kolmogorov V., Rother C.* Minimizing non-submodular functions with graph cuts - a review // TPAMI. — 2007. — V. 29, N. 7. — Pp. 1274–1279.

[10] *Kolmogorov V., Zabih R.* What Energy Functions can be Minimized via Graph Cuts? // TPAMI. — 2004. — V. 26, N. 2. — Pp. 147–159.

[11] *Hammer P. L., Hansen P., Simeone B.* Roof duality, complementation and persistency in quadratic 0-1 optimization // Mathematical Programming. — 1984. — V. 28, N. 2. — Pp. 121–155.

Классификация последовательностей, порождённых скрытыми марковскими моделями, при наличии шума

Гультяева Т. А., Попов А. А.
gult_work@mail.ru, alex@fpm.ami.nstu.ru
Новосибирск, Новосибирский Государственный Технический Университет

В докладе представлены результаты исследований возможности повышения дискриминирующих свойств скрытых марковских моделей с использования признаков, порождаемых этими моделями. В предлагаемом пространстве признаков используется классификатор k ближайших соседей. Рассматриваются только близкие по одному из параметров модели при условии искажения последовательностей шумом.

Скрытые марковские модели (СММ) являются мощным средством моделирования различных процессов и распознавания образов. По своей природе марковские модели позволяют непосредственно учитывать пространственно-временные характеристики последовательностей, и поэтому получили широкое применение [1–2]. Однако, имея хорошие описательные способности, СММ не всегда демонстрируют необходимые дискриминирующие свойства, важные для задач классификации.

Классификация последовательностей

Скрытые марковские модели. СММ полностью описывается следующими параметрами.

1. Вектор вероятностей начальных состояний $\Pi = \{\pi_i\}_{i=1}^N$, где $\pi_i = P\{q_1 = i\}$, q_1 — скрытое состояние в начальный момент времени $t = 1$, N — количество скрытых состояний в модели.
2. Матрица вероятностей переходов $A = (a_{ij})_{N\times N}$, где $a_{ij} = P\{q_t = j | q_{t-1} = i\}$. Последовательность скрытых состояний, моделируемая такой цепью, обозначается как $Q = \{q_1, q_2, \ldots, q_T\}$, где T — длина наблюдаемой последовательности.
3. Вероятностей наблюдаемых символов выглядит следующим образом: $b_i(t) = P\{o_t | q_t = i\}$, где o_t — символ, наблюдаемый в момент времени $t = 1, \ldots, T$, $i = 1, \ldots, N$. В данной работе рассматривается случай, когда функция распределения вероятностей наблюдаемых символов описывается смесью нормальных распределений:

$$b_i(t) = \sum_{m=1}^{M_i} c_i^m (\sqrt{2\pi}\sigma_i^m)^{-1} e^{-(o_t-\mu_i^m)^2/2(\sigma_i^m)^2},$$

где μ_i^m и σ_i^m — это параметры нормального распределения: математическое ожидание и среднеквадратичное отклонение соответственно, $m = 1, \ldots, M_i$, $i = 1, \ldots, N$.

Таким образом, СММ полностью описывается матрицей вероятностей переходов, а также вероятностями наблюдаемых символов и вероятностями начальных состояний: $\lambda = (A, B, \pi)$.

Традиционно при использовании СММ используется классификатор, основанный на отношении логарифмов функций правдоподобия: последовательность наблюдений $O = \{o_1, o_2, \ldots, o_T\}$ считается порождённым моделью λ_1, если выполняется:

$$\ln L(O|\lambda_1) > \ln L(O|\lambda_2). \qquad (1)$$

Иначе считается, что последовательность порождена моделью λ_2. Параметры моделей λ_1 и λ_2, как правило, неизвестны, и поэтому сначала должна быть произведена их оценка (например, с использованием алгоритма Баум-Велша [3]). Если конкурирующие модели близки по параметрам, а наблюдаемые сигналы не являются чисто гауссовскими последовательностями, то традиционная техника классификации с применением (1) далеко не всегда даёт приемлемые результаты. В качестве альтернативы рассмотрим схему, в которой при построении классификатора используются признаки, извлекаемые из обученных скрытых марковских моделей.

Схема классификации.

Этап 1. Для каждой обучающей последовательности $O_l^{learn_i}$, $i = 1, \ldots, K_l$, $l = 1, 2$, где K_l — число обучающих последовательностей для класса с номером l, формируется характеристический вектор. Этот вектор состоит из признаков, зависящих в общем случае от вероятности порождения наблюдаемой последовательности той или иной скрытой марковской моделью. В итоге характеристический вектор для обучающей последовательности $O_l^{learn_i}$, сгенерированной моделью λ_1, в рассматриваемом нами случае двухклассовой классификации будет состоять из двух подвекторов: $V_l^{learn_i} = \left(Z(O_l^{learn_i},\lambda_1)\ Z(O_l^{learn_i},\lambda_2) \right)^{\mathsf{T}}$, где для первого подвектора используются признаки, инициированные моделью λ_1, а для второго — моделью λ_2.

Этап 2. Аналогичным образом для тестовой последовательности O^{test} вычисляется характеристический вектор.

Этап 3. С использованием некой метрику $\rho(x, y)$ вычисляется мера сходства между характеристический векторами, полученными для обучающих и тестовой последовательности.

Этап 4. С использованием некоторого метрического классификатора выясняется, к какому классу принадлежит O^{test}. В работе использовался простейший, но достаточно хорошо себя зареко-

мендовавший классификатор — k ближайших соседей (k nearest neighbors — kNN). Это метрический классификатор, основанный на оценивании сходства объектов. Классифицируемый объект относится к тому классу, которому принадлежит большинство из его соседей — k ближайших к нему объектов обучающей выборки.

В результате проведённых экспериментов было выяснено, что метод взвешенных ближайших соседей даёт наилучшие результаты. Каждому соседу приписывается вес, убывающий с ростом ранга соседа, т. е. чем дальше находится объект из обучающей выборки от тестируемого объекта, тем меньше его вес.

Оптимальное значение параметра k определялось по критерию скользящего контроля с исключением объектов по одному.

Пространство признаков. Ряд авторов (например, [4, 5]) предлагают в качестве пространства признаков, в котором классифицируются последовательности, использовать пространства так называемых вторичных признаков: например, прямых, обратных переменных, используемых для вычисления вероятности порождения последовательности моделью. В качестве признаков используются также первые производные от логарифма функции правдоподобия по различным параметрам конкурирующих моделей. Авторы [5] предлагают включать также саму исходную последовательность O^{test} в вектор признаков. В работе рассматривается возможность проведения классификации в пространстве первых производных логарифма функции правдоподобия по элементам матрицы вероятностей переходов.

Возможности классификаторов целесообразно исследовать в условиях близости конкурирующих моделей. Скрытые марковские модели определяются достаточно большим числом параметров, и, произвольно задавая им значения, мы не можем априори фиксировать степень близости конкурирующих моделей. Один из подходов может состоять в параметризации различия моделей.

Вычислительные эксперименты

Исследования проводились при следующих условиях. Модели λ_1 и λ_2 определены на одинаковых по структуре скрытых марковских цепях и различались только в матрицах переходных вероятностей:

$$\mathbf{A}_{\lambda_1} = \begin{pmatrix} 0.1 & 0.7 & 0.2 \\ 0.2 & 0.2 & 0.6 \\ 0.8 & 0.1 & 0.1 \end{pmatrix},$$

$$\mathbf{A}_{\lambda_2} = \begin{pmatrix} 0.1+d & 0.7-d & 0.2 \\ 0.2 & 0.2+d & 0.6-d \\ 0.8-d & 0.1 & 0.1+d \end{pmatrix}.$$

Вероятности начальных состояний также совпадали: $\pi = (1, 0, 0)$. Параметр d, который можно варьировать в определённых пределах, фактически определяет степень близости конкурирующих моделей. Далее в работе приведены результаты при $d = 0.1$.

Обучающие и тестовые последовательности моделировались по методу Монте-Карло. Для проведения экспериментов было сгенерировано по 5 обучающих наборов последовательностей. К каждому набору этих последовательностей моделировалось по 5 тестовых наборов. Каждый набор содержал по 100 последовательностей для каждого класса. Таким образом, всего было смоделировано $10K_{learn}$ обучающих и 1000 тестовых последовательностей. Здесь K_{learn} — это количества обучающих последовательностей в каждом классе. Результаты классификации усреднялись.

В данной работе рассматривается искажение истинной последовательности с использованием аддитивного наложения на последовательности шумовой составляющей, распределённой по некоторому вероятностному закону. Обозначим смоделированную последовательность по модели λ через u. Тогда, при наложении на эту последовательность шума e согласно следующей формуле, получаем зашумлённую последовательность y с аддитивным шумом:

$$y = (1 - \omega)u + \omega e,$$

где параметр ω определяет степень искажения сигнала шумом.

Рассмотрим вариант, когда каждое скрытое состояние описывается смесью из одного распределения, то есть $M_i = 1,\ i = 1, \ldots, N$.

Параметры гауссовских распределений для моделей λ_1 и λ_2 выбирались одинаковыми:

$\mu_1^1 = 0,\ \mu_2^1 = 5,\ \mu_3^1 = 10,\ \sigma_1^1 = \sigma_2^1 = \sigma_3^1 = 1.$

Далее на всех рисунках график, отражающий результаты классификации для kNN, имеет пунктирную линию, а график для традиционного подхода — сплошную линию.

На рис. 1 приведены результаты сравнения классификации при шуме, распределённом по нормальному закону: $e \succ \mathcal{N}(0, 25)$. Как видно из данного эксперимента, классификатор kNN не даёт улучшений в сравнении с традиционным классификатором. Одновременно с этим можно говорить и об устойчивости (робастности) классификатора на основе отношения функций правдоподобия в условиях данного зашумления. Это может объясняется тем, что шум и истинный сигнал имеют одно и тоже нормальное распределение (но с различными параметрами), и алгоритм Баум-Велша, используемый для оценки параметров, также настроен именно для оценки параметров функции нормального распределения вероятностей наблю-

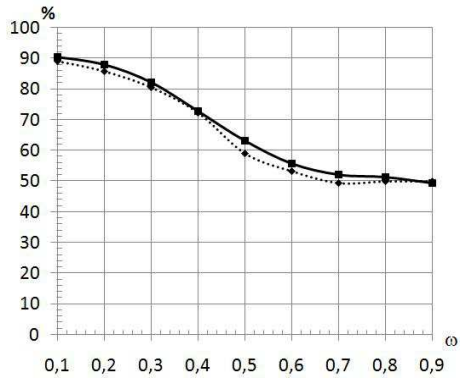


Рис. 1. Зависимость процента верно классифицированных последовательностей от уровня шума при распределении ошибок, подчиняющемся нормальному закону распределения при $M_i = 1,\ i = 1, \ldots, N$.

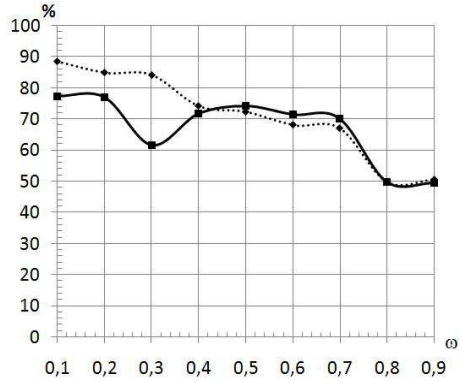


Рис. 2. Зависимость процента верно классифицированных последовательностей от уровня шума при распределении ошибок, подчиняющемся закону распределения Коши при $M_i = 1,\ i = 1, \ldots, N$.

даемых символов. На рис. 2 приведены результаты сравнения классификации при шуме, распределённому по закону Коши: $e \succ C(0, 0.1)$. Здесь наблюдается совершенно другая ситуация: при малом и среднем уровне шума, т. е. при $\omega \leqslant 0.4$, классификатор kNN даёт стабильно лучшие результаты. Таким образом, мы можем говорить о потери робастности традиционного классификатора к засоряющему распределению с тяжёлыми хвостами типа распределения Коши.

Теперь посмотрим, как будут вести себя классификаторы, когда каждое скрытое состояние описывается смесью из трёх распределений. В данном случае $M_i = 3,\ i = 1, \ldots, N$. Параметры гауссовских распределений для моделей λ_1 и λ_2 выбирались одинаковыми:

$c_i^1 = 1,\ c_i^2 = c_i^3 = 0\ ,\ i = 1, \ldots, N;$
$\mu_1^m = 0,\ \mu_2^m = 5\ ,\ \mu_2^m = 10,\ m = 1, 2, 3;$
$\sigma_i^m = 1,\ m = 1, \ldots, M_i, i = 1, \ldots, N.$

В этом пример последовательности, моделируемые СММ, реально имеют те же параметры, что и в предыдущем примере. Дополнительные компоненты смесей введены для того, чтобы при оценивании параметров моделей была возможность учесть и шум.

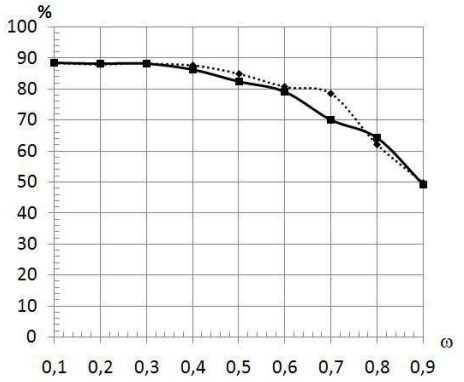


Рис. 3. Зависимость процента верно классифицированных последовательностей от уровня шума при распределении ошибок, подчиняющемся закону распределения Коши при $M_i = 3,\ i = 1, \ldots, N$; каждое скрытое состояние фактически описывается одной компонентой смеси.

Результаты для этого примера сравнения классификации при шуме, распределённому по закону Коши $e \succ C(0, 0.1)$ приведены на рис. 3. Сравнивая результаты, приведённые на этом рисунке с результатами на рис. 2, можно сделать вывод, что введение дополнительных компонент смеси позволило повысить робастность как традиционного классификатора, так и классификатора, основанного на kNN. Кроме того, последний классификатор почти при любых значениях степени искажения сигнала ω даёт результаты не хуже, чем традиционный.

Далее рассмотрим пример, когда каждое скрытое состояние описывается явно тремя компонентами смеси. Параметры гауссовских распределений для моделей λ_1 и λ_2 выбирались одинаковыми:

$c_i^1 = 0.3,\ c_i^2 = 0.5,\ c_i^3 = 0.2\ ,\ i = 1, \ldots, N;$
$\mu_1^1 = 0,\ \mu_1^2 = 5\ ,\ \mu_1^3 = 10,$
$\mu_2^1 = 3,\ \mu_2^2 = 8\ ,\ \mu_2^3 = 13,$
$\mu_3^1 = 6,\ \mu_3^2 = 11\ ,\ \mu_3^3 = 16;$
$\sigma_i^m = 1,\ m = 1, \ldots, M_i, i = 1, \ldots, N.$

На рис. 4 приведены результаты сравнения классификации при шуме, распределённому по закону Коши: $e \succ C(0, 0.1)$. Здесь также видно преимущество использования классификатора kNN перед традиционным классификатором при наличии шума.

По проведённым экспериментам можно сделать вывод, что традиционный классификатор не проявляет свойство робастности к шуму, распределённому по закону Коши. Необходимо отметить, что данный вид распределения помехи как некий индикатор широко используется в исследованиях, связанных с робастными методами оценивания параметров. Объясняется это тем, что данный вид

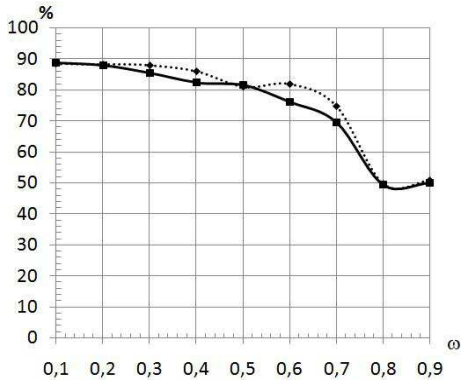


Рис. 4. Зависимость процента верно классифицированных последовательностей от уровня шума при распределении ошибок, подчиняющемся закону распределения Коши при $M_i = 3,\ i = 1, \ldots, N$; каждое скрытое состояние описывается тремя компонентами смеси.

распределения помехи относится к распределениям с тяжёлыми хвостами, и в этих условиях оценки параметров базового нормального распределения, полученные по методу максимального правдоподобия, становятся смещёнными. В нашем случае метод максимального правдоподобия задействуется на этапе оценки параметров СММ (алгоритм Баум-Велша). При этом оценки дисперсии имеют не просто смещение, а тенденцию к неограниченному росту, что приводит к эффекту сближения моделей в части параметров эмиссионного процесса. Однако, не очевидный, но экспериментально подтверждённый факт говорит о том, что рассматриваемые классификаторы по-разному реагируют на этот момент.

Выводы

Исследования показали, что если использовать классификатор k ближайших взвешенных соседей, то при близких по своим параметрам моделях и последовательностях, искаженных помехой, имеющей распределение Коши, удаётся повысить качество классификации.

Литература

[1] *Моттль В. В., Мучник И. Б.* Скрытые марковские модели в структурном анализе сигналов — Москва: Физматлит, 1999. — 351 с.

[2] `www.perso.telecom-paristech.fr/~cappe/docs/hmmbib.html`. — Ten years of HMM — 2001.

[3] *Rabiner L. R.* A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition // Proceedings of the IEEE, 1989. — Vol. 77, No. 2. — Pp. 257–285.

[4] *Aran O., Akarun L.* Recognizing two handed gestures with generative, discriminative and ensemble methods via Fisher kernels // In LNCS: Multimedia Content Representation. Classification and Security International Workshop, 2006. — Vol. 4015. — Pp. 159–166.

[5] *Chen L., Man H.* Combination of Fisher Scores and Appearance Based by Features For Face // Proceedings of the IEEE of the 2003 ACM SIGMM Workshop on Biometrics Methods and Applications, 2003. — Pp. 74–81.

Нечёткое оценивание параметров формы сигналов с учётом априорной информации в задаче инфразвукового мониторинга атмосферы*

Демин Д. С., Чуличков А. И., Куличков С. Н.
dmitryy.demin@gmail.com
Москва, Институт физики атмосферы им. А. М. Обухова РАН

Предлагается теоретико-возможностный метод оценки азимутального угла прихода и следовой скорости фронта инфразвуковой волны, учитывающий априорную информацию о сигнале, по данным регистрации акустического давления массивом пространственно разнесённых микрофонов.

Введение

При проведении инфразвукового мониторинга атмосферы регистрируются сигналы от удалённого источника с помощью группы пространственно разнесённых микрофонов (см. рис. 1). Базовой задачей мониторинга является определение азимутального направления прихода акустической волны, а также скорости её следа [2]. Эти параметры могут быть легко выражены через величины относительных временных задержек сигналов в различных каналах.

Рассмотрим массив, состоящий из m микрофонов. Каждый микрофон определяется своим расположением $\boldsymbol{r_j}$, $j = 1, \ldots, m$, на земной поверхности. Обозначим $\boldsymbol{r}_{j_1,j_2} = \boldsymbol{r}_{j_2} - \boldsymbol{r}_{j_1}$, $j_1, j_2 = 1, \ldots, m$. Каждым микрофоном через равные промежутки времени измеряется акустическое давление ($\boldsymbol{\xi}$). Результатом измерения является массив $\boldsymbol{\xi} \in \mathbb{R}^{n\times m}$ значений сигналов $\boldsymbol{\xi}_{\cdot,j} \in \mathbb{R}^n$, $j = 1, \ldots, m$:

$$\boldsymbol{\xi} = \boldsymbol{h} + \boldsymbol{\nu},\ \boldsymbol{h}, \boldsymbol{\nu} \in \mathbb{R}^{n\times m},\ |\nu_{i,j}| \leqslant \delta \qquad (1)$$

$$\boldsymbol{h} \in \mathbb{V}_{mf}(\boldsymbol{\lambda}),\ \boldsymbol{\lambda} \in \Lambda \qquad (2)$$

$$\mathbb{V}_{mf}(\boldsymbol{t}) = \big\{\boldsymbol{g} \in \mathbb{R}^{n\times m} : \boldsymbol{g}_{\cdot,j} \in \mathbb{R}^n, \\ \boldsymbol{g}_{\cdot,j} = \mathrm{B}_{t_j} \circ F_j \circ \boldsymbol{f},\ \boldsymbol{f} \in R^n\big\} \subset \mathbb{R}^{n\times m}, \quad (3)$$

где $\boldsymbol{\nu}$ — шум, свойства которого мы конкретизируем далее, $\mathbb{V}_{mf}(\boldsymbol{t})$ — множество, называемое формой сигнала [1], заданной параметрически с вектором $\boldsymbol{t} = (t_1, \ldots, t_m)$ параметров — временных задержек t_j, $F_j(\cdot)$ — j-е монотонное преобразование:

$$F_j(\cdot)\colon \mathbb{R}^1 \to \mathbb{R}^1\colon\ \forall x_1, x_2 \in \mathbb{R}^1, \\ x_1 \geqslant x_2 \Rightarrow F_j(x_1) \geqslant F_j(x_2), \quad (4)$$

оператор $B_{t_j}\colon \mathbb{R}^n \to \mathbb{R}^n$, $(B_{t_j} \circ \boldsymbol{f})_i = f_{i+t_j}$ — параметрически заданный обратимый оператор временной задержки сигнала на целое число интервалов дискретизации t_j. Для оценивания относительных временных задержек сигналов мы воспользуемся методами, предложенными в [5].

Работа выполнена при финансовой поддержке РФФИ, проекты № 11-07-00338-а, № 11-05-00890.

Таким образом, по результатам измерений $\boldsymbol{\xi}$ требуется оценить азимутальное направление прихода звуковой волны φ и скорость следа c_{hor} фронта возмущения. Их связь с t_j будет показана в следующем параграфе.

Мы будем использовать методы теории возможности [3] для построения оценок максимальной возможности и учёта априорной информации о значениях параметров. Теория возможностей, по существу, является теорией аддитивных мер со значениями в полукольце $[0, 1]$, в котором операция сложения определена как максимум, а умножения — как минимум, и теорией интегрирования по этим мерам. Конструкции меры возможности применяются в морфологическом анализе сигналов и изображений для моделирования нечёткости и неточности представлений о сцене и условиях регистрации её изображения [1]. Мера возможности задаёт порядок, указывающий, какие события возможны, какие менее возможны, а какие невозможны вообще.

Связь оцениваемых параметров и регистрируемых сигналов

Из физики известно, что величины, которые требуется оценить, связаны с относительными задержками сигналов в различных каналах следующей системой соотношений [2]:

$$\Delta t_{j_1,j_2} = t_{j_2} - t_{j_1} = (\boldsymbol{q}, \boldsymbol{r}_{j_1,j_2}), \qquad (5)$$

$$\varphi = \operatorname{arctg} 2(q_y, q_x),\ \cos(\theta) = c\sqrt{q_x^2 + q_y^2}, \quad (6)$$

$$c_{hor} = \frac{c}{\cos(\theta)} = \frac{1}{\sqrt{q_x^2 + q_y^2}}, \qquad (7)$$

где $\Delta t_{j_1,j_2}$ — относительная задержка сигналов в каналах j_1 и j_2, c — скорость звука в воздухе, φ — азимутальный угол, θ — угол наклона волнового вектора к горизонтали, c_{hor} — скорость следа волнового фронта на поверхности, $\boldsymbol{q} = \frac{\boldsymbol{k}}{\omega}$ — отношение волнового вектора и круговой частоты звуковой волны.

Из (5)–(7) получаются следующие переходные возможности:

$$\pi(c_{hor},\varphi\,|\boldsymbol{q}) = \begin{cases}1 & (6),(7),\\ 0 & \text{иначе}\end{cases} \quad (8)$$

$$\pi(\boldsymbol{q}\,|\boldsymbol{\Delta t}) = \mu_1(\|A\boldsymbol{q}-\boldsymbol{\Delta t}\|^2), \quad (9)$$

$$(A)_{k,\cdot} = \boldsymbol{A}^k = \boldsymbol{r}_k^T,\ k=1,\ldots,e, \quad (10)$$

где

$$\mu_1(\cdot)\colon [0,+\infty)\to[0,1],\ \ \mu_1(0)=1,\ \lim_{x\to+\infty}\mu_1(x)=0$$

— монотонно убывающая функция, каждое значение индекса k соответствует единственная пара значений индексов (j_1,j_2), $j_1,j_2=1,\ldots,m$, $e=C_m^2$, строками $(A)_{k,\cdot}=\boldsymbol{A}^k$ матрицы A являются векторы $\boldsymbol{r}_k=\boldsymbol{r}_{j_1,j_2}$.

Заметим, что $\pi(\boldsymbol{q}\,|\boldsymbol{\Delta t})$ можно было бы задать несколько по-другому:

$$\pi(\boldsymbol{q}\,|\boldsymbol{\Delta t}) = \delta\left(\|\boldsymbol{q}-A^-\boldsymbol{\Delta t}\|\right),\ \delta(x)=\begin{cases}1, & x=0,\\ 0, & \text{иначе},\end{cases}$$

но именно вид (9) учитывает априорную информацию о том, что с ростом значения невязки $\|A\boldsymbol{q}-\boldsymbol{\Delta t}\|$ оценка $\boldsymbol{\Delta t}$ по результату наблюдения $\boldsymbol{\xi}$ всё менее соответствует предположениям физической модели формирования сигнала (1)–(5).

Возможность, порождённая сравнением сигналов по форме

В [5] приведены способы сравнения по форме сигналов, порождённых монотонно неубывающими преобразованиями из одного прообраза.

Введём распределение $\mathfrak{dt}(\cdot):\mathbb{T}^e\to[0,1]$, возможности на прямом произведении $\mathbb{T}^e=\bigotimes\limits_{k=1}^{e=C_m^2}\mathbb{T}_k$ множества относительных задержек сигналов $\mathbb{T}_k=[-\Delta t_{k,max},\Delta t_{k,max}]$.

$$\mathfrak{dt}(\boldsymbol{\Delta t})=\mathfrak{dt}(\Delta t_1,\ldots,\Delta t_e)=\mu_2\left(D_{\boldsymbol{\xi}}(\boldsymbol{\Delta t})\right), \quad (11)$$

где $\boldsymbol{\Delta t}\in\mathbb{T}^e$ — вектор относительных задержек сигналов, $\Delta t_{j_1,j_2}=\Delta t_k$ — относительная задержка j_1 и j_2 сигналов,

$$\Delta t_{k,max}=\Delta t_{j_1,j_2,max}=\frac{\|\boldsymbol{r}_{j_1,j_2}\|}{c},$$

$\mu_2(\cdot)$ — монотонно убывающая функция меры различия смещённых сигналов $D_{\boldsymbol{\xi}}(\boldsymbol{\Delta t})$,

$$\mu_2(\cdot)\colon [0,+\infty)\to[0,1],\ \ \mu_2(0)=1,\ \lim_{x\to+\infty}\mu_2(x)=0.$$

Мера различия смещённых сигналов $D_{\boldsymbol{\xi}}(\boldsymbol{\Delta t})$ может быть задана любым из способов, изложенных в [5], например, порождённым равномерной метрикой.

Нечёткая модель интерпретации измерений. Построение оценки максимальной возможности

Объединяя (8)-(10), получим нечёткую модель интерпретации измерений:

$$P(c_{hor},\varphi)=\sup_{\boldsymbol{q},\boldsymbol{\Delta t}}\pi(c_{hor},\varphi\,|\boldsymbol{q})\bullet\pi(\boldsymbol{q}\,|\boldsymbol{\Delta t})\bullet\mathfrak{dt}(\boldsymbol{\Delta t})=$$

$$=\sup_{\boldsymbol{q},\boldsymbol{\Delta t}}\min\left\{\delta\left(\varphi-\operatorname{arctg}2(q_y,q_x)\right),\ \mu_2\left(D_{\boldsymbol{\xi}}(\boldsymbol{\Delta t})\right),\right.$$

$$\left.\delta\left(c_{hor}-\frac{1}{\sqrt{q_x^2+q_y^2}}\right),\ \mu_1(\|A\boldsymbol{q}-\boldsymbol{\Delta t}\|^2)\right\}.$$

Оценка максимальной возможности даётся выражением

$$(c_{hor,0},\varphi_0)=\arg\sup_{c_{hor},\varphi}P(c_{hor},\varphi). \quad (12)$$

Заметим, что в силу чёткости модели (8), задачу (12) можно заменить на оценивание вектора $\boldsymbol{q}$:

$$\boldsymbol{q}_0=\arg\sup_{\boldsymbol{q}}P(\boldsymbol{q})=$$

$$=\arg\sup_{\boldsymbol{q},\boldsymbol{\Delta t}}\min\left\{\mu_1(\|A\boldsymbol{q}-\boldsymbol{\Delta t}\|^2),\right.$$

$$\left.\mu_2\left(D_{\boldsymbol{\xi}}(\boldsymbol{\Delta t})\right)\right\},$$

которую, в свою очередь, можно разделить на задачу относительно $\boldsymbol{q}$

$$\boldsymbol{q}_0=A^-\boldsymbol{\Delta t}_0=(A^*A)^{-1}A^*\boldsymbol{\Delta t}_0, \quad (13)$$

в которой параметром будет решение $\boldsymbol{\Delta t}_0$ другой задачи уже относительно $\boldsymbol{\Delta t}$:

$$\boldsymbol{\Delta t}_0=\arg\sup_{\boldsymbol{\Delta t}}\min\left\{\mu_1(\|(AA^--I)\boldsymbol{\Delta t}\|^2),\right.$$

$$\left.\mu_2\left(D_{\boldsymbol{\xi}}(\boldsymbol{\Delta t})\right)\right\}, \quad (14)$$

являющейся задачей на условный максимин, которая может быть решена численно.

Подход, основанный на понятиях соотношения сигнал-шум и морфологической корреляции

В методах теории ИВС и морфологического анализа [1, 4] предполагается, что все данные о физическом объекте, получаемые при регистрации его свойств, можно разделить на две части: информативную часть, которая интересует исследователя и о математическая модель которой считается известной, а также шумовую составляющую, которая соответствует той части информации об объекте, которая не интересует исследователя и часто моделирует всевозможные погрешности, сопровождающие процесс регистрации.

Конструктивно информативная часть данных задаётся либо моделью, связывающей параметры объекта с результатом регистрации в случае «идеального измерительного прибора» в теории ИВС, множеством возможных результатов наблюдения за объектом при меняющихся условиях регистрации («формой») в морфологическом анализе сигналов и изображений.

Таким образом, той части информации об объекте, которая интересует исследователя, можно поставить в соответствие результат действия на полученные при регистрации данные оператора проецирования на множество, задаваемое моделью «идеальных» данных (формой). В случае, когда данные об объекте $\boldsymbol{\eta} \in \aleph$ представляют собой вектор линейного нормированного пространства (как в данной работе), результат действия оператора проецирования задаётся как решение задачи

$$P_V \boldsymbol{\eta} = \arg \inf_{\boldsymbol{g} \in V} \|\boldsymbol{g} - \boldsymbol{\eta}\|,$$

наилучшего приближения результата наблюдения $\boldsymbol{\eta}$ векторами формы V.

Соотношением сигнал-шум будем называть отношение квадратов нормы информативной части данных и нормы неинформативной части данных об объекте:

$$SNR_V(\eta) = \frac{\|P_V \boldsymbol{\eta}\|^2}{\|\boldsymbol{\eta} - P_V \boldsymbol{\eta}\|^2} : \aleph \to [0, +\infty)$$

Для анализа качества модели и её соответствия результатам эксперимента более удобен коэффициент морфологической корреляции, связанный с соотношением сигнал-шум выражением, аналогичным классическому, связывающему коэффициент корреляции и соотношение сигнал-шум [7]:

$$\rho_V^2(\eta) = \frac{SNR_V(\eta)}{SNR_V(\eta) + 1} : \aleph \to [0, 1]$$

где $\rho_V(\eta)$ — коэффициент морфологической корреляции. Заметим, что зависимость $\rho_V^2(\eta)$ от $SNR_V(\eta)$ является монотонно возрастающей.

Вычисление меры различия смещённых сигналов $D_{\boldsymbol{\xi}}(\boldsymbol{\Delta t})$, используемой в (11), подразумевает построение оператора проецирования на множество сигналов (параметризованное $\boldsymbol{\Delta t}$), которые могут быть получены в соответствии с моделью (1)–(4) [5, 6] при заданном векторе $\boldsymbol{\Delta t}$ временных сдвигов в различных каналах. Положим, таким образом, $\mathfrak{dt}(\boldsymbol{\Delta t}) = \rho_{\boldsymbol{\Delta t}}(\boldsymbol{\xi})$.

Априорная информация о временных задержках, задаваемая выражением (9), также может быть интерпретирована в терминах «формы»: в этом случае формой назовем множество векторов (13)–(14)

$$\tilde{\mathbb{T}}^e = \left\{ \boldsymbol{\Delta t} \colon P_{\tilde{\mathbb{T}}^e} \boldsymbol{\Delta t} = \boldsymbol{\Delta t} \right\} \subset \mathbb{T}^e,$$
$$P_{\tilde{\mathbb{T}}^e} \boldsymbol{\Delta t} = AA^- \boldsymbol{\Delta t}.$$

Поэтому можно положить в (14)

$$\mu_1(\|(AA^- - I)\boldsymbol{\Delta t}\|^2) = \rho_{\tilde{\mathbb{T}}^e}(\boldsymbol{\Delta t}). \tag{15}$$

Тогда оценка максимальной возможности (14) соответствует оценке максимального соотношения сигнал-шум или максимального коэффициента морфологической корреляции:

$$\boldsymbol{\Delta t}_0 = \arg \sup_{\boldsymbol{\Delta t}} \Upsilon(\boldsymbol{\Delta t}),$$
$$\Upsilon(\boldsymbol{\Delta t}) = \min \left\{ \rho_{\boldsymbol{\Delta t}}(\boldsymbol{\xi}), \rho_{\tilde{\mathbb{T}}^e}(\boldsymbol{\Delta t}) \right\}.$$

Так как Υ имеет известный физический смысл соотношения сигнал-шум или коэффициента морфологической корреляции, характеризующего адекватность модели, можно построить интервальные оценки величин $\boldsymbol{\Delta t}$, $\boldsymbol{q}$, φ и c_{hor}, гарантирующие заданный уровень адекватности модели.

В ряде случаев, когда соотношение шага сетки (зависящего от частоты дискретизации сигналов по времени), на которой вычисляются задержки $\boldsymbol{\Delta t}$ и расстояний между микрофонами $\boldsymbol{r}_{j_1,j_2}$ допускает достаточное число векторов $\boldsymbol{\Delta t}$ (координаты которых пропорциональны шагу дискретизации по времени), при которых $\|(AA^- - I)\boldsymbol{\Delta t}\|^2 < \varepsilon$, можно также сделать переходную возможность в (9) чёткой, положив

$$\mu_1(x) = \begin{cases} 1, & |x| < \varepsilon, \\ 0, & \text{иначе.} \end{cases} \tag{16}$$

Такой подход позволяет дополнительно уточнить решение.

Вычислительный эксперимент

Для трёх микрофонов, в вершинах треугольника со сторонами 30, 30 и 45 метров (см. рис. 1) было произведено моделирование прихода инфразвукового сигнала (см рис. 2) в соответствие с моделью (1)–(4), имевшего исходную форму $|\frac{\sin(t)}{t}|$. Погрешности измерения моделировались нормально распределённым шумом с нулевым математическим ожиданием и стандартным отклонением, равным 0,1. Азимутальный угол был положен равным $\varphi = -120°$, скорость следа — равной $c_{hor} = 343$ м/с.

На рисунке 3 показана зависимость распределения возможности $\mathfrak{dt}(\boldsymbol{\Delta t}) = \rho_{\boldsymbol{\Delta t}}(\boldsymbol{\xi})$ от азимутального угла и угла наклона, соответствующих вектору задержек $\boldsymbol{\Delta t}$.

На рисунке 4 показана зависимость распределения возможности $\Upsilon(\boldsymbol{\Delta t})$, учитывающее априорную информацию о сигнале в форме (15).

На рисунке 5 показана зависимость распределение возможности $\Upsilon_{det}(\boldsymbol{\Delta t})$, учитывающее априорную информацию о сигнале в форме (16) при значении $\varepsilon = 1 \cdot 10^{-8}$. Заметно существенное сокращение

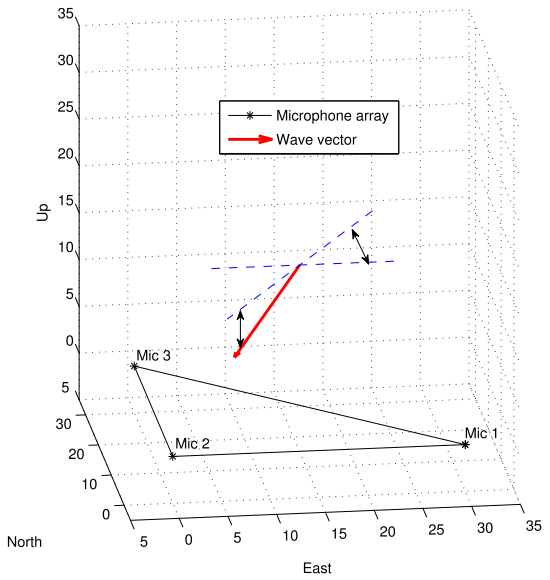


Рис. 1. Схема эксперимента по регистрации сигнала.

меры подмножества пар значений (φ, c_{hor}), для которых значение морфологической корреляции превышает, например, 0,3. Также учёт априорной информации привел к тому, что оценка максимальной возможности для φ была уточнена $-133°$ до $-121°$.

Выводы

В работе предложен метод построения оценок максимальной возможности для азимутального направления прихода и скорости следа инфразвуковой волны по сигналам, получаемым с пространственно разнесённых микрофонов.

Литература

[1] *Пытьев Ю. П., Чуличков А. И.* Морфологические методы анализа изображений. — Москва: Физматлит, 2010. — 336 с.

[2] *J. V. Olson, C. A. L. Szuberla* The Least Squares Estimation of the Azimuth and Velocity of Plane Waves // Inframatics. — June 2004. — № 06 — С. 9–12.

[3] *Пытьев Ю. П.* Возможность как альтернатива вероятности. Математические и эмпирические основы. — Москва: Физматлит, 2007. — 464 с.

[4] *Пытьев Ю. П.* Методы математического моделирования измерительно-вычислительных систем. — Москва: Физматлит, 2002.

[5] *Демин Д. С., Чуличков А. И.* Сравнение формы нескольких сигналов, порожденных нелинейным монотонным преобразованием из неизвестного прообраза и оценивание параметров их формы //тезисы 8-й Международной конференции «Интеллектуализация обработки информации», 2010".

[6] *Чуличков А. И., Куличков С. Н., Демин Д. С.* Морфологический анализ инфразвуковых сигналов в акустике. — Москва: Новый Акрополь, 2010. — 132 с.

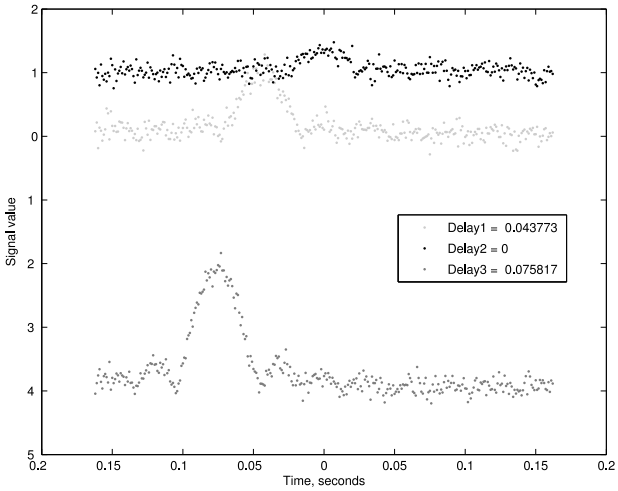


Рис. 2. Сигналы, анализируемые в вычислительном эксперименте.

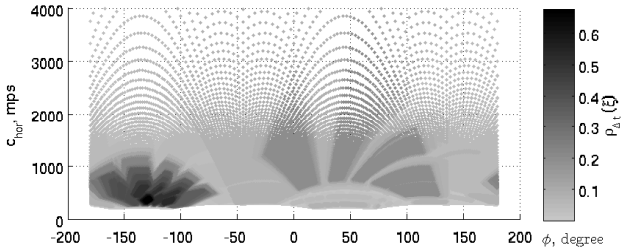


Рис. 3. Зависимость $\rho_{\Delta t}(\xi)$ от φ и c_{hor}, полученная в вычислительном эксперименте.

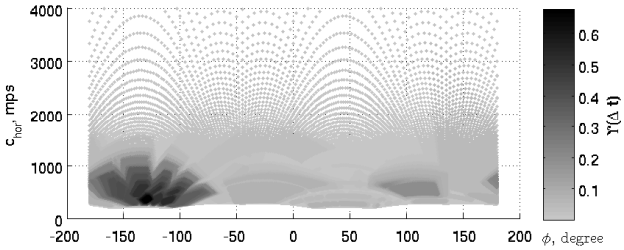


Рис. 4. Зависимость $\Upsilon(\Delta t)$ от φ и c_{hor}, полученная в вычислительном эксперименте.

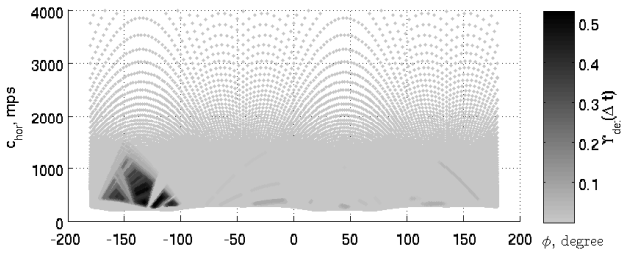


Рис. 5. Зависимость $\Upsilon_{det}(\Delta t)$ от φ и c_{hor}, полученная в вычислительном эксперименте.

[7] *J. Benesty, J. Chen, Y. Huang* Microphone Array Signal Processing. — Berlin: Springer-Verlag, 2008.

Моделирование произношения в речевой технологии*

Чучупал В. Я.
chuchu@ccas.ru
Москва, Вычислительный центр им. А. А. Дородницына РАН

Важным этапом моделирования речевого потока при автоматическом распознавания речи является вычисление фонемных транскрипций слов и словосочетаний. Обычно это осуществляется посредством использования специальных программ — автоматических транскрипторов, которые используют экспертные правила для определения фонем по орфографической записи. В статье представлены алгоритмы автоматического транскрибирования произвольных текстов на русском языке, которые основаны на анализе обучающей выборки методом построения деревьев решений, и, таким образом, не предполагают наличия экспертных знаний.

Задача произносительного моделирования

Основной единицей моделирования речевого потока в современной технологии распознавания речи является т. н. «фон»: контекстно-зависимая реализация фонемы языка. Набор параметров или образ, который соответствует произнесению фона, порождается акустической моделью, как правило, скрытой марковской моделью (СММ). Образ слова при этом определяется как последовательность образов составляющих это слово фонов, а образ высказывания — как последовательность образов слов и межсловных промежутков.

Для корректной работы систем распознавания и синтеза речи нужно знать произносительные (фонемные) транскрипции слов, входящих в словарь системы. Транскрипции можно найти в специальных словарях, однако использование этих словарей в системах распознавания проблематично. Фонемные транскрипции в словарях имеют т. н. каноническую форму, т. е. соответствуют одной определенной норме произношения (московской, петербургской). В словарях отсутствуют фонемные транскрипции многих слов. Кроме этого, при обработке слитной речи существенным оказывается наличие моделей межсловных переходов, ибо артикуляции конца слова и начала последующего слова взаимосвязаны [1]. Наконец, для разговорной речи нужно учитывать вариативность фактического произношения, которая приводит к существованию нескольких фонемных транскрипций у одного и того же слова.

Под произносительным моделированием в речевой технологии обычно понимают набор моделей и методов для определения структуры моделей звуков и фонемных транскрипций слов [2]. При использовании канонических произносительных моделей слов (а это фактически стандартно для существующих систем распознавания речи) перевод текстовой записи в последовательность марковских моделей звуков осуществляется в два шага:

Преобразование T2P («буквы → фонемы»), расчёт фонемных транскрипций по тексту.

Преобразование P2P («фонема → фон»), выбор последовательности СММ по фонемной транскрипции.

Настоящая работа посвящена алгоритмам осуществления первого шага, преобразования T2P. Естественным, альтернативным к использованию словарей способом определения фонемных транскрипций является является подход к построению преобразования T2P на основе синтеза набора правил, т. е. экспертных знаний. Этот подход успешно используется при создании систем синтеза речи. Для некоторых современных языков (например, финского) фонемная транскрипция фактически совпадает с орфографической записью, что упрощает задачу, однако для других, например, русского, орфографическая запись только приблизительно соответствует транскрипции. Один из наиболее известных и удачных транскрипторов для русского языка создан на филологическом факультете МГУ [3]. Он содержит около 600 правил на специальном языке описания преобразований, которые определяют контекстно-свободную грамматику для преобразования T2P.

Как экспертные системы методы преобразования T2P на основе правил имеют свои достоинства и недостатки, последние во многих случаях оказываются существенными. Правила могут оказаться противоречивыми или недостаточными, поэтому для сопровождения системы нужен специалист. Для систем распознавания важным обстоятельством является то, что имеющиеся на сегодняшний день T2P определяют по орфографической записи только один, канонический вариант произнесения.

Альтернативным методом реализации T2P может стать использование методов анализа данных, тем более, что на сегодняшний день для многих языков собраны и аннотированы корпуса данных, включающие орфографическую и фонемную транскрипции.

Можно ли, имея в наличии достаточно большую выборку данных, которые состоят из про-

Работа выполнена при финансовой поддержке РФФИ, проект №11-01-00900а.

изнесенных текстов и соответствующих им фонемных транскрипций, построить алгоритм автоматического транскрибирования, который бы был в некотором смысле «эквивалентен» или превосходил по эффективности системы транскрибирования, основанные на экспертных правилах? В качестве возможного ответа на это вопрос в работе приведен алгоритм автоматического преобразования T2P.

Установление соответствия между орфографической и фонемной записями

Определим следующие величины.

1. Алфавит букв $\mathcal{A} = \{t_i\}$, $i = 1, \dots R$.

2. Алфавит фонем $\mathcal{P} = \{p_i\}$, $i = 1, \dots S$.

3. Набор попарных расстояний между буквами и фонемами $\{d_{tp}(t_i, p_j)\}$.

4. Набор попарных расстояний между фонемами $\{d_{pp}(p_i, p_j)\}$.

Расстояния d_{tp} и d_{pp} основаны на степени акустического сходства между фонемами, они определяются эвристически и могут быть заданы весьма приблизительно, например, принимать значения, соответствующие полному совпадению, неопределенности и полному различию.

Пусть заданы две конечные последовательности, графем $T_1^N = t_1, t_2, \dots, t_N$ и фонем $P_1^M = p_1, p_2, \dots, p_M$.

Назовем соответствием между этими последовательностями любой набор целочисленных пар индексов $(\varphi_t, \varphi_p) = \{(\varphi_t(i), \varphi_p(i))\}$, $i = 1, \dots, L$, который удовлетворяет условиям:

Граничные условия:

$$\begin{cases} \varphi_t(1) = 1, & \varphi_p(1) = 1, \\ \varphi_t(L) = N, & \varphi_p(L) = M. \end{cases}$$

Условия монотонности

$$\begin{cases} \forall i, j : i > j \to \varphi_t(i) \geqslant \varphi_t(j), \\ \forall i, j : i > j \to \varphi_p(i) \geqslant \varphi_p(j). \end{cases}$$

Условия непрерывности

$$\begin{cases} \forall i, \ : \varphi_t(i+1) - \varphi_t(i) \leqslant 1, \\ \forall i, \ : \varphi_p(i+1) - \varphi_p(i) \leqslant 1. \end{cases}$$

Определим меру близости $d_\varphi(T_1^N, P_1^M)$ между T_1^N и P_1^M для $\{\varphi_t, \varphi_p\}$ как

$$d_{(\varphi_t, \varphi_p)}(T_1^N, P_1^M) = \sum_{i=1}^{L} d_{tp}(t_{\varphi_t(i)}, p_{\varphi_p(i)}). \qquad (1)$$

Пусть Φ — множество всех соответствий $\{(\varphi_t, \varphi_p)\}$ между T_1^n и P_1^M. Определим меру близости $d(T_1^N, P_1^M)$ между последовательностями T_1^n и P_1^M как

$$d(T_1^N, P_1^M) = \min_{(\varphi_t, \varphi_p) \in \Phi} d_\varphi(T_1^N, P_1^M). \qquad (2)$$

Будем считать оптимальным такое соответствие между последовательностями, мера близости (1) для которого равна мере близости (2) между ними.

Аналогичным образом определим оптимальное соответствие между конечными последовательностями из фонем.

Для вычисления расстояния и определения оптимального соответствия между двумя символьными последовательностями используем алгоритм динамического программирования:

Алгоритм 1. Определение соответствия между двумя символьными последовательностями.

Вход: Последовательности $T_1^N = t_1, \dots, t_N$ и фонем $P_1^M = p_1, \dots, p_M$

Выход: Соответствие (φ_t, φ_p) между T_1^N и P_1^M

1: инициализация: $\forall m : d(T_1^0, P_1^m) = 0$;
$\forall n : d(T_1^n, P_1^0) = 0$;
$\forall n : L(n, m) = 0$;
$d(T_1^1, P_1^1) = 0$;
$L(1, 1) = 1$;

2: **для всех** $n = 1, \dots N$

3: **для всех** $m = 1, \dots M$

4: Выберем
$d(T_1^i, P_1^j) = \min(d(T_1^n, P_1^{m-1}), d(T_1^{n-1}, P_1^{m-1}))$;

5: вычислим меру близости
$d(T_1^n, P_1^m) = d_{tp}(t_n, p_m) + d(T_1^i, P_1^j)$;

6: длину оптимального пути L
$L(n, m) += L(i, j)$;

7: и значения соответствия
$l = L(n, m); \psi_t(l) = i, \ \psi_p(l) = j$;

8: Восстановим функцию соответствия:
$L = L(N, M)$; $\varphi_t(L) = N$; $\varphi_p(L) = M$;

9: **для всех** $i = L - 1, \dots 1$
$\varphi_t(i) = \psi_t(i + 1)$;
$\varphi_p(i) = \psi_p(i + 1)$;

Применим этот алгоритм для установления оптимального соответствия (φ_t, φ_p) между последовательностью букв T_1^N и последовательностью фонем P_1^M.

Удобно на основе соответствия (φ_t, φ_p) сделать однозначную функцию, φ, выполняющую ту же функцию, что и соответствие. Для этого, в случаях вставки, когда графема t соответствует двум фонемам p_i, p_j, т. е. существует несколько пар типа (t, p_i) (t, p_j), заменяем вторую пару на $(0, p_j)$, где символ 0 — пустая графема. Это эквивалентно добавлению графемы в последовательность графем. Аналогично, в случаях пропуска, когда две графемы t_1, t_2 соответствуют одной фонеме, т. е. встре-

чаются пары (t_1, p) и (t_2, p), заменяем вторую пару на $(t_2, 0)$. В результате получим однозначное соответствие между последовательностью графем обучающей выборки и соответствующей последовательностью фонем.

Для каждого вхождения графемы можно определить набор признаков, связанных со свойствами как самой этой графемы, так и контекста, в котором она встретилась. В принципе можно использовать любые признаки; априори полезными будут те, которые применяются при построении экспертных правил преобразования Т2Р.

Например, для графемы t_i такими признаками будут:

- идентификатор графемы t_i,
- идентификаторы предшествующих и последующих графем $t_{i-1}, t_{i-2}, t_{i+2}, t_{i+2}$,
- код позиции по отношению к ударению (предударный, заударный, ударный),
- код позиции по отношению к началу-концу слова (начало, конец, середина),
- код позиции по отношению к началу-концу фразы (начало, конец, середина),
- идентификатор фонемы p_i,
- идентификатор предшествующей фонемы $p_{t_{i-1}}$,
- идентификатор слова.

Следующий алгоритм на основе полученной функции соответствия определяет дерево бинарных решений для процедуры вычисления автоматической фонемной транскрипции по тексту.

Алгоритм синтеза дерева решений

Для каждой графемы $t \in \mathcal{A}$ создаём отдельное дерево, корень которого пометим идентификатором этой графемы. Занесем в этот корень вектора признаков, полученные для всех появлений графемы t в корпусе данных. Обозначим полученное множество прецедентов через $X = x_1^t, \ldots, x_N^t$. Таким образом, каждый корень (а далее, по построению и каждая вершина и лист) дерева ассоциируется с множеством наблюдений графемы, представленных векторами признаков.

Пусть вычислена и сделана однозначной функция соответствия между последовательностью графем T_1^N обучающей выборки и соответствующей последовательностью фонем P_1^N, так что графеме t_i соответствует фонема p_i. Определим меру однородности L для множества наблюдений X как:

$$L(X) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} d_{pp}(p_i, p_j). \qquad (3)$$

Составим список всевозможных допустимых бинарных (с ответами «true» и «false») вопросов к признакам графем. Вопросы можно перебрать все, т. к. число признаков и их значений невелико. На практике, однако, чтобы избежать больших вычислительных затрат, используются вопросы, существенно опирающиеся на экспертный опыт (например, «Совпадает ли предыдущая графема t_{i-1} с некоторой заданной?») При этом в качестве заданной перебираются все возможные графемы, широкие фонетические категории фонем, близких к данным графемам по мере d_tp, а также возможные варианты их контекстов, позиции по отношению к ударению («Соответствует ли графема гласной? Если да, соответствует она ударной гласной либо заударной или предударной?») и т. п.

Пусть $Q_1^R = q_1, \ldots, q_R$ — список составленных вопросов, $q(x)$ — ответ на вопрос q для образа x графемы t.

Алгоритм 2. Построение дерева решений для преобразования Т2Р.

Вход: графемы $t_1, \ldots, t_N$ с признаками $x_1, \ldots, x_N$, список вопросов $\{q_i\}, i = 1, \ldots, R$

Выход: Деревья решений для преобразования Т2Р

1: **для всех** $t \in \mathcal{A}$
2: создаём дерево с корнем t, множеством листьев $L_t = \{t\}$, множеством прецедентов в каждом листе t — $X_t = \varnothing$
3: **для** $i = 1, \ldots, N$
4: **если** x_i — вектор признаков графемы t, **то** занесем x_i в X_t
5: **для всех** $t \in L_t$
6: **для всех** $q_i : i = 1, \ldots, R$
7: создаём временно листья $t(q, l)$ и $t(q, r)$, $X_{t(q,l)} = X_{t(q,l)} = \varnothing$
8: **для всех** $x \in X_t$
9: **если** $q_i(x) = 1$ **то**
10: $x \to X_{t(q,l)}$;
11: **иначе**
12: $x \to X_{t(q,r)}$;
13: Вычислим по (3) изменение L от замены t на $t(q, l)$ и $t(q, r)$: $dL(q, t) = L(t) - (L(t(q, l)) + L(t(q, r)))$
14: Определим dL_{max} и пару $(\hat{t}, \hat{q})$: $dL_{max} = \max\limits_{(t,q)} dL(t, q)$; $(\hat{t}, \hat{q}) = \arg\max\limits_{(t,q)} dL(t, q)$;
15: Запомнить вопрос $\hat{q}$ в листе $\hat{t}$
16: Листья $t(\hat{q}, l), t(\hat{q}, r)$ сделать постоянными
17: **если** $L_{max} < \text{Threshold}_{dL}$ **то** Переход на 2.
18: Выполнить шаги (6–20) для всех корней t.

Если порог Threshold_{dL} мал, то по завершении работы алгоритма в листах дерева будут находится наблюдения, описывающие совпадающие по расстоянию d_{pp} фонемы. В другом случае после завершения работы алгоритма в каждом листе t дерева будет находиться набор прецедентов, которые определяют несколько фонем, которым может соответ-

ствовать графема с данными признаками. В этом случае преобразование T2P неоднозначно и выбор фонемы можно делать несколькими способами, например, выбирать одну наиболее вероятную фонему, либо порождать несколько возможных гипотез (в этом случае фонемная транскрипция будет представлена в нескольких вариантах).

Обозначим через p_t наиболее вероятную фонему для листа t некоторого построенного по алгоритму (2) дерева решений. Искомое преобразование T2P в таком случае осуществляется прохождением дерева решений снизу вверх, т. е. в соответствии со следующим алгоритмом:

Алгоритм 3. Преобразование T2P.

Вход: Дерево решений, с множеством листьев $\mathcal{L}$ графема t и её вектор признаков x

Выход: Фонема p, соответствующая t

1: Выбрать в качестве текущей вершины t корень дерева, который имеет идентификатор данной графемы
Пусть p_t — наиболее вероятная фонема для вершины t;
2: **если** $t \in \mathcal{L}$ **то**
3: $p = p_t$; Выход;
4: **иначе**
5: **повторять**
6: выберем дочерние для t вершины $t_{left}t_{right}$ и разделяющий вопрос $\hat{q}_t$
7: **если** $\hat{q}_t(x) = true$ **то**
8: $t_{left} -> t$;
9: **иначе**
10: $t_{right} -> t$;
11: **пока** $t \notin \mathcal{L}$;
12: $p = p_t$; Выход.

Выводы

В статье представлены алгоритмы автоматического фонемного транскрибирования произвольных текстов на русском языке, которые основаны на анализе обучающей выборки методом построения и последующего использования деревьев бинарных решений.

Литература

[1] *Hamalainen A., Bosch L., Boves L.* Modelling pronunciation variation with singlepath and multipath syllable models: Issues to consider // Speech Communication, 2009. — Vol. 51, Issue 2. — Pp. 130–150.

[2] *Caraclar M., Khudanpur S.* Pronunciation change in conversational speech and its implications for automatic speech recognition // Computer Speech and Language, 2004. — Vol. 18, Issue 4. — Pp. 375–395.

[3] *Кривнова О. Ф., Захаров Л. М., Строкин Г. С.* Многофункциональный автоматический транскриптор русских текстов // Труды Международного конгресса исследователей русского языка, Москва, 2001.

Разработка алгоритмов распознавания эмоционального состояния человека по паралингвистическим особенностям речи

Кальян В. П.
vkalyan@mail.ru
Москва, Вычислительный Центр им. А. А. Дородницына РАН

Описывается эксперимент по созданию алгоритмов автоматического распознавания эмоционального состояния человека на основании анализа некоторых паралингвистических элементов речи по динамике спектра, высоты и интенсивности звука в речевом сигнале.

Исследователи уже несколько десятилетий разрабатывают аналитические методы распознавания эмоций в речи для определения искренности говорящего и выявления правдивости сказанного. Задача эксперта заключается в распознавании и маркировке тех моментов в высказывании, где проявляется волнение. Оно в дальнейшем может быть интерпретировано как нервозность, раздражение, неуверенность, страх, презрение, недоверие, сомнение, возмущение, гнев, обида, горе, или наоборот — восхищение, радость, удивление, признательность, интерес, надежда, удовлетворение.

Наиболее информативными параметрами для определения темпоральной организации речи являются долготы (длительности) звуков. Акцентуация на уровне слова и фразы в значительной степени достигается увеличением длительности звуков; в частности, лингвисты отмечают увеличение длительности в ударных гласных и «растяжение» некоторых начальных, пред- и послеударных (сонорных и щелевых) согласных в слове как одно из возможных проявлений эмоций в речи.

В данной работе мы остановились лишь на некоторых тональных и темпоральных признаках акцентуации, параметрах оценки темпа и ритма, полученных при исследовании ударных и безударных гласных.

Часть содержательных признаков мы полагаем рассмотреть в следующих работах.

Параметры

В поле нашего зрения в настоящем исследовании оказались такие параметры, как время, высота и интенсивность звука, траектории максимумов первых трёх формант, характеристики частотного спектра в отдельных сегментах или контрольных точках.

Время представлено последовательностью отсчётов с частотой 20 мсек. Временные отсчёты функций $A(t_i)$ и $P(t_k)$ получены с одним и тем же шагом $t = 0.02$ секунд так, что для каждого $k = i = 1, \ldots, n$ отсчёты A и P синхронны.

Мелодический контур мы получали с помощью преобразования траектории основного тона из линейной шкалы частот W в логарифмическую с основанием 2, по формуле:

$$P_t = 12 * \log_2(W_t) + C,$$

где W_t — частота основного тона в момент времени t, C — некоторая константа; величина P_t отображает высоту звука в музыкальном восприятии в момент времени t. Значение 52 по оси P, например, соответствует высоте ноты «до» первой октавы по стандарту MIDI. Значение шага между ближайшими целыми значениями высоты звука (52, 53, 54 и т. д.) соответствует полутону темперированного строя диатоники. Точкам разрыва для простоты работы алгоритма присваивается нулевое значение.

Описание эксперимента

Для экспериментов был взят массив телевизионных и радиоинтервью, в которых основные «базовые» эмоции собеседников проявлялись наиболее очевидным и ярким образом.

Построенные алгоритмы предполагалось испытывать и совершенствовать в последующих комплексных аудио визуальных исследованиях менее очевидных и спорных случаев проявления эмоций в речи.

На основании спектральных и амплитудо-высотных параметров речевого сигнала предполагалось произвести:

1. вычисление локальных экстремумов функций высоты и интенсивности звукового сигнала;
2. определение границ аллофонов и пауз на основании их спектродинамических характеристик;
3. вычисление функций распределения плотности вероятности для функции высоты звука, длительности аллофонов;
4. установление темпоральных соотношений сегментов;
5. выявление признаков, характеризующих акцентуацию;
6. расчёт характеристик для определения ритма и темпа речи;
7. определение признаков для распознавания основных интонационных конструкций;
8. определение значимых признаков для определения концовки фразы;
9. распознавание дрожания/пропадания голоса;

10. распознавание элементов неуверенной речи — заикания, мычания, блеяния;
11. определение признаков эмоций в речевом высказывании;
12. транскрипцию выявленных эмоциональных признаков.

Сегментация

Для распознавания отдельных звуков и измерения их длительности мы применили метод двухуровневой сегментации речевого сигнала. На первом уровне сегментация выполнялась по экстремумам кривых интенсивности (алгоритм был подробно описан в предыдущих работах автора, см. [5–7], на втором происходило уточнение границ звуков по наличию основного тона, траекториям формант и характеристикам высоких, средних и низких частот динамического спектра.

Из найденных на первом этапе $P_{\max}$, $P_{\min}$, $A_{\max}$, $A_{\min}$ были сформированы два массива

$$\mathtt{P}_{\min\max}(t_{P_{\min}}, P_{\min}, t_{P_{\max}}, P_{\max}),$$
$$\mathtt{I}_{\min\max}(t_{A_{\min}}, A_{\min}, t_{A_{\max}}, A_{\max})$$

значений локальных экстремумов и соответствующих им временных значений

$$t_{P_{\min}}(k),\ t_{P_{\max}}(k),\ t_{A_{\min}}(i),\ t_{A_{\max}}(i).$$

За основу первичного разбиения брались точки временной последовательности столбца $t_{A_{\min}}$, т. к. они предположительно должны соответствовать либо центрам согласных звуков, либо начальным или конечным точкам пауз, и, возможно, с меньшей вероятностью, минимумам внутри акцентируемых гласных.

Маркировка и структуризация

На втором этапе производилась идентификация и уточнение границ аллофонов, пауз и ряда экстралингвистических элементов речи. На обучающей выборке с применением табличных данных проводилась ручная разметка основных речевых событий и установление их признаков, являющихся основанием для сегментации, маркировки, структуризации сегментов — объединения их последовательности в синтагмы и фразы.

На участках между соседними экстремумами с точками этой последовательности анализировались основной тон, характеристики спектра и траектории формант и уточнялись границы аллофонов. Границы сонорного участка определялись по присутствию основного тона в точке k и его отсутствию в соседней, т. е., k — пограничная точка, если

$$P(t_k) > 0 \text{ и } (P(t_{k-1}) = 0 \text{ или } P(t_{k+1}) = 0).$$

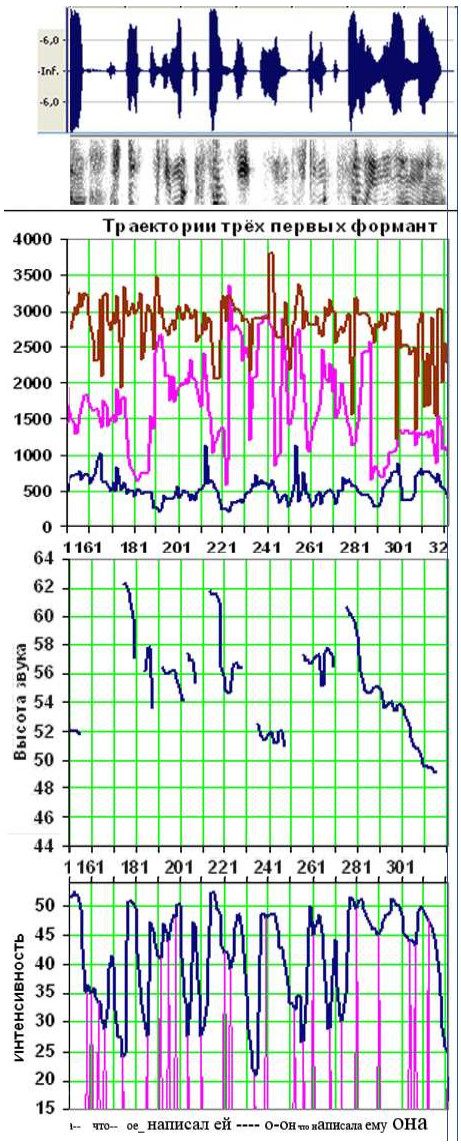


Рис. 1. Синхронизированные графики звукового сигнала, динамического спектра, траекторий трёх первых формант, высоты звука, интенсивности звука с отметками локальных минимумов и масштабированной подтекстовкой для средней части высказывания «Что написал».

Идентификация и уточнение границ аллофонов в сегменте производилось по табличным признакам, содержащим частотное положения пиков формант для гласных и сонорных согласных; соотношение значений интенсивности отдельных спектральных полос — для щелевых согласных; то же в динамике — для взрывных. Так, например, граница между сонорной согласной и гласной уточнялась по порогам интенсивности и положениям формант.

На рисунке 1 изображены графики функций (сверху вниз расположены одна под другой: звуковая волна, спектрограмма, траектории трёх первых формант, высота звука, интенсивность) для речевого высказывания «Что написал». Полный текст

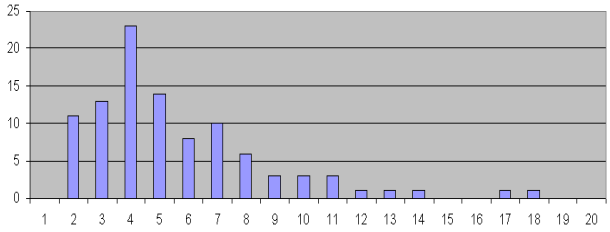


Рис. 2. Плотность вероятности длительности сегментов в высказывании «Что написал». Ось X – шкала длительностей, цена деления — 20 мсек, Y — плотность вероятности.

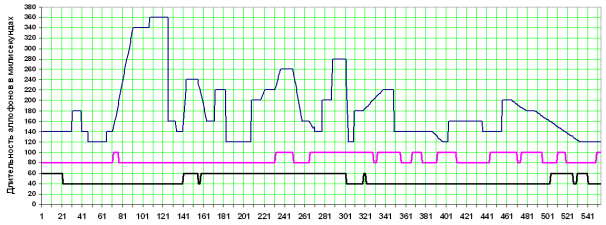


Рис. 3. График зависимости длительностей гласных, сонорных, щелевых и взрывных согласных в высказывании «Что написал» от времени раздельно.

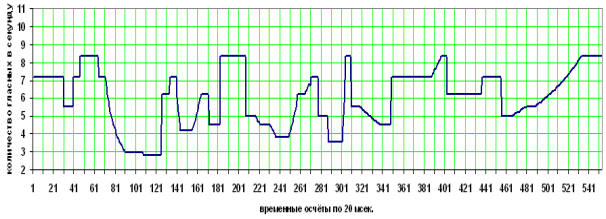


Рис. 4. Динамика темпа произнесения гласных.

отображён в подтекстовке графика интенсивности (самая нижняя функция). Буквы текста расположены под соответствующим этим звукам экстремумам функции интенсивности. Для наглядности и сохранения непрерывности текста звуки более протяженные изображены более крупными буквами, короткие — более мелкими.

Тональный и темпоральный анализ

После уточнения границ аллофонов был сформирован массив $t_{segm}\,(i, t_{in}, t_{out}, \Delta t, P_i, \text{type})$, где i — номер сегмента, t_{in}, t_{out} — его пограничные точки, Δt — длительность в единицах временных отсчётов по 20 миллисекунд, P_i — средняя высота сегмента, type — целое число, идентификатор типа: 1 — гласный, 2 — согласный не взрывной (щелевой, сонорный), 3 — согласный взрывной, 4 — пауза.

По столбцу Δt вычислялась функция распределения плотности вероятности длительности аллофонов в высказывании «Что написал». На рис. 2 изображена гистограмма длительностей сегментов в данном речевом фрагменте, где наглядно отображён диапазон изменения «длиннот» звуков.

Экспертная оценка показала, что длительность гласных в данном речевом фрагменте попадает в диапазон 3–18 отсчётов, что соответствует диапазону длительностей 40–360 миллисекунд.

Безударные гласные внутри слова имеют длительность в диапазоне от 40 до 140 миллисекунд, ударные и начальные гласные в слове — от 120 до 360 миллисекунд.

Гласные, определяющие фразовую акцентуацию, имеют два максимума и состоят из двух сегментов первичной разбивки, каждый длительностью порядка 100–240 миллисекунд. Здесь имеет место и удлинение предударных щелевых или сонорных согласных, некоторые из них имеют по два минимума кривой интенсивности и также состоят из двух сегментов первичной разбивки.

Длительность аллофонов как функция от времени в наглядном (в отношении ритма и темпа виде — разделённом на три группы данных — для гласных и согласных по признаку type) в том же высказывании представлена на рис. 3.

Кривая, соответствующая длительностям гласных, представляет динамику величины обратной темпу речи, а соотношения между всеми тремя зависимостями от времени дают возможность анализировать ритмические конструкции разных уровней.

Используя интерполированную функцию длительности гласных от времени, в речевом высказывании можно построить в первом приближении график динамики темпа произнесения гласных (см. рис. 4).

Для выявления признаков фразового акцента и построения алгоритма его распознавания дополним темпоральную информацию тональной и рассмотрим закономерности амплитудо-высотных соотношений в речевом фрагменте.

На рис. 5 приведена гистограмма функции распределения плотности вероятности высоты звука в высказывании «Что написал».

Звуковые высоты данного речевого фрагмента находятся в диапазоне от высоты 45 (соответствует «фа» малой октавы) до высоты 62 (соответствует «си бемоль» первой октавы) и распадается на два поддиапазона: небольшой нижний — от высоты 45 до высоты 46, и основной верхний — от высоты 49 до высоты 62. Кроме того, можно выделить по крайней мере три подзоны основного диапазона 49–55, 55–60 и 60–62.

Экспертная оценка показала, что все фразовые акценты находятся в подзоне 49–55 и основной акцент предложения целиком расположен в зоне 45–46. Таким образом, высота звука фразовых акцентов в данном примере оказывается в самых нижних зонах и подзонах общего высотного диапазона звучания фразы. Можно ожидать, что в других

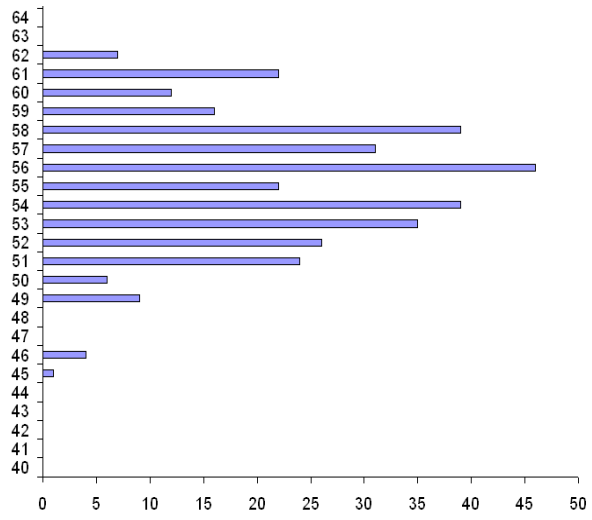


Рис. 5. Функция распределения плотности вероятности высоты звука в высказывании «Что написал».

речевых примерах акцентуация в звуковысотной области будет проявляться прямо противоположным образом — оказываться в самой верхней части общего диапазона. Здесь можно сослаться на мнение многих исследователей, которые обращали внимание на то, что в эмоциональной речи может присутствовать контрастно-регистровое интонирование. Здесь важно зафиксировать сам факт контраста звуковысотных диапазонов разных участков эмоционального речевого высказывания и особое периферийное положение звуковысотной зоны концовок фраз и предложений.

Темпоральный анализ ранее показал, что ударные гласные в словах, на которые приходятся фразовые акценты, имеют длительности от 240 до 360 миллисекунд. Таким образом, для установления фразовых акцентов данного речевого фрагмента проявились отчётливые признаки — периферийные зоны в функциях распределения плотности вероятности длительности аллофонов и высоты звука в них.

Обозначим зону длительностей от 240 до 360 миллисекунд (см. рис. 5), в которую попадают только длительности гласных, на которые приходится фразовые акценты, как Z_t, а звуковысотную зону, в которой находятся высоты звуков концовок фраз — как Z_p, тогда гласная может быть распознана как фразовый акцент (принадлежит некоторому множеству $AccentPhrase$), если выполняются следующие условия:

$$t_{segm}(i=1) \in AccentPhrase, \text{ если } \Delta t \in Z_t \text{и } P_i \in Z_p. \quad (1)$$

Рассмотрим нахождение признаков для границ и акцентов фраз (синтагм). Для этого произведем сегментацию данного высказывания на фразы (синтагмы) на основании признаков (1). Для каждой фразы аналогичным образом вычисляем зоны на границе диапазона длительности и высоты Y_t и Y_p и устанавливаем принадлежность обоим периферийным фразовым зонам ударных гласных для каждой гласной во фразе:

$$t_{segm}(i=1) \in AccentSyntagma, \text{ если } \Delta t \in Y_t \text{ и } P_i \in Y_p.$$

Заключение

В докладе описан эксперимент по созданию алгоритмов автоматического распознавания эмоционального состояния человека на основании анализа некоторых паралингвистических элементов речи. В настоящей работе описаны промежуточные результаты построения алгоритмов распознавания, выделены параметрические признаки лишь по пяти содержательным — удлинению ударных гласных, наличию контрастно-регистрового интонирования, изменению темпа речи. В дальнейших работах предполагается продолжение работ по построению алгоритмов на основе содержательных признаков и комплексное аудио-визуальное исследование в данной области.

Литература

[1] *Кальян В. П.* Музыка, речь и компьютер. — Москва: ВЦ РАН, 1998. — 40 с.

[2] *Златоустова Л. В.* Общая и прикладная фонетика, 2-е изд., дополн. и перераб. — Москва: Изд-во МГУ, 1997. — 416 с.

[3] *Златоустова Л. В., Банин А. А.* Об иерархии уровней ритмического компонента русской речи // Вестник МГУ, сер. 9: Филология. — 1978. — Вып. 2. — С. 35–43.

[4] *Кальян В. П.* Интонационный анализ народных исполнителей // Методы и модели распознавания речи, Москва: ВЦ РАН им. А. А. Дородницына, 2007. — С. 48–56.

[5] *Кальян В. П.* Об алгоритмах сегментации для системы автоматической нотной транскрипции музыкального фольклора. // Докл. 14-й Всеросс. конф. «Математические методы распознавания образов», Москва, 2009.

[6] *Кальян В. П.* О настройке алгоритмов сегментации на задачи нотной транскрипции музыкального фольклора. // Модели и методы распознавания речи. Москва: ВЦ РАН им. А. А. Дородницына, 2009.

[7] *Кальян В. П.* Построение алгоритмов распознавания эмоционального состояния человека по пара и экстралингвистическим особенностям речи. // Модели и методы распознавания речи. Москва: ВЦ РАН им. А. А. Дородницына, 2010.

Анализ голосовых данных человека при гиперградитационном воздействии

Алябушев А. А., Карпушин М. А., Кузьмин А. В., Куликов А. И., Левин С. Г.
kulikov@nmsf.sscc.ru
Новосибирск, ИВМиМГ СО РАН, Новосибирский государственный университет, ООО «Сигнатек»

Предложен алгоритм, основанный на выделении в спектре звукового сигнала устойчивой структуры частот. Алгоритм применён к задачам определения высоты основного тона сигнала, выделения определённого источника в композитном сигнале, распознавания речи. Показана применимость к задаче анализа голосовых данных человека при гиперградитационном воздействии.

Задача

Кратковременное импульсное гиперградитационное воздействие на тело человека вызывает динамическое изменение скорости воздушного потока в дыхательных путях с частотой 30-40 Гц. Проведённые исследования показали, что возникающая при этом интенсивная стимуляция проприорецепторов скелетных мышц человека управляет механизмами контроля нейромышечной системы, дыхания, кровообращения, скорости метаболизма, функциональным состоянием костной и иммунной систем. Выявлены нейрогенные реакции релаксации гладких мышц мелких дыхательных путей в этих условиях, а также обнаружены возможности развития голоса при пении и речи [1].

Представляемая работа направлена на изучение влияния гипергравитационных тренировок на голосовые данные человека, и в первую очередь на вокальную речь. Практическая часть исследования состоит в анализе голосовых данных. В подготовке данных участвовало несколько дикторов, они пропевали звуки определённой высоты в покое и на вибрационной платформе Power Plate.

Звуковой сигнал вокальной речи можно разделить на мелодическую составляющую, имеющую в вокальной речи большую энергию, и собственно речь. Причём важным в данном случае является не определение текста, который произносит диктор, а выделение участков сигнала, на которых произносятся гласные и сонорны звуки — звуки, несущие в себе мелодическую информацию на значительном интервале времени. Однако тот факт, что младшие форманты, как правило, оказываются в одной частотной области с мелодическими структурами сигнала, осложняет формантный анализ такого сигнала.

Существует ряд подходов к моделированию речевого сигнала [2, 3, 4]. Среди них следует выделить параметрические — базирующиеся на параметрической модели голосового тракта человека; статистические — в которых голосовой сигнал моделируется по общим для акустических сигналов правилам, а специфические характеристики оцениваются статистически; психоакустические — основанные на модели восприятия звука человеком, используются главным образом в задачах обработки звука, связанных с его представлением конечному слушателю.

Наиболее адекватными данному исследованию представляются параметрические модели речевого сигнала. Параметрами в таких моделях могут являться, например, скорость экспираторного потока воздуха в дыхательных путях человека или площадь его голосовой щели. Измерение таких показаний требует использования специального медицинского оборудования.

Была изучена возможность использования психоакустической модели голосового сигнала, лежащей в основе популярных звуковых форматов сжатия с потерями, в решении поставленной задачи. Основываясь на этой модели, было предложено выделить в акустическом сигнале следующую структуру.

Метод

Предлагается выделить в спектре сигнала структуру, названную *акустическим ядром* — последовательность значений энергий гармоник, лепестки которых содержат значение частоты, кратное некоторому фиксированному её значению — базовой частоте.

Введём некоторые обозначения. Пусть $2n$ — длина входного вектора, F_s — частота дискретизации. Тогда $n+1$ — число значащих гармоник (с учётом зеркального эффекта). Пусть также $f_i = i\Delta f$, где $i = 0, 1, \ldots, n$, $\Delta f = \frac{F_s}{2n}$. Через $\lfloor a \rfloor$ будем обозначать целую часть числа a.

Акустическим ядром частоты $f \in [f_1, f_n]$ называется последовательность пар $\{(A_j, n_j)\}_{j=1}^{K_f}$, в которой $A_j \geqslant 0$ — амплитуда соответствующей гармоники, $n_j = \left\lfloor \frac{jf}{\Delta F} \right\rfloor$ для всех $j = 1, \ldots, K_f$, $K_f = \max\left\{k \in \mathbb{N}\colon \left\lfloor \frac{kf}{\Delta f} \right\rfloor \leqslant n\right\}$.

Последовательность $M_f = \left\{\left\lfloor \frac{jf}{\Delta f} \right\rfloor\right\}_{j=1}^{K_f}$ назовём *набором частоты* $f \in [f_1, f_n]$. Величину K_f назовём *размером набора*. Заметим: если $f \geqslant g$, то $K_f \leqslant K_g$.

Будем говорить, что $M_f \geqslant M_g$, если для всех $m = 1, \ldots, K_g$ верно $\left\lfloor \frac{mf}{\Delta f} \right\rfloor \geqslant \left\lfloor \frac{mg}{\Delta f} \right\rfloor$. Заметим, что

из того, что $f = g$ или $f \geqslant g$, следует, что $M_f = M_g$ или $M_f \geqslant M_g$ соответственно. Обратное утверждение в обоих случаях, вообще говоря, неверно.

В предложенной терминологии легко доказывается следующее утверждение.

Утверждение 1. *Пусть* $f_1 \leqslant \varphi < \psi \leqslant f_n$.

Тогда существует последовательность точек $\varphi = y_0 < y_1 < \cdots < y_L \leqslant \psi$, *такая, что для любого* $k \in \mathbb{N}$, $0 \leqslant k \leqslant L-1$ *и любых* f, g *из полуинтервала* $[y_k, y_{k+1})$ *или из отрезка* $[y_l, \psi]$ *имеем:* $M_f = M_g$.

При этом, если $k > 0$, *то* $y_k = \frac{c_k \Delta f}{m_k}$ *для некоторых целых чисел* c_k *и* m_k, *удовлетворяющих условиям:* $1 \leqslant m_k \leqslant K_\varphi$, $\frac{m_k \varphi}{\Delta f} < c_k \leqslant \min\{n, \frac{m_k \psi}{\Delta f}\}$.

Точки $y_0, y_1, \ldots, y_L$ назовём *переходными точками*, а полуинтервалы $[y_k, y_{k+1})$ для $k = 0, \ldots, L-1$ и отрезок $[y_L, \psi]$ — *областями неразличимости*. Если $y_L = \psi$, то отрезок вырождается, и последней областью неразличимости становится одна точка ψ.

Приведём теперь описание алгоритма выделения акустических ядер в некотором частотном диапазоне. Алгоритм состоит из двух этапов: подготовительного и основного.

На подготовительном этапе частотный диапазон разбивается на области неразличимости. Разбиение осуществляется согласно Утверждению 1. Это нужно для того, чтобы в ходе основного этапа можно было быстро перебрать все акустические ядра. Для этого достаточно взять по одному значению частоты из каждой области неразличимости и вычислить соответствующий ей набор. В качестве таких точек могут быть использованы переходные точки.

Уже было отмечено, что любая последовательность гармоник с кратными частотами есть акустическое ядро, однако далеко не каждая такая конструкция несёт содержательные сведения. Таким образом, требуется представить некоторую величину, характеризующую степень «адекватности» или «полезности» данного акустического ядра. Для этого введём *функцию релевантности* акустического ядра. Ей может быть произвольная функция $r\colon [\varphi, \psi] \times \mathbb{R}_+^{n+1} \to \mathbb{R}$. По заданным частоте и набору амплитуд всех гармоник в спектре буфера эта функция определяет *релевантность* соответствующего носителя, т. е. степень его содержательности. Данная функция подбирается эмпирически и для решения разных задач имеет различный вид.

На основном этапе алгоритма для заданного входного вектора выделяются всевозможные акустические ядра с частотами из указанного диапазона, и из них, согласно выбранной функции релевантности, выбираются один или несколько «значимых».

Примером функции релевантности может служить следующее:

$$r_A(f, A_0, \ldots, A_n) = \sum_{i=1}^{K_f} A_{\lfloor \frac{if}{\Delta f} \rfloor} / K_f.$$

Данная функция является амплитудной функцией релевантности и определяет акустическое ядро наибольшей средней амплитуды. Её полезно использовать в случае, когда известно, что входной сигнал от одного источника, например в реализации инструментального тюнера или программы-тренера. Построенные на приведённом алгоритме программные средства работают в режиме реального времени на современных массовых вычислительных машинах при разрядности 512–1024 полос.

Применение разработанного алгоритма не ограничивается определением мелодической точности сигнала. Если предложить следующую функцию релевантности:

$$r_P(f, A_0, \ldots, A_n) = \sum_{i=1}^{K_f} p_i A_{\lfloor \frac{if}{\Delta f} \rfloor},$$
$$p_1 + \cdots + p_{K_f} = 1, p_i \geqslant 0, i = 1, \ldots, K_f,$$

появляется возможность выделять отдельные источники в композитном сигнале. Изменяемый набор P позволяет выбрать источник необходимого тембра. Следует заметить, что при построении вектора P следует выбирать $p_1 \leqslant p_2 \leqslant \ldots \leqslant p_{K_f}$, т. к. гармоники больших частот, как правило, имеют меньшую энергию в спектре, но на определение тембральных характеристик звукового сигнала оказывают решающую роль.

Алгоритм с такими свойствами будет полезен в задачах вычленения какого-либо продолжительного звука в композитном сигнале, а также в задачах распознавания речи и сжатия с потерями.

Для этой версии так же было реализовано программное средство, показавшее свою эффективность при выделении синтезированных звуков и звуков, звучащих без пауз.

Ещё одним свойством акустических ядер является соотношение между энергиями гармоник в спектре. Был проведён ряд экспериментов: делались записи произнесений гласных и сонорных звуков на определённых частотах различными дикторами (нота «до» первой октавы для дикторов-женщин, базовая частота 261,6 Гц, и нота «до» малой октавы для дикторов-мужчин, базовая частота 130,8 Гц). Было установлено, что для одного диктора на одних фонемах сохраняются соотношения между энергиями гармоник в акустическом ядре. Это позволяет определять с некоторой точностью произносимые фонемы, если иметь предварительно записанные образцы для конкретного диктора. Ра-

Таблица 1. Соотношения между энергиями последовательных гармоник в акустическом ядре для различных звуков.

	А	И	О
E_2/E_1	1,874067	3,604912	0,717089
E_3/E_2	1,170366	4,497671	2,537819
E_4/E_3	1,355672	3,438823	1,985327
E_5/E_4	2,271577	3,461974	20,96945
	У	Ы	Э
E_2/E_1	1,326708	4,305435	0,78562
E_3/E_2	1,807619	5,112831	3,670975
E_4/E_3	15,81204	4,254243	7,665905
E_5/E_4	4,441692	2,030272	2,562511
	Л	М	Н
E_2/E_1	2,242848	10,89833	9,362077
E_3/E_2	2,607236	0,370299	5,635355
E_4/E_3	0,964622	8,180517	1,519818
E_5/E_4	14,36909	2,557141	0,600061

Таблица 2. Вероятность распознавания различных звуков предложенным методом.

А	И	О	У	Ы	Э
55%	60%	80%	75%	55%	60%

бота алгоритма была проверена на четырёх различных дикторах: трёх девушках и одном мужчине. В таблице 1 приводится пример таких соотношений для одного диктора-девушки. Через E_k в таблице обозначается энергия k-й гармоники в ядре. Значения для конкретных произнесений могут отличаться от указанных на 30%.

В таблице 2 приведена эффективность предложенного алгоритма при распознавании различных звуков.

Также были проведены эксперименты по анализу голоса человека во время гипергравитационной нагрузки и в покое. Одни и те же дикторы при вибрационной стимуляции и в покое произносили одни и те же фразы при одной и той же высоте звука. Было установлено, что спектры записей, сделанных при стимуляции, отличаются от спектров записей, сделанных в покое, большей энергией гармоник и дополнительным пиком на частоте стимуляции, при этом общая форма спектра сохраняется. Отсюда можно заключить, что анализ сигнала в спектральной области при заранее известной частоте стимуляции можно производить по тем же алгоритмам, что и в покое.

Результаты

Разработан алгоритм анализа звуковых данных, основанный на выделении структуры частот в Фурье-спектре сигнала. Программная реализация позволяет в режиме реального времени определять высоту основного тона. При других параметрах алгоритм позволяет выделять в композитном сигнале звук от источника с определёнными тембральными характеристиками. Так же свойства полученной структуры позволяют распознавать в голосовом сигнале гласные и сонорные звуки с эффективностью 55–80%. Полученные результаты могут применяться как для поставленной изначально задачи — анализа голосовых данных при гипергравитационной стимуляции, так и для других задач обработки звуковых сигналов.

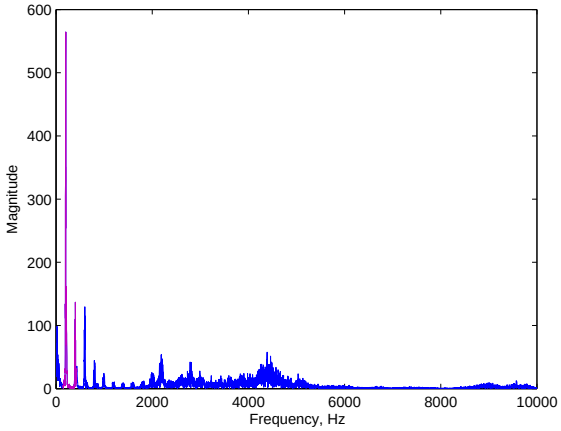


Рис. 1. Спектр сигнала в покое. Слог «ней», женский голос.

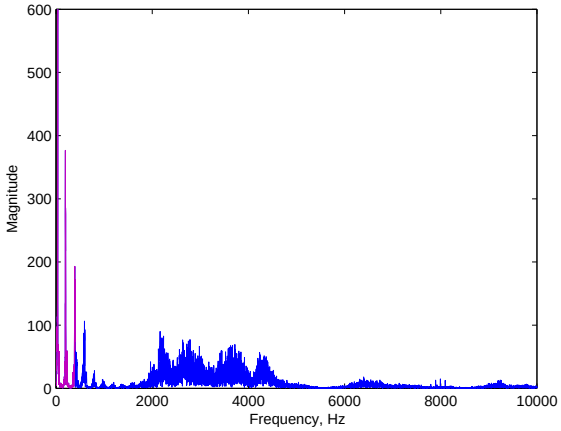


Рис. 2. Спектр сигнала при гипергравитационной стимуляции. Слог «ней», женский голос.

Разработанные средства используются в реализации проекта «Вокальный тренер», основанного на методике обучения вокалу Сэта Риггза [5].

Литература

[1] *Пятин В. Ф., Широлапов И. В.* Однократная вибрационная нагрузка значительно увеличивает скорость экспираторного воздушного потока у челове-

ка // Вестник ТГУ. Серия «биология и экология», 2009. № 1. — С. 38–42.

[2] *Чучупал В. Я., Чичагов А. С., Маковкин К. А.* Цифровая фильтрация зашумлённых звуковых сигналов. — Москва: ВЦ РАН, 1998. — 51 с.

[3] *Сорокин В. Н., Макаров И. С.* Обратная задача для голосового источника // Информационные процессы, — 2006. — Т. 6, № 4. — С. 375–395.

[4] *Алдошина И. А.* Основы психоакустики // Звукорежиссёр. — 1999, № 6 — 2002, № 8.

[5] *Риггз С.* Пойте как звёзды // Сост. и ред. Дж. Д. Каррателло. — Санкт-Петербург: Питер, 2007. — 120 с.

Оценка адекватности вычислительных моделей дискретного преобразования Гильберта*

Чичагов А. В.
mail2chi@ya.ru
Москва, Вычислительный центр им. А. А. Дородницына РАН

В работе рассматривается критерий оценки адекватности (т. е. степени соответствия идеалу) вычислительных моделей преобразований цифровых сигналов. Приводятся результаты вычислительных экспериментов по оценке адекватности вычислительных моделей дискретного преобразования Гильберта.

Вычислительная модель представляет аппроксимацию нетривиального математического преобразования класса сигналов. Очевидно существует множество возможностей для выбора конструктивной аппроксимации (алгоритма) математического преобразования сигнала. Кроме этого, практически любая реализация вычислительной модели, т. е. программа или процедура, имеет несколько аргументов (параметров), значения которых требуется задать при её использовании. Таким образом, пользователь путём выбора как самой процедуры, так и значений её параметров может в определённой степени управлять «качеством» или точностью цифровой обработки сигнала (ЦОС).

Для оценки погрешности цифровой обработки сигнала обычно используют метод вычисления среднего квадратического отклонения (СКО). При этом на практике часто ограничиваются общей оценкой погрешности цифровой обработки, что иногда может оказаться недостаточно. Ниже рассматривается критерий оценки адекватности вычислительных моделей преобразований цифровых сигналов на конкретном примере дискретного преобразования Гильберта (ДПГ).

Критерий оценки адекватности вычислительных моделей *ЦОС*

Выберем семейство комплексных гармонических функций $z_s(t) = A_0 \exp(2\pi j F_s t)$, где A_0 — амплитуда, F_s – частота гармоники, t — время и $j = \sqrt{-1}$, в качестве семейства тестовых сигналов. Оценка мощности/дисперсии комплексной гармоники $\overline{\sigma_s^2} = \overline{|z_s(t)|^2} \equiv A_0^2$ является величиной независимой как от значения частоты гармоники, так и «длины» интервала оценивания, что достаточно удобно для вычисления оценок.

Набор исходных (*source*) тестовых выборок данных или, иначе, набор *цифровых сигналов источника* $\{\mathbf{z}_s^{(src)}\}$, сформируем в виде набора пар временных рядов:

$$x_s^{(src)}(i) = A_0 \cos(2\pi(s/2L_0)i),$$

$$y_s^{(src)}(i) = A_0 \sin(2\pi(s/2L_0)i),$$

где $i = \lfloor F_D^{(src)} t \rfloor$ — индекс (смещение по времени) элемента выборки данных $(0 \leqslant i < N_0)$, $s/2L_0 \sim F_s/F_D^{(src)}$ — нормализованная частота гармоники, s — спектральный индекс (смещение по частоте) гармоники сигнала источника, $F_D^{(src)}$ — частота дискретизации сигнала источника.

По условию Найквиста $F_s/F_D^{(src)} < 1/2$, поэтому полагаем $s \in 0, 1, \ldots, L_0 - 1$. Очевидно, что тестовые цифровые сигналы источника имеют одинаковую мощность/дисперсию $\overline{\sigma_s^2} = \|\mathbf{z}_s^{(src)}\|^2 \equiv A_0^2$ для всех $s \in 0, 1, \ldots, L_0 - 1$.

Набор ожидаемых (*expected*) тестовых цифровых сигналов $\{\mathbf{z}_s^{(exp)}\}$, соответствующий исследуемому преобразованию, в данном случае, преобразованию Гильберта [1] набора исходных тестовых цифровых сигналов, сформируем в виде набора пар временных рядов:

$$x_s^{(exp)}(n) = -A_0(1 - \delta(s)) \cos(2\pi K_D^{-1}(s/2L_0)n),$$

$$y_s^{(exp)}(n) = A_0(1 - \delta(s)) \sin(2\pi K_D^{-1}(s/2L_0)n),$$

где n — целочисленный сдвиг (смещение во времени) элемента выборки данных, $K_D = F_D^{(dst)}/F_D^{(src)}$ — заданный коэффициент редискретизации, $F_D^{(dst)}$ — частота дискретизации преобразованного сигнала, $\delta(s)$ — символ Кронекера.

Элементы набора ожидаемых тестовых цифровых сигналов также имеют одинаковую мощность/дисперсию, равную $\overline{\sigma_s^2} = \|\mathbf{z}_s^{(exp)}\|^2 \equiv A_0^2$ для всех $s \in 1, \ldots, L_0 - 1$, кроме случая $s = 0$, для которого, очевидно, $\overline{\sigma_{s=0}^2} = \|\mathbf{z}_{s=0}^{(exp)}\|^2 \equiv 0$.

Набор фактических (*actual*) тестовых цифровых сигналов $\{\mathbf{z}_s^{(act)}\}$, представляющий результат цифровой обработки набора исходных тестовых цифровых сигналов $\{\mathbf{z}_s^{(src)}\}$ исследуемой процедурой **DSP** цифровой обработки сигналов, в нашем случае процедурой ДПГ, сформируем в виде:

$$\mathbf{x}_s^{(act)} = \mathbf{DSP}(\boldsymbol{p}, \mathbf{x}_s^{(src)}), \;\; \mathbf{y}_s^{(act)} = \mathbf{DSP}(\boldsymbol{p}, \mathbf{y}_s^{(src)}),$$

где $\boldsymbol{p}$ — вектор параметров процедуры.

Для оценки адекватности, т. е. качества (точности) исследуемого цифрового преобразования сигнала предлагается оценка частотной кривой адекватности, которая строится на основе фазосогласованных или минимальных среднеквадратических отклонений (МСКО):

Работа выполнена при финансовой поддержке РФФИ, проект №11-01-00900.

$$\overline{\varepsilon_s^2} = \min_{\Delta\varphi} \|\mathbf{z}_s^{(act)} - e^{-j\Delta\varphi}\mathbf{z}_s^{(exp)}\|^2,$$

где s — спектральный индекс (номер гармоники), $\Delta\varphi$ — возможный фазовый сдвиг между гармониками ожидаемого и фактического тестовых сигналов, который может быть обусловлен реализацией цифрового преобразования.

Оптимальное значение фазового сдвига легко найти; подставив его в приведённое выше выражение для набора величин МСКО, получим:

$$\overline{\varepsilon_s^2} = \|\mathbf{z}_s^{(act)}\|^2 + \|\mathbf{z}_s^{(exp)}\|^2 - \\ - 2\sqrt{\langle \mathbf{z}_s^{(act)}|\mathbf{z}_s^{(exp)}\rangle\langle \mathbf{z}_s^{(exp)}|\mathbf{z}_s^{(act)}\rangle},$$

где

$$\langle \mathbf{z_1}|\mathbf{z_2}\rangle = \frac{1}{(N_2 - N_1)}\sum_{n=N_1}^{N_2-1} z_1(n)z_2^*(n)$$

— скалярное произведение комплексных векторов, N_1, N_2 — соответственно, начальная и конечная границы фрагментов цифровых сигналов, по которым оценивается МСКО. При этом границы фрагментов выбирают так, чтобы исключить влияние краевых эффектов вычислительной модели преобразования цифрового сигнала.

Чтобы можно было легко сравнивать наборы МСКО, соответствующие различным вычислительным моделям или разным значениям параметров моделей цифровых преобразований, необходимо определить унифицированную шкалу оценок. Удобно в качестве такой шкалы использовать шкалу децибел и представить набор МСКО или оценку *частотной кривой адекватности* (оценку «ЧКА») вычислительной модели ЦОС в виде:

$$\overline{\zeta_s^2} = 10\lg(\overline{\varepsilon_s^2}/\overline{\sigma_s^2}),$$

где в качестве нормировочного коэффициента выбрана мощность сигнала источника.

Предложенная оценка ЧКА представляет набор значений («кривую») нормализованных спектральных погрешностей, которая соответствует тройке информационных объектов описывающих вычислительную модель преобразования сигнала, а именно:

- **имени** «преобразования» (естественно-языковому идентификатору, который требуется для использования в информационном «мире» или библиографическом пространстве),
- **декларации** преобразования (определению ожидаемых результатов преобразования для класса сигналов),
- **реализации** преобразования (алгоритму и конструкции специализированного вычислителя выполняющего преобразование цифровых сигналов «более или менее удовлетворительно»).

Как показывает практика, оценка ЧКА является простым и достаточно информативным метрологическим показателем качества или, более точно, адекватности вычислительной модели (реализации) определению (декларации) преобразования сигнала.

Алгоритм дискретного преобразования Гильберта

Пусть $\mathbf{u} = u[iT_D^{(s)}]$ и $\mathbf{v} = v[nT_D^{(d)}]$ — входная и выходная выборки цифровых сигналов, связанные, как априори предполагается, преобразованием Гильберта [1], а $T_D^{(s)} = 1/F_D^{(s)}$, $T_D^{(d)} = 1/F_D^{(d)}$ — шаг и частота дискретизации соответственно входной и выходной выборок цифровых сигналов. Дискретное преобразование Гильберта (ДПГ) с редискретизацией цифрового сигнала формально можно представить в виде:

$$v[nT_D^{(d)}] = \sum_{i=-\infty}^{\infty} g\Big(nT_D^{(d)} - iT_D^{(s)}\Big)\, u\big[iT_D^{(s)}\big],$$

где

$$g(F_D^{(s)}\tau) = \frac{1 - \cos\pi F_D^{(s)}\tau}{\pi F_D^{(s)}\tau}$$

— весовая функция ядра или относительный вклад элемента $u[iT_D^{(s)}]$ входной выборки в текущее значение элемента $v[nT_D^{(d)}]$ выходной выборки (является асимптотически убывающей функцией модуля аргумента).

Определим индекс $\tilde{i}[n]$ элемента входной выборки, который наиболее близко «по физическому времени» соответствует индексу n элемента выходной выборки

$$\tilde{i}[n] = (nF_D^{(s)})/F_D^{(d)}$$

и представим аппроксимацию ДПГ в виде:

$$v[nT_D^{(d)}] = \sum_{m=-M}^{M} g'\Big(nT_D^{(d)} - \big(\tilde{i}[n] + m\big)T_D^{(s)}\Big)\cdot \\ \cdot u\big[\,(\tilde{i}[n] + m)T_D^{(s)}\big],$$

где $m = i - \tilde{i}[n]$ — относительный индекс элементов входной выборки или смещение относительно центра окна или сегмента аппроксимации, $M = M_a/2$ — полуширина сегмента, M_a — ширина или апертура окна аппроксимации ДПГ,

$$g'(F_D^{(s)}\tau) = g(F_D^{(s)}\tau)\cos^{\beta}\Big(\frac{\pi F_D^{(s)}\tau}{2M+1}\Big)$$

— сглаженная весовая функция ядра, $\cos^{\beta}(\dots)$ — формирующая функция окна аппроксимации, $\beta = 0, 1, 2$ — параметр «гладкости» окна.

Введем коэффициент редискретизации цифрового сигнала $K_D = F_D^{(d)}/F_D^{(s)}$ и перепишем аппроксимационную формулу ДПГ в виде простого «академического» алгоритма:

Вход: $\boldsymbol{p} = (M_a, K_D, \beta)$, $\boldsymbol{u} = \{u[i]\}$;
1: инициализация: $n := 0$; $i := 0$;
2: **пока** $i < \boldsymbol{u}.len$ **повторять**
3: $\quad i := \lfloor n/K_D \rfloor$;
4: $\quad v[n] := \sum_{m=-M}^{M} u[i+m] g'(n/K_D - i - m)$;
5: $\quad n := n + 1$;
Выход: $\boldsymbol{v} = \{v[n]\}$;

Рассмотренный «академический» алгоритм ДПГ предполагает значительные вычислительные затраты, так как при вычислении каждого текущего значения выходной выборки цифрового сигнала требуется пересчёт «ядра» преобразования. Этих вычислений, однако, можно избежать, если модифицировать структуру рассмотренного алгоритма.

Действительно, коэффициент редискретизации цифрового сигнала $K_D = F_D^{(d)}/F_D^{(s)}$ можно аппроксимировать *рациональным* числом или дробью $K_D = L/N$, где числитель L можно интерпретировать как коэффициент интерполяции, а знаменатель N — как коэффициент децимации.

В этом случае множество возможных значений минимальных временных сдвигов $\tau = (n/K_D - -i)T_D^{(s)}$ между ближайшими текущими элементами входной и выходной выборок «вырождается» в конечный набор значений, количество элементов которого определяется величиной L.

Из этого следует практическая реализуемость вычисления *полного набора* возможных значений весовой функции ядра для заданного значения коэффициента редискретизации. Вычисленные значения весовой функции затем можно использовать в процессе цифровой обработки сигнала.

Иными словами, рассмотренный выше «академический» алгоритм ДПГ можно представить в виде «автокоммутируемого» набора линейных цифровых фильтров с постоянными коэффициентами или, иначе, полифазного фильтра. Таким образом, «практический» алгоритм ДПГ состоит из двух процедур [2]:

— процедуры вычисления коэффициентов набора линейных цифровых фильтров,
— процедуры полифазной линейной цифровой фильтрации данных.

Процедуру вычисления коэффициентов набора линейных цифровых фильтров легко модифицировать. В частности, можно использовать другие формирующие функции окна аппроксимации, отличные от рассмотренных выше.

Апробация вычислительных моделей ДПГ

Апробация вычислительных моделей ДПГ представляет серию вычислительных экспериментов, в которых исследовалось поведение оценки ЧКА от значений величины апертуры и параметра «гладкости» окна аппроксимации. Для вычисления оценок ЧКА использовалось контрольно-измерительное приложение [3], а для построения графиков оценок ЧКА — пакет OpenOffice.

Сценарий вычислительного эксперимента можно представить в виде следующей последовательности действий:

1. определить значения параметров текщей сессии вычислительного эксперимента, т. е.
   — задать значения параметров процедуры синтеза набора тестовых сигналов,
   — задать значения параметров процедуры цифровой обработки сигнала,
   — задать значения параметров процедуры представления информационного отчёта;
2. вычислить набор исходных *(sources)* тестовых цифровых сигналов,
3. вычислить набор ожидаемых *(expected*) тестовых цифровых сигналов,
4. вычислить набор фактических *(actual)* тестовых цифровых сигналов,
5. вычислить оценку ЧКА, соответствующую заданным значениям параметров,
6. скомпилировать информационный отчёт с результатами текущей сессии вычислительного эксперимента для последующего анализа и обсуждения.

Информационный отчёт должен содержать информацию о текущей сессии вычислительного эксперимента, в частности, следующее:

— значения параметров процедуры синтеза набора тестовых сигналов $\boldsymbol{q} = (A_0, N_0, L_0)$,
— значения параметров процедуры дискретного преобразования Гильберта $\boldsymbol{p} = (M_a, K_D, \beta)$,
— оценку ЧКА $\overline{\zeta_s^2}(\boldsymbol{q}, \boldsymbol{p})$ соответствующую указанным значениям параметров процедуры синтеза набора тестовых сигналов и процедуры ДПГ.

Здесь A_0 — величина амплитуды сигнала источника («амплитуда» гармоники), N_0 — объём тестовой выборки данных («длина» фрагмента сигнала источника), L_0 — размерность набора тестовых выборок данных (мощность множества/спектра гармоник), M_a — апертура окна аппроксимации ДПГ, K_D — коэффициент редискретизации цифрового сигнала, β — параметр «гладкости» окна.

Результаты отдельных вычислительных экспериментов обычно группируются в серию и в дальнейшем анализируются совместно. Результатом

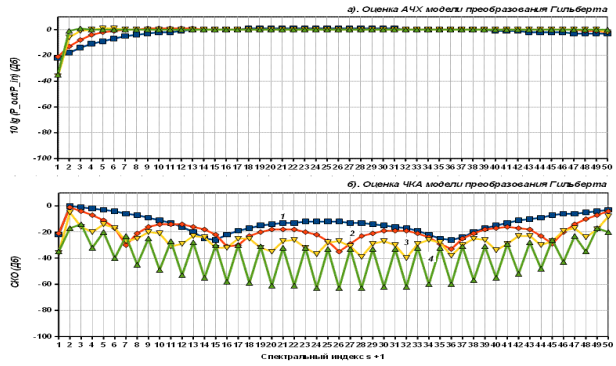


Рис. 1. Графики оценок АЧХ (а) и ЧКА (б) модели ДПГ ($\beta = 0$) при различных значениях апертуры. Кривые $1, \ldots, 4$ соответствуют апертуре $M_a = 5, 10, 25, 50$, при $K_D = 33.3, \quad L_0 = 50, \quad A_0 = 10000$.

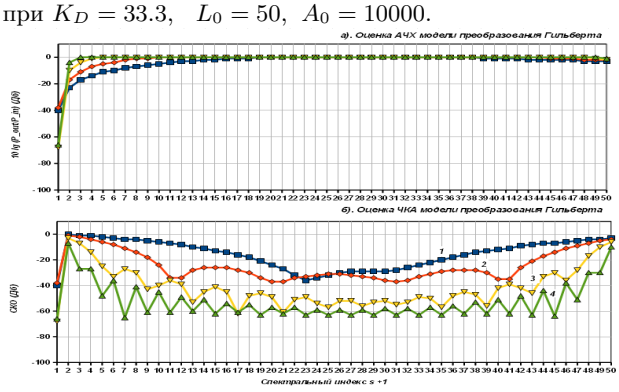


Рис. 2. Графики оценок АЧХ (а) и ЧКА (б) модели ДПГ ($\beta = 1$) при различных значениях апертуры. Кривые $1, \ldots, 4$ соответствуют апертуре $M_a = 5, 10, 25, 50$, при $K_D = 33.3, \quad L_0 = 50, \quad A_0 = 10000$.

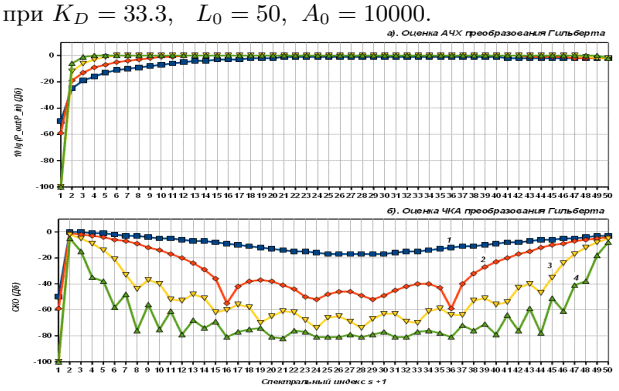


Рис. 3. Графики оценок АЧХ (а) и ЧКА (б) модели ДПГ ($\beta = 2$) при различных значениях апертуры. Кривые $1, \ldots, 4$ соответствуют апертуре $M_a = 5, 10, 25, 50$, при $K_D = 33.3, \quad L_0 = 50, \quad A_0 = 10000$.

анализа такой серии вычислительных экспериментов может быть таблица или диаграмма зависимостей оценок ЧКА от значений параметров процедуры ДПГ или процедуры синтеза тестовых цифровых сигналов.

На рис. 1–3 показаны графики оценки амплитудно-частотной характеристики (АЧХ) в логарифмическом масштабе и оценки ЧКА для рассмотренных выше моделей ДПГ при различных значениях величины апертуры окна аппроксимации $M_a = 5, 10, 25, 50$ и параметра гладкости $\beta = 0, 1, 2$. Можно отметить, что приведённые кривые нормализованных спектральных погрешностей вычислительной модели ДПГ являются квазипериодическими функциями нормализованной частоты, причём период этих функций зависит от величины апертуры окна аппроксимации. Причиной такого характера поведения оценки ЧКА является явление Гиббса, которое проявляется при «обрезании» медленно затухающей импульсной характеристики линейного фильтра, в данном случае, ядра преобразования Гильберта.

Использование формирующих окон, повышающих «скорость затухания» исходной импульсной характеристики линейного фильтра, позволяет существенно уменьшить величину этого эффекта и повысить точность цифрового преобразования в средней полосе спектра частот. Заметим, что обнаружить подобный характер поведения вычислительных моделей ДПГ по графикам оценки АЧХ не так просто.

Выводы

В работе предложена оценка частотной кривой адекватности (ЧКА) вычислительных моделей преобразований цифровых сигналов. Показано, что оценка ЧКА является достаточно информативным метрологическим показателем адекватности вычислительной модели («реализации») преобразования определению («декларации») преобразования сигнала.

Рассмотрен ряд вычислительных моделей дискретного преобразования Гильберта с редискретизацией цифрового сигнала. Изложен сценарий вычислительного эксперимента и приведены результаты серии вычислительных экспериментов по оценке адекватности рассмотренных моделей ДПГ. Показано, что графики оценок ЧКА вычислительных моделей преобразований сигналов предоставляют значительно больше информации о поведении алгоритма/программы преобразования цифровых сигналов, чем интегральная (средняя по спектру) величина погрешности.

Литература

[1] *Рабинер Л., Гоулд Б.* Теория и применение цифровой обработки сигналов. — Москва: Мир, 1978. — 848 с.

[2] *Чичагов А. В.* Метод оценки качества редискретизации цифровых сигналов. // Речевые технологии/Speech Technology. — 2009. № 4. — С. 26–39.

[3] *Чичагов А. В.* Программа оценки точности дискретного преобразования Гильберта (версия 1.0) // Свидетельство РФ о государственной регистрации программы для ЭВМ № 2010614943. Зарегистрировано 29 июля 2010 г.

Новые трёхфазные и пятифазные последовательности с одноуровневой периодической автокорреляционной функцией*

Леухин А. Н., Парсаев Н. В.
code@marstu.net
Йошкар-Ола, Марийский государственный технический университет

Проведён анализ известных методов синтеза p-фазных последовательностей с одноуровневой периодической автокорреляционной функцией, структура которых определяется элементом поля $GF(p)$. Получены аналитические выражения для синтеза трёхфазных и пятифазных последовательности с одноуровневой периодической автокорреляционной функцией.

Унимодулярные (фазокодированные) последовательности с идеальной периодической автокорреляционной функцией (ПАКФ) находят широкое применение в системах связи, синхронизации, радиолокации и радионавигации [1–5]. С точки зрения технической реализации большой практический интерес представляет построение унимодулярных дельта-коррелированных последовательностей с малым объёмом фазового алфавита.

Построение бинарных последовательностей с одноуровневой ПАКФ основано на теории циклических разностных множеств с параметрами (ν, k, λ) [4]. Однако свойство нулевой автокорреляции не может быть достигнуто для бинарных последовательностей за исключением единственной последовательности вида $\{1, 1, 1, -1\}$. Голомбом первым было показано, что разностные множества типа Адамара с параметрами $(4t-1, 2t-1, t-1)$ приводят к построению бинарных последовательностей с одноуровневой ПАКФ и уровнем боковых лепестков $a = -1$. В книге [4] приводится полная классификация известных на сегодняшний день разностных множеств типа Адамара.

Помимо бинарных последовательностей с малым уровнем боковых лепестков ПАКФ большой интерес представляют p-фазные последовательности, построение которых основано на замене символов p-ичной последовательности, определённой над полем $GF(p^m)$, символами из алфавита, образованного делением круга на p равных частей [5]. Данный класс p-фазных последовательностей, имеют такой же уровень боковых лепестков ПАКФ, как и бинарные последовательности типа Адамара. В работе [5] рассмотрены методы синтеза p-ичных m-последовательностей, а в работах [6, 7] рассмотрены методы синтеза p-ичных GMW-последовательностей.

Синтезируемая p-ичная последовательность длины $N = p^m - 1$ определится на основании рекуррентного выражения:

$$c_n = \sum_{j=0}^{m-1} c_{n-j-1} \cdot f_{m-j-1} \pmod{p},$$

$$n = m, m+1, ..., N-1, \qquad (1)$$

где начальные отсчёты последовательности имеют вид $c_0 = c_1 = ... = c_{m-2} = 0$, $c_{m-1} = 1$, коэффициенты $f_i = p - a_i$, $i = 0, 1, \ldots, m-1$ задаются примитивным нормированным многочленом степени m над полем $GF(p)$, $a_i \in GF(p)$:

$$F(x) = x^m + a_{m-1}x^{m-1} + ... + a_1 x + a_0.$$

Бифазные последовательности с одноуровневой ПАКФ

Впервые в 1971 году, в книге Амиантова [8] было показано, что если в бинарной последовательности $\{b_n\}_0^{N-1}$, построенной на основании разностного множества типа Адамара, произвести замену символов $b_n = -1$ на символы:

$$b_n = \exp(i\varphi),\ \cos(\varphi) = \frac{N-1}{N+1}, \qquad (2)$$

то будет получена бифазная дельта-коррелированная последовательность.

В работе [9] полученный результат был обобщен для построения бифазных последовательностей с заданным уровнем a боковых лепестков одноуровневой ПАКФ из диапазона вещественных чисел $a \in [a_{\min}; N]$. Следуя работе [9], угол φ для заданного уровня боковых лепестков a и минимально возможный уровень боковых лепестков $a_{\min}$ определяются на основании выражения:

$$\varphi = \pi \pm \arccos\left(\frac{N^2 + 2k^2 - 2kN + a - N - Na}{2k(N-k)}\right),$$

$$a_{\min} = \frac{N^2 + 4k^2 - 4kN - N}{N-1}, \qquad (3)$$

где k — параметр разностного множества.

Трёхфазные последовательности с одноуровневой ПАКФ

В работе [5], было показано, что для конструирования унимодулярных дельта-коррелиро-

Работа выполнена при финансовой поддержке гранта Президента РФ №МД-5418.2010.9, в рамках ФЦП «Научные и научно-педагогические кадры инновационной России» на 2009-2013 годы ГК №02.740.11.0838 и ГК №П 783, гранта РФФИ №09-07-00072-а.

ванных последовательностей можно воспользоваться не только известными бинарными последовательностями, но и p-ичными последовательностями, заданными над полем $GF(p)$. В частности, в работе [5] было получено выражение для построения трёхфазных унимодулярных дельта-коррелированных последовательностей, заданных над полем $GF(3)$. Если символам 3-ичной последовательности $\{c_n\}_0^{N-1}$ длины $N = 3^m - 1$ с элементами из $GF(3)$ поставить в соответствие следующие символы:

$$b_n = \begin{cases} 1, \text{если } c_n = 0, \\ \exp(i\beta_1), \text{если } c_n = 1, \\ \exp(i\beta_2), \text{если } c_n = 2, \end{cases} \quad (4)$$

где $\beta_1 = \beta_2 \pm \arccos(d)$, $\beta_2 = \arccos\left(d\sqrt{\frac{2}{1+d}}\right) + \frac{1}{2}\arccos(d)$, $d = \frac{1-3^{m-1}}{2\cdot 3^{m-1}}$, то $\{b_n\}_0^{N-1}$ будет унимодулярной дельта-коррелированной последовательностью.

Кроме рассмотренной последовательности, аналогичными свойствами автокорреляции будут обладать последовательности вида

$$b1_n = \begin{cases} 1, \text{если } c_n = 0, \\ \exp(i\beta_1), \text{если } c_n = 1, \\ \exp(i\beta_2), \text{если } c_n = 2, \end{cases}$$

и сопряжённые им последовательности $\{b_n^*\}$ и $\{b1_n^*\}$. Отметим, что все последовательности $\{b_n\}$, $\{b1_n\}$, $\{b_n^*\}$ и $\{b1_n^*\}$ являются эквивалентными.

В данной работе полученный результат был обобщен для построения унимодулярных последовательности с заданным уровнем боковых лепестков a ПАКФ из диапазона вещественных чисел $a \in [-1; N]$. Значения аргументов последовательности $\{b_n\}_0^{N-1}$, построенной по правилу (4), для заданного уровня боковых лепестков a будут определяться на основании выражения:

$$\beta_1 = \beta_2 \pm \arccos(d),$$

$$\beta_2 = \arccos\left(d\sqrt{\frac{2}{1+d}}\right) + \frac{1}{2}\arccos(d), \quad (5)$$

$$d = \frac{1 - 3^{m-1} + a}{2\cdot 3^{m-1}}.$$

Пятифазные последовательности с одноуровневой ПАКФ

В данной работе было получено аналитическое выражение для нахождения значений фаз пятифазной последовательности $\{b_n\}_0^{N-1}$, приводящих к построению унимодулярных последовательностей с одноуровневой ПАКФ.

Поставим в соответствие символам 5-ричной последовательности $\{c_n\}_0^{N-1}$ длины $N = 5^m - 1$ с элементами из $GF(5)$ следующие символы:

$$b_n = \begin{cases} 1, \text{если } c_n = 0, \\ z_1, \text{если } c_n = 1, \\ z_2, \text{если } c_n = 2, \\ z_3, \text{если } c_n = 3, \\ z_4, \text{если } c_n = 4, \end{cases} \quad (6)$$

где $|z_1| = |z_2| = |z_3| = |z_4| = 1$.

Существует два решения, приводящие к построению унимодулярных последовательностей с одноуровневой ПАКФ.

Первое решение существует для уровня боковых лепестков a, принимающих значения из диапазона $a \in [-1; 5^{m-1}-1]$. Значения элементов кодовой последовательности для заданного уровня боковых лепестков a определяются на основании выражения:

$$e = \sqrt{14\cdot 5^m a + 14\cdot 5^m + 5^{2m} - 30a - 15a^2 - 15},$$

$$f = a - 5^{m-1} + 1,$$

$$g = -\frac{1}{4\cdot 5^m}\sqrt{\frac{10\left(\sqrt{5}\left(3f + 4\cdot 5^{m-1}\right) - e\right)}{3e}}\times$$

$$\times\sqrt{\left(-e^2 - 4\cdot 5^m\left(f - 4\cdot 5^{m-1}\right)\right)} + \frac{5f - \sqrt{5}e}{4\cdot 5^m},$$

$$h = \frac{gf}{5^{m-1}} - g^2 - 1,$$

$$l = \sqrt{\frac{h + \sqrt{h^2 - 4g^2)}}{2}},$$

$$z_4 = \frac{l\left(f - \sqrt{(a+1)\left(f - 4\cdot 5^{m-1}\right)}\right)}{5^{m-1}(l-1)(g-l)}, \quad (7)$$

$$z_3 = -z_4 l, z_2 = -z_4\frac{g}{l}, z_1 = z_4 g.$$

Кроме рассмотренной последовательности, аналогичными свойствами автокорреляции будут обладать последовательности вида

$$b1_n = \begin{cases} 1, \ if\, c_n = 0, \\ z_4, \ if\, c_n = 1, \\ z_2, \ if\, c_n = 2, \\ z_3, \ if\, c_n = 3, \\ z_1, \ if\, c_n = 4. \end{cases}$$

$$b2_n = \begin{cases} 1, \ if\, c_n = 0, \\ z_2, \ if\, c_n = 1, \\ z_1, \ if\, c_n = 2, \\ z_4, \ if\, c_n = 3, \\ z_3, \ if\, c_n = 4. \end{cases}$$

$$b3_n = \begin{cases} 1, \ if\, c_n = 0, \\ z_3, \ if\, c_n = 1, \\ z_1, \ if\, c_n = 2, \\ z_4, \ if\, c_n = 3, \\ z_2, \ if\, c_n = 4. \end{cases}$$

и сопряжённые им последовательности $\{b_n^*\}$, $\{b1_n^*\}$, $\{b2_n^*\}$ и $\{b3_n^*\}$. Рассмотренные последовательности формируют две группы эквивалентности. В первую входят последовательности $\{b_n\}$, $\{b_n^*\}$, $\{b1_n\}$ и $\{b1_n^*\}$, во вторую — $\{b2_n\}$, $\{b3_n\}$, $\{b2_n^*\}$ и $\{b3_n^*\}$.

При подстановке в выражение (7) значения $a = 0$ последовательность $\{b_n\}_0^{N-1}$, построенная по правилу (7), будет являться унимодулярной дельта-коррелированной последовательностью.

Второе решение существует для уровня боковых лепестков a, принимающих значения из диапазона $a \in [5^{m-1} - 1; N]$. Значения элементов кодовой последовательности для заданного уровня боковых лепестков a определяются на основании выражений:

$$s = \sqrt{(f - 4 \cdot 5^{m-1})(a+1)},$$

$$t = \frac{5^{m-1}(2f - 5^{m-1}) + (1+a)^2}{2 \cdot 5^{m-1}(3f + 4 \cdot 5^{m-1})} -$$

$$- \frac{(f + 4 \cdot 5^{m-1})s}{2 \cdot 5^{m-1}(3f + 4 \cdot 5^{m-1})},$$

$$z_1 = t - \sqrt{t^2 - 4t - \frac{11f + 4 \cdot 5^{m-1} - 8s}{(3f + 4 \cdot 5^{m-1})}}, \quad (8)$$

$$z_3 = t + \sqrt{t^2 - 4t - \frac{11f + 4 \cdot 5^{m-1} - 8s}{(3f + 4 \cdot 5^{m-1})}},$$

$$z_2 = \frac{f - s}{5^{m-1}} - 3z_1, \ z_4 = \frac{f - s}{5^{m-1}} - 3z_3.$$

Последовательности данного типа по существу являются трёхфазными, так как задаются выражениями вида:

$$b1_n = \begin{cases} 1, \ if\, c_n = 0, \\ z_4, \ if\, c_n = 1, \\ z_3, \ if\, c_n = 2, \\ z_3, \ if\, c_n = 3, \\ z_3, \ if\, c_n = 4. \end{cases} \quad b2_n = \begin{cases} 1, \ if\, c_n = 0, \\ z_3, \ if\, c_n = 1, \\ z_4, \ if\, c_n = 2, \\ z_3, \ if\, c_n = 3, \\ z_3, \ if\, c_n = 4. \end{cases}$$

$$b3_n = \begin{cases} 1, \ if\, c_n = 0, \\ z_3, \ if\, c_n = 1, \\ z_3, \ if\, c_n = 2, \\ z_4, \ if\, c_n = 3, \\ z_3, \ if\, c_n = 4. \end{cases} \quad b4_n = \begin{cases} 1, \ if\, c_n = 0, \\ z_3, \ if\, c_n = 1, \\ z_3, \ if\, c_n = 2, \\ z_3, \ if\, c_n = 3, \\ z_4, \ if\, c_n = 4. \end{cases}$$

$$b5_n = \begin{cases} 1, \ if\, c_n = 0, \\ z_2, \ if\, c_n = 1, \\ z_1, \ if\, c_n = 2, \\ z_1, \ if\, c_n = 3, \\ z_1, \ if\, c_n = 4. \end{cases} \quad b6_n = \begin{cases} 1, \ if\, c_n = 0, \\ z_1, \ if\, c_n = 1, \\ z_2, \ if\, c_n = 2, \\ z_1, \ if\, c_n = 3, \\ z_1, \ if\, c_n = 4. \end{cases}$$

$$b7_n = \begin{cases} 1, \ if\, c_n = 0, \\ z_1, \ if\, c_n = 1, \\ z_1, \ if\, c_n = 2, \\ z_2, \ if\, c_n = 3, \\ z_1, \ if\, c_n = 4. \end{cases} \quad b8_n = \begin{cases} 1, \ if\, c_n = 0, \\ z_1, \ if\, c_n = 1, \\ z_1, \ if\, c_n = 2, \\ z_1, \ if\, c_n = 3, \\ z_2, \ if\, c_n = 4. \end{cases}$$

Заключение

Рассмотрены вопросы синтеза p-фазных последовательностей с одноуровневой периодической автокорреляционной функцией, структура которых определяется на основании p-ичных последовательностей, заданных над полем $GF(p)$. Показано, что если элементам p-ичной последовательности, заданной над расширенным полем $GF(p^m)$ поставить в соответствие комплекснозначные символы, то полученная последовательность может иметь одноуровневую, в том числе идеальную, ПАКФ.

Для трёхфазных последовательностей длины $N = 3^m - 1$ получено аналитическое выражение, приводящие к построению унимодулярных последовательностей с заданным уровнем боковых лепестков a ПАКФ из диапазона вещественных чисел $a \in [-1; N]$.

Для пятифазных последовательностей длины $N = 5^m - 1$ найдены два решения, приводящие к построению унимодулярных последовательностей с заданным уровнем боковых лепестков a ПАКФ из диапазонов $a \in [-1; 5^{m-1} - 1]$ для первого возможного решения и $a \in [5^{m-1} - 1; N]$ для второго возможного решения, соответственно.

Литература

[1] *Гантмахер В. Е., Быстров Н. Е., Чеботарёв Д. В.* Шумоподобные сигналы. Анализ, синтез, обработка. — Санкт-Петербург: Наука и техника, 2005. — 400 с.

[2] *Ипатов В. П.* Периодические дискретные сигналы с оптимальными корреляционными свойствами. — Москва: Радио и связь, 1992. — 152 с.

[3] *Ипатов В. П.* Широкополосные системы и кодовое разделение сигналов. Принципы и приложения. — Москва: Техносфера, 2007.

[4] *Golomb S. W., Gong G.* Signal design for good correlation for wireless communication, cryptography, and radar. — New York: Cambridge University Press, 2005. — 438 p.

[5] *Fan P., Darnell M.* Sequences Design for Coramunicational Applications. — Taunton, Somerset, England: RSP Ltd, 1996. — 493 p.

[6] *Antweiler M., Bomer L.* Complex sequences over $GF(p^m)$ with a two-level autocorrelation function and a large linear span // IEEE Trans Inform Theory. — Vol. 38. — 1992. — Pp. 120–130.

[7] *Gong G.* Q-ary cascaded GMW sequences. // IEEE Trans Inform Theory. — Vol. 42. — 1996. — Pp. 263–267.

[8] *Амиантов И. Н.* Избранные вопросы статистической теории связи. — Москва: Сов. радио, 1971. — 416 с.

[9] *Леухин А. Н., Тюкаев А. Ю., Бахтин С. А., Корнилова Л. Г.* Новые фазокодированные последовательности с хорошими корреляционными характеристиками // Электромагнитные волны и электронные системы. — 2007. — № 6. — С. 51–54.

Теория вейвлет-подобных преобразований типа Хаара над конечными полями

Жарких А. А.
zharkikh090107@mail.ru
Мурманск, Мурманский государственный технический университет

В работе представлена теория вейвлет-подобных преобразований типа Хаара (ВППХ) над конечными полями. Вейвлет-подобные преобразования типа Хаара получаются из вейвлет-преобразований Хаара (ВПХ) путем замены матрицы преобразований на матрицу более общего вида. Арифметика конечного поля позволяет реализовать точно обратимые семейства этих преобразований без ошибок округления. Абсолютная точность таких преобразований, и их линейная сложность при однократном применении, позволяют прогнозировать возможность их эффективного использования для обработки изображений, аудио и текста без потерь. Целесообразно использовать эти преобразования для сжатия данных и для формирования признаков в автоматизированных системах распознавания образов.

Цель данной работы — представление теории вейвлет-подобных преобразований над простыми конечными полями и их расширениями.

Критерии оценки качества работы любой информационной системы субъективны. Тем не менее, можно выделить два основных аспекта в оценке её качества:

1) качество передачи, приема, хранения и обработки информации;
2) качество восприятия информации человеком.

Эти аспекты оценки качества не тождественны. Если конечным потребителем информации является человек, то в систему необходимо включать интерфейс адаптации информации к человеческому восприятию. Передача, прием, хранение и обработка информации могут осуществляться с потерями и без потерь. Человеку безразлично, существуют потери или нет, если он не обнаруживает их с помощью своих органов чувств. При автоматическом функционировании системы без участия человека важно свести любые потери к минимуму, или вообще устранить некоторые из них. Можно выделить два основных вида потерь:

1) потери, обусловленные каналами передачи и элементами памяти;
2) потери, обусловленные природой и несовершенством алгоритмов обработки.

В силу этого растет актуальность технических задач связанных с хранением в памяти, вычислительной обработкой и передачей по каналам связи информации различной физической природы. Хранение, обработка и передача информации неразрывно связаны с понятием сигнала. С математической точки зрения сигнал — это однозначная или многозначная функция одной или нескольких переменных. Среди однозначных сигналов одной переменной в зависимости от области определения и области возможных значений различают аналоговые ($\mathbb{R} \to \mathbb{R}$), дискретные ($\mathbb{Z} \to \mathbb{R}$), квантованные ($\mathbb{R} \to \mathbb{Z}$) и цифровые сигналы ($\mathbb{Z} \to \mathbb{Z}$). Возможны сигналы и смешанного вида. В силу ограниченности ресурсов компьютера во многих приложениях можно считать, что цифровой сигнал реализует отображение $\mathbb{Z}_M \to \mathbb{Z}_N$ при некоторых натуральных $M, N > 1$. С вычислительной точки зрения функция $f : \mathbb{Z}_M \to \mathbb{Z}_N$ может быть реализована различными способами:

1) в арифметике поля характеристики 0 ($\mathbb{Q}$ или $\mathbb{R}$);
2) в арифметике кольца $\mathbb{Z}_N$, N — составное число;
3) в арифметике поля $\mathbb{Z}_P = GF(P)$, P — простое число;
4) в арифметике поля $GF(P^K)$, P — простое число, $K > 1$ натуральное число;
5) в арифметике более сложных алгебраических структур.

Первый способ имеет ошибки округления (при использовании $\mathbb{R}$) или требует существенного увеличения объёма памяти (при использовании $\mathbb{Q}$). Второй способ плохо согласуется с «обычной» арифметикой (существуют делители нуля, не каждый ненулевой элемент имеет обратный), что не позволяет адекватно трактовать физические результаты. Третий и четвертый способы свободны от недостатков первых двух. Однако второй, третий и четвёртый способы не всегда отражают реальные результаты физических наблюдений.

Вейвлет-преобразование в базисе Хаара над произвольным полем

Практически важные вейвлеты традиционно определяются как функции одной вещественной переменной с вещественными значениями. В зависимости от математической модели (структуры области определения, структуры области возможных значений и вида преобразований) различают дискретные и непрерывные вейвлеты. Особенностью так определённых вейвлетов являются ошибки вычислений в вещественной арифметике с плавающей точкой.

В работах [1, 2] можно ознакомиться с теорией и практическими применениями различных вейвлетов с вещественным и значениями. В работах [3, 4] можно ознакомиться с тем, как вейвлеты

используются для обработки изображений и сигналов. Из книг [5, 6] можно получить исчерпывающее представление о полях Галуа и точных алгоритмах вычисления над полями Галуа. Работы [7–10] достаточно подробно освещают современное состояние теории вейвлетов над конечными полями (полями Галуа). Статьи [11–13] показывают перспективы использования вейвлетов для сжатия изображений. Вейвлет-преобразования в базисе Хаара определяются для линейных пространств размерности $N = 2^n$ над произвольным полем любой характеристики соотношениями (1), (2). Очевидно что характеристика поля должна быть отлична от 2.

$$\begin{aligned} C_s^{(k+1)} =& \frac{C_{2s}^{(k)} + C_{2s+1}^{(k)}}{2}, \\ C_{s+N_{k+1}}^{(k+1)} =& \frac{C_{2s}^{(k)} - C_{2s+1}^{(k)}}{2}, \\ s =& 0, \ldots, N_{k+1}. \end{aligned} \tag{1}$$

$$C_s^{(k+1)} = C_s^{(k)}, s = 2N_{k+1}, \ldots, N-1. \tag{2}$$

В представленных формулах (1), (2) $N = 2^n$, $N_0 = N$, $N_{k+1} = N_k/2 = 2^{n-k+1}, k = 0, \ldots, n-1$. k — номер шага вейвлет-преобразования. При $k = 0$ не используются соотношения (2).

Согласно установившейся терминологии, в формулах (1) полусуммы формируют высокочастотные компоненты различных масштабов и положений, а полуразности — низкочастотные компоненты различных масштабов и положений. Таким образом, исходный одномерный вектор отображается в двумерный образ, отражающий содержание составляющих разных масштабов и положений в исходном векторе. Вейвлет-преобразование Хаара существует в поле характеристики 0 и в поле $GF(P^K)$ нечётной характеристики, так как в любом таком поле обратим элемент равный 2. Однако поле $GF(P^K)$ нечётной характеристики плохо согласуется с компьютерными форматами данных. Вычисления в поле характеристики 0 имеют ошибки округления. Поэтому на основе анализа свойств простейшего вейвлет-преобразования Хаара вводятся два семейства вейвлет-подобных преобразований. Одно из них двухпараметрическое несимметрическое, а второе — однопараметрическое симметрическое. Второе семейство является подмножеством первого. Эти вейвлет-подобные преобразования существуют в конечном поле произвольной характеристики кроме поля $GF(2)$. Далее предлагается выбрать поле $GF(2^K)$, $K > 1$, для которого вейвлет-подобные преобразования существуют и очень хорошо согласуются с машинными форматами данных.

Преобразование подобное вейвлет преобразованию в базисе Хаара над произвольным полем

Вейвлет-преобразования в базисе Хаара обладают двумя полезными свойствами, связанными с вычислениями высокочастотных и низкочастотных составляющих:

1) если два элемента совпадают, то один из результатов преобразования равен любому из них;
2) если два элемента совпадают, то второй результат преобразования равен нулю.

Эти свойства оказываются полезными в обработке изображений. Как правило, изображения содержат достаточно большие области с одинаковыми или близкими значениями кодов пикселей. Поэтому, если после преобразования большое число пикселей изображения обращаются в нуль или невелики по модулю, то такое изображение можно представить более коротким кодом, чем исходное, то есть, фактически, сжать. Заменим в преобразованиях (1) полусумму и полуразность любыми преобразованиями, обладающими указанными двумя свойствами. В результате получим преобразования (3) и (4). Будем называть преобразование, определяемое выражениями (3) и (4), преобразованием подобным вейвлет-преобразованию Хаара (ВППХ).

$$\begin{aligned} D_s^{(k+1)} =& \beta D_{2s}^{(k)} + (1-\beta) D_{2s+1}^{(k)}, \\ D_{s+N_{k+1}}^{(k+1)} =& \gamma D_{2s}^{(k)} - \gamma D_{2s+1}^{(k)}, \\ s =& 0, \ldots, N_{k+1}. \end{aligned} \tag{3}$$

$$D_s^{(k+1)} = D_s^{(k)},\ s = 2N_{k+1}, \ldots, N-1. \tag{4}$$

Преобразования (3), (4) могут быть выполнены над элементами произвольного поля. Для практики интересен случай невырожденности матрицы $L = \begin{pmatrix} \beta & 1-\beta \\ \gamma & -\gamma \end{pmatrix}$. В этом случае реализуется двухпараметрическое семейство ВППХ. Если же учесть ещё одно свойство ВПХ — симметрию матрицы $H = \begin{pmatrix} +1/2 & +1/2 \\ +1/2 & -1/2 \end{pmatrix}$, то получим однопараметрическое симметрическое семейство ВППХ с матрицей $S = \begin{pmatrix} 1-\gamma & \gamma \\ \gamma & -\gamma \end{pmatrix}$.

Преобразование подобное вейвлет преобразованию в базисе Хаара над конечным полем

Напомним, что простое поле $GF(P)$ представляет собой кольцо классов вычетов по модулю простого числа, а расширение поля $GF(P^K)$ — факторкольцо полиномов по модулю неприводимого полинома над $GF(P)$. Число нормированных неприводимых полиномов K-й степени над $GF(P)$ [5] равно:

$$N(K,P) = \frac{1}{K} \sum_{d|K} \mu(d) P^{\frac{K}{d}}. \tag{5}$$

Здесь $\mu(\bullet)$ — функция Мёбиуса из теории чисел. Любой из этих неприводимых полиномов может быть использован для вычислений в арифметике расширения поля Галуа. Однако более привлекательными оказываются вычисления на основе нормированных неприводимых примитивных полиномов K-й степени над $GF(P)$. Вычисления упрощаются за счёт использования свойств мультипликативной группы поля. Число нормированных, неприводимых, примитивных полиномов [5] равно:

$$M(K,P) = \frac{\varphi(P^K - 1)}{K}. \qquad (6)$$

Здесь $\varphi(\bullet)$ — это функция Эйлера из теории чисел. В практических приложениях к обработке любых данных лучше использовать конечные поля характеристики 2 $GF(2^K)$, $K > 1$. Это прежде всего связано с компьютерной арифметикой. К примеру, формулы (7) — выражения для неприводимых примитивных полиномов над $GF(2)$ из таблиц второго тома книги [5]. Их можно использовать для вычислений в расширении поля Галуа $GF(2^K)$ с $K = 4, 8, 16, 24, 32$.

$$\begin{aligned} f_4(x) &= x^4 + x + 1; \\ f_8(x) &= x^8 + x^7 + x^6 + x + 1; \\ f_{16}(x) &= x^{16} + x^{12} + x^3 + x + 1; \\ f_{24}(x) &= x^{24} + x^7 + x^2 + x + 1; \\ f_{32}(x) &= x^{32} + x^{22} + x^2 + x + 1. \end{aligned} \qquad (7)$$

Для таких значений K элемент поля $GF(2^K)$ требует для хранения соответственно полубайт, байт, слово, 3 байта, двойное слово.

Выводы

На основе приведённого исследования можно сделать следующие выводы.

1) Ослабление требований к структуре матрицы ВПХ расширяет класс этих преобразований. Получающийся класс преобразований можно назвать ВППХ.
2) Можно выделить два больших семейства ВППХ — образованное невырожденными матрицами и образованное симметрическими невырожденными матрицами. Второе семейство является подсемейством первого;
3) ВППХ может быть реализовано над любым полем любой характеристики за исключением $GF(2)$.
4) ВППХ над конечным полем реализует взаимнооднозначное отображение векторного пространства над этим полем без ошибок округления.
5) В качестве приложений к компьютерной обработке данных наибольший интерес представляют собой ВППХ над полем $GF(2^K)$, $K > 1$.

Приведённые выводы позволяют заключить, что ВППХ над конечными полями интересны как математический объект и как инструмент для обработки мультимедийной информации (аудио, изображения, видео, текст и др.) без ошибок округления.

Литература

[1] *Чуи Ч.* Введение в вэйвлеты. — Москва: Мир, 2001. — 412 с.

[2] *Яковлев А. Н.* Введение в вейвлет-преобразования. — Новосибирск: НГТУ, 2003. — 104 с.

[3] *Столниц Э. Р., ДеРоуз Т., Салезин Д.* Вейвлеты в компьютерной графике. — Ижевск: НИЦ Регулярная и хаотическая динамика, 2002. — 272 с.

[4] *Малла С.* Вэйвлеты в обработке сигналов. — Москва: Мир, 2005. — 671 с.

[5] *Лидл Р., Нидеррайтер Г.* Конечные поля. В 2-х томах. — Москва: Мир, 1988. — 820 с.

[6] *Блейхут Р.* Быстрые алгоритмы цифровой обработки сигналов. — Москва: Мир, 1989. — 448 с.

[7] *Fekri F., Mersereau R. M., Schafer R. W.* Theory of wavelet transform over finite fields // Proceedings of the Acoustics, Speech, and Signal Processing (ICASSP), 1999. — Vol. 03. — Pp. 1213–1216.

[8] *Fekri F., Mersereau R. M., Schafer R. W.* Theory of paraunitary filter banks over fields of characteristic two // IEEE Transactions on Information Theory, 2002. — Vol. 48. — Pp. 2964–2979.

[9] *de Oliveira H. M., Falk T. H., Tavora R.* Wavelet Decomposition over Finite Fields // Journal of the Brazilian Telecom.Society. — 2007. — Vol. 17, №1. — Pp. 38–47.

[10] *Phoong See-May., Vaidyanathan P. P.* Paraunitary Filter Banks Over Finite Fields // IEEE Transactions on Signal Processing, 1997. — Vol. 45. — Pp. 1443–1457.

[11] *Pan H., Siu W.-C., Law N.-F.* Lossless image compression using binary waveletTransform // IET Image Process, 2007. — Vol. 1(4). — Pp. 353–362.

[12] *Chengyi Xiong, Jinwen Tian, Jian Liu* Efficient Architectures for Two-Dimensional Discrete Wavelet Transform Using Lifting Scheme // IEEE TRANSACTIONS ON IMAGE PROCESSING, 2007. — Vol. 16, No. 3.

[13] *Bhuyan M. S.,Nowshad Amin, Md.Azrul Hashi Madesa, Md.Shabiul Islam* FPGA Realization of Lifting Based Forward Discrete Wavelet Transform for JPEG 2000 // INTERNATIONAL JOURNAL OF CIRCUITS, SYSTEMS AND SIGNAL PROCESSING Issue 2, 2007. — Vol. 1.

Аддитивный спектральный подход к выявлению нечетких кластеров по матрице связи*

Миркин Б. Г.[1]***, Насименто С. А.***[2]
[1]bmirkin@hse.ru [2]snt@di.fct.unl.pt
[1]Департамент информатики, Биркбек Колледж Лондонского Университета, Лондон, Великобритания и Отделение прикладной математики и информатики, НИУ Высшая школа экономики, Москва, Российская Федерация [2] Департамент информатики и Центр искусственного интеллекта ЦЕНТРИА, Новый университет Лиссабона, Капарика, Португалия

Разработан новый метод выявления нечетких кластеров. Метод основан на кластерной модели, представляющей собой обобщение известного разложения симметричной матрицы по собственным векторам и значениям. Основная процедура метода — последовательное извлечение кластеров, что позволяет использовать естественные правила ее остановки и, тем самым, определения числа кластеров. Сравнение метода с известными популярными алгоритмами, как на реальных, так и сгенерированных данных, показывает его высокую конкурентоспособность.

Введение

Имеется довольно много алгоритмов нечеткого кластер-анализа матриц связи [1, 2, 3, 4, 10]. Однако большинство из них основаны на эвристических критериях и, более того, их параметры, такие как число кластеров, должны задаваться пользователем «вручную» по интуиции, поскольку не существует никаких методик, помогающих при их выборе. Мы используем модель аддитивных кластеров и спектральный подход к кластер-анализу, чтобы вывести новый метод нечеткого кластер-анализа, который ассоциирован с набором вытекающих из модели характеристик, которые помогают выбрать параметры метода.

Предполагается, что данные представлены в виде матрицы связи $W = (w_{tt'})$, $t, t' \in T$, где величина $w_{tt'}$ характеризует уровень связи между объектами t, t' из множества T. Мы рассматриваем эти связи как измеряемое выражение неизвестных нам образов, представимых в виде нечетких кластеров. Нечеткий кластер, согласно этому подходу, характеризуется двумя элементами:

1) вектором принадлежности $\mathbf{u} = (u_t)$, $t \in T$, такой что $0 \leqslant u_t \leqslant 1$ для всех $t \in T$;

2) интенсивностью $\mu > 0$, которая выражает степень выраженности образа, представленного кластером.

Интенсивность применяется к $\mathbf{u}$ как масштабирующий множитель, так что именно произведение $\mu\mathbf{u}$, а не индивидуальные множители, учитываются в матрице связи. При заданном произведении μu_t, нельзя отделить μ от u_t, так что мы используем традиционное правило. Мы ограничиваем шкалу вектора принадлежности неким постоянным значением, т. е. условием типа того что $\sum_t u_t = 1$ или $\sum_t u_t^2 = 1$. Тогда частное от деления значения суммы на это постоянное значение будет равно значению μ. Мы используем вторую нормировку, евклидовой нормой, поскольку она явно связана с используемой далее моделью нечеткого кластер-анализа.

Чтобы резче проявить скрытую структуру кластеров на матрице связи, мы применяем подход спектрального кластер-анализа, в котором исходная матрица связи W подвергается так называемому нормализованному преобразованию Лапласа, которое в спектральном подходе привязывается к популярному комбинаторному критерию минимизации нормализованного разреза [6]. Этот критерий связан с минимальным ненулевым собственным числом матрицы Лапласа. Чтобы иметь дело с максимальным собственным вектором, как этого требует наша модель, мы используем псевдообратную матрицу для матрицы Лапласа, имеющую те же собственные вектора; собственные же числа получаются обращением, что увеличивает расстояния между ними и тем самым улучшает применимость нашего подхода.

Модель и метод

Предлагаемая аддитивная модель предполагает существование K нечетких кластеров, которые воспроизводят псевдообратные Лапласовы связи $a_{tt'}$ с точностью до аддитивных невязок:

$$a_{tt'} = \sum_{k=1}^{K} \mu_k^2 u_{kt} u_{kt'} + e_{tt'}, \qquad (1)$$

где $\mathbf{u}_k = (u_{kt})$ — вектор принадлежности кластера k, а μ_k — его интенсивность.

Слагаемое $\mu_k^2 u_{kt} u_{kt'}$ — это произведение $\mu_k u_{kt}$ и $\mu_k u_{kt'}$, выражающих участие объектов t и t' в кластере k. Сумма этих величин должна равняться связи $a_{tt'}$ между объектами t и t'. Величина μ_k^2 выражает вклад интенсивности и далее называется весом кластера.

Чтобы найти разложение (1), мы используем критерий наименьших квадратов $e_{tt'}^2$ и стратегию

[1] Работа поддержана грантом PTDC/EIA/69988/2006 Фонда науки и техники Португалии, 2007–2011. [2] Работа частично финансировалась Лабораторией анализа и выбора решений НИУ ВШЭ, Москва, РФ.

метода главных компонент, согласно которой слагаемые суммы в (1) отыскиваются по одному. Это означает, что решается задача минимизации

$$E = \sum_{t,t' \in T} (b_{tt'} - \xi u_t u_{t'})^2 \qquad (2)$$

относительно неизвестных положительных весов ξ, так что $\mu = \sqrt{\xi}$, и нечетких векторов принадлежности $\mathbf{u} = (u_t)$, при заданной матрице связи $B = (b_{tt'})$.

В самом начале, B равна A. Но после каждого шага найденный нечеткий кластер вычитается из B, так что остаточная матрица связи, используемая на следующем шаге, определяется как $B - \mu^2 \mathbf{u}\mathbf{u}'$ где μ и $\mathbf{u}$ — это интенсивность и вектор принадлежности только что найденного кластера. В результате A будет аддитивно представлено в виде (1).

Нетрудно доказать, что оптимальное значение ξ при заданном $\mathbf{u}$ равно

$$\xi = \frac{\mathbf{u}'B\mathbf{u}}{(\mathbf{u}'\mathbf{u})^2}, \qquad (3)$$

что, очевидно, неотрицательно, если B положительно полуопределенная матрица.

Подставив это значение ξ в (2), получим

$$E = S(B) - \xi^2 (\mathbf{u}'\mathbf{u})^2,$$

где $S(B) = \sum_{t,t' \in T} b_{tt'}^2$ — это так называемый разброс данных. Обозначим последнее слагаемое через

$$G(\mathbf{u}) = \xi^2 (\mathbf{u}'\mathbf{u})^2 = \left(\frac{\mathbf{u}'B\mathbf{u}}{\mathbf{u}'\mathbf{u}}\right)^2, \qquad (4)$$

что дает разложение $S(B) = G(\mathbf{u}) + E$ разброса данных на две части, $G(\mathbf{u})$, объясненную кластером $(\mu, \mathbf{u})$, и E, необъясненную. Это означает, что оптимальный кластер максимизирует объясненную часть $G(\mathbf{u})$ в разложении (4). Другими словами, кластер максимизирует

$$g(\mathbf{u}) = \xi \mathbf{u}'\mathbf{u} = \frac{\mathbf{u}'B\mathbf{u}}{\mathbf{u}'\mathbf{u}}. \qquad (5)$$

А это — не что иное как отношение Рэлея, хорошо известное в теории матриц: его максимум равен максимальному собственному значению матрицы B, достигаемому на соответствующем собственном векторе, конечно, если на вектор u не наложено никаких ограничений.

Таким образом, рассматриваемая проблема может решаться с помощью спектрального подхода, согласно которому сначала отыскивается максимальное собственное значение λ и соответствующий нормированный собственный вектор z для матрицы B, $[\lambda, z] = \Lambda(B)$, после чего рассчитывается его проекция на множество векторов, допустимых в качестве векторов принадлежности.

Остановка процесса последовательного извлечения нечетких кластеров и, соответственно, выбор числа кластеров K, производится, когда справедлив хотя бы один из следующих критериев:

1. Оптимальное значение ξ (3) для спектрального нечеткого кластера отрицательно;
2. Вклад извлеченного нечеткого кластера стал слишком небольшим, меньше, чем фиксированное значение$\tau > 0$;
3. Разброс остаточных связей стал слишком небольшим, меньше, чем фиксированное значение ε; например, меньше 5% первоначального разброса связей.

Некоторые эксперименты

Мы называем этот «Fuzzy ADDItive-Spectral» метод извлечения кластеров FADDIS (ФАДДИС). Была проведено экспериментальное сравнение работы ФАДДИСа с работой других подходов на нескольких типах данных, таких как

(a) обыкновенные невзвешенные графы (сети) для обнаружения сообществ,

(b) данные связей аффинности, получаемые из обычных данных объект-признак с помощью кернел-функций,

(c) небольшие данные о близости, часто используемых для сравнения алгоритмов в литературе, и

(d) реальные данные о связях между объектами, включая полученные в нашем собственном исследовании исследовательской активности [7].

Рассмотрим два из экспериментов:

(1) по выявлению сообществ на обыкновенных графах и

(2) по сравнению ФАДДИС с различными версиями известного метода с-средних для нечеткого кластер-анализа [1].

Выявление сообществ. Исследования по этой тематике недавно были возоблены Ньюманом и др., прежде всего, с использованием так называемого критерия модулярности и спектрального подхода (см., например, [8, 6]). Обыкновенный граф на множестве вершин T представляют симметричной матрицей связи $A = (a_{tt'})$ на T, так что $a_{tt'} = 1$ для t и t', связанных ребром графа, и $a_{tt'} = 0$ в остальных случаях. Затем матрицу A симметризуют преобразованием $(A + A')/2$, после чего диагональные элементы зануляются, $a_{tt} = 0$ для всех $t \in T$. Компоненты связности анализируются по отдельности.

Спектральная релаксация задачи [8] использует вычитание «фоновых случайных связей» из матрицы $A = (a_{tt'})$. Эта работа была перенесена на случай нечетких кластеров в [11]. Поскольку матрица A неотрицательна, ее первый собственный вектор тоже неотрицателен и, как правило, положителен, как хорошо известно в матричной алгебре. Это означает, что соответствующий первый кластер,

фактически, совпадает с этим собственным вектором и выражает не столько кластерную структуру связей, сколько факт общей взаимосвязанности вершин компоненты графа. Поэтому его следует отбросить, а в качестве элементов решения рассматривать только последующие нечеткие кластеры. При этом переход к остаточной матрице связи путем вычитания связей по первому собственному вектору аналогичен вычитанию фоновых случайных связей в методах, основанных на модулярности, но обладает тем преимуществом, что учитывает не только парные, но и косвенные связи между объектами.

Кроме того, ожидается, что число положтельных собственных значений матрицы графа невелико, что должно быстро привести к остановке процесса ФАДДИС.

Опишем применение алгоритма ФАДДИС к популярному графу, характеризующему отношения между членами секции карате Захарий, который содержит 34 вершины (по числу участников) и 78 ребер (см. [8, 11], где, в частности, можно найти ссылки на страницы веба, содержащие этот граф). Члены секции разделены по их отношению к двум ее руководителям: администратору и тренеру. Поэтому данная сеть считается разделенной на два сообщества, из 18 и 16 членов соответственно, по тому, с кем из руководителей они ассоциируют себя. В применении к этим данным ФАДДИС порождает всего 3 нечетких кластера, поскольку вклад четвертого кластера составляет всего 2,4% от разброса исходных связей, а это меньше, чем $1/34$ — средний вклад одного индивидуума. При этом первый кластер, как отмечено выше, на самом деле — фоновый, а два других кластера в точности соответствуют ожидаемой структуре сообществ в сети. Некоторые характеристики этих кластеров содержатся в таблице 1.

Таблица 1. Характеристики кластеров, найденных по графу секции карате алгоритмом ФАДДИС.

Кластер	Вклад, %	λ_1	Вес	Интенс.
I	29,00	3,36	3,36	1,83
II	4,34	2,49	1,30	1,14
III	4,19	2,00	0,97	0,98

Метод, предложенный в [11] находит три кластера, два из которых в основном соответствуют ожидаемым группировкам, хотя и со значительным перекрыванием между кластерами. Третий же, меньший кластер, состоит из пяти индивиддумов 5, 6, 7, 11, 17, которые все принадлежат одной из группировок [11], стр. 487. По нашему мнению, этот третий кластер соответствует третьему собственному вектору матрицы графа, отражающему число смежных вершин: именно эти вершины имеют максимальное число соседей в графе. По крайней мере, очевидно, что на этих данных ФАДДИС работает лучше, чем метод из [11].

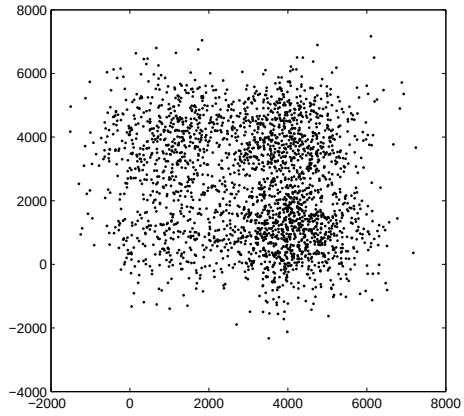


Рис. 1. Данные четырех Гауссовых двумерных распределений [2].

Вычленение кластеров, сильно перекрывающихся в пространстве. Здесь, на самом деле, речь идет о продолжении эксперимента, описанного в [2]. А именно, множество двумерных точек порождается как смесь четырех сферических Гауссовых распределения с стандартным отклонением 950 и центрами в точках $(1000, 1000)$, $(1000, 4000)$, $(4000, 1000)$ и $(4000, 4000)$ (см. Рис. 1).

Брауэр [2] преобразовал эти данные в матрицу D Евклидовых расстояний между точками и применил к ним пять версий метода нечетких с-средних. Три из этих методов переносят методику с-средних на данные о близостях: Roubens [9], Windham [10] и NERFCFM [4]. Два метода комбинируют сам метод нечетких с-средних с процедурами, преобразующими расстояния в формат объект-признак, FastMap и SMACOF [2]. Ощутимо наилучшие результаты из этих пяти получены применением метода нечетких с-средних к пятимерному множеству, извлеченному из D с помощью метода FastMap [2], который дает среднюю точность воспроизведения исходных кластеров по модифицированному индексу Рэнда (МИР) [5] равную 0,67 [2]. ФАДДИС был применен к 10 случайным версиям данных, порожденных из теми же Гауссовыми кластерами. Из каждого ФАДДИС извлекал 5 последовательных нечетких кластеров, останавливаясь по критерию 2: «слишком мал вклад следующего кластера». Затем первый кластер удалялся как просто отражающий общую связь между объектами, а оставшиеся 4 превращались в разбиения так, что каждый объект приписывался к тому кластеру, на котором значение его принадлежности было

максимально. Средние значения модифицированного индекса Рэнда (МИР) [5]: 0,70 (0,03) на данных с 500 и 1000 сгенерированными объектами, и МИР=0,73 (0,01) на данных для 2500 сгенерированных объектах (в скобках — значения стандартного отклонения по 10 множествам), что превышает наилучшие значения в экспериментах, описанных в [2].

Литература

[1] *Bezdek J., Keller J., Krishnapuram R., Pal T.* Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. — Kluwer Academic Publishers, 1999.

[2] *Brouwer R.* A method of relational fuzzy clustering based on producing feature vectors using FastMap // Information Sciences. — 2009. — V. 179. — Pp. 3561–3582.

[3] *Davé R., Sen S.* Robust fuzzy clustering of relational data // IEEE Transactions on Fuzzy Systems. — 2002. — V. 10. — Pp. 713–727.

[4] *Hathaway R. J., Bezdek J. C.* NERF c-means: Non-Euclidean relational fuzzy clustering // Pattern Recognition. — 1994. — V. 27. — Pp. 429–437.

[5] *Hubert L. J., Arabie P.* Comparing partitions // Journal of Classification. — 1985. — V. 2. — Pp. 193–218.

[6] *von Luxburg U.* BibTitleA tutorial on spectral clustering // Statistics and Computing. — 2007. — V. 17. — Pp. 395–416.

[7] *Mirkin B., Nascimento S.* Analysis of Community Structure, Affinity Data and Research Activities using Additive Fuzzy Spectral Clustering // Technical Report 6, School of Computer Science, Birkbeck University of London, 2009.

[8] *M. Newman, M. Girvan.* Finding and evaluating community structure in networks // Phys. Rev. E. — 2003. — V. 69, N. 2.

[9] *Roubens M.* Pattern classification problems and fuzzy sets // Fuzzy Sets and Systems. — 1978. — V. 1. — Pp. 239–253.

[10] *Windham M. P.* Numerical classification of proximity data with assignment measures // Journal of Classification. — 1985. — V. 2. — Pp. 157–172.

[11] *S. Zhang, R.-S. Wang, X.-S. Zhang.* Identification of overlapping community structure in complex networks using fuzzy c-means clustering // Physica A. — 2007. — V. 374. — Pp. 483–490.

Кластеризация разнотипных данных, содержащих пропуски, с применением ансамблевого подхода*

Бериков В. Б.
berikov@math.nsc.ru
Новосибирск, Институт математики им. С. Л. Соболева СО РАН

Рассматривается задача автоматической группировки разнотипных данных, содержащих пропуски. Предлагается метод решения, основанный на ансамблевом подходе. Исследуется вероятностная модель ансамбля. Описывается алгоритм, строяший ансамблевое решение на основе таксономических решающих деревьев. Приводятся результаты экспериментов, подтверждающие эффективность алгоритма.

Кластеризация объектов по схожести их характеристик необходима для компактного описания больших массивов информации, обнаружения закономерностей, выделения наиболее типичных представителей объектов исследования. Известные подходы к решению этой задачи зависят от способа понимания похожести и различия объектов. Так, например, в вероятностном подходе считается, что наблюдаемые объекты принадлежат различным классам, причем каждый класс характеризуется вероятностным распределением с неизвестными параметрами. Другой, «геометрический» подход использует аналогии с классификацией, которую проводит человек при анализе изображений на плоскости или в трехмерном пространстве.

В настоящее время наблюдается тенденция распространения методов и подходов, разработанных в области распознавания образов и прогнозирования, на задачу кластерного анализа. Несмотря на большое число существующих методов и алгоритмов кластерного анализа, теоретическая проработка данной области, глубина и охват решаемых вопросов намного отстает от других разделов теории интеллектуального анализа данных и машинного обучения. Это можно объяснить сложностью формализации этой проблемы, обусловленной тем, что процесс группировки в значительной степени носит субъективный характер.

Во многих приложениях требуется решать задачу кластерного анализа в случае, когда исходный набор переменных, описывающих группируемые объекты, содержит как количественные, так и качественные элементы. При этом часть наблюдений содержит «пропуски» (неизмеренные значения). Задачи описанного типа нередко возникают при анализе археологических данных, в социологии, медицинской статистике и т. д.

В описываемом случае применение классических методов, основанных на определении расстояний между объектами и поиске оптимального разбиения множества объектов на группы затруднительно: в разнотипном пространстве переменных, при введении расстояний, возникают методологические трудности [1]. Наличие пропусков также осложняет задачу, так как в этом случае не для каждой пары объектов возможно вычислять расстояние, используя весь исходный набор переменных.

В предлагаемой работе для решения поставленной задачи используется ансамблевый подход. Этот подход (называемый также коллективным, комитетным и т. п.) успешно применяется в задачах распознавания и прогнозирования. Он основан на нахождении коллективного решения (например, с помощью процедуры голосования) по отдельным решениям, полученным различными алгоритмами или одним алгоритмом, но с разными параметрами работы; по различным подпространствам переменных, подвыборкам и т. п. Такое рассмотрение задачи с различных точек зрения, взаимное «усиление» устойчивых закономерностей и, наоборот, «ослабление» случайных, неустойчивых позволяет значительно повысить качество итогового решения.

В данной работе качестве базового элемента ансамбля используется алгоритм построения таксономического решающего дерева [2]). Этот алгоритм позволяет проводить группировку объектов в разнотипном пространстве переменных, формирует иерархическую логическую модель группировки, определяет наиболее информативные факторы. В алгоритме предусмотрен анализ таблиц с пропущенными значениями. Объекту, который из-за пропусков не попал ни в одну их сформированных групп, присваивается номер такого кластера, который достижим из вершины дерева, соответствующей данному объекту, и для которого число элементов максимально. Однако такой способ может приводить к формированию «ложных» кластеров, особенно когда процент пропусков велик.

В ансамблевом подходе используется более естественный принцип *голосования* для объектов, содержащих пропуски.

Работа имеет следующий план. В первом параграфе дается постановка задачи. Второй параграф посвящен вероятностной модели кластерного ансамбля, которая вводится как модель распознавания образов со скрытыми (латентными) классами. С помощью модели проводится теоретическое

Работа выполнена при финансовой поддержке РФФИ, проекты №11-07-00346а, 10-01-00113а.

обоснование алгоритма ансамблевого кластерного анализа. В следующих разделах кратко описывается алгоритм кластеризации, основанный на ансамблевом подходе, приводятся результаты численных экспериментов, делаются выводы.

Постановка задачи

Пусть имеется множество $s = \{o^{(1)}, \ldots, o^{(N)}\}$ некоторых объектов, выбранных из генеральной совокупности. Каждый объект описывается с помощью набора переменных $X_1,\ldots,X_m$. Этот набор может включать переменные разных типов (количественные и качественные, под которыми будем понимать номинальные и булевы, а также порядковые). Пусть D_j обозначает множество значений переменной X_j (некоторый интервал числовой оси в случае количественной переменной; конечный набор значений (имен) в случае качественной переменной). Пусть $D = \prod\limits_j D_j$. Обозначим через

$$x = x(o) = \big(x_1(o), \ldots, x_m(o)\big)$$

набор наблюдений переменных для объекта o, где $x_j(o)$ есть значение переменной X_j для данного объекта. Часть значений может быть неизвестно; этот факт будем обозначать как $x_j(o) = U$, где U — код пропуска. Соответствующий множеству объектов набор наблюдений будем представлять в виде таблицы данных с N строками и m столбцами.

Требуется разбить объекты на некоторое число K кластеров ($K << N$) так, чтобы критерий качества группировки принял бы оптимальное значение. Число классов может быть как выбрано заранее, так и не задано (в последнем случае оптимальное количество кластеров должно быть определено автоматически).

Предположим, что имеется некоторая скрытая (непосредственно ненаблюдаемая) переменная Y, которая задает принадлежность каждого объекта к некоторому из $K \geqslant 2$ классов. Каждый класс характеризуется определенным законом условного распределения

$$p(x \,|\, Y = k) = p_k(x),$$

$k = 1, \ldots, K$. Рассмотрим следующую вероятностную модель генерации данных. Пусть для каждого объекта определяется класс, к которому он относится, в соответствии с априорными вероятностями

$$\mathbb{P}(Y = k) = P_k,$$

где $\sum_{k=1}^{K} P_k = 1$. Затем в соответствии с распределением $p_k(x)$ определяется значение x. Указанная процедура проводится независимо для каждого объекта. После этого проводится случайный отбор наблюдений, которым присваивается код пропуска. Доля таких наблюдений определяется заданным параметром P_U. Будем считать, что $0 \leqslant P_U \leqslant 1/2$.

Пусть с помощью некоторого алгоритма кластерного анализа α строится разбиение заданного подмножества объектов $s' \subseteq s$ на K подмножеств. Под группировочным решением для объекта $o' \in s'$ будем понимать приписанный ему номер кластера $\alpha(o')$. Если некоторый объект o ($o \in s$) не рассматривался алгоритмом (из-за наличия пропусков в описании объекта), то условимся приписывать ему код пропуска: $\alpha(o) = U$. Группировочной решающей функцией назовем отображение

$$f\colon s \to \{1, \ldots, K, U\}.$$

Поскольку нумерация кластеров не играет роли, удобнее рассматривать отношение эквивалентности, т. е. определять, относит ли алгоритм α каждую пару наблюдений в один и тот же класс, либо в разные классы. Рассмотрим произвольную пару a, b различных объектов выборки s. Обозначим соответствующие наблюдения через x^a и x^b. Определим для пары объектов a, b, таких, что $\alpha(x^a) \neq U$ и $\alpha(x^b) \neq U$ величину

$$h_{\alpha,x^a,x^b} = I\left(\alpha(x^a) \neq \alpha(x^b)\right),$$

где $I(\cdot)$ — индикаторная функция. Если же $\alpha(x^a) = U$ или $\alpha(x^b) = U$, то положим $h_{\alpha,x^a,x^b} = U$.

Пусть

$$P_Y^{a,b} = \mathbb{P}(Y(a) \neq Y(b))$$

— вероятность события, заключающегося в том, что объекты относятся к различным классам. Обозначим вероятность ошибки, которую может совершить алгоритм α при классификации x^a и x^b через $P_{er,\alpha}^{a,b}$, где

$$P_{er,\alpha}^{a,b} = \begin{cases} P_Y^{a,b}, & \text{если } h_{\alpha,x^a,x^b} = 0; \\ 1 - P_Y^{a,b}, & \text{если } h_{\alpha,x^a,x^b} = 1. \end{cases}$$

(предполагается, что $h_{\alpha,x^a,x^b} \neq U$).

Модель ансамбля

Предположим, что алгоритм α зависит от случайного вектора параметров $\Theta \in \mathbf{\Theta}$, где $\mathbf{\Theta}$ — некоторое допустимое множество параметров: $\alpha = \alpha(\Theta)$. Например, в алгоритме k-средних результаты работы зависят от случайного исходного разбиения выборки на K подмножеств. Далее, для краткости, будем обозначать $h_{\alpha(\Theta),x^a,x^b}$ через $h^{a,b}(\Theta)$.

Пусть в результате L-кратного применения алгоритма α со случайно и независимо отобранными параметрами $\Theta_1 = \theta_1$, …, $\Theta_L = \theta_L$ получен набор группировочных решающих функций

$$\mathbf{f} = \{f(\theta_1), \ldots, f(\theta_L)\}.$$

Здесь через $\Theta_1, \ldots, \Theta_L$ обозначены независимые статистические копии случайного вектора Θ. Таким образом, для пары объектов a, b получим набор решений

$$h^{a,b}(\theta_1), \ldots, h^{a,b}(\theta_L).$$

Обозначим через $L^{a,b}$ число решений, для которых выполняется:

$$h^{a,b}(\theta_l) \neq U, \quad l = 1, \ldots, L.$$

Коллективным (ансамблевым) решением для a, b по большинству голосов будем называть функцию

$$H^{a,b}(\theta_1, \ldots, \theta_L) = I\Big(\sum\nolimits_{l:h^{a,b}(\theta_l)\neq U} \frac{h^{a,b}(\theta_l)}{L^{a,b}} < \frac{1}{2}\Big). \quad (1)$$

Интересно исследовать поведение коллективного решения в зависимости от числа элементов ансамбля L. Обозначим через $P^{a,b}_{er}(\Theta_1, \ldots, \Theta_L)$ вероятность ошибки, которую может совершать ансамблевый алгоритм при классификации объектов a и b. Рассмотрим поведение вероятности ошибки для коллективного решения. Воспользуемся следующей априорной информацией об алгоритме кластерного анализа. Будем считать, что ожидаемая вероятность ошибочной классификации $\mathbb{E}_\Theta P^{a,b}_{er}(\Theta) < \frac{1}{2}$. Это означает, что, как ожидается, алгоритм α проводит классификацию с лучшим качеством, нежели алгоритм случайного равновероятного выбора.

Теорема 1. *Если для произвольной пары объектов a, b выполняется $\mathbb{E}_\Theta P^{a,b}_{er}(\Theta) < \frac{1}{2}$, то при увеличении числа элементов ансамбля ожидаемая вероятность ошибочной классификации уменьшается, стремясь в пределе к $min(P^{a,b}_Y, 1 - P^{a,b}_Y)$.*

Таким образом, можно сделать вывод о том, что при выполнении указанного условия применение ансамбля позволяет улучшить качество кластеризации.

Построение ансамбля

При построении ансамбля будем рассматривать различные случайные подсистемы переменных, в пространстве которых проводится группировка (предполагается, что число переменных достаточно велико). При этом для кластеризации отбираются объекты, не содержащие пропуски по переменным, входящим в подсистему. Пусть получен набор группировочных решающих функций, которые формируются в результате применения некоторого алгоритма кластерного анализа для выбранной подсистемы переменных (в данной работе, при проведении экспериментов, используется алгоритм построения таксономического решающего дерева [2]).

Для выбора наилучшей согласующей функции используется принцип, основанный на нахождении согласованной матрицы S различия объектов. Пусть $S = \big(S(i,j)\big)$, где

$$S(i,j) = H^{o^{(i)},o^{(j)}}$$

находится в соответствии с (1).

Близкое к нулю значение величины $S(i,j)$ означает, что данные объекты имеют большой шанс попадания в одну и ту же группу. Близкое к единице значение этой величины говорит о том, что шанс оказаться в одной группе у объектов незначителен.

После вычисления согласованной матрицы подобия, для нахождения итогового варианта группировки будем применять стандартный иерархический метод построения дендрограммы, который в качестве входной информации использует попарные расстояния между объектами [3].

Экспериментальное исследование

Для определения качества предложенного ансамблевого алгоритма была разработана процедура статистического моделирования. Процедура состоит в многократном повторении следующих шагов:

— генерировании случайных выборок в соответствии с заданным распределением для каждого класса;
— построении с помощью алгоритма согласованного группировочного решения для каждой выборки;
— определении качества группировки.

После выполнения заданного числа повторов определяется усредненный по всем выборкам показатель качества. Качество группировки определяется как частота правильной попарной классификации объектов (индекс Ранда). Величина индекса, близкая к 1, говорит о высокой степени согласованности построенных алгоритмом решений с «истинной» принадлежностью к кластерам. Усреднение проводилось по 100 случайным выборкам, являющимися реализациями смеси заданных распределений. Ниже даны результаты моделирования для тестового примера.

Пример. Распределение для каждого из $K = 2$ классов является многомерным нормальным с одной и той же диагональной ковариационной матрицей $\Sigma = \sigma I$, где $\sigma = 0.3$. Вектор математических ожиданий для каждого класса выбирается случайно из множества вершин единичного гиперкуба. Из 100 переменных 60 являются количественными, остальные — булевыми (их номера определяются случайным образом). Для булевых переменных, исходные значения, полученные с помощью датчика случайных чисел, округляются до ближайшего целого из множества $\{0; 1\}$. Объем выборки для первого и второго класса равен 25. Доля

пропусков в таблице данных определяется параметром P_U, принимающем значения из интервала $[0,0{,}5]$ с шагом 0,1. Число деревьев в ансамбле задано равным $L = 7$; каждое дерево строится в случайно выбранном подпространстве размерности 3. На рис. 1 приведены значения полученных усредненных показателей качества для ансамблевого алгоритма и для отдельного алгоритма построения дерева. На графиках также отмечены соответствующие 95% доверительные интервалы. Как видно из рисунка, применение ансамбля позволяет существенно улучшить качество группировки по сравнению с отдельным алгоритмом.

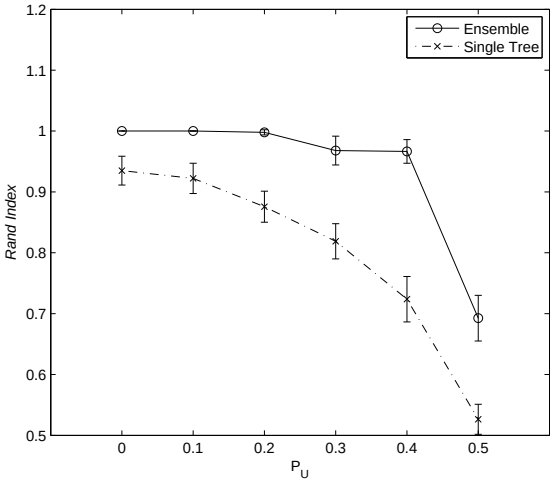


Рис. 1. Результаты работы ансамблевого алгоритма и отдельного алгоритма построения дерева, в зависимости от доли пропусков P_U.

Заключение

В работе рассмотрена задача кластеризации разнотипных данных, при условии, что часть наблюдений содержит пропуски. Для решения задачи использован ансамблевый подход. Введена вероятностная модель анасамбля, основанная на предположении о том, что существует скрытая переменная, определяющая принадлежность к кластерам. Исследование модели показало, что при выполнении определенных условий применение ансамбля позволяет уменьшать вероятность ошибки. Предложен алгоритм, строяший ансамбль на основе таксономических решающих деревьев. Численный эксперимент показал значимо более высокую точность классификации разработанного алгоритма, по сравнению с отдельным таксономическим решающим деревом, при достаточно высокой доле пропусков в таблице данных.

Литература

[1] *Лбов Г. С., Бериков В. Б.* Устойчивость решающих функций в задачах распознавания образов и анализа разнотипной информации. — Новосибирск: Изд-во Ин-та математики, 2005. — 218 с.

[2] *Бериков В. Б.* Построение ансамбля деревьев решений в кластерном анализе // Вычислительные технологии. — 2010. — Т. 15. № 1. — С. 40–52.

[3] *Дуда Р., Харт П.* Распознавание образов и анализ сцен. — М.: Мир, 1976. — 559 с.

Задача диагонализации матрицы связей и задача кластер-анализа*

Двоенко С. Д.
dsd@tsu.tula.ru
Тула, Тульский государственный университет

В случае, когда отсутствует признаковое пространство, множество элементов представлено только результатами их парных сравнений в виде матрицы близостей или расстояний. Для корректной обработки таких данных необходимо модифицировать алгоритмы кластер-анализа и распознавания. Предложен вариант известного алгоритма *k-средних*, в котором не требуется вычислять собственно средние по кластерам. Новый алгоритм сравнивается с алгоритмом диагонализации (агрегирования) матрицы связей. Показано, что алгоритм диагонализации положительно полуопределенной матрицы связей является эвристической версией алгоритма *k-средних*.

Введение

В задаче кластер-анализа объекты $\omega_i \in \Omega$, $i = 1, \ldots, N$ обычно представлены как векторы $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{in})$ в n-мерном пространстве признаков и образуют матрицу данных $X(N, n)$. В соответствии с гипотезой компактности объекты образуют локальные сгущения в виде K кластеров (классов, таксонов).

Хорошо известные алгоритмы типа *k-средних* [1] основаны на идее несмещенного разбиения [2]. В соответствии с ней, каждый кластер Ω_k, $k = 1, \ldots, K$, представлен своим «представителем» $\tilde{\boldsymbol{x}}_k$, а центр кластера представлен средним $\bar{\boldsymbol{x}}_k$.

Если окажется, что для всех кластеров представители и центры совпадают $\tilde{\boldsymbol{x}}_k = \bar{\boldsymbol{x}}_k$, то получена несмещенная кластеризация, а противном случае — смещенная. Тогда необходимо назначить центры (средние объекты) в качестве новых представителей, заново расклассифицировать объекты по минимуму расстояния до представителей и вычислить новые центры кластеров.

В случае, когда признаковое пространство нам недоступно, средний объект $\omega(\bar{\boldsymbol{x}}_k)$ не представлен в матрице расстояний $D(N, N)$ как центр соответствующего кластера. Поэтому обычно применяют некорректную «наивную» версию алгоритма *k-средних*, где вместо центра кластера $\bar{\omega}_k$ в таком качестве используют объект, ближайший ко всем остальным в кластере.

Тогда в общем случае при выполнении всех условий $\tilde{\omega}_k = \bar{\omega}_k$ может быть получена смещенная кластеризация, т. к. при погружении данного множества в соответствующее признаковое пространство окажется, что центр кластера $\boldsymbol{x}(\bar{\omega}_k)$ может не совпадать со средним объектом $\bar{\boldsymbol{x}}_k$.

Кластеризация относительно центров кластеров

Как известно, среднее арифметическое, используемое в качестве центра кластера, минимизирует его дисперсию и, как результат, дисперсию всей кластеризации [1]. Дисперсия кластера представлена квадратами отклонений объектов от центра кластера, т. е. квадратами расстояний

$$\sigma_k^2 = \frac{1}{N_k} \sum_{i=1}^{N_k} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)^2 = \frac{1}{N_k} \sum_{i=1}^{N_k} d^2(\boldsymbol{x}_i, \bar{\boldsymbol{x}}_k)\,.$$

Данный критерий минимизирует среднее квадратов расстояний до центра кластера и средневзвешенную дисперсию кластеризации в целом

$$J(K) = \frac{1}{N} \sum_{k=1}^{K} N_k \sigma_k^2 = \sum_{k=1}^{K} \frac{N_k}{N} \sigma_k^2\,.$$

В отсутствие признаков средние объекты $\bar{\omega}_k$ обеспечивают несмещенную кластеризацию, также минимизируя дисперсии кластеров

$$\sigma_k^2 = \frac{1}{N_k} \sum_{i=1}^{N_k} d^2(\omega_i, \bar{\omega}_k)$$

и значение критерия $J(K)$ в целом.

Если множество Ω будет помещено в соответствующее пространство признаков, где объекты $\boldsymbol{x}(\bar{\omega}_k)$ и $\bar{\boldsymbol{x}}_k$ совпадут, то два критерия

$$J^X(K) = \min_{\bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}}_K} J(K) \text{ и } J^D(K) = \min_{\bar{\omega}_1, \ldots, \bar{\omega}_K} J(K)$$

окажутся одинаковыми $J^X(K) = J^D(K)$. Очевидно, что в общем случае $J^D(K) \geqslant J^X(K)$.

Построим алгоритм для получения несмещенной кластеризации.

Для некоторого элемента $\omega_l \in \Omega$, взятого как начало координат, и пары ω_i, ω_j их скалярное произведение $c_{ij} = (d_{li}^2 + d_{lj}^2 - d_{ij}^2)/2$ вычисляется на основе расстояний $d_{pq} = d(\omega_p, \omega_q)$, где $c_{ii} = d_{li}^2$ при $i = j$.

Следовательно, элементы главной диагонали матрицы $C_l(N, N)$ представляют собой квадраты расстояний от начала координат $\omega_l \in \Omega$ до остальных объектов. Удобно поместить начало координат в центр тяжести множества $\omega_i \in \Omega$, $i = 1, \ldots, N$ [3].

*Работа выполнена при финансовой поддержке РФФИ, проекты №08-01-12023, 09-01-08151, 09-07-00394, 11-07-00728.

Как показано в [4–6], центр кластера $\bar{\omega}_k$ будет представлен своими расстояниями до всех объектов $\omega_i \in \Omega,\ i = 1, \ldots, N$, без необходимости восстановления неизвестного нам признакового пространства, где N_k — число объектов в кластере Ω_k:

$$d^2(\omega_i, \bar{\omega}_k) = \frac{1}{N_k}\sum_{p=1}^{N_k} d_{ip}^2 - \frac{1}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2.$$

Дисперсия кластера вычисляется [4–6] как

$$\sigma_k^2 = \frac{1}{N_k}\sum_{i=1}^{N_k} d^2(\omega_i, \bar{\omega}_k) = \\ = \frac{1}{N_k}\sum_{i=1}^{N_k}\left(\frac{1}{N_k}\sum_{p=1}^{N_k} d_{ip}^2 - \frac{1}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2\right) = \\ = \frac{1}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2\,. \quad (1)$$

Известно, что алгоритм *k-средних* можно представить в разных вариантах в зависимости от способа пересчета средних. Представим данный алгоритм для расстояний в нужном нам виде, где пересчет центров выполняется сразу после очередного переноса:

Шаг 0. Взять в качестве центров $\bar{\omega}_k^0,\ k = 1, \ldots, K$, например, K наиболее удаленных друг от друга объектов и назначить их представителями $\tilde{\omega}_k^0$, $k = 1, \ldots, K$.

Шаг s. Распределить все объекты по кластерам:

1. Переместить объект ω_i в кластер $\omega_i \in \Omega_k^s$, если для всех остальных кластеров при $\omega_i \in \Omega_j^s$ выполнено условие $d(\omega_i, \bar{\omega}_k^s) \leqslant d(\omega_i, \bar{\omega}_j^s)$, где $j = 1, \ldots, K,\ j \neq k$.
2. Пересчитать, если требуется, центры $\bar{\omega}_k^s$, $k = 1, \ldots, K$ и представить их своими расстояниями до всех объектов $d(\omega_i, \bar{\omega}_k^s),\ i = 1, \ldots, N$.
3. Переместить следующий $i = i + 1$ объект ω_i.
4. Стоп, если ни один объект не был перемещен в другой кластер, т. е. получена несмещенная кластеризация, где $\tilde{\omega}_k^s = \bar{\omega}_k^s$, $k = 1, \ldots, K$, иначе $\tilde{\omega}_k^{s+1} = \bar{\omega}_k^s$ и перейти к следующему шагу $s = s + 1$.

Кластеризация без центров

Вычислим среднее квадратов расстояний между объектами в кластере. С учетом расстояний до себя получим выражение

$$\eta'_k = \frac{1}{N_k^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k}(\boldsymbol{x}_i - \boldsymbol{x}_j)^2 = \frac{1}{N_k^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k} d^2(\boldsymbol{x}_i, \boldsymbol{x}_j)\,.$$

Из выражения (1) для σ_k^2 немедленно следует, что $\eta'_k = 2\sigma_k^2$. Обозначим $\eta_k = \sigma_k^2 = \eta'_k/2$.

Следовательно, для всех кластеров минимизация взвешенных квадратов расстояний между объектами в кластерах приводит к минимизации средневзвешенной дисперсии кластеризации в целом

$$\tilde{J}(K) = \frac{1}{N}\sum_{k=1}^{K} N_k\eta_k = \sum_{k=1}^{K}\frac{N_k}{N}\eta_k\,.$$

Следовательно, критерии $\tilde{J}(K)$ и $J(K)$ совпадают $\tilde{J}(K) = J(K)$. Если множество Ω будет помещено в соответствующее пространство признаков, где объекты $\boldsymbol{x}(\bar{\omega}_k)$ и $\bar{\boldsymbol{x}}_k$ совпадут, то два критерия

$$\tilde{J}^X(K) = \min_{\Omega_1,\ldots,\Omega_K \in X} \tilde{J}(K);$$
$$\tilde{J}^D(K) = \min_{\Omega_1,\ldots,\Omega_K \in D} \tilde{J}(K)$$

также совпадут $\tilde{J}^X(K) = \tilde{J}^D(K)$. Также очевидно, что в общем случае $\tilde{J}^D(K) \geqslant \tilde{J}^X(K)$.

Построим алгоритм кластеризации без центров. Очевидно, что такая кластеризация должна быть несмещенной, если для нее вычислить центры кластеров:

Шаг 0. Взять $\Omega_k^0,\ k = 1, \ldots, K$, например, как K наиболее наиболее компактных в некотором смысле подмножеств.

Шаг s. Распределить все объекты по кластерам:

1. Переместить объект ω_i в кластер $\omega_i \in \Omega_k^s$ и принять $\tilde{J}^s(K) = \tilde{J}_k^s(K)$, если для всех остальных кластеров при $\omega_i \in \Omega_j^s$ выполнено условие $\tilde{J}_k^s(K) < \tilde{J}_j^s(K)$, где $j = 1, \ldots, K,\ j \neq k$.
2. Переместить следующий $i = i + 1$ объект ω_i.
3. Стоп, если ни один объект не был перемещен в другой кластер, т. е. получена несмещенная кластеризация. Иначе перейти к следующему шагу $s = s + 1$.

Кластеризация по близостям

Положительно полуопределенная матрица близостей $S(N, N)$ с элементами $s_{ij} = s(\omega_i, \omega_j) \geqslant 0$ может рассматриваться как матрица скалярных произведений в метрическом пространстве размерности не выше N. Относительно некоторой точки $\omega_k \in \Omega$, взятой как начало координат, где $s_{ij} = (d_{ki}^2 + d_{kj}^2 - d_{ij}^2)/2,\ s_{ii} = d_{ki}^2$, расстояния определяются как $d_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij}$.

Тогда центр кластера $\bar{\omega}_k$ представлен своими близостями к объектам $\omega_i \in \Omega,\ i = 1, \ldots, N$, где N_k — число объектов в кластере Ω_k :

$$s(\omega_i, \bar{\omega}_k) = \frac{1}{N_k}\sum_{p=1}^{N_k} s_{ip}\,.$$

Компактность кластера представлена как средняя близость центра к объектам в кластере

$$\delta_k = \frac{1}{N_k}\sum_{i=1}^{N_k} s(\omega_i, \bar{\omega}_k) = \frac{1}{N_k^2}\sum_{i=1}^{N_k}\sum_{p=1}^{N_k} s_{ip};\ \omega_i, \omega_p \in \Omega_k.$$

Несмещенная кластеризация минимизирует дисперсии σ_k^2 и максимизирует компактности δ_k кластеров

$$\sigma_k^2 = \frac{1}{2N_k^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k} d_{ij}^2 = \\ = \frac{1}{2N_k^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k}(s_{ii}+s_{jj}-2s_{ij}) = \\ = \frac{1}{N_k}\sum_{i=1}^{N_k} s_{ii} - \frac{1}{N_k^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k} s_{ij} = \\ = \frac{1}{N_k}\sum_{i=1}^{N_k} s_{ii} - \delta_k\,,$$

где для всех кластеров

$$J(K) = \sum_{k=1}^{K}\frac{N_k}{N}\sigma_k^2 = \sum_{k=1}^{K}\frac{N_k}{N}\left(\frac{1}{N_k}\sum_{i=1}^{N_k} s_{ii} - \delta_k\right) = \\ = \frac{1}{N}\sum_{i=1}^{N} s_{ii} - \sum_{k=1}^{K}\frac{N_k}{N}\delta_k = c - \sum_{k=1}^{K}\frac{N_k}{N}\delta_k\,.$$

Обозначим средневзвешенную компактность кластеризации как

$$I(K) = \sum_{k=1}^{K}\frac{N_k}{N}\delta_k\,, \quad \text{где } I(K) = c - J(K)\,.$$

Мы немедленно получим две модификации алгоритма *k-средних* для близостей: с вычислением центров кластеров и без них. Построим алгоритм кластеризации с вычислением центров:

Шаг 0. Взять в качестве центров $\bar\omega_k^0$, $k = 1,\ldots,K$, например, K наименее близких друг к другу объектов и назначить их представителями $\tilde\omega_k^0$, $k = 1,\ldots,K$.

Шаг s. Распределить все объекты по кластерам:

1. Переместить объект ω_i в кластер $\omega_i \in \Omega_k^s$, если для всех остальных кластеров при $\omega_i \in \Omega_j^s$ выполнено условие $s(\omega_i,\bar\omega_k^s) \geqslant s(\omega_i,\bar\omega_j^s)$, где $j = 1,\ldots,K,\ j \neq k$.
2. Пересчитать, если требуется, центры $\bar\omega_k^s$, $k = 1,\ldots,K$ и представить их своими близостями ко всем объектам $s(\omega_i,\bar\omega_k^s)$, $i = 1,\ldots,N$.
3. Переместить следующий $i = i+1$ объект ω_i.
4. Стоп, если ни один объект не был перемещен в другой кластер, т. е. получена несмещенная кластеризация, где $\tilde\omega_k^s = \bar\omega_k^s$, $k = 1,\ldots,K$, иначе $\tilde\omega_k^{s+1} = \bar\omega_k^s$ и перейти к следующему шагу $s = s+1$.

Построим алгоритм кластеризации по близостям без центров:

Шаг 0. Взять Ω_k^0, $k = 1,\ldots,K$, например, как K наиболее наиболее компактных в некотором смысле подмножеств.

Шаг s. Распределить все объекты по кластерам:

1. Переместить объект ω_i в кластер $\omega_i \in \Omega_k^s$ и принять $I^s(K) = I_k^s(K)$, если для всех остальных кластеров при $\omega_i \in \Omega_j^s$ выполнено условие $I_k^s(K) > I_j^s(K)$, где $j = 1,\ldots,K$, $j \neq k$.
2. Переместить следующий $i = i+1$ объект ω_i.
3. Стоп, если ни один объект не был перемещен в другой кластер, т. е. получена несмещенная кластеризация. Иначе перейти к следующему шагу $s = s+1$.

Диагонализация как выделение кластеров

Рассмотрим задачу диагонализации [7, 8] произвольной матрицы связей $A(N,N)$ с элементами $a_{ij} \geqslant 0$, которая заключается в получении т. н. диагональной структуры квадратной матрицы связей.

Пусть множество представлено значениями парных связей между его элементами, которые естественным образом концентрируются в K компактных подмножеств, образуя агрегаты.

Чтобы выявить диагональную структуру, нужно одновременно переупорядочить строки и столбцы матрицы связей так, чтобы ее элементы образовали блоки, расположенные на главной диагонали и содержащие значения связи между элементами одного агрегата. Очевидно, что значения связи в диагональных блоках должны быть выше, чем значения в остальных элементах.

Процедура поиска максимизирует внутриагрегатные связи, представленные функционалом

$$F(K) = \sum_{k=1}^{K}\frac{N_k}{N}\left(\frac{1}{N_k(N_k-1)}\sum_{i,j\in\Omega_k, i\neq j} a_{ij}\right).$$

Рассмотрим подробнее критерий $I(K)$:

$$I(K) = \sum_{k=1}^{K}\frac{N_k}{N}\delta_k = \sum_{k=1}^{K}\frac{N_k}{N}\left(\frac{1}{N_k^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k} s_{ij}\right) = \\ = \sum_{k=1}^{K}\frac{N_k}{N}\left(\frac{1}{N_k^2}\sum_{i=1}^{N_k} s_{ii} + \frac{1}{N_k^2}\sum_{i=1}^{N_k}\sum_{\substack{j=1\\ i\neq j}}^{N_k} s_{ij}\right) = \\ = \frac{1}{N}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{i=1}^{N_k} s_{ii} + \sum_{k=1}^{K}\frac{N_k}{N}\left(\frac{1}{N_k^2}\sum_{i=1}^{N_k}\sum_{\substack{j=1\\ i\neq j}}^{N_k} s_{ij}\right).$$

Если матрица близостей $S(N,N)$ приведена к нормализованному виду с диагональными элементами $s_{ii} = 1$, $i = 1,\ldots,N$, например, преобразованием $s_{ij}/\sqrt{s_{ii}s_{jj}}$, то получим

$$I(K) = \frac{K}{N} + \sum_{k=1}^{K}\frac{N_k}{N}\left(\frac{1}{N_k^2}\sum_{i=1}^{N_k}\sum_{\substack{j=1\\ i\neq j}}^{N_k} s_{ij}\right).$$

Следовательно, функционал $F(K)$ является эвристической версией критерия $I(K)$, в которой не учитываются N диагональных элементов.

В результате, для положительно полуопределенной матрицы связей $A(N,N)$ процедура поиска агрегатов, максимизирующая функционал $F(K)$, является эвристической версией алгоритма *k-средних* при кластеризации по близостям без вычисления центров.

Обсуждение

В современном анализе данных часто оказывается, что множество, предъявленное для обработки, представлено лишь результатами парных сравнений его элементов, характеризующими различие или, наоборот, сходство элементов множества.

Например, сходство аминокислотных последовательностей определяется на основе их парного выравнивания специальными алгоритмами [9]. Необходимо отметить, что в задачах анализа таких объектов результаты наблюдений, как правило, не могут быть представлены в традиционной форме матрицы данных, т. е. виде измерений соответствующих признаков [10].

Положительно полуопределенная матрица сходства может быть рассмотрена как матрица скалярных произведений в некотором неизвестном нам метрическом (обычно евклидовом) пространстве, размерность которого не превышает числа элементов множества. Такая матрица преобразуется в матрицу расстояний и наоборот. В результате, соответствующая матрица различий может быть рассмотрена как матрица расстояний в некотором неизвестном пространстве.

Алгоритм *k-средних* очень популярен, т. к. результаты его работы интуитивно понятны и обычно легко объяснимы. Тем не менее, его наивная версия некорректна из-за смещенности кластеризаций.

Для получения несмещенных кластеризаций в отсутствие исходного пространства признаков необходимо иметь корректные версии этого и других алгоритмов кластеризации и распознавания.

Применение алгоритма *k-средних* без вычисления центров кластеров также позволяет найти корректное решение задачи кластеризации, как и в случае с их явным вычислением.

Это позволяет в отсутствие признакового пространства не вводить понятие «среднего» объекта. Часто оказывается, что такой объект невозможно получить в явном виде, что, как правило, вызывает затруднения в его содержательной интерпретации.

Заключение

В данной работе предложены новые версии алгоритма *k-средних* для случая, когда пространство исходных признаков нам неизвестно. Рассмотрены две модификации алгоритма, когда средние по кластерам вычисляются и когда не вычисляются. Данный алгоритм сравнивался с эвристическим алгоритмом диагонализации произвольной матрицы связей.

Показано, что алгоритм диагонализации положительно полуопределенной матрицы связей является эвристической версией алгоритма *k-средних*.

Литература

[1] *Duda R. O., Hart P. E., Stork D. G.* Pattern Classification. — N.Y.: Wiley, 2001. — 654 p.

[2] *Шлезингер М. И.* О самопроизвольном различении образов // Читающие автоматы и распознавание образов. Киев: Наукова думка, 1965. — С. 38–45.

[3] *Torgenson W. S.* Theory and Methods of Scaling.— N.Y.: Wiley, 1958. — 460 p.

[4] *Двоенко С. Д.* Кластеризация элементов множества на основе взаимных расстояний и близостей // ММРО-13, М.: МАКС Пресс, 2007. — С. 114–117.

[5] *Двоенко С. Д.* Кластеризация множества, описанного парными расстояниями и близостями между его элементами // Сибирский журнал индустриальной математики. — 2009. — Т. 12, № 1. — С. 61–73.

[6] *Dvoenko S. D.* Clustering and separating of a set of members in terms of mutual distances and similarities // Transactions on Machine Learning and Data Mining. — 2009. — V. 2, N. 2. — Pp. 80–99.

[7] *Браверман Э. М., Дорофеюк А. А., Лумельский В. Я., Мучник И. Б.* Диагонализация матрицы связей и выявление скрытых факторов // Проблемы расширения возможностей автоматов. М.: ИПУ АН СССР, 1971. — С. 42–79.

[8] *Браверман Э. М., Мучник И. Б.* Структурные методы обработки эмпирических данных. — М.: Наука, 1983. — 464 с.

[9] *Pearson W. R.* Rapid and sensitive sequence comparisons with FASTP and FASTA. // Methods in Enzymology. Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences. N.Y. Acad. Press. — 1990. — N. 183. — Pp. 63–98.

[10] *Mottl V. V., Dvoenko S. D., Seredin O. S., Kulikowski C. A., Muchnik I. B.* Featureless pattern recognition in an imaginary Hilbert space and its application to protein fold classification // 2nd Int. Workshop on Machine Learning and Data Mining in Pattern Recognition, Leipzig: Springer, 2001. — Pp. 322–336.

Прогнозирование связности графа*

Дьяконов А. Г.
djakonov@mail.ru
Москва, Московский государственный университет имени М. В. Ломоносова

Описана проблема прогнозирования связности графа (Link Prediction Problem), которая достаточно популярна на западе, но даже не упоминается в русскоязычной литературе. Её популярность объясняется приложениями в исследованиях стремительно развивающихся социальных сетей и теории анализа графов (graph mining). Сделан обзор работ, посвященных этой проблеме, приведены результаты экспериментов по решению реальной прикладной задачи.

В последнее время в анализе данных появляется большое число задач, которые не укладываются в стандартные постановки задач классификации и регрессии. Они отличаются от «классических», во-первых, способом задания исходных данных: оно, как правило, не признаковое (XML-документы, графы, мультимножества терминов, сигналы и т. д.). Во-вторых, в этих задачах особые требования к ответу (например, необходимо не классифицировать Web-страницы, а упорядочить по степени релевантности). В-третьих, каждая задача имеет свою специфику, связанную с областью приложения (в том числе, это объемы данных, дополнительные ограничения на структуру и т. д.).

К сожалению, на Всероссийских конференциях практически отсутствуют доклады по таким новым задачам: докладываются теоретические результаты в классических предположениях или новая задача не подвергается тщательному анализу (не делается полный обзор литературы, не проводятся экспериментальные исследования достаточно большого числа моделей алгоритмов), а сводится к классической постановке и решается стандартными алгоритмами, «чтобы удовлетворить заказчика».

В данной работе описана одна из таких «новых» задач — прогнозирование связности графа (LPP, в англоязычной литературе принят термин «Link Prediction Problem» [1]).

Задача LPP

В самом общем виде постановка задачи следующая: дан граф $G(t) = \big(V(t), E(t)\big)$ в дискретные моменты времени $t = t_1, \ldots, t_n$ (т. е. n, вообще говоря, различных графов). Необходимо установить, какой он будет в момент времени t_{n+1}, т. е. $G(t_{n+1})$.

Например, вершины графа — представители социальной сети, а рёбра — отношения дружбы между ними (или членство в одной организации, присутствие на одних мероприятиях, наличие общих статей, участие в общих проектах, если речь о научной социальной сети, и т. д.), тогда задача построения рекомендательной системы, которая предлагает Вам «зафрендить» кого-то, сводится к задаче LPP. Другой пример: вершины графа — пользователи сотовой сети и услуги сети, пользователь соединен ребром с услугой, если он ей пользуется. Задача предсказания поведения пользователя (к каким услугам он подключится, а какие посчитает ненужными) также сводится к LPP (заметим, что здесь граф двудольный). Ясно, что аналогичная задача возникает при анализе распространения информации в сети Интернет, распространения инфекций среди населения, в коллаборативной фильтрации, при автоматической генерации гиперссылок и т. д. (см. ссылки в [1], [2]).

Как правило, задачу LPP рассматривают при фиксированном множестве вершин: $V(t) = V$, кроме того, на практике чаще возникают задачи, в которых рёбра могут только появляться: $V(t_1) \subseteq \ldots \subseteq V(t_n)$ и требуется предсказать граф в следующий момент времени: $t_1 < \ldots < t_n < t_{n+1}$. Часто также рассматривают задачу в «невременной постановке»: вместо «настоящего» графа (V, E^*) дан граф $G = (V, E)$, $E \subseteq E^*$, и множество E', $E' \supseteq (E^* \setminus E)$. Необходимо определить, какие элементы из E' являются ребрами нашего графа (т. е. лежат в E^*), а какие нет. Именно для такой постановки будут рассмотрены эксперименты в этой статье.

Есть также похожие задачи (см. [3]), в которых граф надо восстановить по дополнительной информации.

Методы решения LPP

Несмотря на существенно «непризнаковые» данные, задачу LPP почти всегда решают сведением к признаковой постановке, часто даже удается построить отдельные признаки, по которым неплохо ранжируется множество E' (вверху списка оказываются реально существующие рёбра, а внизу — не существующие). Ниже опишем признаки — функции, которые паре $\{x, y\} \in E'$ ставят в соответствие вещественное число (см. также обзоры [1, 4]).

1. Число общих соседей $|\Gamma(x) \cap \Gamma(y)|$, где $\Gamma(x) = \{z \mid \{x, z\} \in E\}$ — множество соседей вершины x. Математически корректнее использовать термин «множество смежных вершин», но мы бу-

Работа выполнена при финансовой поддержке РФФИ (№ 10-07-00609-а) и гранта Президента РФ (МД-757.2011.9).

дем использовать термины теории социальных сетей, где вершина интерпретируется как пользователь. Смысл признака понятен: если x «дружит» с z, а z «дружит» с y, то x, наверное, «дружит» с y, или схематично

$$\{x,z\},\ \{z,y\}\ \Longrightarrow \{x,y\}. \tag{1}$$

Таким образом, каждый общий друг x и y прибавляет шансы «дружбы» x и y. Часто также используют **коэффициент предпочтительности** (Preferential Attachment): $|\Gamma(x)|\cdot|\Gamma(y)|$.

2. Коэффициент Жаккара

$$|\Gamma(x)\cap\Gamma(y)|/|\Gamma(x)\cup\Gamma(y)|$$

часто используется в информационном поиске [5] (из [5] взяты и некоторые термины, используемые в настоящей работе). Здесь происходит нормировка предыдущего признака: число общих друзей учитывается по отношению ко всем друзьям x и y.

3. Коэффициент Адамик/Адар (Adamic/Adar)

$$\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{1}{\log|\Gamma(z)|}$$

учитывает, что не всем общим друзьям надо доверять одинаково. Например, если есть пользователь социальной сети, который «френдит всех подряд», то его как общего друга надо учитывать с небольшим весом.

4. Katz$_\beta$ $=\sum_{l=1}^{\infty}\beta^l\mathrm{paths}_{x,y}(l)$,
где $\mathrm{paths}_{x,y}(l)$ — число путей длины l между вершинами x и y. Этот признак учитывает «друзей друзей» (через которых потенциально тоже можно «подружиться»). Есть также многочисленные его модификации. На практике суммируют не по всем l или пользуются формулой $(I-\beta M)^{-1}-I$ — в xy-м элементе этой матрицы стоит значение признака для ребра $\{x,y\}$, где M — матрица смежности.

4. Анализ достижимости. Чем вероятнее во время случайного блуждания попасть из вершины x в вершину y, тем скорее они соединены ребром (здесь также учитываются «друзей друзей»). При простом случайном блуждании часто используют признак **hitting time** (равен среднему времени прибытия в вершину y), при возвращении на каждом шаге в вершину x с вероятностью α и выборе случайного соседа с вероятностью $(1-\alpha)$ — **PageRank** (равен вероятности попадания в вершину y; хороший обзор по этому признаку см. в [6]).

5. Рекуррентные признаки

$$\mathrm{sim}(x,y)=\frac{\gamma}{|\Gamma(x)|\cdot|\Gamma(y)|}\sum_{a\in\Gamma(x)}\sum_{b\in\Gamma(y)}\mathrm{sim}(a,b),$$

$\mathrm{sim}(x,x)=1$, $\gamma\in(0,1]$.

При вычислении признаков основная трудность — работа с гигантской матрицей смежности графа, поэтому часто используют её аппроксимации, например k-ранговую, с помощью сингулярного разложения (SVD). В [7] предложен быстрый метод анализа достижимости с помощью SVD.

Вероятностные методы. Опишем метод **Generative Model** (необходимые ссылки даны в [4]). Допустим, что наш граф получается в результате следующего процесса: с вероятностью $P(i)$ выбирается i-я вершина, затем выбирается латентный класс z с вероятностью $P(z\,|\,i)$, затем порождается ребро $\{i,j\}$ с вероятностью $P(j\,|\,z)$, т. е. вероятность появления этого ребра равна

$$P(i,j)=\sum_z P(i)P(z\,|\,i)P(j\,|\,z). \tag{2}$$

Оцененная вероятность появления ребра является ответом алгоритма. Все вероятности в (2) оцениваются с помощью EM-алгоритма, максимизируя логарифм правдоподобия

$$\sum_{\{i,j\}\in E}\log P(i,j).$$

Результаты экспериментов

Опишем эксперименты по решению задачи LPP на данных соревнования «IJCNN Social Network Challenge» компании KAGGLE [8]. Из графа социальной сети Flickr в фиксированной момент времени изъяли 4480 ребер, участникам соревнования был предложен получившийся граф и перечень 8960 пар вершин, из которых 4480 являются теми самыми изъятыми рёбрами, а остальные 4480 — не являются рёбрами графа. Необходимо упорядочить список этих 8960 пар (по вероятности того, что пара является ребром). Решения оценивались по функционалу AUC-ROC [9]. Число строк в матрице смежности графа — 1 113 547, число ненулевых элементов в ней — 7 237 983.

Особенность задачи — граф ориентированный, т. е. множество ребер (для ориентированных графов используется термин «дуга») — множество упорядоченных пар. Кроме того, множество вершин графа V разбивается на два подмножества V_1 и V_2, $|V_1|=37\,689$ и $|V_2|=1\,133\,518$, дуги исходят только из вершин множества V_1. Нетрудно видеть, что при такой специфике многие признаки становятся бесполезными для решения задачи, поэтому ниже предложены новые. Качество признаков по умолчанию указано в виде значения функционала AUC-ROC на публичной тестовой выборке (10% от всей контрольной выборки).

Пусть $\Gamma(x,*)=\{z\mid(x,z)\in E\}$, $\Gamma(*,y)=\{z\mid(z,y)\in E\}$. Рассмотрим два подхода для генерации признаков.

Первый подход эксплуатирует идею «друг друга», простейший признак (качество — 84,18%)

здесь

$$\frac{|(\Gamma(x,*)\times\Gamma(*,y))\cap E|}{|(\Gamma(x,*)|\cdot|\Gamma(*,y)|+1}, \qquad (3)$$

т. е. он оценивает, насколько много вершин множества $\Gamma(*,y)$, из которых дуги идут в y, соединено с множеством вершин $\Gamma(x,*)$, в которые дуги идут из x (единица в знаменателе добавляется для избежания деления на ноль). Вместо стандартного пополнения ребрами (1) (которое показало низкие результаты $\approx 74\%$) признак пополняет дугами по схеме

$$(x,y_1),\ (x_1,y_1),\ (x_1,y)\implies(x,y).$$

При использовании последовательностей дуг большей длины (5, 7) качество заметно падало.

Признак (3) очень просто записывается в терминах матрицы смежности, например, в нотации системы MatLab —

$$\begin{gathered}\text{I1 = (M(:, y)==1); I2 = (M(x, :)==1);}\\ \text{nnz(M(I1, I2)) / (sum(I1)*sum(I2) + 1)}\end{gathered} \qquad (4)$$

Все эксперименты делались в этой системе, поскольку следующие модификации признака получаются простым изменением формулы. Улучшение качества до 86% даёт следующее обобщение (3):

$$\frac{\sum\limits_{\substack{a\in\Gamma(*,y)\\ b\in\Gamma(x,*)}}\frac{|\Gamma(a,*)\cap\Gamma(x,*)|\cdot|\Gamma(*,b)\cap\Gamma(*,y)|}{\sqrt{|\Gamma(a,*)|\cdot|\Gamma(*,b)|}}}{|(\Gamma(x,*)|\cdot|\Gamma(*,y)|)+1},$$

здесь каждый единичный элемент подматрицы M(I1, I2) из (4) учитывается с весом.

Признак f_1 первого подхода, который использовался для решения задачи, получался перемножением признака

$$\frac{\sum\limits_{\substack{a\in\Gamma(*,y)\\ b\in\Gamma(x,*)}}\frac{1}{\sqrt{(|\Gamma(a,*)\setminus\Gamma(x,*)|+1)\cdot(|\Gamma(*,b)\setminus\Gamma(*,y)|+1)}}+c_1}{|\Gamma(x,*)|^{c_2}\cdot|\Gamma(*,y)|^{c_3}}$$

и признака

$$1+\frac{c_4}{|\Gamma(x,*)|}\sum_{b\in\Gamma(x,*)}\frac{1}{|\Gamma(*,b)|}+\frac{c_5}{|\Gamma(*,y)|}\sum_{a\in\Gamma(*,y)}\frac{1}{|\Gamma(a,*)|}.$$

Значения констант были получены в результате оптимизации: $c_1 = 1{,}25\cdot10^{-5}$, $c_2 = 0{,}93$, $c_3 = 0{,}98$, $c_4 = 24$, $c_5 = 1{,}97$. Качество признака — 87,51%.

Также был построен признак f_2:

$$\frac{1}{|\Gamma(x,*)|}\sum_{b\in\Gamma(x,*)}\frac{|(\Gamma(*,b)\times\Gamma(x,*))\cap E|}{|\Gamma(*,b)|\cdot|\Gamma(x,*)|+1},$$

который показывал высокое качество в линейной комбинации с другими признаками, например, с первым — 90,7%. Этот признак оценивает, насколько «дружелюбны друзья» x (и не зависит от y).

Второй подход эксплуатирует идею «вершины соединены дугой, если соединены похожие на них». Таким образом, если мы нашли множество X вершин, похожих на x, Y — похожих на y, то простейший признак второго подхода:

$$|(X\times Y)\cap E|/(|X|\cdot|Y|+1).$$

Первый вопрос — как определять схожесть вершин? Вершинам множества X соответствуют строки матрицы смежности, а вершинам множества Y — столбцы. Поэтому надо найти строки, похожие на x-ю строку, и столбцы, похожие на y-й столбец. Похожесть между строками/столбцами измерялась (все формулы приведены для строк) с помощью функций:

скалярное произведение строк (лучшее качество — 87,16% достигается при $|X| = 8$, $|Y| = 80$):

$$|\Gamma(x,*)\cap\Gamma(a,*)|$$

(вершины похожи, если у них много одинаковых соседей).

скалярное произведение «с довеском» (88,36% достигается при $|X| = 9$, $|Y| = 40$):

$$|\Gamma(x,*)\cap\Gamma(a,*)|-\frac{1}{2+|\Gamma(a,*)|-|\Gamma(x,*)\cap\Gamma(a,*)|}$$

устраняет недостатки предыдущей функции сходства, которая для многих пар вершин принимала одинаковые значения, при этом качество повышается и требуется меньшее число столбцов. Заметим, что довесок нетривиален, поскольку противоречит логике, которая лежит в основе построения классических признаков, например коэффициента Жаккара.

евклидова метрика (85,95% при $|X| = 40$, $|Y| = 300$):

$$|\Gamma(x,*)\cap\Gamma(a,*)|/(\sqrt{|\Gamma(a,*)|+1}).$$

Если убрать 1 в знаменателе и ещё разделить на $\sqrt{|\Gamma(x,*)|}$, то максимизация полученной величины эквивалентна минимизации евклидовой метрики между нормированными строками матрицы смежности. Множитель $\sqrt{|\Gamma(x,*)|}$ отсутствует, поскольку вершина x фиксирована (при вычислении признака для конкретной пары (x,y)).

l_1-метрика (81,74% при $|X| = 310$, $|Y| = 900$)

$$|\Gamma(x,*)\cap\Gamma(a,*)|/(|\Gamma(a,*)|+1).$$

В последних двух случаях также использовались довески, но существенного выигрыша они не дали. Признаки, полученные с помощью этих четырех функций, являются попарно некоррелированными. Даже те, которые не обладают высоким

качеством, могут быть полезными, поскольку являются «осторожными»: редко принимают большие значения (этим и объясняется низкое качество — они почти всегда нулевые), но если принимают — соответствующие ребра принадлежат графу. Однако, классификаторы, которые строились по всем этим признакам, часто переобучались.

Отметим, что функция качества от переменных $|X|$ и $|Y|$ оказалась «достаточно гладкой» с ярко выраженным экстремумом, который, однако, зависит от контрольной выборки. Ниже показаны значения функции (при вычислении похожести с помощью скалярного произведения с довеском) на специально сгенерированной контрольной выборке:

$\|Y\| =$	30	35	40	50
$\|X\| = 3$	88,19%	88,23%	88,29%	88,66%
$\|X\| = 5$	88,48%	88,51%	88,54%	88,61%
$\|X\| = 7$	88,69%	88,70%	**88,73%**	88,66%
$\|X\| = 9$	88,64%	88,65%	88,68%	88,61%

Второй вопрос второго подхода — как учитывать «соединенность» похожих вершин, кроме простейшего числа связей $|(X \times Y) \cap E|$? Можно учитывать, что ребро (a, b), которое больше похоже на пару (x, y), должно считаться с большим весом. Вершины из X, т. е. соответствующие строки, упорядочим по убыванию похожести, ниже показано, какие веса им приписываются. Аналогично происходит со столбцами. В итоге используется признак

$$\frac{1}{|X| \cdot |Y| + 1} \sum_{a \in X} \sum_{b \in Y} w(a) \cdot w'(b),$$

где $w(a)$ — вес a-й строки, $w'(b)$ — вес b-го столбца. Были использованы различные способы приписывания весов:

$$\begin{aligned}
1, \ldots, 1 \quad & (88{,}73\%) \\
|X|, \ldots, 2, 1 \quad & (88{,}84\%) \\
\sqrt{|X|}, \ldots, \sqrt{2}, \sqrt{1} \quad & (88{,}85\%) \\
\log(|X|+1), \ldots, \log(3), \log(2) \quad & (88{,}84\%) \\
1, \ldots, 2/|X|, 1/|X| \quad & (88{,}84\%) \\
2^1, \ldots, 2^{2/|X|}, 1^{2/|X|} \quad & (88{,}91\%) \\
20, 2^{(|X|-1)/|X|}, \ldots, 2^{2/|X|}, 1^{2/|X|} \quad & (89{,}07\%)
\end{aligned}$$

В результате оптимизации весовых коэффициентов при фиксированных $|X|$ и $|Y|$ был построен признак f_3 второго подхода с качеством $90,72\%$.

Итоговое решение, которое заняло 7 место на соревновании (среди 119 участников — ученых и научных коллективов), было получено линейной комбинацией признаков f_1, f_2 и f_3, качество — 92,6%, результат, показанный на всей контрольной выборке — 92,46%.

Выводы

В настоящей работе сделана попытка восполнить недостаток русскоязычных ресурсов, связанный с отсутствием информации по актуальной проблеме LPP: сделан обзор «классических» признаков, предложены методы прогнозирования связности ориентированного графа социальной сети. Метод базируется на признаках, которые раскрывают некоторые особенности задачи. Например, при различных (естественных!) формализациях похожести вершин (строк/столбцов матрицы смежности) получаются не только разные по качеству признаки (что вполне ожидаемо), но и разные по требованиям к их вычислению: некоторым достаточно знать 8 похожих вершин, некоторым — более 300. Это позволяет строить «адекватные» меры похожести. К недостаткам работы, конечно, относится отсутствие исследования качества работы различных моделей алгоритмов, как это сделано, например, в [2]. Их можно было тестировать в многомерном пространстве признаков, полученных с помощью описанных подходов. Кроме того, совершенно отсутствует попытка генерации признаков, основанных на блужданиях в графах. Качество таких признаков в описанной задаче — около 93%, а при использовании с построенными выше — возрастает до 95%. В настоящий момент все эти недостатки устраняются.

Литература

[1] *Liben-Nowell D., Kleinberg J.* The link-prediction problem for social networks // Society for Information Science and Technology, 2007. — V. 58, N. 7. — Pp. 1019–1031.

[2] *Hasan M. A, Chaoji V., Salem S., Zaki M.* Link Prediction using Supervised Learning // IAM Workshop on Link Analysis, Counterterrorism and Security with SIAM Data Mining Conference, Bethesda, MD, 2006.

[3] *Leroy V., Cambazoglu B. B., Bonchi F.* Cold Start Link Prediction // Proceedings of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2010.

[4] *Huang Z., Lin D.* The Time Series Link Prediction Problem with Applications in Communication Surveillance // INFORMS Journal on Computing, 2009. — V. 21, N. 2. — Pp. 286–303.

[5] *Маннинг К. Д., Рагхаван П., Шютце Х.* Введение в информационный поиск. — М.: ООО «И. Д. Вильямс», 2011. — 528 с.

[6] *Langville A. N., Meyer C.D.* Deeper Inside PageRank // Journal Internet Mathematics, 2004. — V. 1, N. 3. — Pp. 335—380.

[7] *Tong H., Faloutsos C., Pan J-Y.* Fast Random Walks with Restarts and Its Applications // ICDM, 2006. — Pp. 613–622.

[8] `http://www.kaggle.com` — Сайт соревнований в анализе данных — 2011.

[9] `http://en.wikipedia.org/wiki/ROC_plot` — Статья в Википедии — 2011.

Разработка данных систем совместного пользования ресурсами: от трипонятий к трикластерам*

Игнатов Д. И., Кузнецов С. О., Пульманс Й.
dignatov@hse.ru
г. Москва, НИУ Высшая школа экономики

В работе предложен новый подход к трикластеризации трехмерных бинарных данных. Трикластер определен в терминах триадического анализа формальных понятий (Triadic Formal Concept Analysis) как плотное тримножество тернарного отношения Y между объектами, признаками и условиями. Такое определение является ослаблением определения трипонятия и дает возможность найти все трикластеры и трипонятия, содержащиеся в трикластерах больших наборов данных. Данный подход обобщает аналогичные исследования, проведенные нами для случая бикластеризации, основанной на формальных понятиях.

Термин «бикластер» был введен Б. Г. Миркиным в 1996 [15] и появление трикластеризации и n-кластеризации оставалось только делом времени. Сходный подход, называемый прямой кластеризацией (direct clustering), был предложен в начале 1970-х в работе Дж. Хертигана [10]. В анализе формальных понятий (АФП), предложенном в 1982 г. Р. Вилле [9, 17], используется частный случай бикластера, а именно формальное понятие. Триадический анализ формлаьных понятий (ТАФП) был разработан Леманом и Вилле [14] в 1995 г. как расширение АФП для случая трехмерных бинарных данных. Термины «формальные понятия» и «трипонятия» описывают полезные паттерны в бинарных данных, которые однородны и замкнуты (максимальны) в алгебраическом смысле. По причине жесткой структуры формальных понятий и вычислительной сложности порождающих их алгоритмов (экспоненциальной от размера входа) были предложены различные ослабления определения формального понятия для диадического случая (релевантные и плотные бимножества [3], методы факторизации на понятиях [1], плотные бикластеры [11]) и для триадического случая (триадическая факторизация на понятиях [2]). Существует несколько подходов для поиска только релевантных понятий, например «решетки-айсберги» и индексы устойчивости. Необходимость масштабируемых и эффективных алгоритмов трикластеразции очевидна в связи с возрастающей популярностью и размерами систем совместного пользования ресурсами (social resource tagging systems). Трехмерные данные вида «user-tag-resource», так называемые фолксономии, являются ключевой структурой данных в таких системах. TRIAS — один из хорошо известных алгоритмов для разработки данных фолксономий [12]. Есть также многообещающий подход для разработки данных n-арных отношений [5, 6]; его реализация (DataPeleer) основана на замкнутых множествах и превосходит аналогичные алгоритмы, такие как CubeMiner [13] для разработки данных замкнутых тримножеств. Некоторые исследователи пошли дальше и активно применяют замкнутые тримножества для разработки данных сложных признаковых зависимостей в трехмерных данных, например триадических импликаций [8].

Работа выполнена при финансовой поддержке РФФИ, проект №08-07-92497-НЦНИЛ_а.

Основные определения

Триадический контекст $\mathbb{K} = (G, M, B, Y)$ состоит из множеств G (*объектов*), M (*признаков*), и B (*условий*), а также тернарного отношения $Y \subseteq G \times M \times B$. Запись $(g, m, b) \in Y$ показывает, что объект g обладает признаком m при условии b.

Для удобства записи триадический контекст обозначают (X_1, X_2, X_3, Y). Триадический контекст $\mathbb{K} = (X_1, X_2, X_3, Y)$ порождает следующие диадические контексты: $\mathbb{K}^{(1)} = (X_1, X_2 \times X_3, Y^{(1)})$, $\mathbb{K}^{(2)} = (X_2, X_2 \times X_3, Y^{(2)})$, $\mathbb{K}^{(3)} = (X_3, X_2 \times X_3, Y^{(3)})$, где $gY^{(1)}(m, b) :\Leftrightarrow mY^{(1)}(g, b) :\Leftrightarrow :\Leftrightarrow bY^{(1)}(g, m) :\Leftrightarrow (g, m, b) \in Y$. *Операторы штрих* (операторы, формирующие понятия), порождаемые контекстом $\mathbb{K}^{(i)}$, обозначаются $(.)^{(i)}$. Для каждого порожденного диадического контекста мы имеем два типа таких операторов, т. е. для $\{i, j, k\} = \{1, 2, 3\}$ с $j < k$, $Z \subseteq X_i$ и $W \subseteq X_j \times X_k$ (i)-е операторы штрих определяются следующим образом: $Z \mapsto Z^{(i)} = \{(x_j, x_k) \in X_j \times X_k | x_i, x_j, x_k$ связаны Yдля всех $x_i \in Z\}$, $W \mapsto W^{(i)} = \{x_i \in X_i | x_i, x_j, x_k$ связаны Yдля всех $(x_j, x_k) \in W\}$. Формально *триадическое понятие* триадического контекста $\mathbb{K} = (X_1, X_2, X_3, Y)$ является тройкой вида (A_1, A_2, A_3), где $A_1 \subseteq X_1, A_2 \subseteq X_2, A_3 \subseteq X_3$ и для для любых $\{i, j, k\} = \{1, 2, 3\}$ при условии $j < k$ мы имеем $A_i^{(i)} = (A_j \times A_k)$. Для трипонятия (A_1, A_2, A_3) его компоненты A_1, A_2 и A_3 называются *объемом*, *содержанием* и *модусом* соответственно. Заметим, что интерпратиция $\mathbb{K} = (X_1, X_2, X_3, Y)$ как трехмерной таблицы инциденций согласно нашему определению при подходящей перестановке строк, столбцов и слоев этой таблицы влечет интерпретацию триадического понятия (A_1, A_2, A_3) как максимального трехмерного параллелепипеда, заполненного крестиками. Множество всех триадических понятий контекста $\mathbb{K} =$

$= (X_1, X_2, X_3, Y)$ называется *трирешеткой понятий* и обозначается $\mathfrak{T}(X_1, X_2, X_3, Y)$.

Поиск плотных трикластеров

Операторы штрих и бокс-операторы одноэлементных множеств. Для упрощения обозначений мы используем $(.)'$ для всех операторов штрих, как это обычно и делается в АФП. Для наших целей мы рассмотрим триадический контекст $\mathbb{K} = (G, M, B, Y)$ и введем штрихи и *бокс операторы* для конкретных элементов множеств G, M, B соответственно. В дальнейшем мы будем писать g' вместо $\{g\}'$ для 1-множества $g \in G$ и аналогично для $m \in M$ и $b \in B$: m' and b'.

Операторы, формирующие понятия для 1-множеств:

$$\begin{aligned} m' &= \{\ (g,b)\ |(g,m,b) \in Y\}; \\ g' &= \{\ (m,b)\ |(g,m,b) \in Y\}; \\ b' &= \{\ (g,m)\ |(g,m,b) \in Y\}. \end{aligned}$$

Мы не используем операторы двойного штриха, а только введеные нами бокс-операторы:

$$\begin{aligned} g^{\square} &= \{\ g_i \mid (g_i, b_i) \in m' \text{ or } \ (g_i, m_i) \in b'\}; \\ m^{\square} &= \{\ m_i \mid (m_i, b_i) \in g' \text{ or } \ (g_i, m_i) \in b'\}; \\ b^{\square} &= \{\ b_i \mid (g_i, b_i) \in m' \text{ or } \ (m_i, b_i) \in g'\}. \end{aligned}$$

Пусть $\mathbb{K} = (G, M, B, Y)$ — триадический контекст. Для тройки $(g, m, b) \in Y$ назовем *трикластером* $T = (g^{\square}, m^{\square}, b^{\square})$.

Плотность трикластера (A, B, C) триадического контекста $\mathbb{K} = (G, M, B, Y)$ определяется как доля всех троек Y в трикластере, т. е. $\rho(A, B, C) = |I \cap A \times B \times C|/|A||B||C|$.

Трикластер $T = (A, B, C)$ называется *плотным*, если его плотность больше некоторого минимального порога, т. е. $\rho(T) \geqslant \rho_{\min}$. Для данного триконтекста $\mathbb{K} = (G, M, B, Y)$ мы обозначим $\mathbf{T}(G, M, B, Y)$ множество всех его (плотных) трикластеров.

Утверждение 1. *Для любого трипонятия (A, B, C) триадического контекста $\mathbb{K} = (G, M, B, Y)$ и непустых множеств A, B и C мы имеем $\rho(A, B, C) = 1$.*

Утверждение 2. *Для любого трикластера (A, B, C) триадического контекста $\mathbb{K} = (G, M, B, Y)$ и непустных множеств A, B и C мы имеем $0 \leqslant \rho(A, B, C) \leqslant 1$.*

Теорема 1. *Пусть $\mathbb{K} = (G, M, B, Y)$ будет триадическим контекстом и $\rho_{\min} = 0$. Для любого $T_c = (A_c, B_c, C_c) \in \mathfrak{T}(G, M, B, Y)$ существует трикластер $T = (A, B, C) \in \mathbf{T}(G, M, B, Y)$ такой, что $A_c \subseteq A, B_c \subseteq B, C_c \subseteq C$.*

Таблица 1. Фрагмент данных Bibsonomy

	t_1	t_2	t_3
u_1		×	×
u_2	×	×	×
u_3	×	×	×

r_1

	t_1	t_2	t_3
u_1	×	×	×
u_2	×		×
u_3	×	×	×

r_2

	t_1	t_2	t_3
u_1	×	×	×
u_2	×	×	×
u_3	×	×	

r_3

Пример 1. Для примера табл. 67 получим $3^3 = 27$ формальных понятий, 24 с $\rho = 1$ и 3 трипонятия с $\rho = 0$ (они имеют либо пустое множество пользователей, либо ресурсов или тегов). Хотя данные небольшие, мы получили 27 паттернов для анализа (максимальное количество формальных понятий для контекста размера $3 \times 3 \times 3$) — результат того, что мы имеем дело со степенным триадическим контекстом. Однако мы можем заключить, что пользователи u_1, u_2 и u_3 использовали почти одинаковые множества ресурсов и их пометки. Таким образом, они очень сходны в терминах (tag, resource) общих пар и целесообразно уменьшить число паттернов, описывающих данные, с 27 до 1. Трикластер $T = (\{u_1, u_2, u_3\}, \{t_1, t_2, t_3\}, \{r_1, r_2, r_3\})$ с $\rho = 0{,}89$ есть в точности искомый паттерн, а его плотность несколько меньше 1. Каждое из трипонятий триконтекста в $\mathfrak{T} = \{(\varnothing, \{t_1, t_2, t_3\}, \{r_1, r_2, r_3\}), (\{u_1\}, \{t_2, t_3\}, \{r_1, r_2, r_3\}), \ \ldots, \ (\{u_1, u_2, u_3\}, \{t_1, t_2\}, \{r_3\})\}$ содержится (в смысле покомпонентного вложения множеств) в T.

Мы предложили эвристику для вычисления $\rho(T)$, основанную на проверке только небольшого количества случайно выбранных троек, содержащихся в заданном трикластере T. Для трикластера $T = (A, B, C)$ мы провели оценку плотности $\hat{\rho}(T) = |P|/|N|$, где $P = \{(g, m, b) | (g, m, b) \in N \cap Y\}$, N — множество размера $|N|$ случайно выбранных элементов трикластера. Параметр $|N|$ может быть выбран относительно небольшим, скажем $0{,}1|A||B||C|$.

Реальные данные и эксперименты

В наших экспериментах мы анализировали свободно доступные данные популярной системы социальных закладок Bibsonomy [4]. Мы запускали алгоритм TRICL на части данных, состоящих из всех пользователей, ресурсов и присвоенных тегов для выявления сообщества пользователей, имеющих сходное поведение при тегировании.

Результирующая фолксономия (бибсономия) состит из $|U| = 2\ 337$ пользователей, $|T| = 67\ 464$ тегов и $|R| = 28\ 920$ ресурсов (закладок или bibtex описаний), которые связаны $|Y| = 816\ 197$ тройками. Отметим, что мы имеем дело с параллелепипедом, состоящим из 4 559 624 602 560 ячеек.

Алгоритм 1. Алгоритм поиска трикластеров TRICL

Вход: $K = (G, M, B, Y)$ – триконтекст;
$\rho_{\min}$ – порог плотности;
Выход: $\mathbf{T} = \{(A_k, B_k, C_k) | (A_k, B_k, C_k)$ – плотный трикластер$\}$.

1: **для всех** $(g, m, b) \in Y$
2: **если** g not in PrimesObj **то**
3: PrimesObj$[g] = g'$;
4: **если** m not in PrimesAttr **то**
5: PrimesAttr$[m] = m'$;
6: **если** b not in PrimesCond **то**
7: PrimesCond$[b] = b'$;
8: **если** g not in BoxesObj **то**
9: BoxesObj$[g] = g^{\square}$;
10: **если** m not in BoxesAttr **то**
11: BoxesAttr$[m] = m^{\square}$;
12: **если** b not in BoxesCond **то**
13: BoxesCond$[b] = b^{\square}$;
14: **для всех** $(g, m, b) \in Y$
15: $T = (\text{BoxesObj}[g], \text{BoxesAttr}[m], \text{BoxesCond}[b])$
16: Tkey $= \text{hash}(T)$
17: **если** Tkey not in **T то**
18: **если** $\rho(T) \geqslant \rho_{\min}$ **то**
19: **T**[Tkey] $= (T)$;

Мы исследовали статистическое распределение перед применением алгоритма TRICL. Мы вычислили и построили гистограммы для пользователей и числа пар (tag, document) (рис. 1), аналогичные гистограммы для тегов и числа пар (user, document) и для документов и их пар (user, tag). Мы обнаружили, что данные следуют степенному закону распределения $p(x) = Cx^{-\alpha}$ с $\alpha = 3,6778$ и дисперсией $\sigma = 0,0001$ в случае документов и количества пар (user, tag). Для пользователй и тегов мы получили $\alpha = 2{,}13$ и $\alpha = 1{,}8$ соответственно. Мы вычислили α, используя оценку максимального правдоподобия, как описано в [16], и проверили результаты с помощью программного обеспечения, описанного в [7].

Это наблюдение позволит нам использовать жадную стратегию поиска, если мы хотим искать большие и относительно плотные трикластеры, по той причине, что только небольшая часть пользователей совершает присваивание (tag, user) (аналогичные выводы для распределения тегов документов).

Мы измерили производительность нашей реализации (Python 2.7.1) на системе Pentium Core Duo с тактовой частотой процессора 2 ГГц и 2 Гб ОЗУ. Мы использовали реализацию алгоритма TRIAS на Java из работы [12] для построения всех трипонятий заданного контекста. Результаты экспериментов представлены в табл. 2. Два последних

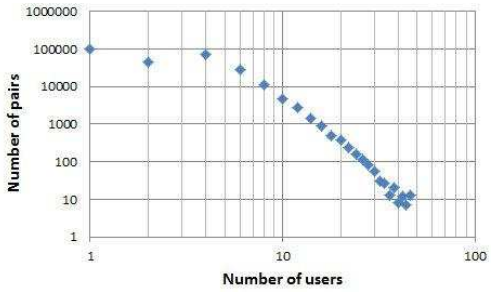


Рис. 1. Гистограмма количества пар (document, tag) для 800 000 записей Bibsonomy

Таблица 2. Экспериментальные результаты для k первых троек набора данных tas с $\rho_{\min} = 0$

k	$\lvert U\rvert$	$\lvert T\rvert$	$\lvert R\rvert$	$\lvert\mathfrak{T}\rvert$	$\lvert\mathbf{T}\rvert$
100	1	47	52	57	1
1000	1	248	482	368	1
10000	1	444	5193	733	1
100000	59	5823	28920	22804	4462
200000	340	14982	61568	—	19053

Trias, с	TriclEx, с	TriclProb, с
0,2	0,2	0,2
1	1	1
2	46,7	47
3386	10311	976
> 24 ч	> 24 ч	3417

Таблица 3. Распределение плотности трикластеров для 200 000 первых троек набора данных tas с $\rho_{\min} = 0$

Нижняя граница ρ	Верхняя граница ρ	Число трикластеров
0	0,05	18617
0,05	0,1	195
0,1	0,2	112
0,2	0,3	40
0,3	0,4	20
0,4	0,5	10
0,5	0,6	8
0,6	0,7	1
0,7	0,8	1
0,8	0,9	0
0,9	1	49

столбца показывают среднее время работы Tricl с полной (TriclEx) и вероятностной (TriclProb) стратегией вычислений.

В наших экспериментах оценка $\hat{\rho}$ имела абсолютную ошибку 0б13 для параметров $|N| = 1/10$, $\rho_{\min} = 0$ и 200 000 троек данных Бибсономии. Алгоритм работает значительно быстрее, чем Trias и TriclExб в случае нашей вероятностной стратегии.

В табл. 3 представлено распределение плотности трикластеров для 200 000 первых троек набора данных bibsonomy дано.

Заключение

Мы предложили АФП-подход к трикластеризации и показали, что:

— (плотная) трикластеризация — неплохая альтернатива для ТАФП, т. к. общее число трикластеров для некоторого набора данных значительно меньше числа формальных понятий;
— (плотная) трикластеризация способна справиться с большим количеством трипонятий в худшем случае триконтекстов (или плотных кубоидов в них), когда их главная диагональ пуста, и рассматривает такие кубоиды как целые трикластеры. Это очень релевантное свойство для исследования трисообществ в системах социальных закладок;
— предложенный алгоритм имеет хорошую масштабируемость на реальных данных, особенно когда используется подход жадного покрытия и оптимизированная версия процедуры вычисления плотности.

Мы продолжаем нашу работу над трикластеризацией в следующих направлениях:

— исследование различных, основанных на ограничениях подходах к трикластеризации (например, анализ данных плотных трикластеров и затем частых множеств тримножеств в них);
— поиск лучших стратегий оценки плотности трикластеров;
— разработка обобщенной теоретической формализации трикластеризации на основе замкнутых множеств;
— учет природы реальных данных с целью оптимизации алгоритмов (разреженность данных, распределение значений и т. д.).

Литература

[1] *Belohlavek, R., Vychodil, V.* Factor analysis of incidence data via novel decomposition of matrices // In: Ferre, S., Rudolph, S. (eds.) ICFCA. Lecture Notes in Computer Science, vol. 5548, pp. 83–97. Springer (2009).

[2] *Belohlavek, R., Vychodil, V.* Factorizing three-way binary data with triadic formal concepts // In: Setchi, R., Jordanov, I., Howlett, R., Jain, L. (eds.) Knowledge-Based and Intelligent Information and Engineering Systems, Lecture Notes in Computer Science, vol. 6276, pp. 471–480. Springer Berlin / Heidelberg (2010).

[3] *Besson, J., Robardet, C., Boulicaut, J.F.* Mining a new fault-tolerant pattern type as an alternative to formal concept discovery // In: Scharfe, H., Hitzler, P., Ohrstrom, P. (eds.) Conceptual Structures: Inspiration and Application, Lecture Notes in Computer Science, vol. 4068, pp. 144–157. Springer Berlin / Heidelberg (2006).

[4] `bibsonomy.org` — Сервис библиографических закладок.

[5] *Cerf, L., Besson, J., Robardet, C., Boulicaut, J.F.* Data peeler: Contraint-based closed pattern mining in n-ary relations // In: SDM. pp. 37–48. SIAM (2008).

[6] *Cerf, L., Besson, J., Robardet, C., Boulicaut, J.F.* Closed patterns meet -ary relations TKDD 3(1) (2009).

[7] *Clauset, A., Shalizi, C.R., Newman, M.E.J.* Power-law distributions in empirical data // SIAM Review 51(4), 661–703 (2009).

[8] *Ganter, B., Obiedkov, S.* Implications in triadic formal contexts // In: Wolff, K., Pfeiffer, H., Delugach, H. (eds.) Conceptual Structures at Work, Lecture Notes in Computer Science, vol. 3127, pp. 237–237. Springer Berlin / Heidelberg (2004).

[9] *Ganter, B., Wille, R.* Formal concept analysis: Mathematical foundations // Springer, Berlin-Heidelberg (1999).

[10] *Hartigan, J.A.* Direct clustering of a data matrix. Journal of the American Statistical Association // 67(337), 123–129 (March 1972).

[11] *Ignatov, D.I., Kaminskaya, A.Y., Kuznetsov, S.O., Magizov, R.A.* A concept-based biclustering algorithm // In: Proceedings of the Eighth International conference on Intelligent Information Processing (IIP-8), pp. 140–143. MAKS Press (2010), in Russian.

[12] *Jäschke, R., Hotho, A., Schmitz, C., Ganter, B., Stumme, G.* Trias — an algorithm for mining iceberg tri-lattices // In: ICDM, pp. 907–911. IEEE Computer Society (2006).

[13] *Ji, L., Tan, K.L., Tung, A.K.H.* Mining frequent closed cubes in 3D datasets // In: Dayal, U., Whang, K.Y., Lomet, D.B., Alonso, G., Lohman, G.M., Kersten, M.L., Cha, S.K., Kim, Y.K. (eds.) VLDB, pp. 811–822. ACM (2006).

[14] *Lehmann, F., Wille, R.* A triadic approach to formal concept analysis // In: Ellis, G., Levinson, R., Rich, W., Sowa, J. (eds.) Conceptual Structures: Applications, Implementation and Theory, Lecture Notes in Computer Science, vol. 954, pp. 32–43. Springer Berlin / Heidelberg (1995).

[15] *Mirkin, B.* Mathematical Classification and Clustering. — Kluwer, 1996.

[16] *Newman, M.E.J.* Power laws, pareto distributions and Zipf's law // Contemporary physics 46(5), 323–351 (2005).

[17] *Wille, R.* Restructuring lattice theory: An approach based on hierarchies of concepts // In: Rival, I. (ed.) Ordered Sets, pp. 445–470. Boston (1982).

Методы улучшения сходимости EM-алгоритма в вероятностном латентном семантическом анализе*

Лексин В. А.
vleksin@gmail.com
Московский физико-технический институт (государственный университет)

Получена оценка скорости сходимости EM-алгоритма в вероятностном латентном семантическом анализе. Предложены способы улучшения скорости сходимости и задания начального приближения параметров.

В основе вероятностного латентного семантического анализа (Probabilistic Latent Semantic Analysis, PLSA) [1] лежит EM-алгоритм [2], который позволяет оценить скрытые предпочтения клиентов (покупателей, пользователей Интернет, держателей пластиковых карт, абонентов мобильной связи и т. д.) по отношению к некоторому набору объектов (товаров, сайтов, услуг, документов и т. д.) по наблюдаемому протоколу транзакций (действий клиентов), без анализа содержимого (контента) объектов. EM-алгоритм формирует компактные *тематические профили* (векторы вероятностей тем) для всех клиентов и объектов, встречающихся в транзакционных данных. Он позволяет обрабатывать огромные объёмы транзакционных данных, что крайне актуально для приложений. В данной работе исследуются факторы, влияющие на сходимость EM-алгоритма, представлена оценка скорости сходимости и условие сходимости за конечное число итераций, предложен алгоритм генерации начального приближения профилей. Эксперименты проводились как на модельных данных, так и на реальных данных о поведении клиентов Интернет-магазина.

Оценка сходимости EM-алгоритма

Введём основные обозначения:

$\mathcal{U}$ — множество номеров клиентов, $|\mathcal{U}| = U$;

$\mathcal{R}$ — множество номеров объектов, $|\mathcal{R}| = R$;

$\mathcal{Y}$ — пространство описаний транзакций (единичных взаимодействий клиента и объекта);

$\mathcal{D} = (u_i, r_i, y_i)_{i=1}^N \subseteq \mathcal{U}\times\mathcal{R}\times\mathcal{Y}$ — протокол транзакций.

$F = \|f_{ur}\|$ — матрица кросс-табуляции, формируемая по протоколу $\mathcal{D}$, $f_{ur} = \operatorname{aggr}\{(u, r, y_i) \in \mathcal{D}\}$, где aggr — некоторая операция агрегирования, например, $f_{ur} = \sum_{(u,r,y_i)\in\mathcal{D}} y_i$.

$\mathcal{T}$ — множество тем, которыми могут интересоваться клиенты $|\mathcal{T}| = T$; обычно предполагается, что $T \ll R$ и $T \ll U$.

Вводится вероятностная модель данных: вероятность $p(u, r)$ того, что клиент $u \in \mathcal{U}$ выберет ресурс $r \in \mathcal{R}$, согласно формуле полной вероятности,

$$p(u,r) = \sum_{t\in\mathcal{T}} p_u p_{tu} q(r \mid t) = p_u q_r \sum_{t\in\mathcal{T}} \frac{q_{tr} p_{tu}}{p_t}, \qquad (1)$$

где $p_{tu} = p(t \mid u)$ — вероятность интереса клиента u к теме t; $q_{tr} = q(t \mid r)$ — вероятность того, что объект r относится к теме t; $p_u = p(u)$ — априорная вероятность клиента u; $q_r = q(r)$ — априорная вероятность объекта r; $p_t = p(t)$ — априорная вероятность темы t; $q(r \mid t) = q_{tr} q_r / p_t$ — апостериорная вероятность объекта r. Здесь неизвестными являются векторы $\mathbf{p}_u = (p_{tu} : t \in \mathcal{T})$ — *профиль клиента* u, и $\mathbf{q}_r = (q_{tr} : t \in \mathcal{T})$ — *профиль объекта* r.

Задача вероятностного латентного семантического анализа — оценить по матрице F профили клиентов и объектов. Априорные вероятности клиентов и объектов оцениваются по протоколу следующим образом: $p_u = S(u)/S$; $q_r = S(r)/S$, где $S(u) = \sum_{r\in\mathcal{R}} f_{ur}$, $S(r) = \sum_{u\in\mathcal{U}} f_{ur}$, $S = \sum_{(u,r)\in\mathcal{U}\times\mathcal{R}} f_{ur}$.

Для нахождения неизвестных параметров модели p_{tu}, q_{tr} и p_t воспользуемся принципом максимума взвешенного правдоподобия:

$$\begin{cases} \mathscr{L} = \sum\limits_{(u,r)\in\mathcal{U}\times\mathcal{R}} f_{ur} \ln p(u,r) \to \max\limits_{\{p_{tu}, q_{tr}, p_t\}}; \\ \sum\limits_{t\in\mathcal{T}} p_{tu} = 1,\ u \in \mathcal{U}; \quad \sum\limits_{t\in\mathcal{T}} q_{tr} = 1,\ r \in \mathcal{R}; \\ \sum\limits_{u\in\mathcal{U}} p_{tu} p_u = \sum\limits_{r\in\mathcal{R}} q_{tr} q_r = p_t, \quad t \in \mathcal{T}; \\ p_{tu} \geqslant 0,\ q_{tr} \geqslant 0, \quad u \in \mathcal{U},\ r \in \mathcal{R},\ t \in \mathcal{T}. \end{cases}$$

Для решения данной задачи используется EM-алгоритм. Вводятся скрытые переменные H — апостериорные вероятности того, что клиент u, выбирая ресурс r, интересовался темой t (E-шаг):

$$H(t \mid u, r) = \frac{p_{tu} q_{tr} / p_t}{\sum\limits_{\tau\in\mathcal{T}} p_{\tau u} q_{\tau r} / p_\tau}. \qquad (2)$$

Профили и вероятности тем выражаются через скрытые переменные следующим образом (M-шаг):

$$\begin{aligned} p_{tu} &= \tfrac{1}{S(u)} \sum_{r\in\mathcal{R}} f_{ur} H(t \mid u, r); \\ q_{tr} &= \tfrac{1}{S(r)} \sum_{u\in\mathcal{U}} f_{ur} H(t \mid u, r); \\ p_t &= \tfrac{1}{S} \sum_{(u,r)\in\mathcal{U}\times\mathcal{R}} f_{ur} H(t \mid u, r). \end{aligned} \qquad (3)$$

Работа поддержана РФФИ (проект №11-07-00480) и программой ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

Теорема 1. *Пусть $\mathbf{p}'_u$ и $\mathbf{q}'_r$ — профили после выполнения M-шага. На каждой итерации справедливы равенства:*

$$\begin{aligned}\mathbf{p}'_u - \mathbf{p}_u &= P_u \frac{\partial \mathscr{L}}{\partial \mathbf{p}_u},\ u \in \mathcal{U},\\ \mathbf{q}'_r - \mathbf{q}_r &= Q_r \frac{\partial \mathscr{L}}{\partial \mathbf{q}_r},\ r \in \mathcal{R},\end{aligned} \tag{4}$$

$$P_u = \tfrac{1}{S(u)}\big(\operatorname{diag}(p_{1u},\ldots,p_{Tu}) - \mathbf{p}_u\mathbf{p}_u^{\intercal}\big);$$
$$Q_r = \tfrac{1}{S(r)}\big(\operatorname{diag}(q_{1r},\ldots,q_{Tr}) - \mathbf{q}_r\mathbf{q}_r^{\intercal}\big).$$

Теорема 2. *Матрицы P_u и Q_r положительно полуопределены для всех $u \in \mathcal{U}$ и $r \in \mathcal{R}$.*

Обозначим вектор всех параметров модели на некоторой k-ой итерации EM-алгоритма

$$\theta = [\mathbf{p}_1^{\intercal},\ldots,\mathbf{p}_U^{\intercal},\mathbf{q}_1^{\intercal},\ldots,\mathbf{q}_R^{\intercal}]^{\intercal}$$

и рассмотрим блочно-диагональную матрицу

$$P = \operatorname{diag}\{P_1,\ldots,P_U,Q_1,\ldots,Q_R\}.$$

Тогда выражения (4) примут вид

$$\theta' - \theta = P\frac{\partial \mathscr{L}}{\partial \theta},$$

где θ' — значения вектора параметров на следующем шаге EM-алгоритма.

Из утверждения теоремы 2 следует, что матрица P положительно полуопределена, что имеет следующую геометрическую интерпретацию: разность векторов $\theta' - \theta$ на каждом шаге EM-алгоритма имеет положительно полуопределенную проекцию на градиент правдоподобия. Это показывает тесную связь EM-алгоритма с градиентным методом, когда на каждом шаге движение идет в направлении градиента с некоторым выбранным шагом, для которого доказана сходимость к глобальному максимуму правдоподобия.

В следующей теореме утверждается, что EM-алгоритм имеет линейную скорость сходимости.

Теорема 3. *Пусть θ^* — локальный максимум $\mathscr{L}(\theta)$, $\theta \to \theta^*$ при $k \to \infty$, где k — номер итерации; в некоторой окрестности θ^* Гессиан $H(\theta) = \frac{\partial^2 \mathscr{L}(\theta)}{\partial\theta\partial\theta^{\intercal}}$ существует и отрицательно определен. Тогда, начиная с некоторой итерации, верны оценки*

$$\|\theta' - \theta^*\| \leqslant r_c\|\theta - \theta^*\|,$$

где $r_c = \|E^{\intercal}\big(I + P(\theta^*)H(\theta^*)\big)\| \leqslant \sqrt{1 + \lambda_M^2 - 2\lambda_m}$, E — произвольный ортонормированный базис линейного подпространства

$$\Theta = \Big\{\theta \;\Big|\; \sum_{t\in\mathcal{T}} p_{tu} = 0,\ u \in \mathcal{U};\ \sum_{t\in\mathcal{T}} q_{tr} = 0,\ r \in \mathcal{R};\ \sum_{u\in\mathcal{U}} p_u p_{tu} = \sum_{r\in\mathcal{R}} q_r q_{tr},\ t \in \mathcal{T}\Big\};$$

λ_M и λ_m ($\lambda_M > \lambda_m > 0$) — минимальное и максимальное собственные значения положительно полуопределенной матрицы $-E^{\intercal}P(\theta^*)H(\theta^*)E$.

Ограничения подпространства Θ — это ограничения нормировки компонент профилей со сдвигом на константный вектор, не зависящий от θ.

Легко показать, что при $\lambda_M < 2$ процесс сходится ($r_c < 1$), а при $\lambda_m = \lambda_M = 1$ он сходится за одну итерацию ($r_c = 0$). Также из теоремы 3 следует, что скорость сходимости критически зависит от числа обусловленности матрицы PH. Если матрица обусловлена плохо, то сходимость не гарантируется.

Следующая теорема утверждает, что если все векторы $\big(H(t\,|\,u,r)\colon t \in \mathcal{T}\big)$, для которых $f_{ur} \neq 0$, содержат строго по одному ненулевому элементу, то алгоритм сходится за конечное число итераций.

Теорема 4. *Пусть $H(t\,|\,u,r) \in \{0,1\}$ для всех $t \in \mathcal{T}$, $u \in \mathcal{U}$ и $r \in \mathcal{R}$ в точке θ^*. Тогда $r_c = 0$, начиная с некоторой итерации.*

Условие $H(t\,|\,u,r) \in \{0,1\}$ имеет вполне понятную содержательную интерпретацию: если событие (u,r) выбора объекта r клиентом u встречается несколько раз в протоколе, то оно всегда должно быть обусловлено интересом клиента u к одной и той же теме $t \in \mathcal{T}$.

Эксперименты и выводы

Алгоритм тестировался на модельных данных и на реальных данных поведения пользователей Интернет-магазина.

Генерация модельных данных. Сначала задаются вероятности $p(u\,|\,t)$, $q(r\,|\,t)$ путем расстановки случайным образом заданного числа m максимумов в каждом из векторов $\big(p(u\,|\,t)\colon t \in \mathcal{T}\big)$ и $\big(q(r\,|\,t)\colon t \in \mathcal{T}\big)$. Затем по этим вероятностям генерируется протокол $\mathcal{D}$ с помощью вероятностной модели (1), где вероятности тем p_t задаются случайным образом. В данном исследовании параметры генерации модельных данных были выбраны следующим образом: $U = 200$; $R = 50$; $T = 8$; $N = 10000$; $m = 3$.

Для оценивания качества восстановления профили на выходе EM-алгоритма сравниваются с исходными, вычисленными по модельному протоколу $\mathcal{D}$. Исходные профили вычисляются непосредственно по формулам M-шага (1), поскольку известно, с какой темой t связана каждая запись (u,r) модельного протокола. Для каждой непустой ячейки $f_{ur} \neq 0$ матрицы кросс-табуляции F скрытые переменные оцениваются как $H(t\,|\,u,r) = N(t,u,r)/f_{ur}$, где $N(t,u,r)$ — количество раз, когда появление пары (u,r) в протоколе было обусловлено темой t. Затем профили клиентов и объектов формируются по формулам (3) через известные значения скрытых переменных.

Расстояние $\rho(\theta_{ini}, \theta)$ между результирующим вектором параметров θ и исходным θ_{ini} определяется строится как среднее расстояние между соответствующими профилями, которое, в свою очередь, определяется как среднее абсолютное отклонение соответствующих компонент профилей. Перед вычислением расстояния $\rho(\theta_{ini}, \theta)$ предварительно определяется оптимальная перестановка тем, одинаковая для всех результирующих профилей, при которой расстояние $\rho(\theta_{ini}, \theta)$ минимально.

Для анализа влияния точности начального приближения профилей на скорость и качество сходимости EM-алгоритма в качестве начального приближения задавались исходные профили, но к каждой компоненте прибавлялось случайное число в диапазоне $[0, \sigma]$ с последующей нормировкой, где σ — заданная амплитуда зашумления начального приближения. На рис. 1 отображены кривые сходимости (k — число итераций) при случайном начальном приближении параметров, и при различных амплитудах зашумления, в том числе нулевой. Эксперимент показывают, что чем ближе начальные приближения профилей к истинным, тем быстрее и надежнее сходится EM-алгоритм.

Для того, чтобы выполнялось условие теоремы 4 $H(t \mid u, r) \in \{0, 1\}$ для всех $u \in \mathcal{U}$, $r \in \mathcal{R}$ и $t \in \mathcal{T}$ при моделировании данных число существенно отличных от нуля элементов выбирается $m = 1$. Тогда в исходных профилях $\mathbf{p}_u^{ini}$ и $\mathbf{q}_r^{ini}$, вычисляемых по формулам $p_{tu}^{ini} = p(u \mid t)p_t/p_u$, $q_{tr}^{ini} = q(r \mid t)p_t/q_r$, будет содержаться единица в одной из компонент и нули во всех остальных. Легко видеть, что в этом случае скрытые переменные $H(t \mid u, r) \in \{0, 1\}$. На рис. 2 показаны кривые сходимости алгоритма при различных значениях m. При $m = 1$ алгоритм сходится за 4 шага. При увеличении m скорость сходимости существенно замедляется.

Протокол реальных данных содержит историю просмотров станиц Интернет-магазина клиентами, причем для тех клиентов, которые попали на на сайт магазина через поисковую машину, известен введенный ими поисковый запрос. В качестве клиентов были выбраны различные запросы поисковой машины, а в качестве объектов — url-адреса страниц Интернет-магазина. Для анализа были отобраны 7675 запросов и 9233 url-страниц (из анализа были исключены слишком редкие и слишком популярные запросы и страницы). Обработанный протокол транзакций содержал 141767 пар запрос – url-страница.

Для оценивания качества работы EM-алгоритма на реальных данных использовалась априорная классификация url-страниц по категориям товаров, описание которых на них размещено. Для 10 ключевых категорий было подобрано по одному наиболее подходящему для них запросу. Критерий качества определялся как средняя доля объектов «чужой» категории среди ближайших 20 объектов к каждому из запросов. Расстояние от запросов до объектов определялось как среднее абсолютное отклонение соответствующих компонент профилей.

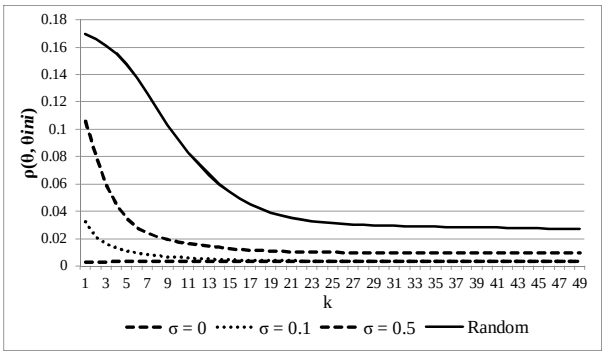


Рис. 1. Амплитуда шума в начальном приближении.

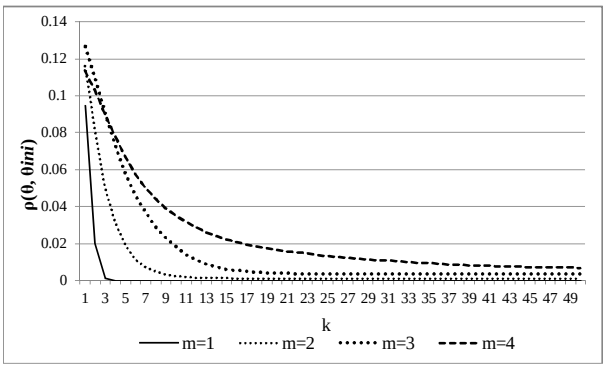


Рис. 2. Число максимумов в модельных данных.

На реальных данных исходные профили неизвестны, поэтому для задания начального приближения разработан следующий алгоритм, позволяющий грубо оценить профили по протоколу транзакций. Сначала выбираются $R_c > T$ объектов, которые назовем множеством кандидатов на базис и обозначим его $\mathcal{R}_c$. Если заранее известна экспертная разбивка $R_e > T$ объектов на темы, то заносим их в $\mathcal{R}_c$. Если нет никакой априорной классификации объектов, то в $\mathcal{R}_c$ выбираем случайным образом $3T$ объектов из $\mathcal{R}$, предварительно отфильтровав сильно популярные и редко используемые по заданным порогам. Далее вычисляется матрица попарных расстояний между объектами в $\mathcal{R}_c$. Функция расстояния определяется по формуле

$$d(r, r') = \begin{cases} \frac{1}{2}\big(1 - \frac{s}{s+d_1+d_2}\big), & d_1 d_2 = 0; \\ \frac{1}{2}\big(\frac{d_1}{s+d_1} + \frac{d_2}{s+d_2}\big), & d_1 d_2 \neq 0; \end{cases}$$

где $s = \sum\limits_{u \in \mathcal{U}} \min(f_{ur}, f_{ur'})$, $d_1 = \sum\limits_{u \in \mathcal{U}} f_{ur}[f_{ur'} = 0]$, $d_2 = \sum\limits_{u \in \mathcal{U}} f_{ur'}[f_{ur} = 0]$, если у объектов r и r' есть хотя бы один клиент, выбравший оба объекта, т. е. $s \neq 0$. Если таких клиентов нет, то расстояние опре-

деляется по формуле $d_0(r,r') = \sum\limits_{\rho \in A(r,r')} \frac{d(r,\rho)d(\rho,r')}{|A(r,r')|}$, где $A(r,r')$ — множество таких объектов, что среди клиентов, которые ими воспользовались, есть хотя бы по одному, кто также пользовался объектами r и r'. По матрице попарных расстояний выбираются T самых удаленных объектов из $\mathcal{R}_c$, назовем их базисными. Для этого сначала находятся два самых удаленных объекта, затем по очереди выбираются остальные $T-2$, сумма расстояний от которых до уже выбранных максимальна. В компоненты профилей объектов заносится значение близости $q_{tr} = 1 - d(r, e_t)$, где e_t — t-ый базисный объект, с последующей нормировкой. В компоненты профилей клиентов p_{tu} заносятся средние значения близости $1 - d(u,u')$ до тех клиентов u', которые посещали соответствующий базисный объект, с последующей нормировкой. Если известны тематики базисных объектов, то можно дать содержательную интерпретацию соответствующих компонент профилей.

На рис. 3 и 4 показаны кривые сходимости EM-алгоритма на модельных и на реальных данных в двух случаях: начальное приближение генерируется случайным образом и по описанному алгоритму. Использование описанного алгоритма для задания начального приближения существенно улучшает скорость сходимости и точность результата по сравнению со случайным.

Можно искусственно обнулять малые значения $H(t \mid u,r)$, p_{tu} и q_{tr} по заданному порогу для того, чтобы сократить объем хранимых в памяти данных, храня только максимальные элементы. На рис. 5 показаны результаты сходимости в трех случаях: без обнуления, с обнулением $H(t \mid u,r)$ и обнулением компонент профилей на модельных данных. Из рисунка видно, что обнуление $H(t \mid u,r)$ дает то же результирующее качество восстановления профилей, что и без него, но за меньшее число итераций. Обнуление компонент профилей не дает улучшения в скорости сходимости, однако улучшает качество восстановления профилей.

Выводы

На сходимость EM-алгоритма в PLSA влияют точность выбора начального приближения профилей и разреженность (среднее число существенно отличных от нуля компонент) векторов скрытых переменных $\big(H(t \mid u,r) \colon t \in \mathcal{T}\big)$, $u \in \mathcal{U}$, $r \in \mathcal{R}$.

Для задания начального приближения профилей предложен алгоритм, позволяющий ускорить сходимость, а также учесть априорную информацию об объектах.

Оценка скорости сходимости и результаты экспериментов позволяют выдвинуть гипотезу: чем менее выраженные максимумы содержат векторы скрытых переменных, тем медленнее сходится алгоритм. На основе этой гипотезы предложен алгоритм обнуления малых компонент скрытых переменных $H(t \mid u,r)$ и профилей p_{tu} и q_{tr}. Это позволяет увеличить скорость сходимости, уменьшить объем хранимых в памяти данных и сократить среднюю ошибку восстановления профилей объектов и клиентов.

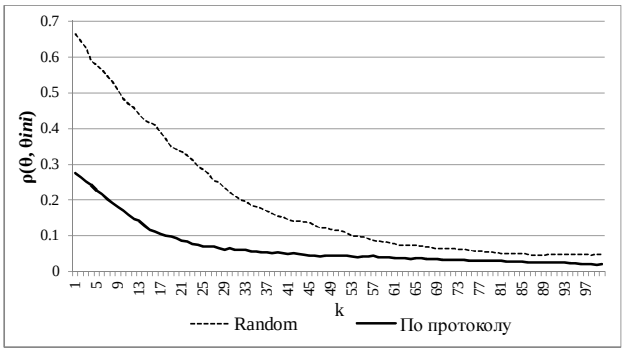


Рис. 3. Начальное приближение на модельных данных.

Рис. 4. Начальное приближение на реальных данных.

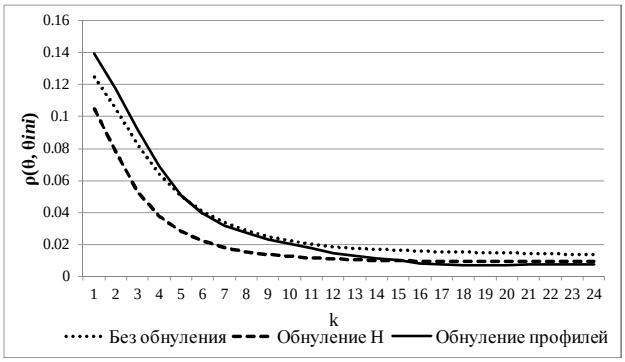


Рис. 5. Обнуление $H(t \mid u,r)$ и профилей.

Литература

[1] *Leksin V. A.* Symmetrization and overfitting in probabilistic latent semantic analysis // *Pattern Recognition and Image Analysis.* — 2009. — Vol. Volume 19, Number 4 / Декабрь 2009 г. — Pp. 565–574.

[2] *Ma J., Xu L., Jordan M. I.* Asymptotic convergence rate of the em algorithm for gaussian mixtures // *Neural Comput.* — December 2000. — Vol. 12. — Pp. 2881–2907.

Инкрементные методы коллаборативной фильтрации для больших разреженных порядковых данных*

Полежаева Е. А.
lena_polejaeva@mail.ru
Московский государственный университет им. М. В. Ломоносова, факультет ВМК

Методы коллаборативной фильтрации, применяемые в современных рекомендующих системах, должны эффективно обрабатывать сверхбольшие динамически пополняемые массивы данных. Дополнительная сложность для использования стандартных методов матричных разложений возникает в тех задачах, где исходные данные представляют собой рейтинги, то есть задаются в порядковой шкале. Стандартные методы применимы исключительно к данным в количественных шкалах. В данной работе предлагается инкрементный градиентный метод для построения матричных разложений для больших массивов порядковых данных. Экспериментально исследуются вопросы скорости сходимости, выбора оптимальной размерности представления, зависимости точности представления данных от объема данных. Сравнивается точность методов, учитывающих и не учитывающих порядковую природу данных.

Методы коллаборативной фильтрации (CF) используются в рекомендующих системах и системах управления взаимоотношениями с клиентами (CRM) для автоматического формирования персональных предложений.

Исходными данными является матрица Y, строки которой соответствуют n клиентам, столбцы — d объектам. В зависимости от приложения объектами могут быть товары, услуги, текстовые документы, и т. п. Каждая заполненная ячейка матрицы содержит информацию об использовании данным клиентом данного объекта. Это может быть отметка о посещении, выставленный клиентом рейтинг, заплаченная им сумма, и т. д. Задача состоит в том, чтобы для произвольного клиента спрогнозировать оценки предпочтительности объектов по всем незаполненным ячейкам в строке матрицы Y. Обычно для этого выделяется множество клиентов со схожими предпочтениями (коллаборация).

В практических приложениях к методам CF предъявляется следующая совокупность требований: возможность эффективного пересчёта решения при добавлении строк, столбцов и ячеек в Y (инкрементность); возможность построения модели, когда матрица Y содержит порядковые данные (рейтинги); возможность формирования компактных описаний (профилей) клиентов и объектов и эффективного вычисления оценок сходства по ним; учёт сильной разреженности матрицы Y; линейный рост вычислительной сложности при увеличении числа заполненных ячеек.

В методах матричных разложений матрица Y аппроксимируется матрицей меньшего ранга, представимой в виде произведения двух матриц $\hat{Y} = UR$. Каждая строка матрицы $U_{n\times L}$ — профиль клиента, каждый столбец $R_{L\times d}$ — профиль объекта, L — число признаков. Матрицы, допускающие такое разложение, имеют ранг не более, чем L.

Работа выполнена при финансовой поддержке РФФИ, проект № 10-07-00609-а.

Известны инкрементные методы CF, основанные на сингулярном разложении разреженных матриц, в которых в матрицу Y эффективно добавляются объекты и ячейки [1, 2] или только клиенты [3]. Целью данной работы является обобщение этих методов на случай, когда данные порядковые, разреженные, имеют большой размер, и допускаются произвольные модификации матрицы Y.

Основные обозначения

$Y_{n\times d} = (y_{ij})_{i=1}^{n}{}_{j=1}^{d}$ — разреженная матрица исходных данных;

$\hat{Y}_{n\times d} = (\hat{y}_{ij})_{i=1}^{n}{}_{j=1}^{d}$ — матрица небольшого ранга, аппросимирующая матрицу Y;

L — число признаков;

$U_{n\times L} = (u_{il})_{i=1}^{n}{}_{l=1}^{L}$ — матрица, отвечающая профилям (сжатым описаниям) клиентов;

$R_{L\times d} = (r_{lj})_{l=1}^{L}{}_{j=1}^{d}$ — матрица, отвечающая профилям объектов;

$\Omega \subseteq \{1,\ldots,n\} \times \{1,\ldots,d\}$ — множество индексов непустых элементов Y.

Обобщенный алгоритм Хебба

Для получения сингулярного разложения разреженной матрицы Y, содержащей данные, измеренные в количественной шкале, в [4] предложен обобщенный алгоритм Хебба GHA (Generalized Hebbian Algorithm). Для получения матриц U, R решается задача минимизации среднеквадратичной ошибки [3]:

$$\text{MSE} = \frac{1}{|\Omega|}\sum_{(i,j)\in\Omega}(y_{ij} - \hat{y}_{ij})^2 \to \min_{U,R},$$

где $\hat{y}_{ij}$ — оценка, полученная алгоритмом:

$$\hat{y}_{ij} = \sum_{l=1}^{L} u_{il} r_{lj} = u_i r_j, \qquad (1)$$

u_i — i-я строка U, r_j — i-й столбец R. В случае, когда $\Omega = \{1,\ldots,n\} \times \{1,\ldots,d\}$, имеем классическую задачу сингулярного разложения матрицы Y, не содержащей пропущенных значений.

В случае порядковых данных предлагается перейти к более сильной количественной шкале. Пусть, без ограничения общности, элементы матрицы Y принимают значения из множества $M = \{1, \dots, \bar{m}\}$. Поставим каждому $m \in M$ в соответствие число $\beta_m \in \mathbb{R}$, потребовав $\beta_1 \leqslant \dots \leqslant \beta_{\bar{m}}$.

Для получения матриц U, R будем решать ту же задачу минимизации среднеквадратичной ошибки, но определять ещё и совокупность значений $\beta = (\beta_1, \dots, \beta_{\bar{m}})$:

$$\text{MSE} = \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} (\beta_{y_{ij}} - u_i r_j)^2 \to \min_{U,R,\beta}.$$

Чтобы исключить тривиальное решение $\beta = 0$, $U = 0$, $R = 0$, введём дополнительные ограничения нормировки:

$$\frac{1}{n}\sum_{i=1}^{n} \|u_i\|^2 = 1; \quad \frac{1}{d}\sum_{j=1}^{d} \|r_j\|^2 = 1.$$

Оптимальное значение β_m легко находится аналитически:

$$\beta_m = \frac{\sum\limits_{(i,j)\in\Omega} [y_{ij} = m] u_i r_j}{\sum\limits_{(i,j)\in\Omega} [y_{ij} = m]}.$$

Введем двойственные переменные λ_1 и λ_2, соответствующие ограничениям нормировки, и запишем оптимизационную задачу для Лагранжиана:

$$\sum_{(i,j)\in\Omega} (\beta_{y_{ij}} - u_i r_j)^2 + \lambda_1 \sum_{i=1}^{n} (\|u_i\|^2 - 1) + \\ + \lambda_2 \sum_{j=1}^{d} (\|r_j\|^2 - 1) \to \max_{\lambda_1,\lambda_2} \min_{u,r,\beta}.$$

Для оптимизации λ_1 и λ_2 используем перебор по сетке значений. Для оптимизации матриц U, R используем метод стохастического градиентного спуска. На каждом шаге итерационного процесса будем выбирать случайную пару $(i, j) \in \Omega$ и выполнять градиентный шаг для всех $l = 1, \dots, L$:

$$u'_{il} = (1 - \eta\lambda_1) u_{il} + \eta \sum_{j\colon (i,j)\in\Omega} r_{lj}(\beta_{y_{ij}} - u_i r_j),$$

$$r'_{lj} = (1 - \eta\lambda_2) r_{lj} + \eta \sum_{i\colon (i,j)\in\Omega} u_{il}(\beta_{y_{ij}} - u_i r_j),$$

где η — темп обучения, u'_{il}, r'_{lj} — новые значения переменных u_{il}, r_{lj}.

Для вычисления среднеквадратичной ошибки в исходной шкале значений M введём функцию, которая отображает получаемые оценки $\hat{y}_{ij} = u_i r_j$ обратно во множество M:

$$R(\hat{y}_{ij}) = \begin{cases} 1, & \hat{y}_{ij} \leqslant \frac{\beta_1+\beta_2}{2}, \\ m, & \frac{\beta_{m-1}+\beta_m}{2} < \hat{y}_{ij} \leqslant \frac{\beta_m+\beta_{m+1}}{2}, \\ \bar{m}, & \frac{\beta_{\bar{m}-1}+\beta_{\bar{m}}}{2} \leqslant \hat{y}_{ij}. \end{cases} \qquad (2)$$

Алгоритм 1. Метод стохастического градиентного спуска для порядковых данных.

Вход: $Y = \{y_{ij}\} \in M^{n\times d}$, $(i,j) \in \Omega$; λ_1, λ_2, η;
Выход: U, R;

1: инициализировать $\beta_m = m$ для всех $m \in M$; инициализировать U, R небольшими значениями из интервала $[0, 0.7]$;
2: $\bar{y}_m = \sum\limits_{(i,j)\in\Omega} [y_{ij} = m]$;
3: **для** $t = 1, \dots, T$
4: **для всех** $(i,j) \in \Omega$ в случайном порядке
5: **для всех** $l = 1, \dots, L$
6: $u_{il} := (1 - \eta\lambda_1) u_{il} + \eta \sum\limits_{j} r_{lj}(\beta_{y_{ij}} - u_i r_j)$;
7: $r_{lj} := (1 - \eta\lambda_2) r_{lj} + \eta \sum\limits_{i} u_{il}(\beta_{y_{ij}} - u_i r_j)$;
8: $\beta_m := (\bar{y}_m)^{-1} \sum\limits_{(i,j)\in\Omega} [y_{ij} = m] u_i r_j$;
9: вычислить RMSE.
10: **если** RMSE не уменьшился на последних двух итерациях, **то выход**;
11: добавить новые элементы (i,j).

Качество разложения будем оценивать величиной RMSE, равной квадратному корню из среднеквадратичной ошибки:

$$\text{RMSE}^2 = \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} (y_{ij} - R(\hat{y}_{ij}))^2.$$

Преимущество градиентного алгоритма в том, что он позволяет добавлять новые значения в матрицу Y непосредственно во время итераций, что предусмотрено в Алгоритме 1 на шаге 11.

Описание эксперимента

Алгоритм 1 тестировался на реальных данных MovieLens (943 клиента, 1682 объекта, 10^5 заполненных ячеек). В ходе итераций производились модификации матрицы Y: добавлялись строки (клиенты) и столбцы (объекты), добавлялись и заменялись значения ячеек. Эксперименты показали, что при увеличении объема данных среднеквадратичная ошибка уменьшается.

При увеличении числа клиентов (строк Y) растет скорость сходимости алгоритма (от 1000 итераций при 600 клиентах до 40 на 940-м клиенте).

Наименьшее значение RMSE = 0.92 достигается при $L = 16$ признаках, при этом одновременно относительный размах значений β_m оказывается максимальным, $\frac{\beta_5-\beta_1}{\beta_1} = 13.8$, см. рис. 1. Остальные параметры: $\lambda_1 = \lambda_2 = 0.18$, $\eta = 0.25$.

Ошибка Алгоритма 1 (RMSE = 0.92 при 5 итерациях), немного лучше, чем у алгоритма из [3] (RMSE = 0.93 при 7 итерациях), не учитывающего, что данные порядковые, см. рис. 2.

Результаты тестирования Алгоритма 1 в инкрементном режиме показаны на рис. 3. Поведение

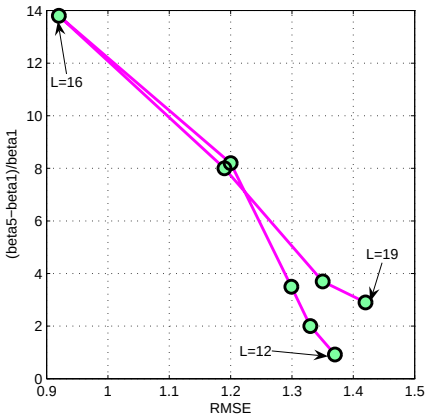


Рис. 1. Зависимость RMSE и относительного размаха оценок $\frac{\beta_5-\beta_1}{\beta_1}$ от числа признаков L.

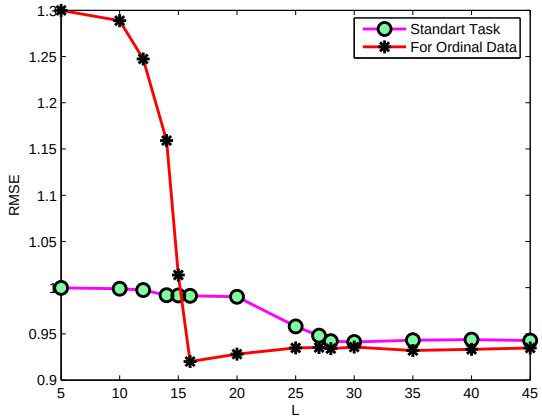


Рис. 2. Зависимость RMSE от числа признаков L для Алгоритма 1 и алгоритма из [3], не учитывающего, что данные порядковые.

RMSE сильно зависит от типа модификации и итерации, во время которой происходит модификация.

При выбранном критерии остановки (RMSE не уменьшается на последних двух итерациях) при добавлении отдельных элементов (i, j) ошибка возрастает немного, так как в выборке уже есть i-й клиент и j-й объект, и для них профили u_i и r_j уже настроены на предыдущих итерациях.

При добавлении клиентов (строк) RMSE возрастает сильнее, а при добавлении объектов (столбцов) — ещё сильнее. Тем не менее, алгоритм сходится, даже если модификации происходят на каждой итерации. Поскольку постоянно происходит добавление новых клиентов и объектов с малым числом заполненных ячеек, доля заполненных ячеек во всей матрице постоянно уменьшается (убывающая пунктирная кривая на рис. 3).

Измерения времени выполнения показывают, что предложенный метод хорошо масштабируется и может применяться к большим объёмам данных.

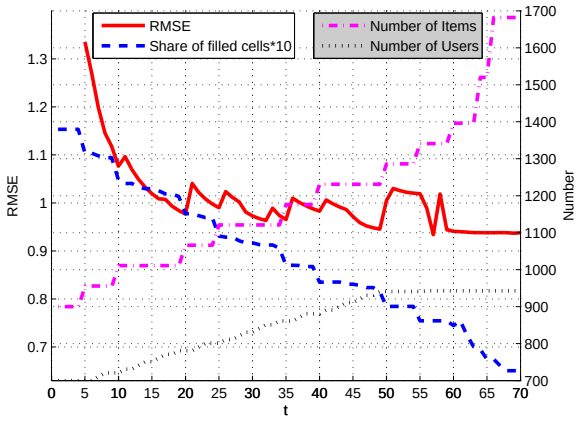


Рис. 3. Зависимость RMSE, доли заполненных ячеек, числа объектов и клиентов от номера итерации t в режиме инкрементных модификаций матрицы исходных данных.

Выводы

Градиентные методы вычисления матричных разложений могут применяться для решения практических задач коллаборативной фильтрации, в которых исходные данные имеют большой размер, разрежены, динамически пополняются и модифицируются, измеряются в количественной или порядковой шкале. Показано, что учёт шкалы измерения позволяет немного повысить точность разложения.

Литература

[1] *Brand M.* Fast online SVD revisions for lightweight recommender systems // In SIAM International Conference on Data Mining. — 2003. — Pp. 37–46.

[2] *Brand M.* Fast Low-rank modifications of the thin singular value decomposition // Linear Algebra and Its Applications. — 2006. — Vol. 415, No. 1. — Pp. 20–30.

[3] *Takacs G., Pilaszy I., Nemeth B., Tikk D.* Scalable Collaborative Filtering Approaches for Large Recommender Systems // The Journal of Machine Learning Research. — 2009. — Vol. 10, — Pp. 623–656.

[4] *Gorrell G.* Generalized Hebbian Algorithm for Incremental Singular Value Decomposition in Natural Language Processing // Proceedings of Interspeech. — 2006.

NP-полнота некоторых задач кластеризации*

Кельманов А. В.
kelm@math.nsc.ru
Новосибирск, Институт математики им. С. Л. Соболева СО РАН, Новосибирский государственный университет

Доказана NP-полнота нескольких актуальных задач кластеризации конечного множества векторов евклидова пространства.

Объект исследования настоящей работы — проблемы оптимизации в задачах анализа данных и распознавания образов. Предмет исследования — некоторые актуальные задачи кластеризации конечного множества векторов евклидова пространства. Цель работы — анализ алгоритмической сложности этих задач.

Проблемы кластерного анализа исследуются более полувека. Из недавно опубликованного обзора [1] видно, что принципы, критерии, модели, задачи, методы и алгоритмы кластеризации рассматривались в тысячах публикаций. При этом скорость разработки алгоритмов (как правило, эвристических и не имеющих теоретических гарантий по точности) для решения разнообразных прикладных задач значительно опередила скорость изучения алгоритмической сложности редуцированных оптимизационных задач, к которым сводятся прикладные проблемы. Статус сложности многих задач кластерного анализа до настоящего времени остается невыясненным, хотя интуитивно и гипотетически они считаются NP-трудными. Между тем, выяснение сложностного статуса позволяет решить вопросы о существовании, как точного полиномиального алгоритма решения редуцированной экстремальной задачи, так и эффективного алгоритма, гарантирующего оптимальность решения соответствующей прикладной проблемы. Поэтому изучение алгоритмической сложности задач кластерного анализа и их систематизация является важным направлением исследований.

В настоящей работе анализируется алгоритмическая сложность нескольких задач кластеризации. Мотивацией исследований послужил тот факт, что статус сложности этих задач ранее не был установлен. В постановочном плане они близки к хорошо известной [2–8] задаче MSSC (Minimum-Sum-of-Squares Clustering) — кластеризации множества векторов евклидова пространства по критерию минимума суммы квадратов расстояний от элементов кластеров до их центров. По своей постановке они близки также к тем задачам кластерного анализа, NP-полнота которых была установлена в [9–12].

Задача MSSC и её близкие аналоги

Задача MSSC в форме верификации свойств имеет следующую формулировку.

Задача MSSC.

Дано: множество $\mathcal{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ векторов из $\mathbb{R}^q$, натуральное число $J > 1$ и положительное число A.

Вопрос: существует ли разбиение множества $\mathcal{Y}$ на непустые подмножества (кластеры) $\mathcal{C}_1, \ldots, \mathcal{C}_J$ такое, что имеет место неравенство

$$\sum_{j=1}^{J} \sum_{\boldsymbol{y} \in \mathcal{C}_j} \|\boldsymbol{y} - \overline{\boldsymbol{y}}(\mathcal{C}_j)\|^2 \leqslant A\,, \qquad (1)$$

где $\overline{\boldsymbol{y}}(\mathcal{C}_j) = (1/|\mathcal{C}_j|) \sum_{\boldsymbol{y} \in \mathcal{C}_j} \boldsymbol{y}$, $j = 1, \ldots, J$, — центр j-го кластера?

В некоторых публикациях (см., например, [1, 3, 7]) эта же задача фигурирует под названием k-Means. Напомним, что одномерный вариант этой задачи разрешим за полиномиальное время [4]. Четыре возможных случая задачи (обусловленных тем, что размерность пространства и число кластеров могут являться и не являться частью входа задачи) проанализированы в [5–8]. Среди них полиномиально разрешимым является случай, когда размерность пространства и число кластеров не являются частью входа [5]. Оставшиеся три случая являются NP-полными задачами [6–8].

К числу слабо изученных в алгоритмическом плане относятся задачи кластеризации по критерию минимума суммы квадратов расстояний, сформулированные ниже.

Задача MSSC-Case.

Дано: множество $\mathcal{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ векторов из $\mathbb{R}^q$, натуральное число $M > 1$ и положительное число A.

Вопрос: существует ли разбиение множества $\mathcal{Y}$ на $J = N - M + 1$ непустых кластеров $\mathcal{C}_1, \ldots, \mathcal{C}_J$ такое, что мощность одного из этих кластеров равна M и справедливо неравенство (1)?

Задача MSSC-N.

Дано: множество $\mathcal{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ векторов из $\mathbb{R}^q$, натуральное число J и положительное число A.

Работа выполнена при финансовой поддержке РФФИ, проекты №09-01-00032, №10-07-00195; целевой программы №2 Президиума РАН, проект №227; целевой программы СО РАН, интеграционный проект №44; целевой программы АВЦП Рособразования, проект №2.1.1/3235, а также федеральной целевой программы «Научные и научно-педагогические кадры инновационной России», гос. контракт №14.740.11.0362.

Вопрос: существует ли разбиение множества $\mathcal{Y}$ на непустые кластеры $\mathcal{C}_1, \ldots, \mathcal{C}_J$ и $\mathcal{B} = \mathcal{Y} \backslash (\cup_j \mathcal{C}_j)$ такое, что справедливо неравенство

$$\sum_{j=1}^{J} \sum_{\boldsymbol{y} \in \mathcal{C}_j} \|\boldsymbol{y} - \overline{\boldsymbol{y}}(\mathcal{C}_j)\|^2 + \sum_{\boldsymbol{y} \in \mathcal{B}} \|\boldsymbol{y}\|^2 \leqslant A\,, \qquad (2)$$

где $\overline{\boldsymbol{y}}(\mathcal{C}_j) = (1/|\mathcal{C}_j|) \sum_{\boldsymbol{y} \in \mathcal{C}_j} \boldsymbol{y}$, $j = 1, \ldots, J$, — центр j-го кластера?

Задача MSSC-F.

Дано: множество $\mathcal{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ векторов из $\mathbb{R}^q$, натуральные числа $M_1, \ldots, M_J$ и положительное число A.

Вопрос: существует ли разбиение множества $\mathcal{Y}$ на непустые кластеры $\mathcal{C}_1, \ldots, \mathcal{C}_J$ и $\mathcal{B} = \mathcal{Y} \backslash (\cup_j \mathcal{C}_j)$ такое, что имеет место неравенство (2), при ограничениях $|\mathcal{C}_j| = M_j$, $j = 1, \ldots, J$, на мощности кластеров?

Символы F и N в названиях задач образованы от английских слов Fixed и Nonfixed для обозначения двух возможных вариантов одной общей проблемы. Эти варианты соответствуют наличию (Fixed) или отсутствию (Nonfixed) ограничений на мощности искомых кластеров.

В задачах MSSC-N и MSSC-F, в отличие от задачи MSSC, центр одного из кластеров определять не требуется. В приложениях, связанных с анализом таблиц данных, этот кластер ассоциируется с шумом, помехой или «мусором», который может содержаться в таблице. NP-полнота задач MSSC-Case, MSSC-N и MSSC-F была установлена совсем недавно в [9–12].

Актуальные задачи кластеризации

Все задачи, сформулированные ниже, индуцируются одной оптимизационной моделью содержательной проблемы анализа данных, которая, судя по множеству публикаций естественно-научного и технического плана, характерна и актуальна для широкого спектра приложений. В этих задачах требуется, используя критерий минимума суммы квадратов расстояний, разбить совокупность векторов на три семейства кластеров так, что:

1) первое семейство состоит из кластеров, каждый из которых содержит «близкие» между собой векторы (также, как и в задаче MSSC);
2) второе семейство состоит из кластеров, каждый из которых содержит векторы, «похожие» на заданный вектор;
3) третье семейство состоит из единственного кластера, центр которого определять не требуется (также, как и в задачах MSSC-N и MSSC-F).

Задача MSSC-NN.

Дано: множество $\mathcal{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ и алфавит $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K\}$ векторов из $\mathbb{R}^q$, натуральное число J и положительное число A.

Вопрос: существует ли разбиение множества $\mathcal{Y}$ на непустые кластеры $\mathcal{C}_1, \ldots, \mathcal{C}_J$, $\mathcal{B}_1, \ldots, \mathcal{B}_K$ и $\mathcal{B} = \mathcal{Y} \backslash ((\cup_j \mathcal{C}_j) \cup (\cup_k \mathcal{B}_k))$ такое, что имеет место неравенство

$$\sum_{j=1}^{J} \sum_{\boldsymbol{y} \in \mathcal{C}_j} \|\boldsymbol{y} - \overline{\boldsymbol{y}}(\mathcal{C}_j)\|^2 + \sum_{k=1}^{K} \sum_{\boldsymbol{y} \in \mathcal{B}_k} \|\boldsymbol{y} - v_k\|^2 + \sum_{\boldsymbol{y} \in \mathcal{B}} \|\boldsymbol{y}\|^2 \leqslant A\,, \qquad (3)$$

где $\overline{\boldsymbol{y}}(\mathcal{C}_j) = (1/|\mathcal{C}_j|) \sum_{\boldsymbol{y} \in \mathcal{C}_j} \boldsymbol{y}$, $j = 1, \ldots, J$, — центр j-го кластера?

Задача MSSC-FN.

Дано: множество $\mathcal{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ и алфавит $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K\}$ векторов из $\mathbb{R}^q$, натуральные числа $M_1, \ldots, M_J$ и положительное число A.

Вопрос: существует ли разбиение множества $\mathcal{Y}$ на непустые кластеры $\mathcal{C}_1, \ldots, \mathcal{C}_J$, $\mathcal{B}_1, \ldots, \mathcal{B}_K$ и $\mathcal{B} = \mathcal{Y} \backslash ((\cup_j \mathcal{C}_j) \cup (\cup_k \mathcal{B}_k))$ такое, что справедливо неравенство (3), при ограничениях $|\mathcal{C}_j| = M_j$, $j = 1, \ldots, J$, на мощности кластеров?

Задача MSSC-NF.

Дано: множество $\mathcal{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ и алфавит $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K\}$ векторов из $\mathbb{R}^q$, натуральные числа J, $N_1, \ldots, N_K$ и положительное число A.

Вопрос: существует ли разбиение множества $\mathcal{Y}$ на непустые кластеры $\mathcal{C}_1, \ldots, \mathcal{C}_J$, $\mathcal{B}_1, \ldots, \mathcal{B}_K$ и $\mathcal{B} = \mathcal{Y} \backslash ((\cup_j \mathcal{C}_j) \cup (\cup_k \mathcal{B}_k))$ такое, что справедливо неравенство (3), при ограничениях $|\mathcal{B}_k| = N_k$, $k = 1, \ldots, K$, на мощности кластеров?

Задача MSSC-FF.

Дано: множество $\mathcal{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ и алфавит $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K\}$ векторов из $\mathbb{R}^q$, натуральные числа $M_1, \ldots, M_J$, $N_1, \ldots, N_K$ и положительное число A.

Вопрос: существует ли разбиение множества $\mathcal{Y}$ на непустые кластеры $\mathcal{C}_1, \ldots, \mathcal{C}_J$, $\mathcal{B}_1, \ldots, \mathcal{B}_K$ и $\mathcal{B} = \mathcal{Y} \backslash ((\cup_j \mathcal{C}_j) \cup (\cup_k \mathcal{B}_k))$ такое, что справедливо неравенство (3), при ограничениях $|\mathcal{C}_j| = M_j$, $j = 1, \ldots, J$, и $|\mathcal{B}_k| = N_k$, $k = 1, \ldots, K$, на мощности кластеров?

Комбинации последних двух символов — FF, NF, FN и NN — в кратких названиях сформулированных выше задач обозначают четыре возможных варианта одной общей проблемы. Они обусловлены наличием или отсутствием ограничений на мощности элементов из пары семейств искомых кластеров.

Основным результатом настоящей работы является доказательство NP-полноты сформулированных выше задач.

Анализ сложности

Для доказательства факта труднорешаемости задач напомним следующие NP-полные [13, 14] задачи выбора подмножеств. Символы SVS в кратких названиях этих задач образованы от английского словосочетания Searching Vector Subsets,

а комбинации двух последних символов, как и ранее, обозначают возможные варианты проблемы, обусловленные ограничениями на мощности искомых подмножеств.

Задача SVS-NN.

Дано: множество $\mathcal{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ и алфавит $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K\}$ векторов из $\mathbb{R}^q$, натуральное число J и положительное число D.

Вопрос: существуют ли во множестве $\mathcal{Y}$ непустые непересекающиеся подмножества $\mathcal{Y}_j^1 \subset \mathcal{Y}$, $j = 1, \ldots, J$, и $\mathcal{Y}_k^2 \subset \mathcal{Y}$, $k = 1, \ldots, K$, такие, что имеет место неравенство

$$\sum_{j=1}^{J} \frac{1}{|\mathcal{Y}_j^1|} \Big\| \sum_{\boldsymbol{y} \in \mathcal{Y}_j^1} \boldsymbol{y} \Big\|^2 + \sum_{k=1}^{K} \sum_{\boldsymbol{y} \in \mathcal{Y}_k^2} \big\{ 2(\boldsymbol{y}, \boldsymbol{v}_k) - \|\boldsymbol{v}_k\|^2 \big\} \geqslant D \,? \quad (4)$$

Задача SVS-FN.

Дано: множество $\mathcal{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ и алфавит $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K\}$ векторов из $\mathbb{R}^q$, натуральные числа $M_1, \ldots, M_J$ и положительное число D.

Вопрос: существуют ли во множестве $\mathcal{Y}$ непустые непересекающиеся подмножества $\mathcal{Y}_j^1 \subset \mathcal{Y}$, $j = 1, \ldots, J$, и $\mathcal{Y}_k^2 \subset \mathcal{Y}$, $k = 1, \ldots, K$, такие, что имеет место неравенство (4), при ограничениях $|\mathcal{Y}_j^1| = M_j$, $j = 1, \ldots, J$, на мощности подмножеств?

Задача SVS-NF.

Дано: множество $\mathcal{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ и алфавит $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K\}$ векторов из $\mathbb{R}^q$, натуральные числа $J, N_1, \ldots, N_K$ и положительное число D.

Вопрос: существуют ли во множестве $\mathcal{Y}$ непустые непересекающиеся подмножества $\mathcal{Y}_j^1 \subset \mathcal{Y}$, $j = 1, \ldots, J$, и $\mathcal{Y}_k^2 \subset \mathcal{Y}$, $k = 1, \ldots, K$, такие, что имеет место неравенство

$$\sum_{j=1}^{J} \frac{1}{|\mathcal{Y}_j^1|} \Big\| \sum_{\boldsymbol{y} \in \mathcal{Y}_j^1} \boldsymbol{y} \Big\|^2 + 2 \sum_{k=1}^{J} \sum_{\boldsymbol{y} \in \mathcal{Y}_k^2} (\boldsymbol{y}, \boldsymbol{v}_k) \geqslant D \,, \quad (5)$$

при ограничениях $|\mathcal{Y}_k^2| = N_k$, $k = 1, \ldots, K$ на мощности подмножеств?

Задача SVS-FF.

Дано: множество $\mathcal{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ и алфавит $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K\}$ векторов из $\mathbb{R}^q$, натуральные числа $M_1, \ldots, M_J$, $N_1, \ldots, N_K$ и положительное число D.

Вопрос: существуют ли во множестве $\mathcal{Y}$ непустые непересекающиеся подмножества $\mathcal{Y}_j^1 \subset \mathcal{Y}$, $j = 1, \ldots, J$, и $\mathcal{Y}_k^2 \subset \mathcal{Y}$, $k = 1, \ldots, K$, такие, что имеет место неравенство (5), при ограничениях $|\mathcal{Y}_j^1| = M_j$, $j = 1, \ldots, J$, и $|\mathcal{Y}_k^2| = N_k$, $k = 1, \ldots, K$, на мощности подмножеств?

Теорема 1. *Задачи MSSC-NN, MSSC-FN, MSSC-NF и MSSC-FF NP-полны.*

При доказательстве показано, что к рассматриваемым задачам кластерного анализа полиномиально сводятся сформулированные выше NP-полные задачи выбора подмножеств векторов. Идею доказательства продемонстрируем для одной из задач.

Установим связь, например, между целевыми функциями задач MSSC-NN и SVS-NN. Для целевой функции задачи MSSC-NN имеем:

$$\sum_{j=1}^{J} \sum_{\boldsymbol{y} \in \mathcal{C}_j} \|\boldsymbol{y} - \overline{\boldsymbol{y}}(\mathcal{C}_j)\|^2 + \sum_{k=1}^{K} \sum_{\boldsymbol{y} \in \mathcal{B}_k} \|\boldsymbol{y} - v_k\|^2 + \sum_{\boldsymbol{y} \in \mathcal{B}} \|\boldsymbol{y}\|^2 = \sum_{\boldsymbol{y} \in \mathcal{Y}} \|\boldsymbol{y}\|^2 - \left\{ \sum_{j=1}^{J} \frac{1}{|\mathcal{C}_j|} \Big\| \sum_{\boldsymbol{y} \in \mathcal{C}_j} \boldsymbol{y} \Big\|^2 + \sum_{k=1}^{K} \sum_{\boldsymbol{y} \in \mathcal{B}_k} \big\{ 2(\boldsymbol{y}, \boldsymbol{v}_k) - \|\boldsymbol{v}_k\|^2 \big\} \right\}. \quad (6)$$

Первый член в правой части полученного равенства — константа, а второй член — выражение в фигурных скобках — целевая функция задачи SVS-NN от подмножеств $\mathcal{Y}_j^1 = \mathcal{C}_j$, $j = 1, \ldots, J$, и $\mathcal{Y}_k^2 = \mathcal{B}_k$, $k = 1, \ldots, K$. Отсюда следует, что NP-полная задача SVS-NN полиномиально сводится к задаче MSSC-NN. Действительно, из (3), (4) и (6) легко видеть, что разбиение множества $\mathcal{Y}$ на кластеры $\mathcal{C}_1, \ldots, \mathcal{C}_J$, $\mathcal{B}_1, \ldots, \mathcal{B}_K$ и $\mathcal{B} = \mathcal{Y} \backslash ((\cup_j \mathcal{C}_j) \cup (\cup_k \mathcal{B}_k))$ в задаче MSSC-NN существует тогда и только тогда, когда в задаче SVS-NN при $D = \sum_{\boldsymbol{y} \in \mathcal{Y}} \|\boldsymbol{y}\|^2 - A$ существуют соответствующие непустые подмножества $\mathcal{Y}_j^1$, $j = 1, \ldots, J$, и $\mathcal{Y}_k^2$, $k = 1, \ldots, K$.

NP-полнота остальных задач доказывается аналогично.

Подчеркнем, что задачи остаются NP-полными даже в простейшем случае, когда $K = J = 1$.

Заключение

В работе установлена NP-полнота нескольких актуальных задач кластеризации. Установленный факт представляется значимым дополнением как к результатам, опубликованным в [13, 14], так и к систематизации труднорешаемых задач кластерного анализа. Полученные результаты могут служить в качестве полезного инструмента для анализа алгоритмической сложности других задач кластеризации.

Какие-либо эффективные алгоритмы с оценками точности для решения оптимизационных вариантов рассмотренных задач на сегодняшний день неизвестны. Относительно возможных алгоримических решений заметим следующее.

Фактически, задачи кластерного анализа, рассмотренные в настоящей работе, и задачи выбора непересекающихся подмножеств, изучавшиеся в [13, 14], являются парами взаимно противоположных задач. Взаимная противоположность здесь

трактуется в том смысле, что минимизация целевых функций в задачах кластеризации эквивалентна максимизации целевых функций соответствующих задач поиска подмножеств. Эквивалентность следует из того, что сумма двух целевых функций (от одинаковых переменных) взаимно противоположных задач равна константе (что было установлено при доказательстве теоремы 1). Поэтому по точному или асимптотически точному решению задач поиска подмножеств (или кластеризации) можно найти такое же по точности решение противоположных задач кластеризации (или поиска подмножеств). Однако из-за связи целевых функций противоположных задач через аддитивную константу для алгоритмического решения задачи поиска подмножеств (или кластеризации), имеющего константную оценку точности, будет весьма проблематично найти константную оценку точности соответствующей задачи кластеризации (или поиска подмножеств). Потребуются индивидуальные подходы к построению алгоритмов с константными оценками точности для решения взаимно противоположных задач.

Литература

[1] *Anil K. Jain K.* Data Clustering: 50 Years Beyond k-Means // Pattern Recognition Letters. — 2010. — Vol. 31. — Pp. 651–666.

[2] *Aloise D., Hansen P.* On the Complexity of Minimum Sum-of-Squares Clustering // Les Cahiers du GERAD, G-2007-50. — 2007. — 12 p.

[3] *MacQueen J. B.* Some Methods for Classification and Analysis of Multivariate Observations // Proc. 5-th Berkeley Symp. of Mathematical Statistics and Probability. — 1967. — Vol. 1. — Pp. 281–297.

[4] *Rao M.* Cluster Analysis and Mathematical Programming // J. Am. Stat. Assoc. — 1971. — Vol. 66. — Pp. 622–626.

[5] *Inaba M., Katch N., Imai H.* Applications of Weighted Voronoi Diagrams and Randomization to Variance-Based Clustering // Proc. Annual Symp. on Comput. Geom. — 1994. — Pp. 332–339.

[6] *Aloise D., Deshpande A., Hansen P., Popat P.* NP-Hardness of Euclidean Sum-of-Squares Clustering // Les Cahiers du GERAD, G-2008-33. — 2008. — 4 p.

[7] *Mahajan M., Nimbhorkar P., Varadarajan K.* The Planar k-means Problem is NP-Hard // Lecture Notes in Computer Science. — 2009. — Vol. 5431. — Pp. 284–285.

[8] *Долгушев А. В., Кельманов А. В.* К вопросу об алгоритмической сложности одной задачи кластерного анализа // Дискретный анализ и исследование операций. — 2010. — Т. 17, № 2. — С. 39–45.

[9] *Кельманов А. В., Пяткин А. В.* О сложности одного из вариантов задачи выбора подмножества «похожих» векторов // Докл. РАН. — 2008. — Т. 421, № 5. — С. 590–592.

[10] *Кельманов А. В., Пяткин А. В.* Об одном варианте задачи выбора подмножества векторов // Дискретный анализ и исследование операций. — 2008. — Т. 15, № 5. — С. 25–40.

[11] *Кельманов А. В., Пяткин А. В.* О сложности некоторых задач поиска подмножеств векторов и кластерного анализа // Журн. вычисл. математики и мат. физики. — 2009. — Т. 49, № 11. — С. 2059–2067.

[12] *Кельманов А. В., Пяткин А. В.* NP-полнота некоторых задач выбора подмножества векторов // Дискретный анализ и исследование операций. — 2010. — Т. 17, № 5. — С. 37–45.

[13] *Кельманов А. В.* О сложности некоторых задач анализа данных // Журн. вычисл. математики и мат. физики. — 2010. — Т. 50, № 11. — С. 2045–2051.

[14] *Кельманов А. В.* NP-полнота некоторых задач поиска подмножеств векторов // Тр. Ин-та математики и механики УрО РАН. — 2010. — Т. 16, № 3. — С. 121–129.

Алгоритмы с оценками для некоторых задач поиска подмножества векторов и кластерного анализа*

Кельманов А. В., Романченко С. М.
kelm@math.nsc.ru, semenr@bk.ru
Новосибирск, Институт математики им. С. Л. Соболева СО РАН, Новосибирский государственный университет

Анализируются некоторые NP-трудные задачи кластеризации и поиска в заданном множестве векторов евклидова пространства подмножества векторов фиксированной мощности. К этим задачам сводится одна из проблем анализа данных по критерию минимума суммы квадратов. Обоснованы эффективные 2-приближенные алгоритмы решения этих задач, а также псевдополиномиальные алгоритмы, гарантирующие отыскание оптимума, в случае, когда компоненты векторов имеют целочисленные значения и размерность пространства фиксирована.

Объект исследования настоящей работы — проблемы оптимизации в задачах анализа данных и распознавания образов. Предмет исследования — некоторые труднорешаемые экстремальные задачи поиска подмножества векторов евклидова пространства и кластерного анализа. Цель работы — обоснование алгоритмов для решения этих задач.

Одна из возможных содержательных трактовок проблемы, которая приводит к решению рассматриваемых задач, состоит в следующем. Имеется таблица, содержащая результаты измерения набора числовых информационно значимых характеристик для совокупности некоторых материальных объектов. Часть объектов из этой совокупности идентичны и имеют одинаковые характеристики. Число идентичных объектов известно. Оставшиеся объекты различны и имеют отличающиеся характеристики. В каждом результате измерения, представленном в таблице, имеется ошибка, причем соответствие между объектом и набором неизвестно. Требуется, используя критерий минимума суммы квадратов расстояний, найти подмножество наборов, соответствующих идентичным объектам, и оценить по результатам измерения набор характеристик этих объектов (учитывая, что данные содержат ошибку измерения).

В работе [1] было установлено, что эта проблема сводится к четырем тесно связанным между собой труднорешаемым задачам оптимизации квадратичных функций. Эти задачи близки к хорошо известной задаче MSSC (Minimum-Sum-of-Squares Clustering) в постановочном плане, но не эквивалентны ей. Одна из этих задач является задачей на максимум, а остальные — на минимум. Формулировки задач приведены в следующем параграфе. Здесь лишь отметим, что в форме верификации свойств эти задачи NP-полны в сильном смысле [1]. Поэтому, как известно [2], псевдополиномиальные алгоритмы для их решения не существуют (в предположении справедливости гипотезы $\mathsf{P} \neq \mathsf{NP}$). Эти факты послужили стимулом как к обоснованию эффективных приближенных алгоритмов, так и к поиску и выделению таких подклассов этих задач, для которых возможно построение точных или приближенных алгоритмов, имеющих полиномиальную или псевдополиномиальную сложность.

Работа выполнена при финансовой поддержке РФФИ, проекты №09-01-00032, №10-07-00195; целевой программы №2 Президиума РАН, проект №227; целевой программы СО РАН, интеграционный проект №44; целевой программы АВЦП Рособразования, проект №2.1.1/3235; а также федеральной целевой программы «Научные и научно-педагогические кадры инновационной России», гос. контракт №14.740.11.0362.

Задачи поиска подмножества векторов и кластерного анализа

Четыре сформулированные ниже задачи поиска подмножества векторов и кластерного анализа индуцируются [1] одной и той же математической моделью содержательной проблемы анализа данных, приведенной во введении.

Задача MSSC-Case.

Дано: множество $\mathcal{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ векторов из $\mathbb{R}^q$ и натуральное число $M > 1$.

Найти: разбиение множества $\mathcal{Y}$ на $J = N - M + 1$ непустых кластеров $\mathcal{C}_1, \ldots, \mathcal{C}_J$ такое, что мощность одного из этих кластеров равна M и

$$S(\mathcal{C}_1, \ldots, \mathcal{C}_J) = \sum_{j=1}^{J} \sum_{\boldsymbol{y} \in \mathcal{C}_j} \|\boldsymbol{y} - \overline{\boldsymbol{y}}(\mathcal{C}_j)\|^2 \to \min, \quad (1)$$

где $\overline{\boldsymbol{y}}(\mathcal{C}_j) = (1/|\mathcal{C}_j|) \sum_{\boldsymbol{y} \in \mathcal{C}_j} \boldsymbol{y}$, $j = 1, \ldots, J$, — центр j-го кластера?

Задача VS-1 (Vector Subset 1).

Дано: множество $\mathcal{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ векторов из $\mathbb{R}^q$, натуральное число $M > 1$.

Найти: подмножество $\mathcal{C} \subseteq \mathcal{Y}$ векторов такое, что целевая функция

$$Q(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \left\| \sum_{\boldsymbol{y} \in \mathcal{C}} \boldsymbol{y} \right\|^2 + \sum_{\boldsymbol{y} \in \mathcal{Y} \setminus \mathcal{C}} \|\boldsymbol{y}\|^2$$

максимальна, при ограничении $|\mathcal{C}| = M$ на мощность подмножества $\mathcal{C}$.

Задача VS-2 (Vector Subset 2).

Дано: множество $\mathcal{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ векторов из $\mathbb{R}^q$ и натуральное число $M > 1$.

Найти: подмножество $\mathcal{C} \subseteq \mathcal{Y}$ векторов такое, что целевая функция

$$F(\mathcal{C}) = \sum_{\boldsymbol{y} \in \mathcal{C}} \|\boldsymbol{y} - \overline{\boldsymbol{y}}(\mathcal{C})\|^2 , \qquad (2)$$

где $\overline{\boldsymbol{y}}(\mathcal{C}) = (1/|\mathcal{C}|) \sum_{\boldsymbol{y} \in \mathcal{C}} \boldsymbol{y}$ минимальна, при ограничении $|\mathcal{C}| = M$ на мощность искомого подмножества.

Задача VS-3 (Vector Subset 3).

Дано: множество $\mathcal{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ векторов из $\mathbb{R}^q$, натуральное число $M > 1$.

Найти: подмножество $\mathcal{C} \subseteq \mathcal{Y}$ векторов такое, что целевая функция

$$H(\mathcal{C}) = \sum_{\boldsymbol{y} \in \mathcal{C}} \sum_{\boldsymbol{z} \in \mathcal{C}} \|\boldsymbol{y} - \boldsymbol{z}\|^2$$

минимальна, при ограничении $|\mathcal{C}| = M$ на мощность искомого подмножества.

Между целевыми функциями задач VS-1, VS-2 и VS-3 имеется однозначное соответствие [1]:

$$F(\mathcal{C}) = \sum_{\boldsymbol{y} \in \mathcal{Y}} \|\boldsymbol{y}\|^2 - Q(\mathcal{C}) = \frac{1}{2|\mathcal{C}|} H(\mathcal{C}) . \qquad (3)$$

Кроме того, если считать, что в задаче MSSC-Case мощность, например, j-го кластера $\mathcal{C}_j$ зафиксирована и равна M, то имеет место равенство [1]

$$S(\mathcal{C}_1, \ldots, \mathcal{C}_j, \ldots, \mathcal{C}_J) = F(\mathcal{C}_j) , \qquad (4)$$

т. к. из условий задачи следует, что мощности кластеров из совокупности $\{\mathcal{C}_1, \ldots, \mathcal{C}_j, \ldots, \mathcal{C}_J\} \setminus \mathcal{C}_j$ равны 1. Поэтому задачи MSSC-Case и VS-2 эквивалентны.

Формулы (3) и (4), связывающие целевые функции задач, используются далее при построении эффективных приближенных и точных псевдополиномиальных алгоритмов. В качестве базовой рассматривается задача VS-2. По её решению отыскиваются решения остальных задач.

2-приближенные алгоритмы

Обозначим через $\mathcal{C}^*$ — оптимальное решение задачи VS-2. Положим $\boldsymbol{c}^* = \overline{\boldsymbol{y}}(\mathcal{C}^*)$, где $\overline{\boldsymbol{y}}(\cdot)$ — функция, определенная в формулировке этой задачи. Свойство оптимального решения устанавливает

Лемма 1. *Для любого вектора $\boldsymbol{y} \in \mathcal{C}^*$ и для любого вектора $\boldsymbol{z} \in \mathcal{Y} \setminus \mathcal{C}^*$ справедливо неравенство*

$$\|\boldsymbol{y} - \boldsymbol{c}^*\| \leqslant \|\boldsymbol{z} - \boldsymbol{c}^*\| .$$

Доказательство этого и других ниже следующих утверждений, обосновывающих алгоритмические решения, приведено в работах [3, 4].

Лемма 1 показывает, что оптимальное решение задачи VS-2 — подмножество $\mathcal{C}^* \subseteq \mathcal{Y}$ — состоит из векторов, ближайших к вектору $\boldsymbol{c}^*$ по расстоянию. Она устанавливает необходимое условие минимума и указывает на изложенный ниже возможый подход к решению задачи. Реализация этого подхода опирается на решение следующей вспомогательной задачи.

Задача VSVN (Vector and Subset of Vectors which are Nearest to this vector).

Дано: множество $\mathcal{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ векторов из $\mathbb{R}^q$ и натуральное число $M > 1$.

Найти: подмножество $\mathcal{B} \subseteq \mathcal{Y}$ векторов мощности M и вектор $\boldsymbol{b} \in \mathcal{Y}$ такие, что целевая функция

$$G(\mathcal{B}, \boldsymbol{b}) = \sum_{\boldsymbol{y} \in \mathcal{B}} \|\boldsymbol{y} - \boldsymbol{b}\|^2 \qquad (5)$$

минимальна.

Обозначим через $\mathcal{B}^*$ и $\boldsymbol{b}^*$ подмножество и вектор, доставляющие минимум G^* целевой функции G. Построим алгоритм решения этой задачи.

Алгоритм $\mathcal{A}_1$.

Шаг 1. Для каждого вектора $\boldsymbol{y} \in \mathcal{Y}$ найдем множество $\mathcal{B}(\boldsymbol{y})$, состоящее из вектора $\boldsymbol{y}$ и $M - 1$ векторов множества $\mathcal{Y}$, ближайших по расстоянию к вектору $\boldsymbol{y}$. Вычислим значение целевой функции $G(\mathcal{B}(\boldsymbol{y}), \boldsymbol{y})$.

Шаг 2. Среди найденных на шаге 1 множеств выберем в качестве решения то множество $\mathcal{B}^*$ и вектор $\boldsymbol{b}^*$, для которых значение целевой функции G минимально. Если минимальному значению целевой функции соответствует несколько решений, то в качестве окончательного решения выберем любое из этих решений.

Оценку сложности и точности алгоритма устанавливает

Лемма 2. *Алгоритм $\mathcal{A}_1$ находит оптимальное решение задачи VSVN за время $\mathcal{O}(qN^2)$.*

Сформулируем вспомогательные утверждения.

Лемма 3. *Пусть $\mathcal{Z}$ — непустое конечное множество векторов из $\mathbb{R}^q$, а $\overline{\boldsymbol{z}}(\mathcal{Z}) = (1/|\mathcal{Z}|) \sum_{\boldsymbol{z} \in \mathcal{Z}} \boldsymbol{z}$. Тогда, если вектор $\boldsymbol{x} \in \mathbb{R}^q$ удовлетворяет условиям*

$$\|\boldsymbol{x} - \overline{\boldsymbol{z}}\| \leqslant \|\boldsymbol{z} - \overline{\boldsymbol{z}}\| , \qquad \forall \boldsymbol{z} \in \mathcal{Z} ,$$

то имеет место неравенство

$$\sum_{\boldsymbol{z} \in \mathcal{Z}} \|\boldsymbol{z} - \boldsymbol{x}\|^2 \leqslant 2 \sum_{\boldsymbol{z} \in \mathcal{Z}} \|\boldsymbol{z} - \overline{\boldsymbol{z}}\|^2 .$$

Лемма 4. *Пусть $\mathcal{B}^*$, b^* — оптимальное решение задачи VSVN, а $\mathcal{C}^*$ — оптимальное решение задачи VS-2. Тогда $F(\mathcal{B}^*) \leqslant 2F(\mathcal{C}^*)$.*

Опираясь на лемму 4, представим алгоритм решения задачи VS-2.

Алгоритм $\mathcal{A}_2$.

Шаг 1. По заданному множеству $\mathcal{Y}$ и числу M находим оптимальное решение $\mathcal{B}^*, b^*$ вспомогательной задачи VSVN с помощью алгоритма $\mathcal{A}_1$.

Шаг 2. Подмножество $\mathcal{B}^*$ объявляем решением задачи VS-2.

Из лемм 1–4 следует

Теорема 5. *Алгоритм $\mathcal{A}_2$ находит приближённое решение задачи VS-2 с гарантированной оценкой точности 2 за время $\mathcal{O}(qN^2)$. Оценка 2 точности алгоритма достижима и неулучшаема.*

Изложим алгоритм приближённого решения задачи MSSC-Case.

Шаг 1. По заданному множеству $\mathcal{Y}$ и числу M находим приближённое решение $\mathcal{B}^*$ задачи VS-2 с помощью алгоритма $\mathcal{A}_2$.

Шаг 2. Решением задачи MSSC-Case объявляем кластер $\mathcal{C}_1 = \mathcal{B}^*$ мощности M и совокупность $\{\mathcal{C}_2, \ldots, \mathcal{C}_J\} = \mathcal{Y} \backslash \mathcal{B}^*$ одноэлементных кластеров.

В силу (4) леммы 4 и теоремы 5 этот алгоритм гарантирует эффективное отыскание 2-приближённого решения задачи MSSC-Case.

Алгоритм $\mathcal{A}_2$ можно использовать для эффективного приближённого решения задачи VS-1. Однако найденное с помощью алгоритма $\mathcal{A}_2$ решение задачи VS-1 не будет иметь каких-либо теоретических гарантий по точности решения из-за того, что, согласно (3), целевые функции задач VS-2 и VS-1 связаны через аддитивную константу — сумму квадратов норм векторов из множества $\mathcal{Y}$.

Наконец, построенный алгоритм можно использовать для эффективного 2-приближённого решения задачи VS-3. Действительно, возьмем в качестве приближённого решения этой задачи подмножество $\mathcal{B}^*$, найденное с помощью алгоритма $\mathcal{A}_2$. Для этого решения значение целевой функции задачи VS-3 равно $H(\mathcal{B}^*) = 2M \cdot F(\mathcal{B}^*)$. Поэтому в силу (3), леммы 4 и теоремы 5 для точности решения задачи имеем оценку $H(\mathcal{B}^*)/H(\mathcal{C}^*) = F(\mathcal{B}^*)/F(\mathcal{C}^*) \leqslant 2$.

Точные псевдополиномиальные алгоритмы

Допустим, что компоненты векторов из множества $\mathcal{Y}$ имеют целочисленные значения. Положим

$$B = \max_{\boldsymbol{y} \in \mathcal{Y}} \max_{j \in \{1,\ldots,q\}} |(\boldsymbol{y})^j| \,, \tag{6}$$

где $(\boldsymbol{y})^j$ — j-я компонента вектора $\boldsymbol{y}$. Определим множество

$$\mathcal{B} = \big\{\boldsymbol{b} \big| \, (\boldsymbol{b})^j = \frac{1}{M}(\boldsymbol{v})^j, \;\; (\boldsymbol{v})^j \in \mathbb{Z}, \; |(\boldsymbol{v})^j| \leqslant MB, j = 1, \ldots, q\big\} \tag{7}$$

векторов из $\mathbb{R}^q$. Заметим, что

$$|\mathcal{B}| = (2MB + 1)^q \,. \tag{8}$$

Суть подхода состоит в следующем. Сначала вычисляется семейство значений функции

$$G(\boldsymbol{b}) = \sum_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{b})} \|\boldsymbol{y} - \boldsymbol{b}\|^2, \;\; \boldsymbol{b} \in \mathcal{B} \,, \tag{9}$$

где $\mathcal{Y}(\boldsymbol{b})$ — подмножество множества $\mathcal{Y}$, состоящее из M векторов, ближайших к вектору $\boldsymbol{b}$ по расстоянию. Затем в качестве решения задачи выбирается подмножество $\mathcal{Y}(\boldsymbol{b}) \subseteq \mathcal{Y}$, для которого значение функции (9) имеет наименьшее значение. Иными словами, алгоритмическое решение ищется в виде

$$\mathcal{C}_A = \mathcal{Y}(\boldsymbol{b}_A) \,,$$

где

$$\boldsymbol{b}_A = \arg\min_{\boldsymbol{b} \in \mathcal{B}} G(\boldsymbol{b}) \,.$$

Допустим, что векторы из множества $\mathcal{B}$ упорядочены, например, в лексикографическом порядке. Изложим алгоритм решения задачи VS-2.

Алгоритм $\mathcal{A}_3$.

Шаг 1. Найдём значение B и мощность $|\mathcal{B}|$ множества $\mathcal{B}$ по формулам (6) и (8). Положим $\mathcal{C}_A = \varnothing$, $G_{\min} = +\infty$, $i = 0$.

Шаг 2. $i := i + 1$; в соответствии с (7) сформируем i-й элемент $\boldsymbol{b}_i$ множества $\mathcal{B}$, учитывая лексикографический порядок его элементов, и положим $\boldsymbol{b} = \boldsymbol{b}_i$.

Шаг 3. Найдем множество $\mathcal{Y}(\boldsymbol{b})$ ближайших к $\boldsymbol{b}$ векторов из множества $\mathcal{Y}$.

Шаг 4. Вычислим значение $G(\boldsymbol{b})$ функции (9).

Шаг 5. Если $G(\boldsymbol{b}) \leqslant G_{\min}$, то положим $\boldsymbol{b}_A = \boldsymbol{b}$, $G_{\min} = G(\boldsymbol{b}_A)$, $\mathcal{C}_A = \mathcal{Y}(\boldsymbol{b}_A)$; иначе переходим к следующему шагу.

Шаг 6. Если $i < |\mathcal{B}|$, то переходим на шаг 2; иначе — к следующему шагу.

Шаг 7. Подмножество $\mathcal{C}_A$ и значение $F_A = F(\mathcal{C}_A)$ объявляем результатом работы алгоритма.

Свойство алгоритмического решения устанавливает

Теорема 6. *Пусть в условиях задачи VS-2 компоненты всех векторов из множества $\mathcal{Y}$ имеют целочисленные значения в интервале $[-B, B]$. Тогда алгоритм $\mathcal{A}_3$ находит оптимальное решение этой задачи за время $\mathcal{O}(qN(2MB + 1)^q)$.*

Таким образом, для алгоритмического и оптимального решений имеем равенство $\mathcal{C}^* = \mathcal{C}_A$.

Ключевым элементом в доказательстве теоремы 6 является пара свойств оптимального решения. Первое свойство сформулировано в лемме 1, а второе, легко устанавливаемое свойство, состоит в том, что оптимальный вектор $\boldsymbol{c}^*$ лежит во множестве $\mathcal{B}$.

Замечание 1. Псевдополиномиальность алгоритма $\mathcal{A}_3$ при фиксированной размерности q пространства следует из того, что временная сложность алгоритма зависит от значений числовых данных (а именно, от B) на входе задачи.

Так как задача MSSC-Case эквивалентна задаче VS-2, алгоритм отыскания её оптимального решения заключается в следующем.

Шаг 1. По заданному множеству $\mathcal{Y}$ и числу M находим оптимальное решение $\mathcal{C}^*$ задачи VS-2 с помощью алгоритма $\mathcal{A}_3$ и значение $F(\mathcal{C}^*)$ целевой функции F.

Шаг 2. Решением задачи MSSC-Case объявляем кластер $\mathcal{C}_1^* = \mathcal{C}^*$ мощности M и совокупность $\{\mathcal{C}_2^*, \ldots, \mathcal{C}_{N-M+1}^*\} = \mathcal{Y} \backslash \mathcal{C}^*$ одноэлементных кластеров. Для оптимального значения целевой функции S полагаем $S^* = F(\mathcal{C}^*)$ в соответствии с (4).

Алгоритмы поиска оптимального решения задач VS-1 и VS-3 аналогичны. Они содержат по 2 шага и отличаются от изложенного алгоритма только на втором шаге. Решением этих задач объявляем подмножество $\mathcal{C}^*$, найденное на первом шаге с помощью алгоритма $\mathcal{A}_3$. На втором шаге алгоритмов решения задач VS-1 и VS-3 по найденному подмножеству $\mathcal{C}^*$ вычисляем оптимальные значения $Q(\mathcal{C}^*)$ и $H(\mathcal{C}^*)$ целевых функций соответствующих задач, используя формулу (3).

Заключение

В работе построены 2-приближенные эффективные и точные псевдополиномиальные алгоритмы решения некоторых NP-трудных задач, к которым сводится оптимизационная модель одной из проблем анализа данных. Рассмотренные задачи по своей сути являются задачами кластеризации заданного множества векторов евклидова пространства и поиска в этом множестве подмножества векторов фиксированной мощности по критерию минимума суммы квадратов расстояний.

Поскольку рассмотренные задачи относятся к числу слабо изученных в алгоритмическом плане, исследование вопросов их аппроксимируемости, а также обоснование алгоритмов другого типа (асимптотически точных, рандомизированных и др.) для их решения представляется делом ближайшей перспективы.

Литература

[1] *Кельманов А. В., Пяткин А. В.* NP-полнота некоторых задач выбора подмножества векторов // Дискретный анализ и исследование операций. — 2010. — Т. 17, № 5. — С. 37–45.

[2] *Garey M. R., Johnson D. S.* Computers and Intractability: A Guide to the Theory of NP-Completeness. San Francisco: Freeman. — 1979. — 314 p.

[3] *Кельманов А. В., Романченко С. М.* Приближённый алгоритм для решения одной задачи поиска подмножества векторов // Дискретный анализ и исследование операций. — 2011. — Т. 18, № 1. — С. 61–69.

[4] *Кельманов А. В., Романченко С. М.* Псевдополиномиальные алгоритмы для некоторых труднорешаемых задач поиска подмножества векторов и кластерного анализа // Автоматика и телемеханика. — 2011 (принята в печать).

Об одной задаче поиска и идентификация векторных наборов в последовательности*

Кельманов А. В., Михайлова Л. В., Хамидуллин С. А.
kelm@math.nsc.ru, mikh@math.nsc.ru, kham@math.nsc.ru
Новосибирск, Институт математики им. С. Л. Соболева СО РАН, Новосибирский государственный университет

Рассматривается дискретная экстремальная задача, к которой сводится одна из проблем поиска и идентификации векторных наборов в последовательности. Построен полиномиальный off-line алгоритм, гарантирующий оптимальность решения этой задачи и имеющий меньшую трудоемкость по сравнению с известным аналогом.

Объект исследования настоящей работы — проблемы оптимизации в области анализа данных и распознавания образов. Предмет исследования — дискретная экстремальная задача, к которой сводится одна из проблем помехоустойчивого обнаружения и идентификации векторных наборов в последовательности. Цель работы — построение эффективного алгоритма, позволяющего получить оптимальное решение задачи за существенно меньшее по сравнению с известным алгоритмом время.

Одна из возможных трактовок рассматриваемой проблемы состоит в следующем. Каждому символу (букве) естественного языка поставим в соответствие ненулевой вектор из конечномерного евклидова пространства, а пробелу — нуль-вектор. Тогда каждому слову языка соответствует комбинация векторов (или векторный набор). Представим себе текст на естественном языке, устроенный так, что между символами этого текста вставлены пробелы, число которых неизвестно, но ограничено сверху и снизу некоторыми константами. Этому тексту будет соответствовать векторная последовательность, в которой ненулевые векторы перемежаются с нуль-векторами. Допустим, что эта векторная последовательность искажена аддитивной помехой. Предположим, что в нашем распоряжении имеется совокупность (словарь) векторных наборов. Задача состоит в том, чтобы в искаженной шумом последовательности, включающей ненулевые векторы, перемежающиеся с нуль-векторами, найти и идентифицировать векторные наборы. На языке символьных последовательностей эта задача соответствует поиску и идентификации слов в тексте, который имеет специальную (описанную выше) структуру. Задача актуальна для ряда приложений, связанных с помехоустойчивым анализом данных и распознаванием сигналов (см., например, [1–4] и цитированные там работы).

Работа выполнена при финансовой поддержке РФФИ, проекты № 09-01-00032, № 10-07-00195 и № 11-01-00696; целевой программы № 2 Президиума РАН, проект № 227; целевой программы СО РАН, интеграционный проект № 44; целевой программы АВЦП Рособразования, проект № 2.1.1/3235; а также федеральной целевой программы «Научные и научно-педагогические кадры инновационной России», гос. контракт № 14.740.11.0362.

Дискретная экстремальная задача, к которой сводится эта проблема, рассматривалась в [4]. В этой работе был обоснован полиномиальный off-line алгоритм, гарантирующий оптимальность ее решения по критерию минимума суммы квадратов. В настоящей работе предложен новый точный полиномиальный алгоритм решения задачи, имеющий существенно меньшую трудоемкость.

Формальная постановка задачи

Рассмотрим следующую модель анализируемых данных.

Допустим, что последовательность $\boldsymbol{x}_n \in \mathbb{R}^q$, $n \in \mathcal{N} = \{1, \ldots, N\}$, порождена векторным набором $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_J)$, в котором $\boldsymbol{w}_j = (\boldsymbol{u}^j_1, \ldots, \boldsymbol{u}^j_{L_j})$, $j = 1, \ldots, J$, и имеет следующую структуру:

$$\boldsymbol{x}_n = \begin{cases} \boldsymbol{u}^j_i, & \text{если } n = n_{s_{j-1}+i}, \quad j = 1, \ldots J, \\ & \quad i = 1, \ldots, \min\{L_j, M - s_{j-1}\}; \\ 0, & \text{если } n \in \mathcal{N} \backslash \{n_1, \ldots, n_M\}, \end{cases} \quad (1)$$

где $s_0 = 0$, $s_j = \sum_{k=1}^{j} L_k$, L_k — размерность k-го векторного набора, а элементы подмножества $\{n_1, \ldots, n_M\} \subseteq \mathcal{N}$ номеров удовлетворяют ограничениям

$$1 \leqslant T_{\min} \leqslant n_m - n_{m-1} \leqslant T_{\max} \leqslant N - 1, \quad m = 2, \ldots, M, \quad (2)$$

где $T_{\min}$ и $T_{\max}$ — натуральные числа.

Определим совокупность (словарь)

$$\mathcal{W} \subset \big\{ \boldsymbol{w} \,|\, \boldsymbol{w} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_L),\ \boldsymbol{u}_i \in \mathbb{R}^q, \ \|\boldsymbol{u}_i\| \neq 0,\ \ i = 1, \ldots, L,\ \ L \leqslant L_{\max} \big\}, \quad (3)$$

$|\mathcal{W}| = K$, векторных наборов. В формуле (3) $L_{\max}$ — максимальная размерность слова в словаре.

Предположим, что набор $\boldsymbol{w}_j$, $j = 1, \ldots, J$, из совокупности $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_J)$, порождающей последовательность (1), является элементом словаря $\mathcal{W}$. Допустим, что для наблюдения доступна последовательность

$$\boldsymbol{y}_n = \boldsymbol{x}_n + \boldsymbol{e}_n,\ n \in \mathcal{N}, \quad (4)$$

где $\boldsymbol{e}_n$ — вектор помехи (ошибки измерения), независимый от вектора $\boldsymbol{x}_n$.

Учитывая зависимость элементов последовательности (1) от набора $(n_1, \ldots, n_M)$ и совокупности $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_J)$, положим

$$S(n_1, \ldots, n_M, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_J) = \sum_{n \in \mathcal{N}} \|\boldsymbol{y}_n - \boldsymbol{x}_n\|^2 . \quad (5)$$

Рассмотрим модель анализа данных в форме следующей оптимизационной задачи.

Дано: последовательность $\boldsymbol{y}_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, совокупность $\mathcal{W}$ наборов векторов из $\mathbb{R}^q$ и натуральные числа $T_{\min}$ и $T_{\max}$.

Найти: последовательность $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_J$ векторных наборов из совокупности $\mathcal{W}$, а также натуральное число M и подмножество $\{n_1, \ldots, n_M\} \subseteq \mathcal{N}$ номеров такие, что целевая функция (5) минимальна, при ограничениях (2) на элементы набора $\{n_1, \ldots, n_M\}$ и при условии, что структура последовательности описывается формулами (1)–(4).

Из (1) видно, что последовательность $\boldsymbol{x}_n$, $n \in \in \mathcal{N}$, содержит J участков, каждый из которых соответствует слову. Элементы набора $\{n_1, \ldots, n_M\}$ соответствуют номерам ненулевых векторов в последовательности (1). Переменная s_j, $j = 1, \ldots, J$, обозначает порядковый номер вектора в последовательности (1), завершающего j-й векторный набор w_j; последний J-й набор w_J может лишь частично входить в состав последовательности (1). Неравенства (2) ограничивают сверху и снизу интервал между номерами двух последовательных ненулевых векторов в последовательности (1).

Редуцированная задача

Следуя [4], раскроем квадрат нормы в формуле (5) и, с учетом (1), получим

$$S = \sum_{n \in \mathcal{N}} \|\boldsymbol{y}_n\|^2 - \sum_{j=1}^{J} \sum_{m=s_{j-1}+1}^{\min\{s_j, M\}} \left(2\langle \boldsymbol{y}_{n_m}, \boldsymbol{u}^j_{m-s_{j-1}} \rangle - \right. \\ \left. - \|\boldsymbol{u}^j_{m-s_{j-1}}\|^2\right),$$

где $\langle \cdot, \cdot \rangle$ — скалярное произведение векторов.

Первое слагаемое в правой части полученного выражения — константа. Поэтому минимизация функционала $S(\cdot)$ сводится к максимизации двойной суммы в правой части этого выражения.

Обозначим через $L(\boldsymbol{w})$ размерность набора $\boldsymbol{w} \in \in \mathcal{W}$. Положим $\boldsymbol{w} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{L(\boldsymbol{w})})$. Определим функцию

$$g(\boldsymbol{w}, i, n) = 2\langle \boldsymbol{y}_n, \boldsymbol{u}_i \rangle - \|\boldsymbol{u}_i\|^2 ,$$

где $i \in \{1, \ldots, L(\boldsymbol{w})\}$, $n \in \mathcal{N}$.

В соответствии с введенными обозначениями имеем следующую оптимизационную задачу.

Задача SIVTS (Searching and Identification of Vector Tuples in a Sequence).

Дано: последовательность $\boldsymbol{y}_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, совокупность $\mathcal{W}$ наборов векторов из $\mathbb{R}^q$ и натуральные числа $T_{\min}$ и $T_{\max}$.

Найти: последовательность $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_J$ векторных наборов из совокупности $\mathcal{W}$, натуральное число M и подмножество $\{n_1, \ldots, n_M\} \subseteq \mathcal{N}$ номеров такие, что целевая функция

$$G = \sum_{j=1}^{J} \sum_{m=s_{j-1}+1}^{\min\{s_j, M\}} g(\boldsymbol{w}_j, m - s_{j-1}, n_m) ,$$

где $s_0 = 0$, максимальна, при ограничениях (2) на элементы набора $\{n_1, \ldots, n_M\}$.

Легко установить, что если $\boldsymbol{e}_n$ в формуле (4) есть выборка единичного объема из q-мерного нормального распределения с параметрами $(0, \sigma^2 I)$, где I единичная матрица, а в качестве критерия решения задачи использовать максимум функционала правдоподобия, то статистический подход к рассматриваемой проблеме обнаружения и идентификации приводит к такой же экстремальной задаче.

Алгоритм решения задачи

Определим множество

$$\omega_m(M) = [(m-1)T_{\min} + 1, N - (M-m)T_{\min}] , \\ m \in \{1, \ldots, M\} , \qquad M \in \{M^-, \ldots, M^+\} ,$$

где

$$M^- = 1 ; \\ M^+ = \lfloor (N-1)/T_{\min} + 1 \rfloor .$$

Множество $\omega_m(M)$ задает область допустимых значений переменной n_m из набора $\{n_1, \ldots, n_M\}$.

Опираясь на это определение, положим

$$\omega^+ = \bigcup_{M=M^-}^{M^+} \omega_1(M) ; \qquad \omega^- = 7 \bigcup_{M=M^-}^{M^+} \omega_M(M) ;$$

$$\omega^1 = \bigcup_{M=M^-}^{M^+} \bigcup_{m=2}^{M} \omega_m(M) ;$$

$$\gamma^-(n) = \left\{\omega^+ \cup \omega^1\right\} \cap [n - T_{\min}, n - T_{\max}], \quad n \in \omega^1 .$$

Справедлива следующая

Лемма 1. *Максимум целевой функции G в задаче SIVTS определяется по формуле*

$$G_{\max} = \max_{\boldsymbol{w} \in \mathcal{W}} \max_{i \in \{1, \ldots, L(\boldsymbol{w})\}} \max_{n \in \omega^-} G(\boldsymbol{w}, i, n) ,$$

а значения функции $G(\boldsymbol{w}, i, n)$, $\boldsymbol{w} \in \mathcal{W}$, $i \in \in \{1, \ldots, L(\boldsymbol{w})\}$, $n \in \omega^-$, находятся по рекуррентным формулам:

$$G(\boldsymbol{w}, i, n) = -\infty ,$$

для всех $\boldsymbol{w} \in \mathcal{W}$, $i \in \{1, \ldots, L(\boldsymbol{w})\}$, *если* $n \notin \omega^+ \cup \omega^1$;

$$G(\boldsymbol{w}, i, n) = \begin{cases} g(\boldsymbol{w}, 1, n), \quad \textit{если} \quad i = 1, n \in \omega^+ \setminus \omega^1; \\ -\infty, \quad \textit{если} \quad i = 2, \ldots, L(\boldsymbol{w}), n \in \omega^+ \setminus \omega^1; \\ g(\boldsymbol{w}, 1, n) + \max\{0, \max\limits_{\boldsymbol{v} \in \mathcal{W}} \max\limits_{j \in \gamma^-(n)} G(\boldsymbol{v}, L(\boldsymbol{v}), j)\}, \\ \qquad \textit{если} \quad i = 1, n \in \omega^+ \cap \omega^1; \\ g(\boldsymbol{w}, i, n) + \max\limits_{j \in \gamma^-(n)} G(\boldsymbol{w}, i-1, j), \\ \qquad \textit{если} \quad i = 2, \ldots, L(\boldsymbol{w}), n \in \omega^+ \cap \omega^1; \\ g(\boldsymbol{w}, 1, n) + \max\limits_{\boldsymbol{v} \in \mathcal{W}} \max\limits_{j \in \gamma^-(n)} G(\boldsymbol{v}, L(\boldsymbol{v}), j), \\ \qquad \textit{если} \quad i = 1, n \in \omega^+ \setminus \omega^1; \\ g(\boldsymbol{w}, i, n) + \max\limits_{j \in \gamma^-(n)} G(\boldsymbol{w}, i-1, j), \\ \qquad \textit{если} \quad i = 2, \ldots, L(\boldsymbol{w}), n \in \omega^+ \setminus \omega^1, \end{cases}$$

для всех $\boldsymbol{w} \in \mathcal{W}$, $i \in \{1, \ldots, L(w)\}$, *если* $n \in \omega^+ \cup \omega^1$.

Для вычисления оптимальных значений $\hat{M}$ и $\hat{J}$, а также значений компонент оптимальных наборов $(\hat{n}_1, \ldots, \hat{n}_{\hat{M}})$ и $(\hat{\boldsymbol{w}}_1, \ldots, \hat{\boldsymbol{w}}_{\hat{J}})$ определим три вспомогательных функции:

$$\boldsymbol{t}(n) = \arg\max_{\boldsymbol{w} \in \mathcal{W}} \Big\{ \max_{j \in \gamma^-(n)} G(\boldsymbol{w}, L(\boldsymbol{w}), j) \Big\}, \quad n \in \omega^1,$$

$$I(\boldsymbol{w}, i, n) = \begin{cases} \arg\max\limits_{j \in \gamma^-(n)} G(\boldsymbol{t}(n), L(\boldsymbol{t}(n)), j), \quad \text{если} \quad i = 1; \\ \arg\max\limits_{j \in \gamma^-(n)} G(\boldsymbol{w}, i-1, j), \\ \qquad \text{если} \quad i = 2, \ldots, L(\boldsymbol{w}), \end{cases}$$
$$\boldsymbol{w} \in \mathcal{W}, \quad i = 1, \ldots, L(\boldsymbol{w}), \quad n \in \omega^1,$$

$$r(\boldsymbol{w}, i, n) = \begin{cases} 1, \quad \text{если} \ \ n \in \omega^1 \setminus \omega^+, i = 1, \ldots, L(\boldsymbol{w}); \\ 1, \quad \text{если} \ \ n \in \omega^1 \cap \omega^+, i = 2, \ldots, L(\boldsymbol{w}); \\ 0, \quad \text{если} \ \ n \in \omega^1 \cap \omega^+, i = 1; \\ \qquad \max\limits_{\boldsymbol{v} \in \mathcal{W}} \max\limits_{j \in \gamma^-(n)} G(\boldsymbol{v}, L(\boldsymbol{v}), j)) < 0; \\ 1, \quad \text{если} \ \ n \in \omega^1 \cap \omega^+, i = 1; \\ \qquad \max\limits_{\boldsymbol{v} \in \mathcal{W}} \max\limits_{j \in \gamma^-(n)} G(\boldsymbol{v}, L(\boldsymbol{v}), j)) \geqslant 0; \\ 0, \quad \text{если} \ \ n \in \omega^1 \setminus \omega^+, i = 1, \end{cases}$$
$$\boldsymbol{w} \in \mathcal{W}, \quad i = 1, \ldots, L(\boldsymbol{w}), \quad n \in \omega^1 \cup \omega^+.$$

Кроме того, положим

$$\hat{\boldsymbol{v}}_1 = \arg\max_{\boldsymbol{w} \in \mathcal{W}} \left\{ \max_{i \in \{1, \ldots, L(\boldsymbol{w})\}} \max_{n \in \omega^-} G(\boldsymbol{w}, i, n) \right\};$$

$$\hat{k}_1 = \arg\max_{i \in \{1, \ldots, L(\hat{\boldsymbol{v}}_1)\}} \left\{ \max_{n \in \omega^-} G(\hat{\boldsymbol{v}}_1, i, n) \right\};$$

$$\hat{s}_1 = \arg\max_{n \in \omega^-} G(\hat{\boldsymbol{v}}_1, \hat{k}_1, n)$$

и

$$\hat{s}_m = I(\hat{\boldsymbol{v}}_{m-1}, \hat{k}_{m-1}, \hat{s}_{m-1});$$

$$\hat{\boldsymbol{v}}_m = \begin{cases} \hat{\boldsymbol{v}}_{m-1}, & \text{если} \quad \hat{k}_{m-1} > 1; \\ \boldsymbol{t}(\hat{s}_{m-1}), & \text{если} \quad \hat{k}_{m-1} = 1; \end{cases}$$

$$\hat{k}_m = \begin{cases} \hat{k}_{m-1} - 1, & \text{если} \quad \hat{k}_{m-1} > 1; \\ L(\hat{\boldsymbol{v}}_m), & \text{если} \quad \hat{k}_{m-1} = 1, \end{cases}$$

где $m = 2, \ldots, M^*$, а

$$M^* = \min\{m \in \mathbb{N} \,|\, r(\hat{\boldsymbol{v}}_m, \hat{k}_m, \hat{s}_m) = 0\}.$$

Следствие 1. *1) Компоненты оптимального набора* $(\hat{n}_1, \ldots, \hat{n}_{\hat{M}})$ *и его размерность* $\hat{M}$ *определяются по правилу:*

$$\hat{M} = M^*,$$

$$\hat{n}_m = s_{\hat{M}-m+1}, \qquad m = 1, \ldots, \hat{M};$$

2) элементы оптимального набора $(\hat{\boldsymbol{w}}_1, \ldots, \hat{\boldsymbol{w}}_{\hat{J}})$ *и его размерность* $\hat{J}$ *находятся по формулам:*

$$\hat{J} = \big|\{m \,|\, m \in \{1, \ldots, \hat{M}\}, \hat{k}_m = 1\}\big|,$$

$$\hat{\boldsymbol{w}}_1 = \hat{\boldsymbol{v}}_{\hat{M}}, \ \hat{\boldsymbol{w}}_j = \hat{\boldsymbol{v}}_{\hat{M} - L(\hat{\boldsymbol{w}}_1) - \ldots - L(\hat{\boldsymbol{w}}_{j-1})}, \ j = 2, \ldots, \hat{J}.$$

Изложим алгоритм решения задачи.

Алгоритм $\mathcal{A}$.

Шаг 1 (прямой ход алгоритма). Вычислим оптимальное значение целевой функции по формулам леммы 1.

Шаг 2 (обратный ход алгоритма). Найдем оптимальное решение по формулам следствия 1.

Справедлива следующая

Теорема 2. *Алгоритм* $\mathcal{A}$ *находит оптимальное решение задачи SIVTS за время* $\mathcal{O}(KN(T_{\max} - T_{\min} + 1)(L_{\max} + K))$.

Оптимальность решения следует из леммы 1 и следствия 1. Временная сложность алгоритма определяется рекуррентными формулами динамического программирования.

Замечание 1. В оценку временной сложности алгоритма входят числовые значения $(T_{\max} - T_{\min} + 1)$ и $L_{\max}$. Эти значения ограничены сверху размером N входа задачи. Поэтому в общем случае трудоемкость алгоритма есть величина $\mathcal{O}(KN^2(N + K))$, т. е. алгоритм полиномиален.

Напомним, что в [4] был обоснован алгоритм, гарантирующий оптимальное решение задачи SIVTS за время $\mathcal{O}(KN^4(N+K))$. Из теоремы 2 и замечания 1 следует, что новый алгоритм, предложенный в настоящей работе, имеет в N^2 раз меньшую трудоемкость.

Литература

[1] *Кельманов А. В., Окольнишникова Л. В.* Апостериорное совместное обнаружение и различение подпоследовательностей в квазипериодической последовательности // Сиб. журн. индустриальной математики. — 2000. — Т. 3, № 2(6). — С. 115–139.

[2] *Кельманов А. В., Михайлова Л. В., Хамидуллин С. А.* Оптимальное обнаружение в квазипериодической последовательности повторяющегося набора эталонных фрагментов // Сиб. журн. вычисл. математики. — 2008. — Т. 11, № 3. — С. 311–327.

[3] `http://math.nsc.ru/~serge/qpsl/` — Система QPSLab для решения задач компьютерного анализа и распознавания числовых последовательностей с квазипериодической структурой. — 2008.

[4] *Кельманов А. В., Михайлова Л. В., Хамидуллин С. А.* Об одной задаче обнаружения и идентификации векторных наборов в последовательности // Интеллектуализация обработки информации: 8-я международная конференция: Сборник докладов. — М.: МАКС Пресс, 2010. — С. 270–273.

2-приближенный алгоритм для одной задачи поиска в векторной последовательности совокупности «похожих» элементов*

Кельманов А. В., Романченко С. М., Хамидуллин С. А.
kelm@math.nsc.ru, semenr@bk.ru, kham@math.nsc.ru
Новосибирск, Институт математики им. С. Л. Соболева СО РАН, Новосибирский государственный университет

Анализируется одна из труднорешаемых проблем поиска в заданной последовательности векторов евклидова пространства набора элементов «близких» между собой по критерию минимума суммы квадратов расстояний. Предложен 2-приближенный эффективный алгоритм решения этой задачи.

Объект исследования работы — проблемы оптимизации в задачах анализа данных. Предмет исследования — дискретная экстремальная задача, к которой сводится одна из проблем помехоустойчивого обнаружения в последовательности, состоящей из результатов измерения набора информационно значимых числовых характеристик физических объектов, совокупности «похожих» между собой элементов (наборов). Цель работы — обоснование приближенного эффективного алгоритма решения этой задачи.

На содержательном уровне рассматриваемую проблему можно трактовать как разновидность так называемой проблемы «обучения» компьютера распознаванию образов. Мотивацией исследований послужил тот факт, что до настоящего времени отсутствовали какие-либо эффективные алгоритмы с гарантированными оценками точности для решения экстремальной задачи, к которой сводится содержательная проблема.

Модель анализа данных

Рассмотрим следующую структуру данных, представленных в виде совокупности векторов евклидова пространства.

Пусть векторная последовательность $\boldsymbol{x}_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, где $\mathcal{N} = \{1, \ldots, N\}$, обладает свойством

$$\boldsymbol{x}_n = \begin{cases} \boldsymbol{w}, & n \in \mathcal{M}; \\ \boldsymbol{v}_n, & n \in \mathcal{N}\backslash\mathcal{M}, \end{cases} \qquad (1)$$

где $\mathcal{M} = \{n_1, \ldots, n_M\} \subseteq \mathcal{N}$.

Допустим, что для обработки доступна последовательность

$$\boldsymbol{y}_n = \boldsymbol{x}_n + \boldsymbol{e}_n, \quad n \in \mathcal{N}, \qquad (2)$$

где $\boldsymbol{e}_n$ — вектор помехи (ошибки измерения), независимый от вектора $\boldsymbol{x}_n$. Учитывая зависимость элементов последовательности (1) от множеств и векторов, положим

$$S(\mathcal{M}, \boldsymbol{w}, \{\boldsymbol{v}_i, i \in \mathcal{N}\backslash\mathcal{M}\}) = \sum_{n \in \mathcal{N}} \|\boldsymbol{y}_n - \boldsymbol{x}_n\|^2 \qquad (3)$$

и рассмотрим модель анализа данных в виде следующей экстремальной задачи.

Дано: последовательность $\boldsymbol{y}_n$, $n \in \mathcal{N}$, векторов из $\mathbb{R}^q$ и натуральные числа $T_{\min}$, $T_{\max}$ и $M > 1$.

Найти: непустое подмножество $\mathcal{M} \subseteq \mathcal{N}$ номеров мощности M, вектор $\boldsymbol{w}$ и совокупность $\{\boldsymbol{v}_i, i \in \in \mathcal{N}\backslash\mathcal{M}\}$ векторов, минимизирующих $S(\cdot)$, при условии, что структура последовательности описывается формулами (1) и (2), при следующих ограничениях на элементы подмножества $\mathcal{M}$:

$$1 \leqslant T_{\min} \leqslant n_m - n_{m-1} \leqslant T_{\max} \leqslant N - 1, \quad m = 2, \ldots, M. \qquad (4)$$

В этой модели совокупность элементов последовательности $\boldsymbol{y}_n$, $n \in \mathcal{N}$, ассоциируется с набором наблюдаемых или имеющихся в распоряжении данных, в которых скрыта ненаблюдаемая последовательность $\boldsymbol{x}_n$, $n \in \mathcal{N}$. Эта последовательность в соответствии с (1) включает вектор w, повторяющийся M раз, который ассоциируется с информационно значимым набором характеристик одного или нескольких идентичных объектов. Остальные элементы — $\boldsymbol{v}_n$, $n \in \mathcal{N}\backslash\mathcal{M}$, — последовательности $\boldsymbol{x}_n$, $n \in \mathcal{N}$, трактуются как возможный «мусор», который, как правило, имеется в совокупности обрабатываемых данных.

Если номера членов последовательностей интерпретировать как равномерные дискретные отсчеты времени, то параметры $T_{\min}$ и $T_{\max}$ из модели соответствуют минимальному и максимальному интервалам времени между двумя последовательными повторами неизвестного информационно значимого набора. Эти параметры на практике часто бывают априори известны. Случай $T_{\min} = T_{\max}$ соответствует самой простой ситуации, когда повторы периодичны. В случае $T_{\min} \neq T_{\max}$ подслучай $T_{\min} = 1$ и $T_{\max} = N - 1$ соответствует другой крайней ситуации, когда объём информации о времени между двумя последовательными повторами минимален (энтропия максимальна).

Работа выполнена при финансовой поддержке РФФИ, проекты №09-01-00032, №10-07-00195; целевой программы №2 Президиума РАН, проект №227; целевой программы СО РАН, интеграционный проект №44; целевой программы АВЦП Рособразования, проект №2.1.1/3235; а также федеральной целевой программы «Научные и научно-педагогические кадры инновационной России», гос. контракт №14.740.11.0362.

Легко установить, что если $\boldsymbol{e}_n$ в формуле (2) есть выборка единичного объема из q-мерного нормального распределения с параметрами $(0, \sigma^2 I)$, где I — единичная матрица, а в качестве критерия решения задачи использовать максимум функционала правдоподобия, то статистический подход к рассматриваемой проблеме анализа данных приводит к задаче минимизации функционала (3). Статистические аспекты проблемы находятся вне рамок данной работы.

Редуцированная задача

Учитывая (1), для функционала (3) имеем

$$S = \sum_{n\in\mathcal{M}} \|\boldsymbol{y}_n - \boldsymbol{w}\|^2 + \sum_{n\in\mathcal{N}\setminus\mathcal{M}} \|\boldsymbol{y}_n - \boldsymbol{v_n}\|^2 .$$

Минимум этого функционала по неизвестным векторам $\boldsymbol{w}$ и $\boldsymbol{v}_n$, $n \in \mathcal{N}\setminus\mathcal{M}$, находится аналитически. Нетрудно убедиться, что для любого непустого подмножества $\mathcal{M} \subseteq \mathcal{N}$ этот минимум доставляется векторами $\bar{\boldsymbol{w}} = (1/|\mathcal{M}|)\sum_{n\in\mathcal{M}} \boldsymbol{y}_n$, $\bar{\boldsymbol{v}}_n = \boldsymbol{y}_n$, $n \in \mathcal{N}\setminus\mathcal{M}$, и равен

$$F(\mathcal{M}) = \sum_{n\in\mathcal{M}} \|\boldsymbol{y}_n - \bar{\boldsymbol{w}}\|^2 . \qquad (5)$$

Поэтому минимизация функционала (3) сводится к минимизации целевой функции (5). Отсюда получаем следующую задачу.

Задача VSS (searching Vector Subsequence in a Sequence).

Дано: набор $\mathcal{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)$ векторов из $\mathbb{R}^q$, натуральные числа $T_{\min}$, $T_{\max}$ и $M > 1$.

Найти: подмножество $\mathcal{M} = \{n_1, \ldots, n_M\} \subseteq \mathcal{N}$ номеров элементов набора $\mathcal{Y}$ такое, что целевая функция (5) минимальна, при ограничениях (4) на элементы искомого подмножества $\mathcal{M}$.

Замечание 1. Если значения $T_{\min}$ и $T_{\max}$ неизвестны (не заданы на входе задачи), то в ограничениях (4) полагаем $T_{\min} = 1$ и $T_{\max} = N - 1$.

Формула (4) задает ограничения на порядок выбора векторов из набора (последовательности) $\mathcal{Y}$. Если эти ограничения отсутствуют или в этих ограничениях $T_{\min} = 1$ и $T_{\max} = N - 1$, то задача VSS совпадает с NP-трудной задачей VS-2 (Vector Subset 2) [1]. В [2] для решения задачи VS-2 предложен эффективный 2-приближенный алгоритм, имеющий временную сложность $\mathcal{O}(qN^2)$. Его, очевидно, можно без изменений применить для приближенного решения указанного подслучая задачи VSS. Ясно также, что в подслучае, когда $T_{\min} = T_{\max}$, задача VSS разрешима за полиномиальное время.

В общем случае, когда $T_{\min} \neq T_{\max}$ или $T_{\min} \neq 1$ и $T_{\max} \neq N-1$, задача VSS интуитивно кажется сложнее задачи VS-2. Скорее всего, в этом общем случае задача VSS относится к числу труднорешаемых. Однако гипотеза о её труднорешаемости пока не доказана. Тем не менее, ниже предложен приближенный эффективный алгоритм решения задачи VSS. Суть подхода состоит в замене решения задачи VSS решением более простой вспомогательной задачи и последующей оценкой точности этой замены.

Вспомогательная задача

Рассмотрим следующую задачу.

Задача VSVS (searching Vector Subsequence and Vector in a Sequence).

Дано: набор $\mathcal{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)$ векторов из $\mathbb{R}^q$, натуральные числа $T_{\min}$, $T_{\max}$ и $M > 1$.

Найти: подмножество $\mathcal{M} = \{n_1, \ldots, n_M\} \subseteq \mathcal{N}$ номеров элементов набора $\mathcal{Y}$ и вектор $\boldsymbol{b} \in \mathcal{Y}$ такие, что целевая функция

$$G(\mathcal{M}, \boldsymbol{b}) = \sum_{n\in\mathcal{M}} \|\boldsymbol{y}_n - \boldsymbol{b}\|^2$$

минимальна, при ограничениях (4) на элементы подмножества $\mathcal{M}$.

Положим

$$g(n, \boldsymbol{b}) = \|\boldsymbol{y}_n - \boldsymbol{b}\|^2,\ n \in \mathcal{N},\ \boldsymbol{b} \in \mathcal{Y} .$$

Тогда

$$G(\mathcal{M}, \boldsymbol{b}) = \sum_{n\in\mathcal{M}} g(n, \boldsymbol{b}) .$$

Легко заметить, что

$$\min_{\mathcal{M},\boldsymbol{b}} G(\mathcal{M}, \boldsymbol{b}) = \min_{\boldsymbol{b}} \min_{\mathcal{M}} G(\mathcal{M} \,|\, \boldsymbol{b}) . \qquad (6)$$

Поэтому решение задачи VSVS можно найти в 2 этапа. На первом этапе для каждого фиксированного $\boldsymbol{b} \in \mathcal{Y}$ находим

$$G_{\min}(\boldsymbol{b}) = \min_{\mathcal{M}} G(\mathcal{M} \,|\, \boldsymbol{b}) ; \qquad (7)$$

$$\mathcal{M}(\boldsymbol{b}) = \arg\min_{\mathcal{M}} G(\mathcal{M} \,|\, \boldsymbol{b}) . \qquad (8)$$

На втором этапе находим оптимальные вектор, подмножество и значение целевой функции по следующему правилу:

$$G_{\min} = \min_{\boldsymbol{b}} G_{\min}(\boldsymbol{b}) ; \qquad (9)$$

$$\hat{\boldsymbol{b}} = \arg\min_{\boldsymbol{b}} G_{\min}(\boldsymbol{b}), \quad \widehat{\mathcal{M}} = \mathcal{M}(\hat{\boldsymbol{b}}) . \qquad (10)$$

В приведеном ниже алгоритме вычисление значения $G_{\min}(\boldsymbol{b})$ условного минимума при каждом фиксированном $\boldsymbol{b}$ реализуется схемой динамического программирования, обоснованной в [3].

Алгоритм $\mathcal{A}_1$.

Шаг 1. Положим $i = 0$; $G_{\min} = \infty$.

Шаг 2. $i := i + 1$; положим $\boldsymbol{b} = \boldsymbol{y}_i$.

Шаг 3. Используя рекуррентные формулы, для каждого фиксированного $\boldsymbol{b}$ вычислим значения

$$G_m(n, \boldsymbol{b}) = \begin{cases} g(n \,|\, \boldsymbol{b}), & \text{если } n \in \omega_1,\ m = 1; \\ g(n \,|\, \boldsymbol{b}) + \min\limits_{j \in \gamma^-_{m-1}(n)} G_{m-1}(j \,|\, \boldsymbol{b}), \\ \qquad \text{если } n \in \omega_m,\ m = 2, \ldots, M, \end{cases}$$

где

$$\omega_m = \big\{n \,|\, 1 + (m-1)T_{\min} \leqslant n \leqslant N - (M-m)T_{\min}\big\}$$

— область допустимых значений переменной n_m, а

$$\gamma^-_{m-1}(n) = \big\{j \mid \max\{1 + (m-2)T_{\min}, n - T_{\max}\} \leqslant j \leqslant n - T_{\min}\big\}, \quad n \in \omega_m,$$

— область допустимых значений переменной n_{m-1} при условии, что значение переменной n_m фиксировано и равно n.

Шаг 4. Найдем значение условного минимума целевой функции G по формуле

$$G_{\min}(\boldsymbol{b}) = \min_{n \in \omega_M} G_M(n \,|\, \boldsymbol{b}),$$

где $G_M(n \,|\, \boldsymbol{b})$ — семейство значений, вычисленных на шаге 3.

Шаг 5. Вычислим значения компонент условно-оптимального (при фиксированном $\boldsymbol{b}$) набора $\{\hat{n}_1(\boldsymbol{b}), \ldots, \hat{n}_M(\boldsymbol{b})\}$ по следующим формулам:

$$\hat{n}_M(\boldsymbol{b}) = \arg\min_{n \in \omega_M} G_M(n \,|\, \boldsymbol{b}),$$
$$\hat{n}_{m-1}(\boldsymbol{b}) = I_m(\hat{n}_m \,|\, \boldsymbol{b}),\ m = M, M-1, \ldots, 2,$$

где

$$I_m(n, \boldsymbol{b}) = \arg\min_{j \in \gamma^-_{m-1}(n)} G_{m-1}(j \,|\, \boldsymbol{b}), \quad n \in \omega_m.$$

Шаг 6. Если $G_{\min} > G_{\min}(\boldsymbol{b})$, то положим $\hat{\boldsymbol{b}} = \boldsymbol{b}$, $\hat{n}_m = \hat{n}_m(\boldsymbol{b})$, $m = 1, \ldots, M$; $G_{\min} := G_{\min}(\boldsymbol{b})$, иначе переходим на Шаг 7.

Шаг 7. Если $i < N$ переходим на Шаг 2, иначе — выход.

Результатом работы алгоритма объявляем вектор $\hat{\boldsymbol{b}}$, набор $\widehat{\mathcal{M}} = \{\hat{n}_1, \ldots, \hat{n}_M\}$ и значение $G_{\min}$.

Замечание 2. Если значения $T_{\min}$ и $T_{\max}$ неизвестны, то в формулах для вычислений на шаге 3 полагаем $T_{\min} = 1$ и $T_{\max} = N - 1$ в соответсвии с замечанием 1.

Лемма 1. *Алгоритм $\mathcal{A}_1$ находит оптимальное решение задачи VSVS за время $\mathcal{O}(N(MN(T_{\max} - T_{\min} + 1) + q)$.*

Оптимальность решения следует из формул (6)–(10) и формул шагов 3–5 алгоритма, гарантирующих оптимальность [3] процедуры отыскания условного минимума. Сложность алгоритма определяется формулами динамического программирования.

Замечание 3. В оценку временной сложности алгоритма $\mathcal{A}_1$ входят числовые значения M и $(T_{\max} - T_{\min} + 1)$. Эти значения ограничены сверху размером N входа задачи. Поэтому в общем случае трудоемкость алгоритма есть величина $\mathcal{O}(N(N^3 + + q))$, т. е. алгоритм полиномиален.

Приближенный алгоритм

Изложим алгоритм решения задачи VSS.

Алгоритм $\mathcal{A}$.

Шаг 1. По заданной последовательности $\mathcal{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)$ векторов из $\mathbb{R}^q$, натуральным числам $T_{\min}$, $T_{\max}$ и $M > 1$ находим оптимальное решение — $\hat{\boldsymbol{b}}$ и $\widehat{\mathcal{M}} = \{\hat{n}_1, \ldots, \hat{n}_M\}$ — вспомогательной задачи VSVS с помощью алгоритма $\mathcal{A}_1$.

Шаг 2. Приближенным решением задачи объявляем набор $\widehat{\mathcal{M}}$.

Оценку вектора $\boldsymbol{w}$ в модели (1) можно найти по формуле $\widehat{\boldsymbol{w}} = (1/|\widehat{\mathcal{M}}|) \sum_{n \in \widehat{\mathcal{M}}} \boldsymbol{y}_n$ в соответствии с (5).

Теорема 2. *Алгоритм $\mathcal{A}$ находит 2-приближенное решение задачи VSS за время $\mathcal{O}(N(N^3 + q))$. Оценка 2 точности алгоритма достижима и неулучшаема.*

Идея и техника доказательства теоремы 2 базируется на результатах работы [2].

Заключение

В работе обоснован 2-приближенный эффективный алгоритм для решения задачи, к которой сводится оптимизационная модель одной из актуальных из проблем анализа данных.

Поскольку рассмотренная задача относится к числу практически неизученных в алгоритмическом плане, исследование вопросов её аппроксимируемости, а также обоснование алгоритмов другого типа (асимптотически точных, рандомизированных и др.) для её решения представляется делом ближайшей перспективы.

Литература

[1] *Кельманов А. В., Пяткин А. В.* NP-полнота некоторых задач выбора подмножества векторов // Дискретный анализ и исследование операций. — 2010. — Т. 17, № 5. — С. 37–45.

[2] *Кельманов А. В., Романченко С. М.* Приближённый алгоритм для решения одной задачи поиска подмножества векторов // Дискретный анализ и исследование операций. — 2011. — Т. 18, № 1. — С. 61–69.

[3] *Кельманов А. В., Хамидуллин С. А.* Апостериорное обнаружение заданного числа одинаковых подпоследовательностей в квазипериодической последовательности // Журнал вычислительной математики и математической физики. — 2001. — Т. 41, № 5. — С. 807–820.

Аппроксимационная схема для одной задачи поиска подмножества векторов*

Шенмайер В. В.
shenmaier@mail.ru
Новосибирск, Институт математики им. С. Л. Соболева СО РАН

Рассматривается следующая задача кластеризации: среди заданного множества векторов найти подмножество мощности k, обладающее минимальным квадратичным отклонением от своего среднего. Расстояния между векторами определяются евклидовой метрикой. Предлагается аппроксимационная схема (PTAS), позволяющая решать данную задачу с произвольной относительной погрешностью ε за время $O(n/\varepsilon)^{O(1/\varepsilon)}$.

Рассматривается следующая задача. Пусть X — конечное множество векторов в евклидовом пространстве и k — некоторое число, $1 \leqslant k \leqslant n$, где $n = |X|$. Требуется найти такое подмножество (кластер) $K \subseteq X$ мощности k, что сумма квадратов расстояний от векторов множества K до вектора $\overline{c}(K) = \sum_{x \in K} x/k$ (центра кластера), минимальна:

$$\min_K \sum_{x \in K} \|x - \overline{c}(K)\|^2 ,$$

$$K \subseteq X, \quad |K| = k .$$

Содержательный смысл задачи может быть следующим. Имеется множество результатов измерений характеристик некоторых объектов. Каждый результат измерения представляет из себя многомерный вещественный вектор. Измерения объектов имеют ошибку, и соответствие между объектами и результатами измерений неизвестно. Но при этом известно, что k результатов измерений соответствуют одному и тому же объекту. Требуется, используя критерий минимума суммы квадратов расстояний, найти подмножество результатов измерений, вероятнее всего соответствующих данному неизвестному объекту, и оценить набор его характеристик.

В работе [1] установлено, что задача NP-трудна в сильном смысле. Отсюда следует, что при условии P $\neq$ NP не существует ни полиномиального алгоритма для ее решения, ни псевдополиномиального алгоритма, ни полностью полиномиальной аппроксимационной схемы (FPTAS). В работе [2] предложен полиномиальный приближенный алгоритм, имеющий оценку точности 2. В работе [3] предложен псевдополиномиальный алгоритм для частного случая, когда размерность пространства фиксирована (трудоемкость данного алгоритма экспоненциально зависит от размерности пространства).

Предлагается приближенный алгоритм, имеющий относительную погрешность $1/t + 8\sqrt{t-1}/s$ и трудоемкось $O(n^{t+2}s^{t-1})$, где t и s — произвольные целочисленные параметры ($t, s \geqslant 1$). В частности, при выборе $t = 2/\varepsilon$, где $\varepsilon > 0$, и $s = 8t^{3/2}$ имеем аппроксимационную схему (PTAS), позволяющую решать задачу с относительной погрешностью ε за время $O(n^{2/\varepsilon+2}(8/\varepsilon)^{3/\varepsilon})$.

Работа выполнена при финансовой поддержке РФФИ, проекты № 09-01-00032, № 10-07-00195; целевой программы № 2 Президиума РАН, проект № 227; а также целевой программы СО РАН, интеграционный проект № 44.

Идея алгоритма

Заметим, что если центр оптимального кластера известен, то сам кластер восстанавливается с помощью выбора k ближайших к центру векторов множества X. Идея алгоритма состоит в поиске данного центра среди векторов линейных оболочек всех наборов векторов множества X мощности t. При этом для дискретизации алгоритма на каждой из этих линейных оболочек рассматривается $(t-1)$-мерная сетка с шагом h, где h — константа, вычисляемая на предварительном шаге алгоритма.

Пусть $f(y, K) = \sum_{x \in K} \|x - y\|^2$ — значение целевой функции в точке y на кластере K. Легко показать, что при фиксированном кластере K минимальное значение f достигается в точке $\overline{c}(K)$, при этом

$$f(y, K) = OPT + k\,\|y - \overline{c}(K)\|^2 , \qquad (1)$$

где $OPT = f(\overline{c}(K), K)$. Обозначим через $\varepsilon(y, K)$ относительную погрешность целевой функции в точке y:

$$\varepsilon(y, K) = \frac{f(y, K) - OPT}{OPT} .$$

Геометрическое обоснование алгоритма дает следующая

Теорема 1. *Пусть K — произвольное множество векторов и $1 \leqslant t \leqslant |K|$. Тогда линейная оболочка одного из подмножеств K мощности t содержит вектор y_t такой, что $\varepsilon(y_t, K) \leqslant 1/t$.*

Вывод ее основывается на следующем факте:

Лемма 2. *Пусть x — произвольный вектор евклидового пространства и $y = y(x, K)$ — ближайший к вектору $\overline{c}(K)$ вектор, лежащий на лучах, проведенных из x во все векторы кластера K. Тогда*

$$\varepsilon(y, K) \leqslant \frac{\varepsilon(x, K)}{1 + \varepsilon(x, K)} .$$

Доказательство леммы. Не нарушая общности, будем считать, что вектор $\overline{c}(K)$ совпадает с началом координат O, $O = (0, \ldots, 0)$. Пользуясь евклидовостью рассматриваемого пространства, можно также считать, что координатные оси повернуты таким образом, что вектор x лежит на первой из них: $x = (x(1), 0, \ldots, 0)$, для определенности $x(1) \geqslant 0$. Аналогично можно считать, что вектор y находится в плоскости, образованной первой и второй координатными осями.

Поскольку вектор O — центр кластера K, то среди векторов кластера найдутся такие, что лежат в полупространстве $X(1) \leqslant 0$. Следовательно, луч xy пересекает гиперплоскость $X(1) = 0$ в некоторой точке $A = (0, a, 0, \ldots, 0)$. При этом если одно из чисел $x(1)$ или a равно нулю, то вектор y совпадает с оптимальным решением и, следовательно, утверждение леммы очевидно. Поэтому, не нарушая общности, будем предполагать, что $a > 0$ и $x(1) > 0$.

Заметим, что согласно выбору луча xy все векторы кластера K находятся вне конуса C, образованного вектором x и шаром радиуса a, находящимся в гиперплоскости $X(1) = 0$, и с центром в начале координат.

Согласно неравенству (1) имеем

$$\varepsilon(x, K) = \frac{\|x\|^2\, k}{OPT} = \frac{x(1)^2\, k}{OPT}\,, \tag{2}$$

где $OPT = f(O, K)$. Из геометрических соображений

$$\varepsilon(y, K) = \frac{\|y\|^2 k}{OPT} = \frac{a^2\, x(1)^2}{a^2 + x(1)^2}\,\frac{k}{OPT}\,. \tag{3}$$

Оценим величину $f(O, K)$. Применим для этого технику упрощений, заключающуюся в переходе к более простой геометрической ситуации с сохранением оцениваемой величины. Заметим, что $f(O, K) = f(O', K')$, где O' — двумерный нулевой вектор, а множество K' состоит из двумерных векторов вида $u' = (u(1), \sqrt{\|u\|^2 - u(1)^2})$, где $u \in K$. Действительно, каждый вектор u' совпадает по первой координате с исходным вектором u, а вторая координата u' равна расстоянию от u до первой оси координат. Таким образом, $\|u'\| = \|u\|$.

Далее в силу того, что среднее квадратов не меньше квадрата среднего, имеем $f(O', K') \geqslant k\,\|\overline{u'}\|^2$, где $\overline{u'} = \sum_{u' \in K'} u'/k$. Но поскольку векторы исходного кластера K лежат вне конуса C, то векторы множества K' лежат над прямой, соединяющей двумерные векторы $x' = (x(1), 0)$ и $A' = (0, a)$. Следовательно, средний вектор $\overline{u'}$ также лежит над данной прямой. При этом, поскольку $\overline{u'}(1) = 0$, получаем, что $\|\overline{u'}\| \geqslant a$. Отсюда

$$f(O, K) \geqslant k\, a^2\,. \tag{4}$$

Подставим оценку (4) в равенство (3):

$$\varepsilon(y, K) \leqslant \frac{x(1)^2}{a^2 + x(1)^2}\,. \tag{5}$$

С другой стороны, объединяя равенства (2) и (3), получим

$$\varepsilon(y, K) = \frac{a^2\, \varepsilon(x, K)}{a^2 + x(1)^2}\,. \tag{6}$$

Рассмотрим выражения (5) и (6) как функции от аргумента a. Первая из них монотонно убывает от 1 до 0, вторая монотонно возрастает от 0 до $\varepsilon(x, K)$. Следовательно, минимум из этих двух функций достигает максимума в точке их пересечения, определяемой соотношением $x(1)^2 = a^2\, \varepsilon(x, K)$. Таким образом,

$$\varepsilon(y, K) \leqslant \frac{a^2\, \varepsilon(x, K)}{a^2 + a^2\, \varepsilon(x, K)} = \frac{\varepsilon(x, K)}{1 + \varepsilon(x, K)}\,.$$

Лемма 2 доказана. □

Доказательство теоремы 1 проводится индукцией по t.

База индукции: $t = 1$. Поскольку вектор y_1 — ближайший к центру из всех векторов кластера K, то $f(\overline{c}(K), K) \geqslant k\,\|y_1 - \overline{c}(K)\|^2$. Следовательно, $f(y_1, K) = f(\overline{c}(K), K) + k\,\|y_1 - \overline{c}(K)\|^2 \leqslant 2 f(\overline{c}(K), K)$. Таким образом, $\varepsilon(y_1, K) \leqslant 1$.

Индуктивный переход. Рассмотрим в качестве вектора y_{t+1} вектор $y(y_t, K)$ (см. лемму 2). Согласно данной лемме

$$\varepsilon(y_{t+1}, K) \leqslant \frac{\varepsilon(y_t, K)}{1 + \varepsilon(y_t, K)}\,.$$

По индукции $y_t \leqslant 1/t$, следовательно

$$\varepsilon(y_{t+1}, K) = \frac{1}{1/\varepsilon(y_t, K) + 1} \leqslant \frac{1}{t+1}\,.$$

Теорема доказана. □

Замечание 1. Доказанная оценка относительной погрешности является достижимой при любых t. Для того чтобы убедиться в этом, достаточно рассмотреть в качестве множества K набор из k единичных орт пространства $\mathbb{R}^k$ и устремить величину k к бесконечности.

Теорема 1 гарантирует, что если в качестве центра искомого кластера рассмотреть векторы линейных оболочек всех t-векторников из X, то один из них приведет к решению с относительной погрешностью $1/t$. При этом, если $t \geqslant k$, то данное свойство также справедливо, поскольку центр оптимального кластера K принадлежит линейной оболочке всех k векторов кластера.

Дискретизация алгоритма

Заметим, что если K — оптимальный кластер, а векторы $y_1, \ldots, y_t$ последовательно построены с помощью леммы 2, то из геометрических соображений $\|y_t - \overline{c}(K)\| \leqslant \|y_1 - \overline{c}(K)\| \leqslant A$, где $A = \sqrt{f(y_1, K)/k}$. Следовательно, для нахождения вектора y_t достаточно рассматривать окрестности радиуса $2A$ первых векторов рассматриваемых t-векторников.

Для каждого выбранного набора векторов $x_1, \ldots, x_t$ из X рассмотрим $(t-1)$-мерную сетку с шагом $h > 0$ на линейной оболочке данного набора. В качестве базиса сетки возьмем ортонормированный базис, получаемый по правилам линейной алгебры из векторов $x_2 - x_1, \ldots, x_t - x_1$. Поскольку евклидов шар радиуса $2A$ покрывают не более чем s^{t-1} элементов сетки, где $s = 4A/h$, то вектор y_t попадет в одну из s^{t-1} ее ячеек. Пусть y'_t — ближайший к y_t узел рассматриваемой сетки.

Лемма 3. $\varepsilon(y'_t, K) \leqslant \varepsilon(y_t, K) + 8\sqrt{t-1}/s$.

Доказательство. Расстояние от вектора y_t до y'_t в $(t-1)$-мерном евклидовом пространстве не превосходит величины $v = \sqrt{t-1}h/2$. Следовательно, $f(y'_t, K) - f(y_t, K) \leqslant k\left((a+v)^2 - a^2\right)$, где a — расстояние от вектора y_t до центра кластера K. В первом приближении разность $(a+v)^2 - a^2$ не превосходит величины $2av = \sqrt{t-1}ah$. Но поскольку $a \leqslant A$, имеем

$$f(y'_t, K) - f(y_t, K) \leqslant k\sqrt{t-1}ah \leqslant 4k\sqrt{t-1}A^2/s = \\ = 4\sqrt{t-1}f(y_1, K)/s \leqslant 8f(\overline{c}(K), K)\sqrt{t-1}/s\,.$$

Поделив данное выражение на $f(\overline{c}(K), K)$, получим требуемое неравенство. □

Теорема 1 и лемма 3 гарантируют, что алгоритм, заключающийся в переборе $n^t s^{t-1}$ кандидатов на роль центра искомого кластера, приведет к решению, имеющему относительную погрешность $1/t + 8\sqrt{t-1}/s$. Поскольку восстановление кластера по его центру занимает не более n^2 действий, трудоемкость данного алгоритма оценивается величиной $O(n^{t+2}s^{t-1})$.

Следствие 1. *При $t = 1$ алгоритм имеет относительную погрешность* 1 (*или, другими словами, относительную точность* 2) *и трудоемкость $O(n^3)$, что совпадает с результатом, полученным в работе* [2].

Следствие 2. *При $t = 2$ алгоритм имеет относительную погрешность $1/2 + \varepsilon$ и трудоемкость $O(n^4/\varepsilon)$.*

Следствие 3. *При $t = 2/\varepsilon$, где $\varepsilon > 0$, и $s = 8t^{3/2}$ алгоритм позволяет решать задачу с относительной погрешностью ε за время $O(n^{2/\varepsilon+2}(8/\varepsilon)^{3/\varepsilon})$. Таким образом, получена аппроксимационная схема (PTAS).*

Литература

[1] *Кельманов А. В., Пяткин А. В.* NP-полнота некоторых задач выбора подмножества векторов // Дискрет. анализ и исслед. опер. — 2010. — Т. 17, №5. — С. 37—45.

[2] *Кельманов А. В., Романченко С. М.* Приближённый алгоритм решения одной задачи поиска подмножества векторов // Дискретн. анализ и исслед. опер. — 2011. — Т. 18, №1. — С. 61–69.

[3] *Кельманов А. В., Романченко С. М.* Алгоритмы с оценками для некоторых задач поиска подмножества векторов и кластерного анализа. Настоящий сборник.

Построение и исследование полиномиальных алгоритмов для задач логического анализа данных в распознавании*

Дюкова Е. В., Колесниченко А. С.
edjukova@mail.ru, whestt@gmail.com
Москва, Вычислительный центр РАН, МГУ им. М В. Ломоносова

Доклад посвящен вопросам вычислительной сложности дискретных перечислительных задач, возникающих при логическом анализе данных в распознавании. Рассматриваемые задачи решаются на этапе поиска информативных фрагментов в признаковых описаниях объектов и могут быть сформулированы как задачи преобразования нормальных форм логических функций. Для перечисления максимальных конъюнкций булевой функции F от n переменных, заданной 2-КНФ с m элементарными дизъюнкциями, построен и экспериментально исследован алгоритм с полиномиальной задержкой $O(qm^2)$, где $q = \min(m, n)$. Ранее для важного частного случая, а именно, когда F- монотонная булева функция, в [6] был построен алгоритм с задержкой $O(n^3)$.

Введение

Конструирование логических процедур распознавания по прецедентам основано на построении дизъюнктивных нормальных форм логических функций. Центральной задачей является дуализация. Существует несколько вариантов формулировки этой задачи. Приведем их.

1) Дана конъюнктивная нормальная форма (КНФ) из m элементарных дизъюнкций, реализующая монотонную булеву функцию F от n переменных. Требуется построить (перечислить) все максимальные конъюнкции функции F.
2) Дана булева матрица L размера $m \times n$. Требуется найти все неприводимые покрытия матрицы L.
3) Дан гиперграф G с n вершинами и m рёбрами. Требуется найти все минимальные вершинные покрытия гиперграфа G.

Покажем связь задач 1) и 2).

Определение 1. *Набор столбцов H матрицы L называется покрытием, если каждая строка матрицы L в пересечении хотя бы с одним из столбцов, входящих в H , даёт 1. Покрытие называется неприводимым, если никакое его собственное подмножество не является покрытием.*

Пусть задана монотонная КНФ от n переменных $\mathfrak{K} = D_1 \& D_2 \& \ldots \& D_m$, где $D_i, i \in \{1, \ldots, m\}$, — элементарная дизъюнкция, не содержащая отрицаний переменных. Сопоставим КНФ $\mathfrak{K}$ булеву матрицу $L_{\mathfrak{K}} = (a_{ij})_{m \times n}$, где $a_{ij} = 1$, если $x_j \in D_i$ и $a_{ij} = 0$ в противном случае. Имеет место

Утверждение 1. *Конъюнкция $x_{j_1}, \ldots, x_{j_r}$ является максимальной для $\mathfrak{K}$ тогда и только тогда, когда набор столбцов матрицы $L_{\mathfrak{K}}$ с номерами $j_1, \ldots, j_r$ является неприводимым покрытием.*

Число искомых решений для сформулированных выше задач 1)–3) экспоненциально растет с ростом размера задачи. Эффективность алгоритмов для таких задач принято оценивать сложностью шага [6]. Рассмотрим два основных подхода к оценке сложности шага алгоритма, перечисляющего неприводимые покрытия булевой матрицы L.

Через $P(L)$ обозначим множество всех неприводимых покрытий матрицы L.

Подход 1. Будем говорить, что алгоритм строит $P(L)$ с *(квази)полиномиальной задержкой*, если на каждом шаге строится в точности один набор из $P(L)$ и при этом выполняется не более d элементарных операций, где d ограничено сверху (квази)полиномом от m, n. Под элементарной операцией понимается просмотр одного элемента матрицы L.

Алгоритм с квазиполиномиальной задержкой до сих пор не построен, и неизвестно, существует ли он.

Подход 2. Будем говорить, что алгоритм строящий $P(L)$, является *инкрементально (квази)полиномиальным*, если если на каждом шаге строится в точности один набор из $P(L)$ и при этом выполняется не более d элементарных операций, где d ограничено сверху (квази)полиномом от m, n и числа неприводимых покрытий, найденных на предыдущих шагах.

В [5] построен инкрементально квазиполиномиальный алгоритм (ИКП-алгоритм) для дуализации. В [4] проведено экспериментальное исследование этого алгоритма на случайных данных.

В [1, 2] предложен ещё один подход к оценке эффективности алгоритмов для перечислительных задач, основанный на понятии *асимптотически оптимального* алгоритма с полиномиальной задержкой. В отличие от «точного» алгоритма с полиномиальной задержкой (см. подход 1) асимптотически оптимальному алгоритму разрешено делать лишние полиномиальные шаги, но число таких шагов для почти всех матриц данного размера должно иметь более низкий порядок роста,

Работа выполнена при финансовой поддержке РФФИ, проект №10-01-00770 и гранта президента РФ по поддержке ведущих научных школ НШ №7950.2010.1

чем число всех шагов алгоритма с ростом размера матрицы.

К настоящему моменту построен ряд асимптотически оптимальных алгоритмов поиска неприводимых покрытий булевой матрицы (в основном для случая, когда матрица вытянута по горизонтали). В этих алгоритмах встречаются лишние шаги двух типов. Первый тип лишнего шага — построение набора столбцов, не являющегося неприводимым покрытием, второй тип — построение набора столбцов, который уже был построен ранее. Среди построенных асимптотически оптимальных алгоритмов есть алгоритмы с лишними шагами только одного из указанных типов и есть алгоритмы, в которых присутствуют лишние шаги обоих типов. Обзор результатов, полученных в области синтеза асимптотически оптимальных методов логического анализа данных, можно найти в [3].

В данной работе на примере задачи 2) проведено экспериментальное сравнение ИКП-алгоритма из [5] с асимптотически оптимальным алгоритмом ОПТ, описанным в [3]. Варьировались параметры m, n, и максимальное число единиц k в строке матрицы L. Результаты счёта показали, что ИКП-алгоритм очень чувствителен к росту числа столбцов в матрице и работает быстро лишь на матрицах с небольшими значениями параметра k при условии $m \geqslant n$. Алгоритм ОПТ ведёт себя полностью противоположным образом и работает значительно быстрее во всех случаях, кроме одного, а именно, когда $k = 2$ и $m \geqslant n$.

В [6] построен алгоритм с полиномиальной задержкой $O(n^3)$ для задачи нахождения максимальных независимых множеств графа G, которая двойственна к задаче 3). Для задачи 1) это случай, когда F — монотонная булева функция, заданная 2-КНФ. Для задачи 2) это случай, когда матрица L имеет не более двух единичных элементов в каждой строке. Алгоритм из [6] так же, как и ИКП-алгоритм Гурвича и Хачияна, очень чувствителен к росту параметра n.

Основной целью данной работы является построение алгоритма с полиномиальной задержкой в меньшей степени зависящей от роста n для частного случая задачи дуализации, рассмотренного в [6]. Доказана

Теорема 1. *Задача 2) в случае, когда матрица L имеет не более двух единичных элементов в каждой строке, может быть решена с полиномиальной задержкой $O(qm^2)$, где $q = \min(m, n)$.*

Доказательство теоремы 1 основано на построении алгоритма с указанной задержкой (алгоритма PL2). В основе алгоритма PL2 лежит идея построения с полиномиальной задержкой перестановочных подматриц матрицы L, порождающих неприводимые покрытия. Ранее эта идея использовалась в одном из вариантов асимптотически оптимального алгоритма с повторяющимися шагами (в алгоритме АО2 [1]). Незначительные дополнительные построения позволяют в условиях теоремы 1 полностью избавиться от повторно построенных неприводимых покрытий.

Алгоритм PL2 модифицирован для поиска максимальных конъюнкций булевой функции, заданной 2-КНФ. Формулировка этого результата приведена ниже в теореме 2.

Основные понятия. Схема работы алгоритма АО2

Пусть $L = (a_{ij}), i = i, \ldots, m, j = 1, \ldots, n$ – булева матрица. Элементу a_{ij} матрицы L присвоим номер $N[i, j] = (j - 1)m + 1$.

Через $P(L)$ обозначим множество всех неприводимых покрытий матрицы L.

Определение 2. *Квадратную подматрицу Q матрицы L назовем перестановочной, если в каждой строке и в каждом столбце подматрицы Q в точности один элемент равен 1.*

Определение 3. *Перестановочную подматрицу назовем максимальной, если она не содержится ни в каких других перестановочных подматрицах.*

Перестановочная подматрица полностью определяется набором ее единичных элементов.

Пусть перестановочная подматрица Q имеет порядок r и пусть $\{a_{i_1 j_1}, \ldots, a_{i_r j_r}\}$ — набор её единичных элементов. Будем говорить, что набор $\{a_{i_1 j_1}, \ldots, a_{i_r j_r}\}$ образует (порождает) подматрицу Q. Будем пользоваться записью $Q = \{a_{i_1 j_1}, \ldots, a_{i_r j_r}\}$. Будем считать, что

$$N[i_{u+1}, j_{u+1}] > N[i_u, j_u]$$

при $u = 1, \ldots, r - 1$. Через $H(Q)$ будем обозначать набор столбцов матрицы L с номерами $j_1, \ldots, j_r$.

Далее, через $L_1(Q)$ обозначим подматрицу матрицы L, образованную столбцами с номерами не меньшими j_1. Через $L_u(Q)$, $u \in \{2, \ldots, r + 1\}$, обозначим подматрицу матрицы L, образованную столбцами с номерами большими j_{u-1} и имеющими 0 в пересечении со строками с номерами $i_1, \ldots, i_{u-1}$, а также строками, имеющими 0 в пересечении со столбцами с номерами $j_1, \ldots, j_{u-1}$.

Пусть a_{ij} — единичный элемент. Справедливо

Утверждение 2. *Набор $\{a_{i_1 j_1}, \ldots, a_{i_{u-1} j_{u-1}}, a_{ij}\}$, $u \in \{2, 3, \ldots, r + 1\}$, порождает перестановочную подматрицу Q тогда и только тогда, когда $a_{ij} \in L_u(Q)$.*

Определение 4. *Будем говорить, что строка с номером p_1 охватывает строку с номером p_2 матрицы $L_u(Q)$, $u \in \{1, 2, \ldots, r\}$, если $a_{p_1 j} \geqslant a_{p_2 j}$ при всех j таких, что столбец с номером j принадлежит матрице $L_u(Q)$.*

Определение 5. *Единичный элемент $a_{p_1 j}$ матрицы $L_u(Q)$, $u \in \{1,2,\ldots,r{+}1\}$, назовем запрещенным, если строка с номером p_1 охватывает строку с номером p_2 матрицы $L_u(Q)$ и $a_{p_2 j} = 0$.*

Определение 6. *Перестановочную подматрицу $Q = \{a_{i_1 j_1}, \ldots, a_{i_r j_r}\}$ назовем правильной, если единичный элемент $a_{i_u j_u}$ не является запрещенным в $L_u(Q)$ при $u \in \{1,2,\ldots,r\}$.*

Через $S(L)$ обозначим множество всех правильных перестановочных подматриц матрицы L.

Утверждение 3. *Набор столбцов $H(Q)$ является неприводимым покрытием в L тогда и только тогда, когда Q – максимальная правильная перестановочная подматрица в L.*

Пусть $Q = \{a_{i_1 j_1}, \ldots, a_{i_r j_r}\}$ – перестановочная подматрица. Через $\hat{L}_u(Q)$, $u \in \{1,\ldots,r{+}1\}$, обозначим подматрицу, полученную из $L_u(Q)$, удалением нулевых столбцов и охватывающих строк. Так как $\hat{L}_u(Q)$ не содержит охватывающих строк, то любой ее единичный элемент не является запрещенным.

Через $S^*(L)$ обозначим множество всех таких подматриц $Q = \{a_{i_1 j_1}, \ldots, a_{i_r j_r}\}$ из $S(L)$, для которых выполнено: $a_{i_u j_u} \in \hat{L}_u(Q)$ при $u \in \{1,\ldots,r\}$.

Легко видеть, что если в подматрице $Q = \{a_{i_1 j_1}, \ldots, a_{i_r j_r}\}$ элемент $a_{i_u j_u}$ не является запрещенным в $L_u(Q)$, строка с номером i_u охватывает строку с номером i в $L_u(Q)$ и $a_{i j_u} = 1$, то набор единичных элементов

$$\{a_{i_1 j_1}, \ldots, a_{i_{u-1} j_{u-1}}, a_{i j_u}, a_{i_{u+1} j_{u+1}}, \ldots, a_{i_r j_r}\}$$

образует перестановочную подматрицу, в которой элемент $a_{i j_u}$ также не является запрещенным. Из сказанного следует

Утверждение 4. *Набор столбцов H матрицы L является неприводимым покрытием в том и только в том случае, если в $S^*(L)$ существует максимальная перестановочная подматрица Q, такая, что $H = H(Q)$.*

Определение 7. *Подматрицу $Q = \{a_{i_1 j_1}, \ldots, a_{i_r j_r}\}$ из $S^*(L)$ назовем верхней, если для любого $u \in \{1,\ldots,r\}$ и любого $i > i_u$, $i \in \{1,\ldots,m\}$, набор $\{a_{i_1 j_1}, \ldots, a_{i_{u-1} j_{u-1}}, a_{i j_u}, a_{i_{u+1} j_{u+1}}, \ldots, a_{i_r j_r}\}$ не образует перестановочную подматрицу из $S^*(L)$.*

Через $V(L)$ обозначим множество всех верхних подматриц в $S^*(L)$.

При работе с матрицей L алгоритм АО2 строит дерево решений D_L. Вершинам дерева (кроме корня) сопоставлены правильные перестановочные подматрицы из $S^*(L)$.

Пусть $q = \min(m,n)$. На шаге t, $t \geqslant 1$, за время $O(qm^2)$ алгоритм АО2 строит ветвь дерева D_L (строит висячую вершину). Подматрица Q из $S^*(L)$, соответствующая висячей вершине ветви, является максимальной и, следовательно, согласно утверждению 3, набор столбцов $H(Q)$ – неприводимое покрытие.

Неприводимое покрытие строится столько раз, сколько подматриц из $S^*(L)$ содержит H. Таким образом, алгоритм АО2 делает «лишние» шаги. Шаг считается лишним, если висячей вершине, построенной на этом шаге, соответствует подматрица из $S^*(L)$, не являющаяся верхней. Проверка, является ли построенная правильная перестановочная подматрица порядка r верхней, требует просмотра не более, чем $m^2 r$ элементов матрицы L.

Отметим, что в случае $t > 1$ первая вершина ветви находится на ярусе, номер которого меньше номера яруса последней из построенных висячих вершин не более, чем на единицу.

Описание алгоритма PL2

В данном разделе приводится доказательство теоремы 1.

Пусть $Q \in S^{'}(L)$, $Q = \{a_{i_1 j_1}, \ldots, a_{i_r j_r}\}$.

Строку матрицы L с номером i назовем *конкурентной* для Q, если эта строка имеет в точности один единичный элемент $a_{i j_u}$ в пересечении со столбцами с номерами $j_1, \ldots, j_r$, $i > i_u$ и $a_{i j_u} \in \hat{L}_u(Q)$.

Очевидно, что $Q \in V(L)$ тогда и только тогда, когда множество строк, конкурентных для Q, пусто. Очевидно также, что $H(Q) \in P(L)$ тогда и только тогда, когда подматрица $\hat{L}_{r+1}(Q)$ пуста.

Пусть $H(Q) \in P(L)$, $r > 1$, $Q \in V(L)$, $u \in \{1,2,\ldots,r\}$, $Q_u = \{a_{i_1 j_1}, \ldots, a_{i_u j_u}\}$, $a_{ij} \in \hat{L}(Q_u)$. Положим $\hat{L}(a_{ij}) = \hat{L}_{u+1}(Q_{u-1} \cup \{a_{ij}\})$.

Пусть далее $\hat{L}(a_{ij})$ не пуста и J- множество номеров столбцов подматрицы $\hat{L}(a_{ij})$. Обозначим через $M(a_{ij})$ подматрицу матрицы L, образованную столбцами с номерами из J и строками, конкурентными для $Q_{u-1} \cup \{a_{ij}\}$.

Предположим, что $M(a_{ij})$ не пуста (Q имеет конкурентные строки) и не содержит нулевых строк. Тогда, очевидно, набор ненулевых столбцов H матрицы $M(a_{ij})$ является неприводимым покрытием, в каждой строке которого в точности один элемент равен 1. Пусть J_{ij} — набор номеров столбцов, образующих H.

Определение 8. *Элемент a_{ij} называется узловым в $\hat{L}(a_{ij})$, если имеет место один из следующих случаев:*

1) *подматрица Q_u не имеет конкурентных строк;*
2) *подматрица Q_u имеет конкурентные строки и J_{ij} содержит перестановочную подматрицу.*

Утверждение 5. *Перестановочная подматрица Q, $Q = \{a_{i_1 j_1}, \ldots, a_{i_r j_r}\}$, принадлежит $V(L)$ тогда и только тогда, когда при любом $u \in \{1,\ldots,r\}$ элемент a_{ij} является узловым.*

Приведем описание алгоритма PL2.

При работе с матрицей L алгоритм PL2 строит дерево решений D_L^*. Вершинам дерева (кроме корня) сопоставлены подматрицы из $V(L)$. Корень дерева есть пустая вершина.

На шаге t, $t \geqslant 1$, за время $O(qm^2)$ алгоритм строит ветвь дерева D_L^*, представляющую собой последовательность вершин $Q_1^t, Q_2^t, \ldots, Q_p^t, \ldots, Q_{p_t}^t$, в которой две соседние вершины соединены ребром, а $Q_{p_t}^t$ — висячая вершина.

Первый шаг алгоритма PL2 совпадает с первым шагом алгоритма АО2.

Шаг 1. Строится первая внутренняя вершина дерева $Q_1^1 = \{a_{ij}\}$, где a_{ij} — единичный элемент с наименьшим номером в $\hat{L}$. Очевидно, $Q_1^1 \in V(L)$.

Пусть построена вершина $Q_p^1 = \{a_{i_1 j_1}, \ldots, a_{i_r j_r}\}$, $1 \leqslant p \leqslant p_1$, $Q_p^1 \in V(L)$.

Если $\hat{L}_{r+1}(Q_p^1) \neq \varnothing$, то, согласно утверждению 2, подматрица Q_p^1 не является максимальной, и набор столбцов $H(Q)$ не является неприводимым покрытием. В этом случае строим следующую вершину первой ветви. Полагаем $Q_{p+1}^1 = Q_p^1 \cup \{a_{ij}\}$, где a_{ij} — единичный элемент с наименьшим номером в $\hat{L}_{r+1}(Q_p^1)$. Очевидно, $Q_{p+1}^1 \in V(L)$.

В случае $\hat{L}_{r+1}(Q_p^1) = \varnothing$ набор столбцов $H(Q_p^1)$ является неприводимым покрытием. Переходим к шагу 2.

Шаг $t+1$, $t \geqslant 1$. Пусть на шаге t, построена висячая вершина $Q_{p_t}^t = \{a_{i_1 j_1}, \ldots, a_{i_r j_r}\}$, $r \geqslant 1$, $Q_{p_t}^t \in V(L)$.

Если $r = 1$, то из $\hat{L}$ удаляем столбцы, номера которых меньше или равны j_r. Если в полученной матрице L' есть нулевые строки, то конец работы алгоритма. В противном случае, с помощью алгоритма АО2 строим новую ветвь исходящую из вершины $\{a_{ij}\}$, где a_{ij} — единичный элемент с минимальным номером в L' (см. шаг 1).

Если $r > 1$, то в $\{1, \ldots, r\}$ ищем максимальное число u такое, что в $\hat{L}_u(Q_{p_t}^t)$ есть узловые элементы, следующие за $a_{i_u j_u}$. Если такое u существует, то полагаем $Q_1^{t+1} = \{a_{i_1 j_1}, \ldots, a_{i_{u-1} j_{u-1}, a_{ij}}\}$, где a_{ij} — узловой элемент, непосредственно следующий за $a_{i_u j_u}$ в $\hat{L}_u(Q_{p_t}^t)$.

Если же указанное u не существует, то из $\hat{L}$ удаляем столбцы, номера которых меньше или равны j_r. Если в получившейся подматрице L' есть нулевые строки, то конец работы алгоритма. В противном случае, с помощью алгоритма АО2 строим новую ветвь, исходящую из вершины a_{ij}, где a_{ij} — единичный элемент с минимальным номером в $\hat{L}$ (см. шаг 1).

Теорема 2. *Множество всех максимальных конъюнкций булевой функции F от n переменных, заданной 2-КНФ с m элементарными дизъюнкциями, может быть построено с задержкой $O(qm^2)$.*

Доказательство Теоремы 2 основано на полиномиальном преобразовании исходной 2-КНФ с отрицаниями переменных в 2-КНФ без отрицаний переменных.

Результаты экспериментов

Проведено тестирование алгоритма PL2 на случайных матрицах, содержащих не более двух единиц в строке. PL2 сравнивался с алгоритмом из [6]. Результаты счета представлены в таблице, в которой t_1 — время работы алоритма из [6], t_2 — время работы алгоритма PL2.

Эксперименты показали, что время работы PL2 сравнимо со временем работы алгоритма из [6] в случае $m = n$ (отношение t_2/t_1 близко к единице).

В случае $m > n$ алгоритм из [6] работает быстрее алгоритма PL2.

В случае $n \gg m$, алгоритм PL2 обгоняет алгоритм из [6] ($t_2/t_1 > 1$), и разрыв значительно увеличивается с ростом n.

Размеры матриц $m \times n$	t_1, с	t_2, с	t_2/t_1
50×50	4,3	4,95	1,15
100×30	3,2	5,2	1,54
30×100	6,4	4,1	0,64
40×120	15,3	6,31	0,41
50×150	27,4	9,4	0,34
70×200	111,3	15,6	0,14

Литература

[1] *Дюкова Е. В* Об асимптотически оптимальном алгоритме построения тупиковых тестов // ДАН СССР. — 1977. — Т. 233, № 4. — С 527–530.

[2] *Дюкова Е.В.* О сложности реализации дискретных (логических) процедур распознавания // ЖВМ и МФ, 2004. — Т. 44, № 3. — С. 551–561.

[3] *Дюкова Е.В., Инякин А.С.* Асимптотически оптимальное построение тупиковых покрытий целочисленной матрицы // Математические вопросы кибернетики. — 2008. — N. 17. — С. 235–246.

[4] *Boros E., Elbassioni K., Gurvich V., Khachiyan L* An Efficient Implementation of a QuasiPolynomial Algorithm for Generating Hypergraph Transversals and its Application in Joint Generation // Discrete Appl. Math. — 2006. — V. 154, N. 16. — Pp. 2350–2372.

[5] *Gurvich V., Khachiyan L.* On Generating the Irredundant Conjunctive and Disjunctive Normal Forms of Monotone Boolean Functions // Discrete Appl. Math. — 1999. — V. 96–97, N. 1–3. — Pp. 363–373.

[6] *Jonson D.S., Yannakakis M., Papadimitriou C.H.* On Generating All Maximal Independent Sets Information Processing Letters. — 1988. — V. 27. — Pp. 119–123.

О построении сокращенных множеств неприводимых покрытий булевой матрицы*

Инякин А. С.
andre_w@mail.ru
Москва, Учреждение Российской академии наук Вычислительный центр им. А.А. Дородницына Российской академии наук

В статье рассматривается стохастический подход к построению сокращенных множеств элементарных классификаторов в задачах распознавания и классификации.

Одними из наиболее сложных в вычислительном плане задач, решаемых при конструировании дискретных (логических) процедур распознавания и классификации являются задачи поиска покрытий булевых и целочисленных матриц. Задачи поиска покрытий возникают, например, при построении множества элементарных классификаторов в таких алгоритмах распознавания, как тестовые алгоритмы, алгоритмы голосования по представительным и антипредставительным наборам, алгоритмы голосования по покрытиям класса; при синтезе корректных процедур распознавания; при решении задач кластерного анализа; при построении корректных перекодировок и т.д. [1]–[8]. Особенно трудоемкой является задача поиска неприводимых покрытий булевых матриц, которая формулируется следующим образом.

Пусть L — булева матрица размера $m \times n$. Набор H из r, различных столбцов матрицы L, называется покрытием, если в подматрице L^H матрицы L, образованной столбцами набора H, не содержится нулевой строки. Покрытие называется неприводимым (или тупиковым), если никакое его собственное подмножество не является покрытием. Очевидно, покрытие является неприводимым тогда и только тогда, когда в подматрице L^H содержится единичная (перестановочная) подматрица порядка r. Положим $P(L)$ — множество всех неприводимых покрытий матрицы L. Требуется найти $P(L)$.

Данная задача может быть также сформулирована как задача построения максимальных конъюнкций двузначной логической функции, заданной множеством нулей. Поиски эффективных алгоритмов ее решения ведутся с середины 1950-х годов [9]–[10].

В настоящее время для решения указанной задачи существует ряд эффективных в теоретическом и практическом плане алгоритмов.

В [1]–[5] рассмотрен случай, когда число строк m булевой матрицы L имеет более низкий порядок роста, чем число столбцов n, при условии, что $n \to \infty$. Для этого случая построен асимптотически оптимальный алгоритм поиска неприводимых покрытий (алгоритм АО1). Данный алгоритм строит с задержкой, не превосходящей $O(mn)$ приближенное решение, в качестве которого рассматривается совокупность наборов столбцов, содержащих единичные подматрицы. Каждый такой набор алгоритм АО1 строит столько раз, сколько единичных подматриц он содержит. Показано, что если $m^\alpha \leqslant n \leqslant 2^{m^\beta}$, $\alpha > 1$, $\beta < 1$, то при $n \to \infty$ число шагов данного алгоритма, равное числу единичных подматриц, почти всегда (для почти всех матриц размера $m \times n$) асимптотически равно мощности $P(L)$. При конструировании алгоритма АО1 в качестве приближенного решения может быть рассмотрена совокупность наборов столбцов, содержащих максимальные единичные подматрицы (каждый такой набор столбцов строится столько раз, сколько максимальных единичных подматриц он содержит). Тогда алгоритм будет делать меньшее число шагов и работать с задержкой не превосходящей $O(qmn)$, здесь и далее $q = \min(m, n)$. Данная модификация алгоритма АО1 обычно используется при его реализации на ЭВМ [12].

В [6], [7] построен алгоритм, основанный на переборе с задержкой не превосходящей $O(qm^2n)$ неприводимых покрытий матрицы L (алгоритм АО2). Однако данный алгоритм строит каждый набор из $P(L)$ столько раз, сколько единичных подматриц он содержит. Из сказанного выше следует, что указанный недостаток не существенен при $m^\alpha \leqslant n \leqslant 2^{m^\beta}$, $\alpha > 1$, $\beta < 1$ (в этом случае число шагов алгоритма АО2, равное числу единичных подматриц, порождающих неприводимые покрытия, почти всегда при $n \to \infty$ асимптотически равно мощности $P(L)$).

Отметим, что проверка на повторяемость построенного на очередном шаге набора столбцов в алгоритмах АО1 и АО2 требует просмотра не более $O(qm)$ элементов матрицы L.

В [13] построен алгоритм (обозначаемый далее алгоритм УОС), в основе которого лежит следующее простое утверждение: в неприводимом покрытии не существует пары столбцов такой, что один столбец охватывает другой. На каждом шаге работы алгоритма строится наборов столбцов, не содержащих в себе охватывающих. За счет этого удается существенно сократить перебор. Алгоритм явля-

Работа выполнена при финансовой поддержке РФФИ, проект № 10-07-00673-а.

ется «универсальным», т.е. эффективным при любых соотношениях числа строк и столбцов исходной матрицы, и, кроме того, его применение позволяет существенно сократить время построения множества неприводимых покрытий по сравнению с алгоритмами АО1, АО2 и другими.

В [14]–[15] построен алгоритм поиска неприводимых покрытий булевой матрицы (алгоритм ОПТ), асимптотически оптимальный при тех же условиях, что и алгоритмы АО1 и АО2. Данный алгоритм основан на переборе с задержкой не превосходящей $O(qmn(m+q))$ наборов столбцов матрицы, содержащих единичные подматрицы и удовлетворяющих некоторым дополнительным условиям. Результаты численных экспериментов со случайными матрицами при различных соотношениях между m и n показали, что практически во всех случаях алгоритм ОПТ делает существенно меньшее число шагов и поэтому работает существенно быстрее по сравнению с алгоритмами АО1 и АО2, а также другими известными алгоритмами поиска неприводимых покрытий булевой матрицы.

Несмотря на то, что к настоящему времени построены достаточно эффективные алгоритмы поиска всех неприводимых покрытий, на практике их применение ограничено ввиду высокой вычислительной сложности и плохой интерпретируемости получаемого решения. Уже для относительно небольшой задачи со 100 признаками и 60 объектами число элементарных классификаторов, порождаемых неприводимыми покрытиями доходит до десятков миллионов. Поэтому актуальной является задача построения сокращенных и, в некотором смысле, «представительных» подмножеств множества $P(L)$.

В дальнейших рассуждениях будет использоваться наиболее простой с описательной точки зрения алгоритм поиска неприводимых покрытий будевой матрицы УОС.

Алгоритм поиска неприводимых покрытий

Введем ряд определений. Пусть $L = (a_{ij})$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, — булева матрица. Столбец матрицы L будем называть единичным, если все его элементы равны единице, и нулевыми, если все его элементы равны нулю. Остальные столбцы будем называть смешанными. Будем говорить, что столбец с номером j_1 охватывает столбец с номером j_2, если $a_{ij_1} \geqslant a_{ij_2}$ для любых $i = 1, \ldots, m$. Будем говорить, что столбец с номером j покрывает строку с номером i, или строка с номером i покрывает столбец с номером j, если $a_{ij} = 1$.

Суть алгоритма поиска неприводимых покрытий УОС состоит в следующем.

В процессе решения задачи, алгоритм УОС осуществляет односторонний обход дерева, вершинам которого соответствуют пары $\left(H, L^H\right)$, где H — некоторый упорядоченный набор номеров столбцов матрицы L и L^H — некоторая подматрица матрицы L, которая специальным образом строится по набору столбцов H. Обход дерева начинается с корневой вершины, которой соответствует пара $(\varnothing, L)$.

В вершине $(\varnothing, L)$ последовательно выполняются следующие действия. Выписываются неприводимые покрытия матрицы L, порождаемые единичными столбцами. Осуществляется последовательный обход вершин вида $\left(\{j_1\}, L^{\{j_1\}}\right)$, где j_1 — смешанный столбец матрицы L, $L^{\{j_1\}}$ — матрица, состоящая из смешанных столбцов матрицы L, не охватывающих столбец j_1, и строк матрицы L, не покрытых столбцом j_1, и $L^{\{j_1\}}$ не пуста.

При обходе вершин дерева вида $\left(H, L^H\right)$, где $H = \{j_1, \ldots, j_{r-1}\}$ набор из $r-1$ столбцов, $r > 1$, последовательно выполняются следующие действия (рис. 1).

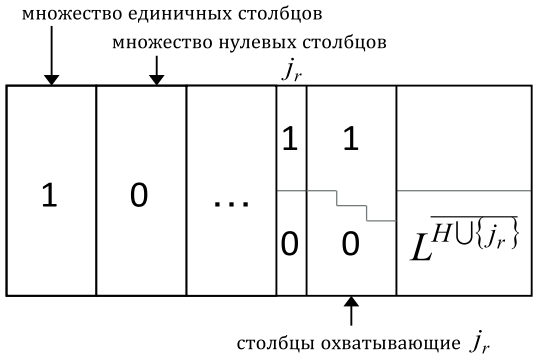


Рис. 1. Матрица L^H

1. Просматриваются все наборы столбцов матрицы L вида $H \cup \{j_r\}$, где j_r — единичный столбец матрицы L^H, по построению являющиеся покрытиями матрицы L. Каждый из этих наборов столбцов такой, что матрица $L^{H\cup\{j_r\}}$ содержит единичную подматрицу порядка r, является неприводимым покрытием.
2. Осуществляется последовательный обход вершин дерева вида $\left(H \cup \{j_r\}, L^{H\cup\{j_r\}}\right)$, где j_r смешанный столбец матрицы L^H, $L^{H\cup\{j_r\}}$ — матрица, состоящая из смешанных столбцов матрицы L^H, не охватывающих столбец j_r и строк матрицы L^H, не покрытых столбцом j_r, и $L^{H\cup\{j_r\}}$ не пуста.
3. Осуществляется возврат к вершине

$$\left(H\backslash\{j_{r-1}\}, L^{H\backslash\{j_{r-1}\}}\right).$$

По окончании обхода дерева будет построено множество $P(L)$ всех неприводимых покрытий матрицы L.

Пусть $G(L)$ — множество наборов столбцов матрицы L таких, что любые два столбца из набора со-

держат единичную подматрицу порядка 2. Таким образом, в процессе работы алгоритма просматриваются наборы столбцов из некоторого подмножества множества $G(L)$, содержащего множество всех неприводимых покрытий, причем построение таких наборов осуществляется с полиномиальной задержкой.

Стохастический алгоритм поиска неприводимых покрытий

Пусть в каждой вершине дерева $\left(H, L^H\right)$ определен функционал $\mathscr{S}(H, j) \in \{0, 1\}$, $j \in \{1, \ldots, n\}$, $j \notin H$.

В алгоритме УОС модифицируем правила обхода дерева следующим образом.

При обходе вершин дерева вида $\left(H, L^H\right)$, где $H = \{j_1, \ldots, j_{r-1}\}$ набор из $r-1$ столбцов, $r > 1$, последовательно выполняются следующие действия.

1. Просматриваются все наборы столбцов матрицы L вида $H \cup \{j_r\}$, где j_r — единичный столбец матрицы L^H, и $\mathscr{S}(H, j_r) = 1$, по построению являющиеся покрытиями матрицы L. Каждый из этих наборов столбцов такой, что матрица $L^{H \cup \{j_r\}}$ содержит единичную подматрицу порядка r, является неприводимым покрытием.
2. Осуществляется последовательный обход вершин дерева вида $\left(H \cup \{j_r\}, L^{H \cup \{j_r\}}\right)$, где j_r — смешанный столбец матрицы L^H, $\mathscr{S}(H, j_r) = 1$, $L^{H \cup \{j_r\}}$ — матрица, состоящая из смешанных столбцов матрицы L^H, не охватывающих столбец j_r и строк матрицы L^H, не покрытых столбцом j_r, и $L^{H \cup \{j_r\}}$ не пуста.

По окончании обхода дерева будет построено сокращенное множество $P_{\mathscr{S}}(L)$ неприводимых покрытий матрицы L.

Выбор функционала $\mathscr{S}(H, j)$ определяет мощность множества $P_{\mathscr{S}}(L)$, распределение неприводимых покрытий из $P_{\mathscr{S}}(L)$ по длинам и столбцам матрицы L.

Аналогичным образом, для построения сокращенного множества $P_{\mathscr{S}}(L)$ неприводимых покрытий матрицы L, могут быть модифицированны алгоритмы АО1, АО2, ОПТ. Данный подход может быть также применен для построяния сокращенных множеств тупиковых σ-покрытий целочисленной матрицы.

Выводы

Таким образом, применения стохастического подхода для построении элементарных классификаторов позволяет существенно сократить затрачиваемое на решение прикладных задач время, при этом не снижая, а при соответствующем выборе функционала $\mathscr{S}(H, j)$, увеличивая точность решения.

Литература

[1] *Дюкова Е.В.* Об асимптотически оптимальном алгоритме построения тупиковых тестов // ДАН СССР, 1977. — Т. 233, № 4. — С. 527–530.

[2] *Дюкова Е.В.* Асимптотически оптимальные тестовые алгоритмы в задачах распознавания // Пробл. кибернетики. — М.: Наука, 1982. — Вып. 39. — С. 165–199.

[3] *Дюкова Е.В.* О сложности реализации некоторых процедур распознавания // ЖВМ и МФ, 1987. — Т. 27, № 1. — С. 114–127.

[4] *Дюкова Е.В.* Алгоритмы распознавания типа Кора: сложность реализации и метрические свойства // Распознавание, классификация, прогноз (матем. методы и их применение). — М.: Наука, 1989. — Вып. 2. — С. 99–125.

[5] *Дюкова Е.В., Журавлёв Ю.И.* Дискретный анализ признаковых описаний в задачах распознавания большой размерности // ЖВМ и МФ, 2000. — Т. 40, № 8. — С. 1264–1278.

[6] *Djukova E.V.* Discrete (Logical) Recognition Procedures: Principles of Construction, Complexity of Realization and Basic Models // Patt. Recogn. and Image Anal., 2003. — Vol. 13, No. 3. — P. 417–425.

[7] *Дюкова Е.В.* О сложности реализации дискретных (логических) процедур распознавания // ЖВМ и МФ, 2004. — Т. 44, № 3. — С. 550–572.

[8] *Дюкова Е.В., Инякин А.С.* О процедурах классификации, основанных на построении покрытий классов // ЖВМ и МФ, 2003. — Т. 43, № 12. — С. 1910–1921.

[9] *Яблонский С.В.* Введение в дискретную математику. М.: Наука, 1986.

[10] *Чегис И.А., Яблонский С.В.* Логические способы контроля электрических схем // Труды Матем. ин-та АН СССР, 1958. — Т. 51. — С. 270–360.

[11] *Jonson D.S., Yannakakis M., Papadimitriou C.H.* On General All Maximal Independent Sets // Inform. processing Letters, 1988. — V. 27. — P. 119–123.

[12] *Дюкова Е.В.* Об одном алгоритме построения тупиковых тестов для бинарных таблиц // Сб. работ по дискретной математике. М.: ВЦ АН СССР, 1976. — Вып. 1. — С. 167–185.

[13] *Инякин А.С.* Об одном алгоритме поиска тупиковых σ-покрытий // Интеллектуализация обработки информации: тезисы докладов Международной конференции, Симферополь, 2002. — С. 47–48.

[14] *Дюкова Е.В., Инякин А.С.* Построение неприводимых покрытий булевой матрицы с полиномиальной задержкой // Доклады Академии наук, 2007. — Т. 413, № 5. — С. 596–598.

[15] *Дюкова Е.В., Инякин А.С.* О сложности решения задачи построения неприводимых покрытий булевой матрицы. М.: ВЦ РАН, 2006. 23 с.

[16] *Инякин А. С.* Алгоритмы поиска неприводимых покрытий булевых матриц. // М.: ВЦ РАН, 2004. 25 с.

Приближенный метод решения одной оптимизационной задачи в теории распознавания*

Катериночкина Н. Н.
nnkater@yandex.ru
Москва, Вычислительный центр им. А. А. Дородницына РАН

В теории распознавания возникает ряд оптимизационных задач. Одной из таких задач является выделение максимальной совместной подсистемы из системы линейных неравенств. Известны методы ее решения, основанные на переборе множества узловых подсистем заданной системы линейных неравенств. Это может быть полный перебор, дающий точное решение, или частичный направленный перебор, находящий приближенное решение. В данной работе предложен приближенный метод решения указнной задачи, принципиально отличный от упомянутых выше. Метод основан на ряде соображений геометрического характера.

Решение оптимизационных задач в теории распознавания является одним из важнейших этапов синтеза высокоточных алгоритмов классификации, распознавания и прогноза. Центральными оптимизационными задачами здесь являются задачи поиска максимальных совместных подсистем систем неравенств, поиска минимальных покрытий матриц, синтеза минимальных формул реализации слабо-определенных булевских и конечнозначных функций, построения алгебраических корректоров минимальной степени и др.

В процессе оптимизации некоторых моделей алгоритмов распознавания (например, алгоритмов вычисления оценок — АВО) требуется как можно точнее удовлетворить определенной системе условий, которая описывается большим числом линейных неравенств и в целом может быть противоречивой. В общем случае решение этой проблемы сводится к выделению максимальной по числу неравенств совместной подсистемы из заданной системы линейных неравенств.

Постановка задачи

Рассмотрим систему S линейных неравенств:

$$\sum_{j=1}^{n} a_{ij}x_j \leqslant b_i, \quad i = 1, \ldots, m. \tag{1}$$

Система S вообще говоря не совместна. Требуется найти максимальную совместную подсистему (МСП) данной системы линейных неравенств.

Известен подход к решению данной задачи, использующий метод свертки системы линейных неравенств (см. [1]). Однако этот метод требует быстро растущих объемов памяти и непригоден для задач большой размерности.

Ранее автором были построены несколько алгоритмов решения этой задачи, основанных на переборе специального множества подсистем системы линейных неравенств (см. [2, 3]).

Пусть система S имеет ранг r, $r > 0$.

Работа выполнена при финансовой поддержке РФФИ, проект № 11-01-00585.

Определение 1. *Подсистему мощности r и ранга r системы S назовем r-подсистемой.*

Определение 2. *Пусть P — r-подсистема системы S. Узловым решением подсистемы P назовем такое ее решение, которое обращает все ее неравенства в равенства.*

Упомянутые алгоритмы опираются на теорию узловых решений систем линейных неравенств (см. [1]). Они основаны на переборе множества r-подсистем заданной системы линейных неравенств. Для каждой такой подсистемы требуется найти одно узловое решение, подставить его во все неравенства системы и выделить из нее совместную подсистему. При этом точные методы требуют полного перебора всех таких подсистем, что бывает слишком трудоемко для задач большой размерности. Поэтому на практике чаще используют приближенные методы, которые осуществляют частичный направленный перебор r-подсистем.

В данной работе представлен приближенный метод поиска МСП, принципиально отличный от описанных выше. Этот метод основан на ряде соображений геометрического характера. Производится сравнение двух приближенных методов: предложенного метода и приближенного метода, основанного на переборе r-подсистем.

Описание метода

Систему S линейных неравенств будем иногда для краткости записывать в виде:

$$l_i(\boldsymbol{x}) \leqslant b_i, \quad i = 1, \ldots, m, \tag{2}$$

где $l_i(\boldsymbol{x}) = (\boldsymbol{a}_i, \boldsymbol{x}) \equiv \sum\limits_{j=1}^{n} a_{ij}x_j$.

Определение 3. *Неравенство $(\boldsymbol{a}, \boldsymbol{x}) \leqslant b$ назовем несобственным, если вектор $\boldsymbol{a}$ нулевой.*

Несобственное неравенство в зависимости от знака величины b либо противоречиво, либо тождественно. Очевидно, что несобственные неравенства можно не учитывать в процессе построения МСП системы. Противоречивые неравенства

не войдут ни в одну совместную подсистему, а тождественные можно добавить к любой совместной подсистеме.

Известно следующее утверждение (см., например, [4]).

Утверждение 1. *Любая несовместная система линейных неравенств, не содержащая несобственных неравенств, может быть разбита на две совместные подсистемы.*

Действительно, разобьем систему (2) на две подсистемы следующим образом. Обозначим через I_+ множество тех индексов $i \in \{1, \ldots, m\}$, для которых первый из отличных от нуля коэффициентов $a_{i1}, \ldots, a_{in}$ положителен. Положим $I_- = \{1, \ldots, m\} \setminus I_+$. Система (2) распадается на две подсистемы:

$$l_i(\boldsymbol{x}) \leqslant b_i, \quad i \in I_+; \tag{2.1}$$

$$l_i(\boldsymbol{x}) \leqslant b_i, \quad i \in I_-. \tag{2.2}$$

Легко показать, что эти подсистемы совместны, и найти их частные решения.

Опишем общую схему работы алгоритма поиска МСП системы S вида (2).

Этап 0. Исходная система разбивается на две совместных подсистемы: (2.1) и (2.2). Для каждой из этих подсистем находится по одному частному решению: $\boldsymbol{y}$ и $\boldsymbol{z}$, соответственно.

Далее идет цикл по i, $i = 1, \ldots, m$. Причем i-я итерация содержит несколько этапов.

Этап i1. Заметим, что каждому неравенству $l_i(\boldsymbol{x}) \leqslant b_i$ исходной системы линейных неравенств соответствует гиперплоскость $l_i(\boldsymbol{x}) = b_i$. Сначала построим точки $p(\boldsymbol{y}, i)$ и $p(\boldsymbol{z}, i)$, — проекции точек $\boldsymbol{y}$ и $\boldsymbol{z}$ на гиперплоскость $l_i(\boldsymbol{x}) = b_i$.

Этап i2. Для прямой, проходящей через точки $p(\boldsymbol{y}, i)$ и $p(\boldsymbol{z}, i)$, находим точки ее пересечения со всеми остальными гиперплоскостями, кроме i-й. Пусть это будут точки $\boldsymbol{u}^1, \ldots, \boldsymbol{u}^l$ $(l \leqslant m-1)$.

Этап i3. Каждую из полученных точек $\boldsymbol{u}^1, \ldots, \boldsymbol{u}^l$ подставляем поочередно во все неравенства исходной системы (2) и подсчитываем число неравенств, которым эта точка удовлетворяет. Эти неравенства образуют заведомо совместную подсистему. Мощность этой подсистемы сравниваем с рекордной мощностью, полученной на предыдущих шагах. При этом запоминаем текущее максимальное значение мощности ρ совместной подсистемы и соответствующую точку $\boldsymbol{v}$.

В результате после всех итераций запоминается рекордное значение мощности ρ и соответствующее решение $\boldsymbol{v}$. По этому решению восстанавливается совместная подсистема, которая принимается в качестве приближения МСП системы S.

Оценим порядок числа операций при работе описанного алгоритма. Алгоритм содержит m итераций, на каждой из которых мы находим не более m точек. Каждую точку подставляем в m неравенств от n переменных. Всего получаем порядка $m^3 n$ операций.

О параметрах предложенного алгоритма

На начальном этапе система (2) разбивается на две совместные подсистемы, определенные выше. Для каждой из этих подсистем надо найти одно частное решение. Очевидно, что частные решения подсистем (2.1) и (2.2) можно находить не однозначно. Рассмотрим подсистему (2.1). Матрицу $\boldsymbol{A}_+$ коэффициентов этой подсистемы путем перестановки строк можно привести к ступенчатому виду. При этом первые ненулевые столбцы каждой ступени будут состоять из положительных элементов. Пусть приведенная матрица имеет k ступеней. Обозначим через i_j, $j = 1, \ldots, k$, номер первого ненулевого столбца j-й ступени, а через l_j — номер верхней строки j-й ступени. Будем искать частное решение подсистемы, начиная с неравенств, соответствующих k-й ступени приведенной матрицы. Очевидно, что переменным x_j при $i_k < j \leqslant n$ можно придать любые значения. Тогда подсистема, соответствующая k-й ступени, примет вид:

$$a_{ji_k} x_{i_k} \leqslant c_j, \quad j = l_k \ldots, m, \quad a_{ji_k} > 0.$$

Следовательно, выбор значения переменной x_{i_k} нужно подчинить условию:

$$x_{i_k} \leqslant \min_{l_k \leqslant j \leqslant m} \frac{c_j}{a_{ji_k}}.$$

Подставляя найденные значения переменных в предыдущие неравенства, найдем по аналогии значения остальных переменных.

По такой же схеме строится частное решение подсистемы (2.2), с тем отличием, что множество значений соответствующей переменной будет ограничено не сверху, а снизу.

Таким образом, при поиске частных решений одним переменным (свободным) можно придавать любые значения, а другим — произвольные значения из бесконечного множества.

Следовательно, эти переменные можно считать параметрами. Варьируя эти параметры, будем получать различные частные решения. Этим фактом можно пользоваться в процессе применения представленного метода. Меняя соотношения между параметрами, можно попытаться улучшить первоначальный результат. (Настроиться на задачу в диалоговом режиме).

Сравнение двух методов решения рассматриваемой задачи

Было проведено сравнение представленного метода (метод 2) и приближенного метода, основанного на частичном переборе множества r-подсистем заданной системы линейных неравенств (метод 1). Этими методами было решено около пяти десятков задач. При этом на 48% задач оба указанных метода получили приближение для МСП одинаковой мощности, на 24% задач лучший результат дал метод 1, а на 28% задач лучше сработал метод 2.

Заключение

Представленный в данной работе метод поиска максимальной совместной подсистемы системы линейных неравенств можно применять наряду с другими. На некоторых задачах он может сработать лучше. Кроме того, его можно настраивать на задачу, варьируя параметры при поиске частных переменных.

Литература

[1] *Черников С. Н.* Линейные неравенства. — Москва: Наука, 1968. — 488 с.

[2] *Катериночкина Н. Н.* Методы выделения оптимальной совместной подсистемы системы линейных неравенств // Математические методы распознавания образов. Доклады 12-й Всероссийской конференции (ММРО-12), 2005. — С. 122–125.

[3] *Katerinochkina N. N.* Decision methods for some discrete extreme problems in recognition theory // Pattern Recognition and Image Analysis. — 2008. — V. 18, N. 4. — Pp. 584–587.

[4] *Еремин И. И.* Линейная оптимизация и системы линейных неравенств. — Москва: Издательский центр «Академия», 2007. — 249 с.

Вопросы комитетной полиэдральной отделимости конечных множеств*

Поберий М. И.
maschas_briefen@mail.ru
Екатеринбург, ИММ УрО РАН

Статья посвящена исследованию вычислительной сложности и аппроксимируемости задач комбинаторной оптимизации, связанных с комитетной полиэдральной отделимостью конечных множеств, а также исследованию вопроса обоснования обобщающей способности комитетных кусочно-линейных решающих правил в терминах класической теории Вапника-Червоненкиса.

С 80-х годов прошлого века у исследователей возник интерес к изучению вычислительной сложности комбинаторных задач, связанных с процедурой обучения распознаванию образов. Задача о минимальном по числу элементов аффинном разделяющем комитете (MASC), занимающая центральное место в настоящей статье, порождена оптимальными процедурами обучения распознаванию образов в классе коллективных кусочно-линейных решающих правил комитетного типа. Таким образом, актуальность исследования задачи MASC подтверждается ее тесной связью с процедурой обучения распознаванию обра зов и важностью построения оптимальных (с точки зрения теории структурной минимизации риска) комитетных кусочно-линейных решающих правил.

Введение

Задачей обучения распознаванию образов называется оптимизационная задача

$$\min_{\alpha\in\Lambda} P(\alpha) = \min_{\alpha\in\Lambda} \int\limits_{X\times\Omega} \big(f(x,\alpha)-\omega\big)^2 dP(x,\omega), \quad (1)$$

где X — пространство результатов измерений, $\Omega=\{0,1\}$ — множество названий образов (классов), P — вероятностная мера, заданная с точностью до конечной выборки $(x_1,\omega_1),\ldots,(x_l,\omega_l)$, $\mathcal{F}=\{f(\cdot,\alpha)\colon X\to\Omega : \alpha\in\Lambda\}$ — класс решающих правил, в котором происходит обучение. Для аппроксимации неполностью формализованной задачи (1) традиционно рассматривается *задача минимизации эмпирического риска*:

$$\min_{\alpha\in\Lambda}\big\{\nu(\alpha)\equiv\frac{1}{l}\sum_{i=1}^{l}\big(\omega_i-f(x_i,\alpha)\big)^2\big\}. \quad (2)$$

Как известно [1], точность аппроксимации монотонно убывает с ростом емкости VCD класса решающих правил. В рамках алгебраического подхода к решению задач распознавания исследуются классы, содержащие корректные на выборке решающие правила. Для таких классов повышение качества обучения может быть связано с минимизацией емкости класса, то есть с решением задачи

$$\min\big\{\mathrm{VCD}(\mathcal{F}') : \min\{\nu : f\in\mathcal{F}'\}=0,\ \mathcal{F}'\subset\mathcal{F}\big\}. \quad (3)$$

Исследуемая в статье задача о минимальном аффинном разделяющем комитете (MASC) является частным случаем задачи (3), в котором $\mathcal{F}$ — класс комитетных кусочно-линейных решающих правил.

Постановка задачи и известные результаты

Возникновение комитетных конструкций обусловлено необходимостью обобщения классического понятия решения на случай несовместных систем. Данный подход активно применяется в теории голосования, оптимизации и классификации, распознавании образов и математическом программировании.

Определение 1. *Конечная последовательность функций $K=(f_1,\ldots,f_q)$, $f_i(x)=c_i^Tx-d_i$, называется аффинным комитетом, разделяющим множества $A,B\subset\mathbb{R}^n$, если выполнено условие*

$$|\{i\in\mathbb{N}_q:\ f_i(a)>0\}|>q/2 \qquad (a\in A);$$

$$|\{i\in\mathbb{N}_q:\ f_i(b)<0\}|>q/2 \qquad (b\in B),$$

при этом q называется числом элементов (членов) комитета K.

Как известно [2], множества A и B отделимы аффинным комитетом тогда и только тогда, когда $A\bigcap B=\varnothing$. Тем не менее по ряду причин особый интерес представляют разделяющие комитеты с наименьшим (для данных множеств) числом элементов, называемые *минимальными*.

Задача 1. *«Минимальный аффинный разделяющий комитет»* (MASC). Заданы конечные множества $A,B\subset\mathbb{Q}^n$, $A=\{a_1,\ldots,a_{m_1}\}$ и $B=\{b_1,\ldots,b_{m_2}\}$. Требуется указать аффинный комитет с наименьшим числом элементов, разделяющий множества A и B.

В общем случае задача MASC является труднорешаемой и трудноаппроксимируемой [3].

Работа выполнена при финансовой поддержке РФФИ, проект № 10-07-00134, и президиума УрО РАН, проекты № 09-П-1-1001 и № 09-С-1-1010.

Известно, что многие NP-трудные в общем случае задачи комбинаторной оптимизации становятся полиномиально (или псевдополиномиально) разрешимыми при дополнительных ограничениях: при фиксации размерности пространства, числа ограничений и т. п. Также известно, что задача MASC, заданная в одномерном пространстве, может быть решена за полиномиальное время [2]. Поэтому интерес вызывает исследование вычислительной сложности и аппроксимируемости задачи о минимальном аффинном разделяющем комитете в пространствах фиксированной размерности, большей единицы MASC(n).

В работе [4] показано, что задача MASC(n) остается труднорешаемой при произвольном фиксированном $n > 1$, однако для обоснования этого факта рассматриваются частные случаи задачи MASC(n), в которых разделяемые множества находятся не в общем положении, то есть доказательство существенным образом опирается на вырожденность разделяемых множеств. Для того, чтобы исключить из рассмотрения подобные частные случаи задачи MASC(n), далее будет введено дополнительное ограничение на разделяемые множества.

Труднорешаемость задачи MASC-GP(n)

Данный раздел посвящен исследованию вычислительной сложности задачи о минимальном аффинном разделяющем комитете в пространствах фиксированной размерности, большей единицы, при условии общности положения разделяемых множеств — назовем ее MASC-GP(n).

Определение 2. *Множество $D \subset \mathbb{R}^n$, $|D| > n$, находится в общем положении, если для каждого подмножества $D' \subseteq D$ мощности $n+1$ справедливо соотношение* $\dim \operatorname{aff}(D') = n$.

Очевидно, что доказательство труднорешаемости задачи MASC-GP(n) при $n > 1$ достаточно провести на плоскости, для чего обоснуем полиномиальную сводимость к задаче MASC-GP(2), сформулированной в виде задачи верификации свойства (назовем ее PASC-GP), известной NP-полной (в сильном смысле) задачи о покрытии конечного множества точек плоскости множеством прямых (PC).

Задача 2. *«Покрытие прямыми конечного множества точек плоскости»* (PC). Заданы конечное множество точек плоскости $P = \{p_1, \ldots, p_k\}$ с целочисленными координатами и число $s \in \mathbb{N}$. Существует ли покрытие L множества P, не превосходящее по мощности s?

Н. Мегиддо и А. Тамиром доказано, что в общем случае задача PC NP-полна в сильном смысле [5].

Пусть далее условие частной задачи PC определяется конечным множеством P из k целочисленных точек и некоторым натуральным числом s. Определим числа ρ и ε по формулам

$$\rho = \max\bigl\{\|p\|_2 : p \in P\bigr\}, \quad \varepsilon = \frac{1}{6(2\rho+1)+1}. \quad (4)$$

Зафиксируем двумерные векторы σ и τ так, что $\|\sigma\|_2 = |\tau\|_2 = 1$, $\sigma^T\tau = 0$ и для любых $\{i, j\} \subset \mathbb{N}_k$ пары отрезков $[p_i - \varepsilon\sigma, p_i + \varepsilon\sigma]$ и $[p_j - \varepsilon\sigma, p_j + \varepsilon\sigma]$, $[p_i - \varepsilon\tau, p_i + \varepsilon\tau]$ и $[p_j - \varepsilon\tau, p_j + \varepsilon\tau]$ не лежат на одной прямой. Сопоставим исходной задаче PC частную задачу PASC-GP, определяемую соотношениями $A = \{p \pm \frac{\varepsilon(p)}{M}\tau : p \in P\}$, $B = \{p \pm \varepsilon(p)\sigma : p \in P\}$, $t = 2s + 1$.

Числа $\varepsilon(p) \in (0, \varepsilon)$ и $M > 0$ выберем так, чтобы выполнялось условие $\max\limits_{p\in P} \frac{\varepsilon(p)}{M} < \min\limits_{p\in P} \varepsilon(p)$ и множество $A \cup B$ находилось в общем положении.

Описанный переход от задачи PC к задаче PASC-GP может быть осуществлен за полиномиальное время. Для завершения обоснования полиномиальной сводимости достаточно показать, что задача PC и поставленная ей в соответствие задача PASC-GP имеют положительные или отрицательные ответы одновременно.

Теорема 1. *Множество $P = \{p_1, \ldots, p_k\} \subset \mathbb{Z}^2$ обладает покрытием из s прямых в том и только в том случае, когда соответствующие ему множества $A = \{p \pm \frac{\varepsilon(p)}{M}\tau : p \in P\}$ и $B = \{p \pm \varepsilon(p)\sigma : p \in P\}$ отделимы аффинным комитетом из $2s+1$ элемента.*

Следствие 1. *Задача* PASC-GP *является* NP*-полной в сильном смысле. Задачи* MASC-GP(n) *и* MASC-GP *являются* NP*-трудными в сильном смысле.*

Аппроксимируемость задачи MASC-GP(n)

Традиционный для теории вычислительной сложности подход к исследованию NP-трудных задач комбинаторной оптимизации предполагает анализ аппроксимационных свойств задачи, разработку и обоснование приближенных алгоритмов для ее решения. В этом разделе представлен полиномиальный приближенный алгоритм решения задачи MASC-GP(n). При этом процесс построения решения описывается в терминах графа максимальных аффинно разделимых подмножеств множества $Z = A \cup B \subset \mathbb{Q}^n$, определяющего условие частной задачи MASC-GP(n).

Определение 3. *Подмножество $Z' = A' \cup B'$, где $A' \subseteq A$, $B' \subseteq B$, называется аффинно разделимым подмножеством (множества Z), если существует вектор $c \in \mathbb{R}^n$ и число $d \in \mathbb{R}$ такие, что*

$$\begin{cases} c^T a - d > 0, & (a \in A'); \\ c^T b - d < 0, & (b \in B'). \end{cases} \quad (5)$$

Алгоритм 1. Алгоритм посторения графа G_Z.

1: полагаем $V = \varnothing$ и $E = \varnothing$;
2: **для всех** подмножества $\zeta \subset Z$, $|\zeta| = n$ (т. к. Z находится в общем положении, такое подмножество ζ существует)
3: строим гиперплоскость $H(x) = 0$, содержащую подмножество ζ; ввиду общности положения множества Z такая гиперплоскость единственная и $H(x) \neq 0 \quad (x \in Z \setminus \zeta)$;
4: определяем множества X_1 и X_2:
$X_1 = \zeta \cup \{a \in A\colon H(a) > 0\} \cup \{b \in B\colon H(b) < 0\}$;
$X_2 = \zeta \cup \{a \in A\colon H(a) < 0\} \cup \{b \in B\colon H(b) > 0\}$;
5: **для всех** $i \in \{1, 2\}$
6: исключаем из V все элементы $Y \in V$ такие, что $Y \subset X_i$;
7: **если** не существует элементов $Y \in V$ таких, что $X_i \subseteq Y$ **то**
8: полагаем $V = V \cup \{X_i\}$;
9: **для всех** пар $e = \{Z'_1, Z'_2\} \subset V\colon Z'_1 \cup Z'_2 = Z$
10: переопределяем множество $E = E \cup \{e\}$.

Множество решений системы (5) обозначим через $\mathfrak{S}(Z')$. При этом аффинно разделимое подмножество Z' называется *максимальным (по включению) аффинно разделимым подмножеством* множества Z, если для каждого $z \in Z \setminus Z'$ справедливо $\mathfrak{S}(Z' \cup \{z\}) = \varnothing$. Через $\mathfrak{M}(Z)$ обозначим множество всех максимальных аффинно разделимых подмножеств множества Z.

Определение 4. *Конечный граф $G_Z = (V, E)$ называется графом максимальных аффинно разделимых подмножеств множества Z, если $V = \mathfrak{M}(Z)$ и для каждой пары $\{Z'_1, Z'_2\} \subset V$,*

$$\{Z'_1, Z'_2\} \in E \iff Z'_1 \cup Z'_2 = Z.$$

Для каждого множества $Z = A \cup B \subset \mathbb{Q}^n$, $|Z| = m$, находящегося в общем положении, граф G_Z может быть построен при помощи алгоритма 1, корректность которого следует из принципа граничных решений для систем линейных неравенств. В частности, если множество Z аффинно разделимо, то построенный граф имеет вид $G_Z = (\{Z\}, \varnothing)$.

Алгоритм 2 содержит полиномиальный приближенный алгоритм для решения исследуемой задачи MASC-GP(n).

Предположение 1. Для аффинно неразделимого множества $Z = A \cup B$ и некоторого t существуют подмножества $Z'_0, \ldots, Z'_{2t} \in V$ (не обязательно попарно различные) такие, что $\{Z'_{2j-1}, Z'_{2j}\} \in E$ $(j \in \mathbb{N}_t)$ и для произвольных $(c_i, d_i) \in \mathfrak{S}(Z'_i)$, $i = 0, \ldots, 2t$, последовательность

$$K = (c_0^T x - d_0,\ c_1^T x - d_1, \ldots, c_{2t}^T x - d_{2t})$$

является минимальным аффинным разделяющим комитетом для множеств A и B.

Алгоритм 2. Жадный алгоритм.

1: строим граф $G_Z = (V, E)$;
2: **если** множество $V = \{Z\}$ **то**
определяем конечную последовательность $K = (Z)$, $q_{\min} = 1$ и переходим на шаг 12;
3: **иначе**
4: полагаем $q_{\min} = \infty$;
5: **для всех** вершины $\zeta \in V$
6: определяем конечную последовательность $K(\zeta)$, множество J и число $q(\zeta)$ равенствами: $K(\zeta) = (\zeta)$, $J = Z \setminus \zeta$, и $q(\zeta) = 1$;
7: **пока** $J \neq \varnothing$
8: находим $\{Z', Z''\} = \arg\max\{|X_1 \cap X_2 \cap J| : \{X_1, X_2\} \in E\}$;
9: добавляем множества Z' и Z'' к последовательности $K(\zeta)$, переопределяем $J = J \setminus (Z' \cap Z'')$ и $q(\zeta) = q(\zeta) + 2$;
10: **если** $q(\zeta) < q_{min}$ **то**
11: множество $K = K(\zeta)$ и $q_{min} = q(\zeta)$;
12: пусть $K = (Z'_1, \ldots, Z'_{q_{\min}})$;
13: **для всех** $i \in \mathbb{N}_{q_{\min}}$
14: определяем функцию $f_i(x) = c_i^T x - d_i$, где (c_i, d_i) — произвольный элемент $\mathfrak{S}(Z'_i)$;
15: Конечная последовательность $Q = (f_1, \ldots, f_{q_{\min}})$ образует искомый аффинный комитет, разделяющий множества A и B.

Условие предположения кажется слишком сильным, однако следует отметить, что примеры задачи MASC, используемые в известных доказательствах труднорешаемости данной задачи (в частности, в доказательстве теоремы 1), удовлетворяют ему.

Утверждение 1. *Пусть множество $Z = A \cup B \subset \mathbb{Q}^n$, $|Z| = m$, определяет частную постановку задачи* MASC-GP(n). *Тогда Жадный алгоритм обладает вычислительной сложностью $O\left(\binom{m}{n}^3 + \Theta m\right)$, где Θ — вычислительная сложность решения совместной системы из не более чем m линейных неравенств от $n+1$ переменной. При этом алгоритм имеет точность $O(m/n)$.*

Если для множества Z выполняется предположение 1, то алгоритм обладает точностью $O(\log m)$.

Max-SNP-трудность задач MinPC и MASC-GP(n)

Класс задач Max-SNP был впервые определен в совместной работе Х. Пападимитриу и М. Яннакакиса [6]. В этой же работе введено понятие Max-SNP-полноты и доказано, что задачи Max-3SAT и Max-3SAT(t)[1] являются Max-SNP-полными. В данном разделе мы покажем,

[1] Задача Max-3SAT(t) — это задача Max-3SAT при дополнительном условии, что каждая переменная может входить в булеву формулу не более t раз.

что задача о минимальном покрытии конечного множества точек плоскости множеством прямых (MinPC)[2] и задача MASC-GP(n) (при произвольном фиксированном $n > 1$) Max-SNP-трудны.

Теорема 2. *Существует схема полиномиального сведения задачи* Max-3SAT(t) *к задаче* MinPC, *преобразующая булеву формулу* φ *к частной постановке задачи* MinPC *так, что*

- *если* $OPT(\varphi) = m$, *то* $OPT(\mathrm{PC}) = nt$;
- *если* $OPT(\varphi) = m' < (1-\varepsilon)m$, *то* $OPT(\mathrm{PC}) > nt + \lceil \varepsilon n/6 \rceil$,

где $\varphi = E_1 \wedge \ldots \wedge E_m$ — *булева формула от* n *переменных,* $\varepsilon > 0$.

Поскольку задача Max-3SAT(t) Max-SNP-полна, то теорема доказывет, что задача MinPC Max-SNP-трудна.

Теорема 3. *Существует схема полиномиального сведения задачи* Max-3SAT(t) *к* MASC-GP(2), *преобразующая булеву формулу* φ *к частной постановке задачи* MASC-GP(2) *так, что*

- *если* $OPT(\varphi) = m$, *то* $OPT\big(\text{MASC-GP}(2)\big) = 2nt + 1$,
- *если* $OPT(\varphi) = m' < (1-\varepsilon)m$, *то* $OPT\big(\text{MASC-GP}(2)\big) > 2nt + \lceil \varepsilon n/3 \rceil + 1$,

где $\varphi = E_1 \wedge \ldots \wedge E_m$ — *булева формула от* n *переменных,* $\varepsilon > 0$.

Последняя теорема доказывает Max-SNP-трудность задачи MASC-GP(2), а следовательно, и задачи MASC-GP(n) при произвольном $n > 1$.

Принадлежность классу Max-SNP-трудных задач MinPC и MASC-GP(n) влечет невозможность построения для них полиномиальной аппроксимационной схемы, если P $\neq$ NP.

Емкость класса комитетных кусочно-линейных решающих правил

Обозначим через $\mathcal{F}_q$ класс аффинных комитетных решающих правил, состоящих из не более чем q функций. Справедлива следующая теорема, дающая верхнюю и нижнюю оценки емкости данного класса.

Теорема 4. *Емкость класса* $\mathcal{F}_q$ *комитетных решающих правил, состоящих из не более чем* q *аффинных функций, удовлетворяет соотношениям*

$$q \leqslant 2\left\lfloor \frac{\lfloor (\mathrm{VCD}(\mathcal{F}_q) - n + 1)/2 \rfloor}{n} \right\rfloor;$$
$$\mathrm{VCD}(\mathcal{F}_q) \leqslant q(n+1)$$

и, следовательно, $\mathrm{VCD}(\mathcal{F}_q) = O(qn)$.

[2] MinPC — оптимизационная версия задачи PC.

Таким образом, из теоремы следует, что задача минимизации эмпирического риска в классе аффинных комитетных решающих правил, состоящих из не более чем q функций, эквивалентна задаче о минимальном аффинном разделяющем комитете (MASC).

Выводы

В статье показано, что задача о минимальном аффинном разделяющем комитете, сформулированная в пространстве фиксированной размерности, большей единицы, при дополнительном условии общности положения разделяемых множеств (MASC-GP(n)) является NP-трудной (в сильном смысле). Представлен полиномиальный приближенный алгоритм для ее решения, обладающий в общем случае точностью аппроксимации $O(m/n)$, а при справедливости некоторого естественнго предположения — точностью $O(\log m)$.

Следующий круг вопросов, рассмотренный в статье, посвящен классу задач Max-SNP и обоснованию Max-SNP-трудности задач MinPC и MASC-GP(n), откуда следует невозможность построения для них полиномиальной аппроксимационной схемы, если P $\neq$ NP.

В завершении сформулирована теорема, содержащая верхнюю и нижнюю оценки емкости класса комитетных решающих правил с ограниченным числом элементов, совпадающие по порядку величины, что подтверждает эквивалентность задачи минимизации эмпирического риска в классе аффинных комитетных решающих правил, состоящих из не более чем q функций, задаче MASC.

Литература

[1] *Вапник В. Н., Червоненкис А. Я.* Теория распознавания образов. — Москва: Наука, 1974. — 416 с.

[2] *Мазуров Вл. Д.* Комитеты систем неравенств и задача распознавания // Кибернетикаю.— 1971. — № 3. — С. 140–146.

[3] *Хачай М. Ю.* О вычислительной и аппроксимационной сложности задачи о минимальном аффинном разделяющем комитете // Таврический вестник информатики и математики, — 2006. — № 1. — С. 34–43.

[4] *Khachai M. Yu.* Computational and Approximational Complexity of Combinatorial Problems Related to the Committee Polyhedral Separability of Finite Sets // Pattern Recognition and Image Analysis. — 2008. — V. 18, N. 2. — Pp. 237–242.

[5] *Megiddo N., Tamir A.* On the complexity of locating linear facilities in the plane // Operations research letters. — 1982. — V. 1, N. 5. — Pp. 194–197.

[6] *Papadimitriou C., Yannakakis M.* Optimization, approximation, and complexity classes // J. Comput. System Sci. — 1991. — V. 43, N. 3. — Pp. 425–440.

Об одном методе редукции выборки для задачи обучения в классе комитетных решающих правил*

Кобылкин К. С.
kobylkinks@gmail.com
Екатеринбург, Институт математики и механики УрО РАН

Рассматривается метод сведения задачи обучения в классе комитетных кусочно-линейных решающих правил с выборками A и B в $\mathbb{R}^n$ из двух классов к той же задаче, имеющей разделяемые выборки меньшей мощности. Его вычислительная сложность является полиномом по $|A \cup B|$, но зависит экспоненциально от n. Приводятся результаты вычислительного эксперимента, показывающего эффективность метода на одном классе задач.

Нахождение минимального комитета аффинных функций, разделяющего два конечных множества точек A и B в $\mathbb{R}^n$, $n > 1$, является интересной задачей комбинаторной оптимизации, связанной [1] с задачей обучения в классе коллективных решающих правил на основе голосования большинством голосов, а также с задачей обучения простейшего двухслойного персептрона. Доказана труднорешаемость этой задачи [2] в случае конечных множеств A и B в $\mathbb{R}^2$, находящихся в общем положении.

В работе исследуется эффективность одного метода редукции задачи поиска комитета q аффинных функций, разделяющего множества A и B, где $\operatorname{conv} A \cap \operatorname{conv} B \neq \varnothing$, к той же задаче с множествами A' и B', причем $m' = |A' \cup B'| \leqslant m$. Метод основан на следующем понятии сходства: точки c и c' из $A \cup B$ близки, если они входят в одни и те же *максимальные по включению аффинно разделимые подсистемы* (МРП) системы точек $A \cup B$. Сложность этого метода равна $O\left(C_m^{n+1}(mn^3 + m^2) + m\Theta\right)$, что меньше сложности приближенного «жадного» метода поиска минимального разделяющего комитета аффинных функций [2] с оценками точности $O\left(\frac{m}{n}\right)$ и сложности $O((C_m^n)^3 + \Theta m)$, где $m = |A \cup B|$, а Θ — сложность решения совместной системы не более, чем m строгих однородных линейных неравенств в $\mathbb{R}^{n+1}$.

Для оценки эффективности метода для задач комитетной отделимости двух конечных множеств в $\mathbb{R}^2$ был проведен вычислительный эксперимент. В нем задачи отделимости задавались системами 100 строгих однородных линейных неравенств в $\mathbb{R}^3$. Векторы коэффициентов для этих систем выбирались из трехмерных гауссовых распределений с центрами в 10 случайно выбранных точках на единичной сфере S с центром в 0 и ковариационной матрицей $\sigma^2 I$, где $\sigma = 0{,}1$ и $0{,}2$. Из таких систем метод в среднем удалял $10 - 15\%$ и $1 - 2\%$ неравенств соответственно, при этом число *максимальных по включению совместных подсистем* (МСП) в этих системах не уменьшалось. Во второй части эксперимента векторы коэффициентов 100 неравенств системы выбирались из равномерного распределения на S, при этом метод не удалил ни одного неравенства для 19 систем из 20.

Кроме того, при $n = 2$ исследовалась эффективность метода редукции для систем строгих неоднородных линейных неравенств $\mathbb{R}^n$, использующего похожее понятие сходства: i-е неравенство системы близко к j-му, если всякая МСП этой системы, включающая j-е неравенство, содержит также i-е неравенство. Такие системы возникают при добавлении к системе строгих однородных линейных неравенств в $\mathbb{R}^{n+1}$ ограничения $x \in H$, где H — гиперплоскость в $\mathbb{R}^{n+1}$, причем $0 \notin H$. Сложность метода равна $O\left(C_m^n(mn^3 + m^2) + m\Theta\right)$, где Θ — сложность решения системы не более, чем m линейных неравенств в $\mathbb{R}^n$, что меньше сложности аналога «жадного» алгоритма для систем линейных неоднородных неравенств [3]. Эксперимент с 10 системами 100 линейных неравенств в $\mathbb{R}^2$, коэффициенты которых выбирались из равномерного распределения на отрезке $[-1, 1]$ показал, что число неравенств в таких системах в среднем сокращается на 60%, а число МСП на 80%.

Определения и постановка задачи

Пусть $\mathbb{R}^n$ — n-мерное евклидово пространство, а $N_m = \{1, \ldots, m\}$. Под *аффинной* функцией понимается функция $f : \mathbb{R}^n \to \mathbb{R}$ вида $f(x) = c^{\intercal}x + d$, где $c \in \mathbb{R}^n$, а $d \in \mathbb{R}$.

Определение 1. *Комитетом [4] аффинных функций, разделяющим два конечных множества A и B в $\mathbb{R}^n$, называется такая конечная последовательность $Q = (f_1, \ldots, f_q)$ (с возможными повторениями) аффинных функций, что во всякой точке $a \in A$ (соответственно, в любой точке $b \in B$) более, чем $q/2$ аффинных функций этой последовательности (с учетом повторений) принимают положительное (соответственно, отрицательное) значение. При этом длина q последовательности Q называется числом элементов комитета.*

Комитет с наименьшим для данных множеств A и B числом функций называется *минимальным*.

Работа выполнена при финансовой поддержке РФФИ (проекты №10-01-00273, 10-07-00134), Президиума УрО РАН (проекты №09-П-1-1001, 09-С-1-1010, 10-1-НП-367).

Задача 1. Найти комитет из q аффинных функций (соответственно, минимальный комитет), разделяющий два подмножества $A = \{a_1, \ldots, a_{m_1}\}$ и $B = \{b_1, \ldots, b_{m_2}\}$ в $\mathbb{R}^n$, имеющих пересекающиеся выпуклые оболочки, где $q \in \mathbb{N}$ — заданное число.

Определение 2. *Комитетом [4] несовместной системы строгих линейных неравенств*

$$c_j^\intercal x > b_j, \quad j = 1, \ldots, m,\ c_j, x \in \mathbb{R}^n,\ b_j \in \mathbb{R}, \quad (1)$$

называется такая конечная последовательность $Q = (x_1, \ldots, x_q)$ (с возможными повторениями) векторов из $\mathbb{R}^n$, что каждому неравенству этой системы удовлетворяет более половины элементов в Q с учетом кратности.

Аналогично дается определение минимального комитета системы (1).

Задача 2. Найти комитет из q элементов (соответственно, минимальный комитет) системы (1).

Справедливо следующее утверждение [4]:

Утверждение 1. *Задача* 1 *эквивалентна задаче* 2 *для системы строгих однородных линейных неравенств в $\mathbb{R}^{n+1}$*

$$c_j^\intercal x > 0, \quad j = 1, \ldots, m, \quad (2)$$

в которой $c_j = [a_j, 1],\ j = 1, \ldots, m_1$, а также $c_{m_1+j} = [-b_j, -1],\ j = 1, \ldots, m_2,\ m = m_1 + m_2$.

Описание метода

Сформулируем метод в общем виде для системы строгих линейных неравенств

$$c_j^\intercal x > b_j, \quad j = 1, \ldots, m,\ c_j, x \in \mathbb{R}^n,\ b_j \in \mathbb{R}, \quad (3)$$

в которой любая подсистема из n неравенств имеет ранг n и $m > n$.

Определение 3. *Назовем i-е неравенство системы* (3) *несущественным в ней по отношению к j-му неравенству, если всякая ее МСП, включающая j-е неравенство, содержит также i-е неравенство. Система* (3) *называется несократимой, если она не содержит несущественных в ней неравенств.*

Утверждение 2. *Для того чтобы i-е неравенство системы* (3) *было несущественно в ней по отношению к j-му неравенству, необходимо и достаточно, чтобы были выполнены следующие два условия:*

1. *совместность произвольной ее подсистемы $\mathcal{J}$ мощности $n+1$, включающей j-е неравенство, влечет за собой совместность подсистемы той же мощности, полученной из $\mathcal{J}$ удалением j-го и добавлением i-го неравенства;*
2. *совместность всякой подсистемы системы* (3) *мощности $n+1$, включающей i-е и j-е неравенства.*

Алгоритм 1. Прямой ход метода: редукция.

Вход: векторы $\{c_j\}_{j=1}^m$ и коэффициенты $\{b_j\}_{j=1}^m$.
Выход:
1. индексное множество J' некоторой подсистемы системы (3);
2. множества $S(j)$, состоящие из номеров неравенств, исключаемых из (3) как несущественных по отношению к j-му неравенству, $j = 1, \ldots, m$;

для $j = 1, \ldots, m$
  $S(j) := \varnothing$;
для всех $i, j \in N_m$, таких, что $i \neq j$
  $z_{ij} := 0$
$l := 0$;
// инициализация завершена
для всех $J \subset N_m$, таких, что $|J| = n$
  $J_+ = J;\ J_- = \varnothing$;
  для $j \in N_m \backslash J$
    если FEASIBLE$(J \cup \{j\})$ **то**
      $J_+ := J_+ \cup \{j\}$;
    иначе
      $J_- := J_- \cup \{j\}$;
  если $J_- \neq \varnothing$ **то**
    $l := l + 1$;
    для всех $i, j \in N_m$, таких, что $i \neq j$
      если $i \in J_-$ и $j \in J_+$ **то**
        $v_{ijl} := 1;\ z_{ij} := z_{ij} + 1$;
      иначе
        $v_{ijl} := 0$;
    для $j = 1, \ldots, m$
      если $j \in J$ **то**
        $w_{jl} := 1$;
      иначе
        $w_{jl} := 0$;
$J' := N_m;\ U_0 := N_l$;
повторять
  $I := \varnothing$;
  для $i = 1, \ldots, m$
    для $j = 1, \ldots, m,\ j \neq i$
      если $z_{ij} = 0$ и $i, j \in J'$ **то**
        $J' := J' \backslash \{i\};\ I := I \cup \{i\}$;
        $S(j) := S(j) \cup \{i\}$;
  если $I \neq \varnothing$ **то**
    $U := \bigcup\limits_{i \in I} \{t \in U_0 : w_{it} = 1\};\ U_0 := U_0 \backslash U$;
    для всех $i, j \in J'$, таких, что $i \neq j$
      $z_{ij} := z_{ij} - \sum\limits_{t \in U} v_{ijt}$;
пока $I \neq \varnothing$;
return J'.

Утверждение 3. *Пусть в системе* (3) $b_j = 0$ *для любого $j \in N_m$. Для того чтобы i-е неравенство системы* (3) *было несущественно в ней по отношению к j-му неравенству, необходимо и достаточно, чтобы была совместна всякая ее подсистема мощности $n+1$, включающая i-е и j-е неравенства.*

Алгоритм 2. Обратный ход метода: комитет.

Вход:
1. индексное множество J' подсистемы системы (3), полученной в результате прямого хода метода для (3);
2. множества $S(j)$, $j = 1, \ldots, m$, также являющиеся результатом выполнения прямого хода метода;
3. комитет $Q' = (x'_1, \ldots, x'_q)$ из q элементов подсистемы с индексным множеством J' системы (3);

Выход: комитет $Q = (x_1, \ldots, x_q)$ для (3);

для $i = 1, \ldots, q$
  $J_i := \{j \in J' : c_j^{\intercal} x_i > b_j\}$; $J := \varnothing$;
  для всех $j \in J_i$
    $J := J \cup \mathrm{ADD}\big(j, \{S(p)\}\big)$;
  $J_i := J_i \cup J$;
  $x_i := \mathrm{SOLVE}(J_i)$;
return $Q = (x_1, \ldots, x_q)$.

Предлагаемый метод состоит из двух этапов. На первом этапе, называемом *прямым ходом* метода (см. алгоритм 1), в результате сокращения произвольной системы (3) получается некоторая ее подсистема $\mathcal{J}'$, являющаяся несократимой системой. На втором этапе, который называется *обратным ходом* метода (алгоритм 2), в предположении, что существует и известен некоторый комитет из q элементов системы $\mathcal{J}'$, строится комитет из q элементов системы (3), где $q \in \mathbb{N}$.

В описании метода используются три процедуры FEASIBLE, SOLVE и ADD. Первая из них выдает сообщение `TRUE` (соответственно, `FALSE`) о совместности (соответственно, несовместности) произвольной подсистемы из $n + 1$ неравенства системы (3), заданной своим *индексным множеством*, под которым понимается множество номеров входящих в эту подсистему неравенств из (3). Вторая процедура выдает некоторое решение произвольной совместной подсистемы системы (3), получив на входе индексное множество этой подсистемы. Описание третьей процедуры дано в алгоритме 3.

Замечание 1. Процедура FEASIBLE реализуема на основе критерия Школьника [5] за время $O(n^3)$.

Пусть в результате выполнения прямого хода метода для системы (3) получено индексное множество J' некоторой ее подсистемы, а также множества $S(j)$, $j = 1, \ldots, m$. Предположим, что для этой подсистемы каким-либо из методов [3, 4] (методом [2] в случае однородной системы (3)) найден комитет Q' из q элементов. Поскольку неравенства, входящие в (3), являются строгими, можно считать, что члены Q' попарно различны. В алгоритме 2 дается описание второго этапа метода.

Теорема 1. *Справедливы утверждения:*

Алгоритм 3. Процедура ADD.

Вход:
1. $i \in N_m$;
2. множества $S(j)$, $j = 1, \ldots, m$;

Выход: подмножество $J \subset N_m$;

инициализация: $J := S(i)$;
для всех $j \in S(i)$
  $J := J \cup \mathrm{ADD}\big(j, \{S(p)\}\big)$;
return J.

1. *подсистема с индексным множеством J' системы (3), полученная при прямом ходе метода для (3), является несократимой системой;*
2. *последовательность Q, выдаваемая в результате обратного хода метода, является комитетом из q элементов системы (3);*
3. *суммарная сложность прямого и обратного хода метода равна $O\left(C_m^n(mn^3 + m^2) + m\Theta\right)$, где Θ — сложность нахождения решения совместной системы не более, чем m линейных неравенств в $\mathbb{R}^n$.*

Следствие 1. *Если существует комитет системы (3), то система (3) и ее подсистема с индексным множеством J' имеют одинаковое число членов минимального комитета.*

Вычислительный эксперимент

Для исследования эффективности предложенного метода редукции был проведен вычислительный эксперимент для задач комитетной отделимости на плоскости. При этом задачи генерировались в форме систем 100 строгих однородных линейных неравенств в $\mathbb{R}^3$. Векторы коэффициентов первых 10 неравенств каждой системы выбирались случайно из равномерного распределения на единичной сфере S с центром в 0. Далее из трехмерных гауссовых распределений с центром в каждой из этих 10 точек, и ковариационной матрицей $\sigma^2 I$, выбиралось еще 9 векторов, где $\sigma = 0{,}1$ и $0{,}2$. Результаты эксперимента показаны в табл. 1 и 2 соответственно. Здесь m (соответственно, m') обозначает число неравенств, m_0 (соответственно, m'_0) — мощность наибольшей по числу неравенств МСП, k_0 (соответственно, k'_0) — число таких МСП, а k (соответственно, k') — число всех МСП системы (соответственно, ее несократимой подсистемы, полученной в результате прямого хода метода).

При увеличении σ процент исключаемых неравенств уменьшается с $10{-}15\%$ до $1{-}2\%$. Сравнивая последние два столбца в табл. 1 и 2, можно видеть, что исходная система и ее несократимая подсистема имеют одинаковое число МСП. Таким образом, задача поиска минимального комитета несократимой подсистемы не становится существенно более простой по сравнению с таковой для исходной си-

Таблица 3. Параметры случайных неоднородных систем в $\mathbb{R}^2$ и их несократимых подсистем.

$m = 100$						
m'	m_0	m'_0	k_0	k'_0	k	k'
39	62	20	1	39	424	64
38	62	21	1	1	409	62
43	59	22	1	44	408	79
47	56	25	5	1	405	108
42	64	23	3	3	400	94
42	64	22	1	22	371	76
37	65	20	3	1	392	80
38	63	20	2	20	385	72
50	67	27	1	4	375	130
30	68	16	5	17	343	44

Таблица 1. Параметры случайных однородных систем в $\mathbb{R}^3$ и их несократимых подсистем для эксперимента с $\sigma = 0{,}1$.

$m = 100$						
m'	m_0	m'_0	k_0	k'_0	k	k'
37	91	28	1	1	76	76
86	77	63	3	3	231	231
81	78	59	2	2	325	325
76	81	57	1	1	230	230
88	78	66	2	2	321	321
9	98	7	1	2	10	10
88	80	68	2	2	317	317
45	90	36	1	1	66	66
88	85	73	1	1	271	271
97	76	73	2	2	328	328

Таблица 2. Параметры однородных систем в $\mathbb{R}^3$ и их несократимых подсистем для эксперимента с $\sigma = 0{,}2$.

$m = 100$						
m'	m_0	m'_0	k_0	k'_0	k	k'
100	70	70	1	1	566	566
94	80	74	1	1	328	328
94	86	80	3	3	265	265
100	79	79	1	1	418	418
100	82	82	1	1	411	411
87	84	71	1	1	262	262
99	76	75	1	1	526	526
88	89	77	1	1	198	198
98	81	79	1	1	415	415
100	78	78	1	1	444	444

стемы в смысле сложности «жадного» алгоритма поиска комитета [2].

Во второй части эксперимента векторы коэффициентов 100 неравенств однородной системы выбирались случайно из равномерного распределения на сфере S. В 19 из 20 случаев системы оказывались несократимыми, мощность m_0 их наибольших МСП лежала в диапазоне от 60 до 65, а общее число k МСП — в диапазоне от 700 до 800, что больше числа k для систем, описанных в табл. 1 и 2.

Кроме того, вычислительный эксперимент проводился для оценки производительности метода редукции в случае системы неоднородных линейных неравенств. В табл. 3 даны результаты эксперимента для систем неоднородных линейных неравенств в $\mathbb{R}^2$, коэффициенты которых выбраны случайно из равномерного распределения на отрезке $[-1, 1]$. В этом случае, в среднем на 60% уменьшается число неравенств, а также на 80% сокращается число МСП. Тем самым уменьшается сложность алгоритма поиска комитета, данного в [3], зависящая полиномиально от числа МСП.

Выводы

Предложен метод редукции несовместной системы линейных неравенств, основанный на одном понятии сходства неравенств, и сводящий задачу поиска ее комитета к той же задаче меньшего размера. Вычислительный эксперимент показал, с одной стороны, относительно небольшую эффективность его непосредственного применения к задачам комитетной отделимости; с другой стороны, его эффективность оказалась неплохой для одного класса задач размерности $n = 2$.

Интересны обобщения понятия близости точек, например, на основе ослабления требований критерия из утверждения 3, с целью отыскания точек «сгущения» обучающей выборки.

Литература

[1] *Хачай М. Ю.* О длине обучающей выборки для комитетного решающего правила // Искусственный интеллект. — 2000. — № 2. — С. 219–223.

[2] *Khachay M. Yu., Poberii M. I.* Complexity and Approximability of Committee Polyhedral Separability of Sets in General Position // Informatica. — 2009. — V. 20, N. 2. — P. 217–234.

[3] *Hachai M. Yu., Rybin A. I.* A New Estimate of the Number of Members in a Minimum Committee of a Linear Inequalities System // Pattern Recognition and Image Analysis. — 1998. — V. 8, N. 4. — P. 491–496.

[4] *Мазуров Вл. Д.* Метод комитетов в задачах оптимизации и классификации. — М.: Наука, 1990. — 248 с.

[5] *Черников С. Н.* Линейные неравенства. — М.: Наука, 1968. — 488 с.

Сложность задачи отбора эталонов в методе ближайшего соседа*

Зухба А. В.
a__l@mail.ru
Московский физико-технический институт (государственный университет)

Задача отбора эталонов заключается в том, чтобы выделить в обучающей выборке подмножество объектов минимальной мощности, обеспечивающее оптимальное или хотя бы близкое к оптимальному значение заданного функционала качества классификации. В данной работе рассматриваются функционалы частоты ошибок и полного скользящего контроля для метода ближайшего соседа. Показывается, что задача отбора эталонов является NP-полной. Выводятся верхние оценки минимального числа эталонов для задач, удовлетворяющих гипотезе компактности.

Введение

Рассматривается задача классификации в стандартной постановке. На множестве объектов задана функция расстояния. Метод ближайшего соседа запоминает обучающую выборку и строит алгоритм классификации, который относит произвольный объект к тому классу, которому принадлежит ближайший обучающий объект. *Задача отбора эталонных объектов* заключается в том, чтобы отобрать как можно меньше обучающих объектов, достаточных для надежной классификации. Отбор эталонов преследует сразу несколько целей: сокращение объёма хранимых данных, повышение скорости классификации, выделение нетипичных объектов и повышение качества классификации.

Методы последовательного (жадного) добавления эталонов Stolp и FRiS-Stolp [5, 2] неплохо зарекомендовали себя на практике. Однако они оставляют открытыми теоретические вопросы: минимизируют ли они какой-либо функционал, почему они обладают хорошей обобщающей способностью, и почему эвристические жадные алгоритмы неплохо справляются с задачей отбора эталонов.

В [4, 9] предложены алгоритмы отбора эталонов, минимизирующие функционал полного скользящего контроля CCV [10], который характеризует обобщающую способность метода обучения. Хотя задача ставится как оптимизационная, для её решения опять-таки предлагаются эвристические жадные алгоритмы последовательного добавления или удаления эталонов, приводящие к результатам, сопоставимым с FRiS-Stolp.

В [11] показано, что задача отбора эталонов путём минимизации CCV является NP-полной для обоих «естественных» вариантов определения CCV по отношению к вхождению эталонов в контрольные подвыборки. Доказательство основано на сведении известной NP-полной задачи о вершинном покрытии графа к задаче отбора эталонов в некоторой специально построенной выборке.

Работа поддержана РФФИ (проект №11-07-00480) и программой ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

В данной работе рассматривается ещё одна NP-полная задача — о минимальном покрытии множества системой его подмножеств, которая также сводится к задаче отбора эталонов. При этом упрощается конструкция выборки и выводится верхняя оценка числа эталонов для случая «достаточно хороших» выборок, удовлетворяющих гипотезе компактности.

Определения и обозначения

Задано множество объектов $\mathbb{X} = \{x_1, \ldots, x_L\}$, конечное множество классов Y, и существует целевая функция $y\colon X \to Y$. Обозначим $y_i = y(x_i)$, $i = 1, \ldots, L$. Задача обучения по прецедентам состоит в том, чтобы по заданной обучающей выборке $X \subset \mathbb{X}$ построить алгоритм классификации — функцию $a\colon \mathbb{X} \to Y$, которая приближала бы целевую функцию $y(x)$ на всём множестве $\mathbb{X}$.

Методом обучения называется отображение μ, которое произвольной выборке $X \subset \mathbb{X}$ ставит в соответствие некоторый алгоритм $a\colon \mathbb{X} \to Y$.

Индикатором ошибки алгоритма a на объекте x_i называется функция $I(x_i, a) = \big[y(x_i) \neq a(x_i)\big]$.

Частота ошибок алгоритма на выборке X определяется как $\nu(a, X) = \frac{1}{|X|} \sum\limits_{x \in X} I(x, a)$.

Малая частота ошибок на обучающей выборке X еще не гарантирует, что построенный алгоритм будет столь же редко ошибаться на новых (контрольных) объектах $\bar{X} = \mathbb{X} \setminus X$. Для оценивания обобщающей способности метода обучения вводятся функционалы качества, основанные на принципе скользящего контроля [3]. Рассматривается множество всех C_L^ℓ разбиений множества объектов $\mathbb{X} = X \sqcup \bar{X}$ на две выборки — наблюдаемую обучающую X длины ℓ и скрытую контрольную $\bar{X}$ длины $k = L - \ell$. Если предположить, что все разбиения реализуются с равными вероятностями, то функционал полного скользящего контроля (complete cross validation, CCV) [10] определяется как математическое ожидание частоты ошибок на контрольной подвыборке, следовательно, характеризует обобщающую способность метода обучения:

$$Q(\mu, \mathbb{X}) = \frac{1}{C_L^\ell} \sum_X \nu(\mu(X), \bar{X}).$$

При при $k = 1$ функционал CCV переходит в другой известный функционал — скользящий контроль с одним отделяемым объектом (leave-one-out, LOO). Однако существенное ограничение на длину контрольной выборки и малое число разбиений $C_L^\ell = L$ приводят к тому, что LOO оказывается не достаточно надёжной характеристикой качества обучения со слишком большой дисперсией.

Оптимизационные постановки задачи отбора эталонов

Пусть $\rho(x, x')$ — функция расстояния на множестве $\mathbb{X}$, вообще говоря, не обязательно метрика. Для произвольного $x_i \in \mathbb{X}$ положим $x_i \equiv x_{i0}$ и обозначим через $x_{i1}, \ldots, x_{i,L-1}$ последовательность всех объектов выборки $\mathbb{X}$, упорядоченную по возрастанию расстояний $\rho(x_i, x_{im})$, $m = 1, \ldots, L-1$.

Профилем компактности выборки $\mathbb{X}$ называется функция $P(m)$, выражающая долю объектов, у которых m-й сосед лежит в другом классе [3]:

$$P(m) = \frac{1}{L}\sum_{i=1}^{L}\big[y(x_{im}) \neq y_i\big]; \quad m = 1, \ldots, L-1.$$

Следующая теорема позволяет эффективно вычислять CCV для метода ближайшего соседа [3].

Теорема 1. *Для метода ближайшего соседа μ справедлива формула*

$$Q(\mu, \mathbb{X}) = \sum_{m=1}^{k} P(m)\frac{C_{L-1-m}^{\ell-1}}{C_{L-1}^{\ell}}. \qquad (1)$$

Отношение биномиальных коэффициентов в (1) убывает экспоненциально по m, следовательно, только при малых m значения $P(m)$ могут давать существенный вклад в CCV. Таким образом, задача классификации является «хорошей», если близкие объекты часто лежат в одном классе. Это условие обычно вводится неформально и называется *гипотезой компактности* [1, 5]. Теорема 1 формально связывает количественную оценку обобщающей способности (CCV) с количественной характеристикой (профилем) компактности выборки.

Теперь рассмотрим более сложный метод обучения μ_Ω, который запоминает не всю обучающую выборку, а лишь подмножество эталонных объектов $\Omega \subseteq \mathbb{X}$. На стадии классификации используется тот же алгоритм ближайшего соседа, но теперь ближайшие соседи выбираются только из Ω. Обозначим этот алгоритм через a_Ω.

Простым естественным критерием для отбора эталонов Ω является минимизация частоты ошибок на всех остальных (неэталонных) объектах:

$$\nu(a_\Omega, \mathbb{X}\backslash\Omega) \to \min_{\Omega}. \qquad (2)$$

Не менее естественной представляется минимизация функционала $Q(\mu_\Omega, \mathbb{X})$, характеризующего обобщающую способность метода μ_Ω. При этом возникают два возможных варианта определения CCV в зависимости от того, позволять ли эталонам входить в контрольную выборку.

Вариант 1. Будем разбивать $\mathbb{X}$ на X и $\bar{X}$ так, чтобы множество эталонных объектов Ω всегда находилось целиком в обучении X. Число таких разбиений $C_{L-|\Omega|}^{k}$. Обозначим модифицированный таким образом функционал CCV через $Q^*(\mu_\Omega, \mathbb{X})$.

Теорема 2. *Если $\mu_\Omega(X) = a_\Omega$ для любой обучающей выборки X, содержащей Ω, то модифицированный функционал CCV совпадает с частотой ошибок на неэталонных объектах* (2)*:*

$$Q^*(\mu_\Omega, \mathbb{X}) = \frac{1}{C_{L-|\Omega|}^{k}}\sum_{X:\ \Omega\subset X}\frac{1}{k}\sum_{x\in\bar{X}} I(x_i, a_\Omega) = \nu(a_\Omega, \mathbb{X}\backslash\Omega).$$

Вариант 2. Теперь разрешим эталонным объектам из Ω попадать как в обучение, так и в контроль. Потребуем, чтобы длина контрольной выборки была меньше числа эталонных объектов, чтобы гарантировать $X \cap \Omega \neq \varnothing$.

Введём обозначение: $x^O mega_{im}$ — m-ый объект из множества эталонов Ω, если упорядочить их по возрастанию расстояний до объекта x_i. Обратим внимание, что если $x_i \in \Omega$, то $m = 1, \ldots, |\Omega| - 1$, а если $x_i \in \mathbb{X}\backslash\Omega$, то $m = 1, \ldots, |\Omega|$.

Профилем Ω-компактности выборки $\mathbb{X}$ называется функция $P^\Omega(m)$, выражающая долю объектов, у которых m-й сосед из множества Ω лежит в другом классе [11]:

$$P^\Omega(m) = \frac{1}{L}\sum_{i=1}^{L}\big[y(x_{im}^\Omega) \neq y_i\big]; \quad m = 1, \ldots, L-1.$$

Теорема 3. *Для метода ближайшего соседа μ_Ω со множеством эталонов Ω справедлива формула*

$$Q(\mu_\Omega, \mathbb{X}) = \sum_{m=1}^{k} P^\Omega(m)\frac{C_{L-1-m}^{\ell-1}}{C_{L-1}^{\ell}}.$$

Далее рассматривается вычислительная сложность задач минимизации функционалов $Q^*(\mu_\Omega, \mathbb{X})$ и $Q(\mu_\Omega, \mathbb{X})$ по множеству эталонов Ω при дополнительном условии минимизации числа эталонов.

NP-полнота задачи отбора минимального множества эталонов

Покажем связь некоторых NP-полных задач с задачами отбора эталонов по критериям минимума функционалов $Q^*(\mu_\Omega, \mathbb{X})$ и $Q(\mu_\Omega, \mathbb{X})$.

Задача о минимальном вершинном покрытии графа. *Вершинным покрытием* неориентированного графа $G = (V, E)$ называется подмножество его вершин S, такое, что у каждого ребра графа хотя бы один из концов принадлежит S. Задача о минимальном вершинном покрытии состоит в том, чтобы для заданного графа G найти вершинное покрытие минимальной мощности $|S|$.

Данная задача является NP-полной [6].

Теорема 4. *Задача поиска минимального вершинного покрытия произвольного графа G сводится к задаче выбора из некоторой искусственной выборки $\mathbb{X}_G$ множества эталонов Ω минимальной мощности, по которому классификация алгоритмом ближайшего соседа a_Ω даёт $\nu(a_\Omega, \mathbb{X}_G \backslash \Omega) = 0$. Причем выборка $\mathbb{X}_G$ строится по G за полиномиальное время и имеет полиномиальное количество объектов относительно $|V| + |E|$.*

Следствие 1. *Задача поиска множества эталонов Ω минимальной мощности, при котором $\nu(a_\Omega, \mathbb{X} \backslash \Omega) = 0$, является NP-полной.*

Заметим, что задача поиска множества эталонов Ω, не обязательно минимальной мощности и дающего $\nu(a_\Omega, \mathbb{X}) = 0$, не является NP-полной, поскольку имеет тривиальное решение $\Omega = \mathbb{X}$.

Теорема 5. *Задача поиска минимального вершинного покрытия произвольного графа G сводится к задаче выбора из некоторой искусственной выборки $\mathbb{X}_G$ множества эталонов Ω, $|\Omega| \geqslant k + 1$, минимизирующего функционал $Q(\mu_\Omega, \mathbb{X}_G)$. Причем выборка $\mathbb{X}_G$ строится по G за полиномиальное время и имеет полиномиальное количество объектов относительно $|V| + |E|$.*

Следствие 2. *Для любого k задача поиска множества эталонов Ω, $|\Omega| \geqslant k + 1$, минимизирующего функционал $Q(\mu_\Omega, \mathbb{X})$, является NP-полной.*

Задача о минимальном покрытии множества системой его подмножеств. Пусть U — конечное множество, F — семейство его подмножеств. *Покрытием множества U подмножествами* называют множество подмножеств $C \subset F$, объединение которых совпадает с U.

Задача поиска покрытия минимальной мощности является NP-полной [6].

Теорема 6. *Задача поиска минимального покрытия $C \subseteq F$ множества U сводится к задаче выбора из некоторой искусственной выборки $\mathbb{X}_{U,F}$ множества эталонов Ω минимальной мощности, по которому классификация алгоритмом ближайшего соседа a_Ω даёт $\nu(a_\Omega, \mathbb{X}_{U,F} \backslash \Omega) = 0$. Причем выборка $\mathbb{X}_{U,F}$ строится по задаче о покрытии множества за полиномиальное время и имеет полиномиальное количество объектов относительно $|U| + |F|$.*

Теорема 7. *Задача поиска минимального покрытия $C \subseteq F$ множества U сводится к задаче выбора из некоторой искусственной выборки $\mathbb{X}_{U,F}$ множества эталонов Ω, такого, что в Ω входит хотя бы по одному объекту из каждого класса, минимизирующего функционал $Q(\mu_\Omega, \mathbb{X}_{U,F})$. Причем выборка $\mathbb{X}_{U,F}$ строится по U и F за полиномиальное время и имеет полиномиальное количество объектов относительно $|U| + |F|$.*

Заметим, что условие $|\Omega| \geqslant k + 1$ действительно необходимо, иначе функционал Q невозможно будет записать в виде, указанном в теореме 3.

NP-полнота этих задач оправдывает применение приближённых эвристических алгоритмов, выбирающих Ω так, чтобы минимизируемый функционал принимал значение, близкое к минимальному, и/или мощность множества Ω была близка к минимальной.

Приближенные жадные алгоритмы

Определение 1. *Выборка $\mathbb{X}$ называется k-идеальной, если все профили компактности до k включительно равны нулю (следовательно, $Q(\mu, \mathbb{X}) = 0$).*

Для любого объекта ближайший к нему эталон может многократно меняться в процессе жадного добавления или удаления эталонов. Однако если выбрать эталоны так, чтобы для каждого объекта нашлось не менее k эталонов, находящихся ближе, чем ближайший объект чужого класса, то это гарантировало бы, что $Q(\mu_\Omega, \mathbb{X}) = 0$. Будем называть минимальное по мощности множество эталонов, обладающее данным свойством, *оптимальным*.

Задача построения оптимального множества Ω эталонов сводится к задаче k-кратного покрытия множества $\mathbb{X}$ подмножествами F. Семейство F состоит из подмножеств M, которые строятся следующим образом: для каждого объекта $x_i \in \mathbb{X}$ строится подмножество $M \subset \mathbb{X}$, состоящее из объектов того же класса, к которым x_i находится ближе ближайшего объекта чужого класса.

Простой жадный алгоритм поиска эталонов в терминах подмножеств состоит в следующем. На каждом шаге выбирается такое подмножество M, которое покрывает максимальное количество непокрытых k раз объектов. Алгоритм останавливается, когда каждый из объектов оказывается покрыт хотя бы k раз.

Теорема 8. *Количество объектов, отобранных жадным алгоритмом в Ω, превышает оптимальное не более чем в $H\big(\max\{|M|\colon M \in F\}\big)$ раз, где $H(n) = \sum\limits_{k=1}^{n} \frac{1}{k}$ — гармоническое число.*

Для выборки, не являющейся k-идеальной, для каждого объекта $x_i \in \mathbb{X}$ упорядочим всех его соседей по возрастанию расстояний. Найдем такое чис-

ло m_i, что первые m_i соседей объекта x_i принадлежат тому же классу, а (m_i+1)-й сосед — чужому. Семейство подмножеств F строится по выборке $\mathbb{X}$ так же, как и в случае k-идеальной выборки.

Простой жадный алгоритм поиска эталонов для выборки, не являющейся k-идеальной, состоит в следующем. На каждом шаге выбирается такое подмножество M, которое покрывает максимальное количество непокрытых $\min\{k, m_i\}$ раз объектов. Алгоритм останавливается, когда каждый из объектов оказывается покрыт хотя бы $\min\{k, m_i\}$ раз.

Для данного жадного алгоритма также верна оценка Теоремы 8.

О субквадратичных алгоритмах отбора эталонов. Алгоритм называется субквадратичным, если время его работы в наихудшем случае составляет $o(n^2)$, где n — длина входных данных.

Следующая теорема указывает на проблему, возникающую при оценке компактности больших выборок.

Теорема 9. *Если для функции расстояния $\rho(x, x')$ нет возможности оценить расстояние межу объектами (например, при помощи неравенства треугольника), непосредственно не вычисляя его, то невозможно построить субквадратичный алгоритм вычисления $P(1)$, гарантирующий точность выше $0,5$.*

Идея доказательства состоит в следующем. В общем случае в каждом из классов находится $O(n)$ объектов. Если алгоритм субквадратичный, то точная информация о том, какому классу принадлежит ближайший сосед, имеется не более чем для $o(n^2)/n$, т. е. для $o(n)$ объектов.

Рассмотрим значение $P(1)$, оцениваемое алгоритмом.

Пусть $P(1) \geqslant 0,5$. Назначим невычисленные расстояния следующим образом: все расстояния между объектами одного класса меньше наименьшего из вычисленных, все расстояния между объектами разных классов больше наибольшего из вычисленных. Тогда для более чем $n/2$ объектов алгоритм будет давать ответ, что ближайший сосед принадлежит чужому классу, и из этих ответов только $o(n)$ будут верными.

Пусть теперь $P(1) < 0,5$. Назначим невычисленные расстояния следующим образом: все расстояния между объектами одного класса больше наибольшего из вычисленных, все расстояния между объектами разных классов меньше наименьшего из рассмотренных. Тогда для более чем $n/2$ объектов алгоритм будет давать ответ, что ближайший сосед принадлежит тому же классу, и из этих ответов только $o(n)$ будут верными.

Следовательно, ошибка с ростом n может оказаться сколь угодно близка к $0,5$.

Выводы

Доказано, что задачи выбора оптимального множества эталонов как по минимуму частоты ошибок, так и по минимуму полного скользящего контроля, являются NP-полными. Тем самым оправдывается применение субоптимальных жадных эвристических методов отбора эталонов.

Получены оценки точности некоторых семейств жадных алгоритмов.

Показано, что применение субквадратичных алгоритмов к задаче оценки профиля компактности выборки без привлечения дополнительной информации о расстояних не дают гарантии точности выше $0,5$.

Литература

[1] *Бонгард М. М.* Проблема узнавания. — М.: Наука, 1967.

[2] *Борисова И. А., Дюбанов В. В., Загоруйко Н. Г., Кутненко О. А.* Сходство и компактность // Всеросс. конф. Математические методы распознавания образов-14. — М.: МАКС Пресс, 2009. — С. 89–92.

[3] *Воронцов К. В.* Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики / под ред. О. Б. Лупанова. — М.: Физматлит, 2004. — Т. 13. — С. 5–36.

[4] *Воронцов К. В., Колосков А. О.* Профили компактности и выделениеопорных объектов в метрических алгоритмах классификации // Искусственный Интеллект. — 2006. — С. 30–33.

[5] *Загоруйко Н. Г.* Прикладные методы анализа данных и знаний. — Новосибирск: ИМ СО РАН, 1999.

[6] *Кормен Т., Лейзерсон Ч., Ривест Р., Штайн К.* Алгоритмы: построение и анализ, 2-е издание, М.: Издательский дом «Вильямс», 2005.

[7] *Bermejo S., Cabestany J.* Learning with nearest neighbour classifiers // Neural Processing Letters, 2001. — Vol. 13, No. 2. — Pp. 159–181.

[8] *Burges C. J. C.* A tutorial on support vector machines for pattern recognition // Data Mining and Knowledge Discovery, 1998. — Vol. 2, No. 2. — Pp. 121–167.

[9] *Ivanov M. N.* Prototype sample selection based on minimization of the complete cross validation functional // *Pattern Recognition and Image Analysis.* — 2010. — Vol. 20, no. 4. — Pp. 427–437.

[10] *Mullin M., Sukthankar R.* Complete cross-validation for nearest neighbor classifiers // Proceedings of International Conference on Machine Learning, 2000. — Pp. 639–646.

[11] *Zukhba A. V.* NP-completeness of the problem of prototype selection in the nearest neighbor method // Pattern Recognition and Image Analysis, 2010. — Vol. 20, No. 4. — Pp. 484–494.

О параметрах некоторых частично упорядоченных множеств*

Гуров С. И.
sgur@cs.msu.ru
Москва, ф-т ВМиК МГУ им. М. В. Ломоносова

Представлены формулы для вычисления числа элементов решёток порядковых идеалов и размеров частично упорядоченных множеств специального вида, предложен метод оценки размера частично упорядоченных множеств с использованием принципа согласованности. Дана верхняя оценка наибольшего значения множества Парето ч. у. множества.

Основные обозначения. Операции над ч. у. множествами

Рассматриваются конечные частично упорядоченные (ч. у.) множества, т. е. пары $\langle P, \leqslant \rangle = \mathbf{P}$, где P — непустое конечное множество (носитель), а $\leqslant$ — порядок (рефлексивное, антисимметричное и транзитивное бинарное отношение) на нём.

Для ч. у. множества $\mathbf{P}$ с носителем P и $a, b \in P$ используем следующие обозначения: $a \lessdot b$ означает, что b покрывает a (т. е. a непосредственно предшествует b); несравнимость a и b обозначаем $a \nsim b$. Совокупность максимальных элементов ч. у. множества называется его множеством Парето; $\mathbf{2}^n$ — булев n-мерный куб.

Ч. у. множества будем задавать диаграммами Хассе. Ч. у. множества, состоящие из $n > 2$ элементов $\{v_1, \ldots, v_n\}$ с отношениями включения $v_{2i-1} \lessdot v_{2i}$ и $v_{2i} \gtrdot v_{2i+1}$ (последнее включение при чётном n и $i = n/2$ отсутствует) и двойственные им назовём заборами и будем обозначать $\mathbf{Z}_n$. Обычно элементы нижней и верхней долей множества $\mathbf{Z}_n$ обозначают соответственно символами a и b с индексами. Зигзаг $\mathbf{Z}_8$ изображён на Рис. 1.

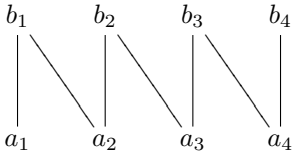


Рис. 1. Зигзаг $\mathbf{Z}_8$.

Если в $2n$-элементном заборе при $n \geqslant 3$ добавить условие «последний элемент покрывает первый», то получим ч. у. множество, которое назовём (малой) короной $\mathbf{s}_n$. Малая корона $\mathbf{s}_4$ изображена на Рис. 2. Понятно, что корона $\mathbf{s}_n$ изоморфна упорядоченной по включению совокупности всех одноэлементных и двухэлементных подмножеств вида $\{v_1, v_2\}, \{v_2, v_3\}, \ldots, \{v_n, v_1\}$ n-элементного множества $\{v_1, \ldots, v_n\}$.

Короны и заборы часто появляются при исследовании конечных ч. у. множеств [1]. Короны играют важную роль при изучении симметрии в частичных порядках, а заборы возникают при изучении их связности. Эти ч. у. множества рассмотрены Д. Даффусом и И. Ривалом [2] в их теории порядковых многообразий. Алгебраические свойства малых корон и заборов изучаются в [3] и [4]. В [5] представлены точные формулы и асимптотики чисел порядковых эндоморфизмов для заборов $1 \lessdot 2 \gtrdot 3 \lessdot \ldots \gtrdot 2n-1 \lessdot 2n$ и корон $1 \lessdot 2 \gtrdot 3 \lessdot \ldots \gtrdot 2n-1 \lessdot 2n \gtrdot 1$.

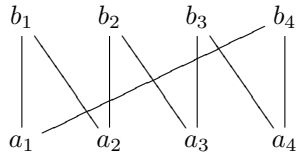


Рис. 2. Малая корона $\mathbf{s}_4$.

Большая корона $\mathbf{S}_n$ — это $2n$-элементное (всегда предполагают, что $n \geqslant 3$) двудольное ч. у. множество, которого $A = \{a_1, \ldots, a_n\}$ — множество минимальных, а $B = \{b_1, \ldots, b_n\}$ — множество максимальных элементов, причём для элементов $a_i \in A$ и $b_i \in B$ полагают $a_i \leqslant b_j$ для всех $i \neq j$, $i, j = 1, \ldots, n$ (и, естественно, $\leqslant$ рефлексивен). На Рис. 3 изображена корона $\mathbf{S}_5$. Очевидно,

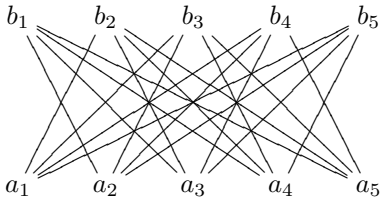


Рис. 3. Большая корона $\mathbf{S}_5$.

упорядоченное по включению множество одноэлементных и $(n-1)$-элементных подмножеств n-элементного $(n \geqslant 3)$ множества изоморфно $\mathbf{S}_n$.

В. Троттер [6] ввёл ч. у. множество, которое мы будем называть обобщённой короной. Это $2(n+k)$-элементное $(n \geqslant 3,\ k \geqslant 0)$ двудольное ч. у. множество $\mathbf{S}_n^k = A \cup B$, где $A = \{a_1, \ldots, a_{n+k}\}$ — множество минимальных, а $B = \{b_1, \ldots, b_{n+k}\}$ — множество максимальных элементов. Порядок $\leqslant$ на S_n задаётся следующим образом: $b_i \nsim \{a_i, a_{i+1}, \ldots, a_{i+k}\}$

Работа выполнена при финансовой поддержке РФФИ, проект №10-01-00131-а и компании А/О «Интел».

для $i = 1, 2, \ldots, n+k$, $a \lessdot b$ для остальных пар элементов из A и B (т. е. $a_j \lessdot b_i$ для $j = i+k+1, i+k+2, \ldots, i+k+n-1$). При этом границы индексов интерпретируются циклически ($n+k+1$ заменяется на 1, $n+k+2$ — на 2 и т. д.). Очевидно, $|A \cap \langle b \rangle| = n-1$ для любого $b \in B$. Согласно определениям, $\mathbf{S}_3^{n-3} = \mathbf{s}_n$ и $\mathbf{S}_n = \mathbf{S}_n^0$.

Кратко результаты данной работы опубликованы в [13].

Решётка порядковых идеалов

Напомним, что порядковым идеалом ч. у. множества $\langle P, \leqslant \rangle$ называется подмножество J его элементов такое, что $x \in J \,\&\, y \leqslant x \;\Rightarrow\; y \in J$. Согласно определению, пустое множество всегда является порядковым идеалом. Главный идеал, порождённый элементом a ч. у. множества $\langle P, \leqslant \rangle$ — множество $\{ x \in P \mid x \leqslant a \}$ — обозначаем $\langle a \rangle$. Идеал, порождённый совокупностью элементов $D = \{ x_1, \ldots, x_n \}$ обозначаем $\langle x_1, \ldots, x_n \rangle$ или $\langle D \rangle$.

Многие комбинаторные свойства конечного ч. у. множества $\mathbf{P}$ имеют простые интерпретации в терминах (дистрибутивной) решётки его порядковых идеалов $J(\mathbf{P})$. Значение $|J(\mathbf{P})|$ удалось определить лишь для немногих ч. у. множеств.

Числа Люка определяются как $L_n = F_{n+1} + F_{n-1}$, где F_n — n-е число Фибоначчи.

Теорема 1. $|J(\mathbf{s}_n)| = L_{2n}$.

Теорема 2. $|J(\mathbf{S}_n)| = 2^{n+1} + n - 1$.

Размер корон

Согласно теореме Шпильрайна [7], любой частичный порядок может быть продолжен до линейного на том же множестве. Совокупность всех таких продолжений (линеаризаций) $\{ \mathbf{L}_1, \ldots, \mathbf{L}_k \}$ ч. у. множества $\mathbf{P}$ обозначим $\mathcal{R}(\mathbf{P})$.

Дашник и Миллер [8] установили, что каждый порядок может быть представлен в виде пересечения своих линейных продолжений, т. е. $\mathbf{P} = \bigcap\limits_{\mathbf{L} \in \mathcal{R}(\mathbf{P})} \mathbf{L}$.

Мощность множества $\mathcal{R}(\mathbf{P})$ называют размером ч. у. множества $\mathbf{P}$ и обозначают $e(\mathbf{P})$. Оно может интерпретироваться как некоторая оценка сложности $\mathbf{P}$. Значение $e(\mathbf{P})$ известно лишь для немногих типов ч. у. множеств, а задача его вычисления NP-полна [9]. Имеющиеся формулы [10, 11], требуют перебора по некоторым связанным с $\mathbf{P}$ множествам и не позволяют определять значения размера $e(\mathbf{P})$ в явном виде[1].

Известны несколько принципиальных способов определения размера ч. у. множества. Один из них [11] основан на том, что величина $e(\mathbf{P})$ равна числу $\mathcal{R}(J(\mathbf{P}))$ максимальных цепей в $J(\mathbf{P})$ от $\varnothing$ до $\langle P \rangle$. Этим методом были получены результаты, приведённые в нижеследующих теоремах.

Теорема 3. $e(\mathbf{S}_n) = (n-1)!(n+1)!$.

Для зигзагов $\mathbf{Z}_n$ давно известно представление

$$\sum_{n \geqslant 0} \frac{e(\mathbf{Z}_n)}{n!}\, x^n = \tan x + \sec x\,,$$

при этом значения $e(\mathbf{Z}_n)$ при чётных n называют числами секанса, а при нечётных — числами тангенса и совпадают с числом up-down перестановок (такие перестановки $(i_1, \ldots, i_n)$ первых n натуральных чисел, что $i_1 < i_2 > i_3 < i_4 > \ldots$. При этом явной формулы для $e(\mathbf{s}_n)$ известно не было.

Теорема 4.

$$\sum_{n \geqslant 1} \frac{e(\mathbf{s}_n)}{n!}\, x^n = \frac{x}{\cos^2 x} = x \tan' x\,.$$

Числа $e(\mathbf{s}_n)$ и, для сравнения, числа секанса $e(\mathbf{Z}_{2n})$, для первых значений n приведены в нижеследующей таблице.

n	2	3	4	5	6
$e(\mathbf{s}_n)$	4	48	1 088	39 680	2 122 752
$e(\mathbf{Z}_{2n})$	5	61	1 385	50 521	2 702 765

Другой возможный способ [11] вычисления размера ч. у. множеств основан на том факте, что для $P = \{ v_1, \ldots, v_n \}$ имеет место равенство $e(\mathbf{P}) = n! \cdot \mathrm{vol}(\mathcal{P})$, где $\mathrm{vol}(\mathcal{P})$ — объём многогранника

$$\mathcal{P} = \mathcal{P}(\mathbf{P}) = \{ (x_1, \ldots, x_n) \in \mathbb{R}^n \mid 0 \leqslant x_i \leqslant 1,\ v_i \leqslant v_j \;\Rightarrow\; x_i \leqslant x_j \}\,.$$

Ясно, что $\mathcal{P}$ — выпуклый многогранник, заключённый в куб E^n с единичным ребром в $\mathbb{R}^n$ и ограниченный гиперплоскостями $x_i = x_j$ для всех сравнимых элементов v_i и v_j ч. у. множества $\mathbf{P}$. Заметим, что, вообще говоря, $\mathrm{vol}(\mathcal{P}) \ll 1$.

Оценки $\mathrm{vol}(\mathcal{P})$ вероятностного типа получим с помощью метода Монте-Карло. Для этого зафиксируем ч. у. множество $\mathbf{P}$, сгенерируем $N \gg 1$ псевдослучайных точек, равномерно распределённых в E^n и подсчитаем их количество m, попавших в многогранник $\mathcal{P}(\mathbf{P})$. Умножая отношение $p = \frac{m}{N}$ на $n!$, получим точечную оценку искомой величины $e(\mathbf{P})$.

Для получения интервальных оценок $\mathrm{vol}(\mathcal{P})$ воспользуемся принципом согласованности, ориентированным, прежде всего, на получение уточнённых оценок редких событий. Для нахождения точечных оценок вероятностей p редких событий

[1] В англоязычной литературе такие формулы называют messy — неудобные, неуклюжие.

в рамках байесовского подхода Э. Леман [12] предложил в качестве плотности априорного распределения использовать **B**(бетта)-распределение (распределение Бернулли)

$$Be_p(a,\,b) \;=\; \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\,p^{a-1}(1-p)^{b-1}$$

с параметром $a = 1$ и достаточно большим положительным b. Однако не было приведено ни обоснования данному выбору, ни каких-либо указаний на возможный способ определения параметра b. Принцип согласованности обосновывает эту идею Э. Лемана и позволяет определить b исходя из идеи совпадения оценок, получаемых в рамках байесовского, классического частотного и фишеровского фидуциального подходов [14, 15].

Согласованный доверительный интервал (p^-, p^+) с коэффициентом доверия $\eta = (P_\eta + 1)/2$ находится из уравнений

$$\begin{cases} I_{p^-}(m+a-1,\,N-m+b) \;=\; 1 \,-\, P_\eta; \\ I_{p^+}(m+a,\,N-m+b-1) \;=\; P_\eta, \end{cases}$$

где $I_p(\cdot,\cdot)$ — отношение неполной **B**-функции к полной **B**-функции с соответствующими параметрами, а величины a и b определяются условиями

$$\begin{cases} a \;=\; 1\,,\; b \;=\; \dfrac{N-m}{m}\,, & \text{если } 1 \leqslant m \leqslant \dfrac{N}{2}\,; \\ a \;=\; \dfrac{m}{N-m}\,,\; b \;=\; 1\,, & \text{если } \dfrac{n}{2} < m \leqslant N-1; \end{cases}$$

(в нашем случае будет реализовываться первый вариант)[2].

Была составлена компьютерная программа, реализующая данный подход и проведены контрольные расчёты для величин $e(\mathbf{Z}_n)$, точные значения которых известны. Например, для $n = 12$ при точном значении $e = 2\,702\,765$ получена точечная оценка $\widehat{e} = 2\,710\,622{,}154$ при сокращении длины доверительного интервала с 20 620 до 20 610 (число испытаний $N = 20 \cdot 10^6$, коэффициент доверия $\eta = 0{,}95$).

Проблема Ногина

Представление ч. у. множества в виде пересечения всех своих возможных линеаризации (как правило, очень сильно) избыточно. Говорят, что совокупность $\mathcal{R}$ цепей $\{\,\mathbf{L}_1,\,\ldots,\,\mathbf{L}_r\,\}$ мощности r цепей реализует ч. у. множество $\mathbf{P}$, если $\mathbf{P} = \bigcap\limits_{\mathbf{L}\in\mathcal{R}} \mathbf{L}$, при этом r называют рангом $\mathcal{R}$. Дашником и Миллером [8] введено понятие размерности ч. у. множества.

[2]Заметим, что классические доверительные интервалы могут быть вычислены по указанным формулам, если положить $a = b = 1$, они будут иметь бóльшую, чем согласованные, длину, а совпадение будет только при $m = N/2$.

Определение 1. *Размерностью* $\mathbf{P}$ *(символически* $\dim(\mathbf{P})$*) ч. у. множества* $\mathbf{P}$ *называется наименьшее натуральное* d *такое, что для совокупности* $\{\mathbf{L}_1, \ldots, \mathbf{L}_d\}$ *линейных расширений* $\mathbf{P}$ *справедливо* $\mathbf{P} = \bigcap\limits_{i=1}^{d} \mathbf{L}_i$.

Известно [16], что $\dim(\mathbf{P})$ совпадает с наименьшим числом k линейных порядков C_i таких, существует вложение $P \hookrightarrow C_1 \times \ldots \times C_k$. Задача вычисления размерности ч. у. множества NP-полна [17]. Проблеме размерности ч. у. множеств посвящена монография [18].

Определение 2. *Конечное ч. у. множество называют* d*-несводимым, если его размерность равна* $d \geqslant 2$ *и удаление любого его элемента приводит к ч. у. множеству с размерностью, меньшей* d*.*

Ч. у. множество несводимо, если оно d*-несводимо для некоторого* d*.*

Из результатов Хирагучи [19] сразу следует, что ч. у. множество $\mathbf{P}$ d-несводимо для некоторого $d \geqslant 2$, если и только если $\dim(\mathbf{P} - x) = d - 1$ для всех $x \in P$.

Несводимые ч. у. множества моделируют наиболее сложные в структурном отношении задачи векторной оптимизации и играют важную роль при анализе и синтезе алгоритмов решения этих экстремальных задач.

Все 3-несводимые порядки описаны, например, в [20]; в частности, имеется 3 шестиэлементных и 21 семиэлементных 3-несводимых ч. у. множества.

Любое ч. у. множество размерности не меньше d содержит d-несводимое подмножество. Известно, что d-несводимое ч. у. множество при $d > 3$ не может содержать $2d + 1$ элементов. Однако, если $d \geqslant 4$ и $n > 2d + 1$, то n-элементное d-несводимое множество существует всегда [21].

Теорема 5 ([6, 18]; [22]).

1. *Если* $n + k = q(k + 2) + 1$, *то* $\mathbf{S}_n^k$ *—* $(2q + 1)$*-несводимо.*
2. *Пусть* $n+k = q(k+2)+\lfloor\frac{1}{2}(k+2)\rfloor+1$. *Тогда* $\mathbf{S}_n^k$ *—* $(2q + 2)$*-несводимо если и только если* k *либо равно 0, либо нечётно.*
3. *Корона* $\mathbf{S}_n^k$ *несводима, если и только если она описывается двумя вышепреведёнными случаями.*
4. $\mathbf{S}_n$ *— единственное* n*-несводимое ч. у. множество с* $2n$ *элементами.*

В 1990 г. В. Д. Ногиным [23, 24] поставлена проблема: *Каково наибольшее значение* $\mu(d, n)$ *числа максимальных элементов* n*-элементного* d*-несводимого ч. у. множества при* $d \geqslant 4$*?*

Со времени постановки данной проблемы в решении её не было каких-либо продвижений. Ясно

только, что результаты теоремы 5 позволяют установить нижнюю границу $\mu(d,n)$ для чётных n и некоторых значений d. Нижеследующая теорема определяет верхнюю границу величины множества Парето.

Теорема 6. $\mu(d,n) \leqslant n-d$.

Следствие 1. *Из данной теоремы и результатов [25] следует, что для любых целых* $d>3$, $q>1$

$$\mu(d,2d)=d;$$

т. к. единственное ч. у. множество данного типа — $\mathbf{S}_d$*;*

$$\mu(2q+1,2k(q+2)+2) \geqslant k(q+2)+1 \quad \text{при } k \geqslant 0;$$

$$\mu(2q+1,2q(k+2)+k+3) \geqslant q(k+2)+\frac{k+3}{2}$$

при k *нечётном.*

Выводы

В статье даны формулы для вычисления числа элементов решёток порядковых идеалов и размеров частично упорядоченных множеств специального вида. Также предложен метод оценки размера частично упорядоченных множеств с использованием принципа согласованности и дана верхняя оценка наибольшего значения множества Парето. В целом проблема Ногина остаётся открытой.

Литература

[1] *Demetrovics J., Rónyai L.* Algebraic Properties of Crowns and Fences // Order. — 1989. — N. 6. — Pp. 91–99.

[2] *Duffus D., Rival I.* A structure theory of ordered sets // Discrete Math. — 1981. — N. 36. — Pp. 53–18.

[3] *Quackenbush R. W., Rival I., Rosenberg I. G.* Near unanimity orders. Clones, Order Varieties, Near Unanimity Functions and Holes // Order. — 1990. — N. 7. — Pp. 239–247.

[4] *Fiiredi Z., Rosenberg I. G.* Orders admitting an isotone majority operation // Preprint CRM Universite de Montreal. — 1985.

[5] *Currie J. D., VisentiN. T. I.* The number of order-preserving maps for fences and crowns // Order. — 1991. — V. 8, N. 3. — Pp. 133–142.

[6] *Trotter W. T., Jr.* Dimension of the crown $\mathbf{S}_n^k$ // Discrete Math. — 1974. — N. 8. — Pp. 85–103.

[7] *SzpilrajN. E.* Sur l'extension de l'ordere parttiel // Fund. Math. — 1930. — N. 16. — Pp. 368–389.

[8] *Dushnik B., Miller E. W.* Partially ordered sets // Amer. J. Math. — 1948. — N. 63. — Pp. 600–610.

[9] *Brightwell G., Winkler P.* Counting linear extensions // Order. — 1991. — N. 8. — Pp. 225–242.

[10] *Stanley R.* Ordered structures and partitions // Memories Amer. Math. So. — 1972. — N. 119.

[11] *Стенли Р.* Перечислительная комбинаторика (Volume I). — М.: Мир, 1990. — 440 с.

[12] *Леман Э.* Теория точечного оценивания. — М.: Мир, 1991. — 445 с.

[13] *Гуров С. И.* Нахождение некоторых характеристик частично упорядоченных множеств // Дискретные модели в теории управляющих систем: VIII Международная конференция, Москва, 6-9 апреля, 2009 г.: Труды. Издательский отдел факультета ВМиК МГУ им. М. В. Ломоносова; МАКС Пресс, 2009. — С. 70–75.

[14] *Гуров С.И.* Принцип согласованности и бейесовское интервальное оценивание // Таврический вестник информатики и математики. — 2003. — № 2. — С. 14–27.

[15] *Гуров С.И.* Интервальное оценивание на основе принципа согласованности // Вестник Тверского государственного университета. Серия «Прикладная математика». — 2008. — Т. 74, № 14. — С. 77–93.

[16] *Оре О.* Теория графов. — М.: Наука, 1980. — 336 с.

[17] *Yannakakis M.* The complexity of the partional order dimension promlem // SIAM J. Alg. Disc. Math. — 1982. — N. 3. — Pp. 351–358.

[18] *Trotter W. T.* Combinatorics and partially ordered sets: dimension theory. — Baltimore and Loondon: The Johnes Hopkins University Press, 1992. — 308 p.

[19] *Hiraguchi T.* On the dimension of partially ordered sets // Sci. Rep. Kanazawa Univ. — 1951. — N. 1. — Pp. 77–94.

[20] *Kelly D.* The 3-irreducible partially ordered sets // Canad. J. Math. — 1977. — N. 29. — Pp. 363–383.

[21] *Kimble R. J.* External problems in dimension theory of partially ordered sets // Ph. D. thesis., M.I.T. — 1973.

[22] *Trotter W. T.* Maximal dimansional partially ordered sets II: characterization of $2n$-element posets with dimension n // Discrete Math. — 1973. — N. 5. — Pp. 33–43.

[23] *Ногин В. Д.* Критериальная размерность задач векторной оптимизации // Модели и методы оптимизации. — ВНИИСИ, Москва. — 1997. — Т. 38, № 7. — С. 55–60.

[24] *Ногин В. Д.* Проблема. — 1990. — `http://pmpu.ru/vf/NogiN.\_Vladimir\_Dmitrievich/Problem`.

[25] *Trotter W.T.* Dimension of the crown $\mathbf{S}_n^k$ // Discrete Math. — 1974. — N. 8. — Pp. 85–103.

Оптимальный метод перераспределения общей памяти для двух последовательных циклических FIFO-очередей*

Аксенова Е. А., Соколов А. В.
aksenova@krc.karelia.ru, avs@krc.karelia.ru
Петрозаводск, Институт прикладных математических исследований Карельского научного центра РАН

В данной работе предложена математическая модель работы с двумя последовательными циклическими FIFO-очередями, расположенными в памяти размера m единиц, и алгоритм, который позволяет обрабатывать ситуацию переполнения одной из FIFO-очередей. Решается задача оптимального перераспределения памяти между двумя очередями после переполнения одной из очередей. В качестве математической модели предложено случайное блуждание по целочисленной решетке в прямоугольной области на плоскости.

Во многих приложениях требуется работа с несколькими FIFO-очередями, расположенными в общем пространстве памяти. Для этого применяют различные программные или аппаратные решения [1–3]. В работах [4–7] предлагались математические модели для последовательного циклического и связанного способов представления очередей. На основе предложенных моделей решались задачи оптимального начального распределения памяти для очередей при различных критериях оптимальности. Если в качестве критерия оптимальности рассматривалась минимальная доля потерянных элементов при бесконечном времени работы очередей [5–7], то предполагалось, что при переполнении очереди все последующие элементы, поступающие в нее, отбрасываются до тех пор, пока не появится свободная память. Если в качестве критерия оптимальности рассматривалось максимальное среднее время до переполнения памяти [4], то предполагалось, что при переполнении очереди работа программы заканчивается. Такой метод работы с очередями необходим, когда FIFO-очереди используются в программных системах, и потери элементов очередей недопустимы.

В рассмотренных моделях распределение памяти между очередями выполняется один раз в зависимости от заданного распределения вероятностей (предполагается, что для данного приложения заранее проведен необходимый статистический анализ) выполнения операций включения и исключения элементов очередей.

Математическая модель

Рассмотрим две FIFO-очереди, расположенные в памяти размера m единиц. В каждый момент времени может произойти включение элемента (информации) в очередь, исключение элемента из очереди, чтение или отсутствие операции. В очередях хранятся элементы равной длины. Пусть заданы вероятностные характеристики очередей:

- p_i — вероятность включения элемента в i-ю очередь, $i = 1, 2$;
- q_i — вероятность исключения элемента из i-й очереди, $i = 1, 2$;
- r — вероятность того, что очереди не меняют свою длину (чтение или отсутствие операции).

Выделим первой очереди s единиц памяти, тогда второй очереди останется $m - s$ единиц. Обозначим текущие длины очередей x_1 и x_2. В качестве математической модели рассмотрим случайное блуждание по целочисленной решетке в прямоугольной области на плоскости $0 \leqslant x_1 < s + 1$, $0 \leqslant x_2 < m - s + 1$. Блуждание начинается в точке $x_1 = 0,\ x_2 = 0$. Переполнению первой очереди соответствуют точки на прямой $x_1 = s+1$, переполнению второй — точки на прямой $x_2 = m - s + 1$. Прямые $x_1 = -1$ и $x_2 = -1$ — отражающие экраны, т. е. при исключении элемента из пустой очереди работа не завершается.

Требуется оптимально перераспределить свободную память между очередями после переполнения одной из очередей.

В данной задаче предполагаем, что при переполнении одной из очередей происходит перераспределение свободной памяти между очередями и работа с программной системой продолжается. Другими словами, при попытке включения элемента в заполненную очередь, когда $x_1 = s$ (или $x_2 = m - s$), требуется определить новую область блуждания $0 \leqslant x_1 < s^* + 1, 0 \leqslant x_2 < m - s^* + 1$, т. е. такое значение s^*, где $s^* > s$ (или $m - s^* > m - s$), чтобы время до следующего переполнения какой-либо очереди было максимальным. Этот процесс перераспределения можно продолжать до полного исчерпания свободной памяти.

Такой подход оправдан, если при переполнении одной из очередей, в области памяти, выделенной для других очередей, есть достаточно свободной памяти. В классической работе Д. Кнута [1] для работы с n последовательными стеками и очередями рассматривался алгоритм Гарвика. В этом алгоритме при переполнении какой-либо структуры данных приблизительно 10% свободной памяти делится поровну между структурами данных, а оставшиеся 90% делятся пропорционально росту

Работа выполнена при финансовой поддержке РФФИ, проект № 09-01-00330.

размеров структур данных с момента предыдущего распределения памяти.

Алгоритм решения

Математическая постановка задачи сводится к решению задачи нелинейного целочисленного программирования с критерием оптимальности, заданным алгоритмически. Случайное блуждание по целочисленной решетке будем рассматривать как конечную однородную поглощающую цепь Маркова с матрицей вероятностей переходовP [8]. Для решения задачи требуется матрица Q вероятностей переходов из невозвратных состояний в невозвратные (Q — подматрица матрицы P). Вводится нумерация состояний Марковской цепи, это позволяет определить структуру матрицы Q для любого размера памяти m и любого значения s. Среднее время работы до переполнения будем искать с помощью фундаментальной матрицы $N = (E - Q)^{-1}$. Предполагаем, что в самом начале работы, когда обе очереди пустые $x_1 = 0,\ x_2 = 0$, память между ними тоже нужно разделить оптимально, т. е. выбрать такое s, чтобы время работы с очередями до переполнения было максимально.

Алгоритм.

Шаг 1. Ввод вероятностных характеристик очередей, размера памяти m.

Шаг 2. Для каждого значения s, $0 \leqslant s \leqslant m$, генерируем матрицу Q, вычисляем матрицу N.

Шаг 3. Для каждого значения s суммируем элементы матрицы N в строке, соответствующей состоянию $x_1 = 0,\ x_2 = 0$.

Шаг 4. Выбираем такое значение s, которому соответствует максимальная сумма элементов в строке соответствующей матрицы N.

Шаг 5. Для каждого состояния, в котором одна из очередей заполнила выделенную память, т. е. $x_1 = s$ или $x_2 = m - s$, перебираем значения $0 \leqslant s^* \leqslant m$, где $s^* > s$, если $x_1 = s$, и $m - s^* > m - s$, если $x_2 = m - s$. Для каждого s^* генерируем новую матрицу Q, вычисляем матрицу N.

Шаг 6. Для каждого значения s^* суммируем элементы матрицы N в строке, соответствующей рассматриваемому состоянию ($x_1 = s$ или $x_2 = m - s$).

Шаг 7. Выбираем такое значение s^*, которому соответствует максимальная сумма элементов в строке соответствующей матрицы N. Для каждого состояния, соответствующего тому, что одна из очередей заполнила выделенную память, получаем оптимальное значение s^*.

Шаги 5–7 можно повторять до полного исчерпания свободной памяти (если перед каждым новым прераспределением текущее значение s^* обозначить как s).

Вычисление критерия оптимальности в данной задаче является очень ресурсоемким за счет обращения матрицы $(E - Q)$ большого размера. Матрицу N можно представить в виде

$$N = (E - Q)^{-1} = E + Q + Q^2 + \ldots = \sum_{k=0}^{\infty} Q^k .$$

Для данной задачи такое представление матрицы дает выигрыш в размере памяти для вычислений. Поскольку для алгоритма важна сумма элементов в определенной строке матрицы N, то запись в виде ряда позволяет вычислять элементы конкретной строки, не вычисляя остальных элементов матрицы.

Выводы

Предложенный алгоритм реализован на языке C++, проведены численные эксперименты. Для рассматриваемой цепи Маркова можно вычислить вероятности попадания в состояния поглощения на прямых $x_1 = s + 1$ и $x_2 = m - s + 1$, т. е. вероятности переполнения очередей. Было бы интересно построить и проанализировать весь процесс перераспределения до полного исчерпания памяти с учетом вероятностей переполнения, и сравнить с имитационной моделью данного процесса.

Литература

[1] *Кнут Д.* Искусство программирования для ЭВМ. — М.: Вильямс, 2001. — 736 с.

[2] *Седжвик Р.* Фундаментальные алгоритмы на C++. — К.: Диасофт, 2001. — 688 с.

[3] *Боллапрагада В., Мэрфи К., Уайт Р.* Структура операционной системы Cisco IOS. — М.: Вильямс, 2002. — 208 с.

[4] *Аксенова Е. А., Соколов А. В.* Некоторые задачи оптимального управления FIFO-очередями // Труды Второй Всероссийской научной конференции «Методы и средства обработки информации» — М.: Изд. отдел ВМК МГУ им. М. В. Ломоносова, 2005. — С. 318–322.

[5] *Аксенова Е. А.* Оптимальное управление FIFO-очередями на бесконечном времени // Межвузовский сборник «Стохастическая оптимизация в информатике». — СПб.: Изд-во С.-Петербургского университета, 2006. — С. 71–76.

[6] *Аксенова Е. А., Драц А. В., Соколов А. В.* Об оптимальном управлении FIFO-очередями на бесконечном времени // Обозрение прикладной и промышленной математики, 2009. — Т. 16, Вып. 3. — С. 401–415.

[7] *Аксенова Е. А., Драц А. В., Соколов А. В.* Оптимальное управление n FIFO-очередями на бесконечном времени // Информационно-управляющие системы, 2009. — № 6. — С. 46–54.

[8] *Кемени Дж., Снелл Дж.* Конечные цепи Маркова. — М.: Наука, 1970. — 736 с.

Управление двумя FIFO-очередями в случае их движения друг за другом по кругу*

Соколов А. В., Драц А. В.
avs@krc.karelia.ru, adeon88@mail.ru
ИПМИ КарНЦ РАН, ПетрГУ

В статье исследуется метод представления двух FIFO-очередей в виде движения друг за другом по кругу в одноуровневой памяти. В качестве математических моделей предложены случайные блуждания по целочисленной решетке в различных областях трехмерного пространств. Задачи решаются с помощью аппаратов поглощающих и регулярных цепей Маркова.

Существуют два принципиально разных способа организации работы с динамическими структурами данных: последовательный и связанный [1–3]. Для FIFO-очередей при последовательном циклическом представлении вся память делится на несколько частей и каждой очереди выделяется своя область памяти. В этом случае могут возникнуть потери памяти, когда какая-нибудь структура данных полностью исчерпает свою часть памяти. При связанном представлении структура данных хранится в виде списка. Когда нужно добавить элемент, то необходимо найти незанятую область памяти и записать в нее новый элемент.

В данной работе исследуется новый метод представления нескольких последовательных циклических FIFO-очередей в общей памяти, предложенный в [5]. В этом методе очереди двигаются по кругу, друг за другом, начиная с некоторого начального места в памяти. Память здесь заранее не делится между очередями. В случае, если одна из очередей стала пустой, то вторая может занимать всю доступную память, пока не догонит сама себя. Если же оживет вторая очередь, то ее можно пустить с середины свободного участка или с некоторой оптимальной точки (поиск этой точки является нашей задачей), которая зависит от вероятностных характеристик очередей. В качестве критериев оптимальности рассмотрено максимальное среднее время до переполнения памяти и минимальная доля потерянных элементов. В работе предложена и исследуется математическая модель для этого способа работы с двумя FIFO-очередями. Данный метод сравнивается с последовательным и связанным методами представления FIFO-очередей.

Постановка задачи

Рассмотрим две очереди, находящиеся в памяти размера m единиц. Время дискретно и на каждом шаге возможна одна из следующих операций:

— включение элемента в первую очередь с вероятностью p_1;

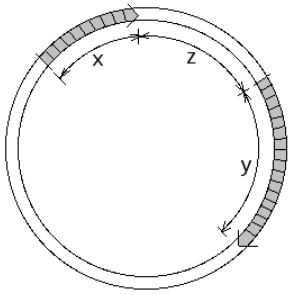


Рис. 1. Движение очередей по кругу

— исключение элемента из первой очереди с вероятностью q_2;
— включение элемента в первую очередь с вероятностью p_2;
— исключение элемента из первой очереди с вероятностью q_2;
— чтение элемента в одной из очередей без его исключения с вероятностью r:

$$p_1 + q_1 + p_2 + q_2 + r = m.$$

Все элементы имеют одинаковый размер, и не происходит завершения работы в случае попытки исключения элемента из пустой очереди. Работа начинается с пустых очередей.

Пусть x, y — длины очередей, z — расстояние от головы первой очереди до хвоста второй (рис. 1).

Если $x = 0$ или $y = 0$, т. е. хотя бы одна из очередей пустая, то z, вообще говоря, неопределено, поэтому будем считать в этом случае $z = 0$. Тройка (x, y, z) однозначно задает положение очередей в памяти. Переполнение возникнет в том случае, если первая очередь догонит вторую (при $z = 0$) или вторая первую (при $x + y + z = m$).

Среднее время работы до переполнения

Будем считать, что в случае переполнения происходит аварийное завершение работы. Критерием оптимальности является максимизация среднего времени работы до переполнения T.

В качестве математической модели рассмотрим блуждание по целочисленной трехмерной пирами-

Работа выполнена при финансовой поддержке РФФИ, грант № 09-01-00330.

де с вершиной $(0,0,0)$, ребрами $x=0,\ y=0,\ z=0$ и основанием $x+y+z=m$.

Поглощающие экраны: $x+y+z=m+1$ и $z=-1$.

Отражающие экраны: $x=-1$ и $y=-1$.

Перенумеруем состояния следующим образом:
$(0,0,0)$, $(1,0,0)$, …, $(M,0,0)$,
$(0,1,0)$, $(1,1,0)$, …, $(M-1,1,0)$,
…
$(0,M-1,0)$, $(1,M-1,0)$,
$(0,M,0)$,
$(0,0,1)$, $(1,0,1)$, …, $(M-1,0,1)$,
$(0,1,1)$, $(1,1,1)$, …, $(M-2,1,1)$,
…
$(0,M-2,1)$, $(1,M-2,1)$,
$(0,M-1,1)$,
…
$(0,0,M-1)$, $(1,0,M-1)$,
$(0,1,M-1)$,
$(0,0,M)$.

Переходы между состояниями:

$$(x,y,z)\xrightarrow{p_1}\begin{cases}(x+1,y,z-1), & \text{если } x,y,z>0;\\ \text{переполнение}, & \text{если } z=0,\ y\neq 0\\ & \text{или } x=m;\\ (1,y,(m-y)/2), & \text{если } x=0,\ y<m;\\ (x+1,0,0), & \text{если } y=0,\ x<m;\end{cases}$$

$$(x,y,z)\xrightarrow{q_1}\begin{cases}(0,y,0), & \text{если } x=1;\\ (x-1,y,z), & \text{иначе}\end{cases}$$

$$(x,y,z)\xrightarrow{p_2}\begin{cases}(x,y+1,z), & \text{если } x,y,z>0;\\ \text{переполнение}, & \text{если } x+y+z=m;\\ (x,1,(m-x)/2), & \text{если } y=0,\ x<m;\\ (0,y+1,0), & \text{если } x=0,\ y<m;\end{cases}$$

$$(x,y,z)\xrightarrow{q_2}\begin{cases}(x,0,0), & \text{если } y=1;\\ (x,y-1,z+1), & \text{иначе}\end{cases}$$

Среднее время работы T можно найти с помощью аппарата поглощающих цепей Маркова. Для этого необходимо вычислить фундаментальную матрицу $N=(E-Q)^{-1}$ [4] (N_{ij} показывает сколько времени процесс находился в состоянии j, если он начался в состоянии i) и просуммировать элементы строки, которая соответствует начальному состоянию $(0,0,0)$.

Управление на бесконечном времени

Переполнение одной из очередей не всегда является аварийной ситуацией. В том случае, если очередь заполнила всю отведенную ей память, то все последующие элементы отбрасываются до тех пор, пока не появится свободная память. Такое поведение очереди называется «сбросом хвоста» [2]. В этом случае критерий оптимальности — минимизация доли времени, которую система проводит в состоянии «сброса хвоста».

В качестве математической модели рассмотрим блуждание по трехмерной пирамиде, как в предыдущем случае. Дополнительно для каждого состояния на плоскостях $z+y+z=0$ и $z=0$ введем состояния «сброса хвоста».

Долю времени, которую система проводит в состоянии «сброса хвоста» P^*, можно найти с помощью аппарата регулярных цепей Маркова. Для этого необходимо найти предельный вектор α и просуммировать его компоненты, соответствующие состояниям «сброса хвоста».

Сравнение различных вариантов представлений

В [6, 7] были исследованы последовательный и связанный способы представления очередей. В этом параграфе сравниваются различные варианты представлений и приводятся численные результаты экспериментов.

В табл. 1 приводятся результаты для критерия максимизации времени работы до переполнения T.

В табл. 2 — результаты для минимизации доли времени P^*.

При связанном представлении часть памяти тратится на указатели. l — отношение размера указателя к размеру информационной части, тогда на хранение информационных частей тратится $M=m/(1+l)$ единиц памяти. Чем меньше l, тем выгоднее использовать связанное представление.

В строках $l=1,\ l=1/2,\ l=1/4,\ l=1/8$ приведены результаты для связанного представления для различных l.

В строках $T_c,\ P_c^*$ — результаты для последовательного представления при оптимальном разбиении памяти.

В строке $T_r,\ P_r^*$ — результаты при движении очередей друг за другом по кругу.

Из приведенных таблиц видно, что для критерия максимизации среднего времени работы до переполнения предпочтительнее использовать представление очередей в виде движения друг за другом по кругу или связанное представление в том случае, если на указатели тратится незначительная часть памяти ($l=1/8$). Для критерия минимизации доли времени, которую система проводит в состоянии «сброса хвоста» предпочтительнее использовать последовательное представление. Связанное представление и представление в виде движения друг за другом по кру-

Таблица 1. Критерий максимизации среднего времени работы до переполнения, $m=20$.

p_1	0,1	0,2	0,3	0,4	0,4	0,3
q_1	0,4	0,3	0,2	0,1	0,1	0,2
p_2	0,25	0,25	0,25	0,25	0,1	0,2
q_2	0,25	0,25	0,25	0,25	0,4	0,3
T_c	550,94	287,53	85,95	43,41	57,56	111,82
$l=1$	225,47	140,23	51,44	26,58	33,29	65,73
$l=1/2$	370,38	241,29	72,35	35,07	43,57	92,93
$l=1/4$	487,01	327,66	87	40,82	50,23	111,73
$l=1/8$	619,63	429,91	102,06	46,62	56,89	130,87
T_r	721,48	343,94	92,08	44,71	62,71	125,58

Таблица 2. Критерий минимизации доли времени, которую система проводит в состоянии «сброса хвоста», $m = 20$.

p_1	0,1	0,2	0,3	0,4	0,4	0,3
q_1	0,4	0,3	0,2	0,1	0,1	0,2
p_2	0,25	0,25	0,25	0,25	0,1	0,2
q_2	0,25	0,25	0,25	0,25	0,4	0,3
P_c^*	0,015	0,022	0,122	0,315	0,3	0,1023
$l = 1$	0,0328	0,049	0,192	0,4875	0,375	0,167
$l = 1/2$	0,026	0,037	0,186	0,4875	0,375	0,167
$l = 1/4$	0,022	0,032	0,185	0,4875	0,375	0,167
$l = 1/8$	0,020	0,028	0,184	0,4875	0,375	0,167
P_r^*	0,017	0,024	0,164	0,475	0,373	0,146

гу предпочтительнее лишь в некоторых случаях, когда вероятности исключения элементов из очередей больше, чем вероятности включения.

Литература

[1] *Кнут Д.* Искусство программирования для ЭВМ. Т. 1 — М.: Вильямс, 2001.

[2] *Боллапрагада В., Мэрфи К., Уайт У* Структура операционной системы Cisco IOS. — М.: Вильямс, 2002. 208 с.

[3] *Кормен Е., Лейзерсон Ч., Ривест Р.* Алгоритмы построение и анализ. — М.: МЦНМО, 2000.

[4] *Кемени Дж., Снелл Дж.* Конечные цепи Маркова. — М.: Наука, 1970. 272 с.

[5] *Соколов А. В.* Математические модели и алгоритмы оптимального управления динамическими структурами данных. — ПетрГУ. Петрозаводск, 2002. 216 с.

[6] *Драц А. В., Соколов А. В.* Оптимальное размещение в памяти одного уровня n стеков и/или очередей //Стохастическая оптимизация в информатике, Вып.5. — СПб.: Изд-во С.-Петербургского университета, 2009. 72–90 с.

[7] *Аксенова Е. А. Драц А. В. Соколов А. В.* Оптимальное управление n FIFO-очередями на бесконечном времени //Информационно-управляющие системмы. — СПб.: Изд-во «Политехника», 2009. 46–55 с.

Имитационная модель единого ресурса алгоритмических схем

Лукьянова Е. А., Дереза А. В.
lukyanovaea@mail.ru
Симферополь, ТНУ им. В. И. Вернадского

Предлагается метод совмещения различных функционирующих схем в один работающий комплекс. Решается задача комбинирования последовательного и параллельного выполнения исходных процессов за счет выполнения APPLY-операции для OBDD, представляющих автоматы соответствующих сетей Петри алгоритмических схем, составляющих работающий комплекс.

Применение методов формальной верификации для проверки правильности проектируемой системы предполагает формализацию реальной системы. Строится модель исследуемой системы. Однако в реальности моделировать приходится сложные системы, которые сами представляют некоторую совокупность двух или нескольких систем. В настоящей работе предлагается метод, позволяющий построить модель сложной многоблоковой реальной системы, организованной до результирующего конечного автомата, реализующего работу всего комплекса. Это дает возможность в дальнейшем провести обширную проверку свойств исходной системы. Предлагаемый метод использует формализм сетей Петри (СП) [1], компактное представление конечных систем, в частности конечных автоматов (КА), упорядоченными бинарными диаграммами решений (УБДР, или OBDD — Ordered Binary Decision Diagrams [2]) и применяет APPLY-операцию к OBDD конечных автоматов сетей Петри, моделирующих исходные алгоритмические схемы.

В статье предлагаемый метод продемонстрирован на работе единого ресурса, который содержит две криптографические схемы Диффи-Хелмана и RSA.

Предварительные сведения

BDD — графическое представление булевой функции, представляющая собой ориентированный ациклический граф $G = (V, E)$, где V — множество вершин графа. Каждая внутренняя (нетерминальная) вершина дерева решений v обозначается символом переменной $var(v)$ и имеет дуги, которые ведут к сыновьям: сын $l(v)$ (дуга, обозначаемая пунктиром) отвечает значению переменной 0 и сын $h(v)$ (дуга — сплошная линия) отвечает значению переменной 1. Все вершины-листки дерева (терминальные вершины) обозначаются константами 0 или 1, и не имеют выходящих дуг. Для заданного распределения логических значений (интерпретаций) переменных, путь в дереве, который отвечает этому распределению, ведёт от корня дерева к вершине – листку, обозначение которой есть значение функции.

Для вычисления булевой функции $F(x_1, x_2, \ldots, x_n)$, BDD используется следующим образом: каждый вход $x = (x_1, x_2, \ldots, x_n) \in \{0,1\}^n$ определяет вычислительный путь через BDD, начинающийся с корня. Если путь достигает нетерминальный узел v — он следует по дуге $l(v)$, если $x_i = 0$, и по дуге $h(v)$, если $x_i = 1$. На всех путях достигается терминальный узел, так как граф направленный и ациклический. BDD с заданным порядком на переменных называется $OBDD$. От выбора подходящего порядка зависит размер и эффективность манипуляций с $OBDD$.

Использование $OBDD$ для представления конечного X-автомата $A = (A, X, f, F)$ основано на представлении его функции перехода f в виде отношения R_f, заданного на множестве кодов состояний A X-автомата, где $X = \{0,1\}$, $f : A \times \times X \longrightarrow A$, $F \subseteq A$. Пусть для некоторых состояний $a, b \in A$ и некоторого $x \in X$ имеем $f(a, x) = b$. Если $\underline{x}, \underline{a}, \underline{b}$ — соответственно коды символа x и состояний $a, b \in A$, то

$$R_f(\underline{x}, \underline{a}, \underline{b}) = \begin{cases} 1, & \text{если } f(a,x) = b; \\ 0, & \text{если } f(a,x) \neq b; \\ d, & \text{в других случаях.} \end{cases}$$

Отношению R_f соответствует булева функция $g_f(x, a, b)$, которая равна 1, тогда и только тогда, когда $R_f(\underline{x}, \underline{a}, \underline{b}) = 1$.

Символьные операции над булевыми функциями с применением $OBDD$ можно рассматривать как соответствующие алгоритмы на графах, которые представляют эти $OBDD$. Причем выполнять можно произвольные как угодно сложные последовательности операций, не нарушая заданного порядка для переменных.

$APPLY$-операция строит булеву функцию $f(op)g$ при помощи применения бинарных операций ($(op) \in \{+, \cdot\}$) к функциям-аргументам f и g и выполняет соединение двух графов-аргументов глубины 1 при использовании двух хеш-таблиц. Первая из этих таблиц служит для повышения эффективности вычислений, а вторая — для построения максимально редактированного графа-результата. $APPLY$-операция основывается на возможности перестановки бинарных операций с расширенной операцией сужения для произвольной переменной x:

$f(op)g = \neg x \cdot (f\mid_{x\longleftarrow 0}(op)g\mid_{x\longleftarrow 0}) + x \cdot (f\mid_{x\longleftarrow 1}(op)g\mid_{x\longleftarrow 1})$.

Данное равенство — рекурсивная процедура вычисления $OBDD$, представляющей функцию $f(op)g$.

Операция сужения для функции f, представленной $OBDD$ с вершиной-корнем r_f, относительно переменной $x < var(r_f)$ вычисляется при помощи следующей процедуры:

$$f\mid_{x\longleftarrow b} = \begin{cases} r_f, & \text{если } x < var(r_f)\,; \\ l(r_f), & \text{если } x = var(r_f) \text{ and } b = 0\,; \\ h(r_f), & \text{если } x = var(r_f) \text{ and } b = 1\,. \end{cases}$$

Применение $APPLY$-алгоритма к $OBDD$ выполняется следующим образом. Каждый шаг вычислений определяется корневыми вершинами графов-аргументов. Пусть функции f и g представлены $OBDD$, корни которых r_f и r_g соответственно. В том случае, когда r_f и r_g конечные вершины-листки, они обозначаются константой. Рекурсия заканчивается и получаем вершину, которую обозначаем этой константой.

Если какая-нибудь из вершин r_f или r_g не являются конечными, то выбираем переменную x, которая связывает эти вершины. Выбор такой переменной выполняется по правилу:

$$x = \min(var(r_f), var(r_g))\,.$$

$OBDD$ для функций $f\mid_{x\longleftarrow 0}(op)g\mid_{x\longleftarrow 0}$ и $f\mid_{x\longleftarrow 1}(op)g\mid_{x\longleftarrow 1}$ вычисляются рекурсивно для сужения функций f и g для значения 0 (пунктирная дуга) и для значения 1 (сплошная дуга) соответственно.

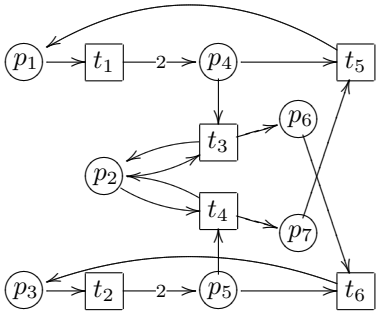
В случае, когда одна из вершин графов-аргументов является листком и представляет доминирующее значение (1 для операции + и 0 для операции $\cdot$) вызов рекурсивной процедуры не нужен, результатом является вершина, обозначенная доминирующим значением.

Для исключения повторных рекурсивных вызовов для одной и той же пары вершин структуру данных организуют в виде хеш-таблицы для сохранения уже построенных пар. При этом вершины представляют собой ключ, при помощи которого в этой таблице находятся пара вершин и построенная вершина-результат для этой пары. При существовании такой хеш-таблицы начало вычислений начинается с проверки для двух аргументов u и v, есть или нет пара с ключом $key(u, v)$ в таблице. Если такая пара уже существует, то результатом будет построенная вершина, которая отвечает этой паре. Если не существует такой пары, то в таблицу записывается пара вершин (u, v) с ключом $key(u, v)$ и выполняются вычисления, описанные выше для пары (u, v), и после окончания вычислений результат записывается в хеш-таблицу с тем же ключом.

Каждый шаг вычислений определяет вершину в строящемся графе. Для того, чтобы сгенерировать редуцированный граф непосредственно, при каждом шаге вычислений необходимо избежать возникновения новой вершины, что достигается путём использования соответствующих правил преобразований $OBDD$.

Функционирующие схемы

Согласно результатам, представленных в [3], работа схемы Диффи-Хелмана моделируется следующей СП:

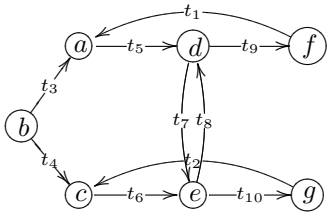


Эта СП имеет конечные множества S и T-инвариантов [3], что позволяет по данной СП построить соответствующую конечную транзиционную систему.

Транзиционная система (ТС) представляет собой четверку $A = (S = \{a, b, c, d, e, f, g\},\ T = \{t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9, t_{10}\}, f, g)$, где S — множество состояний, T — множество переходов, f и g — два отображения из T в S, ставящие в соответствие каждому переходу t из T два состояния $f(t)$ и $g(t)$ — соответственно начало и конец перехода t. Зададим функции $f(t)$ и $g(t)$ с помощью следующей таблицы:

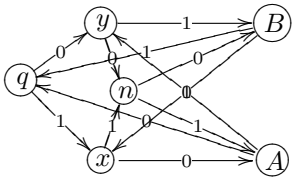
	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
f	f	g	b	b	a	c	d	e	d	e
g	a	c	a	c	d	e	e	d	f	g

Построим конечную ТС:

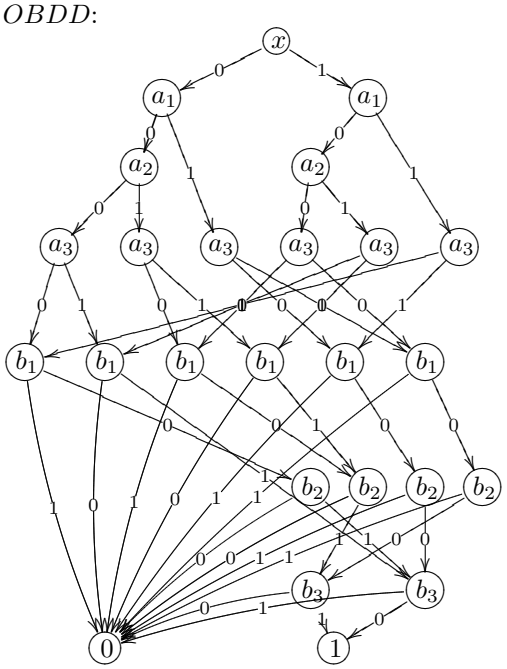


Имея множество состояний S ТС и учитывая таблицу переходов, получим, что построенная ТС есть следующий конечный инициальный бинарный X-автомат (КА).

Граф переходов и выходов автомата имеет вид:



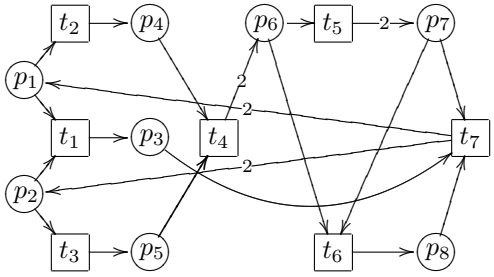
Полученный конечный X-автомат схемы Диффи-Хелмана представляется следующей редуцированной $OBDD$:



Проведя аналогичные [3] и описанные выше исследования для криптографической схемы RSA, получаем ее модель в виде СП, проводим формальный анализ свойств модели на наличие S и T-инвариантов, строим соответствующую ТС, трансформируем ее в КА и строим редуцированную $OBDD$.

На каждом этапе выясняются и проверяются свойства рассматриваемой алгоритмической схемы.

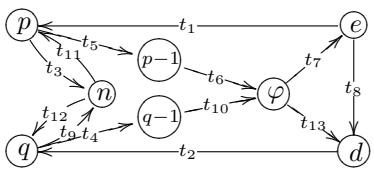
Сеть Петри, моделирующая протокол RSA:



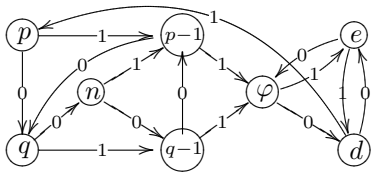
Табличное представление ТС:

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}
f	e	d	p	q	p	$p-1$	φ	e	q	$q-1$	n	n	φ
g	p	q	n	$q-1$	$p-1$	φ	e	d	n	φ	p	q	d

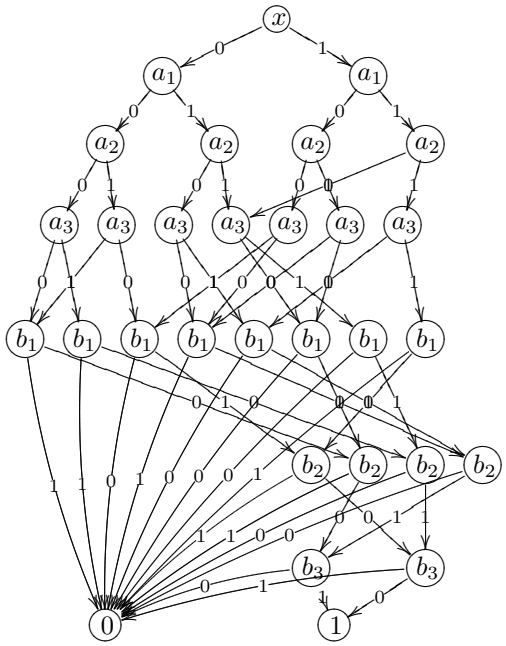
Графовое представление ТС:



Графовое представление инициального X-автомата:



Редуцированная $OBDD$, построенная по КА схемы RSA:



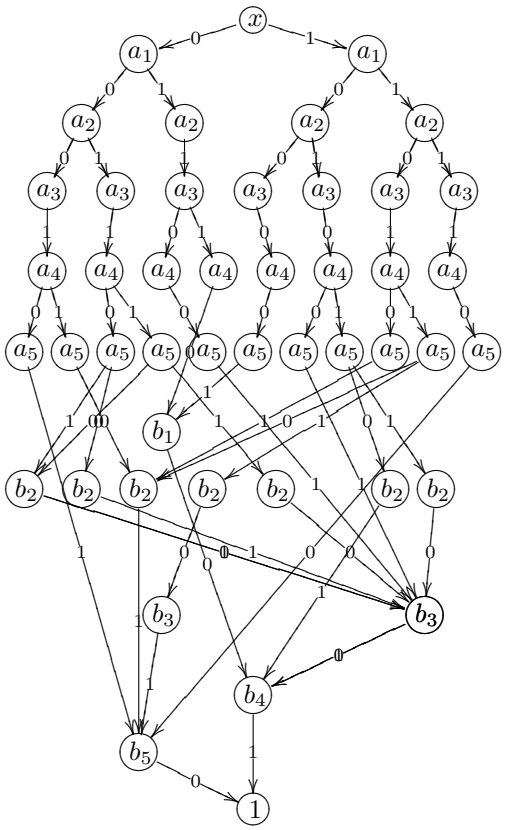
$APPLY$-операция функционирующих схем, результирующий X-автомат

Выполнение $APPLY$-операции для построения $OBDD$ для операции $((op) \in \{+, \cdot\})$, применительно к булевым функциям, сконструированных согласно отношениям R_f относительно кодировки состояний соответствующих X-автоматов, отражающих соответственно работу схем Диффи-Хелмана и RSA, дает результирующий граф, который после применения правил упрощения для $OBDD$ может быть переведет в конечный результирующий X-автомат.

Использование операции $\cdot$ или $+$ отражает либо последовательный характер работы единого ресурса алгоритмических схем, либо параллельное выполнение.

В данной работе демонстрируется результат операции «и».

Результирующая $OBDD$ представляется двенадцати-уровневым графом. Применив к нему правила редуцирования, получили следующую OBDD (дуги, ведущие в листок 0 не указанны, чтобы не загромождать рисунок):



Данная $OBDD$ может быть представлена в виде конечного инициального X-автомата A, насчитывающего 31 состояние. Алфавит входных символов которого представляется множеством $X = \{0, 1\}$. Подмножество $F \subseteq A$ состоит из 15 состояний, называемых заключительными или финальными. Функция переходов $f : A \times X \to A$ задается при помощи таблицы.

Выводы

В работе предложен метод, позволяющий выполнять процедуру сложения или умножения графов различных алгоритмических схем, образующих общий ресурс, что в перспективе дает возможность исследовать свойства всей сложной системы в комплексе. Например, установить важные свойства системы: свойства взаимного исключения (mutex) и справедливости (fairness). В возможностях размера данной статьи работа метода продемонстрирована на примере.

	1	2	3	4	5	6	7	10	12	13	14	15	19	23	26	31
0		8		10	8		8	12			10	8	14		8	
		10		11	9		9	14			11	9	15		9	
		12		12	12		10	18			12	21	27		10	
		14		13	13		11	19			13	29	28		11	
		16		23	21		12	24			23				12	
		17		25	25		12	25			25				13	
		18		30	29		14	26			30				14	
		19		31	30		15	27			31				15	
		20														
		21														
		22														
		23														
		24														
		25														
		26														
		27														
1	8		10			12			12	10				12	8	10
	9		11			14			13	11				14	10	17
	10		14			18			14	23				18	12	22
	11		15			19			15	31				19	14	23
	12		23			24			25					24	16	
	13		27			25			27					25	17	
	14		28			26			28					26	18	
	15		31			27			30					27	19	
															20	
															21	
															22	
															23	
															24	
															25	
															26	
															27	

Литература

[1] *Котов В. Е.* Сети Петри. — М.: Наука, 1984.

[2] *Bryant R.* Symbolic Boolean Manipulation wiht Ordered Binary Decision Diagrams // Shcool of Computer Science, Carnegi Mellon University, Pittsburg. — 1992. — 34 p.

[3] *Лукъянова Е. А.,Дереза А. В.* Об одном процессе верификации алгоритмических схем // International conference “Intelligent Informations Procesing”. IIP-8, Cyprus, Parhos. — 2010.— Pp. 299–302.

Иерархический классификатор на основе древовидно структурированных покрытий*

Ланге М. М., Ганебных С. Н.
lange_mm@mail.ru
Москва, Вычислительный центр им. А. А. Дородницына РАН

Предлагается метод построения классификатора двумерных объектов в пространстве представлений с многоуровневым разрешением. Метод базируется на построении древовидно структурированного покрытия (Tree-Structured Covers) обучающего множества шарами в пространстве представлений с наибольшим уровнем разрешения. Центры покрывающих шаров и их проекции на всех уровнях разрешения порождают многоуровневую сеть эталонов, а сами шары образуют области влияния эталонов на соответствующих уровнях разрешения. Используя многоуровневую структуру сети эталонов, предложен алгоритм иерархического поиска решения, который при большом числе классов c обеспечивает вычислительный выигрыш порядка $2c/\log c$ по сравнению с переборным алгоритмом. Эффективность TSC-классификатора продемонстрирована экспериментальными оценками вероятности ошибок распознавания подписей и жестов руки, а также их сравнением с аналогичными оценками для SVM-классификатора.

Для многих приложений задачу распознавания образов целесообразно рассматривать в терминах соотношения характеристик вычислительной сложности и качества, требуя минимизации вероятности ошибок при заданном ограничении на объём вычислений. В такой постановке задача распознавания может быть решена в пространстве структурных представлений образов с многоуровневым разрешением [1, 2]. Для двумерных объектов с полутоновой окраской, способ построения инвариантных представлений с многоуровневым разрешением предложен в работе [3]. Типичными примерами допустимых объектов для такого способа представления являются подписи, рукописные символы, жесты, лица и силуэты.

В настоящей работе исследуется задача построения и обучения классификатора в пространстве представлений с многоуровневым разрешением, описанном в [3]. Задача обучения сводится к отбору эталонов на основе построения и оптимизации древовидно-структурированного покрытия (TSC) обучающего множества объектов в заданном пространстве представлений. Проведена апробация TSC-классификатора в экспериментах по распознаванию подписей и жестов руки.

Формализация задачи

Многоуровневое представление образов

Пусть $\mathbf{A}$ — множество объектов (образов), в котором каждый образ $A \in \mathbf{A}$ представлен бинарным $(L+1)$-уровневым деревом эллиптических примитивов [3]

$$A^L = (a^0, \ldots, a^l, \ldots, a^L), \qquad (1)$$

где $a^l\colon\ l+1 \leqslant ||a^l|| \leqslant 2^l$ — представление l-го уровня, образованное концевыми вершинами $(l+1)$-уровневого поддерева $A^l = (a^0, \ldots, a^l)$. В общем случае $||a^l|| \leqslant 2^l$, $l = 0, \ldots, L$, все поддеревья $A^l \in A^L$, включая дерево A^L, являются завершенными, а в случае знака равенства — полными. Примеры представлений a^l жеста руки на уровнях $l = 0, 2, 4, 6, 8$ даны на рис. 1.

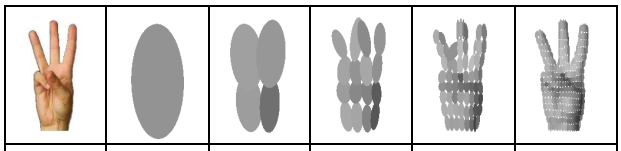


Рис. 1. Представления жеста руки.

Модель классификатора

Будем считать, что множество $\mathbf{A} = \{\mathbf{A}_i\}_{i=0}^c$ содержит объекты, принадлежащие $c+1$ классам, где каждый класс $\mathbf{A}_i$ с номером $i \neq 0$ включает семантически однородные объекты, а класс $\mathbf{A}_0$ объединяет все прочие объекты. На множестве $\mathbf{A}$ вводятся вероятности классов $P_i = P(\mathbf{A}_i),\ i = 0, \ldots, c$, которые определяют априорные вероятности

$$P_{\text{own}} = \sum_{i=1}^{c} P_i,\ P_{\text{alien}} = P_0 = 1 - P_{\text{own}}, \qquad (2)$$

«своих» и «чужих» объектов для классификатора, содержащего описания всех ненулевых классов.

В работе [3] для любой пары объектов $(A, \hat{A}) \in \in \mathbf{A}$ заданных представлениями (1), введена мера различия

$$d_l(A, \hat{A}) = d(A^l, \hat{A}^l),\ l = 0, ..., L, \qquad (3)$$

l-го порядка, которая вычисляется по $(l+1)$-уровневым поддеревьям A^l и $\hat{A}^l$. Мера (3) позволяет определить на множестве $\mathbf{A}$ шар

$$S_l(\hat{A}, D_l(\hat{A})) = \{A|\ d_l(A, \hat{A}) \leqslant D_l(\hat{A})\} \qquad (4)$$

l-го уровня с центром $\hat{A} \in \mathbf{A}$, представленным поддеревом $\hat{A}^l$, и заданным радиусом $D_l(\hat{A}) \geqslant 0$.

Работа выполнена при финансовой поддержке РФФИ, проект № 09-01-00573.

Для обучения используется обучающее множество объектов

$$\mathbf{B} = \big\{\mathbf{B}_i = \{B_{ij}\}_{j=1}^{m_i}\big\}_{i=1}^{c} \subset \mathbf{A} \tag{5}$$

где $\mathbf{B}_i \subset \mathbf{A}_i$ — i-й кластер мощности $m_i = \|\mathbf{B}_i\|$. Пусть

$$\mathbf{S}_l(\hat{\mathbf{B}}, \mathbf{D}_l(\hat{\mathbf{B}})) = \big\{\{S_l(\hat{B}_{ij}, D_l(\hat{B}_{ij}))\}_{j=1}^{\hat{m}_i}\big\}_{i=1}^{c} \tag{6}$$

множество шаров вида (4), выбранных в кластерах с номерами $i \neq 0$ так, что $1 \leqslant \hat{m}_i \leqslant m_i$ и шары l-го уровня являются «проекциями» соответствующих шаров L-го уровня при всех $l = 0, \ldots, L$. В пространстве представлений (1) множество центров

$$\hat{\mathbf{B}} = \big\{\hat{\mathbf{B}}_i = \{\hat{B}_{ij}\}_{j=1}^{\hat{m}_i}\big\}_{i=1}^{c} \tag{7}$$

шаров в (6) образует $(L+1)$-уровневую сеть эталонов, в которой l-й уровень представлен множеством $(l+1)$-уровневых поддеревьев

$$\hat{\mathbf{B}}^l = \big\{\hat{\mathbf{B}}_i^l = \{\hat{B}_{ij}^l\}_{j=1}^{\hat{m}_i}\big\}_{i=1}^{c} \tag{8}$$

и множеством радиусов

$$\mathbf{D}_l(\hat{\mathbf{B}}) = \big\{\mathbf{D}_l(\hat{\mathbf{B}}_i) = \{D_l(\hat{B}_{ij})\}_{j=1}^{\hat{m}_i}\big\}_{i=1}^{c} \tag{9}$$

Используя множества (8) и (9), введем меру сходства l-го порядка между объектом $A \in \mathbf{A}$ и подмножеством эталонов $\hat{\mathbf{B}}_i$:

$$\mu_l(A, \hat{\mathbf{B}}_i) = \max_{j=1}^{\hat{m}_i} 2^{-d_l(A, \hat{B}_{ij})/D_l(\hat{B}_{ij})}, \tag{10}$$

которая даёт положительный голос подмножества представлений $\hat{\mathbf{B}}_i^l$ l-го уровня в пользу принадлежности объекта A классу $\mathbf{A}_i$. Правило классификации строится с использованием значений меры $\mu_L(A, \hat{\mathbf{B}}_i)$, $i = 1, \ldots, c$, вида (8) L-го порядка и пороговых величин Δ_i : $0 < \Delta_i \leqslant \frac{1}{2}$, $i = 1, \ldots, c$, и состоит в вычислении максимума

$$\mu_L^*(A, \hat{\mathbf{B}}_k) = \max_{i=1}^{c} \big(\mu_L(A, \hat{\mathbf{B}}_i)[\mu_L(A, \hat{\mathbf{B}}_i) \geqslant \Delta_i]\big),$$

и номера класса

$$i^* = k[\mu_L^*(A, \hat{\mathbf{B}}_k) > 0], \tag{11}$$

где логическое выражение в квадратных скобках принимает значение 1 или 0. Вероятность ошибки классификатора с решающим правилом (11) определяется математическим ожиданием

$$\varepsilon = \varepsilon_{\text{own}} P_{\text{own}} + \varepsilon_{\text{alien}} P_{\text{alien}} \tag{12}$$

по распределению (2), где ε_{own} и $\varepsilon_{\text{alien}}$ — вероятности ошибочно классифицируемых «своих» и «чужих» объектов.

Задача исследования

Решается задача обучения классификатора на множестве объектов вида (5) в пространстве представлений (1). Обучение состоит в построении множества шаров (6) L-го порядка, которое образует древовидно-структурированное покрытие множества $\mathbf{B}$. Оптимизация TSC-классификатора выполняется по числу покрывающих шаров (эталонов) $\{\hat{m}_i\}_{i=1}^c$ и порогам $\{\Delta_i\}_{i=1}^c$, минимизирующим оценку вероятности ошибок $\varepsilon_{\mathbf{B}}$ вида (12) на множестве $\mathbf{B}$. Для обученного TSC-классификатора вычисляются оценки вероятности ошибки $\varepsilon_{\mathbf{A}\setminus\mathbf{B}}$ вида (12) на тестовом множестве $\mathbf{A} \setminus \mathbf{B}$. Эти оценки сравниваются с аналогичными показателями SVM-классификатора. Рассматриваются два источника объектов: подписи [4] и жесты руки [5] при различных априорных распределениях $(P_{\text{own}}, P_{\text{alien}})$. Исследуется вычислительная сложность алгоритма иерархического поиска решения в многоуровневой базе эталонов. Приводятся сравнительные оценки вычислительной сложности иерархического и переборного алгоритмов.

TSC-классификатор

Древовидно-структурированное покрытие

Покрытие множества $\mathbf{B}_i$, $i = 1, \ldots, c$ формируется на объединении покрытий кластеров $\mathbf{B}_i \in \mathbf{B}$, $i = 1, \ldots, c$, и строится в два этапа. На первом этапе для каждого кластера $\mathbf{B}_i$ определяется оптимальная пара параметров $(\hat{m}_i, \Delta_i)$, доставляющая минимум ошибки классификации объектов обучающего множества $\mathbf{B}$ при выполнении скользящего контроля. На втором этапе для каждого кластера $\mathbf{B}_i$ мощности m_i и найденного оптимального значения $\hat{m}_i \leqslant m_i$ строится подмножество шаров покрытия

$$\mathbf{S}_L(\hat{\mathbf{B}}_i, D_l(\hat{\mathbf{B}}_i)) = \{S_L(\hat{B}_{ij}, D_L(\hat{B}_{ij}))\}_{j=1}^{\hat{m}_i} \tag{13}$$

с использованием дихотомической декомпозиции кластера $\mathbf{B}_i$ на $\hat{m}_i$ непересекающихся сегментов $\mathbf{B}_{ij}$ мощности $m_{ij} = \|\mathbf{B}_{ij}\|$, так что $\mathbf{B}_i = \cup_{j=1}^{\hat{m}_i} \mathbf{B}_{ij}$, $m_i = \sum_{j=1}^{\hat{m}_i} m_{ij}$, и выбора покрывающих шаров для указанных сегментов.

На каждом шаге дихотомии разбиению подвергается сегмент наибольшей мощности. В разбиваемом сегменте $\mathbf{B}_{ij}$ выделяется пара наиболее удалённых друг от друга опорных объектов:

$$(\hat{B}', \hat{B}'') = \arg \max_{(B' \in \mathbf{B}_{ij}, B'' \in \mathbf{B}_{ij})} d_L(B', B''),$$

и $\mathbf{B}_{ij}$ разбивается на два сегмента $\{\mathbf{B}_{ij'}, \mathbf{B}_{ij''}\}$ так, что $\mathbf{B}_{ij'}$ и $\mathbf{B}_{ij''}$ содержат наиболее близкие объекты по заданной мере (3) L-го порядка к соответствующим опорным объектам $\hat{B}'$ и $\hat{B}''$. В результате дихотомической декомпозиции кластера и выбора покрывающих шаров для всех сегментов, строится завершенное бинарное дерево, которое содер-

жит $\hat{m}_i - 1$ промежуточных вершин (делимых сегментов) и $\hat{m}_i$ концевых вершин (неделимых сегментов). Шары покрытия неделимых сегментов образуют подмножество вида (13).

Стратегия выбора центра $\hat{B}_{ij'}$ и радиуса $D_L(\hat{B}_{ij'})$ шара $S_L(\hat{B}_{ij'}, D_L(\hat{B}_{ij'}))$, покрывающего текущий сегмент $\mathbf{B}_{ij'}$ из пары $\{\mathbf{B}_{ij'}, \mathbf{B}_{ij''}\}$, состоит в следующем. Пусть

$$\hat{d}_L(\mathbf{B}_{ij'}, \hat{B}) = \max_{B \in \mathbf{B}_{ij'}} d_L(B, \hat{B})$$

радиус сегмента $\mathbf{B}_{ij'}$ по мере (3) L-го порядка относительно объекта $\hat{B} \in \mathbf{B}_{ij'}$. Центр покрывающего шара для сегмента $\hat{\mathbf{B}}_{ij'}$ выбирается как центр этого сегмента

$$\hat{B}_{ij'} = \arg \min_{\hat{B} \in \mathbf{B}_{ij'}} \hat{d}_L(\mathbf{B}_{ij'}, \hat{B}),$$

а радиус $D_L(\hat{B}_{ij'})$ определяется величиной

$$\max(\hat{d}_L(\mathbf{B}_{ij'}, \hat{B}_{ij'}),\ d_L(B_{ij''}, \hat{B}_{ij'})),$$

где $\hat{d}_L(\mathbf{B}_{ij'}, \hat{B}_{ij'})$ — радиус сегмента $\mathbf{B}_{ij'}$ относительно центра $\hat{B}_{ij'}$, а

$$d_L(B_{ij''}, \hat{B}_{ij'}) = \min_{B \in \mathbf{B}_{ij''}} d_L(B, \hat{B}_{ij'})$$

величина различия по мере (3) между центром $\hat{B}_{ij'} \in \mathbf{B}_{ij'}$ и ближайшим к нему объектом $B_{ij''} \in \mathbf{B}_{ij''}$. Радиус сферы покрытия исходного кластера $\mathbf{B}_i$ определяется его радиусом. Объединение покрытий (13), построенных для c кластеров множества $\mathbf{B}$, даёт древовидно-структурированное множество шаров (4) L-го порядка.

Оптимизация параметров

Поиск оптимальных значений $\{\hat{m}_i, \Delta_i\}_{i=1}^c$ выполняется с помощью процедуры скользящего контроля в модификации «leave-one-out» для каждого кластера $\mathbf{B}_i$, $i = 1, \ldots, c$, в обучающем множестве вида (5). В такой схеме «свои» объекты представлены подмножеством $\mathbf{B}_i$, а «чужие» — подмножеством $\mathbf{B} \setminus \mathbf{B}_i$. В качестве оценки вероятности ошибки скользящего контроля использована функция

$$\varepsilon_{\mathbf{B}}(\hat{m}_i, \Delta_i) = P_{\text{own}}\big(\varepsilon_{\mathbf{B}} + \varepsilon_{\mathbf{B}\setminus\mathbf{B}_i}(c-1)/c\big) + P_{\text{alien}}\varepsilon_{\mathbf{B}\setminus\mathbf{B}_i},$$

где $\varepsilon_{\mathbf{B}_i}$ и $\varepsilon_{\mathbf{B}\setminus\mathbf{B}_i}$ — доли ложных отказов и ложных распознаваний при предъявлении «своих» и «чужих» объектов текущему покрытию i-го кластера с числом эталонов $\hat{m}_i$ при пороге Δ_i. Оптимизация по i-му кластеру, сводится к нахождению пары

$$(\hat{m}_i, \Delta_i) = \arg \min_{1 \leqslant \hat{m}'_i \leqslant m,\ 0 \leqslant \Delta_i \leqslant 1/2} \varepsilon_{\mathbf{B}}(\hat{m}'_i, \Delta'_i), \quad (14)$$

где m — заданное максимальное число эталонов, ограниченное сверху величиной $\min\limits_{i=1}^{c} m_i - 1$. Параметры (14) доставляют минимум оценки вероятности ошибки скользящего контроля

$$\varepsilon_{\mathbf{B}} = \frac{1}{c} \sum_{i=1}^{c} \varepsilon_{\mathbf{B}}(\hat{m}_i, \Delta_i).$$

для TSC-классификатора, построенного на объединении покрытий c кластеров множества $\mathbf{B}$.

Вычислительная сложность

В основе алгоритма иерархического поиска решения лежит стратегия сужения областей поиска на последовательных уровнях разрешения сети эталонов. Для этой цели вводится функция, определяющая число классов, анализируемых на l-м уровне сети эталонов:

$$c_l = \left\lfloor c 2^{-\beta l} \right\rfloor,\ l = 0, \ldots, L,$$

где $\beta = \frac{1}{L}\log(c/c_L) > 0$ и $c_L \geqslant 1$. На l-м уровне иерархический алгоритм вычисляет значения меры сходства (10) для c_l классов множества эталонов $\hat{\mathbf{B}}^l$ и среди них выбирает c_{l+1} классов с наибольшими значениями меры сходства l-го порядка для последующего анализа в множестве $\hat{\mathbf{B}}^{l+1}$. Для получения решения (11), максимум вычисляется по c_L классам множества $\hat{\mathbf{B}}^L$, которые отобраны в $\hat{\mathbf{B}}^{L-1}$. Для сравнения переборный алгоритм поиска решения использует полное множество эталонов $\hat{\mathbf{B}}^L$ L-го уровня.

Вычислительную сложность обоих алгоритмов поиска решения будем измерять числом вершин-примитивов, которые обрабатываются в представлениях эталонов. Тогда верхняя оценка сложности может быть получена для представлений в виде полных $(L+1)$-уровневых деревьев. При заданных параметрах L, $c \geqslant c_L \geqslant 1$, $\{\hat{m}_i\}_{i=1}^c$, и условии $2^L \leqslant c/c_L$, оценки сложности иерархического и переборного алгоритмов имеют вид

$$C_h \leqslant c(1 + \log c) \max_{i=1}^{c} \hat{m}_i$$

$$C_f \leqslant 2c^2 \max_{i=1}^{c} \hat{m}_i.$$

При большом числе классов c полученные оценки демонстрируют вычислительный выигрыш порядка $2c/\log(c)$ для иерархического алгоритма относительно переборного.

Экспериментальные результаты

Апробация TSC-классификатора выполнена для двух источников реальных объектов, взятых из базы изображений подписей [4] и базы изображений жестов руки, соответствующих буквам латинского алфавита [5]. Множество объектов $\mathbf{A}$ от каждого источника многократно разбивалось случайным образом в равных долях на обучающую выборку $\mathbf{B}^{(k)}$ и тестовую выборку $\mathbf{A}\setminus\mathbf{B}^{(k)}$. Множество

Таблица 1. Экспериментальные оценки вероятности ошибки распознавания подписей и жестов.

P_{own}		1	1	0.75	0.5
Классификатор		SVM	TSC	TSC	TSC
`signatures`	ε	0.0061	0.0070	0.0130	0.0116
	σ	0.0040	0.0019	0.0028	0.0035
`gestures`	ε	0.0019	0.0042	0.0086	0.0070
	σ	0.0019	0.0040	0.0050	0.0053

выборок $\{\mathbf{B}^{(k)}\}_{k=1}^{K}$ использовалось для построения серии TSC-классификаторов и вычисления усреднённой доли ошибок обучения $\varepsilon_{\mathbf{B}}$. Тестирование построенных классификаторов и получение для них усреднённой доли ошибок $\varepsilon_{\mathbf{A}\setminus\mathbf{B}}$ выполнено на множестве тестовых выборок $\{\mathbf{A}\setminus\mathbf{B}^{(k)}\}_{k=1}^{K}$.

При заданном априорном распределении вероятностей: P_{own}, $P_{\text{alien}} = 1 - P_{\text{own}}$, kq-я реализация TSC-классификатора $k = 1,\ldots,K$, $q = 1,\ldots,M$ строилась на объединении $c^* = \lfloor cP_{\text{own}}\rfloor \leqslant cP_{\text{own}}$ случайно выбранных покрытий вида (13) взятых с оптимальными значениями соответствующих порогов Δ. В результате каждая kq-я реализация TSC-классификатора определялась тройкой множеств $\hat{\mathbf{B}}^{(kq)}$, $\mathbf{D}_L(\hat{\mathbf{B}}^{(kq)})$, $\{\Delta_i^{(kq)}\}_{i=1}^{c}$. Для каждой реализации подмножество «своих» объектов определялось выбранными c^* классами, а подмножество «чужих» объектов — остальными $(c-c^*)$ классами. Используя подмножества «своих» и «чужих» объектов, для каждой qk-й реализации TSC-классификатора вычислялись доли ошибок $\varepsilon_{\mathbf{A}\setminus\mathbf{B}^{(kq)}}(\hat{B}^{(kq)}, D_L(\hat{B}^{(kq)}))$ вида (12) и их среднее значение

$$\varepsilon_{\mathbf{A}\setminus\mathbf{B}} = \frac{1}{KM}\sum_{k=1}^{K}\sum_{q=1}^{M}\varepsilon_{\mathbf{A}\setminus\mathbf{B}^{(kq)}}(\hat{\mathbf{B}}^{(kq)}, \mathbf{D}_L(\hat{\mathbf{B}}^{(kq)}))$$

при различных априорных вероятностях P_{own}.

Исходные объекты из базы подписей, заданных сигналами пера, были преобразованы в 8-битовые изображения размера 512×512, а объекты из базы жестов были заданы 8-битовыми изображениями размера 256×256 пикселей. Представления объектов от обоих источников строились в форме $(L+1)$-уровневых полных деревьев с параметром $L=10$ для подписей и $L=8$ для жестов. В эксперименте с подписями использовано $\|\mathbf{A}\| = 800$ объектов от $c=40$ персон (классов), по двадцать реализаций в каждом классе. Множество подписей разбивалось на равные по объёму выборки: $\{\mathbf{B}^{(k)}, \mathbf{A}\setminus\mathbf{B}^{(k)}\}_{k=1}^{K}$, $K=7$, которые обеспечивали одинаковые размеры кластеров $m_i = \frac{1}{c}\|B^{(k)}\| = \frac{1}{c}\|\mathbf{A}\setminus\mathbf{B}^{(k)}\|$, $i=1,\ldots,c$. В эксперименте с жестами задействовано $\|\mathbf{A}\| = 750$ объектов, которые соответствуют $c=25$ буквам (классам) латинского алфавита, по тридцать реализаций в каждом классе. Эксперимент также проведен на семи выборках $(K=7)$ с размерами кластеров $m_i = 15$, $i = 1,\ldots,25$. В обоих экспериментах $M=1$ при $P_{\text{own}}=1$, и $M=5$ при $P_{\text{own}}<1$.

Усреднённые доли ошибок $\varepsilon = \varepsilon_{\mathbf{A}\setminus\mathbf{B}}$ и их среднеквадратические отклонения σ для TSC-классификатора представлены в таблице 1 при трёх значениях $P_{\text{own}} = c^*/c$. Для сравнения при $P_{\text{own}}=1$ даны аналогичные показатели SVM-классификатора [6] в пространстве признаков, образованном векторами различий L-го порядка между предъявляемыми объектами и объектами обучающей выборки.

Выводы

Предложен новый метод построения классификатора для объектов, заданных изображениями и представленных древовидными описаниями с многоуровневым разрешением. Центральным звеном разработанного метода является процедура обучения классификатора, которая сводится к отбору эталонов и назначению их сфер влияния. Обучение выполнено на основе построения древовидно-структурированных покрытий кластеров обучающего множества шарами и оптимизации параметров покрытий с использованием процедуры скользящего контроля.

Разработанный TSC-классификатор апробирован в экспериментах по распознаванию подписей и жестов руки. Для различных режимов TSC-классификатора получены оценки вероятности ошибки распознавания. Найденные оценки сопоставимы с аналогичными показателями SVM-классификатора. В перспективе предполагается модифицировать процедуру построения покрытий за счёт оптимизации схемы дихотомической декомпозиции кластеров, а также исследовать TSC-классификатор с другими решающими правилами.

Литература

[1] *Berretti S., Del Bimbo A.* A Multiresolution Spatial Partitioning for Shape Representation // Proceedings of ICPR, 2004. — Vol. 2. — Pp. 775–778.

[2] *Torsello A.* Matching Hierarchical Structures for Shape Recognition. PhD thesis. — York University, 2004. — 197 p.

[3] *Ganebnykh S. N., Lange M. M.* Classification of 2D Grayscale Objects in a Space of Multiresolution Representations // Pattern Recognition and Image Analysis. — 2009. — Vol. 19, No. 4. — Pp. 591–602.

[4] `www.cse.ust.hk/svc2004` — First International Signature Verification Competition — 2004.

[5] `vision.auc.dk/~tbm/Gestures/database.html` — Thomas Moeslund's Gesture Database — 2002.

[6] `opencv.willowgarage.com` — Open Source Computer Vision library — 2010.

Вероятностная модель классификатора на основе древовидно-структурированных гауссовых смесей*

Новиков Н. А.[1]***, Ланге М. М.***[2]
nikknovikov@gmail.com
Москва, [1]Московский государственный институт радиотехники, электроники и автоматики (МИРЭА),
[2]Вычислительный центр им. А. А. Дородницына РАН

Предложен новый подход к построению байесовского классификатора в многомерном пространстве признаков. Основу подхода составляют древовидно-структурированные гауссовы смеси (TSM) в качестве оценок условных по классам плотностей распределения. Разработана процедура оптимизации TSM-классификатора по числам компонент гауссовых смесей и пороговым значениям смесей, введённым для реализации функции отказа. Оценены доли ошибок TSM-классификатора на множестве трёхмерных признаков монохромного изображения. Для сравнения приведены доли ошибок классификаторов, использующих в качестве оценок функций правдоподобия одиночные гауссовы плотности и гауссовы смеси, построенные с помощью EM-алгоритма. Приложение TSM-классификатора иллюстрируется примерами сегментации спутникового изображения, полученного с помощью программы Google Earth.

Введение

В задачах классификации с большими размерами обучающей выборки, как правило, применяются вероятностные классификаторы. При этом каждый класс рассматривается как множество объектов, чьи признаковые описания являются случайными величинами, подчиняющимися некоторой условной по классу плотности распределения вероятностей (функции правдоподобия). Как правило, условные по классам плотности распределения неизвестны, и построение их оценок составляет часть процедуры обучения [1].

Распространённый подход заключается в построении оценок условных плотностей распределений в виде гауссовых плотностей, однако аппроксимация эмпирических условных распределений является в данном случае очень грубой. Компромисс между точностью аппроксимации и переобучением классификатора может быть достигнут путём построения оценок в виде гауссовых смесей. В настоящей работе предложен байесовский классификатор, который строится на основе древовидного разбиения каждого кластера обучающей выборки на непересекающиеся фрагменты и формирования смеси в виде взвешенной суммы гауссовых плотностей, вычисляемых на этих фрагментах [2]. Числа компонентов смесей и соответствующие пороги отказов определяются в результате оптимизации параметров классификатора на этапе обучения.

Постановка задачи

Рассматриваемая модель классификации подразумевает задание множества классов $\Omega=\{\omega_i\}_{i=1}^c$ с априорными вероятностями $P(\omega_i \mid \Omega)$ и множества объектов $\boldsymbol{X}=\{x\}$ с условными по классам плотностями распределения вероятностей $P(x \mid \omega_i)$, где каждый объект x является вектором в многомерном пространстве признаков $\boldsymbol{X}$. Предполагается, что существуют объекты, которые не относятся ни к одному из классов в Ω и, следовательно, такие объекты должны быть отнесены к классу отказов $\omega_0 \notin \Omega$.

В байесовской модели решение относительно объекта x принимается с использованием апостериорных вероятностей:

$$P(\omega_i \mid x) = \frac{P(\omega_i \mid \Omega)P(x \mid \omega_i)}{\sum\limits_k P(\omega_k \mid \Omega)P(x \mid \omega_k)}, \quad i = 1, \ldots, c. \quad (1)$$

Для определения решающего правила, введем для каждой условной по классу $\omega_i \in \Omega$ плотности распределения пороговое значение $\delta_i > 0$, которое допускает принадлежность объекта x к классу ω_i только в том случае, если $P(x \mid \omega_i) \geqslant \delta_i$. Используя апостериорные вероятности (1), можно ввести функцию

$$q(x, \omega_i) = P(\omega_i \mid x)[P(x \mid \omega_i) \geqslant \delta_i],$$

где за $[*]$ обозначен индикатор, принимающий значение 1, если условие в скобках выполняется, и значение 0 в противном случае. Тогда решающее правило по объекту x имеет вид:

$$\omega(x) = \begin{cases} \arg\max\limits_{\omega_i \in \Omega}(q(x, \omega_i)), & \max\limits_{\omega_i \in \Omega}(q(x, \omega_i)) > 0, \\ \omega_0, & \max\limits_{\omega_i \in \Omega}(q(x, \omega_i)) = 0. \end{cases} \quad (2)$$

Условные плотности распределения $P(x \mid \omega_i)$ и пороги δ_i, необходимые для вычисления функции $q(x, \omega_i)$, заранее не известны, поэтому оценки $\hat{P}(x \mid \omega_i)$ и пороги δ_i для всех классов множества Ω должны быть получены на этапе обучения классификатора.

Пусть $X=\{X_i\}_{i=1}^c \subset \boldsymbol{X}$ — обучающее множество объектов, в котором X_i образует кластер объектов, принадлежащих классу ω_i. Для каждого ω_i оценка $\hat{P}(x \mid \omega_i)$ выбирается из некоторого заданного се-

Работа выполнена при финансовой поддержке РФФИ, проект № 09-01-00573-а

мейства функций таким образом, чтобы максимизировать выборочное правдоподобие [1]:

$$L(\omega_i) = \sum_{x \in X_i} \ln\left(\hat{P}(x \,|\, \omega_i)\right) \to \max. \qquad (3)$$

Оценки $\hat{P}(x \,|\, \omega_i)$, $i{=}1, \ldots, c$, дающие хорошие аппроксимации условных по классам эмпирических распределений, могут быть построены в виде гауссовых смесей. Точность аппроксимации определяется числами компонент смеси n_i, которые должны быть найдены на этапе обучения классификатора.

Наиболее известным способом построения гауссовых смесей является EM-алгоритм [3]. Он позволяет найти среди смесей с заданным числом компонент реализацию, доставляющую глобальный максимум выборочного правдоподобия (3), но имеет большую вычислительную сложность. Древовидная структуризация процедуры построения смеси позволяет уменьшить вычислительные затраты, обеспечивая при этом высокую точность аппроксимации эмпирических распределений.

Построение гауссовых смесей и обучение классификатора

Древовидно-структурированные гауссовы смеси. Для заданного класса $\omega_i{\in}\Omega$ предлагаемый TS-алгоритм строит по выборке X_i' из кластера X_i оценку $\hat{P}(x \,|\, \omega_i)$ в виде древовидно структурированной гауссовой смеси (TSM) с заданным числом компонент n_i. Оценка строится в результате оптимальной рекурсивной декомпозиции выборки X_i' на фрагменты и описания фрагментов нормальными плотностями распределения, взвешенная сумма которых порождает древовидно-структурированную гауссову смесь. В такой схеме фрагменты $X_{i,j}'$ выборки X_i' соответствуют узлам завершенного бинарного дерева декомпозиции. Оценка $\hat{P}(x \,|\, \omega_i)$ строится на описаниях n_i концевых узлов этого дерева и имеет вид:

$$\hat{P}(x \,|\, \omega_i) = \sum_{j\,:\,X'_{i,j}-\text{концевой узел}} \alpha_{i,j}\, \mathcal{N}_{i,j}(x), \qquad (4)$$

где $\alpha_{i,j}$ — весовой коэффициент, определяемый долей фрагмента $X_{i,j}'$ в выборке X_i', а $\mathcal{N}_{i,j}(x) = \mathcal{N}(\mu_{i,j}, \Sigma_{i,j})$ — нормальная плотность распределения, в которой вектор математических ожиданий $\mu_{i,j}$ и ковариационная матрица $\Sigma_{i,j}$ образованы соответствующими выборочными средними значениями и ковариационными моментами, вычисленными на подмножестве объектов, принадлежащих фрагменту $X_{i,j}'$.

Переход в дереве декомпозиции от текущего фрагмента $X_{i,j}'$ к паре новых фрагментов $\left\{X'_{i,j'}; X'_{i,j''}\right\}$, $j''{=}j'+1$, следующего уровня осуществляется путём разбиения $X_{i,j}'$ гиперплоскостью, проходящей через точку $\mu_{i,j}$ перпендикулярно собственному вектору матрицы $\Sigma_{i,j}$, которому соответствует наибольшее собственное значение. На каждом шаге декомпозиции выбирается фрагмент $X_{i,j}'$, разбиение которого порождает оценку $\hat{P}(x \,|\, \omega_i)$, доставляющую наибольшее значение выборочному правдоподобию (3) по объектам полной выборки X_i'. Решающее правило (2) с оценками $\hat{P}(x \,|\, \omega_i)$, $i{=}1, \ldots, c$, вида (4), порождает байесовский TSM-классификатор.

В процессе рекурсивной декомпозиции, выполняемой TS-алгоритмом, описание каждого фрагмента выборки строится однократно. С учётом максимизации правдоподобия (3) на каждом шаге TS-алгоритма, число обрабатываемых фрагментов оценивается удвоенным числом узлов бинарного дерева декомпозиции, и составляет величину порядка $4n_i$ для n_i-компонентной смеси. При построении n_i-компонентной смеси с помощью итеративного EM-алгоритма формирование каждой компоненты требует s итераций и в целом sn_i итераций. Поэтому при сопоставимых вычислительных затратах TS-алгоритма на обработку одного фрагмента и затратах EM-алгоритма на одну итерацию на компоненту, выигрыш в вычислительной сложности TS-алгоритма по сравнению с EM-алгоритмом оценивается величиной порядка $s/4$.

Оптимизация числа компонент смеси и порога отказа. Правило принятия решения (2) и алгоритм построения гауссовой смеси вида (4) требуют предварительного задания пар параметров (n_i, δ_i), $i{=}1, \ldots, c$. Во избежание переобучения, настройка параметров должна осуществляться на основе скользящего контроля [4]. Разобьем обучающее множество $X{=}\{X_i\}_{i=1}^{c}$ на два подмножества: $X{=}U{\cup}V$. Подмножество $U{\subset}X$ используется для оптимизации параметров, а подмножество $V{\subset}X$ — для построения и тестирования классификатора с выбранными оптимальными параметрами.

В предлагаемой схеме пары параметров (n_i, δ_i) настраиваются отдельно для каждого класса $\omega_i{\in}\Omega$. Подмножество U разбивается несколькими способами на две равные по объёму выборки: $U^{(k)}$ и $U{\setminus}U^{(k)}$, где $k{=}1, \ldots, N$, а N — число способов разбиения. Для фиксированной текущей пары параметров (n_i, δ_i) и кластера $U_i^{(k)}{\subset}U^{(k)}$ строится гауссова смесь вида (4), которая даёт классификатор $C(U_i^{(k)})$ с решающим правилом (2). Для TSM-классификатора $C(U_i^{(k)})$ вычисляется доля ошибок при предъявлении объектов из выборки $U{\setminus}U^{(k)}$:

$$\varepsilon^{(k)}(n_i, \delta_i) = \frac{1}{\left\|U \setminus U^{(k)}\right\|} N_e\left(U \setminus U^{(k)} \,|\, C\left(U_i^{(k)}\right)\right), \qquad (5)$$

где N_e — число ошибочно классифицированных объектов с учётом ложных отказов по объектам из класса ω_i и ложных решений по объектам из всевозможных классов подмножества $\Omega\backslash\omega_i$. Усреднение доли ошибок (5) по N разбиениям подмножества U даёт среднее значение доли ошибок классификатора $C(U_i^{(k)})$:

$$\varepsilon(n_i, \delta_i) = \frac{1}{N}\sum_{k=1}^{N}\varepsilon^{(k)}(n_i, \delta_i). \qquad (6)$$

В качестве оптимальных параметров выбирается пара (n_i, δ_i), которая доставляет наименьшее значение функции (6).

Построение и тестирование классификатора. Качество классификатора оценивается на основе выборки $V\subset X$. Для обучения используется количество классов, равное c', а элементы остальных $(c - c')$ классов считаются «чужими». Из полного множества классов Ω всевозможными способами выбираются подмножества $\Omega_m,\ m{=}1,\ldots,M$, содержащие c' классов. Для каждого Ω_m множество V разбивается различными способами на две равные по объёму выборки: $V^{(k)}$ и $V\backslash V^{(k)}$, с сохранением относительных вкладов разных классов. Из $V^{(k)}$ выделяется подвыборка $V_m^{(k)}$, включающая только элементы из набора классов Ω_m. Используя параметры, полученные минимизацией ошибок (6), на основе выборки $V_m^{(k)}$ строится TSM-классификатор $C(V_m^{(k)})$ и вычисляется доля ошибок, допускаемых этим классификатором на выборке $V\backslash V^{(k)}$:

$$\varepsilon_m^{(k)} = \frac{1}{\left\|V\backslash V^{(k)}\right\|}N_e\left(V\backslash V^{(k)} \mid C\left(V_m^{(k)}\right)\right), \qquad (7)$$

где N_e — число объектов, на которых была допущена ошибка. Критерий качества классификатора определяется усреднением ошибок (7) по всем M комбинациям из c' классов и по всем K разбиениям для каждой комбинации:

$$\varepsilon_{c'} = \frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K}\varepsilon_m^{(k)}. \qquad (8)$$

Экспериментальные данные

Схема эксперимента. Целью эксперимента является сравнение TSM-классификатора с байесовскими классификаторами, использующими оценки условных по классам плотностей распределений в виде одиночных гауссиан (G-классификатор) и в виде гауссовых смесей, построенных на основе EM-алгоритма (EMM-классификатор).

В качестве обучающих и тестовых объектов используются векторы текстурных признаков, вычисленные для пикселей полутонового снимка земной поверхности, полученного средствами Google Earth. Каждый вектор содержит три признака, вычисленные в окрестности размера 9×9 пикселей: 6-й и 9-й признаки Харалика [5] и коэффициент автокорреляции со сдвигом на два пикселя. Рассматриваются три класса объектов, которые соответствуют различным типам местности: полям (класс 1), лесным массивам (класс 2) и городским застройкам (класс 3). Векторы текстурных признаков пикселей из заданного набора областей образуют исходное множество $X{=}\{X_i\}_{i=1}^{c}$.

Исходное множество разбивалось случайным образом на два равных по размерам подмножества: $X{=}U{\cup}V$, с сохранением относительных долей объектов из различных классов. Подмножество U использовалось для настройки параметров (чисел компонент гауссовых смесей и порогов отказа), а подмножество V — для построения классификаторов и их тестирования.

При настройке параметров TSM-классификатора для каждого класса была вычислена матрица ошибок вида (6) при различных значениях пары параметров (n_i, δ_i) с использованием $N{=}10$ разбиений. Для каждого n_i выбиралось оптимальное значение δ_i, при котором доля ошибок ε_i минимальна. Затем, используя оптимальные значения δ_i, строилась зависимость ошибки от n_i и выбиралось то значение n_i, при котором ошибка переставала существенно меняться. При построении EMM-классификатора использовались те же числа компонент, что и для TSM. Пороги для классификаторов G и EMM настраивались независимо. Для реализации EM-алгоритма применялась библиотека OpenCV [6]. При сравнении классификаторов вычислялись ошибки вида (8) с использованием $K{=}25$ разбиений. Для каждого классификатора вычислялось три значения: ε_1, ε_2 и ε_3, соответствующие случаям обучения на одиночных классах, на парах классов и на всех трех классах одновременно.

Результаты эксперимента. В результате настройки параметров были получены следующие значения:
$(n_1{=}1; d_1{=}1,24)$, $(n_2{=}1; d_2{=}2,14)$, $(n_3{=}1; d_3{=}3,70)$ для G-классификатора;
$(n_1{=}4; d_1{=}1,03)$, $(n_2{=}2; d_2{=}1,24)$, $(n_3{=}4; d_3{=}0,86)$ для TSM-классификатора;
$(n_1{=}4; d_1{=}1,24)$, $(n_2{=}2; d_2{=}0,86)$, $(n_3{=}4; d_3{=}2,57)$ для EMM-классификатора.

Сравнительные оценки качества классификаторов, определяемые долей ошибок вида (8) при значениях $c'{=}1,2,3$, приведены в таблице 1. Данные таблицы показывают, что применение древовидно-структурированных гауссовых смесей дает лучший результат по сравнению с одиночны-

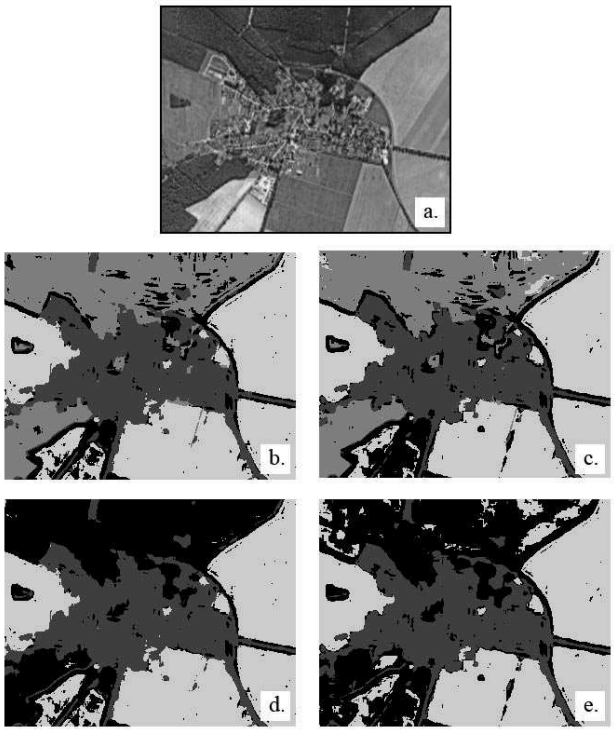


Рис. 1. Результаты сегментации изображения фрагмента земной поверхности (чёрным помечены отказы). (a) Исходное изображение. (b) Результат TSM-классификатора с обучением на 3-х классах. (c) Результат G-классификатора с обучением на 3-х классах. (d) Результат TSM-классификатора с обучением на 2-х классах. (e) Результат G-классификатора с обучением на 2-х классах.

Таблица 1. Оценки качества классификаторов

	ε_1	ε_2	ε_3
G	0.0352	0.0510	0.0490
TSM	0.0174	0.0266	0.0275
EMM	0.0140	0.0219	0.0242

ми гауссовыми плотностями, причём доля ошибок TSM-классификатора близка к доле ошибок EMM-классификатора. Апробация TSM-классификатора выполнена в экспериментах по сегментации полутонового изображения на основе поэлементной классификации пикселей. Выбрана область сегментации, которая не пересекается с обучающими областями. Эксперимент проводился в условиях обучения на всех трёх классах и на паре классов (поля и городские застройки). Для сравнения аналогичные эксперименты проведены с применением G-классификатора. Результаты сегментации представлены на рис. 1. Видно, что G-классификатор при обучении на двух классах даёт большое число ложных распознаваний «чужих объектов» вследствие грубой аппроксимации кластеров. Для TSM-классификатора ложных распознаваний не наблюдается, что свидетельствует о корректности границ кластеров, определяемых построенными оценками условных плотностей распределения и найденными значениями порогов отказа.

Выводы

В работе предложен байесовский TSM-классификатор, использующий оценки условных по классам плотностей распределений, построенные на основе рекурсивной декомпозиции кластеров признакового пространства. Разработанная процедура требует меньших вычислительных затрат по сравнению с EM-алгоритмом. Реализована оптимизация чисел компонент смесей и порогов отказа на основе скользящего контроля. Проведено тестирование TSM-классификатора на множестве трёхмерных векторов текстурных признаков монохромного изображения. Установлено, что оценки качества TSM-классификатора сопоставимы с аналогичными показателями классификатора, построенного с использованием EM-алгоритма. Эксперименты с обучением на неполном множестве классов подтверждают корректность найденных оценок плотностей и порогов отказа. В перспективе планируется улучшить качество оценок условных распределений за счёт усовершенствования алгоритма рекурсивной декомпозиции кластеров. Предполагается исследовать применимость TSM-классификатора для решения задач из стандартных репозиториев.

Литература

[1] *Duda R., Hart P., Stock D.* Pattern classification. Second Edition. — Wiley, 2001.

[2] *Ganebnykh S., Lange M.* Classification of 2D Grayscale Objects in a Space of the Multiresolution Representations // Pattern Recognition and Image Analysis, Nauka-Interperiodica — 2009. — Vol. 19, No. 4.— Pp. 591–602.

[3] *Воронцов К. В.* Курс лекций «Математические методы обучения по прецедентам». Байесовские алгоритмы классификации. `http://www.ccas.ru/voron/download/Bayes.pdf`

[4] *Theodoridis S., Koutroumbas K.* Pattern Recognition. Third Edition. — San-Diego, USA: Academy Press, 2006.

[5] *Miyamoto E., Merryman T., Jr.* Fast calculation of Haralick texture features. `http://ece.cmu.edu/~pueschel/teaching/18-799B-CMU-spring05/material/eizan-tad.pdf`

[6] OpenCV documentation: Expectation-Maximization. `opencv.willowgarage.com/documentation/cpp/expectation-maximization.html`

Распознавание лиц по многослойным древовидным представлениям цветных изображений*

Степанов Д. Ю.[1]***, Ланге М. М.***[2]
lange_mm@ccas.ru
[1] Москва, Московский государственный институт радиотехники, электроники и автоматики
[2] Москва, Вычислительный центр им. А. А. Дородницына РАН

Исследуется новый подход к распознаванию лиц по цветным изображениям в рамках модели классификации двумерных объектов, заданных многоканальными изображениями. Рассматриваемая модель использует пространство многослойных представлений лиц деревьями эллиптических примитивов. Предлагается процедура обучения TSC-классификатора на основе построения древовидно-структурированного покрытия (Tree-Structured Covering) обучающего множества шарами по заданной мере различия многослойных представлений. Эффективность TSC-классификатора продемонстрирована экспериментальными результатами распознавания лиц в пространстве представлений цветных HSI-изображений.

В ряде прикладных задач возникает необходимость в классификации объектов по их образам, полученным от нескольких каналов наблюдения. Примером такой модели является многоканальная система биометрической идентификации личности по изображению лица, форме и рисунку ладони, изображению радужной оболочки, подписи и т. п. [1]. Двумерные объекты, заданные цветными изображениями, могут рассматриваться в рамках трёхканальных моделей RGB и HSI [2], причём вторая модель более предпочтительна для анализа цветных изображений в силу меньшей корреляции каналов H, S и I по сравнению с каналами R, G и B. В частности, модель HSI находит применение в системах распознавания лиц [3].

Наряду с многоканальностью исходных изображений, важным фактором является выбор пространства представлений образов на выходе каналов наблюдения. В данной работе предлагается способ построения многослойных древовидных представлений для объектов, заданных многоканальными изображениями. Такие представления обладают многоуровневым разрешением и ориентированы на снижение вычислительной сложности процедуры распознавания за счёт возможности иерархического поиска решения в многоуровневой базе эталонов. Предлагаемое многослойное древовидное представление применяется для описания лиц, заданных цветными изображениями в модели HSI, и является обобщением древовидного представления, предложенного в [4] для описания полутоновых объектов. На множестве обобщённых древовидных представлений введена мера различия наблюдаемых объектов и построен классификатор по критерию ближайшего эталона, в котором предусмотрена функция отказа. Обучение классификатора базируется на построении и оптимизации древовидно-структурированных покрытий (TSC) семантически однородных кластеров обучающего множества шарами в выбранном пространстве представлений образов. В результате обучения строятся оценки параметров покрытий и формируется множество эталонов, которое образует многоуровневую сеть, удобную для быстрого поиска решающих эталонов.

Экспериментальная апробация TSC-классификатора проведена с использованием базы цветных изображений лиц в модели HSI. Результатами апробации являются сравнительные оценки вероятности ошибок для разработанного TSC-классификатора и SVM-классификатора из библиотеки OpenCV [5].

Работа выполнена при финансовой поддержке РФФИ, проект № 09-01-00573-а.

Модель классификации и задача исследования

Многослойное древовидное представление двумерных объектов

Пусть $\mathbf{A}$ — множество двумерных объектов, в котором каждый объект $A \in \mathbf{A}$ задан набором образов от N каналов наблюдения. Будем считать, что образ от каждого q-го канала представлен в виде $(L+1)$-уровневого бинарного дерева [4], а набор представлений образов по N каналам образует многослойное представление

$$A^L = \{A_q^L = (a_q^0, \ldots, a_q^l, \ldots, a_q^L)\}_{q=1}^N, \qquad (1)$$

в котором q-й слой задан деревом A_q^L и

$$a_q^l = \{Q_n | \, Q_n \text{ - концевой узел в } A_q^l\} \qquad (2)$$

множество концевых вершин в поддереве $A_q^l \subset A_q^L$. Вершины Q_n в (2) образованы эллиптическими примитивами, которые аппроксимируют сегменты образа в q-ом канале. Пример трёхслойного представления вида (1), построенного для каналов цветного изображения лица в модели HSI, дан на рис. 1. В данном представлении слои с номерами $q = 1, 2, 3$ соответствуют каналам H, S, I, и каждый q-й слой представлен набором множеств примитивов a_q^l, $l = 0, \ldots, 8$, мощности $\|a_q^l\| = 2^l$, которые образуют

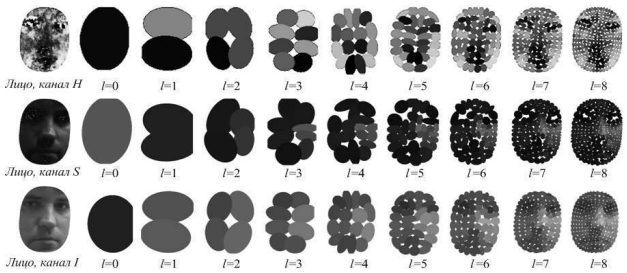


Рис. 1. Пример трёхслойного древовидного представления лица в модели HSI.

полное бинарное дерево A_q^L с параметром $L = 8$. Эллиптические примитивы являются аппроксимациями сегментов, полученных в результате дихотомического разбиения образов в каналах H, S и I.

Для любой пары объектов $A \in \mathbf{A}$ и $\hat{A} \in \mathbf{A}$ примитивы $Q_n \in A_q^L$, $\hat{Q}_n \in \hat{A}_q^L$, находящиеся в вершинах с номером n, считаются соответственными, а подмножество пар соответственных примитивов образует пересечение деревьев q-го слоя

$$A_q^L \cap \hat{A}_q^L = \{(Q_n, \hat{Q}_n)\}.$$

Поскольку описание каждого эллиптического примитива содержит три группы параметров (вектор центра, векторы ориентации и размеров осей, и среднее значение яркости сегмента), то различие соответственных примитивов (Q_n и $\hat{Q}_n$) в [4] определяется тремя функциями

$$\rho_k^{(q)}(Q_n, \hat{Q}_n) \geqslant 0,\ k = 1, 2, 3, \tag{3}$$

выраженными через параметры соответствующей группы. Функции (3) дают три компоненты меры различия пары объектов $(A, \hat{A})$ по q-му слою их представлений вида (1)

$$d_{Lk}^{(q)}(A, \hat{A}) = \sum_{n:(Q_n, \hat{Q}_n) \in (A_q^L \bigcap \hat{A}_q^L)} w_n^{(q)} \rho_k^{(q)}(Q_n, \hat{Q}_n), \tag{4}$$

$k = 1, 2, 3$, где $\{w_n^{(q)} \geqslant 0\}$ — нормированные весовые коэффициенты q-го слоя, определяемые структурой пересечения $A_q^L \bigcap \hat{A}_q^L$. Тогда с учётом (4) мера различия пары объектов $(A, \hat{A})$ по q-му слою их представлений определяется функцией

$$d_L^{(q)}(A, \hat{A}) = \sum_{k=1}^{3} \omega_k^{(q)} d_{lk}^{(q)}(A, \hat{A}), \tag{5}$$

а мера различия этой пары по многослойным представлениям вида (1) — функцией

$$d_L(A, \hat{A}) = \sum_{q=1}^{N} \gamma^{(q)} d_L^{(q)}(A, \hat{A}), \tag{6}$$

где $\{\{\omega_k^{(q)} \geqslant 0\}_{k=1}^3\}_{q=1}^N$ и $\{\gamma^{(q)} \geqslant 0\}_{q=1}^N$ — свободные нормированные коэффициенты, оцениваемые на этапе обучения.

Формализация модели классификации

Будем считать, что множество объектов источника $\mathbf{A} = \{\mathbf{A}_i\}_{i=0}^c$ содержит $c+1$ классов, где каждый класс A_i с номером $i \neq 0$ включает семантически однородные объекты, а класс $\mathbf{A}_0$ объединяет все прочие объекты. Пусть $\{P_i = P(\mathbf{A}_i)\}_{i=0}^c$ — вероятности классов. Тогда вероятности

$$P_{\text{own}} = \sum_{i=1}^{c} P_i,\ P_{\text{alien}} = P_0 = 1 - P_{\text{own}} \tag{7}$$

дают априорное распределение «своих» (own) и «чужих» (alien) объектов на множестве $\mathbf{A}$.

Для обучения используется множество объектов

$$\mathbf{B} = \{\mathbf{B}_i = \{B_{ij}\}_{j=1}^{m_i}\}_{i=1}^c \subset \mathbf{A}, \tag{8}$$

составленное из семантически однородных кластеров $\mathbf{B}_i$ мощности $m_i,\ i \neq 0$. На обучающем множестве (8) отбирается множество эталонов

$$\hat{\mathbf{B}} = \{\hat{\mathbf{B}}_i = \{\hat{B}_{ij}\}_{j=1}^{\hat{m}_i}\}_{i=1}^c \subset \mathbf{B}, \tag{9}$$

в котором $\hat{m}_i \leqslant m_i$ и каждый эталон $\hat{B}_{ij}$ является центром шара

$$S_L(\hat{B}_{ij}, D_L(\hat{B}_{ij})) = \{A|\ d(A, \hat{B}_{ij}) \leqslant D_L(\hat{B}_{ij})\} \tag{10}$$

с радиусом $D_L(\hat{B}_{ij})$, вычисляемым по мере (6).

Используя множество эталонов (9) и радиусы соответствующих шаров вида (10), вводится мера сходства объекта A с подмножеством эталонов $\hat{\mathbf{B}}_i$:

$$\mu(A, \hat{\mathbf{B}}_\mathbf{i}) = \max_{j=1}^{\hat{m}_i} \big(2^{\frac{-d_L(A, \hat{B}_{ij})}{D_L(\hat{B}_{ij})}} [d_L(A, \hat{B}_{ij}) \leqslant D_L(\hat{B}_{ij})]\big), \tag{11}$$

где $[*]$ — индикатор, принимающий значения 1 или 0 при выполнении указанного или обратного неравенства. При этом $\mu(A, \hat{\mathbf{B}}_i) > 0$, если A принадлежит хотя бы одному из шаров вида (10) с центром $\hat{B}_{ij} \in \hat{\mathbf{B}}_i$ и $\mu(A, \hat{\mathbf{B}}_i) = 0$ в противном случае. Мера сходства (11) даёт следующий критерий установления номера класса i^* для объекта A:

$$\mu(A, \hat{\mathbf{B}}_\mathbf{k}) = \max_{i=1}^{c}(\mu(A, \hat{\mathbf{B}}_\mathbf{i})),$$

$$i^* = k[\mu(A, \hat{\mathbf{B}}_\mathbf{k}) > 0], \tag{12}$$

в котором значение $i^* = 0$ соответствует классу отказов $\mathbf{A}_0$.

Задача исследования

Классификатор по критерию (12) полностью определяется множеством эталонов $\hat{\mathbf{B}}$ вида (9) и множеством радиусов

$$\mathbf{D}_L(\hat{\mathbf{B}}) = \big\{\{D_L(\hat{B}_{ij})\}_{j=1}^{\hat{m}_i}\big\}_{i=1}^c$$

шаров вида (10). Качество классификатора будем оценивать средним значением долей ошибок распознавания «своих» и «чужих» объектов

$$\varepsilon(\hat{\mathbf{B}}, \mathbf{D}_L(\hat{\mathbf{B}})) = \varepsilon^{(\text{own})}(\hat{\mathbf{B}}, \mathbf{D}_L(\hat{\mathbf{B}}))P_{\text{own}} + \\ + \varepsilon^{(\text{alien})}(\hat{\mathbf{B}}, \mathbf{D}_L(\hat{\mathbf{B}}))P_{\text{alien}} \qquad (13)$$

по заданному априорному распределению (7). Задача состоит в выборе стратегии построения покрытия обучающего множества $\mathbf{B} \subset \mathbf{A}$ шарами вида (10) и нахождении в рамках предложенной стратегии пары множеств

$$(\hat{\mathbf{B}}, \mathbf{D}_L(\hat{\mathbf{B}})) = \arg \min_{\hat{\mathbf{B}}', \mathbf{D}_L(\hat{\mathbf{B}}')} \varepsilon_{\mathbf{B}}(\hat{\mathbf{B}}', \mathbf{D}_L(\hat{\mathbf{B}}')), \qquad (14)$$

которая доставляет минимум средней доли ошибок $\varepsilon_{\mathbf{B}}(\mathbf{B}, \mathbf{D}_L(\mathbf{B}))$ вида (13) на обучающем множестве объектов $\mathbf{B}$. Для классификатора, определяемого множествами (14), необходимо найти оценки вероятностей ошибок $\varepsilon_{\mathbf{A}\setminus\mathbf{B}}(\mathbf{B}, \mathbf{D}_L(\mathbf{B}))$ на тестовом множестве $\mathbf{A} \setminus \mathbf{B}$ при различных значениях априорных вероятностей P_{own} и P_{alien}.

В решаемой задаче источник объектов задаётся множеством $\mathbf{A}$ цветных изображений лиц в модели HSI. Однако предлагаемое решение может быть использовано для других источников, допускающих многослойные представления вида (1).

Обучение классификатора

Обучение состоит в получении оценок параметров $\{\{\omega_k^{(q)} \geqslant 0\}_{k=1}^3\}_{q=1}^N$ и $\{\gamma^{(q)} \geqslant 0\}_{q=1}^N$ в мерах вида (5) и (6), и построении множеств $(\hat{\mathbf{B}}, \mathbf{D}_L(\hat{\mathbf{B}}))$, удовлетворяющих условию (14). Ниже предлагается способ построения оптимальной пары множеств эталонов и радиусов по заданной мере различия объектов. Способ применим для построения классификаторов по любым (k, q)-составляющим меры (4) и q-составляющим меры (5), что позволяет найти оценки вероятностей ошибок обучения $\varepsilon_{\mathbf{B}}^{(kq)}$ и $\varepsilon_{\mathbf{B}}^{(q)}$ по указанным составляющим меры различия. При фиксированных значениях k и q в качестве оценок параметров в (5) и (6) выбираются величины

$$\hat{\omega}_k^{(q)} = \frac{\log \varepsilon_{\mathbf{B}}^{(kq)}}{\sum_{k=1}^3 \log \varepsilon_{\mathbf{B}}^{(kq)}}, \ k = 1, 2, 3, \qquad (15)$$

$$\hat{\gamma}^{(q)} = \frac{\log \varepsilon_{\mathbf{B}}^{(q)}}{\sum_{q=1}^N \log \varepsilon_{\mathbf{B}}^{(q)}}. \qquad (16)$$

При заданной на множестве $\mathbf{B}$ мере различия $d_L(B, \hat{B})$ пары объектов $(B, \hat{B})$, предлагаемый способ построения множеств $(\hat{\mathbf{B}}, \mathbf{D}_L(\hat{\mathbf{B}}))$ состоит в следующем. Для каждого кластера $\mathbf{B}_i \subset \mathbf{B}$ строится древовидно-структурированное подмножество покрывающих этот кластер шаров путём дихотомического разбиения кластера на $\hat{m}_i$ непересекающихся сегментов

$$\{\mathbf{B}_{ij} \mid \bigcup_{j=1}^{\hat{m}_i} \mathbf{B}_{ij} = \mathbf{B}_i\},$$

и выбора для каждого $\mathbf{B}_{ij}$ шара с центром

$$\hat{B}_{ij} = \arg \min_{\hat{B}_{ik} \in \mathbf{B}_{ij}} \max_{B_{ik'} \in \mathbf{B}_{ij}} d_L(B_{ik'}, \hat{B}_{ik})$$

и радиусом вида

$$D_L(\hat{B}_{ij}) = (1 - \alpha_i) D_L^{(\min)}(\hat{B}_{ij}) + \alpha_i D_L^{(\max)}(\hat{B}_{ij}),$$

где $\alpha_i \in [0, 1]$ — свободный параметр. Величина $D_L^{(\min)}$ определяется расстоянием по заданной мере между центром $\hat{B}_{ij}$ и наиболее удалённым объектом в сегменте $\mathbf{B}_{ij}$; величина $D_L^{(\max)}$ — расстоянием между $\hat{B}_{ij}$ и наиболее удалённым объектом в подмножестве $\mathbf{B} \setminus \mathbf{B}_i$. На каждом шаге дихотомии выполняется разбиение $\mathbf{B}_{ij} \to (\mathbf{B}_{ij'}, \mathbf{B}_{ij''})$ сегмента $\mathbf{B}_{ij}$ наибольшей мощности $\|\mathbf{B}_{ij}\|$. Сегменты $\mathbf{B}_{ij'}, \mathbf{B}_{ij''}$ включают объекты сегмента $\mathbf{B}_{ij}$, которые наиболее близки по заданной мере к наиболее удалённым друг от друга объектам $B_{ij'} \in \mathbf{B}_{ij}$ и $B_{ij''} \in \mathbf{B}_{ij}$. Покрытие кластера $\mathbf{B}_i \subset \mathbf{B}$ даёт пару множеств $(\hat{\mathbf{B}}_i, \mathbf{D}_L(\hat{\mathbf{B}}_i))$. Объединение покрытий кластеров порождает пару $(\hat{\mathbf{B}}, \mathbf{D}_L(\hat{\mathbf{B}})) = \{(\hat{\mathbf{B}}_i, \mathbf{D}_L(\hat{\mathbf{B}}_i))\}_{i=1}^c$.

Оптимизация параметров $(\hat{m}_i, \alpha_i)$ покрытия каждого кластера $\mathbf{B}_i$ производится независимо на основе минимизации доли ошибок скользящего контроля

$$\varepsilon_{\mathbf{B}}^{(\text{cv})}(\hat{m}_i, \alpha_i) = P_i^* \varepsilon_{\mathbf{B}_i}^{(\text{cv})}(\hat{m}_i, \alpha_i) + (1 - P_i^*) \varepsilon_{\mathbf{B}\setminus\mathbf{B}_i}^{(\text{cv})}(\hat{m}_i, \alpha_i), \qquad (17)$$

где

$$\varepsilon_{\mathbf{B}_i}^{(\text{cv})}(\hat{m}_i, \alpha_i) = \text{FRR}(\hat{m}_i, \alpha_i),$$

$$\varepsilon_{\mathbf{B}\setminus\mathbf{B}_i}^{(\text{cv})}(\hat{m}_i, \alpha_i) = \text{FAR}(\hat{m}_i, \alpha_i)$$

— доли ложных отказов и ложных распознаваний при предъявлении покрытию с параметрами $(\hat{m}_i, \alpha_i)$ «своих» $(\mathbf{B}_i)$ и «чужих» $(\mathbf{B} \setminus \mathbf{B}_i)$ объектов из обучающего множества $\mathbf{B}$, а P_i^* и $(1 - P_i^*)$ — оценки вероятностей «своих» и «чужих» объектов среди общего числа предъявлений. В качестве оценок P_i^* выбираются вероятности P_i, $i = 1, \ldots, c$, в (7). При условии, что эти вероятности одинаковы, $P_i^* = P_{\text{own}}/c$ и $1 - P_i^* = 1 - P_{\text{own}}/c$. Оптимизация покрытия кластера $\mathbf{B}_i$ сводится к нахождению пары $(\hat{m}_i, \alpha_i)$, которая доставляет наименьшее значение функции (17). Для оптимизированных покрытий кластеров оценка вероятности ошибок обучения TSC-классификатора определяется средним значением

$$\varepsilon_{\mathbf{B}}(\hat{\mathbf{B}}, \mathbf{D}_L(\hat{\mathbf{B}})) = \frac{1}{c} \sum_{i=1}^c \varepsilon_{\mathbf{B}}^{(\text{cv})}(\hat{m}_i, \alpha_i). \qquad (18)$$

Оценки вида (18), полученные при обучении по (k,q)-составляющим меры (4) и q-составляющим меры (5), использованы в (15) и (16) для вычисления соответствующих параметров меры.

Экспериментальные результаты распознавания лиц

Схема проведения экспериментов заключалась в следующем. Исходные объекты из сформированной базы лиц были преобразованы в HSI-изображения размера 800x600 пикселей. Трёхслойные представления информативных объектов строились с использованием $(L+1)$-уровневых полных деревьев с параметром $L=10$. В эксперименте использовано $\|\mathbf{A}\|=1000$ изображений лиц от $c=25$ персон (классов), по $m_i=40$ объектов в каждом классе. Множество лиц разбивалось пятикратно на равные по объёму обучающие $\mathbf{B}$ и тестовые $\mathbf{A}\backslash\mathbf{B}$ выборки, в которых семантически однородные кластеры имели одинаковые мощности $m_i=20,\quad i==1,\ldots,c$. Априорные вероятности «своих» и «чужих» объектов выбирались равными значениям: $P_{\text{own}}=1;\ 0.75;\ 0.5$ и $P_{\text{alien}}=1-P_{\text{own}}$. При каждом фиксированном значении P_{own}, на пяти обучающих выборках строились TSC-классификаторы с помощью покрытий для $c'=\lfloor cP_{\text{own}}\rfloor$ произвольно выбранных кластеров. Для заданной обучающей выборки $\mathbf{B}$ при каждом значении $c'<c$ строилось несколько реализаций таких классификаторов (при $c'=c$ одна реализация для каждой выборки $\mathbf{B}$). Предъявляя объекты тестовой выборки $\mathbf{A}\setminus\mathbf{B}$ для каждой реализации TSC-классификатора, вычислялась оценка вероятности ошибочных решений (13) и усреднённая оценка вероятности ошибки по всем реализациям классификатора и предъявляемым тестовым выборкам. Полученные для TSC-классификатора доли ошибок распознавания сравнивались с аналогичными показателями SVM-классификатора [5]. В качестве признаков SVM-классификатора использовались попарные расстояния по заданной мере между предъявляемым объектом и объектами обучающей выборки.

Усреднённые оценки вероятностей ошибок для обоих классификаторов представлены в таблице 1. Оценки качества SVM-классификатора приведены для случая $P_{\text{own}}=1$. Из данных таблицы следует, что оба классификатора демонстрируют сопоставимые показатели качества и обеспечивают уменьшения доли ошибок за счёт использования многослойных представлений модели HSI по сравнению с ошибками по отдельным каналам H, S или I. Необходимо отметить, что применение древовидных представлений с многоуровневым разрешением позволяет организовывать в TSC-классификаторе иерархический поиск решений в многоуровневой базе эталонов. При большом числе классов c стратегия иерархического поиска сокращает вычислительные затраты в $c/\log c$ раз по сравнению с переборным поиском.

Таблица 1. Усреднённые оценки вероятности ошибок распознавания лиц для классификаторов типа TSC и SVM.

$P_{(own)}$	1	1	0.75	0.5
Классификатор	TSC	SVM	TSC	TSC
`H`	0.015	0.011	0.021	0.020
`S`	0.026	0.018	0.026	0.023
`I`	0.032	0.078	0.036	0.029
`HSI`	0.012	0.007	0.019	0.016

Выводы

Предложен метод построения классификатора по критерию ближайшего эталона в пространстве многослойных древовидных представлений объектов, заданных многоканальными изображениями. Обучение классификатора состоит в отборе эталонов и реализовано на основе построения древовидно-структурированных покрытий семантически однородных кластеров обучающего множества шарами, центры и радиусы которых вычисляются по заданной мере на множестве представлений.

Эффективность предложенного метода продемонстрирована результатами распознавания цветных изображений лиц в трёхканальной модели HSI. Показано преимущество использования многоканальных изображений по сравнению с монохромными изображениями. Полученные оценки качества распознавания лиц сопоставимы с аналогичными показателями классификатора на основе метода опорных векторов.

В перспективе предполагается исследовать модификации древовидно-структурированных покрытий и другие схемы комплексирования многоканальных данных. Планируется также расширение состава биометрических источников.

Литература

[1] *Phillips P. J, Martin A., Wilson L. C, Przybocki M.* An introduction to evaluating biometric systems // IEEE Computer, 2000. — Vol. 21, No. 2. — Pp. 56–63.

[2] *Гонсалес Р., Вудс Р., Эддинс С.* Цифровая обработка изображений в среде Matlab.— Москва: Техносфера, 2006.

[3] *Zhao W., Chellappa R., Phillips P. J, Rosenfeld A.* Face recognition: a literature survey // ACM Computing Surveys, 2003. — Vol. 35, No. 4. — Pp. 399–458.

[4] *Ganebnykh S. N, Lange M. M,.* Classification of 2D Grayscale Objects in a Space of the Multiresolution Representations // Pattern Recognition and Image Analysis, 2009. — Vol. 19, No. 4. — Pp. 591–602.

[5] `opencv.willowgarage.com` — Open Source Computer Vision Library — 2010.

Применение триплетных признаков распознавания к цветным изображениям*

Федотов Н. Г., Романов С. В., Мокшанина Д. А.
ec@pnzgu.ru
Пенза, Пензенский государственный университет

В данном докладе рассматривается один из возможных подходов к распознаванию цветных изображений. Метод основан на использовании триплетных признаков распознавания изображений. Представлены Трейс-функционалы для обработки цветных изображений. Рассмотрены вопросы эффективности данного подхода. Использование триплетных признаков для обработки цветных изображений увеличивает точность распознавания и расширяет сферу применения алгоритмов на их основе.

На сегодняшний день распознавание изображений нашло широкое применение в различных приложениях. Постоянное увеличение вычислительных мощностей позволило не только разрабатывать программные продукты, предназначенные для распознавания изображений, но и реализовывать данную технологию в законченных устройствах, например, фотоаппаратах, автомобилях и т. д.

Традиционно для распознавания графического образа анализируется его контур и ряд геометрических характеристик. Соответственно, для вычисления признаков распознавания необходимо использовать бинарное представление изображения. Современные средства получения цифрового представления изображения обеспечивают возможность работы с цветными изображениями. Таким образом, во многих задачах для распознавания изображения необходимо провести его предварительную обработку, во время которой осуществляется преобразование цветного изображения в бинарное. Очевидно, что информативность бинарного изображения ниже, чем цветного, за исключением тех случаев, когда объект на изображении изначально не является цветным, например, текст. Уменьшение информативности объекта упрощает процедуру обработки и снижает вычислительную сложность, но, одновременно, снижает точность распознавания. Постоянное стремление увеличить точность алгоритмов распознавания подталкивает к использованию дополнительных признаков и дополнительной информации. Одним из путей решения данной задачи является анализ цвета изображения.

Частично анализ цвета реализуется на этапе предварительной обработки, когда выделение информативных объектов осуществляется на основе цветовых признаков. Для выделенных объектов в ряде систем распознавания изображений успешно применяется дополнительное вычисление яркостных или цветовых признаков на основе анализа гистограмм данных изображений. Несмотря на эффективность данных решений, они не обеспечивают полноценного анализа цветного изображения.

Более полное описание можно получить путём непосредственного анализа цветного изображения без его предварительной обработки. В настоящей работе представлен подход к анализу цветных изображений, основанный на аппарате стохастической геометрии и функционального анализа, обеспечивающий возможность непосредственной работы с подобными объектами. Теория распознавания образов на основе стохастической геометрии и функционального анализа ранее применялась лишь к бинарным и полутоновым изображениям. В настоящей работе представлено дальнейшее развитие данной теории, расширяющее область её применения на цветные изображения.

Триплетные признаки распознавания бинарных и цветных изображений

Структура триплетных признаков в виде композиции трёх функционалов позволяет путём автоматической генерации получить большое количество (десятки тысяч) характеристик исходного изображения, что обеспечивает возможность всестороннего анализа объекта [1]. Данная особенность обеспечивает универсальность метода. Его применение в самых различных областях показало высокую точность распознавания.

Триплетный признак распознавания может быть записан в виде:

$$\Pi(F) = \Theta \cdot P \cdot T(F, \rho, \theta),$$

где F — исходное изображение, $\Pi(F)$ — признак, Θ — диаметральный функционал, P — круговой функционал, $T(F, \rho, \theta)$ — T-функционал.

T-функционал играет ключевую роль в формировании триплетного признака распознавания. Вычисление T-функционала непосредственно связано с понятием сканирующей прямой $l(\rho, \theta)$. Положение на плоскости данной прямой определяется нормальными координатами: ρ — расстояние до начала координат, θ — угол наклона, относительно оси x. Рис. 1 иллюстрирует прохождение сканирующей прямой через изображение. В результате пересечения $F \cup l(\rho, \theta)$ получаем некоторый вектор L,

Работа выполнена при финансовой поддержке РФФИ, проект № 09-07-00089-а.

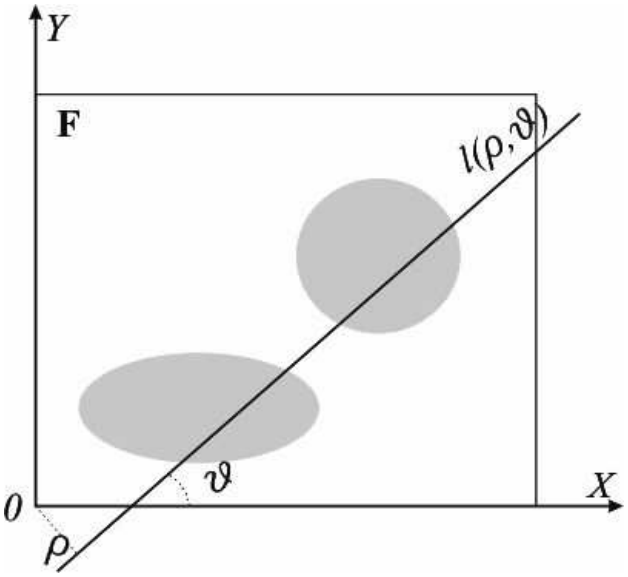


Рис. 1. Пересечение изображения сканирующей прямой.

который в случае бинарного изображения представляет собой набор нулей и единиц, соответствующих пересекаемым объектам на изображении F.

К полученному вектору применяется некоторая функция $f(L)$, ставящая ему в соответствие число. Вычисление -функционала осуществляется для множества сканирующих прямых, полученных изменением параметров ρ и θ. Обычно используется фиксированный набор значений ρ и θ с заданным шагом. В результате вычисления T-функционала для данного набора сканирующих прямых получается трейс-матрица. Каждое число в этой матрице является значением функционала $f(L)$ для заданных ρ и θ.

Функционал P используется для обработки столбцов матрицы. В результате его вычисления трейс-матрица преобразуется в вектор. Дальнейшее применение функционала Θ преобразует данный вектор в число, которое и является результатом вычисления триплетного признака.

Функционалы P и Θ идентичны. Для их вычисления обычно используются стандартные математические или статистические формулы, например, среднее значение, сумма, минимум, максимум и т. д. Можно сказать, что для вычисления подходит любая функция, аргументом которой является вектор, а результат представлен единственным числом. Очевидно, что данные функционалы, используемые только для обработки трейс-матрицы, не связаны с типом изображения.

Для бинарных изображений T-функционал алгоритмически образуется последовательным вычислением двух функций. Первая осуществляет преобразование исходного вектора L в новый вектор L'; суть данного преобразования заключается в переходе от бинарного представления результатов пересечения сканирующей прямой с изображением к некоторым численным характеристикам данного пересечения. Для бинарных изображений традиционно используются две функции: длина пересекаемых участков и количество пересекаемых участков. К полученному вектору L' применяется функция аналогичная, используемой для вычисления функционалов P и Θ. Применение данного подхода позволяет получить оценку большинства геометрических характеристик объекта.

Для обработки цветных изображений необходимо изменить T-функционал. Для случая, когда признак вычисляется только для одной компоненты цвета (независимо от цветового пространства — RGB, HSV или др.) результатом пересечения сканирующей прямой с изображением является вектор L', состоящий из значений яркости соответствующих точек сканирующей прямой для заданной цветовой компоненты. Если признак вычисляется на основе всех компонент цвета, то результат пересечения сканирующей прямой и изображения представлен векторами $L1$, $L2$ и $L3$, каждый из которых представляет отдельную цветовую компоненту. Как и в случае с бинарным изображением, необходимо осуществить переход к вектору L'. Для осуществления данного преобразования можно использовать множество функций, например, вычисление яркости.

Формирование триплетного признака распознавания для цветного изображения

Традиционно используются два подхода к формированию триплетного признака распознавания: экстракция и автоматическая компьютерная генерация [1]. Суть экстракции признака заключается в выделении неких графических или геометрических особенностей объекта и последующему подбору или расчёту функционалов, которые бы давали устойчивую числовую оценку заданным особенностям. Многочисленные эксперименты позволили подобрать для бинарных изображений триплетные признаки распознавания, которые характеризуют площадь объекта, его периметр, диаметр, количество углов и т. д. Обычно экстракция признаков отличается высокой сложностью и трудоемкостью, так как этот процесс мало формализован и, зачастую, носит эмпирический характер. Дополнительно, возникает необходимость точно сформулировать особенности объектов, которые будут использоваться для их распознавания. На данном этапе появляется определённая субъективность в выборе признаков и сокращение информации об объекте.

Вторым подходом к формированию триплетного признака является автоматическая компьютерная генерация. Структура триплетного при-

знака в виде композиции функционалов позволяет получить очень большое количество признаков путём их комбинаторного перебора. Конструирование большого числа признаков с последующим сокращением признакового пространства позволяет быстро подобрать необходимое количество признаков, которые объективно наиболее точно характеризуют объекты обучающего множества. Для минимизации признакового пространства используются специальные алгоритмы, основанные на анализе компактности класса и расстояний между классами или использующие теорию решёток. К недостаткам второго метода можно отнести неизвестность геометрических характеристик объекта, которые легли в основу вычисления признаков.

Для цветных изображений экстракция признаков отличается большей сложностью. Были получены положительные результаты в определении геометрических характеристик объектов с заданными цветовыми характеристиками. Фактически, в данном случае, процедуры предварительной обработки изображения и последующее вычисление признаков были объединены в процессе вычисления признаков. С одной стороны, данный подход упрощает обработку изображения и позволяет проще реализовать адаптивные алгоритмы предварительной обработки изображений. С другой стороны, отсутствие результатов предварительной обработки не позволяет оценить ошибки на данном этапе. Следует отметить, что скорость обработки увеличилась в 1,37 раза. Учитывая данные особенности, можно использовать следующий подход к разработке системы распознавания образов. На этапе разработки системы осуществляется предварительная обработка изображения, выявляются наилучшие алгоритмы фильтрации изображения. Наличие результатов предварительной обработки позволяет быстро скорректировать алгоритмы и получить эффективные решения. Затем алгоритмы, полученные на предыдущем этапе, переносятся в T-функционал. В результате достигается увеличение производительности, а корректность работы алгоритма можно осуществлять путём сравнения работы непосредственной обработки цветных изображений с вычислением признаков для бинарных изображений после предварительной обработки.

Прямое перенесение генерации признаков на цветные изображения принесло положительные результаты только при распознавании текстур [2]. В данном случае для формирования признаков фактически использовалась только яркость точки. Применение триплетных признаков позволило осуществить распознавание текстур с значительно большей точностью по сравнению с классическими методами анализа.

Анализ полученных результатов, позволил выявить причину низкой точности распознавания цветных изображений. Суть ошибки заключается в неполном переборе функционалов. Если вычисление признаков для бинарных изображений позволяло дать оценку только геометрическим характеристикам объекта, то для цветных изображений возможно, как минимум, дать геометрическую оценку характеристик объекта после его цветовой фильтрации и характеризовать процессы изменения цвета объекта.

Осуществить перебор всех возможных функционалов невозможно ввиду огромной вычислительной сложности. Поэтому был опробован подход объединяющий экстракцию и генерацию признаков. Суть его заключается в выборе T-функционала на основе эффективного алгоритма предварительной обработки цветного изображения и переборе оставшихся функционалов. Для задачи анализа гистологических изображений первичное применение подобного алгоритма привело к увеличению точности распознавания в 1,17 раза [3].

Можно сделать следующий вывод: генерация триплетных признаков распознавания является эффективным инструментом разработки систем распознавания образов, но для цветных изображений необходимо выбрать методы цветовой фильтрации. Классический подход, предполагающий предварительную обработку изображения, позволяет непосредственно использовать автоматическую генерацию признаков для вычисления цветовых характеристик объекта, но при вычислении геометрических характеристик отличается худшей точностью по сравнению с обработкой бинарных изображений.

Выводы

Непосредственное распознавание цветных изображений является перспективным направлением развития данной области знаний. Потенциально вычисление признака цветного изображения позволяет получить более высокую точность распознавания и большее быстродействие алгоритма, обусловленное однократной обработкой изображения. Триплетные признаки распознавания изображений были совсем недавно применены к цветным изображениям, и многие вопросы ещё требуют глубокой проработки и формализации. Тем не менее, уже сегодня можно отметить, что данный подход позволяет упростить и частично унифицировать предварительную обработку изображений; дать численную характеристику изменениям цвета объекта; обеспечить вычисление всех геометрических признаков, которые могут быть получены при обработке бинарных изображений.

Применение триплетных признаков распознавания к цветным изображениям может быть эффективным при решении большого количества задач.

Литература

[1] *Федотов Н. Г.* Теория признаков распознавания образов на основе стохастической геометрии и функционального анализа. — Москва: ФИЗМАТЛИТ, 2009. — 304 с.

[2] *Федотов Н. Г., Мокшанина Д. А.* Распознавание изображений со сложной полутоновой текстурой // Измерительная техника. — 2010. — № 11. — С. 27–31.

[3] *Федотов Н. Г., Романов С. В., Мокшанина Д. А.* Сегментация гистологических изображений. Выделение фолликулов и ядер // Математические методы распознавания образов. — Москва: МАКС Пресс, 2009. — С. 611–613.

[4] *Sonka M., Hlavac V., Boyle R.* Image Processing, Analysis, and Mashine Vision // Brooks and Cole Publishing, 1998.

[5] *Прэтт У. К.* Цифровая обработка изображений: В 2 т. — Москва: Мир, 1982.

Теоретические основы корреляционно-экстремальных контурных методов распознавания*

Лебедев Л. И.
lebedev@pmk.unn.ru
Нижний Новгород, НИИ прикладной математики и кибернетики Нижегородского государственного национально-исследовательского университета им. Н. И. Лобачевского

Анонсируется корреляционно-экстремальный контурный подход в задачах распознавания объектов изображений, базирующийся на вычислении оценок сходства, инвариантных относительно аффинных преобразований. Приводится теоретическое обоснование с вероятностной точки зрения выбранного критерия сходства. Показывается, что нахождение оценок сходства сводится к решению оптимизационной задачи по параметрам аффинного преобразования. Приводятся в аналитическом виде оценки сходства, инвариантные относительно ортогональных преобразований, а также ортогональных преобразований и масштабирования. Даётся оценка сложности вычисления оценок сходства.

Исходным описанием изображений принято считать растровую модель представления информации. Для бинарных изображений растровому описанию эквивалентной является векторная модель представления информации, которая для решения задач распознавания во многом является более удобной. В рамках этой модели изображение представляется набором «раскрашенных» контуров и кривых. Огромная масса вводимых документов имеет черно-белый формат или выполнена с использованием фиксированного числа цветов (графические документы). Таким образом, с учётом возможностей цветоделения такие документы можно представить совокупностью бинарных изображений и, следовательно, реализовать решение задач распознавания на базе контурного описания.

Выбор критерия сходства

Один из наиболее распространённых подходов к построению правил опознавания изображений базируется на интуитивном ассоциировании понятий «сходство» и «близость» и выражается в предположении о правомерности сопоставления степени сходства изображений со степенью близости соответствующих точек множеств. Доведение такой ассоциации до логического завершения приводит к предположению о том, что «очень близким» точкам должны соответствовать «почти неотличимые» изображения, то есть столь сходные изображения, что они почти всегда принадлежат к одному классу (гипотеза «компактности» [1]). Объявляя одно из изображений класса эталонным, указанный подход к опознаванию изображений определяет группу методов сравнения с эталонами, отличающимися по существу только заданием вида критерия близости.

Пусть описание эталона $\mathbf{S}$ задано последовательностью точек на плоскости, равномерно расположенных вдоль контура $\mathbf{w^s} = \{\mathbf{w}_1^s, \mathbf{w}_2^s, \ldots, \mathbf{w}_n^s\}$, где $\mathbf{w}_i^s = (x_i^s, y_i^s)^\intercal$. Зададим следующую модель формирования объектов класса эквивалентности $K(\mathbf{S})$, порождённого эталоном $\mathbf{S}$. Описание контура эталона $\mathbf{w^s}$ преобразуем в соответствии с невырожденным аффинным преобразованием и произведем округление координат точек до заданной точности. В результате будем иметь множество контуров, полученных при различных параметрах аффинного преобразования. Эта схема формирования объектов различных классов эквивалентности моделирует процедуру сканирования хаотически расположенных эталонных объектов и векторизации полученных изображений. Рассмотрим объект $\mathbf{O} \in K(\mathbf{S})$, описание которого задано последовательностью точек $\mathbf{w} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n\}$. Так как он является образом эталона $\mathbf{S}$, то его описание может быть задано формулой

$$\mathbf{w} = \mathbf{G} * \mathbf{w^s} + \xi = \widehat{\mathbf{w}}^{\mathbf{s}} + \xi, \qquad (1)$$

где $\mathbf{G}$ — оператор аффинного преобразования, а ξ — случайные помехи, отражающие результаты округлений. Таким образом, точки $\mathbf{w_i}$, $i = 1, \ldots, n$ контура объекта $\mathbf{O}$ можно рассматривать как множество значений, принимаемых системой случайных независимых величин $\mathbf{W_i}$, распределённых по нормальному закону с плотностью $\mathbf{p_i} \sim \mathbf{N}(\widehat{\mathbf{w}}_\mathbf{i}^\mathbf{s}, \mathbf{K_{xy}})$. Следовательно, для установления факта принадлежности объекта $\mathbf{O}$ классу $K(\mathbf{S})$ необходимо подобрать такие параметры аффинного преобразования $\mathbf{G}$, чтобы полученная совокупность значений $\widehat{\mathbf{w}}_\mathbf{i}^\mathbf{s} = \mathbf{G} * \mathbf{w}_\mathbf{i}^\mathbf{s}$ могла рассматриваться как математическое ожидание такого же распределения. Принято считать, что погрешности округления характеризуются круговым рассеиванием $\mathbf{K_{xy}} = \sigma^2\mathbf{E}$. Из принципа максимального правдоподобия с учётом последнего замечания параметры аффинного преобразования $\mathbf{G}$ должны тогда удовлетворять условию

$$\sum_{i=1}^{n} (\mathbf{w_i} - \mathbf{G} * \mathbf{w_i^s})^\intercal (\mathbf{w_i} - \mathbf{G} * \mathbf{w_i^s})_{\overrightarrow{\mathbf{G}}} \min. \qquad (2)$$

Работа выполнена при финансовой поддержке РФФИ, проект № 10-07-00330.

Формула (2) задаёт математическое совмещение образа эталона с объектом (наложение на объект) и возможность получения оценки сходства для различных преобразований. Для определённого типа преобразований их параметры и оценка сходства получены в аналитическом виде. Для остальных преобразований, рассматриваемых в работе, будет использован комбинированный способ нахождения параметров оптимизационной задачи [2], когда часть параметров получена в аналитическом виде, а для нахождения других используются итерационные методы.

Критерий сходства при ортогональных преобразованиях

Рассмотрим основы корреляционно-экстремального контурного подхода, связанные с разработкой методов вычисления оценок сходства в зависимости от группы аффинного преобразования. Вначале предположим, что класс эквивалентности $K(\mathbf{S})$ порождается ортогональным преобразованием. В этом случае правую часть в формуле (2) можно представить в виде

$$\varepsilon(\mathbf{A}, \Delta\mathbf{w^s}) = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{w_i} - \mathbf{A}\mathbf{w_i^s} - \Delta\mathbf{w^s})\|^2, \qquad (3)$$

где

$$\mathbf{A} = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}, \Delta\mathbf{w^s} = \begin{pmatrix} \Delta x^s \\ \Delta y^s \end{pmatrix}. \qquad (4)$$

Отсюда, нахождение параметров ортогонального преобразования и получение оценки сходства сводится к решению оптимизационной задачи вида (3)

$$\varepsilon_m^{\mathbf{A}} = \min_{\alpha, \Delta\mathbf{w^s}} \varepsilon(\mathbf{A}, \Delta\mathbf{w^s}). \qquad (5)$$

Обозначим посредством $\bar{x}$, $\bar{xy}$ статистические начальные моменты первого и второго порядка. Тогда $\bar{\mathbf{w}} = (\bar{x}, \bar{y})^{\mathsf{T}}$, $\mathrm{cov}(\mathbf{x}, \mathbf{x}) = \bar{xy} - \bar{x}\cdot\bar{y}$, $\mathsf{D}\mathbf{x} = \mathrm{cov}(\mathbf{x}, \mathbf{x})$, а $\mathsf{D}\mathbf{w} = \mathsf{D}\mathbf{x} + \mathsf{D}\mathbf{y}$. Пусть

$$\begin{cases} \mathrm{Sn} = \mathrm{cov}(\mathbf{x^s}, \mathbf{y}) - \mathrm{cov}(\mathbf{y^s}, \mathbf{x}) \\ \mathrm{Cs} = \mathrm{cov}(\mathbf{x^s}, \mathbf{x}) + \mathrm{cov}(\mathbf{y^s}, \mathbf{y}), \end{cases} \qquad (6)$$

и $R^2 = \mathrm{Sn}^2 + \mathrm{Cs}^2$.

В дальнейшем приведённые характеристики будут использоваться для получения в аналитической форме оценки сходства эталона с объектом, а также параметров ортогонального преобразования для их совмещения. Это утверждение основывается на доказательстве трёх теорем.

Теорема 1. *Если для описаний $\boldsymbol{w^s}$ эталона $\boldsymbol{S}$ и $\boldsymbol{w}$ объекта $\boldsymbol{O}$, выполняется условие $R^2 \neq 0$, то экстремумы функции $\varepsilon(\boldsymbol{A}, \Delta\boldsymbol{w^s})$ возможны в точках, компоненты которых удовлетворяют условиям*

$$\mathrm{tg}\,\alpha = \frac{\mathrm{Sn}}{\mathrm{Cs}}, \ \Delta\boldsymbol{w^s} = \bar{\boldsymbol{w}} - \boldsymbol{A}\bar{\boldsymbol{w}^s}. \qquad (7)$$

Доказательство. Функция $\varepsilon(\mathbf{A}, \Delta\mathbf{w^s})$ является непрерывной вместе со своими частными производными. Поэтому, для нахождения её экстремумов частные производные по параметрам α и $\Delta\mathbf{w^s}$ приравнивались к нулю. Полученная система уравнений имеет решение, если $R^2 \neq 0$ (без ограничения общности будем считать, что $\mathrm{Cs} \neq 0$), которое может быть задано выражением (7).

Теорема 2. *Минимум функции $\varepsilon(\boldsymbol{A}, \Delta\boldsymbol{w^s})$ достигается только в одной точке, удовлетворяющей условиям*

$$\mathrm{sign}[\sin\alpha] = \mathrm{sign}[\mathrm{Sn}], \quad \mathrm{sign}[\cos\alpha] = \mathrm{sign}[\mathrm{Cs}]. \qquad (8)$$

Доказательство. Условия (7) определяют две точки, подозрительные на экстремум, имеющие по параметру α значения, отличающиеся на величину π, а именно $\widetilde{\alpha}$ и $(\widetilde{\alpha} + \pi)$, где $\widetilde{\alpha} = \mathrm{arctg}(\mathrm{Sn}\cdot\mathrm{Cs}^{-1})$. Для проверки достаточных условий существования экстремумов использовался критерий Сильвестра. Первые два главных минора матрицы $\mathbf{C} = (c_{ij})_{3\times3}$, составленной из производных второго порядка, положительны при любых значениях параметров. Значение определителя матрицы при подстановке только значения $\Delta\mathbf{w^s}$ равно $8(\mathrm{Cs}\cdot\cos\alpha + \mathrm{Sn}\cdot\sin\alpha) = 8\mathrm{Cs}\cdot\cos\alpha(1+\mathrm{tg}^2\,\alpha) = = 8\mathrm{Sn}\cdot\sin\alpha(\mathtt{ctg}^2\alpha+1)$. Из последних двух равенств следует, что определитель будет положителен, если выполняются условия формулы (8) и, следовательно, при этом значении α будет достигнут минимум функции $\varepsilon(\mathbf{A}, \Delta\mathbf{w^s})$. При другом значении α минимума нет, так как $\det\mathbf{C} < 0$.

Теорема 3. *Минимум функции $\varepsilon(\boldsymbol{A}, \Delta\boldsymbol{w^s})$ определяется формулой*

$$\varepsilon_m^{\boldsymbol{A}} = \mathsf{D}\boldsymbol{w^s} + \mathsf{D}\boldsymbol{w} - 2R. \qquad (9)$$

Доказательство. Формула вычисления оценки сходства (9) получается подстановкой значений параметров из (7) в выражение (3) путём эквивалентных преобразований с учётом (8) при выборе ветви функции извлечения квадратного корня.

Критерий сходства для ортогональных преобразований и масштабирования

В соответствии с одной из групп Ли на плоскости, обеспечивающей ортогональное преобразование и масштабирование (ОПМ), образ любой точки находится по исходному описанию по формуле $\widehat{\mathbf{w}}_\mathbf{i}^\mathbf{s} = k\mathbf{A}*\mathbf{w}_\mathbf{i}^\mathbf{s} + \Delta\mathbf{w^s}$, где $\mathbf{B} = k\mathbf{A}$, k — коэффициент масштабирования. Тогда результатом совмещения эталона с объектом в зависимости от параметров ОПМ на основании (2) является величина среднеквадратичной ошибки

$$\varepsilon(\mathbf{B}, \Delta\mathbf{w^s}, k) = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{w_i} - k\mathbf{A}\mathbf{w_i^s} - \Delta\mathbf{w^s})\|^2. \qquad (10)$$

В этом случае параметры ОПМ и сама оценка сходства получается из решения оптимизационной задачи [4]

$$\varepsilon_m^{\mathbf{B}} = \min_{k,\alpha,\Delta\mathbf{w^s}} \varepsilon(k, \mathbf{A}, \Delta\mathbf{w^s}). \qquad (11)$$

Итогом решения задачи (11) являются следующие теоремы о нахождении в аналитическом виде оптимальных параметров ОПМ и оценки сходства.

Теорема 4. *При выполнении неравенства $R^2 > 0$ необходимые условия существования экстремумов функции $\varepsilon(k, \boldsymbol{A}, \Delta\boldsymbol{w^s})$ запишутся следующим образом*

$$\operatorname{tg}\alpha = \frac{\mathrm{Sn}}{\mathrm{Cs}},\ k = \frac{\boldsymbol{\Phi}}{\mathsf{D}\boldsymbol{w^s}},\ \Delta\boldsymbol{w^s} = \bar{\boldsymbol{w}} - k\boldsymbol{A}\bar{\boldsymbol{w}}^{\boldsymbol{s}}, \qquad (12)$$

где $\boldsymbol{\Phi} = \mathrm{Cs}\cdot\cos\alpha + \mathrm{Sn}\cdot\sin\alpha$.

Доказательство. Как и в предыдущем случае, для получения необходимых условий существования экстремумов $\varepsilon(k, \mathbf{A}, \Delta\mathbf{w^s})$ приравнивались к нулю частные производные по параметрам k, α и $\Delta\mathbf{w^s}$. Полученная система уравнений имеет решение, которое можно представить в виде (12), если $R^2 \neq 0$ (при $R^2 = 0$ система уравнений имеет бесконечное множество решений).

Теорема 5. *Минимум функции $\varepsilon(k, \boldsymbol{A}, \Delta\boldsymbol{w^s})$ достигается в обеих точках, полученных при значениях α, равных* $\operatorname{arctg}(\mathrm{Sn}\cdot\mathrm{Cs}^{-1})$, $\operatorname{arctg}(\mathrm{Sn}\cdot\mathrm{Cs}^{-1})+\pi$.

Доказательство. Для доказательства теоремы вновь используем критерий Сильвестра. Аналогично, первые три главных минора матрицы $\mathbf{C} = (c_{ij})_{4\times4}$, составленной из производных второго порядка по параметрам $\Delta\mathbf{x^s}$, $\Delta\mathbf{y^s}$, k и α будут положительны при любых значениях параметров, так равны $M_1 = 2$, $M_2 = 4$ и $M_3 = 8\mathsf{D}\mathbf{w^s}$. Минор четвёртого порядка ($M_4 = \det\mathbf{C}$) при подстановке параметра $\Delta\mathbf{w^s}$ из (12) равен $(4k\mathsf{D}\mathbf{w^s})^2 > 0$, так как из (12) следует, что $\mid k \mid = R(\mathsf{D}\mathbf{w^s})^{-1} > 0$. Так как все главные миноры матрицы $\mathbf{C}$ строго положительны, то значит, что в обоих точках достигается минимум.

Теорема 6. *Минимум функции $\varepsilon(k, \boldsymbol{A}, \Delta\boldsymbol{w^s})$ находится по формуле*

$$\varepsilon_m^{\boldsymbol{B}} = \mathsf{D}\boldsymbol{w} - \frac{R^2}{\mathsf{D}\boldsymbol{w^s}}. \qquad (13)$$

Доказательство. Формула вычисления оценки сходства (13) получается подстановкой значений параметров из (12) в выражение (10) путём эквивалентных преобразований. Минимальные значения функции в обеих точках равны, поэтому при нахождении параметров преобразования естественно остановиться на варианте, когда коэффициент масштабирования $k > 0$, что приводит к необходимости выполнения условия (8) по углу α.

Критерий сходства для аффинных преобразований

Аффинное преобразование (АП) эталона может быть представлено в виде

$$\widehat{\mathbf{w}}^{\mathbf{s}} = \mathbf{G}\mathbf{w^s} + \Delta\mathbf{w^s}, \qquad (14)$$

где

$$\mathbf{G} = \begin{pmatrix} a & b \\ c & d \end{pmatrix},\ \det\mathbf{G} > 0. \qquad (15)$$

Величина невязки в этом случае будет находиться по формуле

$$\varepsilon(\mathbf{G}, \Delta\mathbf{w^s}) = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{w_i} - (\mathbf{G}\mathbf{w^s})_\mathbf{i} - \Delta\mathbf{w^s})\|^2, \qquad (16)$$

где $(\mathbf{G}\mathbf{w^s})_\mathbf{i}$, $i = 1,\ldots,n$ — точки интерполированного с равномерным шагом контура эталона после АП.

Критерий сходства здесь будет иметь вид

$$\varepsilon_m^{\mathbf{G}} = \min_{a,b,c,d,\Delta\mathbf{w^s}} \varepsilon(\mathbf{G}, \Delta\mathbf{w^s}). \qquad (17)$$

Так как при АП расстояния между точками не сохраняются и $(\mathbf{G}\mathbf{w^s})_\mathbf{i} \neq \mathbf{G}\mathbf{w^s_i}$, то решение задачи нахождения оценки сходства не может быть получено напрямую с использованием предыдущих методов. Одним из наиболее подходящих для решения задачи (17) способов является метод сеток. Однако, в задаче нахождения глобального минимума (17) возникают проблемы с локализацией параметров АП. Поэтому, для гарантированного получения оценки сходства, инвариантной относительно АП при приемлемой сложности вычислений в этой работе предлагается представить группу АП последовательностью пяти операций, включающей ортогональные преобразования, проектирование и масштабирование [5].

Предположим, что эталон лежит в плоскости XOY трёхмерного пространства. Осуществим операцию Ψ_β вращения системы координат вокруг оси OX на угол β по часовой стрелке и операцию вращения Ψ_γ на угол γ вокруг оси OY. Спроектируем на плоскость XOY полученную в пространстве фигуру (операция $\Psi_p r$). Получим фигуру $\mathbf{S_{pr}}$, заданную последовательностью точек с координатами, которые находятся на основе исходного описания по формуле

$$\begin{cases} \widetilde{x}^s = x\cos\gamma + y\sin\beta\sin\gamma, \\ \widetilde{y}^s = y\cos\beta. \end{cases} \qquad (18)$$

В плоскости XOY осуществим операции масштабирования и ортогонального преобразования путём умножения на коэффициент k и вращения фигуры $\mathbf{S_{pr}}$ на угол α с параллельным переносом на $\Delta\mathbf{w^s}$

(операции Ψ_k и Ψ_α). Таким образом, в результате действия оператора $\Psi_\alpha(\Psi_k(\Psi_{pr}(\Psi_\gamma(\Psi_\beta))))$ исходное описание эталона с учётом (18) преобразуется по формуле

$$\begin{cases} \widehat{x}^s = k(\widetilde{x}\cos\alpha - \widetilde{y}\sin\alpha) + \Delta x_s, \\ \widehat{y}^s = k(\widetilde{x}\sin\alpha + \widetilde{y}\cos\alpha) + \Delta y_s, \end{cases} \tag{19}$$

которая, очевидно, задаёт АП. Справедливо также и обратное утверждение, что любое АП может быть задано последовательностью указанных операций, то есть для любых a, b, c, d параметров АП существуют k, α, β, γ такие, что формулы (18,19) и (14) определяют эквивалентные преобразования. Это утверждение базируется на следующей теореме.

Теорема 7. *Для любых параметров АП ($\forall a, b, c, d$) нелинейная система уравнений*

$$\begin{cases} k\cos\gamma\cos\alpha = a, \\ k(\cos\beta\sin\alpha + \sin\beta\sin\gamma\cos\alpha) = b, \\ -k\cos\gamma\sin\alpha = c, \\ k(\cos\beta\cos\alpha - \sin\beta\sin\gamma\sin\alpha) = b, \end{cases} \tag{20}$$

имеет решение относительно неизвестных k, α, β, γ.

Доказательство. Из первого и третьего уравнений легко находится значение угла α. Сложение квадратов первого и третьего уравнений, а также сложение второго и четвёртого, взятых с множителями $\sin\alpha$ и $\cos\alpha$, дает формулы нахождения неизвестных β, γ

$$\cos\beta = \frac{ad - bc}{k\sqrt{a^2 + c^2}},\ \cos^2\gamma = \frac{a^2 + c^2}{k^2}. \tag{21}$$

Если во второе уравнение подставить значения, определённые в (21), и избавиться от радикалов возведением обеих частей в квадрат, то получим биквадратное уравнение относительно k

$$k^4 - Qk^2 + (\det\mathbf{G})^2 = 0,$$

где $Q = a^2 + b^2 + c^2 + d^2$. Решение этого уравнения

$$k = \sqrt{0.5(Q + \sqrt{Q - 4(\det\mathbf{G})^2}} \tag{22}$$

определяет значение коэффициента масштабирования в явном виде и даёт возможность показать, что правые части равенств в формуле (21) по модулю не превосходят единицы. Следовательно, решение относительно неизвестных β, γ также существует.

Для предложенной последовательности операций, заменяющей АП в явном виде, легко реализовать декомпозицию вычисления параметров. Оценку сходства эталона $\mathbf{S_{pr}}$ с распознаваемым объектом можно найти непосредственно по формуле (13), как и параметры ортогонального преобразования и масштабирования $\alpha, k, \Delta\mathbf{w^s}$ на плоскости XOY, используя формулу (12). Локализацию местоположения глобального минимума (17) в пространстве параметров β, γ можно осуществить, используя метод сеток. Так как область изменения параметров β, γ ограничена, это позволяет получить оценку сходства, инвариантную к АП с гарантированной сложностью вычислений, а именно $N \cdot O(n)$, где N — количество ячеек в методе сеток.

Выводы

Полученные оценки сходства нетрудно ассоциировать с апостериорной вероятностью принадлежности объекта заданному классу. Это даёт все основания использовать для решения задачи распознавания только оценки сходства с эталонами. Ясно, что для распознавания большого множества объектов необходим только один эталон, что естественным образом оптимизирует вычислительную сложность алгоритма классификации. Степень изменения формы эталона, как категория шумов, легко трансформируется в оценку сходства, которую при определённых условиях, используя формулы (9), (13), можно получить теоретически. При построении классификаторов здесь не возникает проблем и с классом отказов, используя для решения этой задачи пороговое отсечение по оценкам близости с эталонами. Наконец, алгоритм распознавания на базе оценок сходства легко распараллеливается. Наибольшие трудности вычислительного характера возникают при поиске согласованных описаний, так как контур может быть представлен любым из бесконечного числа циклических векторных описаний. Решение этой задачи пока предполагает использование метода локализации глобального минимума с последующим уточнением его местоположения итерационными методами.

Литература

[1] *Файн В. С.* Опознавание изображений. — Москва: Наука, 1970. — 299 с.

[2] *Венцель Е. С.* Теория вероятностей. — Москва: Наука, 1964. — 576 с.

[3] *Васин Ю. Г., Лебедев Л. И., Пучкова О. В.* Контурные корреляционно-экстремальные методы обнаружения и совмещения объектов видеоинформации // Автоматизация обработки сложной графической информации, Горький: ГГУ, 1987. — С. 97–112.

[4] *Васин Ю. Г., Лебедев Л. И.* Инвариантные методы определения сходства плоских форм // Информационные технологии в анализе изображений и распознавании образов, Львов: Физ.-мат.ин-т АН УССР, 1990. — С. 225–228.

[5] *Vasin Ju. G., Lebedev L. I.* Recognition of base of similarity estimations, invariant with respect to affinities // 8-th Int'l. Conf. on Pattern recognition and image analysis: new information technologies, 2007. — Vol. 2. — Pp. 57–60.

Задача нахождения согласованных описаний в корреляционно-экстремальных контурных методах распознавания.*

Васин Ю. Г., Лебедев Л. И.

lebedev@pmk.unn.ru

Нижний Новгород, НИИ прикладной математики и кибернетики Нижегородского государственного национально-исследовательского университета им. Н. И. Лобачевского

Показывается, что задача нахождения согласованных описаний может быть сведена к задаче поиска глобального минимума на множестве циклических векторных описаний контура объекта. Предлагается метод локализации глобального минимума. Предлагаются итерационные методы уточнения местоположения глобального минимума. Даётся функциональная зависимость оценки близости от положения начальной точки описания контура объекта. Даётся оценка сложности нахождения согласованных описаний.

Введение

Основой корреляционно-экстремальных контурных методов распознавания являются методы вычисления оценок сходства (близости), инвариантных относительно различных групп преобразований. Для нахождения самих оценок сходства используется векторная (контурная) модель представления информации. В рамках этой модели контур может быть представлен любым из бесконечного числа циклических векторных описаний. Каждое из этих описаний получается на основе базового сдвигом начальной точки описания по контуру на величину $Q \in [0, S]$, где S — длина контура. Оценка близости с эталоном в зависимости от реализации описания контура будет меняться, то есть она является функцией параметра Q. Эти изменения обусловлены изменениями величин статистических смешанных корреляционных моментов, от значений которых оценка близости зависит функционально.

Постановка задачи

Пусть базовые описания эталона $\mathbf{S}$ и объекта $\mathbf{O}$ заданы последовательностями узловых точек на плоскости $\mathbf{w^s} = \{\mathbf{w}_1^s, \mathbf{w}_2^s, \ldots, \mathbf{w}_m^s\}$ и $\mathbf{w} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k\}$ соответственно, где $\mathbf{w}_j^s = (\dot{x}_j^s, \dot{y}_j^s)^\intercal$, $\mathbf{w}_j = (\dot{x}_j, \dot{y}_j)^\intercal$. Без ограничения общности можно предположить, что длины контуров совпадают и равны S (иначе производится масштабирование, например, объекта). Рассмотрим оценки близости, инвариантные относительно ортогональных преобразований $\varepsilon_m^{\mathbf{A}}$ и ортогональных преобразований и масштабирования $\varepsilon_m^{\mathbf{B}}$ [1,2]

$$\varepsilon_m^{\mathbf{A}} = \mathsf{D}\mathbf{w^s} + \mathsf{D}\mathbf{w} - 2R, \tag{1}$$

$$\varepsilon_m^{\mathbf{B}} = \mathsf{D}\mathbf{w} - \frac{R^2}{\mathsf{D}\mathbf{w^s}}, \tag{2}$$

где $R^2 = \mathrm{Sn}^2 + \mathrm{Cs}^2$, $\mathsf{D}\mathbf{w} = \mathsf{D}\mathbf{x} + \mathsf{D}\mathbf{y}$, $\mathsf{D}\mathbf{x} = \mathrm{cov}(\mathbf{x},\mathbf{y})$,

$$\begin{cases} \mathrm{Sn} = \mathrm{cov}(\mathbf{x^s}, \mathbf{y}) - \mathrm{cov}(\mathbf{y^s}, \mathbf{x}) \\ \mathrm{Cs} = \mathrm{cov}(\mathbf{x^s}, \mathbf{x}) + \mathrm{cov}(\mathbf{y^s}, \mathbf{y}). \end{cases} \tag{3}$$

Начальные статистические моменты и дисперсии $\mathsf{D}\mathbf{w^s}$, $\mathsf{D}\mathbf{w}$ постоянны для любых циклических описаний контура, а смешанные корреляционные статистические моменты зависят от переноса начальной точки описания. Так как эти параметры входят в формулы вычисления оценок близости $\varepsilon_m^{\mathbf{A}}$ и $\varepsilon_m^{\mathbf{B}}$ симметрично, то в дальнейшем, используя обозначение $\varepsilon_m(Q)$, будем подразумевать любую из них. Наиболее наглядно зависимость смешанных корреляционных статистических моментов прослеживается при описании контуров равномерно расположенными точками

$$\mathbf{P^s} = \{\mathbf{P}_1^s, \mathbf{P}_2^s, \ldots, \mathbf{P}_n^s\}, \qquad \mathbf{P}_i^s = (x_i^s, y_i^s)^\intercal;$$
$$\mathbf{P} = \{\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_n\}, \qquad \mathbf{P}_i = (x_i, y_i)^\intercal.$$

Тогда, в общем случае при $q \neq d$

$$\mathrm{cov}_q(\mathbf{x^s}, \mathbf{y}) - \mathrm{cov}_d(\mathbf{x^s}, \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i^s (y_{i+q} - y_{i+d}) \neq 0.$$

Так как оценка близости эталона и объекта должна отражать наименьшую ошибку при совмещении контуров, то, следовательно, для реализации этого утверждения необходимо решить оптимизационную задачу вида

$$\varepsilon_{\min} = \varepsilon_m(Q^*) = \min_{Q \in [0,S]} \varepsilon_m(Q). \tag{4}$$

Определение 1. *Описания эталона и объекта, при которых достигается наименьшее значение ошибки совмещения контуров среди всего множества циклических векторных описаний, будем называть согласованными.*

Таким образом, задача состоит в нахождении такого местоположения точки начала векторного описания объекта Q^* относительно исходного, которая удовлетворяла бы условию (4).

Работа выполнена при финансовой поддержке РФФИ, проект № 10-07-00330.

Методы решения

Задачу получения согласованных описаний разобьем на две задачи — разработку методов грубого определения местоположения глобального минимума и методов уточнения его местоположения. Для локализации местоположения глобального минимума воспользуемся следующим алгоритмом. Возьмём некий базовый эталон, который описывается минимальным количеством узловых точек и не имеет точек и осей симметрии. Эти условия предъявляются для обеспечения быстродействия вычисления оценок сходства и отсечения случаев копирования глобальных минимумов. Таким требованиям удовлетворяет треугольник Пифагора, масштабированный на периметр длины S. Далее по описаниям базового эталона и эталона из распознаваемого списка $\mathbf{S_i}$, $i = 1, \ldots, L$ на исходном описании эталона находится точка Q_i^*, удовлетворяющая условию (4). Эта величина является характеристикой эталона и определяется на этапе его формирования. Аналогично, при распознавании объекта на его контуре находится точка $Q^\star$, для которой при вычислении оценки сходства с базовым эталоном выполняется условие (4). Следовательно, если распознаваемый объект и эталон принадлежат одному классу, то согласование описаний будет заключаться в перемещении начальной точки на контуре объекта на величину $Q^* - Q_i^*$ в направлении обхода контура, так как это будет соответствовать математическому совмещению точек глобального минимума. Если же распознаваемый объект и эталон принадлежат разным классам, то такое совмещение в общем случае не обеспечивает локализации области поиска их согласованных описаний, что однако не приводит к ухудшению качества распознавания, так как в этом случае для любого описания оценка сходства будет ниже. Для получения областей локализации глобального минимума как эталонов, так и объекта используется метод сеток. Уточнение же его местоположения может проводится различными способами. Так как в локализованных областях все рассматриваемые функции являются унимодальными, то для уточнения можно применить метод Кифера, обеспечивающий при заданной погрешности меньшее количество итераций. Этот способ уточнения является эффективным для получения величин $Q^* - Q_i^*$, так как эта процедура не влияет на быстродействие самого алгоритма распознавания. Для уточнения же местоположения глобального минимума функции $\varepsilon_m(Q)$, являющейся оценкой близости эталона с объектом, предлагается более эффективный метод парабол. Этот метод, используя специфику функции $\varepsilon_m(Q)$, даёт более быструю сходимость к точке минимума и является более управляемым. Основой для разработки метода парабол является следующая теорема.

Теорема 1. *Если эталон и объект определяют один и тот же контур, то функция $\varepsilon_m(Q)$, которая является оценкой их близости в зависимости от начальной точки описания объекта, будет чётной относительно точки $Q^o = Q^* - Q_i^*$.*

Доказательство. Доказательство проведем для описаний эталона и объекта последовательностями интерполированных точек. Предположим вначале, что

$$x_i = x_i^s, y_i = y_i^s, \tag{5}$$

$i = 1, \ldots, n$, то есть выполнены условия теоремы при $Q^o = 0$. Тогда, обозначая $j = i - d$, и с учётом (5), последовательно сделав замены $x_{j+d} = x_{j+d}^s$, $y_{j+d} = y_{j+d}^s$, а $x_j^s = x_j,\ y_j^s = y_j$, получим

$$\begin{aligned}\mathrm{Cs}_{(-d)} =& \frac{1}{n}\sum_{i=1}^{n}(x_i^s x_{i-d} + y_i^s y_{i-d}) = \\ =& \frac{1}{n}\sum_{i=1}^{n}(x_i^s x_{i+d} + y_i^s y_{i+d}) = \mathrm{Cs}_{(+d)}.\end{aligned}$$

Аналогично,

$$\begin{aligned}\mathrm{Sn}_{(-d)} =& \frac{1}{n}\sum_{i=1}^{n}(x_i^s y_{i-d} - y_i^s x_{i-d}) = \\ =& \frac{1}{n}\sum_{i=1}^{n}(y_i^s x_{i+d} - x_i^s y_{i+d}) = -\mathrm{Sn}_{(+d)}.\end{aligned}$$

Отсюда, величина R^2 будет одной и той же, и, следовательно, $\varepsilon_m(d) = \varepsilon_m(-d)$ при любом значении d. Легко видеть, что если будут выполняться условия

$$x_i = x_{i+q^o}^s, \quad y_i = y_{i+q^o}^s, \quad i = 1, \ldots, n,$$

то $\varepsilon_m(q^o + d) = \varepsilon_m(q^o - d)$. В пределе при $n \to \infty$ $q^o(S/n) \to Q^o$ и, следовательно,

$$\varepsilon_m(Q^o + Q) = \varepsilon_m(Q^o - Q).$$

Теоретически, местоположение глобального минимума функции $\varepsilon_m(Q)$, имеющей ось симметрии $Q = Q^o$ определить просто, взяв в его окрестности две точки Q_1, Q_2, для которых $\varepsilon_m(Q_1) = \varepsilon_m(Q_2)$, и найдя его значение $Q^o = (Q_1 + Q_2)/2$. Нахождение точек Q_1, Q_2, удовлетворяющих условию $\varepsilon_m(Q_1) = \varepsilon_m(Q_2)$ задача не менее простая, чем исходная задача нахождения местоположения глобального минимума. Однако, можно использовать осесимметрию функции $\varepsilon_m(Q)$, аппроксимируя её параболой, уравнение которой легко находится по значениям, полученным при локализации в методе сеток. Вершина параболы даёт точку уточнённого местоположения глобального минимума, которая используется на последующих шагах уточнения данным методом. Как правило, для достижения необходимой точности требуется всего 2–3 итерации метода

парабол. Контроль осуществляется по результатам, полученным в точке вершины параболы для исходной и аппроксимирующей функций. Данный метод уточнения местоположения глобального минимума применим, если эталон и объект принадлежат одному классу, так как тогда они имеют похожие формы контуров и, поэтому, функция $\varepsilon_m(Q)$ будет близка к осесимметричной. Для более детального исследования функции $\varepsilon_m(Q)$ получим её аналитическое выражение в зависимости от аргумента Q.

В этих целях рассмотрим смешанный статистический начальный момент $\mathbf{Mx^sy}$, полученный по исходным описаниям эталона и объекта. В дальнейшем в зависимости от ситуации будем использовать оба варианта описаний как посредством интерполированных точек, так и с помощью узлов. Переместим начальную точку объекта на одну точку в направлении обхода контура, то есть сформируем описание $\mathbf{P^1} = \{\mathbf{P}_2, \mathbf{P}_3, \ldots, \mathbf{P}_n, \mathbf{P}_1\}$. Тогда

$$\mathbf{M^1x^sy} = \frac{1}{n}\sum_{i=1}^{n} x_i^s y_{i+1} = \mathbf{Mx^sy} - \frac{1}{n}\sum_{i=1}^{n} \Delta x_i^s y_i. \quad (6)$$

Значения Δx_i на отрезках прямых, определяемых узлами постоянны, поэтому

$$\mathbf{M^1x^sy} = \mathbf{Mx^sy} - \frac{1}{n}\sum_{j=1}^{m-1} \Delta x_j^s m_j \bar{y}_j(1), \quad (7)$$

где m_j — количество точек y_i попадающих на отрезок $[\mathbf{w_j^s}, \mathbf{w_{j+1}^s}]$, а $\bar{y}_j(1)$ — среднее значение на этом отрезке при перемещении начальной точки на одну интерполированную точку. Так как $\Delta x_j^s m_j$ при $n \to \infty$ будет стремиться к $(\dot{x}_{j+1}^s - \dot{x}_j^s)$, то выражение (7) можно переписать в виде

$$\mathbf{M^1x^sy} = \mathbf{Mx^sy} - \frac{1}{n}\sum_{j=1}^{m} (\dot{x}_{j+1}^s - \dot{x}_j^s)\bar{y}_j(1). \quad (8)$$

Далее, совместим начальную точку описания эталона с q-ой точкой описания объекта, то есть переместим начальную точку эталона на расстояние Q по направлению обхода контура объекта. Тогда с учётом (8) будем иметь

$$\mathbf{M^qx^sy} = \mathbf{Mx^sy} - \frac{1}{n}\sum_{j=1}^{m} (\dot{x}_{j+1}^s - \dot{x}_j^s) q \frac{1}{q}\sum_{i=1}^{q} \bar{y}_j(i). \quad (9)$$

В пределе при $n \to \infty$ величина $\bar{y}_j(i)$ будет стремиться среднему, вычисляемому по формуле (3)

$$\bar{y} = \frac{1}{2S}\sum_{l=1}^{k} (\dot{y}_{l+1} + \dot{y}_l) S_l, \quad (10)$$

а $q/n \to Q/S$, $\mathbf{M^qx^sy} \to \mathbf{M^Qx^sy}$, и формула вычисления (9) примет следующий вид

$$\mathbf{M^Qx^sy} = \mathbf{Mx^sy} - \frac{Q}{S}\sum_{j=1}^{m} (\dot{x}_{j+1}^s - \dot{x}_j^s)\bar{\bar{y}}_j(Q), \quad (11)$$

где $\bar{\bar{y}}_j(Q)$ — среднее, вычисляемое по значениям средних $\bar{y}_j(i) \to \bar{y}_j(Q_i)$, $Q_i \in [0, Q]$. Учитывая, что для непрерывных функций, к которым относится в предельном варианте $\bar{y}_j(Q_i)$, среднее на промежутке можно вычислять, используя интегральное исчисление, формула (11) нахождения $\mathbf{M^Qx^sy}$ примет вид

$$\mathbf{M^Qx^sy} = \mathbf{Mx^sy} - \frac{1}{S}\sum_{j=1}^{m} \Delta\dot{x}_j^s \int_0^Q \bar{y}_j(\xi)d\xi, \quad (12)$$

где $\Delta\dot{x}_j^s = (\dot{x}_{j+1}^s - \dot{x}_j^s)$. Для окончательного решения задачи нахождения $\mathbf{M^Qx^sy}$, необходимо найти формулу вычисления $\bar{y}_j(\xi)$.

Для этого сместим начальную точку эталона вдоль контура искомого объекта на величину ξ по направлению обхода. Предположим, что начало эталонного участка с номером j будет находиться между узлами $\mathbf{w_r}, \mathbf{w_{r+1}}$ контура эталона о определяться точкой $\hat{\mathbf{w}}_\mathbf{j} = (\hat{x}_j, \hat{y}_j)^\intercal$. Обозначая,

$$B_j = \sum_{i=1}^{r} S_i - \sum_{i=1}^{j-1} S_i^s, \quad \sin\gamma_r = \frac{\dot{y}_{r+1} - \dot{y}_r}{S_r},$$

где S_i, S_i^s расстояния между соответствующими узлами контуров объекта и эталона, будем иметь

$$\hat{y}_j = \dot{y}_{r+1} - B_j \sin\gamma_r + \xi\sin\gamma_r, \quad (13)$$

и на основании (10) первый член суммы при вычислении $\bar{y}_j(\xi)$ будет равен

$$(\hat{y}_j + \dot{y}_{r+1})\widehat{S} = [2\dot{y}_{r+1} - B_j\sin\gamma_r]B_j - \\ - \xi^2\sin\gamma_r + 2\xi(B_j\sin\gamma_r - \dot{y}_{r+1}), \quad (14)$$

где $\widehat{S} = S_r - B_j$.

Аналогично определяется точка $\hat{\mathbf{w}}_\mathbf{j+1} = (\hat{x}_{j+1}, \hat{y}_{j+1})^\intercal$, расположенная между узлами $\mathbf{w_t}$, $\mathbf{w_{t+1}}$ на контуре объекта и соответствующая точке эталона $\mathbf{w_{j+1}^s}$. Компонента $\hat{y}_{j+1}$ этой точки определяется по формуле

$$\hat{y}_{j+1} = \dot{y}_{t+1} - B_{j+1}\sin\gamma_t + \xi\sin\gamma_t, \quad (15)$$

в которой применяются выше принятые обозначения. Таким образом, последний член суммы, входящий в формулу вычисления $\bar{y}_j(\xi)$ будет равен

$$(\hat{y}_{j+1} + \dot{y}_t)(S_t - \widehat{S}) = [\dot{y}_r + \dot{y}_{r+1} - B_j\sin\gamma_t] \\ (S_t - B_j) - [\dot{y}_r + \dot{y}_{r+1} - B_j\sin\gamma_t + \\ + (S_t - B_j)\sin\gamma_t]\xi + \xi^2(S_t - B_j)\sin\gamma_t], \quad (16)$$

Промежуточные члены вычисления $\bar{y}_j(\xi)$ по формуле (10) не зависят от параметра ξ, поэтому среднее $\bar{y}_j(\xi)$ как функция является полиномом второй степени

$$\bar{y}_j(\xi) = A^j\xi^2 + B^j\xi + C^j, \quad (17)$$

где

$$A^j = \sin\gamma_t - \sin\gamma_r,$$
$$B^j = 2(\dot{y}_{t+1} - \dot{y}_{r+1}) + 2B_j\sin\gamma_r - 2B_{j+1}\sin\gamma_t,$$
$$C^j = 2B_j(\dot{y}_{r+1} - B_j\sin\gamma_r) + \sum_{i=r+1}^{t}(\dot{y}_{i+1} + \dot{y}_i) + 2B_{j+1}(\dot{y}_{t+1} + B_j\sin\gamma_t).$$

Следовательно, $\bar{\bar{y}}_j(Q)$ является полиномом третьей степени

$$\bar{\bar{y}}_j(Q) = A^jQ^3/3 + B^jQ^2/2 + C^jQ + D^j.$$

Подставляя найденное значение $\bar{\bar{y}}_j(Q)$ в (12) получаем в окончательном варианте формулу вычисления смешанных моментов второго порядка в зависимости от величины смещения точки точки начала описания объекта относительно базового задания контура

$$\mathbf{M^Q x^s y} = \mathbf{M x^s y} - (AQ^3 + BQ^2 + CQ + D), \quad (18)$$

где

$$\begin{cases} A = \frac{1}{3}\sum\limits_{j=1}^{m-1}(\dot{x}_{j+1} - \dot{x}_j)A^j, \\ A = \frac{1}{2}\sum\limits_{j=1}^{m-1}(\dot{x}_{j+1} - \dot{x}_j)B^j, \\ A = \sum\limits_{j=1}^{m-1}(\dot{x}_{j+1} - \dot{x}_j)C^j, \\ A = \sum\limits_{j=1}^{m-1}(\dot{x}_{j+1} - \dot{x}_j)D^j. \end{cases} \quad (19)$$

Из полученного результата следует, что оценка близости контуров эталона и объекта будет полиномом шестой степени, причём коэффициенты этого полинома будут переменными величинами. Из формул (13)–(17) следует, что изменение коэффициентов полинома будет происходить всякий раз, когда начальная точка описания участка эталона при движении по контуру объекта будет проходить один из его узлов. Так как узлов в описаниях эталона и объекта соответственно m и k, то, следовательно, таких изменений коэффициентов будет mk. Сами контуры как объекта так и эталона, следовательно, можно разбить на такое же число разной длины отрезков, на которых коэффициенты полинома постоянны.

Воспользоваться полученными результатами напрямую не представляется возможным, так как производная от $\varepsilon_m(Q)$ является полиномом пятой степени и найти решение уравнения $\frac{d\varepsilon_m(Q)}{dQ} = 0$ в явном виде затруднительно. Однако, на этапе уточнения местоположения глобального минимума этот результат позволяет более эффективно осуществлять контроль за точностью получаемого решения, а в некоторых ситуациях находить точное решение.

Выводы

Несмотря на удачную параметризацию оценки близости и решения задачи нахождения её функциональной зависимости в явном виде, полученный результат следует считать промежуточным. Конечно, для задачи распознавания объектов изображений можно удовлетвориться предложенными в данной работе методами локализации глобального минимума и методами уточнения его местоположения. Однако, в задачах, где требуется не только факт правильного распознавания, но и восстановления объекта по полученным параметрам, требования к точности определения местоположения глобального минимума резко возрастают. Отсюда следует необходимость увеличения количества итераций в методах уточнения местоположения глобального минимума, что существенным образом сказывается на быстродействии решения поставленной задачи. Поэтому необходимость получения точного решения по прежнему остаётся актуальной.

Литература

[1] *Васин Ю. Г., Лебедев Л. И., Пучкова О. В.* Контурные корреляционно-экстремальные методы обнаружения и совмещения объектов видеоинформации // Автоматизация обработки сложной графической информации, Горький: ГГУ, 1987. — С. 97–112.

[2] *Васин Ю. Г., Лебедев Л. И.* Инвариантные методы определения сходства плоских форм // Информационные технологии в анализе изображений и распознавании образов, Львов: Физ.-мат.ин-т АН УССР, 1990. — С. 225–228.

[3] *Васин Ю. Г., Лебедев Л. И., Пучкова О. В.* Оптимизация вычислительной и емкостной сложности алгоритмов распознавания объектов видеоинформации // Автоматизация обработки сложной графической информации, Нижний Новгород: ННГУ, 1990. — С. 62–86.

Распознавание направления переноса точки на плоскости на фоне случайных гауссовских отклонений

Жарких А. А., Бычкова С. М.
zharkikh090107@mail.ru, lyasnikovasm@yandex.ru
Мурманск, Мурманский государственный технический университет

В работе формулируется задача распознавания направления переноса точки на фоне случайных гауссовских отклонений, как классическая байесовская задача минимизации среднего риска. Специальным выбором элементов платежной матрицы и априорных вероятностей направления движения задача сводится к минимизации средней вероятности ошибки. Для определённых значений параметров задача минимизации средней вероятности ошибки приводит к простому решающему правилу распознавания направления переноса. Получены точные формулы вероятностей правильного распознавания на основе этого решающего правила. Результаты сравниваются с полученными ранее авторами результатами распознавания направления переноса на фоне случайных поворотов.

Цель работы — это получение классического байесовского решения распознавания направления переноса точки на плоскости на фоне изотропных гауссовских отклонений. Классическим разделом математической статистики является теория статистических решений. В этой теории вводится скалярный показатель, называемый средним риском принятия решений. Оптимальным считается решение, минимизирующее средний риск. Понятие риска пришло из экономических приложений. Одним из разделов теории статистических решений является проверка гипотез. Задача проверки гипотез формулируется следующим образом. Пусть существуют m событий. События $H_r, r = 0, \ldots, m-1$ называются гипотезами. В каждом из экспериментов может проявиться лишь одно из этих событий. При принятии решения в конкретном эксперименте статистик может допустить ошибку. Если статистик не допустил ошибку, то считается, что он принял правильное решение. Следовательно, можно ввести условные вероятности $P(T_j \mid H_r), j, r = 0, \ldots, m-1$. Вероятность $P(T_j \mid H_r), j, r = 0, \ldots, m-1$ означает, что в опыте реально проявлялось событие с номером r, а статистик принял его за событие с номером j. Здесь T_j — это j-ая область принятия решения. Если $j = r$, то это соответствует правильному решению. Если $j \neq r$, то эта условная вероятность отражает ошибочное решение. Кроме условных вероятностей, вводятся априорные вероятности $P(H_r), r = 0, \ldots, m-1$. Априорные вероятности показывают распределения гипотез до проведения опыта статистиком. Кроме указанных характеристик, вводятся величины $c_{rj}, j, r = 0, \ldots, m-1$. Эти величины образуют матрицу потерь (или платежную матрицу). На основе всех введённых характеристик записывается выражение среднего риска:

$$c = \sum_{j=0}^{m-1} \sum_{r=0}^{m-1} c_{rj} P(H_r) P(T_j \mid H_r). \qquad (1)$$

Величина среднего риска зависит от выбора элементов платежной матрицы, а также от выбора решающего правила и априорных вероятностей. Априорные вероятности — это те характеристики, на которые статистик повлиять не может. Статистик определяет элементы платежной матрицы, исходя из каких-то дополнительных соображений, а далее минимизирует выражение для среднего риска. В результате процедуры минимизации выбираются также значения $P(T_j \mid H_r)$, которые доставляют среднему риску c минимум. Значения этих условных вероятностей однозначно связаны с разбиением множества решений на подмножества T_j, $j = 0, \ldots, m-1$. Заметим, что изменение элементов платежной матрицы изменяет минимальное значение риска и разбиение пространства решений.

В данной работе рассматривается классическая задача проверки гипотез применительно к некоторому случайному сложному движению точки на плоскости. Сформулирована задача проверки m гипотез как общая задача минимизации риска. Однако приводится решение следующей частной задачи. Если все элементы платежной матрицы, соответствующие ошибочным решениям одинаковы, а все элементы платежной матрицы, соответствующие правильному решению равны 0, то задача сводится к задаче минимизации вероятности ошибки принятия решения. Для данной задачи получено решающее правило, которое показывает, что ошибка минимизируется, когда максимально скалярное произведение вектора выборочного среднего приращений на вектор направления движения. Результаты сравниваются с аналогичными результатами, полученными авторами ранее.

Модель движения точки на плоскости с гауссовскими отклонениями

Движение точки рассматривается в дискретном времени. За единицу времени осуществляется один шаг движения. За этот шаг точка осуществляет параллельный перенос в некотором направлении на величину S с вероятностью p, либо не смещается с вероятностью $q = 1 - p$. Одновременно с указанным переносом на этом же шаге точка откло-

няется на некоторую случайную величину в произвольном направлении. Координаты этого отклонения подчиняются двумерному нормальному распределению, компоненты которого имеют нулевые средние, одинаковые средние квадратические отклонения σ и не коррелируют между собой. Считается, что перенос точки осуществляется в одном из $m, (m \geqslant 2)$ направлений на плоскости, разделённых углами величины $\beta_r = \frac{2\pi}{m} r,\ r = 0, \ldots, m-1$. Наблюдателю известны значение m и одно из направлений. Это позволяет ему определить все направления и выбрать систему координат, ось абсцисс которой совпадает с одним из них. В процессе движения точки её координаты измеряются, записываются и используются для распознавания направления переноса. Предполагается, что точка двигалась до случайного момента начала наблюдений. Это означает, что наблюдатель точно уверен в существовании движения точки, и обнаруживать сам факт этого движения нет необходимости. Вследствие этого неважно, существовало ли это движение бесконечно долго или началось в какой-то предыдущий момент времени. Для удобства обозначим точку, с которой мы начали наблюдение, B_0. Координаты точки через k шагов движения можно записать следующим образом:

$$\begin{cases} x_k = x_{B_0} + m_k S \cos\beta_r + \sum_{i=1}^{k} \xi_i, \\ y_k = y_{B_0} + m_k S \sin\beta_r + \sum_{i=1}^{k} \eta_i. \end{cases} \tag{2}$$

Здесь $(x_{B_0}; y_{B_0})$ — координаты точки, с которой мы начали наблюдение; S — величина параллельного переноса; m_k — случайная величина, которая показывает количество параллельных переносов, совершенных точкой за k шагов, и представляет собой сумму k независимых случайных величин μ_k, каждая из которых подчиняется биномиальному закону; $\beta_r, r = 0, \ldots, m-1$ — угол, задающий истинное направление параллельного переноса; ξ_i, η_i — случайные отклонения с нормальным распределением, имеют нулевые средние, одинаковые дисперсии и не коррелируют между собой. Для удобства будем рассматривать разность между последующей и предыдущей наблюдаемыми координатами точки. Учитывая модель движения точки (2), имеем следующие случайные величины:

$$\begin{cases} \hat{x}_k = x_k - x_{k-1} = S\mu_k \cos\beta_r + \xi_k, \\ \hat{y}_k = y_k - y_{k-1} = S\mu_k \sin\beta_r + \eta_k. \end{cases} \tag{3}$$

Распознавание направления переноса с использованием байесовского решающего правила

Для распознавания направления переноса на основе байесовсого решающего правила [1], нами была выведена $2n$ мерная условная плотность распределения вероятностей случайных величин $\hat{X}_1, \hat{Y}_1; \ldots; \hat{X}_n, \hat{Y}_n$ при условии справедливости гипотезы $H_r,\ r = 0, \ldots, m-1$:

$$P_{\hat{X}_1,\hat{Y}_1;\ldots;\hat{X}_n,\hat{Y}_n}(\hat{x}_1, \hat{y}_1; \ldots; \hat{x}_n, \hat{y}_n \mid H_r) = \prod_{k=1}^{n} \Big(\frac{p}{2\pi\sigma^2} \exp\Big(-\frac{(\hat{x}_k - S\cos\beta_r)^2}{2\sigma^2} - \frac{(\hat{y}_k - S\sin\beta_r)^2}{2\sigma^2} \Big) + \frac{q}{2\pi\sigma^2} \exp\Big(-\frac{\hat{x}_k^2 + \hat{y}_k^2}{2\sigma^2} \Big) \Big). \tag{4}$$

Плотность распределения вероятностей (4) является функцией правдоподобия. Частный случай формулы (4), когда параллельный перенос осуществляется всегда $p = 1$, тогда функция правдоподобия имеет вид:

$$P_{\hat{X}_1,\hat{Y}_1;\ldots;\hat{X}_n,\hat{Y}_n}(\hat{x}_1, \hat{y}_1; \ldots; \hat{x}_n, \hat{y}_n \mid H_r) = \frac{1}{(2\pi)^n \sigma^{2n}} \exp\Big(-\frac{1}{2\sigma^2} \cdot \sum_{k=1}^{n} \big((\hat{x}_k - S\cos\beta_r)^2 + (\hat{y}_k - S\sin\beta_r)^2 \big) \Big). \tag{5}$$

В данной задаче нужно принять решение в пользу одной из гипотез $H_r,\ r = 0, \ldots, m-1$: «движение осуществляется в направлении с номером r, т. е в направлении, которое задано углом β_r». Отношение правдоподобия имеет следующий вид:

$$\Lambda = \frac{P_{\hat{X}_1,\hat{Y}_1;\ldots;\hat{X}_n,\hat{Y}_n}(\hat{x}_1, \hat{y}_1; \ldots; \hat{x}_n, \hat{y}_n \mid H_r)}{P_{\hat{X}_1,\hat{Y}_1;\ldots;\hat{X}_n,\hat{Y}_n}(\hat{x}_1, \hat{y}_1; \ldots; \hat{x}_n, \hat{y}_n \mid H_j)}, \quad j \neq r, j = 0, \ldots, m-1. \tag{6}$$

Нетрудно показать, что для случая $p = 1$ и с учётом (5) отношение правдоподобия примет вид:

$$\Lambda = \exp\Big(\frac{nS}{\sigma^2} (\tilde{x}_n \cos\beta_r + \tilde{y}_n \sin\beta_r - \tilde{x}_n \cos\beta_j - \tilde{y}_n \sin\beta_j) \Big), \tag{7}$$

здесь

$$\tilde{x}_n = \frac{1}{n} \sum_{k=1}^{n} \hat{x}_k, \quad \tilde{y}_n = \frac{1}{n} \sum_{k=1}^{n} \hat{y}_k.$$

Прологарифмируем отношение правдоподобия (7), тогда получим следующее выражение:

$$\ln\Lambda = \frac{nS}{\sigma^2} \Big(\tilde{x}_n \cos\beta_r + \tilde{y}_n \sin\beta_r - \tilde{x}_n \cos\beta_j - \tilde{y}_n \sin\beta_j \Big). \tag{8}$$

Рассмотрим частную задачу, когда минимизация среднего риска сводится к минимуму средней вероятности ошибки. В этом случае порог принятия

решения $\Lambda_0 = \frac{P(H_j)}{P(H_r)}$. Для практических целей интересен случай, когда априорные вероятности гипотез одинаковы. Тогда $\Lambda_0 = 1, \forall j, r$. В этом случае при $\Lambda \geqslant 1 (\ln \Lambda \geqslant 0)$ решение принимается в пользу гипотезы H_r. Тогда для случая $p = 1$, согласно (8), получим следующее правило для разделения $\mathbb{R}^2$ на подмножества, чтобы средний риск оказался минимальным:

$$\tilde{x}_n \cos\beta_r + \tilde{y}_n \sin\beta_r \geqslant \tilde{x}_n \cos\beta_j + \tilde{y}_n \sin\beta_j, \quad j \neq r, j = 0, \ldots, m-1. \qquad (9)$$

Мы будем применять полученное правило (9) и для случая $p \neq 1$ в силу его простоты.

Мы определили вероятность правильного распознавания направления параллельного переноса за n шагов наблюдения, используя следующие соображения. Предположим, что истинный параллельный перенос соответствует направлению, задаваемому углом $\beta_r = 0$, т. е. осуществляется в положительном направлении оси абсцисс. Введем для удобства следующие случайные величины:

$$x = n\tilde{x}_n = \sum_{k=1}^{n} \hat{x}_k, \quad y = n\tilde{y}_n = \sum_{k=1}^{n} \hat{y}_k.$$

Очевидно, что решающее правило не изменится, если заменить в нём вектор $(\tilde{x}_n; \tilde{y}_n)$ на вектор $(x; y)$. На основе элементарных тригонометрических преобразований неравенства (9) можно показать, что $P(T_0 \,|\, H_0)$ соответствует попаданию $(x; y)$ в сектор бесконечного радиуса, задаваемого углами $-\pi/m$ и π/m . Непопадание в этот сектор соответствует ошибке, т. е. выбору неверного направления переноса. Если повернуть систему координат на угол $\beta_l = 2\pi l/m$ против часовой стрелки, то направление $r = 0$ станет направлением l и все номера направлений изменятся согласно циклической перестановке $r_1 = (r+l) \mod m$. Однако это не отразится на мере сектора по сравнению с мерой плоскости. Поэтому можно считать, что перенос осуществляется в положительном направлении оси абсцисс, т. е. $r = 0, \beta_r = 0$. При переносе в положительном направлении оси абсцисс средний риск является минимальным, если $(x; y)$ попадает в сектор бесконечного радиуса:

$$\begin{cases} -x \operatorname{tg}(\frac{\pi}{m}) < y < x \operatorname{tg}(\frac{\pi}{m}), \\ x > 0. \end{cases} \qquad (10)$$

Вероятность попадания в сектор, заданный углами $-\pi/m$ и π/m равна условной вероятности правильного распознавания направления за n шагов и, с учетом симметрии направлений, равна вероятности правильного распознавания направления параллельного переноса за n шагов. Обозначим $P_{XY}(xy)$ плотность распределения вероятностей системы двух случайных величин X и Y. Тогда, согласно системе неравенств (10), вероятность правильного определения направления переноса за n шагов наблюдения:

$$P_{\text{true}}(n) = \begin{cases} \int\limits_0^{+\infty} \int\limits_{-x\operatorname{tg}\frac{\pi}{m}}^{x\operatorname{tg}\frac{\pi}{m}} P_{XY}(x, y \,|\, H_0)\, dydx, \\ \qquad\qquad\qquad\qquad m > 2; \\ \int\limits_0^{+\infty} \int\limits_{-\infty}^{+\infty} P_{XY}(x, y \,|\, H_0)\, dydx, m = 2. \end{cases} \qquad (11)$$

Заметим, что мнимая единица, встречающаяся в приведенных ниже формулах, обозначается через i. Совместная плотность распределения случайных величин X и Y также была получена нами и определяется выражением:

$$P_{XY}(x, y \,|\, H_0) = \frac{1}{(2\pi)^2} \int\limits_{-\infty}^{+\infty} \int\limits_{-\infty}^{+\infty} \theta_{XY}(u, v \,|\, H_0) \times \times \exp(-i(xu + yv))dudv, \qquad (12)$$

здесь $\theta_{XY}(u, v \,|\, H_0)$ — двумерная характеристическая функция пары случайных величин X и Y, определяемая выражением:

$$\theta_{XY}(u, v \,|\, H_0) = \exp\left(-\frac{n\sigma^2}{2}\left(u^2 + v^2\right)\right) \times \times \left(p \cdot \exp(iSu) + q\right)^n. \qquad (13)$$

На основе (11), (12) и (13) можно записать вероятность $P_{true}(n)$ правильного распознавания направления параллельного переноса точки за n шагов наблюдения для различных значений параметров. Например, когда $m = 2$, вероятность правильного определения направления переноса определяется выражением:

$$P_{true}(n) = \frac{1}{2} + \frac{1}{2\pi} \int\limits_{-\infty}^{+\infty} \frac{1}{iu} \exp\left(-\frac{u^2 n}{2}\right) \times \times \left(p \cdot \exp\left(iu\frac{S}{\sigma}\right) + q\right)^n du. \qquad (14)$$

Для $m \neq 2$ аналогичное выражение выглядит следующим образом:

$$P_{true}(n) = \frac{1}{2\pi} \int\limits_{-\infty}^{+\infty} \frac{1}{iu} \exp\left(-\frac{u^2 n}{2}\right) \times \times \left(p \cdot \exp\left(iu\frac{S}{\sigma}\sin\left(\frac{\pi}{m}\right)\right) + q\right)^n du + \frac{1}{2\pi^2} \times \times \int\limits_{-\infty}^{+\infty} \int\limits_{-\infty}^{+\infty} \frac{1}{uw} \exp\left(-\frac{u^2 - 2uw\sin\left(\frac{\pi}{m}\right) + w^2}{2} n\right) \times \times \left(p \cdot \exp\left(i\frac{S}{\sigma}\left(u - w\sin\left(\frac{\pi}{m}\right)\right)\right) + q\right)^n dudw. \qquad (15)$$

Сравнение с задачей распознавания направления случайного переноса на фоне случайных поворотов

В работах [2, 3] нами была рассмотрена аналогичная модель движения точки, но помехой для распознавания направления переноса были случайные повороты движущейся точки относительно её центра вращения. Нами было показано, что на шаге n решающее правило для проверки гипотез о направлении переноса реализуется нахождением максимума (максимум находится по значениям r, $r = 0, \ldots, m-1$ — номер направления):

$$l = \arg\max(\tilde{x}_n \cos\beta_r + \tilde{y}_n \sin\beta_r), \qquad (16)$$

где $(\tilde{x}_n; \tilde{y}_n)$ — выборочные средние координат точки наблюдения. Решающее правило (16) применяется для любых вероятностей p осуществления параллельного переноса. Вывод данного правила был основан на геометрии задачи. Таким образом, мы получили для моделей движения с различными помехами аналогичные решающие правила (9) и (16) для определения направления параллельного переноса. Отличие состоит лишь в том, что в решающем правиле (9) $(\tilde{x}_n; \tilde{y}_n)$ — это выборочные средние от разности координат последующей и предыдущей наблюдаемой точки, а в решающем правиле (16) $(\tilde{x}_n; \tilde{y}_n)$ — это выборочные средние координат точки наблюдения. Также в предыдущих работах нами были выведены вероятности правильного распознавания направления переноса точки на фоне случайных поворотов.Для случая $m = 2$ вероятность правильного распознавания направления имеет вид:

$$P_{true}(n) = \frac{1}{2} + \frac{1}{2\pi} \int\limits_{-\infty}^{+\infty} \frac{1}{iu} J_0(n \cdot u) (J_0(u))^n \times \times \prod_{k=1}^{n} \left(p \cdot \exp\left(iu \frac{S}{R} k \right) + q \right) du. \qquad (17)$$

Для случая $m \neq 2$, вероятность правильного распознавания направления имеет вид:

$$P_{true}(n) = \frac{1}{2\pi} \int\limits_{-\infty}^{+\infty} \frac{1}{iu} J_0(n \cdot u)(J_0(u))^n \times \times \prod_{k=1}^{n} \Big(p \cdot \exp\Big(iu \tfrac{S}{R} k \sin\big(\tfrac{\pi}{m}\big) \Big) + q \Big) du + + \frac{1}{2\pi^2} \int\limits_{-\infty}^{+\infty} \int\limits_{-\infty}^{+\infty} J_0\Big(n \cdot \sqrt{u^2 - 2uw \sin(\tfrac{\pi}{m}) + w^2} \Big) \times \times \frac{1}{uw} \Big(J_0\Big(\sqrt{u^2 - 2uw \sin(\tfrac{\pi}{m}) + w^2} \Big) \Big)^n \times \times \prod_{k=1}^{n} \Big(p \cdot \exp\big(i \big(u - w \sin\big(\tfrac{\pi}{m}\big) \big) \tfrac{S}{R} k \big) + q \Big) du dw. \qquad (18)$$

В формулах (17) и (18) $J_0(\cdot)$ — функция Бесселя нулевого порядка, n — число шагов наблюдения, S — величина смещения, R — радиус вращения движущейся точки относительно её центра.

Таким образом, формулы (14) и (17) (для $m = 2$), а также (15) и (18) (для $m \neq 2$) имеют сходную структуру. Это связано со сходством моделей движения и решающих правил распознавания.

Выводы

Для достижения поставленной цели были решены следующие задачи:

- рассмотрена математическая модель движения точки на плоскости, включающая в себя описание случайных переносов и случайных гауссовских некоррелированных отклонений;
- для данной модели движения сформулирована задача проверки статистических гипотез определения направления переноса на фоне гауссовских случайных отклонений;
- рассмотрен частный случай распознавания в котором задача минимизации среднего риска сводится к задаче минимизации вероятности ошибочного решения;
- проведено сравнение полученного решающего правила с решающим правилом для определения направления переноса на фоне случайных поворотов и установлено сходство полученных решающих правил;
- проведено сравнение вероятностей правильного распознавания направления переноса на фоне гауссовских отклонений и на фоне случайных поворотов.

Литература

[1] *Фукунага К.* Введение в статистическую теорию распознавания образов. — Москва: Наука, 1979. — 368 с.

[2] *Жарких А. А., Бычкова С. М.* Вероятности распознавания направления переноса в одной модели случайного движения точки на плоскости // 8-я Международная конференция «Интеллектуализация обработки информации», Москва: МАКС Пресс, 2010. — С. 346–349.

[3] *Zharkikh A., Bychkova S.* Statistical theory of recognition of direction of random shift of point on a plane // 10-th Int'l. Conf. on Pattern Recognition and Image Analysis: New Information Technologies (PRIA-10-2010), 2010. — Vol. 1, — Pp. 131–134.

Обзор и анализ распознавателей рукопечатных символов*

Дробков А. В., Семенов А. Б.
anatoly.drobkov@gmail.com, semenov@tversu.ru
Тверь, Тверской Государственный Университет

Целью данной работы является построение и анализов распознавателей для рукопечатных символов — печатных символов, вводимых человеком от руки и получаемых с помощью планшета или из отсканированных документов. Исследуется устойчивость распознавателей к изменению начертания символов. Наилучший полученный распознаватель обладает точностью 99.6% для конкретной тестовой базы символов.

Разработка методов для интеллектуального распознавания символов (ICR) из рукописных текстов или фотографий по сей день является актуальной задачей. Для некоторых языков (например, восточных [7]) успехи в данной области являются более скромными по сравнению с латиницей и кириллицей. Исследование, описанное в данной работе, проводится на нескольких собранных авторами базах реальных печатных символов русского языка (т. н. *рукопечатных* символов). Анализируется точность нескольких методов, основанных в большей или меньшей степени на ранее известных.

Проекции символа

Будем задавать символ в виде чёрно-белого растра , т. е. прямоугольной матрицы A размера n на m. *Вертикальной проекцией* растра будем называть массив $H[1..m]$, где $H[i]$ — это число чёрных пикселей, попавших в i-й столбец матрицы A. *Горизонтальной проекцией* будем называть массив $V[1..n]$, где $V[i]$ — это число чёрных пикселей, попавших в i-ю строку матрицы . Максимальный используемый размер рамки символа — 62×62 пикселей.

Рис. 1. Рукопечатные символы.

На рисунке 1 изображены конкретные символы. На рисунках 2, 3 — их горизонтальные и вертикальные проекции.

Построение алгоритма распознавания на основе проекций основывается на подтверждающемся на практике предположении, что символы из одного класса имеют похожие графики соответствующих им вертикальных и горизонтальных проекций (похожий метод описан в [6]). Графики же проекций символов, принадлежащих различным классам, достаточно отличаются, что хорошо видно из рисунков.

Работа выполнена при финансовой поддержке РФФИ, проект №11-01-00783-а.

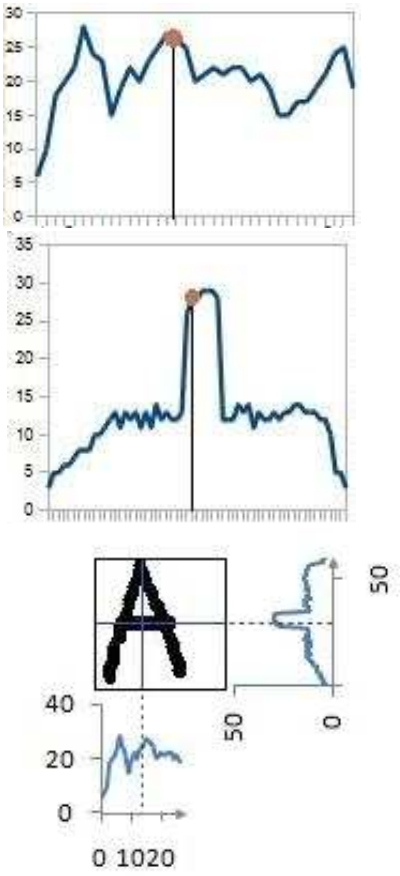


Рис. 2. Вертикальная и горизонтальные проекции символа «А».

Мерой сходства графиков выбирается *расстояние Левенштейна* между соответствующими конечными числовыми последовательностями.

Обозначим $c(w1[i], w2[j])$ функцию стоимости замены символа $w1[i]$ на $w2[j]$ в строках $w1$ и $w2$, и пусть константы $insCost$ и $delCost$ — это стоимости, соответственно, вставки и удаления символа в строке $w1$.

Авторами статьи были подготовлены тестовая (T) и экспертная (E) базы, состоящие из нескольких тысяч символов. Распознавание функционировало по методу ближайшего соседа. Результаты приведены ниже.

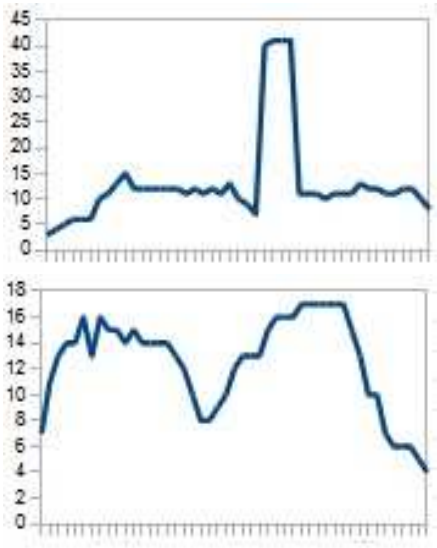


Рис. 3. Вертикальная и горизонтальные проекции символа «Ж».

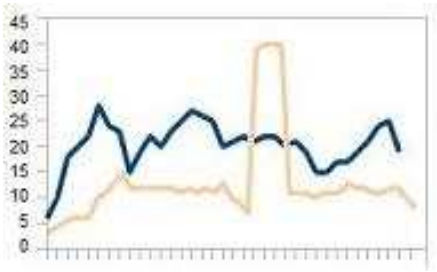


Рис. 4. Сравнение вертикальных проекций символов «А» и «Ж».

ID	$\|T\|$	$\|E\|$	Агрегац Функция	Точн.
1	3581	3603	$V+H$	87,5%
11	1786	5398	$V+H$	91,9%
13	1786	5398	$V+H+$ $+45Grad+$ $+45Grad2$	98%
14	5398	1786	$V+H+$ $+45Grad+$ $+45Grad2$	96%
15	526	526	$V+H+$ $+45Grad+$ $+45Grad2$	93%

Стоимости операций редактирования задавались равными $0,1$ для $insCost$, $0,3$ для $delCost$ и $|x-y|$ для стоимость замены x на y. Отметим также, что расстояние Левенштейна определялось не между самими проекциями, а между их образами. Данные расстояние обозначены в таблице как V, H, $45Grad$ и $45Grad2$ — для вертикальной, горизонтальной и с проекции с поворотом на 45 градусов. *Образ проекции* p задавался следующим образом:

$$I(p) = N(M(p)),$$

где операторы N и M — это процедуры *нормализации* и *фильтрации* соответственно.

Метрика на базе двумерного DCT

Формально двумерное дискретное косинусное преобразование (DCT 2D) задаётся в следующем виде [5]:

$$C(u,v) = \alpha(u)\alpha(v)\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} f(x,y)\cdot \cdot\cos\left[\frac{\pi(2x+1)u}{2N}\right]\cos\left[\frac{\pi(2y+1)v}{2N}\right].$$

Таким образом, применяя его для каждой из проекций (горизонтальной, вертикальной и соответствующей углу поворота α), мы получаем по вектору длины n. Затем данные вектора сравниваются с помощью Манхэттенского расстояния по методу ближайшего соседа (тестовая база содержит 1786 символов, экспертная — 5398).

n	Точность	Скорость (симв./сек)	Проекции
18	99,3%	1,8	$V, H, 45°$
25	99,4%	1,42	$V, H, 45°$
35	99,6%	1,41	$V, H, 45°$

Метод пересечений

Метод пересечений описан в [1] (используется для машинопечатных символов, в отличие от данной статьи); вариация, исследуемая авторами здесь, имеет следующие отличия: другие наборы прямых (см. рис. 5), сформулирован конкретный способ определения расстояния между наборами выходных векторов на основе сумм расстояний Левенштейна (в [1] не обозначен конкретный алгоритм).

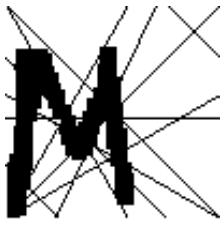


Рис. 5. Символ «М» с конкретным набором прямых метода.

Получена следующая точность распознавания на двух наборах прямых.

$\|T\|$	$\|E\|$	Наб. прямых	Точность	Симв./сек
1786	5398	$K3$	76%	0,69
1786	312	$K3$	82%	11,26
1786	5398	$K1$	89%	0,69

Метод проекции контура

Жирные рукопечатные символы хорошо определяются своей внешней границей. С этой точки зрения информация о внутренних границах и толщине штриха является избыточной для определения класса символа.

Рис. 6. Символ и его внешняя граница.

Среди символов, рассматриваемых в статье, только символ «Ы» имеет более одной границы (символы «Ё» и «Й» приравниваются к «Е» и «И»), все остальные символы — в точности одну. При этом ограничение на *связность* изображения символа и неразрывность его контура является обязательным для корректного определения внешней границы. Анализу будет подвергаться компонента связности на изображении, имеющая максимальное число чёрных пикселей. Итак, пусть $Contour$ — это конечная последовательность точек внешней границы символа, построенная при обходе границы против часовой стрелки, причём первый элемент этой последовательности является верхней точкой изображения, имеющей минимальную абсциссу. Точка $c = (cx, cy)$ — это центр bounding box'а, лежащий внутри внешней границы. Если c не лежит внутри границы, то cx изменяется таким образом, чтобы это условие выполнялось (линейный поиск в строке cy растра). *Опорным вектором* будем называть вектор $support = (c, cRight)$, где $c.cy = cRight.cy$, а — это абсцисса самой правой точки границы, лежащей в строке cy. Выберем угол φ, который будет использоваться при обходе границы, $n = \frac{2*\pi}{\varphi}$ определяет порядок метода (число элементов в результирующем векторе v). Определим последовательность элементов вектора v следующим образом:

$$v_0 = \|support\|, v_{i+1} = \|(cx, cy); (px, py)\|,$$

где px и py — точки пересечения луча $support$, повернутого на $i * \varphi$ радиан с внешней границей (если

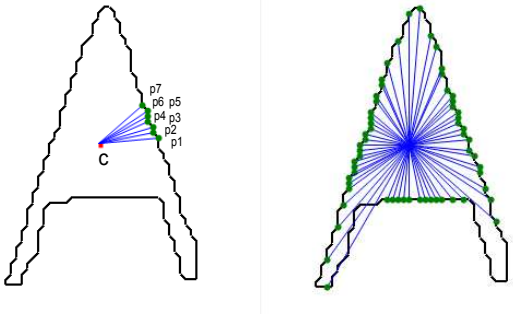


Рис. 7. Первые 7 точек построения проекции (слева), все точки построения проекции (справа).

таких точек несколько, алгоритм выбирает первую точку в порядке обхода контура).

На рисунке 7 показаны первые 7 векторов построения проекции при $n = 60$, точка pi соответствует элементу v_i. Далее вектор v подвергается нормализации:

$$\acute{v}_i = \frac{v_i}{\max\limits_j(v_j)}.$$

Графики функции $\acute{v}_i$ для различных символов представлены на рисунках ниже.

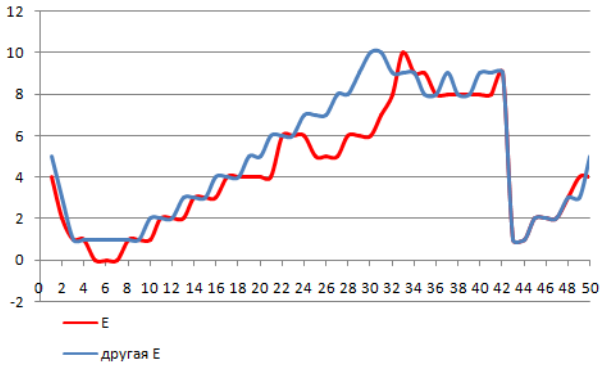


Рис. 8. Графики проекций контура различных букв «Е».

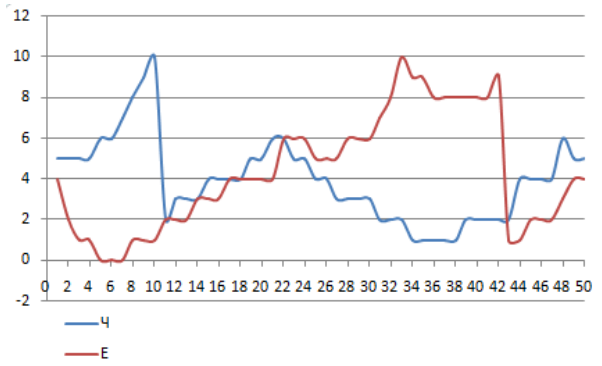


Рис. 9. Графики проекций контура букв «Ч» и «Е».

Из рисунков видно, что графики для символов одного класса имеют сходство, графики символов из разных классов существенно различаются (рис. 9. В качестве расстояния между двумя проекциями (по построению они гарантированно имеют одинаковую длину) используется *Манхэттенское*.

Для баз Testing (254 символа), Expert (312 символов):

n	Точность	Скорость (символов/сек)
20	75%	~ 140
25	76%	
30	78%	
40	80%	
50	80%	
60	83%	

Для баз Testing (1786 символов), Expert (5398 символов):

n	Точность	Скорость (символов/сек)
40	93%	81
50	94,3%	74
60	94,7%	68

Гибридный распознаватель

Итого, были построены следующие методы:

Codename	Распознаватель	Точность
PROJ	Проекция	98%
INT	Пересечения	89%
DCTPROJ	Проекция + DCT2D	99,6%
CONTOUR	Проекция контура	94,7%
NN	Нейронная сеть	87%

Распознаватель с помощью нейронной сети был получен авторами данной статьи в [2] и здесь не описывается. Формируются следующие комитеты, функционируюшие по принципу, изложенному в [4]. Комитет выдвигает набор гипотез, *экспертный* распознаватель их проверяет и выносит решение.

ID	Экс	Ан1	Ан2	Ан3
1	DCTPROJ	INT	CONTOUR	-
2	DCTPROJ	INT	CONTOUR	NN

Результаты распознавания комитетов следующие (nH — число гипотез, (у) — точность *устойчивости* при распознавании символов, полученных с помощью геометрических деформаций [2]):

ID	|E|	|T|	nH	Точн./Уст.	Симв./сек
1	312	1786	4	95,4%	4,2
2	312	1786	5	97,2%	3,63
1	312	1786	6	96,6%	3,78
2	312	1786	6	97,5%	3,53
1	5398	1786	4	96,6%	0,81
2	5398	1786	6	98,6%	0,85
1	5398	1581	6	75%(у)	0,8
2	5398	1581	6	80,7%(у)	0,8

Выводы

В проведённом исследование на реальных рукопечатных символах был достигнут наилучший результат точности 99,6%. Наилучший результат гибридного распознавателя — 98,6%. Вопрос поиска оптимального гибридного распознавателя остаётся открытым.

Литература

[1] *Андреев С. В., Бондаренко А. В.* Алгоритмическое обеспечение прототипа устройства считывания машиночитаемых документов // Москва: Институт прикладной математики им. М. В. Келдыша РАН, 2003. — 32 с.

[2] *Дробков А. В., Семенов А. Б.* Исследование одного метода распознавание рукопечатных символов // Вестник Тверского государственного университета, серия «Прикладная математика». — 2009. № 15. — С. 15–26.

[3] *Хайкин С.* Нейронный сети: полный курс. — Москва: Издательский дом «Вильямс», 2006. — 1104 с.

[4] *Ян Д. Е.* Исследование, развитие и реализация методов автоматического распознавания рукописных текстов в компьютерных системах // Диссертация на соискание ученой степени кандидата-физико математических наук. Москва: МФТИ, 2003.

[5] *Khayam S. A.* The Discrete Cosine Transform (DCT): Theory and Applications. — Michigan State University: Department of Electrical and Computer Engineering, 2003. — 32 p.

[6] *Miciak M.* Character Recognition Using Radon Transformation and Principal Component Analysis in Postal Applications // Sadhana, 2007. — Vol. 32, Part. 5. — Pp. 521–533.

[7] *Sanjeev Kunte R.* A simple and efficient optical character recognition system for basic symbols in printed Kannada text // Proceedings of the International Multiconference on Computer Science and Information Technology, 2008. Vol. 32, Pp. 495–500.

Использование образцов некорректных символов при обучении классификатора*

Сорокин С. В., Грицай А. А., Пономарёв С. А.
sergey@tversu.ru
Тверь, Тверской государственный университет, ООО «Комплексные системы»

В статье рассматривается возможность повышение эффективности систем распознавания текста путём введения дополнительного класса объектов, распознаваемых классификатором. Этот класс представляет объекты, не являющиеся корректными изображениями букв. Рассмотрено два способа построения обучающих примеров для этого класса: случайный и использующий фрагменты букв. Предложено решающее правило для композитных классификаторов, опирающееся на уровень принадлежности объекта этому классу. Показано, что данное правило является более эффективным, чем метод простого голосования.

При создании классификаторов в рамках концепции обучения с учителем системе предъявляют образцы объектов, относящихся к распознаваемым системой классам. Например, для систем распознавания текста такими образцами являются изображения букв. Типичный классификатор в системе распознавания текста, обученный таким образом, вычисляет уровень принадлежности рассматриваемого символа к классам, соответствующим буквам распознаваемого алфавита. Однако, при этом игнорируется ещё один класс возможных изображений — изображения, не являющиеся буквами. Такие изображения могут возникать, например, в случае некорректного определения границ букв на этапе сегментации изображения. Понятно, что эффективность системы распознавания в целом в таком случае будет зависеть от способности выявлять и отбрасывать подобные некорректные разбиения.

Далее мы рассмотрим два возможных варианта введения в обучающую последовательность изображений, соответствующих классу «не букв» и рассмотрим результаты работы полученного классификатора в системе распознавания текста.

Используемая система распознавания и классификаторы

Система распознавания текста, в рамках которой проводились эксперименты, включает этапы предобработки, сегментации, и распознавания.

Этап предобработки включает корректировку поворота изображения страницы, удаления шумов и нормализации изображения. Поскольку рассматриваемые в данной статье эксперименты проводились над искусственно сгенерированным текстом, предобработка не проводилась.

Сегментация выделяет структурные элементы текста, такие как строки, слова и буквы. Выделение соответствующих элементов осуществляется путём анализа минимумов и максимумов функции числа тёмных точек в строках (для поиска строк)

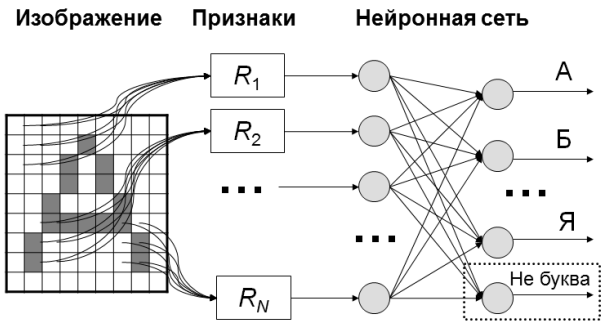


Рис. 1. Простой классификатор.

или столбцах (для поиска слов и букв) изображения. Поскольку точная сегментация на уровне букв является сложной задачей мы не принимаем на этапе сегментации окончательного решения о расположении границ букв внутри слова, а вместо этого формируем список возможных положений таких границ, с различными уровнями уверенности, определяющимися выполнением эвристических условий, основанных на работе [1].

На этапе распознавания анализируются возможные варианты разбиения слова на буквы, формируемые на основании информации от модуля сегментации. Для каждого варианта разбиения проводится классификация предполагаемых фрагментов-букв и вычисляется уровень уверенности системы в данном прочтении, зависящий от значений уровней уверенности использованных границ, выхода классификатора и частот соответствующих буквосочетаний в русском языке.

Использованный классификатор использует подход композиции алгоритмов с использованием простого голосования. Работа базового классификатора осуществляется в два шага (рис. 1). Сначала по исходному изображению вычисляются признаки. Значение каждого признака является функцией от яркостей некоторого подмножества пикселей изображения. В результате получается вектор значений признаков, который поступает на вход нейронной сети. Каждый выход сети соответствует одной из букв алфавита, а получаемое на выходе

Работа выполнена в рамках ФЦП «Научные и научно-педагогические кадры инновационной России» 2009–2013 г.г.

Таблица 1. Сводные данные по расстояниям между векторами признаков букв.

Шрифт	Минимальное	Среднее	Максимальное
Arial	0,057	0,231	0,352
Times	0,064	0,285	0,452
Courier	0,068	0,196	0,307

значение рассматривается как уровень возможности принадлежности анализируемого изображения к классам букв.

Нами использовались базовые классификаторы, у которых число признаков равняется 50, а нейронная сеть не содержит скрытых слоёв. Как показали практические эксперименты, такая архитектура позволяет решать задачу распознавания символов одного начертания.

В классификаторах использовались признаки прямоугольной формы:

$$P\left((x_1,y_1),(x_2,y_2)\right) = \{(x,y)|\, \min\{x_1,x_2\} \leqslant x \leqslant \max\{x_1,x_2\} \quad \min\{y_1,y_1\} \leqslant y \leqslant \max\{y_1,y_2\}\},$$

где (x_1,y_1), (x_2,y_2) — координаты углов, являющиеся параметрами признака.

Значение признака вычисляется по формуле

$$R(P) = \frac{\sum_{p\in P}(255-\mathrm{img}(p))}{255 * |P|},$$

где $\mathrm{img}(p)$ — значение яркости изображения img в точке p (0 — точка чёрная, 255 — белая); $|P|$ — число точек, входящих в P.

Генерация признаков и обучение нейронных сетей для базовых классификаторов осуществлялось автоматически, с помощью алгоритма, описанного в [2].

Случайные образцы

Первая использованный нами метод создания обучающих примеров для класса «не буквы» состоял в генерации случайных обучающих векторов в пространстве признаков. Чтобы исключить коллизии с изображениями букв, вычислялось минимальное расстояние от сгенерированного случайного вектора до векторов, соответствующих изображениям букв из обучающей последовательности. Если это расстояние было меньше заданного порога, то данный вектор отбрасывался и вместо него генерировался другой. Для определения порога были измерены расстояния между векторами, соответствующими буквам различных шрифтов (таб. 1). В качестве значения порога было выбрано близкое к максимальному расстоянию значение 0,3.

Таблица 2. Результаты тестирования метода случайных «не букв».

Правило	Распознано слов	Расстояние
Голосование	171	1,3
Правило (1)	170	1,2

Решающее правило. Для учёта дополнительного выхода базовых классификаторов нами было введено новое решающее правило: сначала выбирается один базовый классификатор, у которого выходное значение для класса «не буква» минимально, и затем выбирается класс, для которого данный классификатор даёт максимальное выходное значение:

$$C = \arg\max_{a\in A}\{b_m(\mathrm{img},a)\}, \qquad (1)$$

$$m = \arg\max_{i\in 1,\ldots,n}\{b_i(\mathrm{img},\text{«не буква»})\},$$

где A — множество распознаваемых букв, n — число базовых классификаторов, $b_i(\mathrm{img},a)$ — отклик i-го базового классификатора для класса a на изображении img.

Методика тестирования. Для оценки эффективности рассматриваемого способа создания примеров для класса «не буква» и решающего правила были созданы 10 базовых классификаторов, обученных распознавать символы шрифта Arial, 14 pt. К буквам шрифта были добавлены 60 случайных примеров для класса «не буква».

Системе распознавания текста предлагалось распознать набор из 300 случайно выбранных из словаря слов, средней длинной 8,65 символов. Изображение каждого слова генерировалось с помощью функции печати текста операционной системы шрифтом Arial, 14 pt. Оценивалось число правильно распознанных слов, а в случае отсутствия правильного варианта — минимальное расстояние Левенштейна до правильного слова.

В таблице 2 приведены результаты, полученные системой распознавания, использующей решающее правило (1) и метод простого голосования. В последнем случае дополнительный выход базовых классификаторов игнорировался.

Как можно видеть, правило (1) в данном случае дало практически одинаковое число распознанных слов, но меньшее число ошибок для тех слов, которые не были точно распознаны. Эксперименты с изменением числа образцов для класса «не буква» показали аналогичные результаты.

Фрагменты букв

Если рассматривать ошибки сегментации в качестве источника появления изображений, не являющихся буквами, то можно сделать вывод, что чаще всего в качестве таких изображений будут вы-

Таблица 3. Результаты тестирования метода фрагментов букв.

Правило	Распознано слов	Расстояние
Голосование	830 (92%)	1,2
Правило (1)	884 (98%)	1,8

ступать фрагменты изображений букв. В такой ситуации представляется естественным использовать в качестве обучающих примеров для класса «не буква» фрагменты изображений тех же букв, которые выступают в роли обычных обучающих примеров.

Таким образом, для создания обучающих примеров в этом случае изображения букв разрезались по вертикали. Для полученных изображений вычислялся вектор отклика признаков, который затем сравнивался с векторами, соответствующими корректным буквам. Слишком близкие вектора отбрасывались, чтобы исключить коллизии (например, из буквы «Т» таким образом несложно получить букву «Г»), а находящиеся на достаточном расстоянии помещались в обучающую последовательность классификатора.

Результаты тестирования. С использованием фрагментов букв нами было создано три набора по 10 базовых классификаторов. Каждый набор тестировался на распознавание изображений 300 слов, полученных аналогичным предыдущему способом. Суммарные результаты тестирования приведены в таблице 3. Результаты таблицы отражают распознавание 900 слов.

В данном случае использование фрагментов букв в качестве обучающих примеров для класса «не буква» совместно с решающим правилом (1) позволило улучшить результаты распознавания по сравнению с методом простого голосования в среднем на 6 процентов.

Иерархическая композиция

Нами также было проверено, как рассматриваемый метод ведёт себя в случае использования в иерархической композиции алгоритмов. Для этого был создан классификатор, на первом уровне которого использовалось по 10 базовых классификаторов, обученных распознавать шрифты Arial 14 pt и Times New Roman 14 pt. Результаты классификаторов, обученных распознавать одинаковый шрифт, комбинировались с помощью решающих правил, и попадали на второй уровень, где комбинировались с помощью другого решающего правила. Более подробно структура использованного многоуровневого классификатора описана в [3].

Проверялись возможности классификаторов, у которых в качестве решающих правил использовались простое голосование и правило (1). Классификатору предлагалось распознавать по триста

Таблица 4. Результаты тестирования двухуровневого классификатора.

1 уровень	2 уровень	Распознано слов
Голосование	Голосование	627 (69%)
Правило (1)	Голосование	744 (82%)
Голосование	Правило (1)	133 (15%)
Правило (1)	Правило (1)	4 ($< 1\%$)

слов написанных шрифтом Arial 14 pt. Было обучено и протестировано по три набора по 10 классификаторов; таким образом, общее число тестов равняется 900. При обучении классификаторов в качестве обучающих примеров для класса «не буква» использовались фрагменты букв.

Результаты тестирования приведены в таблице 4. Как можно видеть, правило (1) оказалось эффективным при применении на нижнем уровне классификатора, но совершенно непригодным для второго.

Выводы

Введение в классификатор дополнительного выхода, соответствующего классу некорректных объектов, позволяет улучшить работу составных классификаторов. При этом для генерации обучающих примеров целесообразно использовать примеры, которые могут возникать в процессе работы системы с большой вероятностью, такие как фрагменты букв при работе системы распознавания текста.

Для комбинирования простых классификаторов может быть использовано правило (1).

При создании иерархических композиций классификаторов правило (1) может быть рекомендовано для использования только на первом уровне классификатора. Использование информации о принадлежности классифицируемого объекта к классу «некорректный вход» на более высоких уровнях остаётся открытым вопросом.

Литература

[1] *Арлазаров В. Л., Куратов П. А., Славин О. А.* Распознавание строк печатных текстов // Сб. трудов ИСА РАН «Методы и средства работы с докуметами», Москва: Эдиториал УРСС, 2000. — С. 31–51.

[2] *Багрова И. А., Грицай А. А., Сорокин С. В., Пономарёв С. А., Сытник Д. А.,* Выбор признаков для распознавания печатных кириллических символов // Вестник Тверского государственного университета. Серия: Прикладная математика. — 2010. — № 18. — С. 59–72.

[3] *Багрова И. А., Сорокин С. В., Пономарёв С. А., Сытник Д. А.,* Комбинирование классификаторов на основе теории нечётких множеств // Программные продукты и системы. — 2010. — № 4 (92). — С. 112–117.

Метод распознавания размытых штрихкодов на мобильных устройствах без автофокусировки

Цымбал Д. А., Чепурной К. В.
Dmitry.Tsymbal@novsu.ru
Великий Новгород, Новгородский Государственный Университет им. Ярослава Мудрого

Предлагается метод распознавания одномерных продуктовых штрихкодов стандартов UPC-A и EAN-13, который может быть использован на мобильных устройствах с камерой без автофокуса. Метод позволяет обнаружить границы штрихкода на изображении и производить распознавание штрихкода, которое основано на сравнении данных с камеры и данных, получаемых путем размытия эталонных данных.

Традиционные методы распознавания штрихкодов основаны на поиске границ между штрихами единичной длины [1, 2]. Такой подход не работает в том случае, когда изображение с камеры получается размытым из-за расфокусировки. Во многих мобильных устройствах линзы сфокусированы на даль, поэтому изображения, полученные с камеры, находящейся на расстоянии 10–15 см, оказываются нерезкими (например, рис. 4, 8). Примером такого устройства служит iPhone 3G от компании Apple.

Стандарты UPC-A и EAN-13

Оба стандарта широко используются в настоящее время для упорядочивания и отслеживания товара в магазинах. Первый используется в США, а второй в странах Европы. Первый появился раньше и является предшественником EAN-13. Коды UPC легко преобразуются в EAN-13 (но не наоборот). Основное отличие этих кодов по внутренней организации — механизм вычисления тринадцатой цифры и почти несущественное изменение в расчёте контрольного числа с учётом этой 13-й цифры.

Штрихкод EAN-13 состоит из 30 вертикальных параллельных темных штрихов, разделенных пробелами (белыми штрихами). Каждый штрих может иметь ширину от 1 до 4 единиц. Общая ширина для одной цифры штрихкода всегда составляет 7 единиц. Битовая комбинация для каждой цифры разработана таким образом, чтобы цифры, насколько это возможно, отличались друг от друга. Общая ширина всего кода всегда равна 95 единицам. В любом коде 29 светлых и 30 тёмных штрихов. Последняя цифра — контрольное число, служит для выявления возможной ошибки при чтении кода сканером или ручного ввода цифр кода с клавиатуры. Также штрихкод содержит сторожевые паттерны, которые размечают границы блоков: левый и правый образцы представляют собой трёхбитный код 101, центральный образец — пятибитный 01010. Общая ширина всего кода равна 95 битам ($12\times7 = 84$ информационных бита и $3+3+5 = 11$ бит в паттернах). Все эти технические решения очень важны для надёжности и простоты сканирования этого кода.

Используется несколько вариантов кодирования цифр:

Цифра	L-код	R-код	G-код
0	0001101	1110010	0100111
1	0011001	1100110	0110011
2	0010011	1101100	0011011
3	0111101	1000010	0100001
4	0100011	1011100	0011101
5	0110001	1001110	0111001
6	0101111	1010000	0000101
7	0111011	1000100	0010001
8	0110111	1001000	0001001
9	0001011	1110100	0010111

Графические отличия L-кода, R-кода и G-кода состоят в следующем. Для каждой цифры это одна и та же комбинация чёрно-белых штрихов, L-код отличается от R-кода лишь фотографически негативным исполнением, а G-код отличается от R-кода реверсивным (зеркальным) исполнением. Все это нужно, чтобы отличить UPC-A от EAN-13.

Схема обработки изображения

Если полагать изображение идеально чётким, то задача распознавания состоит из следующих этапов [1–4]:

1) первичное обнаружение границ штрихкода;
2) подготовка исходных данных (выделение одномерного сигнала);
3) определение размера одного кодового бита;
4) выделение 12 блоков, содержащих информационные биты (12 цифр);
5) нахождение наиболее соответствующего кода для каждого из 12 блоков;
6) проверка контрольной суммы.

Подготовка исходных данных

Первоначально изображение переводится в оттенки серого. Исходные данные для алгоритма распознавания представляют собой одномерный сигнал, построенный как усреднение сканирующих линий, проходящих параллельно на небольшом расстоянии друг от друга (несколько точек). Этот подход позволяет компенсировать ошибки, возникающие по причине шумов и искажений в одной из линий, и в то же время позволяет получать корректное усреднение при перспективных искажениях.

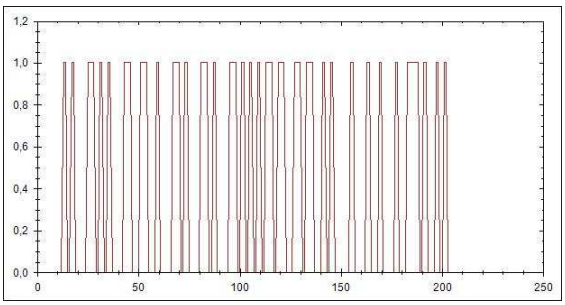


Рис. 1. Одномерный сигнал для идеального бинарного изображения штрихкода.

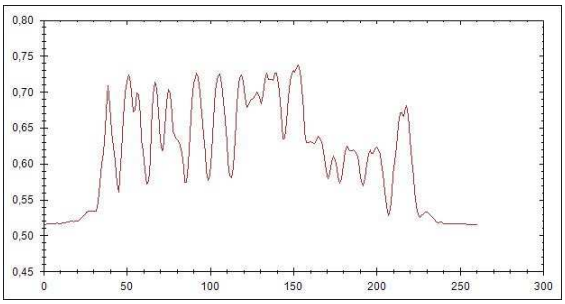


Рис. 2. Одномерный сигнал для идеального реального размытого изображения штрихкода.

Обнаружение границ

Для обнаружения границ штрихкода хорошо подходит широко известный метод выделения границ, предложенный Джоном Канни в 1986 году. В данной работе используются два первых вычислительных шага из данного метода — размытие исходного изображения фильтром Гаусса и применение оператора Собеля к размытой картинке. Размытие позволяет уменьшить влияние шума. Оператор Собеля позволяет найти градиент яркости в точке и направление его изменения. Определение границ начинается с верхней и нижней границы, затем идет обнаружение боковых границ в несколько этапов, рассматриваемых далее.

Размытие фильтром Гаусса. Функция Гаусса для двумерного случая:

$$f(x, y, \delta) = \frac{1}{2\pi\delta^2} \exp\left(-\frac{x^2 + y^2}{2\delta^2}\right).$$

Влияние пикселей друг на друга обратно пропорционально квадрату расстояния между ними. Степень размытия определяется параметром δ. Параметры размытия были подобраны эмпирически: $\delta = 2$, размер окна размывающего фильтра равен 7.

Оператор Собеля. Оператор Собеля позволяет вычислить приближенное значение градиента в каждой точке изображения. Результатом применения оператора является вектор градиента, либо его норма. Ядра фильтра Собеля:

6	5	4	3	2
7				1
8		x		0
9				15
10	11	12	13	14

Рис. 3. Направления градиента.

$$g_x = \begin{pmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{pmatrix}; \quad g_y = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix}.$$

Далее осуществляется свертка с использованием ядер. В дальнейших вычислениях нас будут интересовать как значение градиента в точке изображения, так и его направление. В качестве значения градиента берется доминирующее значение градиента в точке, т. е. $G = \max(Gx, Gy)$

Угол вектора квантуется по 22.5°, получается 16 возможных направлений градиента (рис. 3):

$$\theta = \texttt{round}\left(\tfrac{8}{\pi} \arctan \tfrac{Gy}{Gx}\right) \tfrac{\pi}{8} - \tfrac{\pi}{2}.$$

Ориентация штрихкода. Карта градиентов разбивается на блоки 8×8 пикселов. В каждом таком блоке находится доминирующее направление градиента путем подсчета количества пикселей, которые указывают в одно из первых восьми направлений (рис. 3). Противоположные направления подсчитываются, используя арифметику по модулю 8. Затем ищется прямоугольная зона со штрихкодом: строится некоторое количество линий параллельных направлению градиента и ищется зона где значения интенсивности меняются не значительно. Попутно с этим определяются границы штрихкода.

Поиск сторожевых паттернов. Метод Собеля даёт приблизительное положение границ. Поскольку границы штрихкода на расфокусированном изображении получаются нерезкими, ошибка в нахождении границ может достигать нескольких точек.

Уточнить положение границ можно, применив метод уточняющей аппроксимации. Для этого строится битовая последовательность, представляющая собой сторожевой паттерн и пустое («белое») поле значительной ширины (чтобы не перепутать сторожевой паттерн с такой же кодовой подпоследовательностью 0 101, следующей за ним). Таким образом, для левого паттерна происходит подбор наилучшей позиции для кода 000101, для правого — 101000. В качестве критерия наилучшего положения границы служит минимум ошибки при сравнении реальных данных с построенным эталоном паттерна.

Рис. 4. Исходное изображение.

Рис. 5. Карта градиентов.

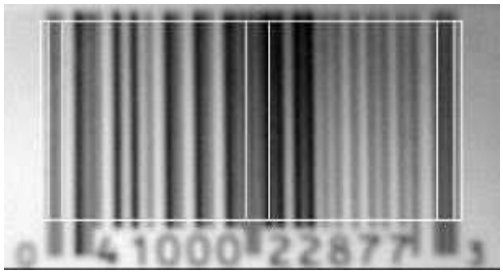


Рис. 6. Изображение штрих-кода с обнаруженными границами (помечены красным) и положением сторожевых шаблонов (помечены синим).

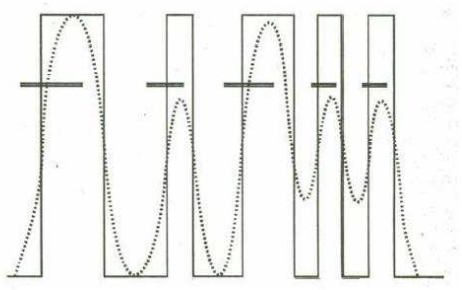


Рис. 7. Идеальный и нерезкий сигналы.

Модель размытия. В случае, когда штрихкод при сканировании находится не в фокусе, изображение получается размытым (нерезким). В этом случае сигнал можно представить следующим образом: $U = \texttt{Blur}(U) + n$, где $\texttt{Blur}(U)$ — функция расфокусировки, U — идеально резкий сигнал, n — шум.

Восстановление резкости такого изображения должно опираться на знание о природе размытия и представляет собой сложную вычислительную задачу (поиск оптимального исходного сигнала).

Общее число возможных кодовых комбинаций достигает 2^n (где n — ширина блока с одной цифрой, $\approx 14 - 16$), и даже при наличии хорошего критерия оптимизации задача поиска наилучшего остаётся весьма трудоёмкой. В то же время, количество корректных комбинаций на самом деле ограничено 10 для каждого блока.

Поэтому более рациональным подходом является применение изложенного выше метода аппроксимации эталонным сигналом, дополненного размывающим фильтром. То есть, реальные данные в каждом блоке сравниваются поочерёдно с десятью эталонами, пропущенными через фильтр. Для эмуляции размытия используется гауссово (нормальное) распределение, как наиболее естественный вид распределения случайных величин. Экспериментально проверено, что реальные данные оказываются весьма близко к эталону, полученному гауссовым размытием идеальных данных. Таким образом, дискретный фильтр Гаусса применяется к каждой точке одномерного сигнала (значение сигнала в точке обусловлено значениями точек в некоторой окрестности; значения накапливаются с соответствующими весами):

$$x'_i = \sum_{k=-r}^{r} \big(x_{i+k} G_{m,\sigma}(k)\big),$$

где $G_{m,\sigma}(x)$ — веса для фильтра, r — радиус окна фильтра. Веса вычисляются следующим образом:

$$G_{m,\sigma}(x) = \begin{cases} 0, & x < -r; \\ \Gamma_{m,\sigma}(x + \frac{1}{2}), & x = -r; \\ \Gamma_{m,\sigma}(x + \frac{1}{2}) - \Gamma_{m,\sigma}(x - \frac{1}{2}), & x \in (-r, r); \\ 1 - \Gamma_{m,\sigma}(x - \frac{1}{2}), & x = r; \\ 0, & x > r; \end{cases}$$

где $\Gamma_{m,\sigma}(x)$ — функция нормального распределения с математическим ожиданием m и дисперсией σ.

Для iPhone 3G были эмпирически подобраны оптимальные параметры фильтра Гаусса. Ширина окна фильтра равна 7 ($r = 3$), $m = 0$, $\sigma = 2$.

Распознавание штрихкода

Вычисление границ блоков. Ширина бита. Как уже было сказано ранее, штрихкод UPC-A состоит из 95 информационных битов. Таким образом, ширина бита вычисляется как

$$\texttt{bitlen} = (\texttt{end} - \texttt{start})/95,$$

где `start` — начальная позиция штрихкода, `end` — конец штрихкода.

Ширина левого и правого образцов равна

$$\texttt{side\_pattern\_size} = 3 \cdot \texttt{bitlen},$$

Ширина центрального образца равна

$$\texttt{center\_pattern\_size} = 5 \cdot \texttt{bitlen}.$$

Ширина блока данных равна

$$\texttt{block\_size} = 7 \cdot \texttt{bitlen}.$$

Таким образом, шесть левых блоков имеют следующие границы ($i = 0, \ldots, 5$):

$$\texttt{block\_start}[i] = \texttt{left\_start} + i \cdot \texttt{block\_size},$$
$$\texttt{block\_stop}[i] = \texttt{block\_start}[i] + \texttt{block\_size},$$

где $\texttt{left\_start} = \texttt{start} + \texttt{side\_pattern\_size}$.

Шесть правых блоков имеют границы:

$$\texttt{block\_start}[i] = \texttt{right\_start} + i \cdot \texttt{block\_size},$$
$$\texttt{block\_stop}[i] = \texttt{block\_start}[i] + \texttt{block\_size},$$

где $\texttt{right\_start} = \texttt{start} + \texttt{side\_pattern\_size} + \texttt{center\_pattern\_size} + 6 \cdot \texttt{block\_size} = \texttt{start} + 50 \cdot \texttt{bitlen}$, $i = 0, \ldots, 5$.

Построение и размытие исходных сигналов. После того, как вычислены границы блоков, для каждого из них производится аппроксимация сигнала эталонными значениями.

Предварительно вычисляются оценочные амплитуды сигнала: максимальная и минимальная. Максимум и минимум выбираются, исходя из характеристик сигнала.

Для этого отбирается несколько значений для максимальной амплитуды сигнала:

- абсолютный максимум (255);
- 90% от наблюдаемого максимума;
- максимальное значение сигнала за выбросом 10% (т. е. отбраковываются 10% наибольших значений, и среди оставшихся берётся максимум); выбирается то из значений, которое даёт меньшую погрешность при аппроксимации сторожевых паттернов.

Затем для каждого из 12 блоков производится ряд итераций сравнения с эталонным сигналом (по числу возможных кодовых комбинаций, для UPC-A это всегда 10 вариантов).

Каждая итерация состоит из следующих шагов:

1) построение очередного эталонного сигнала, масштабируемого по ширине блока;
2) размытие эталонного сигнала фильтром Гаусса;
3) расчёт ошибки для текущей аппроксимации относительно реальных данных.

На первом этапе эталонная кодовая последовательность (7 бит) масштабируется по ширине блока и нормируется по диапазону амплитуд реального сигнала:

$$\texttt{curr}_j[x] = \texttt{Min} + \Delta \cdot \texttt{code}_j \left[\frac{(x - 7 \cdot \texttt{block\_start}[i])}{\texttt{block\_size}} \right],$$

где $\Delta = \texttt{Max} - \texttt{Min}$, $\texttt{Max}$ — максимальная амплитуда, $\texttt{Min}$ — минимальная амплитуда сигнала; $\texttt{code}_j$ — 7-битная кодовая комбинация для цифры j, $x \in \in \big[\texttt{block\_start}[i]; \texttt{block\_stop}[i]\big]$.

На втором этапе фильтр Гаусса применяется к полученным эталонным данным в диапазоне $x \in \in \big[\texttt{block\_start}[i]; \texttt{block\_stop}[i]\big]$:

$$\texttt{curr}_j[x] = \sum_{k=-r}^{r} \big(\texttt{curr}'_j[x+k] \cdot G_{m,\sigma}(k)\big),$$

где $\texttt{curr}'$ — масштабированный эталонный сигнал, нормированный по амплитудам.

На третьем этапе вычисляется ошибка данной аппроксимации, опираясь на критерий наилучшего варианта, изложенный далее.

Критерий поиска наилучшего варианта. Наилучшим вариантом для i-го блока считается вариант, дающий минимальное значение средневзвешенной квадратичной ошибки на интервале $\big(\texttt{block\_start}[i]; \texttt{block\_stop}[i]\big)$, таким образом, $\texttt{best\_code}[i] = k$, где

$$\texttt{diff}[i,k] = \min_j(\texttt{diff}[i,j]), \quad j = 0, \ldots, 9;$$
$$\texttt{diff}[i,j] = \sum_{x=\texttt{block\_start}[i]}^{\texttt{block\_stop}[i]} \big(\texttt{data}[x] - \texttt{curr}_j[x]\big)^2.$$

Таким образом, на i-м блоке вычисляется ошибка $\texttt{diff}[i,j]$ для всех значений j (цифры от 0 до 9). В качестве наилучшего варианта выбирается то j, для которого ошибка минимальна.

Уточнение границ через итеративную аппроксимацию. Уточнение положения границ штрихкода повышает качество распознавания и уменьшает количество ошибок.

Алгоритм повторно уточняет границы штрихкода, производя сдвиг границ в радиусе 2 точек. Таким образом, сдвиг составляет от -2 до 2 относительно текущего положения; на данный момент используется шаг 0,5 точки.

Для каждого из положений границы выполняется аппроксимация штрихкода идеальным сигналом: при поиске левой границы аппроксимируются первые шесть кодовых блоков, при поиске правой — вторые шесть (правая часть).

Границы сдвигаются поочерёдно: сначала производится поиск наилучшего положения для левой

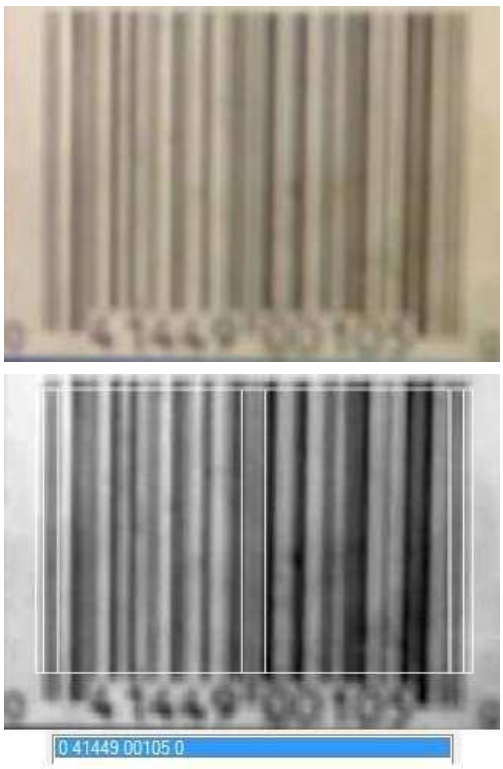


Рис. 8. Расфокусированное изображение UPC-A на iPhone 3G и результат работы алгоритма.

границы, затем для правой. После того, как лучшее положение левой границы найдено, оно устанавливается в качестве текущего, и поиск правой границы производится с учётом найденной левой.

После каждого сдвига границ происходит оценка качества аппроксимации. При сдвиге левой границы оценка выполняется по левой группе данных (первые 6 цифр), при сдвиге правой — по правой группе (цифры с 7 по 12).

Затем выполняется финальная итерация аппроксимации штрихкода, выполняемая для найденных наилучших границ. Целью является получение наиболее достоверного приближения, от которого в дальнейшем отталкивается эвристика для поиска корректного штрихкода.

Заключение

Результатом применения алгоритма к входному изображению (рис. 8) является информация о положении границ баркода и сторожевых шаблонов, а также последовательность цифр в привычном десятичном формате.

Точность метода в значительной мере зависит от точности обнаружения границ баркода. В текущем варианте алгоритм позволяет распознавать штрихкоды размерами от 2 см в ширину на изображениях, получаемых с камеры iPhone 3G или iPad 2 без автофокуса.

Разработанный алгоритм ориентирован на чтение достаточно чистых, равномерно освещённых баркодов и не учитывает возможных искажений баркода (замятие, загрязнение, цилиндрические искажения и т. п.). В то же время, алгоритм способен распознавать штрихкоды с небольшими цилиндрическими искажениями. Перспективные искажения также не влияют на работу метода.

Для распознавания штрихкодов на устройствах с автофокусировкой камеры достаточно не использовать гауссово размытие эталонного сигнала, компенсирующее подобное размытие реальных данных. Для этого требуется реализовать переключатель между режимами работы с автофокусом и без него.

Для каждого отдельного устройств без автофокуса может потребоваться подстройка весов и радиуса фильтра Гаусса, поскольку параметры размытия зависят от характеристик конкретной оптики.

Литература

[1] *Adelmann R., Langheinrich M., Florkemeier C.*. Toolkit for Bar Code Recognition and Resolving on Camera Phones — Jump Starting the Internet of Things // Institute for Pervasive Computing, ETH Zurich, 2006.

[2] *Wang K., Zou Y., Wang H.* 1D Bar Code Reading on Camera Phones. Int'l. Journal of Image and Graphics, 2007. — Vol. 7, No. 3. — Pp. 529–550.

[3] *Красильников Н. Н.* Цифровая обработка 2D- и 3D-изображений. — СПб.: БХВ-Петербург, 2011. — 608 с.

[4] *Canny J. F.* A computational approach to edge detection // IEEE Trans. Pattern Analysis and Machine Intelligence, 1986. — Vol. 8, No. 6. — Pp. 679–698.

Параметризация цветового представления изображения пламени с использованием одноклассового классификатора*

Ларин А. О., Середин О. С.
ekzebox@gmail.com, oseredin@yandex.ru
Тула, Тульский государственный университет

Данная работа посвящена задаче поиска огня на изображении. Основная идея работы заключается в применении одноклассового классификатора Тэкса для параметризации цветового представления изображения пламени. Эффективность выбранного подхода подтверждается экспериментальными исследованиями, в ходе которых было изучено влияние настраиваемого параметра классификатора на результаты параметризации. Также в данной работе рассмотрен подход к оптимизации решающего правила классификатора, позволяющий существенно увеличить скорость распознавания.

Введение

Поиск огня на изображениях (видеопотоках с камер) — довольно сложная задача компьютерного анализа данных, решение которой, как правило, выполняется в несколько этапов. Одним из них является попиксельный анализ изображения с целью поиска правила цветового описания пикселей, которые потенциально могут принадлежать изображению пламени. Данная работа посвящена рассмотрению метода решения именно этой подзадачи, которую мы будем называть «параметризацией пикселей пламени». Суть параметризации заключается в построении некоторого правила, позволяющего с управляемой точностью отделять пиксели огня от всех остальных пикселей на изображении, опираясь только на их цветовое представление.

Пиксель изображения обычно задаётся совокупностью трёх цветовых компонент палитры RGB, и его можно понимать как трёхмерный вектор, каждая компонента которого совпадает со значением одной из составляющих цвета пикселя RGB (красный, зелёный или синий соответственно). Трёхмерное пространство, в котором по каждой из координатных осей отсчитывается значение определённой цветовой компоненты пикселя, условимся называть пространством RGB.

Основная гипотеза поиска пикселей, принадлежащих пламени, заключается в экспертном указании на изображении фрагментов (прямоугольников, полигонов, явного перечисления пикселей), описывающих пламя. Используя только эту информацию и принцип обучения по прецедентам, необходимо построить математическое описание правила, относящего пиксели к одному из двух классов: пламя — не пламя. Традиционно такая задача решается путём параметризации экспериментально полученного множества точек в трёхмерном пространстве RGB. На Рис. 1 представлен пример вполне произвольного указания множества пикселей, составляющих обучающую совокупность. Область обучающей совокупности, взятой

Рис. 1. Кадр, содержащий изображение огня. Прямоугольником выделена область для обучения.

на Рис. 1 образует множество векторов, которое можно представить в пространстве RGB, как показано на Рис. 2.

Широко используемым подходом к решению задачи параметризации является метод аппроксимации экспериментальных данных сферами, предложенный в [2], где рассмотрен алгоритм подбора параметров для некоторого количества вероятностных распределений Гаусса, наилучшим образом, с точки зрения некоторого критерия, описывающих исходные данные (массив пикселей, изначально представляющих изображение пламени). На основании полученных распределений и строятся сферы, описывающие исходные данные. Центром каждой аппроксимирующей сферы является точка, представленная математическим ожиданием соответствующего распределения, а радиусом — удвоенное среднее квадратичное отклонение того же распределения (см. Рис. 3). Пиксель, классифицируемый на основе этой модели, считается принадлежащим изображению огня в случае, если ле-

Работа выполнена при финансовой поддержке РФФИ, проект № 09-07-00394.

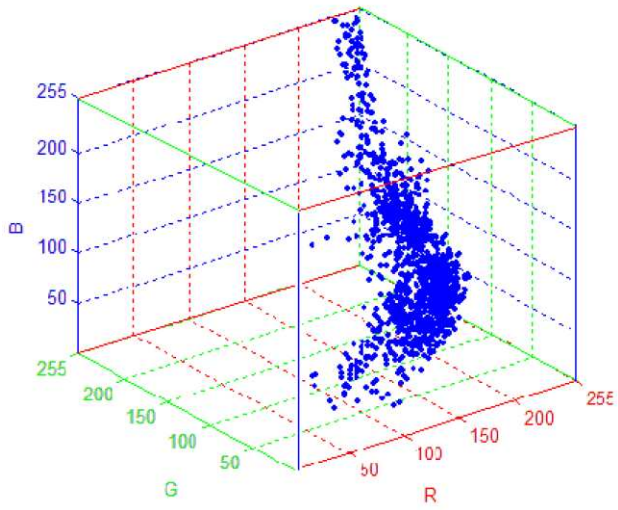


Рис. 2. Множество пикселей, представляющих изображение пламени, в пространстве RGB.

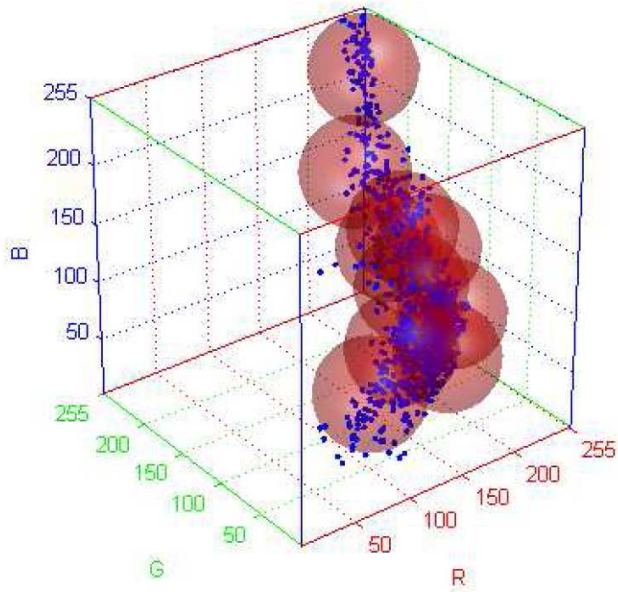


Рис. 3. Аппроксимация сферами исходного набора данных (центр сферы — среднее значение распределения, радиус — два с.к.о.).

жит внутри хотя бы одной из аппроксимирующих сфер.

Важно отметить, что количество сфер должно быть подобрано изначально. В самой работе сделано предположение о том, что для рассматриваемого там конкретного случая необходимо десять сфер, чтобы достаточно хорошо параметризовать пиксели, представляющие изображение пламени.

Если обратить внимание на вид исходного набора данных в пространстве RGB и постановку задачи, то можно предположить, что задачу параметризации можно решить, пользуясь методом одноклассовой классификации. Один из таких методов мы исследовали ранее, используя его в задаче идентификации личности по фотопортрету, где он показал неплохие результаты распознавания, поэтому возникла идея попытаться использовать одноклассовую классификацию для решения задачи параметризации пикселей пламени.

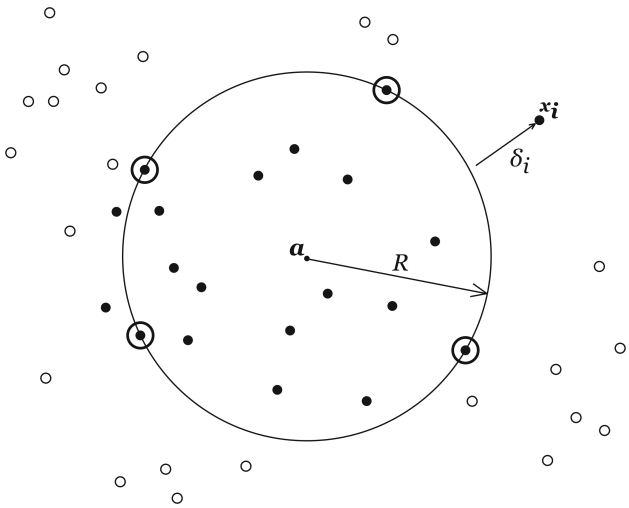


Рис. 4. Сферическая модель описания данных.

Метод описания данных опорными векторами

Метод решения одноклассовых задач распознавания образов, имеющий аналогию с методом опорных векторов В. Н. Вапника, был предложен Д. Тэксом в [1] и получил название Support Vector Data Description (метод описания данных опорными векторами).

Моделью описания набора данных $\boldsymbol{x_i} \in \mathbb{R}^n$, $i = 1, \ldots, N$ в этом методе является гиперсфера, представляющая ближайшую внешнюю границу вокруг данных. Основными параметрами, задающими гиперсферу, являются: центр $\boldsymbol{a} \in \mathbb{R}^n$ и радиус $R \in \mathbb{R}$.

Гиперсфера подбирается таким образом, чтобы её радиус был минимален, но при этом большая часть объектов обучающей совокупности не выходила за её пределы (см. Рис. 4). Объекты же, попадающие за границу гиперсферы, должны быть оштрафованы.

Таким образом, необходимо минимизировать структурную ошибку модели:

$$\begin{cases} R^2 + C\sum\limits_{i=1}^{N} \delta_i \to \min\limits_{R,\boldsymbol{a},\boldsymbol{\delta}}, \\ \|\boldsymbol{x_i} - \boldsymbol{a}\|^2 \leqslant R^2 + \delta_i,\ \delta_i \geqslant 0,\ i = 1, \ldots, N. \end{cases} \tag{1}$$

Двойственная задача по отношению к (1) имеет вид:

$$\begin{cases} \sum\limits_{i=1}^{N} \lambda_i \left(\boldsymbol{x_i} \cdot \boldsymbol{x_i}\right) - \sum\limits_{i=1}^{N} \sum\limits_{j=1}^{N} \lambda_i \lambda_j \left(\boldsymbol{x_i} \cdot \boldsymbol{x_j}\right) \to \min\limits_{\boldsymbol{\lambda}}, \\ \sum\limits_{i=1}^{N} \lambda_i = 1,\ 0 \leqslant \lambda_i \leqslant C,\ i = 1, \ldots, N, \end{cases} \tag{2}$$

где $\boldsymbol{\lambda}$ — множители Лагранжа.

Объекты, лежащие на границе гиперсферы (множители Лагранжа которых принадлежат интервалу $(0, C)$), Тэкс называет опорными, и именно эти объекты используются для описания всей обучающей совокупности.

Новый объект считается принадлежащим классу интереса, когда расстояние от него до центра гиперсферы меньше её радиуса. Из этого следует, что функция одноклассового решающего правила распознавания будет иметь вид индикаторной функции:

$$d\left(\boldsymbol{z}; \boldsymbol{\lambda}, R\right) = I\left(\left\|\boldsymbol{z} - \boldsymbol{a}\right\|^2 \leqslant R^2\right), \tag{3}$$

где

$$\left\|\boldsymbol{z} - \boldsymbol{a}\right\|^2 = \left(\boldsymbol{z} \cdot \boldsymbol{z}\right) - 2 \sum_{i=1}^{N_{SV}} \lambda_i \left(\boldsymbol{z} \cdot \boldsymbol{x_i}\right) + \\ + \sum_{i=1}^{N_{SV}} \sum_{j=1}^{N_{SV}} \lambda_i \lambda_j \left(\boldsymbol{x_i} \cdot \boldsymbol{x_j}\right), \tag{4}$$

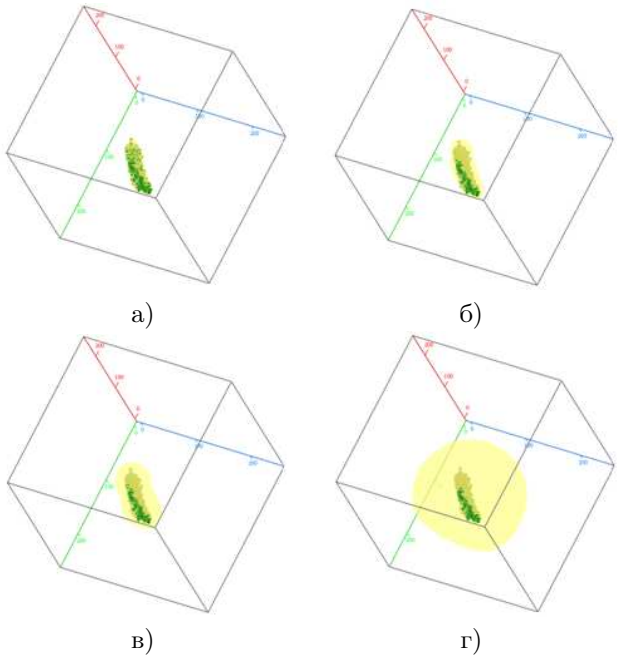
$$R^2 = \left(\boldsymbol{x_k} \cdot \boldsymbol{x_k}\right) - 2 \sum_{i=1}^{N_{SV}} \lambda_i \left(\boldsymbol{x_k} \cdot \boldsymbol{x_i}\right) + \\ + \sum_{i=1}^{N_{SV}} \sum_{j=1}^{N_{SV}} \lambda_i \lambda_j \left(\boldsymbol{x_i} \cdot \boldsymbol{x_j}\right), \tag{5}$$

где N_{SV} — количество опорных объектов, а $\boldsymbol{x_k}$ — любой опорный объект.

Для возможности описания данных более «гибкой формой», нежели сфера, Д. Тэкс использовал идею метода потенциальных функций [4, 5] для перехода в спрямляющее пространство признаков большей размерности. Наиболее часто используемыми потенциальными функциями являются полиномиальная и радиальная базисная функция Гаусса

$$K(\boldsymbol{x_i}, \boldsymbol{x_j}) = \exp\left(\frac{-\left\|\boldsymbol{x_i} - \boldsymbol{x_j}\right\|^2}{s^2}\right). \tag{6}$$

Таким образом, чтобы получить улучшенную модель описания данных по методу Тэкса, необходимо заменить в (2), (4) и (5) операцию вычисления скалярного произведения двух векторов вычислением значения потенциальной функции двух аргументов.



а) б) в) г)

Рис. 5. Результаты построения разделяющей границы с использованием классификатора Тэкса для различных параметров потенциальной функции Гаусса s: а) 10, б) 50, в) 100, г) 500.

Очевидно, что для решения задачи параметризации цветового представления пикселей пламени предпочтительной является потенциальная функция Гаусса, поэтому далее будем использовать именно её.

Результаты экспериментов

Для проверки модели описания данных опорными векторами был использован видеопоток, содержащий изображение пламени.

На Рис. 1 отмечена прямоугольная область, из которой был получен набор пикселей для обучения классификатора Тэкса. Результаты построения разделяющей границы вокруг данных на основании обученного классификатора представлены на Рис. 5. На Рис. 6 проиллюстрировано применение классификатора для поиска пламени на изображении.

Следует отметить, что при уменьшении параметра потенциальной функции Гаусса s область становится более «плотной» и ограничивает в пространстве RGB меньшее количество пикселей вокруг обучающего множества. Таким образом возникает необходимость в процедуре поиска оптимального параметра потенциальной функции.

Подбор оптимального параметра s для потенциальной функции Гаусса

В своей работе Тэкс также предложил алгоритм, позволяющий подбирать оптимальное значение параметра s функции Гаусса. Суть алгоритма заключается в поиске параметра s, соответству-

Рис. 6. Кадр, на котором посредством классификатора Тэкса, обученного по фрагменту, представленному на Рис. 1, были отмечены пиксели, принадлежащие пламени.

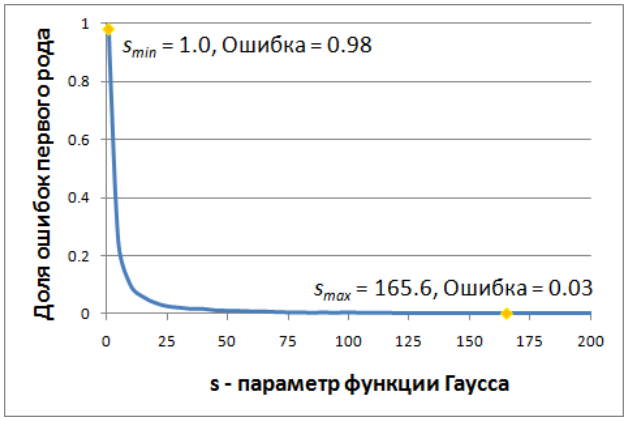


Рис. 7. Зависимость между параметром потенциальной функции и ошибкой классификации на скользящем контроле.

ющего заданной ошибке распознавания, в пределах от минимального значения параметра до максимального. Эти значения могут быть определены по следующим формулам:

$$s_{min} = \min_{i,j} \|\boldsymbol{x_i} - \boldsymbol{x_j}\|, \; i \neq j,$$

$$s_{max} = \max_{i,j} \|\boldsymbol{x_i} - \boldsymbol{x_j}\|.$$

Поиск параметра должен осуществляться по итерационной схеме с заданным шагом до тех пор, пока ошибка распознавания не будет соответствовать желаемой. Оценку ошибки распознавания было предложено делать, используя процедуру скользящего контроля. На Рис. 7 показана зависимость между ошибкой классификации и параметром потенциальной функции Гаусса.

Если обратить внимание на график зависимости между параметром потенциальной функции

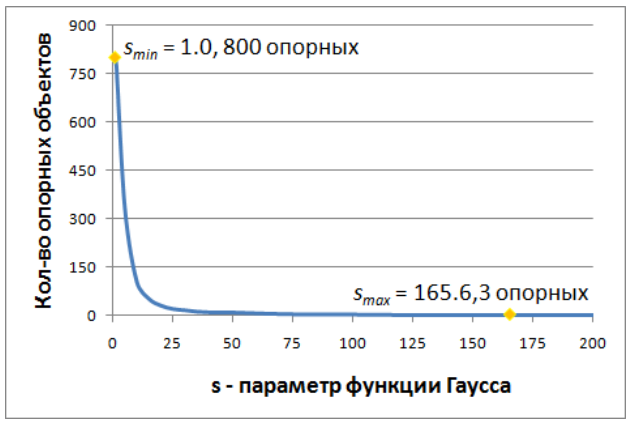


Рис. 8. Зависимость между параметром потенциальной функции и количеством опорных объектов в решающем правиле классификатора.

Гаусса и количеством опорных объектов, участвующих в решающем правиле распознавания (см. Рис. 8), то можно заметить, что количество опорных объектов уменьшается с увеличением параметра потенциальной функции. Таким образом количество опорных объектов в решающем правиле является косвенным признаком, по которому можно грубо оценивать будущую ошибку классификации.

Оптимизация скорости работы классификатора на этапе распознавания

Использование реализованного программно классификатора в системе анализа видеопотока, работающей в реальном времени, показало необходимость ускорения его работы на этапе распознавания. Это обусловлено тем, что процедуру распознавания необходимо запускать для каждого пикселя обрабатываемого изображения несколько раз в секунду.

Если обратить внимание на вид потенциальной функции Гаусса (6), то можно заметить, что она обладает двумя свойствами, следующими из свойств экспоненты:

1) принимает только положительные значения на всей своей области определения,
2) в случае равенства аргументов принимает значение, равное единице.

Воспользовавшись этими утверждениями, можно преобразовать выражение решающего правила (3), получив его упрощённый вид

$$d(\boldsymbol{z}) = I\left(\sum_{i=1}^{N_{SV}} \lambda_i K(\boldsymbol{z}, \boldsymbol{x_i}) - \sum_{i=1}^{N_{SV}} \lambda_i K(\boldsymbol{x_k}, \boldsymbol{x_i}) \geqslant 0\right),$$

где $\boldsymbol{x_k}$ — любой опорный вектор. Ввиду положительности гауссовской потенциальной функции достаточно частичного вычисления первой суммы в выражении (3), а вторая сумма может быть посчитана заранее перед процедурой классификации.

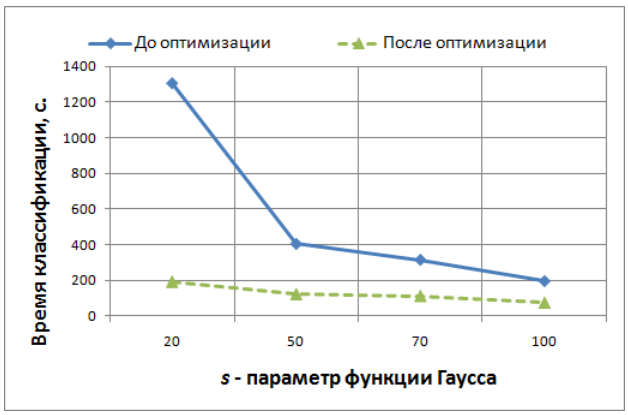


Рис. 9. Результаты оптимизации решающего правила.

Ясно, что для вычисления данного выражения, в сравнении с первоначальным, требуется меньшее число операций. Следовательно, временные затраты на процедуру классификации снизятся.

Хорошо известно, что операция вычисления экспоненты обладает большой вычислительной сложностью, поэтому мы предположили, что уменьшение вызовов этой функции во время выполнения может существенно снизить временные затраты на распознавание пикселей. В нашем случае вектор признаков представлен следующим образом:

$$\boldsymbol{x_i} = (r_i, g_i, b_i)\,. \tag{7}$$

Подставляя (7) в исходное выражение потенциальной функции Гаусса (6) и используя одно из свойств функции экспоненты, мы можем записать выражение для гауссовской потенциальной функции следующим образом:

$$K(\boldsymbol{x_i}, \boldsymbol{x_j}) = \exp\left(-\frac{|r_i - r_j|^2}{s^2}\right)\exp\left(-\frac{|g_i - g_j|^2}{s^2}\right)\cdot \\ \cdot\exp\left(-\frac{|b_i - b_j|^2}{s^2}\right). \tag{8}$$

Введем функцию $f(v)$ вида:

$$f(v) = \exp\left(-\frac{v^2}{s^2}\right) \tag{9}$$

и, выполняя замену в выражении (8), получим

$$K(\boldsymbol{x_i}, \boldsymbol{x_j}) = f(|r_i - r_j|)f(|g_i - g_j|)f(|b_i - b_j|).$$

Обратим внимание на то, что значение каждой цветовой компоненты RGB ограничено множеством $(0, 1, \ldots, 255)$, поэтому модуль разности двух любых цветовых компонент также будет принадлежать этому множеству. Отсюда следует, что область определения функции (9) будет дискретна и ограничена множеством $(0, 1, \ldots, 255)$, поэтому все значения этой функции могут быть вычислены ещё до начала основной процедуры классификации, что позволяет нам полностью отказаться от вычисления функции экспоненты, занимающего большую часть времени выполнения.

Таким образом, было существенно снижено число операций, необходимых для классификации, а также полностью исключено вычисление экспоненты во время выполнения. Сравнительные результаты можно увидеть на Рис. 9.

Эксперимент по измерению времени параметризации пикселей проводился следующим образом. Было выбрано 50 тыс. «своих» пикселей (принадлежащих изображению пламени), и 500 тыс. «чужих» (не принадлежащих изображению пламени). Для каждого объекта из выборки процедура распознавания запускалась 200 раз, чтобы снизить ошибку измерения времени.

Выводы

Метод описания данных опорными векторами (Support Vector Data Description) представляет собой довольно простую математическую модель, позволяющую решать задачу одноклассового распознавания образов. Нами было показано, что такая модель может быть применена к задаче параметризации пикселей пламени на изображении (см. Рис. 6).

Существенным достоинством модели Тэкса является то, что она не требует априорного знания количества сфер, необходимых для аппроксимации исходной выборки данных. Для параметров модели существуют эвристические алгоритмы определения их оптимальных значений, что даёт возможность строить модель описания данных без дополнительного вмешательства в процесс обучения: визуализации данных полученной модели и её дополнительного анализа.

Литература

[1] *Tax D.* One-class classification; Concept-learning in the absence of counterexamples // Ph.D thesis. Delft University of Technology, ASCI Dissertation Series. — 2001. — 146 p.

[2] *Töreyin B.* Fire detection algorithms using multimodal signal and image analysis // Ph.D thesis. — 2009. — 138 p.

[3] *Stauffer C., Grimson W.* Adaptive background mixture models for real-time tracking // Proceedings 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Cat No PR00149, 1999. — Vol. 2, Issue c, Publisher: IEEE Comput. Soc. — Pp. 246-252

[4] *Vapnik V.* Statistical Learning Theory. — New York: J. Wiley, 1998. — 768 p.

[5] *Айзерман М., Браверман Э., Розоноэр Л.* Метод потенциальных функций в теории обучения машин. — Москва: Наука, 1970. — 386 с.

Интерпретация сегментации по Мамфорду-Шаху*

Харинов М. В.
khar@iias.spb.su
Санкт-Петербург, СПИИРАН

Модель Мамфорда-Шаха интерпретируется как метод аппроксимации изображения оптимальными приближениями, которые вычисляются итеративным повторением ограниченного перебора вариантов. Иллюстрируется недостаточность алгоритма вычислений. Оцениваются и сравниваются результаты сегментации по 21 алгоритму разбиения изображения на последовательно возрастающее число сегментов. Обсуждается оптимизация программной реализации сегментации по Мамфорду-Шаху.

Модель Мамфорда-Шаха можно достаточно точно охарактеризовать как метод аппроксимации изображения его оптимальными, точнее, оптимизированными приближениями ступенчатой функцией, которые в ключевых работах [1–4] обозначаются термином «piecewise constant approximations».

В современных приложениях к самым разным изображениям модель обеспечивает «хорошую» сегментацию посредством «медленных» алгоритмов итеративного слияния смежных сегментов. При этом оптимизация выполнения многочисленных итераций по скорости связана с высокой трудоемкостью программирования.

Тем не менее, для того, чтобы запрограммировать и опробовать модель Мамфорда-Шаха на изображениях небольшого размера, квалифицированному программисту достаточно просто ознакомиться с критерием слияния сегментов в версии FLSA (Full λ-Sheduled Algorithm) модели [4–8]. Для развития вычислений могут оказаться полезными излагаемые в докладе дополнительные соображения по поводу использования модели Мамфорда-Шаха в современных программных реализациях.

Недостаточность итеративного слияния сегментов

Построим последовательность оптимальных приближений изображения из четырех пикселей, оценивая оптимальность по среднеквадратичному отклонению приближения от изображения (рис. 1).

На рис. 1 в левом верхнем углу показана матрица изображения. Остальные матрицы демонстрируют приближения изображения, которые состоят из сегментов, заполненных одинаковыми пикселями со средним значением яркости. Для обозначения сегментов границы между объединяемыми пикселями показаны пунктиром. Значения яркости пикселей выписаны в клетках. Под матрицами указаны значения *квадратичной ошибки* SE, равной квадрату среднеквадратичного отклонения StdDev, умноженному на число N пикселей в изображении.

Работа выполнена при финансовой поддержке РФФИ, проект №11-07-00685-а.

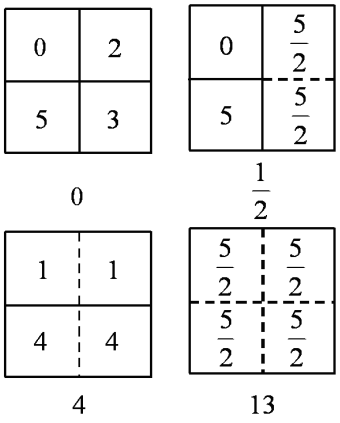


Рис. 1. *Оптимальные* приближения изображения из четырех пикселей четырьмя, тремя, двумя и одним вложенным изображением из одинаковых пикселей при минимальных значениях среднеквадратичного отклонения $\mathrm{StdDev} = \sqrt{\frac{\mathrm{SE}}{N}}$.

Определение 1. *Приближение считается оптимальным, если в сравнении с другими приближениями изображения с тем же самым числом сегментов минимально отличается от изображения по среднеквадратичному отклонению* StdDev *или квадратичной ошибке* SE.

Перебрав возможные варианты, нетрудно убедиться, что на рис. 1 показаны оптимальные приближения. Сравнивая на рис. 1 разбиения на два и на три сегмента, можно заметить, что сегменты из разных разбиений не вложены один в другой, а *перекрываются*. Тогда справедливо следующее.

Утверждение 1. *В общем случае итеративное слияние сегментов уже со второй итерации не обеспечивает получение $n = 1, \ldots, N$ оптимальных приближений изображения из N пикселей.*

Утверждение 2. *Слияние в последовательности оптимальных приближений перекрывающихся сегментов, порождает тривиальную иерархическую последовательность оптимизированных приближений, аппроксимирующих оптимальные приближения.*

При сегментации по Мамфорду-Шаху применяются менее прозрачные способы аппроксимации последовательности оптимальных приближений иерархической последовательностью оптимизированных приближений.

В любом случае оптимизированные приближения представляют интерес для формализации понятия иерархически организованных объектов, зрительно выделяемых на изображении.

Интерпретация вычислений

В модели Мамфорда-Шаха результатом сегментации считается последовательность оптимальных приближений изображения, для аппроксимации которой иерархической последовательностью приближений предусматривается способ итеративного слияния смежных сегментов по определенному алгоритму.

Приближения изображения описываются значениями определенного функционала, в общем случае более сложного, чем квадратичная ошибка SE. *Оптимальные* приближения отличаются от других приближений с тем же числом сегментов минимальными значениями функционала. Рассматривается последовательность оптимальных приближений цифрового изображения из N пикселей, разделяемого на $n = 1, \ldots, N$ *вложенных* в него изображений, которые задаются на связных сегментах.

Утверждение 3. *Приближения изображения из одного и N сегментов являются оптимальными.*

Способ аппроксимации последовательности оптимальных приближений заключается в том, что по оптимальному приближению из N сегментов посредством перебора всех пар смежных сегментов вычисляется оптимальное приближение из $N-1$ сегментов и соответствующее ему значение функционала. Затем изображение рассматривается как исходное, разбитое на $N-1$ суперпикселей, и вычисление очередного «оптимального», а точнее, оптимизированного приближения повторяется до тех пор, пока в качестве очередного приближения не будет вычислено приближение изображения из одного сегмента.

Утверждение 4. *По построению, каждый отличный от пикселя сегмент полученных приближений разделяется на два, и иерархия сегментов оказывается бинарной.*

Помимо тривиальных приближений изображения из одного и N сегментов, способ гарантирует точное вычисление единственного оптимального приближения из $N-1$ пикселей на первой итерации. Остальные последовательно вычисляемые приближения только *аппроксимируют* оптимальные приближения изображения, но не обязательно совпадают с ними. Характерно, что если число пикселей или неделимых *суперпикселей* в изображении превышает три, то в обсуждаемых реализациях модели не гарантируется точного вычисления оптимального приближения из двух сегментов по известному оптимальному приближению из трех сегментов.

Оптимальность приближения в модели Мамфорда-Шаха устанавливается по величине «энергетического» функционала, который в простейшем случае совпадает с квадратичной ошибкой, суммируемой по сегментам изображения. В этом случае минимизируется возрастание функционала при слиянии пары смежных сегментов 1 и 2, которое вычисляется по формуле:

$$\Delta\mathrm{SE}(1,2) \equiv (\Delta I(1,2))^2 \frac{S(1)S(2)}{S(1)+S(2)} = \min, \quad (1)$$

где $\Delta\mathrm{SE}(1,2) \equiv \mathrm{SE}(1\cup 2) - \mathrm{SE}(1) - \mathrm{SE}(2)$ — величина неаддитивной добавки к вычисленным для сегментов квадратичным ошибкам $\mathrm{SE}(1)$ и $\mathrm{SE}(2)$, $\Delta I(1,2)$ — разность средних значений яркости, а $S(1)$ и $S(2)$ — площади сегментов 1 и 2.

Выписанное соотношение задает *критерий слияния* сегментов, учитывающий при объединении сегментов только квадратичную ошибку. В характерных версиях модели Мамфорда-Шаха помимо квадратичной ошибки учитывается длина границ между смежными сегментами, и вместо $\Delta\mathrm{SE}(1,2)$ минимизируется частное от ее деления на длину $l(1,2)$ границы между смежными сегментами 1, 2: $\Delta\mathrm{SE}(1,2)/l(1,2) = \min$ в версии FLSA, или разности $\Delta\mathrm{SE}(1,2)$ и указаной длины с коэффициентом $\lambda \geqslant 0$: $\Delta\mathrm{SE}(1,2) - \lambda l(1,2) = \min$ в версии [3].

Функционалом в [3] служит суммарная квадратичная ошибка SE, к которой с коэффициентом λ добавляется сумма L границ между сегментами. В этом случае при каждом n минимизируется величина:

$$\mathrm{SE} + \lambda L = \min. \quad (2)$$

В практических расчетах значения λ используются как управляющие параметры, устанавливаемые для различных n.

Для исключения управляющих параметров в версии FLSA модели коэффициент λ, называемый «регуляризационным», при каждом n устанавливается так, чтобы при слиянии сегментов компенсировалось изменение функционала (2). При этом поведение λ в зависимости от n служит характеристикой приближений, полученных по критерию слияния сегментов модели FLSA.

Замечание 1. С введением λ, зависящего от n и от изображения, интерпретация в модели FLSA функционала (2) утрачивает очевидность, и результаты «минимизации» (2) не анализируются [5].

Замечание 2. При оценке произвольного приближения изображения посредством функционала (2) в изначальной интерпретации, значения λ полагаются не зависящими от изображения.

В [8] исследуются существенно более сложные критерии слияния с введением нескольких управляющих параметров, что оправдывается моделированием неформализованного зрительного восприятия посредством, так называемой, «перцептивной» или «подпороговой» (perceptual, preattentive) сегментации. Однако, наглядно демонстрируемая автоматическая сегментация изображения является не единственным способом верификации модели Мамфорда-Шаха. Модель позволяет при заданном n формально оценивать и сравнивать оптимизированные приближения по значениям функционалов (1), или (2), не ссылаясь на зрительное восприятие. К сожалению, о результатах подобной оценки часто можно судить лишь по косвенным данным. Так, в [4] обсуждается зависимость квадратичной ошибки SE от суммы границ между сегментами, являющейся монотонной функцией от числа сегментов n, но рассматривается только в контексте задачи автоматического останова сегментации и графически иллюстрируется в зависимости от «регуляризационного» коэффициента λ, который не очевиден для интерпретации.

Анализ StdDev

Для изучения эффекта аппроксимации последовательности оптимальных приближений иерархической последовательностью оптимизированных приближений мы рассчитали среднеквадратичное отклонение StdDev в зависимости от n для двадцати версий модели Мамфорда-Шаха, полученных без линейного комбинирования критериев слияния и сопутствующих управляющих параметров. Мы начали с программной реализации простейшей версии модели (1) без вычисления границ между сегментами, продолжили реализацией с вычислением границ согласно версии FLSA и опробовали ряд других критериев. Слияние сегментов выполнялось по критериям: $\Delta\mathrm{SE}(1,2) = \min$, $\Delta\mathrm{SE}(1,2)/l(1,2) = \min$, $\Delta\mathrm{SE}(1,2)/S(1 \cup 2) = \min$, $\mathrm{SE}(1 \cup 2)/l(1,2) = \min$, $\mathrm{SE}(1 \cup 2) = \min$, $\mathrm{StdDev}(1 \cup \cup 2) = \min$ и некоторым другим, с учетом контуров и прочих признаков сегментов. Для моделирования изменчивости изображения задавались также различные начальные разбиения изображения, получаемые на первых итерациях по одному критерию для дальнейшей обработки по другому критерию слияния сегментов.

Результаты, полученные для стандартного изображения «Лена», приводятся на рис. 2.

На рис. 2 показана зависимость среднеквадратичного отклонения от числа сегментов для двадцати одного алгоритма получения последовательности приближений изображения при различных критериях слияния смежных сегментов и начальных разбиениях изображения. Рис. 2 иллюстрирует поведение StdDev, отложенного по оси ординат, при изменении числа сегментов n от 1 до 1000, в логарифмическом масштабе отложенного по оси абсцисс. Графики на рис. 2 представляют собой монотонные невозрастающие кривые, которые расходятся из точки максимального значения StdDev на оси ординат и сходятся в точке на оси абсцисс, в которой StdDev обращается в ноль. Точка с максимальным значением StdDev описывает изображение, составленное из пикселей со средним по изображению значением яркости, а точка с нулевым значением StdDev отвечает разбиению изображения на сегменты из одинаковых пикселей. Отрезок прямой, соединяющей эти точки, обозначен пунктиром. Кривые, показанные сплошными черными линиями, отвечают критериям, перечисленным выше. Верхняя кривая, выполненная в сером цвете, описывает приближения, полученные для критерия близости, отождествляемого с абсолютной величиной разности средних значений яркости: $|\Delta I(1,2)| = \min$.

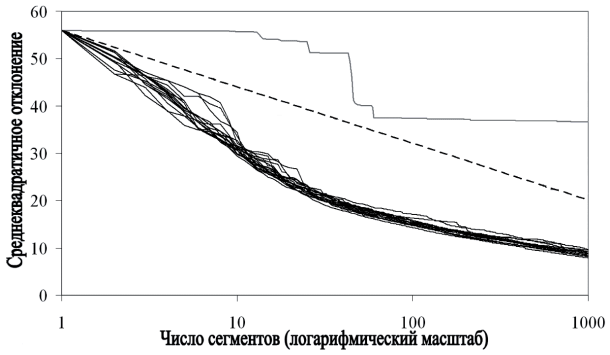


Рис. 2. Зависимость StdDev от n.

Кривая серого цвета иллюстрирует пример неудачного критерия близости смежных сегментов. На графике неудачный выбор критерия близости сегментов проявляется в наличии участков стабилизации среднеквадратичного отклонения при относительно малом числе сегментов, а на изображении при этом выделяются сегменты, отвечающие зрительно незначимым объектам.

Остальные кривые показывают поведение среднеквадратичного отклонения StdDev при более эффективном выборе критерия близости смежных сегментов. При этом с возрастанием числа сегментов наблюдается интенсивное падение StdDev, и кривые, переплетаясь между собой, заполняют некоторую полосу в окрестности нижней границы минимально возможных значений StdDev при каждом n. Если для численной оценки достоверности сегментации значения точек пунктирной прямой

на рис. 2 трактовать как пороговые, то ограничение на число сегментов при заданном среднеквадратичном отклонении выражается в виде:

$$n \leqslant N^{1-\frac{\mathrm{StdDev}}{\mathrm{StdDev}(1)}},$$

где N — число пикселей, а $\mathrm{StdDev}(1)$ — среднеквадратичное отклонение их значений от среднеарифметического, совпадающее с максимальным значением StdDev при $n = 1$.

Согласно предложенной оценке кривые на рис. 2 должны располагаться ниже прямой, обозначенной пунктиром, и для приближений из нескольких сегментов среднеквадратичное отклонение с ростом n должно спадать, по крайней мере, как гипербола.

Выводы

Результаты, демонстрируемые на рис. 2, согласуются с предположением, что иерархическая последовательность оптимизированных приближений изображения с некоторым допуском аппроксимирует последовательность оптимальных приближений из $n = 1, \ldots, N$ сегментов. Однако точного вычисления последовательности оптимальных приближений не достигается в силу того, что последняя не является иерархической. При этом проблема вычисления оптимальных приближений остается актуальной. Для повышения точности вычисления оптимальных приближений в модели Мамфорда-Шаха, вероятно, не следует ограничиваться алгоритмами итеративного слияния сегментов изображения.

По всей видимости, продолжает оставаться актуальной также задача повышения скорости вычислений, которая решается посредством кропотливого отслеживания модификации относительно небольших участков приближения изображения при многократном повторении поиска минимума критерия слияния [4–6]. Выработанные способы оптимизации вычислений, например, метод OLBM (Optimal Locally Best Merging) в [5], подсказывают менее трудоемкие в реализации алгоритмы синхронного слияния сегментов по всему полю изображения для повышения скорости за счет снижения числа итераций. Однако при этом возникает задача запоминания полученных иерархических приближений и их преобразования в приближения с последовательно возрастающим числом сегментов, как в исходной версии многократного поиска минимума критерия слияния. Имеются в виду *изотонные* преобразования одной иерархической последовательности в другую с сохранением порядка вложения сегментов, которые позволяют убедиться в эквивалентности результатов, полученных по ускоренным и исходным алгоритмам, а также оказываются полезными как практические приемы обработки изображений [9].

Наглядные результаты обработки на примере изображения «Лена» опубликованы в [9] и иллюстрируются в докладе на примере из Берклиевской базы сегментированных изображений [10].

Литература

[1] *Mumford D., Shah J.* Boundary detection by minimizing functionals, I // Proc. IEEE Comput. Vision Patt. Recogn. Conf. — San Francisco, 1985. — Pp. 22–26.

[2] *Mumford D., Shah J.* Optimal Approximations by Piecewise Smooth Functions and Associated Variational Problems // Communications on Pure and Applied Mathematics, 1989. — Vol. XLII, № 4. — Pp. 577–685.

[3] *Koepfler G. Lopez C. Morel J. M.* A Multiscale Algorithm for Image Segmentation by Variational Method // SIAM Journal on Numerical Analysis, 1994. — Vol. 31, № 1. — Pp. 282–299.

[4] *Redding N. J., Crisp D. J., Tang D. H., Newsam G. N.* An efficient algorithm for Mumford-Shah segmentation and its application to SAR imagery // Proc. Conf. Digital Image Computing Techniques and Applications (DICTA '99), 1999. — Pp. 35–41.

[5] *Crisp D. J., Tao T. C.* Fast Region Merging Algorithms for Image Segmentation // The 5th Asian Conf. on Computer Vision (ACCV2002), — Melbourne, Australia, 23–25 January 2002. — Pp. 1–6.

[6] *Robinson B. J., Redding N. J., Crisp D. J.* Implementation of a fast algorithm for segmenting SAR imagery // Scientific and Technical Report, Australia: Defense Science and Technology Organization, 01 January 2002. — 42 p.

[7] *Parker B. J., Feng D.* Graph-based Mumford-Shah segmentation of dynamic PET with application to input function estimation // IEEE Transactions on Nucl. Sci., 2005. — Vol. 52, № 1. — Pp. 79–89.

[8] *Marfil R., Sandoval F.* Energy-Based Perceptual Segmentation Using an Irregular Pyramid // LNCS Vol. 5517/2009 Bio-Inspired Systems: Computational and Ambient Intelligence. —Springer–Verlag: Berlin/Heidelberg, 2009. — Pp. 424–431.

[9] *Kharinov M. V.* Adaptive Hierarchical Image Segmentation Technique // Proc. of the 10-th. Int. Conf on Pattern Recognition and Image Analysis: New Information Technologies (PRIA–10–2010), 2010. — Vol. 1, — Pp. 205–208.

[10] *Martin D., Fowlkes C., Tal D., Malik J.* A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics // Proc. 8th Int. Conf. Computer Vision (ICCV), 2001. — Vol. 2, — Pp. 416–423. — `http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/`.

О потенциальной информационной достаточности выявления семантики контента

Цветков О. В., Зайцева А. А.
alexandr@iias.spb.su, cher@iias.spb.su
г. Санкт-Петербург, Учреждение Российской академии наук Санкт-Петербургский институт информатики и автоматизации РАН

В статье описывается подход к семантическому анализу изображений с точки зрения выявления информационной избыточности, приводятся примеры исследований. Показано сравнение энтропийного и семантического критерия информационной избыточности.

Решение задач распознавания изображений в целом сводится к выявлению, селекции семантически узнаваемых фрагментов на изображении. В основе решения этих задач используются методы коллажа и сегментации изображений [1]. Сегментация и фрагментация изображений нацелены на оптимизацию потенциальной информационной достаточности для селекции семантически узнаваемых областей. При видеонаблюдении (распознавание номерных знаков, лиц и др.) узнаваемые элементы имеют определенный размер в поле изображения, отсюда возникает оптимизационная задача минимизации информационного потока. Если стоит задача распознавания, визуализации глазом, то критерий качества восприятия должен учитывать эмоциональную окраску восприятия. В работе рассматриваются методы селекции фрагментов, семантически значимых для узнавания. Следовательно, понятие информационной избыточности принципиально связано с целевыми функциями распознавания и узнавания. Глаз человека, кроме четко определенной спектральной полосы видения, при идентификации и распознавании использует персонифицированные свойства предварительно накопленного знания, интеллекта [2]. «Сравнение в области распознавания изображений однозначно показывает, что компьютер с видеокамерой настолько еще далек от мозга с глазами, как скорость современных космических ракет — от скорости света. Вопрос, однако, состоит в том, насколько точно этот совершенный мозг отражает реальный мир в нашем сознании? Абсолютно точно или с искажениями? Мозг в состоянии скорректировать искажения органов чувств посредством объективных измерительных приборов. Но как быть с искажениями, которые возникают в самом мозге? Когда несколько человек наблюдают какое-нибудь расплывчатое, туманное изображение, то каждый узнает в нем что-то свое» [3].

Постановка задачи

Для решения проблемы нахождения уровня информационной избыточности с точки зрения селекции семантических и идентификации семантически значимых фрагментов были исследованы спутниковые изображения Земли (рис. 1 и 2), например, полученные российским спутником «Электро-Л» (НПО им. С. А. Лавочкина) и американским спутником NASA (опубликованные агентством Рейтер). В [4] описана разница между этим изображениями, заключающаяся в том, что «Электро-Л» делает снимки в трех диапазонах отраженного света: одном красном и двух близких к инфракрасному, что позволяет имитировать виды в стандартном RGB цветовом диапазоне. Аналогичные метеорологические спутники GOES, запущенные NASA, не используют близких к инфракрасным диапазонов, в результате захватывая изображения, более сходные с теми, которые видит глаз человека. Используя методы фрагментарной селекции, можно снижать полосу пропускания и считать эти изображения эквивалентными с точки зрения идентификации неразличимости семантических компонент при узнавании, например контрастных рельефов на спутниковых снимках. Необходимым условием является сохранение контуров береговой линии на изображении. Сенсорная система (рис. 1) фиксирует не только поверхностное отражение, но и захватывает более глубокий слой подводной поверхности. Автоматизированный критерий фрагментации, селекции замкнутых контуров предпочтителен на снимках, ориентированных на визуальное восприятие. Сенсорное восприятие по отраженным волнам несколько различно, а понятие «красоты» — чисто эмоциональный аспект восприятия.

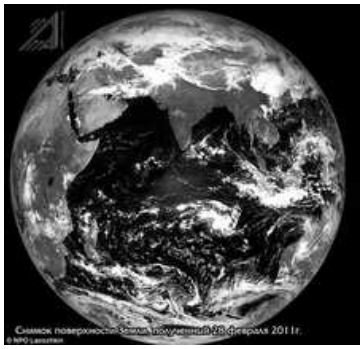


Рис. 1. Изображение Земли, полученное с российского спутника «Электро-Л»

Рис. 2. Изображение Земли, полученное с помощью спутника NASA, опубликованное агентством Reuters.

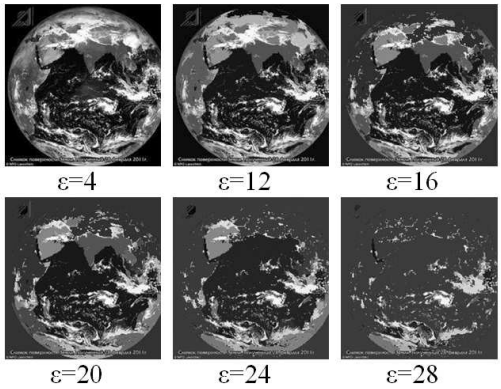


Рис. 3. Результаты сегментации изображения на рис. 1.

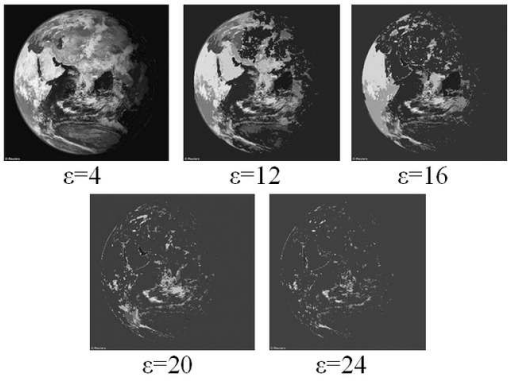


Рис. 4. Результаты сегментации изображения на рис. 2.

Экспериментальные результаты

В качестве исходного материала для эксперимента используются изображения на рис. 1 и 2, изначально отличающиеся по уровню избыточности (т. к. для получения первого использовался только видимый оптический диапазон, а для второго — расширенный оптический диапазон, включающий ИК-область), но искусственно представленных в цветовом пространстве RGB, соответствующем видимому глазом цветовому диапазону. Такое искусственное сужение цветового диапазона дает равное значение формальной информационной избыточности для обоих изображений [5]. В то же время использование расширенного диапазона способствует появлению дополнительных объектов на изображении, идентифицируемых только в этом расширенном диапазоне. Исходные изображения на рис. 1 и 2 последовательно сегментируются, в результате чего строится ряд последовательных слоев изображения (рис. 3, 4). Основным свойством сегментации является изменение разрешающей способности изображения. Процедура сегментации позволяет реализовать метод коллажа [6, 7], т. е. представление исходного изображения либо сегментами, либо фрагментами как совокупностью нескольких сегментов. При этом информационная достаточность определяется на слое, разрешающая способность которого поддерживается сегментацией и фрагментацией.

Предлагается следующий метод семантического анализа: строится график информационной достаточности, где ось абсцисс (горизонтальная) — разрешающая способность, ось ординат — MSE или его эквивалент $PSNR$ от исходного до i-го слоя. Идентификация семантически значимого слоя связана с пересечением послойных значений MSE. Для каждого слоя сегментации производится расчет MSE на каждом уровне сегментации по отношению к исходному (несегментированному) и межслойного $dMSE$ между соседними слоями, а также расчет энтропии $H(n)$ изображения, где n — уровень. В рамках данного исследования энтропия определяется как отношение размера изображения, сжатого некоторым эталонным алгоритмом компрессии без потерь (использовался алгоритм DEFLATE), к размеру исходного изображения на каждом уровне сегментации. Полученные в результате обработки графики приведены на рис. 5 и 6. Величина энтропийной оценки для обоих изображений уменьшается, что соответствует снижению уровня информационной избыточности. Информационная избыточность и достаточность — сугубо инженерные энергетические характеристики для измерения минимальной полосы пропускания [8].

В свою очередь, применение только энтропийной оценки не дает возможности оценить величину избыточности с точки зрения адекватного восприятия изображения глазом. Резкое увеличение энтропийной оценки на первом шаге сегментации связано с технологическими особенностями ее применения, вызывающими появление резких границ между объектами и увеличение количества объектов. По сути, использование сегментации является сменой формата представления изображения, что проявляется в резком увеличении энтропийного критерия. Семантический критерий — точка перегиба на графике $dMSE$ (рис. 5, 6) — изменение избыточности — энергетической составляющей между слоями. Информационная достаточность семантического слоя — минимум рас-

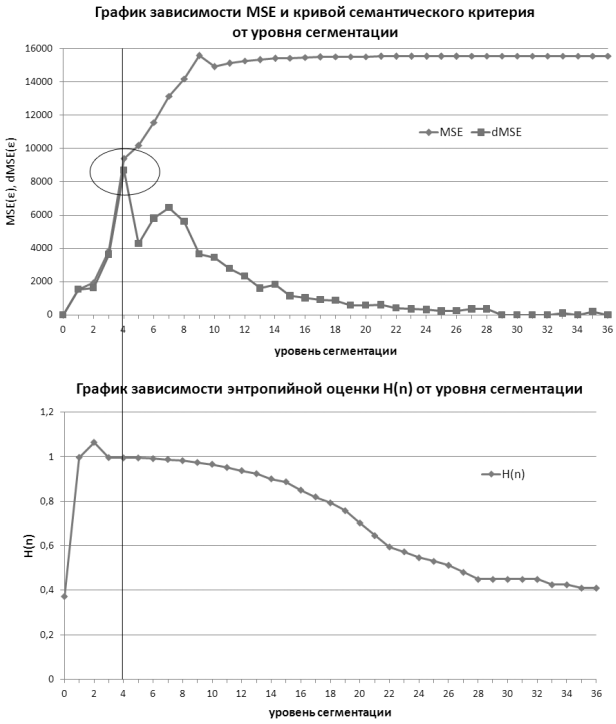


Рис. 5. Результаты обработки сегментированного изображения для рис. 1.

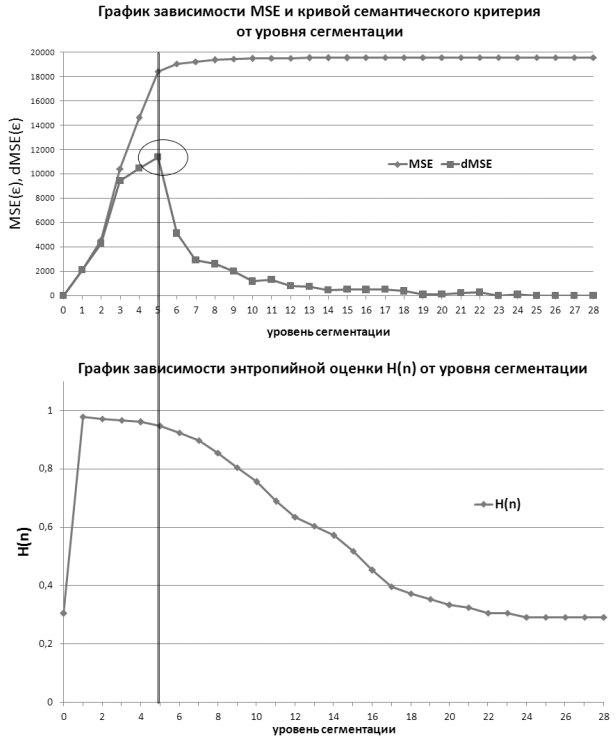


Рис. 6. Результаты обработки сегментированного изображения для рис. 2.

согласования MSE. Слой семантической информационной достаточности определяется минимумом разницы с предыдущим уровнем. Уровень перегиба — значение интегральной MSE на заданном уровне разрешения эквивалентен (сопоставляется) семантическому критерию на этом же уровне, в соответствии с принципом идентификации неразличимости. Информационная достаточность — величина, обратная информационной избыточности. Кроме того, изображение, соответствующее точке перегиба семантического критерия $dMSE$, можно считать характеристическим, соответствующим уровню минимальной информационной избыточности при необходимой информационной достаточности.

Выводы

В статье показан пример работы критерия селективного выделения замкнутых областей и семантической информации, которая не зависит от персонального эмоционального восприятия. Показано сравнение энтропийного и семантического критерия информационной избыточности на примере двух изображений с равным значением формальной информационной избыточности. При этом семантический критерий, в отличие от энтропийного, является чувствительным к семантической составляющей, адекватной визуальному восприятию.

Литература

[1] *Александров В. В., Кулешов С. В., Цветков О. В.* Цифровая технология инфокоммуникации. Передача, хранение и семантический анализ текста, звука, видео. — СПб.: Наука, 2008. — 244 с.

[2] *Александров В. В., Александрова В. В., Зайцева А. А.* Мир образов и мир воображений — от пиктограмм к компьютерной иконике // Труды 9-й Международной конференции «Прикладная оптика — 2010», СПб.: — Т. 3, 2010. — С. 309–312.

[3] *Эстерле О. В.* С какой точностью наш мозг отражает действительность? — 2000. — `http://n-t.ru`.

[4] `http://www.mk.ru` — Запад удивлен: русские снимки Земли лучше американских. — 2011.

[5] *Аксенов А. Ю., Зайцева А. А., Кулешов С. В.* О критерии адекватности цифровых трактов передачи данных // Информационно-измерительные и управляющие системы. — 2010. — Т. 8, № 7. — С. 75–77.

[6] *Кроновер Р. М.* Фракталы и хаос в динамических системах. Основы теории. — М.: Постмаркет, 2000. — 322 с.

[7] *Аксенов А. Ю., Зайцева А. А.* Применение программируемой технологии к обработке сигналов и изображений // Информационно-измерительные и управляющие системы. — 2009. — Т. 7, № 11. — С. 63–66.

[8] *Цветков О. В.* Некоторые граничные оценки информационной избыточности потока видеоданных для плоскостного и объемного телевидения // Информационно-измерительные и управляющие системы. — 2010. — Т. 8, № 11. — С. 5–10.

Ядра на основе интегральных вероятностных метрик для анализа текстурных изображений

Пластинин А. И.
anatoliy.plastinin@gmail.com
Самара, Самарский Государственный Аэрокосмический Университет

Данная работа посвящена разработке ядер для текстурных изображений. Мы предлагаем использовать интегральные вероятностные метрики в качестве меры сходства текстурных изображений, а также показываем, что полученные меры сходства являются отрицательно определёнными ядрами. Этот факт позволяет получить положительно определённые ядра, используя теорему о связи отрицательно и положительно определённых ядер. Проведены исследования предложенных ядер с использованием метода дискриминантного анализа и метода опорных векторов на тестовом наборе изображений.

Текстурный анализ широко распространен в обработке различных типов изображений [1]. Распространённым подходом к классификации текстурных изображений является их описание с помощью набора признаков (например, признаков Харалика) с последующей классификации в пространстве признаков.

Мы предлагаем использовать другой подход: определить функцию ядра на пространстве текстурных изображений, которая является скалярным произведением в некотором гильбертовом пространстве. Это позволяет использовать ряд методов использующих идею переход к ядру, например, метод опорных векторов.

Модель изображения

Областью определения D изображения будем называть конечное связное множество $D \subset \mathbb{Z}^2$. Наиболее часто используются прямоугольный участок двумерной целочисленной решётки $D = L_x \times L_y$, где $L_x = \{1, 2, \ldots, K_x\}$ и $L_y = \{1, 2, \ldots, K_y\}$ — пространственные координаты (соответственно, ширина и высота изображения). Под изображением будем понимать функцию $I\colon D \to G$, где G — множество цветов, например, в случае полутонового изображения $G = [0; 1]$, или для цветного RGB изображения $G = [0; 1]^3$.

Будем считать, что изображение является реализацией марковского случайного поля, т. е. выполнено условие:

$$\mathsf{P}(I_p | I_{D/p}) = \mathsf{P}(I_p | I_{\mathcal{N}_p}),$$

где I_p — значение отсчёта изображения I в пикселе p, $\mathcal{N}_p$ — окрестность пикселя p, $\mathcal{N}_p^k$ — k-ый элемент окрестности, $I_{\mathcal{N}_p} = \left(I_{\mathcal{N}_p^1}, \ldots, I_{\mathcal{N}_p^K}\right)$ — вектор значений отсчётов изображения в окрестности пикселя.

Анализ векторов окрестностей Каждому изображению будет соответствовать конечное множество векторов окрестностей:

$$\mathscr{N}(I) = \big\{(z_0, z_1, \ldots, z_n) | z_0 = I_p, (z_1, \ldots, z_n) = I_{\mathcal{N}_p} \quad p \in L_x \times L_y\big\}. \tag{1}$$

Следовательно, каждое изображение, являющиеся реализацией марковского случайного поля с распределением P, можно рассматривать как реализацию многомерной случайной величины $(z_0, z_1, \ldots, z_n)$, обладающей таким же распределением.

Гильбертовы пространства с воспроизводящим ядром

Гильбертовы пространства с воспроизводящим ядром играют важную роль в теории распознавания образов, так как они позволяют применять линейные методы в нелинейных случаях; ярким примером является метод опорных векторов [2].

Нелинейность достигается за счёт отображения исходного пространства объектов $\mathscr{X}$ в гильбертово пространство $\mathscr{H}$ более высокой размерности (возможно, бесконечной).

Ниже представлены основные положения о гильбертовых пространствах с воспроизводящим ядром [3]. Мы будем рассматривать класс функций (ядер) k, которые соответствуют скалярному произведению в некотором пространстве $\mathscr{H}$ согласно отображению Φ:

$$\Phi\colon \mathscr{X} \to \mathscr{H},$$

то есть

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle.$$

Будем рассматривать только действительно значимые ядра $k\colon \mathscr{X} \times \mathscr{X} \to \mathbb{R}$.

Определение 1 (Положительно определённое ядро). *Пусть $\mathscr{X}$ — непустое множество. Будем называть функцию $k\colon \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ положительно определённым ядром, если*

$$\sum_{i,j=1}^{m} c_i c_j k(x_i, x_j) \geqslant 0$$

для любых $m \in \mathbb{N}$, $x_1, \ldots, x_m \in \mathscr{X}$ и любых $c_1, \ldots, c_m \in \mathbb{R}$.

Определение 2 (Отрицательно определённое ядро). *Пусть $\mathscr{X}$ — непустое множество. Будем называть функцию $k\colon \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ отрицательно*

определённым ядром, если

$$\sum_{i,j=1}^{m} c_i c_j k(x_i, x_j) \leqslant 0$$

для любых $m \geqslant 2$, $x_1, \ldots, x_m \in \mathscr{X}$ *и любых* $c_1, \ldots, c_m \in \mathbb{R}$, *таких что* $\sum_{i=1}^{n} c_i = 0$.

Следует отметить, что для положительно определённого ядра всегда можно построить соответствующее ему гильбертово пространство, то есть построить вложение исходного пространства объектов в гильбертово пространство; детали такого построения можно найти в [2, 3].

Следующие теоремы показывают связь между отрицательно определёнными и положительно определёнными ядрами.

Теорема 1. *Пусть* $\mathscr{X}$ — *непустое множество и* $\psi\colon \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ *является ядром. Тогда* ψ *является отрицательно определённым тогда и только тогда, когда* $\exp(-\gamma\psi)$ *является положительно определённым ядром для любого* $\gamma > 0$.

Теорема 2. *Пусть справедливы условия Теоремы 1. Тогда* ψ^α *также является отрицательно определённым при* $0 < \alpha < 1$.

Теорема 3. *Пусть* $\psi\colon \mathscr{X} \times \mathscr{X} \to \mathbb{R}_+$ *является ядром, тогда* ψ *является отрицательно определённым тогда и только тогда, когда* $(a + \psi)^{-1}$ *(при* $a > 0$*) является положительно определённым.*

Теорема 4. *Пусть* $(\mathscr{X}, \mu)$ — *метрическое пространство.* μ^2 *является отрицательно определённым ядром тогда и только тогда, когда существует Гильбертово пространство* $\mathscr{H}$ *и отображение* $\Phi\colon \mathscr{X} \to \mathscr{H}$, *такое что*

$$\mu(x, x') = \|\Phi(x) - \Phi(x')\|, \quad \forall x, y \in \mathscr{X}.$$

Доказательство теорем 1–4 может быть найдено в [3] в главе 3.

Сравнение текстурных изображений

Рассмотрим другой вид меры расстояния в пространстве вероятностных распределений — интегральные вероятностные метрики [4, 5, 6] — который определяется как:

$$\gamma_{\mathscr{F}}(P, Q) = \sup_{f \in \mathscr{F}} \left| \int_M f\, dP - \int_M f\, dQ \right|,$$

где $\mathscr{F}$ — класс действительных ограниченных измеримых функций на M.

Следует отметить [4], что эта метрика получена из полунормы $\|\mu\|_{\mathscr{F}} = \sup\limits_{f \in \mathscr{F}} |\int f\, d\mu|$, откуда

$$\gamma_{\mathscr{F}}(P, Q) = \|P - Q\|_{\mathscr{F}}. \tag{2}$$

Выбирая соответствующим образом класс функций $\mathscr{F}$, можно получить различные известные меры расстояния:

— Метрика Дадли: $\mathscr{F} = \{f : \|f\|_{BL} \leqslant 1\}$, где $\|f\|_{BL} = \|f\|_\infty + \|f\|_L$, $\|f\|_\infty = \sup\{|f(x)|\colon x \in M\}$ и $\|f\|_L = \sup\{|f(x) - f(y)|/\rho(x, y)\colon x \neq y\}$.

— Метрика Канторовича и расстояние Васерштейна: если использовать $\mathscr{F} = \{f\colon \|f\|_L \leqslant 1\}$, то будет получена метрика Канторовича, которая двойственна расстоянию Васерштейна в случае если M сепарабельно.

— Метрика максимального среднего расхождения (Maximum Mean Discrepancy — MMD [7]): $\mathscr{F} = \{f\colon \|f\|_{\mathscr{H}} \leqslant 1\}$, где $\mathscr{H}$ является гильбертовым пространством с воспроизводящим ядром.

В работе [7] был предложен ряд методов для эмпирической оценки вероятностных метрик. Пусть $C_1 = \{c_1^1, \ldots, c_1^n\}$ и $C_2 = \{c_2^1, \ldots, c_2^m\}$ — независимые одинаково распределённые выборки из распределений $\mathbb{P}$ и $\mathbb{Q}$ соответственно. Тогда эмпирическая оценка $\gamma_{\mathscr{F}}(\mathbb{P}, \mathbb{Q})$ определяется как:

$$\gamma_{\mathscr{F}}(\mathbb{P}_n, \mathbb{Q}_m) = \sup_{f \in \mathscr{F}} \left| \sum_{i=1}^{m+n} y_i f(c^i) \right|,$$

где $\mathbb{P}_n$ и $\mathbb{Q}_m$ являются эмпирическими распределениями $\mathbb{P}$ и $\mathbb{Q}$ соответственно, и

$$y_i = \begin{cases} \frac{1}{n}, & 1 \leqslant i \leqslant n, \\ -\frac{1}{m}, & n < i \leqslant n + m, \end{cases}$$

$$c^i = \begin{cases} c_1^i, & 1 \leqslant i \leqslant n, \\ c_2^{i-n}, & n < i \leqslant n + m. \end{cases}$$

Оценка расстояния Васерштейна Следующая функция является решением для $\mathscr{F} = \mathscr{F}_W$:

$$W(\mathbb{P}_n, \mathbb{Q}_m) = \sum_{i=1}^{m+n} y_i a_i^\star,$$

где $a_i^\star$ является решением следующей задачи линейного программирования:

$$\sum_{i=1}^{m+n} y_i a_i \to \max_{a_1, \ldots, a_{m+n}},$$

$$-\rho(c^i, c^j) \leqslant a_i - a_j \leqslant \rho(c^i, c^j),\ \forall i, j.$$

Оценка метрики Дадли Следующая функция является решением для $\mathscr{F} = \mathscr{F}_\beta$:

$$\beta(\mathbb{P}_n, \mathbb{Q}_m) = \sum_{i=1}^{m+n} y_i a_i^\star$$

где $a_i^\star$ является решением следующей задачи линейного программирования:

$$\sum_{i=1}^{m+n} y_i a_i \to \max_{a_1, \ldots, a_{m+n}, b, c},$$

$$-b\rho(c^i, c^j) \leqslant a_i - a_j \leqslant b\rho(c^i, c^j),\ \forall i, j,$$
$$-c \leqslant a_i \leqslant c,\ \forall i,$$
$$b + c \leqslant 1.$$

Оценка максимального среднего расхождения Следующая функция является решением для $\mathscr{F} = \mathscr{F}_k$:

$$\gamma_k(\mathbb{P}_n, \mathbb{Q}_m) = \sqrt{\sum_{i,j=1}^{m+n} y_i y_j k(c^i, c^j)}.$$

Ядра на пространстве текстурных изображений

Функцию ядра для изображений $k\colon \mathfrak{I} \times \mathfrak{I} \to \mathbb{R}$ можно определить как

$$k(I, I') = \exp\left(-\gamma\mu\left(I, I'\right)\right) \tag{3}$$

где $\gamma > 0$.

Другой способ определить ядро на пространстве изображений:

$$k(I, I') = \left(a + b\mu\left(I, I'\right)\right)^{-1}, \tag{4}$$

где $a > 0$ и $b > 0$.

В ряде источников (например, в [8]) указывается необходимость масштабирования признаков (в случае пространства $\mathbb{R}^n$) в диапазон $[0, 1]$ для улучшения качества классификации. Однако, поскольку у нас нет такой возможности в случае пространства изображений, введем дополнительный масштабирующий коэффициент c_0 при метрике μ. Тогда ядро (3) можно записать в виде

$$k(I, I') = \exp\left(-\gamma c_0\mu\left(I, I'\right)\right), \tag{5}$$

где $\gamma > 0$, c_0 — константа, которая определяется по обучающему множеству, как

$$c_0 = \frac{1}{\max\limits_{i,j} \mu(I_i, I_j)}. \tag{6}$$

А ядро (4) можно представить как

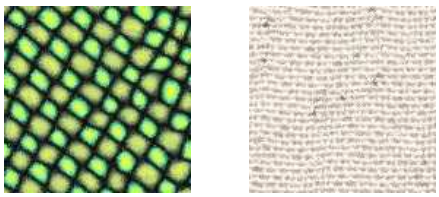
$$k(I, I') = \left(a + bc_0\mu\left(I, I'\right)\right)^{-1}, \tag{7}$$

где $a > 0$ и $b > 0$, а константа c_0 определяется так же, как в (6).

Утверждение 1. *Пусть μ является интегральной вероятностной метрикой, тогда ядра* (3) *и* (5) *являются положительно определёнными.*

Это непосредственно следует из (2) и теорем 4 и 1.

Утверждение 2. *Пусть μ является интегральной вероятностной метрикой, тогда ядра* (4) *и* (7) *являются положительно определёнными.*



(a) (b)

Рис. 1. Примеры тестовых изображений: (a) — 161; (b) — linen.

Это непосредственно следует из (2) и теорем 4 и 1.

Таким образом, пространство текстурных изображений может быть вложено в гильбертово пространство, а следовательно, на пространстве изображений может применяться любой метод, в основе которого используется ядра скалярного произведения, например, метод опорных векторов для классификации и регрессии.

Экспериментальные исследования

Исследования проводились тестовом наборе двух типов изображений, примеры показаны на рисунке 1.

Исследование с использованием дискриминатного анализа В работе [9] было предложено использовать метод перехода к ядру для дискриминантного анализа. Идея классического метода дискриминантного анализа заключается в определении такого линейного отображения исходного пространства в признаковое $\mathbb{R}^m \to \mathbb{R}^n, n < m$, которое максимизирует критерий линейной разделимости. В случае использования перехода к ядру ищется отображение $\mathscr{X} \to \mathbb{R}^n$, которое максимизирует критерий линейной разделимости для образов в пространстве $\mathscr{H}$ индуцированным некоторым ядром k.

Следует отметить, что в случае, если размерность пространства больше количества классов, K центров классов лежат в подпространстве размерности $\leqslant K - 1$.

Воспользуемся этим методом для анализа эффективности применения ядер для текстурных изображений, будем использовать «экспоненциальное» (5) и «обратное» (7) ядра.

Для каждого набора изображений будем строить отображение в пространство $\mathbb{R}^2$ (как было отмечено ранее, для случая двух классов достаточно строить отображение в $\mathbb{R}^1$). Из исходного множества изображений выбирается небольшое подмножество, по которому вычисляются проекторы, используемые для вычисления образов всего исходного множества.

На рисунке 2 представлен результат эксперимента с классами «161» и «linen» (полутоновый вариант изображений представленных на рисунке 1), использовалась метрика Васерштейна с подвыбор-

кой 100 случайных окрестностей, «экспоненциальное» ядро (5).

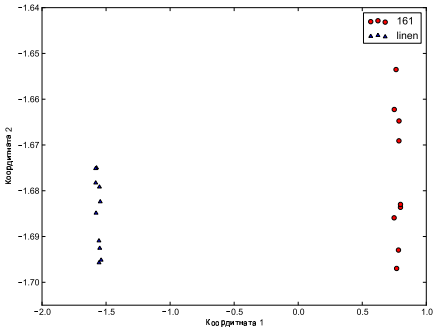


Рис. 2. Дискриминантный анализ изображений «161» и «linen» (метрика Васерштейна, «экспоненциальное» ядро (5)).

На рисунке 3 представлен результат эксперимента с классами «161» и «linen» (полутоновый вариант изображений представленных на рисунке 1), использовалась MMD метрика с подвыборкой 100 случайных окрестностей и $\gamma = 0.1$, «экспоненциальное» ядро (5). Видно, что в признаковом пространстве объекты разных классов являются линейно разделимыми.

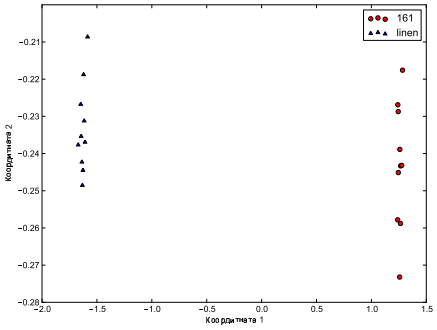


Рис. 3. Дискриминантный анализ изображений «161» и «linen» (MMD метрика $\gamma = 0.1$, «экспоненциальное» ядро (5)).

Исследование классификации по методу опорных векторов В экспериментах использовались различные метрики и ядра, для экспоненциального ядра для параметра γ использовались значения $\gamma = 10^{-3}, 10^{-2}, \ldots, 10^{2}$, аналогично для обратного ядра использовались значения $b = 10^{-3}, 10^{-2}, \ldots, 10^{2}$ при $a = 1$. Следует отметить, что во всех экспериментах вероятность верной классификации составило 100%. Что полностью согласуется с тем, что в дискриминантном анализе были получены линейно разделимые образы объектов.

Выводы

В этой работе мы представили два типа ядер на пространстве текстурных изображений на основе вероятностных метрик. Были рассмотрены вероятностные метрики на основе дивергенции, а также ряд интегральных вероятностных метрик, таких, как метрика Дадли, расстояние Васерштайна и максимальное среднее расхождение. Сперва мы доказали, что функция метрики произвольного метрического пространства является отрицательно определённым ядром, а затем доказали, что представленные ядра являются положительно определёнными. Этот факт позволяет вложить пространство текстурных изображений в гильбертово пространство индуцированное этим ядрами.

Таким образом, стало возможным использовать ряд методов распознавания образов. Сперва мы продемонстрировали возможность отображения пространства текстурных изображений в пространство $\mathbb{R}^2$, затем показали результаты решения задач классификации и регрессии на тестовых наборах изображений.

Результаты показывают, что ядра на основе мер схожести текстурных изображений могут эффективно применяться для решения различных задач анализа текстурных изображений.

Литература

[1] *Mirmehdi M., Xie X., Suri J.* Handbook of Texture Analysis. — London: Imperial College Press, 2008.

[2] *Scholkopf B., Smola A. J.* Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. — Cambridge, MA, USA: MIT Press, 2001.

[3] *Berg C., Christensen J. P. R., Ressel P.* Harmonic Analysis on Semigroups. — Berlin: Springer, 1984.

[4] *Muller A.* Integral probability metrics and their generating classes of functions // Advances in Applied Probability, 1997. — Vol. 29, No. 2. — Pp. 429–443.

[5] *Rachev S. T.* Probability Metrics and the Stability of Stochastic Models New York: John Wiley & Sons, 1991.

[6] *Золотарев В. М.* Вероятностные метрики // Теория вероятностей и ее применения. — 1983. — Т. 28, № 2. — С. 264–287.

[7] *Sriperumbudur B. K., Fukumizu K., Gretton A., Scholkopf B., Lanckriet G.* Non-parametric estimation of integral probability metrics // IEEE International Symposium on Information Theory, 2010. — Pp. 1428–1432.

[8] *Abe S.* Support Vector Machines for Pattern Classification, 2nd ed. — Springer Publishing Company, 2010.

[9] *Mika S., Ratsch G., Weston J., Scholkopf B., Mullers K. R.* Fisher discriminant analysis with kernels // Neural Networks for Signal Processing IX, IEEE, 1999. — Pp. 41–48.

Комбинированный подход к локализации записей на изображениях произведений живописи*

***Мурашов Д. М.*[1], *Березин А. В.*[2], *Иванова Е. Ю.*[2]**
d_murashov@mail.ru, berezin_aleks@mail.ru
[1]Москва, Вычислительный Центр им. А. А. Дородницына РАН
[2]Москва, Государственный Исторический музей

Предлагается подход к локализации областей с нарушенным авторским лако-красочным слоем на изображениях произведений живописи в разных спектральных диапазонах на основе морфологического и локально-адаптивного порогового детекторов, а также теоретико-информационной меры различия изображений.

Рассматривается задача анализа изображений произведений изобразительного искусства, полученных в разных спектральных диапазонах. Многоспектральные изображения широко применяются в музейной практике для целей реставрации и атрибуции. Важным аспектом исследований таких изображений является поиск невидимой для человеческого глаза, но важной для специалистов информации с использованием комбинирования изображений, зафиксированных в ультрафиолетовом (УФ), инфракрасном (ИК), рентгеновском и видимом спектральных диапазонах [1].

На цветных изображениях в УФ лучах области, где делалась запись и ретуширование, выглядят бурыми, а на полутоновых изображениях — тёмными пятнами (см. Рис. 1). Могут быть видны также подлаковые загрязнения. Для формирования задания на реставрационные работы необходимо выделить области с нарушенной авторской живописью на изображении в ультрафиолетовом диапазоне и обозначить контуры выделенных областей на цифровой фотографии в видимом спектральном диапазоне. В литературе имеется ряд

Рис. 1. Изображения фрагментов портрета в видимом и УФ спектральных диапазонах.

публикаций по методам поиска различий на сериях изображений для решения различных задач. В работе [2] представлен метод автоматизированного поиска скрытой информации по фотографии картины и её рентгенограмме. Выявление невидимых глазу объектов на рентгеновском снимке осуществляется сравнением описаний пары изображений, построенных с помощью двух типов иерархических моделей: моделей изображений и модели соответствия деталей изображений. На верхнем уровне изображения сегментируются на области, однородные по яркости. На нижнем уровне каждая из выделенных областей раскладывается на текстурную компоненту и компоненту среднего значения яркости. Анализируются перепады яркости на границах однородных областей, текстурные признаки. Соответствия между деталями изображений описываются линейной регрессионной моделью для текстурной составляющей и результатами сравнения компонент среднего значения яркости. Метод обладает высокой вычислительной сложностью. В работе [3] представлена программная система для сравнения изображений видимых и скрытых слоев живописи. Система предназначена для визуального анализа комбинированных ИК рефлектограмм и цветных цифровых фотографий. Комбинирование осуществляется в цветовом пространстве HSV заменой канала V на изображение рефлектограммы. В [6] предложен метод оценивания визуальных различий изображений на основе модели зрительной системы человека. Мерой различия в точке с заданными координатами является вероятность обнаружения несовпадений двух изображений в этой точке. На основе модификаций меры [6] в [7] вводятся меры визуальных различий для последовательностей многоспектральных изображений. Постановка задачи оценивания визуальных различий изображений не в полной мере соответствует решаемой в данной работе задаче, где требуется выделить объекты, видимые только на одном из предъявляемых изображений.

Исследуемые изображения обладают рядом особенностей, оказывающих влияние на решение задачи. Во-первых, неравномерное освещение при съемке. Во-вторых, объекты интереса — области реставрации и записи авторской живописи, области подлакового загрязнения — имеют различные яркостные профили и контрастность. В-третьих, разме-

Работа выполнена при финансовой поддержке РФФИ, проект № 09-07-00368.

ры объектов могут составлять от нескольких десятков пикселов до нескольких сотен и тысяч. Форма объектов может быть самой разнообразной. В-четвертых, искомые области существенно неоднородны по яркости. Таким образом, области повреждений и вмешательства в авторский красочный слой могут быть разделены на классы по характеру проявления на изображениях, и могут потребоваться разные методы для их локализации.

Для достижения цели работы будет использован следующий подход: (а) с помощью детекторов, ориентированных на определённые классы объектов, на одном из изображений, где проявляются искомые объекты, выделяются области интереса; (б) вводится мера различия изображений; (в) по величине меры различия из найденных на этапе (а) областей выбираются области, соответствующие решаемой задаче.

Детекторы областей интереса

При решении рассматриваемой задачи будут использованы модели изображений и детекторов, соответствующих проявлениям записей авторской живописи.

Пусть функции $U^k = u^k(x,y)$, $(x,y) \in X$, $u^k : X \mapsto \mathbb{Z}^+$, $k = 1,\ldots,K$, определены в некоторой области $X \subset \mathbb{Z}^2$ и описывают полутоновой рельеф на K изображениях сцены, зафиксированных в разных диапазонах спектра при отсутствии дефектов или результатов вмешательства в авторский красочный слой. Все анализируемые изображения предварительно совмещены, и скомпенсирована неравномерность освещённости при съемке. Пусть функции $\xi^k = v^k_{ij}(x,y)$, $\xi^k_{ij} : X \mapsto \mathbb{Z}$, $i = 1,\ldots,N_j$, $j = 1,\ldots,J$ описывают рельеф дефектной области D^k_{ij}, $D^k_{ij} \subset X$. Здесь i — номер дефекта, j — номер класса дефектов. Пусть отображение $\varphi_j : X \times \mathbb{Z} \mapsto \{0,1\}$ описывает детектор дефектов класса j,

$$\varphi_j(u^k(x,y)) = 0 \;\forall (x,y) \in X,$$

$$\varphi_j\Big(u^k(x,y) + \sum_i \sum_j \xi^k_{ij}\Big) = 1 \;\forall (x,y) \in D^k_{ij}.$$

Для поиска областей интереса применяются два детектора. Первый предназначен для выделения крупных объектов и основан на операциях геодезической реконструкции полутоновых изображений [5] и понятии «бассейна» яркостного рельефа глубины h. Детектор включает операции выделения областей, в которых яркость пикселов меньше яркости внешних относительно бассейна пикселов на величину не более, чем h, и получения бинарной маски выделенных областей.

Детектор 1. Бинарная маска формируется следующим образом:

$$M^U_1 = T(U_{bas} - U_{dom}), \tag{1}$$

где $T()$ — операция пороговой бинаризации, U_{dom} — изображение найденных ярких областей («куполов» яркостного рельефа) на U:

$$U_{dom} = U - R^\delta_U(U - g),$$

здесь $R^\delta_U(U-g)$ — результат операции реконструкции геодезической дилатацией маски U из маркера $U - g$, где g — относительная высота выделяемых «куполов». Бассейны с относительной глубиной h на изображении Uв ультрафиолетовом спектральном диапазоне находятся как

$$U_{bas} = R^\delta_U(U + h) - U, \tag{2}$$

где $R^\delta_U(U+h)$ — операция реконструкции геодезической дилатацией маски U из изображения-маркера $U + h$, которое получено из изображения U увеличением яркости на h. Операция поэлементного вычитания $U_{bas} - U_{dom}$ в (1) выполняется для повышения контрастности изображения U_{bas} и повышения точности пороговой бинаризации.

Второй детектор предназначен для выделения небольших фрагментов, отличающихся по уровню серого тона от окружающих областей, и основан на алгоритме локальной адаптивной пороговой бинаризации [4].

Детектор 2. Функция детектора строится следующим образом. Пусть $\bar{u}^k$ — среднее значение функции $u^k(x,y)$ в некоторой области $W \subset X$, $u_m < \bar{u}^k(x,y) < u_M$ для $(x,y) \in W$. Тогда функция детектора имеет вид:

$$\varphi(x,y) = \begin{cases} 1, & \text{если } u(x,y) \geqslant u_M; \\ 0, & \text{если } u(x,y) < u_M, \end{cases} \tag{3}$$

u_M задаётся в виде $u_M = \bar{u}^k(x,y) + q\sigma$, где σ — среднеквадратичное отклонение яркости, вычисленное в скользящем окне W, q — коэффициент. Изображение маски дефектов формируется в виде

$$M^U_2(x,u) = \varphi(x,y). \tag{4}$$

Полученные с помощью детекторов 1 и 2 ((1)–(4)) бинарные маски областей интереса на УФ изображении показаны на Рис. 2. Однако не все выделенные объекты на бинарной маске соответствуют искомым объектам. Необходимо отобрать только те объекты, которые видны на УФ изображении и не видны на фотографии в видимом диапазоне. Для выявления указанных выше различий необходимо ввести меру различия изображений.

Мера различия изображений

В ряде работ для оценивания сходства и различия изображений применяются теоретико-информационные подходы и методы [8]. Для решения задачи совмещения изображений предложена теоретико-информационная мера сходства модельного

Рис. 2. Изображения бинарных масок M_1^U и M_2^U областей интереса на УФ изображении.

и преобразованного входного изображений в виде значения взаимной информации, вычисленного на этих изображениях. В работе [10] введена мера различия изображений в виде суммы их условных энтропий $H(X \mid Y) + H(Y \mid X)$, где X и Y — случайные переменные, характеризующие значения яркости в пикселах изображений. В работе [9] для оценивания качественных показателей результата слияния последовательностей многоспектральных изображений использовались меры на основе взаимной информации и условной энтропии. Условная энтропия $H(X \mid Y)$ интерпретируется как средняя информация, которая требуется для того, чтобы определить X, если известно Y. Для использования теоретико-информационного подхода необходима вероятностная модель связи между изображениями. Пусть значения яркости на изображениях в видимом и ультрафиолетовом диапазонах в точке с координатами x, y описываются дискретными случайными переменными $U(x, y)$ и $V(x, y)$, со значениями u и v, квантованными на конечное число уровней L. Поскольку изображения U и V отображают одну и ту же сцену, то существует связь между переменными $U(x, y)$ и $V(x, y)$. Будет использоваться модель, аналогичная предложенной в [8]:

$$U(Tr(x, y)) = F(V(x, y)) + \eta(x, y), \tag{5}$$

где Tr — преобразование координат (для совмещённых изображений $U(Tr(x, y)) = U(x, y)$); F — функция преобразования яркости, моделирующая связь между двумя изображениями объекта в двух спектральных диапазонах; η — случайная переменная, моделирующая нарушения лакового и красочного слоев, видимые при ультрафиолетовом освещении. Модель (5) можно рассматривать как модель дискретной стохастической информационной системы со входом V и выходом U. Тогда по определению условной энтропии

$$H(U \mid V) = -\sum_{k=1}^{L}\sum_{l=1}^{L} p(u_k, v_l) \log(p(u_k, v_l)/p(v_l)); \tag{6}$$

$$H(V \mid U) = -\sum_{k=1}^{L}\sum_{l=1}^{L} p(u_k, v_l) \log(p(u_k, v_l)/p(u_k)), \tag{7}$$

где $p(u_k)$, $p(v_l)$, $p(u_k, v_l)$ — значения вероятностей появления значений v_l и u_k на входе и выходе системы и их совместной вероятности. В отличие от задачи совмещения изображений, в решаемой задаче требуется мера, позволяющая выделить объекты на изображении U, отсутствующие на V. Предлагается характеризовать отличия изображения U от V значениями условной энтропии $H(U \mid V)$, вычисленными в каждой точке (x, y) изображений U и V. Условия, при которых $H(U \mid V)$ выполняет функции меры отличия даёт следующее утверждение.

Утверждение 1. *Условная энтропия $H(U \mid V)$ является мерой отличия изображения U от изображения V в точках (x, y), в том и только том случае, если*

$$p(v_l) = p(u_k, v_l) \tag{8}$$

в точках (x, y), где нет отличий в содержании изображений и

$$p(v_l) > p(u_k, v_l);\ k = 1, \ldots, L;\ l = 1, \ldots, L \tag{9}$$

в точках (x, y), где изображение U имеет отличия от изображения V.

При реализации оценки вероятностей, входящих в выражения (7) и (6), вычисляются по двумерной гистограмме уровней полутонов, построенной в по изображениям U и V. Выполнение условий (8) и (9) достигается выбором размеров областей, по которым вычисляются оценки вероятностей, и количеством уровней квантования L при построении гистограммы яркости. Значения энтропий в точке с координатами (x, y) вычисляется по некоторой окрестности этой точки. При решении практических задач условие (8), может быть заменено условием:

$$p(u_k, v_l) \gg \sum_{m=1, m\neq k}^{L} p(u_m, v_l)$$

Изображения величин условных энтропий (7) и (6) для полутоновых изображений, полученных из изображений на Рис. 1, показаны на Рис. 3. Вероятности оценивались в окне размера 11×11 при 64 уровнях квантования, а значения $H(U \mid V)$ и $H(V \mid U)$ вычислялись в окрестности 3×3. На изображении $H(U \mid V)$ видны контуры объектов, отсутствующих на изображении V. Для получения маркеров искомых объектов выполняется следующая операция:

$$M^{U \mid V} = T(H(U \mid V) \cdot H(U) - H(V \mid U)), \tag{10}$$

где $T()$ — операция пороговой бинаризации, $H(U)$ — изображение энтропии выхода системы (5), $(\cdot)$ —

Рис. 3. Изображения, построенные по значениям условных энтропий $H(U\,|\,V)$ и $H(V\,|\,U)$, вычисленных для полутоновых версий изображений на Рис. 1.

операция поэлементного умножения изображений. Тогда искомое изображение дефектов, проявляемых в ультрафиолетовом диапазоне, будет найдено с помощью морфологической реконструкции комбинации бинарных масок (1) и (4) (см. Рис. 2):

$$M(U,V) = R^{\delta}_{M^U}(M^{U\,|\,V}),$$

где $M^U = M^U_1 \vee M^U_2$, $\vee$ — операция поэлементного логического «ИЛИ». Изображения результирующей бинарной маски искомых объектов и маски, наложенной на изображение в видимом диапазоне, показаны на Рис. 4. Проведено тестирование предложенной меры различия изображений. Использовались цветные изображения с однородными и текстурными фрагментами (небо, лес, поле), в один из цветовых каналов которых внедрялись от 29 до 92 объектов в виде размытых кругов диаметром от 3 до 7 пикселов и сегментов кривых толщиной от 1 до 7 пикселов. С помощью меры различия найдены 100% внедрённых объектов на однородных фрагментах и 82–93% на фрагментах с текстурой. Наибольшую трудность для обнаружения представляют объекты малых размеров на текстурах с сильными перепадами яркости.

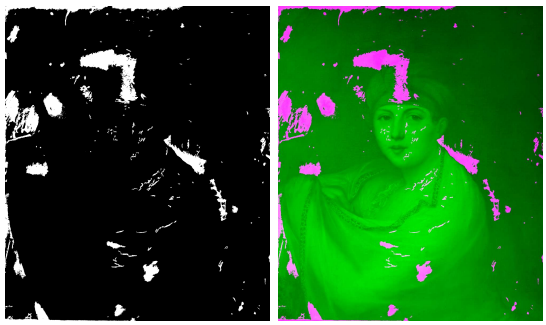


Рис. 4. Результирующая маска дефектов, видимых на УФ изображении, и комбинация маски и цифровой фотографии картины.

Выводы

Предложен подход к локализации записей авторского лако-красочного слоя на изображениях произведений живописи. Объекты, соответствующие по величине яркости участкам вмешательства в авторскую живопись, находятся детекторами на УФ изображении на основе операций полутоновой морфологической реконструкции и пороговой бинаризации с определением порога по локальным характеристикам. Предложена локальная мера отличий изображений в виде значения условной энтропии. Сформулированы условия, при которых значения условной энтропии характеризуют локальные отличия изображений. На основе предложенной меры производится отбор найденных бинарных объектов. Проведённое тестирование показало эффективность предложенного подхода.

Литература

[1] *Kirsh A., Levenson R. S.* Seeing through paintings: Physical examination in art historical studies. — New Haven, CT, Yale U. Press, 2000.

[2] *Heitz F., Maitre H., de Couessin C.* Event Detection in Multisource Imaging: Application to Fine Arts Painting Analysis // IEEE Trans.on acoustics, speech. and signal proc., 1990. — Vol. 38, No. 1. — Pp. 695–704.

[3] *Kammerer P., Hanbury A., Zolda E.* A Visualization Tool for Comparing Paintings and Their Underdrawings // In Proc. of the Conference on Electronic Imaging and the Visual Arts (EVA 2004), 2004. — Pp. 148–153.

[4] *Niblack W.* An Introduction to Digital Image Processing. — Englewood, NJ: Prentice Hall, 1986.

[5] *Soille P.* Morphological Image Analysis: Principles and Applications. — Berlin: Springer-Verlag, 2004. — 391 p.

[6] *Daly S.* The visible differences predictor: an algorithm for the assessment of image fidelity // Digital images and human vision, Cambridge: MIT Press, 1993.

[7] *Petrovic V., Xydeas C.* Evaluation of Image Fusion Performance with Visible Differences // ECCV'2004, LNCS, 2004. — Vol. 3023. — Pp. 380–391.

[8] *Escolano F., Suau P., Bonev B.* Information Theory in Computer Vision and Pattern Recognition. — London: Springer Verlag, 2009.

[9] *Rockinger O., Fechner T.* Pixel-Level Image Fusion: The Case of Image Sequences.— SPIE Proceedings. Signal Processing, Sensor Fusion, and Target Recognition VII, Ivan Kadar, Editor, 1998. — Vol. 3374, — Pp. 378–388.

[10] *Zhang J., Rangarajan A.* Affine image registration using a new information metric. // In: IEEE Computer Vision and Pattern Recognition (CVPR), — 2004. Vol. 1. — Pp. 848–855.

Об условиях устойчивости нахождении осей симметрии зашумленного изображения *

Лепский А. Е.
alex.lepskiy@gmail.com
Москва, национальный исследовательский университет «Высшая школа экономики»

Введен и исследован функционал симметричности зашумленного полутонового изображения. Найдены необходимые и отдельно достаточные условия, при которых оси симметрии, вычисленные с помощью функционала симметричности до и после стационарного некоррелированного зашумления изображения, не меняются. Отдельно исследован случай бинарного изображения.

При анализе изображений часто бывает необходимо определить симметричность объектов, представленных на изображении. Традиционно выделяют два основных класса задач анализа симметрии изображений: определение меры симметрии [1] и определение осей симметрии или параметров симметрии [2, 3]. Значительно меньше в литературе уделено внимания анализу влияния зашумления изображения на меру симметрии и определение параметров симметрии. Ниже будет рассмотрено влияние стационарного некоррелированного зашумления на нахождения оси симметрии. На первый взгляд представляется, что стационарное некоррелированное зашумление не должно влиять на нахождение оси симметрии. Однако, как будет показано ниже (и что подтверждено численным моделированием), при некотором соотношении между интенсивностью зашумления и функцией изображения типа отношения сигнал-шум наблюдается неустойчивость нахождения оси симметрии.

Осесимметричное изображение и мера симметричности

Пусть Ω — некоторая ограниченная область на плоскости R^2, $I \subseteq \Omega$, $f_I(\mathbf{x}) = f_I(x,y)$ — функция изображения (или изображение) объекта на множестве I, причём $f_I(\mathbf{x}) = 0$, если $\mathbf{x} \notin I$ (другими словами, $f_I(\mathbf{x})$ — финитная функция, а I — носитель этой функции). Вообще говоря, как само множество I, так и его граница могут иметь конечное число компонент связности. В общем случае будем считать, что $-\infty < f_I(\mathbf{x}) < \infty$ (отрицательные значения функции яркости можно интерпретировать как значения в логарифмической шкале). Область Ω определяет область изображения, а множество I — область локализации выделенного объекта на изображении. Относительно Ω и I будем предполагать, что это измеримые множества, а функция $f_I(\mathbf{x})$ интегрируема в Ω. Через $L^2(\Omega)$ будем обозначать нормированное пространство измеримых и суммируемых с квадратом функций с нормой $\|f\| = \sqrt{\int_\Omega f^2(\mathbf{x})d\mathbf{x}}$.

Работа выполнена при финансовой поддержке РФФИ, проект № 10-07-00135, 10-07-00478, 11-07-00591.

Рассмотрим понятие симметричного изображения объекта на плоскости. В этой работе будем рассматривать только понятие осесимметричного изображения. Поэтому введем в рассмотрение преобразование симметрии $F_L = F_{\alpha,p}$ относительно прямой $L = L_{\alpha,p}$: $x\cos\alpha + y\sin\alpha = p$, $p \geqslant 0$, $0 < \alpha \leqslant \pi$, как отображение, которое каждой точке $\mathbf{x} \in R^2$ ставит в соответствие точку $F_{\alpha,p}\mathbf{x}$, симметричную к точке $\mathbf{x}$ относительно прямой $L = L_{\alpha,p}$. Заметим, что $F_{\alpha,p}\mathbf{x} = \mathbf{x} + 2\left(p - (\mathbf{n}_\alpha, \mathbf{x})\right)\mathbf{n}_\alpha$, где $\mathbf{n}_\alpha = (\cos\alpha, \sin\alpha)$. Далее будем предполагать, что область Ω и множество I таковы, что для любой прямой L, $L \cap I \neq \varnothing$: $F_L(I) \subseteq \Omega$, где $F_L(I) = \{F_L(\mathbf{x}) : \mathbf{x} \in I\}$ — образ носителя изображения при осевом отображении.

Определение 1. *Назовем изображение $f_I \in L^2(\Omega)$ осесимметричным, если существует такая прямая L, что $\|f_I - f_I(F_L)\| = 0$.*

Норма $\|f_I - f_I(F_L)\|$ в литературе называется преобразованием отражательной симметрии [4]. Рассмотрим функционал осевой симметрии функции изображения f_I (или просто — функционал осевой симметрии изображения) $\Phi(f_I) = \frac{1}{|I|}\inf\limits_L \|f_I - f_I(F_L)\|^2$, где инфимум берется по всем прямым L: $L \cap I \neq \varnothing$.

Заметим, что если изображение f_I является симметричным, то $\Phi(f_I) = 0$. Поэтому, значение функционала $\Phi(f_I)$ будет характеризовать величину осевой «несимметричности» изображения $f_I \in L^2(\Omega)$. В общем случае

$$\Phi(f_I) = \frac{1}{|I|}\inf_L \|f_I - f_I(F_L)\|^2 =$$

$$\frac{1}{|I|}\inf_L \left\{ \|f_I\|^2 + \|f_I(F_L)\|^2 - 2\int\limits_{R^2} f_I(\mathbf{x}) f_I(F_L\mathbf{x}) d\mathbf{x} \right\} =$$

$$= \frac{2}{|I|}\|f_I\|^2 - \frac{2}{|I|}\inf_L \left\{ \int\limits_{R^2} f_I(\mathbf{x}) f_I(F_L\mathbf{x}) d\mathbf{x} \right\}.$$

Другими словами, прямая симметрии может быть найдена из условия максимизации корреляции между функцией изображения и образом этого

изображения при осевой симметрии относительно прямой.

Мера симметричности зашумленного изображения

Будем рассматривать зашумление области изображения объекта I, которое задаётся равенством $\tilde{f}_I(\mathbf{x}) = f_I(\mathbf{x}) + \xi_I(\mathbf{x})$, где $\xi_I(\mathbf{x})$ — случайная функция, удовлетворяющая условиям: 1) $\mathrm{E}[\xi_I(\mathbf{x})] = 0$ для любого $\mathbf{x} \in \Omega$; 2) $\sigma^2[\xi_I(\mathbf{x})] = \begin{cases} \sigma^2, & \mathbf{x} \in I, \\ 0, & \mathbf{x} \notin I; \end{cases}$ 3) $K(\mathbf{x}_1, \mathbf{x}_2) = \mathrm{E}[\xi_I(\mathbf{x}_1)\xi_I(\mathbf{x}_2)] = 0$ для любых $\mathbf{x}_1 \neq \mathbf{x}_2$. Заметим, что можно рассматривать зашумление не области изображения объекта I, а всего изображения Ω. В этом случае результаты существенно не меняются. Тогда мерой симметричности зашумленного изображения $\tilde{I}$ (изображения I с зашумленной функцией изображения $\tilde{f}_I$) назовем функционал

$$\Phi(\tilde{f}_I) = \frac{1}{|I|} \inf_L \int_{R^2} \mathrm{E}\left[\left(\tilde{f}_I(\mathbf{x}) - \tilde{f}_I(F_L\mathbf{x})\right)^2\right] d\mathbf{x}. \quad (1)$$

Рассмотрим свойства функционала симметричности зашумленного изображения. Во-первых, этот функционал можно упростить следующим образом.

Утверждение 1. *Для функционала симметричности $\Phi(\tilde{f}_I)$ зашумленного изображения $\tilde{f}_I$ справедливо равенство*

$$\Phi(\tilde{f}_I) = \sigma^2 + \frac{1}{|I|} \inf_L \left\{ \|f_I - f_I(F_L)\|^2 + \sigma^2 |I \cap F_L(I)| \right\}, \quad (2)$$

где $|A|$ — площадь множества A.

Доказательство. Упростим функционал $\Phi(\tilde{f}_I)$. Имеем

$$\Phi(\tilde{f}_I) = \frac{1}{|I|} \inf_L \int_{R^2} \mathrm{E}\left[\left(\tilde{f}_I(\mathbf{x}) - \tilde{f}_I(F_L\mathbf{x})\right)^2\right] d\mathbf{x} =$$

$$= \frac{1}{|I|} \inf_L \left\{ \|f_I - f_I(F_L)\|^2 + \right.$$

$$+ \int_{R^2} \mathrm{E}\left[\left(\xi_I(\mathbf{x}) - \xi_I(F_L\mathbf{x})\right)^2\right] d\mathbf{x} +$$

$$\left. + 2\int_{R^2} (f_I(\mathbf{x}) - f_I(F_L\mathbf{x}))\, \mathrm{E}[\xi_I(\mathbf{x}) - \xi_I(F_L\mathbf{x})] d\mathbf{x} \right\} =$$

$$= \frac{1}{|I|} \inf_L \left\{ \|f_I - f_I(F_L)\|^2 + \int_{R^2} \mathrm{E}\left[\xi_I^2(\mathbf{x})\right] d\mathbf{x} + \right.$$

$$\left. + \int_{R^2} \mathrm{E}\left[\xi_I^2(F_L\mathbf{x})\right] d\mathbf{x} \right\} =$$

$$= \sigma^2 + \frac{1}{|I|} \inf_L \left\{ \|f_I - f_I(F_L)\|^2 + \sigma^2 |I \cap F_L(I)| \right\},$$

поскольку $\mathrm{E}[\xi_I(\mathbf{x}) - \xi_I(F_L\mathbf{x})] = 0$ для любого $\mathbf{x} \in \mathbf{R^2}$, $\mathrm{E}[\xi_I(\mathbf{x})\xi_I(F_L\mathbf{x})] = 0$ для любого $\mathbf{x} \notin L$ и утверждение доказано.

Для бинарного изображения $f_I(\mathbf{x}) = h$ при $\mathbf{x} \in \mathbf{I}$ и $f_I(\mathbf{x}) = 0$ при $\mathbf{x} \notin I$ равенство (2) можно упростить.

Следствие 1. *Для зашумленного бинарного изображения справедливо равенство*

$$\Phi(\tilde{f}_I) = \sigma^2 + \frac{1}{|I|} \inf_L \left\{ h^2 |I \Delta F_L(I)| + \sigma^2 |I \cap F_L(I)| \right\},$$

где $A\Delta B = (A \cup B) \setminus (A \cap B)$ – операция симметрической разности двух множеств.

Следствие 2. *Для меры симметричности зашумленного изображения f_I верна оценка $\sigma^2 + \Phi(f_I) \leqslant \leqslant \Phi(\tilde{f}_I) \leqslant 2\sigma^2 + \Phi(f_I)$. В частности, если f_I — симметричное изображение, то $\sigma^2 \leqslant \Phi(\tilde{f}_I) \leqslant 2\sigma^2$.*

Доказательство. Имеем $\Phi(\tilde{f}_I) \geqslant \sigma^2 + \frac{1}{|I|} \inf_L \|f_I - f_I(F_L)\|^2 + \frac{\sigma^2}{|I|} \inf_L |I \cap F_L(I)| = \sigma^2 + \Phi(f_I)$, поскольку $\inf_L |I \cap F_L(I)| = 0$ (для такой прямой L, что $L \cap \mathrm{con}(I) = \varnothing$, где $\mathrm{con}(I)$ — выпуклая оболочка множества I). С другой стороны, из (2) следует, что $\Phi(\tilde{f}_I) \leqslant 2\sigma^2 + \Phi(f_I)$. В частности, если f_I — симметричное изображение, то $\Phi(f_I) = 0$ и, следовательно, $\sigma^2 \leqslant \Phi(\tilde{f}_I) \leqslant 2\sigma^2$.

О степени устойчивости оси симметрии к зашумлению изображения

Пусть изображение f_I является симметричным, т. е. $\Phi(f_I) = 0$. Тогда существует такая прямая L_0, что $\|f_I - f_I(F_{L_0})\| = 0$. Подвергнем изображение f_I зашумлению и поставим вопрос о величине «отклонения» прямой симметрии зашумленного изображения от прямой симметрии незашумленного изображения. Некоторым упрощением этого вопроса будет следующая формулировка: пусть изображение f_I является симметричным относительно прямой L_0. Каким условиям должно удовлетворять изображение f_I и уровень зашумления σ^2, чтобы после зашумления прямая

симметрии, доставляющая минимум функционалу (1), не изменилась? Вообще говоря, изображение может иметь множество (и даже бесконечное множество) осей симметрии. Будем говорить, что ось симметрии симметричного изображения после зашумления существенно не изменилась, если $\Phi(\tilde{f}_I) = \frac{1}{|I|}\int\limits_{R^2} \mathrm{E}\left[\left(\tilde{f}_I(\mathbf{x}) - \tilde{f}_I\left(F_{L_0}\mathbf{x}\right)\right)^2\right]d\mathbf{x}$ и $\|f_I - f_I\left(F_{L_0}\right)\| = 0$. В противном случае будем говорить, что ось симметрии существенно изменилась. Другими словами, в последнем случае ось симметричного изображения «перескакивает» в такую позицию, где оси симметрии нет. Для зашумленного изображения $\tilde{f}_I$ при «удалении» прямой L от L_0 первое слагаемое под знаком inf в (2) возрастает, а второе слагаемое при этом убывает ($\max\limits_L |I \cap F_L(I)| = |I \cap F_{L_0}(I)| = |I|$). Из формулы (2) для функционала симметричности зашумленного выражения непосредственно вытекает справедливость следующего утверждения.

Утверждение 2. *Если после зашумления симметричного изображение f_I ось симметрии существенно не изменилась, то $\Phi(\tilde{f}_I) = 2\sigma^2$.*

Найдем условие существенной неизменности оси симметрии зашумленного изображения. Пусть $M_S[\varphi] = \frac{1}{|S|}\int\limits_S \varphi(\mathbf{x})d\mathbf{x}$ — среднеинтегральное значение функции φ в области S. Верно следующее утверждение.

Утверждение 3. *Пусть f_I — симметричное изображение. Если справедлива оценка*

$$\sigma^2 \leqslant 2M_{I\setminus F_L(I)}\left[f_I^2\right] \quad \forall L, \tag{3}$$

то после зашумления изображения f_I ось симметрии существенно не изменится.

Доказательство. Предположим, что f_I — симметричное изображение и верна оценка (3). Тогда (3) можно переписать следующим образом:

$$2\int\limits_{I\setminus F_L(I)} f_I^2(\mathbf{x})d\mathbf{x} \geqslant |I\setminus F_L(I)|\,\sigma^2 \quad \forall L. \tag{4}$$

Так как $\int\limits_{I\cap F_L(I)} f_I^2(\mathbf{x})d\mathbf{x} = \int\limits_{I\cap F_L(I)} f_I^2\left(F_L\mathbf{x}\right)d\mathbf{x}$, то, используя неравенство Коши-Буняковского, из (4) получим

$$2\int\limits_I f_I^2(\mathbf{x})d\mathbf{x} \geqslant \int\limits_{I\cap F_L(I)} f_I^2(\mathbf{x})d\mathbf{x} + |I\setminus F_L(I)|\,\sigma^2 \geqslant$$
$$\geqslant 2\sqrt{\int\limits_{I\cap F_L(I)} f_I^2(\mathbf{x})d\mathbf{x}}\cdot\sqrt{\int\limits_{I\cap F_L(I)} f_I^2\left(F_L\mathbf{x}\right)d\mathbf{x}} +$$
$$+ |I\setminus F_L(I)|\,\sigma^2 \geqslant 2\int\limits_{I\cap F_L(I)} f_I(\mathbf{x})f_I\left(F_L\mathbf{x}\right)d\mathbf{x} +$$
$$+ |I|\,\sigma^2 - |I\cap F_L(I)|\,\sigma^2.$$

Таким образом,

$$2\int\limits_I f_I^2(\mathbf{x})d\mathbf{x} \geqslant 2\int\limits_{I\cap F_L(I)} f_I(\mathbf{x})f_I\left(F_L\mathbf{x}\right)d\mathbf{x} + |I|\,\sigma^2 - |I\cap F_L(I)|\,\sigma^2 \quad \forall L. \tag{5}$$

Так как $\int\limits_{R^2} f_I^2(\mathbf{x})d\mathbf{x} = \int\limits_{R^2} f_I^2\left(F_L\mathbf{x}\right)d\mathbf{x} = \int\limits_I f_I^2(\mathbf{x})d\mathbf{x}$, то (5) можно переписать так

$$|I|\,\sigma^2 \leqslant \|f_I - f_I\left(F_L\right)\|^2 + |I\cap F_L(I)|\,\sigma^2 \quad \forall L.$$

Откуда, учитывая (2), получим, что $\Phi(\tilde{f}_I) \geqslant 2\sigma^2$. Но из следствия 2 утверждения 1 вытекает, что для симметричного изображения справедлива оценка $\Phi(\tilde{f}_I) \leqslant 2\sigma^2$. Таким образом, имеем:

$$\Phi(\tilde{f}_I) = 2\sigma^2. \tag{6}$$

Покажем, что при выполнении условия (3) из (6) следует, что после зашумления изображения f_I ось симметрии существенно не изменится. Действительно, из (2) и (6) вытекает, что ось симметрии должна удовлетворять равенству

$$\inf_L\left\{\|f_I - f_I\left(F_L\right)\|^2 - |I\setminus F_L(I)|\,\sigma^2\right\} = 0$$

или

$$2\|f_I\|^2 = \inf_L\left\{2\int\limits_{I\cap F_L(I)} f_I(\mathbf{x})f_I\left(F_L\mathbf{x}\right)d\mathbf{x} + |I\setminus F_L(I))|\,\sigma^2\right\}.$$

Последнее равенство можно записать так:

$$\inf_L\left\{\int\limits_{I\cap F_L(I)}\left(f_I^2(\mathbf{x}) - f_I(\mathbf{x})f_I\left(F_L\mathbf{x}\right)\right)d\mathbf{x} + \int\limits_{I\setminus F_L(I)}\left(f_I^2(\mathbf{x}) - \tfrac{1}{2}\sigma^2\right)d\mathbf{x}\right\} = 0. \tag{7}$$

Проанализируем равенство (7). Первое слагаемое под знаком inf неотрицательно в силу неравенства Коши-Буняковского. Второе слагаемое также неотрицательно в силу условия (3). Поэтому равенство нулю в (7) возможно только при условии, что ось симметрии после зашумления удовлетворяет равенству

$$\inf_L \left\{ \int\limits_{I \cap F_L(I)} \left(f_I^2(\mathbf{x}) - f_I(\mathbf{x}) f_I\left(F_L\mathbf{x}\right)\right) d\mathbf{x} \right\} = 0,$$

которое равносильно равенству $\inf\limits_L \|f_I - f_I\left(F_L\right)\|^2 = 0$, что и означает существенную неизменность оси симметрии после зашумления. Утверждение доказано.

Следствие 3. *Пусть f_I — бинарное симметричное изображение, т. е. $f_I(\mathbf{x}) = h = const$ для всех $\mathbf{x} \in I$. Если $\sigma^2 \leqslant 2h^2$, то после зашумления изображения его ось симметрии существенно не изменится.*

Достаточное условие устойчивости оси симметрии при зашумлении, сформулированные в последнем утверждении и его следствии, можно частично (а в случае бинарного изображения и полностью) обратить.

Утверждение 4. *Пусть f_I — симметричное изображение. Если для этого изображения и зашумления верна оценка*

$$\sigma^2 > 2M_{I \setminus F_L(I)}\left[f_I^2\right] \quad \forall L, \tag{8}$$

то после зашумления ось симметрии существенно изменится.

Доказательство. Предположим, что f_I — симметричное изображение, выполняется условие (8), но ось симметрии существенно не изменилась. Тогда для такой прямой L^0, что $L^0 \cap \mathrm{con}(I) = \varnothing$ из (2) следует, что

$$\Phi(\tilde{f}_I) \leqslant \sigma^2 + \frac{2}{|I|} \int\limits_{|I|} f_I^2(\mathbf{x}) d\mathbf{x} = \\ = \sigma^2 + 2M_{I \setminus F_{L^0}(I)}\left[f_I^2\right] < 2\sigma^2.$$

Таким образом, $\Phi(\tilde{f}_I) < 2\sigma^2$, что противоречит утверждению 2, если ось симметрии существенно не изменилась. Утверждение доказано.

Из утверждений 3 и 4 вытекает справедливость следующего критерия существенной неизменности оси симметрии для бинарного изображения.

Следствие 4. *Пусть f_I — бинарное симметричное изображение, $f_I(\mathbf{x}) = h = const$ для всех $\mathbf{x} \in I$. Тогда ось симметрии существенно не изменится в том и только том случае, если $\sigma^2 \leqslant 2h^2$.*

Теоретический эффект смещённости оси симметрии при большой интенсивности зашумления был подтвержден и при численном моделировании. На рис. 1 показан результат вычисление оси симметрии бинарного $h = 50$ изображения зашумленного прямоугольника при $\sigma = 50$. В этом случае ось симметрии находится правильно.

Рис. 1.

На рис. 2 показан результат вычисление оси симметрии бинарного $h = 50$ изображения зашумленного прямоугольника при $\sigma = 100$. В этом случае ось симметрии находится неправильно.

Рис. 2.

Выводы

В статье найдены необходимые и достаточные условия типа соотношения «сигнал-шум» для функции полутонового и бинарного изображений и интенсивности некоррелированного стационарного зашумления, при которых оси симметрии, вычисленной с помощью функционала (1), остаются неизменными. В ходе вычислительных экспериментов был подтвержден «эффект неустойчивости» нахождения осей симметрии при достаточно большой интенсивности зашумления. Рассмотренные условия необходимо учитывать при численном нахождении параметров симметричности объектов на зашумленных изображениях. Представляет интерес также исследование влияния нестационарного зашумления на определение параметров симметричности.

Литература

[1] *Kazhdan M., Chazelle B., Dobkin D. et al.* A reflective symmetry descriptor // 7th Europ. Conf. on Comp. Vis. (ECCV 2002) , 2002. — Pp. 642–656.

[2] *Marola G.* On the detection of the axes of symmetry of symmetric and almost symmetric planar images // IEEE Trans. Pattern Anal. Mach. Intell. —1989. — Vol. 11, № 1. Pp. 104–108.

[3] *Горбань А. С.* Методы обнаружения отражательной симметрии полутоновых изображений // Интеллект. сист. принятия решений и приклад. аспекты инф. техн-й. — 2006.— Т. 1. С. 57–61.

[4] *Каркищенко А. Н., Горбань А. С.* К определению мер сходства полутоновых изображений // «Известия ЮФУ. Технические науки».— 2008.— Т. 81, № 4. С. 98–103.

Преобразование симметрии периодических структур в частотной области*

Каркищенко А. Н., Мнухин В. Б.
karkishalex@gmail.com, v.mnukhin@gmail.com
Москва, НИИАС

Предлагается метод построения элементарной ячейки периодической структуры на плоскости, основанный на фильтрации Фурье-образа изображения данной структуры.

Введение

Изучение симметрии изображений является в настоящее время одним из активно развиваемых направлений теоретической информатики. Работы в этой области активно стимулируются такими направления развития информационных технологий как робототехника, искусственное зрение, методы автоматического контроля и др. [1].

Как известно [2, 3], симметрии на плоскости исчерпываются комбинациями отражений, вращений и трансляций (сдвигов), причём трансляции возникают при описании симметрии плоских периодических структур, называемых *орнаментами* и *бордюрами* (фризами). Несмотря на наличие бесконечного числа таких периодических структур, с каждой из них однозначно ассоциируется одна из 24 так называемых *кристаллографических групп*. Эти группы не зависят от размера, поворота, яркости и плотности изображений структур, являясь, тем самым, их сильными дескрипторами. Использование кристаллографических групп в качестве дескрипторов позволяют проводить эффективный поиск изображения в базах данных при условии наличии на данном изображении одного или нескольких периодических фрагментов, что может использоваться, в частности, при автоматическом аннотировании изображений в Интернете. Проблемы распознавания периодической структуры естественно возникают также в рентгеноструктурном анализе, электронной микроскопии, нанотехнологиях, при анализе аэрофотоснимков, текстурном анализе образов [4], анализе походки [5], и т. п.

Реализация подобного подхода требует развитых методов анализа симметрий изображений, позволяющих с высокой степенью надежности классифицировать периодические структуры по их группам [6]. Одним из таких методов является *непрерывное преобразование симметрии*, предложенное в [7, 8, 9]. Использовавшийся вначале для распознавания отражательной и вращательной симметрии, в [10, 11] метод обобщен и на трансляционную симметрию. В частности, с его помощью задача выделения элементарной ячейки изображения периодической структуры на плоскости сводится [10] к поиску экстремума некоторой функции. Для оптимального выбора начальных условий, обеспечивающих эффективность такого поиска, предложено [11] использовать преобразование Фурье, сохраняющее, как известно, симметрии исходного изображения. Более того, можно предложить перенести в частотную область сам метод непрерывного преобразования симметрии, тем самым сняв проблему выбора начальных условий.

В настоящей работе метод непрерывного преобразования симметрии рассматривается в частотной области. Предлагается метод построения элементарной ячейки плоской периодической структуры, основанный на специальной фильтрации Фурье-образа её изображения.

Периодические структуры на плоскости

Уточним понятие периодической структуры на плоскости. Напомним, что группа G движений плоскости называется *дискретной*, если для любой точки x плоскости найдется окрестность, не содержащая других точек из орбиты точки x.

Определение 1. *Связная замкнутая область F плоскости является фундаментальной для дискретной группы G, если любая точка плоскости принадлежит орбите некоторой точки (возможно, граничной) области F, но никакие две внутренние точки области не лежат в одной орбите группы G.*

Таким образом, все образы фундаментальной области под действием группы G различны и заполняют всю плоскость без пропусков и наложений. Важно отметить, что группа может иметь различные фундаментальные области, но их площади обязаны совпадать, поскольку однозначно определяется группой, см. [11].

Определение 2. *Группа G движений плоскости называется двумерной кристаллографической группой, если она имеет ограниченную фундаментальную область.*

Простейшая двумерная кристаллографическая группа, обозначаемая $p1$, порождается сдвигами вдоль двух неколлинеарных векторов. Всякая фундаментальная область группы $p1$ является параллелограммом. Соответствующее разбиение плоскости показано на рис. 1.

Работа выполнена при финансовой поддержке РФФИ, проекты № 10-07-00478-а, № 10-07-00135-а и № 11-07-00591-а.

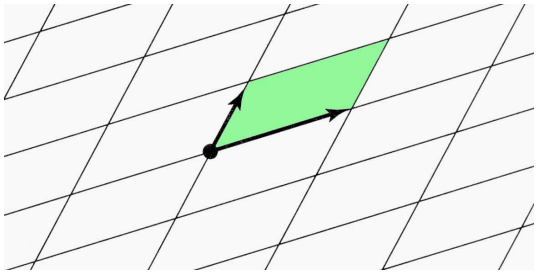


Рис. 1. Фундаментальные области группы $p1$.

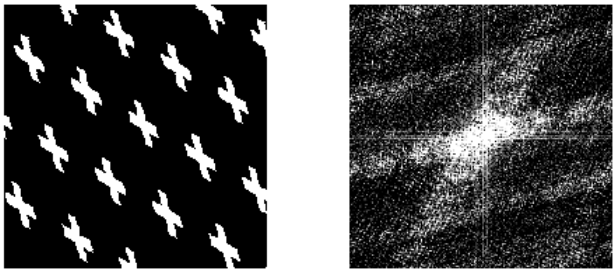


Рис. 2. Спектр периодической структуры.

Поместив внутрь фундаментальной области двумерной кристаллографической группы произвольную асимметричную фигуру, получим под действием группы бесконечную периодическую структуру, обычно называемую *орнаментом*. (Двумерные кристаллографические группы иногда называют *группами орнаментов*. Орнамент с группой $p1$ изображен на рис. 3а.) Аналогично, рассмотрение *бордюров* (т. е. плоских структур, бесконечно повторяющихся вдоль одного направления) приводит к *одномерным кристаллографическим группам*. Доказано [3], что существуют только 7 одномерных и 17 двумерных кристаллографических групп. Таким образом, бесконечное множество периодических структур на плоскости естественно распадается на 24 класса.

Помимо сдвигов кристаллографические группы могут порождаться отражениями и вращениями. Вместе с тем каждая двумерная кристаллографическая группа G содержит абелеву подгруппу H, состоящую из всех сдвигов в G. Можно показать [2], что H всегда имеет тип $p1$.

Определение 3. *Параллелограмм* Φ, *являющийся фундаментальной областью подгруппы сдвигов* H, *называется элементарной ячейкой группы* G.

Другими словами, элементарная ячейка группы есть наименьшая часть плоскости, образующая её «замощение» под действием только сдвигов из G.

Преобразование Фурье изображений периодических структур

Задача выделения элементарной ячейки является важным этапом классификации периодической структуры. Покажем, что её можно свести к определенной фильтрации Фурье-образа изображения данной структуры.

Рассмотрим непрерывный случай. Будем использовать следующее преобразование Фурье:

$$\mathscr{F}[f] = F(u,v) = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} f(x,y)\, e^{-2\pi i(xu+yv)}\,dx\,dy,$$

$$\mathscr{F}^{-1}[F] = f(x,y) = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} F(u,v)\, e^{2\pi i(xu+yv)}\,du\,dv.$$

Как известно [13], справедлива теорема о сдвиге

$$\mathscr{F}[f(x-a,y-b)] = e^{-2\pi i(au+bv)}F(u,v);$$

кроме того, если A — матрица невырожденного линейного преобразования плоскости, то

$$\mathscr{F}[f(\mathbf{x}A)] = \frac{1}{|A|}F\big(\mathbf{u}A^{-\mathrm{T}}\big),$$

где $\mathbf{x} = (x,y)$, $\mathbf{u} = (u,v)$, и $A^{-\mathrm{T}} = \big(A^{\mathrm{T}}\big)^{-1}$. Таким образом, преобразование Фурье сохраняет симметрии исходного изображения, что позволяет использовать его (см. [12, 14]) для анализа периодических структур на плоскости.

Введем определения. Под *изображением* f в некоторой области I действительной плоскости $\mathbb{R}^2$ будем понимать неотрицательную ограниченную функцию $f(x,y)$, $(x,y) \in I$. Не ограничивая общность, можем считать, что $0 \leqslant f(x,y) \leqslant 1$, и полагать I единичным квадратом: $I = [0,1] \times \times [0,1]$. Ограничение изображения на подобласть $X \subseteq I$ обозначим через f_X: $f_X = f(x,y)$ при $(x,y) \in \in X$, и $f_X(x,y) = 0$ при $(x,y) \notin X$. В частности, ограничение f на прямоугольник $X = [a,1-a] \times \times [b,1-b] \subseteq I$, (где $0 \leqslant a \leqslant 1/2$, $0 \leqslant b \leqslant 1/2$), будем обозначать через $f_{a,b}$.

Рассмотрим вначале случай, когда f является фрагментом *прямоугольного* орнамента, т. е. такого, что подгруппа H порождается сдвигами вдоль двух ортогональных векторов, параллельных координатным осям и имеющим длины p и q, $0 < p < 1$, $0 < q < 1$. Пусть $(a,b) \in I$ — некоторая точка такая, что прямоугольник $\Phi = [a, a+p] \times [b, b+q]$ целиком лежит внутри единичного квадрата I. Тогда Φ может быть выбран в качестве элементарной ячейки заданного орнамента. Если $k = \lfloor \frac{1}{p} \rfloor$, $l = \lfloor \frac{1}{q} \rfloor$, то Φ укладывается в I ровно $k \times l$ раз, оставляя «зазоры» с совокупными длинами $\left\{\frac{1}{p}\right\} = 1 - pk$ и $\left\{\frac{1}{q}\right\} = 1 - ql$.

Для произвольных $0 \leqslant a \leqslant 1/2$, $0 \leqslant b \leqslant 1/2$ обозначим через $F_{a,b} = \mathscr{F}[f_{a,b}]$ Фурье-образ изображения f в прямоугольнике $[a,1-a] \times [b,1-b] \subseteq I$.

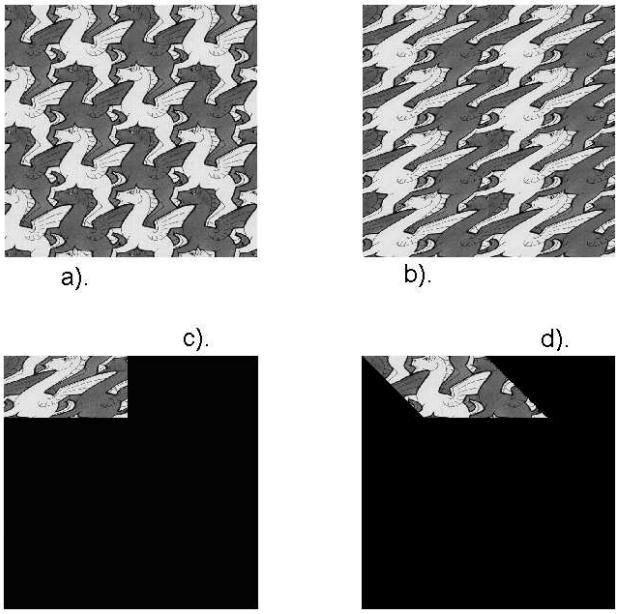


Рис. 3. Выделение элементарной ячейки орнамента.

Пусть $a_0 = \frac{1}{2}\left\{\frac{1}{p}\right\}$, $b_0 = \frac{1}{2}\left\{\frac{1}{q}\right\}$, и

$$H_{a,b}(u,v) = \frac{\sin(\pi up)}{\sin\big(\pi u(1-2a)\big)}\,\frac{\sin(\pi vq)}{\sin\big(\pi v(1-2b)\big)}\cdot \cdot e^{-\pi i[u(p-1)+v(q-1)]}.$$

Тогда справедлив следующий результат:

Теорема 1. *Если f_Φ — изображение элементарной ячейки прямоугольного орнамента f, то*

$$f_\Phi = \lim_{a\to a_0}\lim_{b\to b_0} \mathscr{F}^{-1}\big[F_{a,b}H_{a,b}\big].$$

Доказательство. Пусть $\Phi = \mathscr{F}[f_\Phi]$. Применяя соотношения $pk + 2a_0 = 1$, $ql + 2b_0 = 1$ и теорему о сдвиге, получим

$$\begin{aligned} F_{a_0,b_0} &= \Phi e^{-2\pi i(ua_0+vb_0)}\sum_{s=0}^{k-1}\sum_{t=0}^{l-1} e^{-2\pi i(psu+qtv)} = \\ &= \Phi e^{-2\pi i(ua_0+vb_0)}\,\frac{1-e^{-2\pi iupk}}{1-e^{-2\pi iup}}\,\frac{1-e^{-2\pi ivql}}{1-e^{-2\pi ivq}} = \\ &= \frac{\Phi}{H_{a_0,b_0}}\ , \end{aligned}$$

откуда и вытекает требуемое тождество.

Отметим, что нахождение a_0 и b_0 требует знания длин p и q сторон элементарной ячейки, изначально, как правило, неизвестных. Вместе с тем целые числа k и l могут считаться известными, поскольку они могут быть выявлены статистическим анализом изображения. Таким образом, можно считать известными оценки $0 \leqslant a_0 < 1/2k$, $0 \leqslant b_0 < 1/2l$. Это сводит поиск элементарной ячейки к решению оптимизационной задачи:

$$\mu\Big(f_X, \mathscr{F}^{-1}\big[F_{a,b}H_{a,b}\big]\Big) \longrightarrow \max \quad \text{при} \quad (a,b) \in [0,1/2k]\times[0,1/2l],$$

где μ — мера сходства изображений, $0 \leqslant \mu \leqslant 1$, а $X = [a,(ak-2a+1)/k]\times[b,(bl-2b+1)/l]$.

Рассмотрим теперь случай, когда орнамент f либо не имеет прямоугольных элементарных ячеек, либо стороны его прямоугольной элементарной ячейки не параллельны координатным осям. Заметим, что после подходящего невырожденного линейного преобразования A плоскости орнамент $f^* = f\big((x,y)A\big)$ уже будет обладает требуемыми свойствами (отметим, что при нахождении f^* следует полагать f определённым всюду на $\mathbb{R}^2$). Элементарная ячейка Φ преобразованного орнамента f^* может быть найдена рассмотренным выше методом, после чего элементарная ячейка орнамента f определяется как $f^*_\Phi\big((x,y)A^{-1}\big)$. Указанный процесс показан на рис. 3, где (a) показывает исходный орнамент f, (b) — преобразованный орнамент f^*, (c) — результат выделения Φ, и (d) — элементарную ячейку орнамента f.

Для нахождения преобразования A можно (см. [12, 14]) использовать статистический анализ Фурье-образа F изображения f. Типичный вид спектра $|F|$ периодической структуры (подвергнутый логарифмическому преобразованию для улучшения видимости) показан на рис. 2, по которому нетрудно выделить направления сдвигов, порождающих элементарную ячейку.

Реализация метода для цифровых изображений

Рассмотрим цифровое изображение $f(x,y)$, определённое для дискретных аргументов $x, y \in \mathbb{Z}$ в прямоугольнике $[0,M-1]\times[0,N-1]$, и дискретное преобразование Фурье (ДПФ):

$$F(u,v) = \frac{1}{MN}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)\, e^{-2\pi i\left(x\frac{u}{M}+y\frac{v}{N}\right)}.$$

Величины M и N полагаются фиксированными и оптимизированными для эффективного вычисления БПФ. Заметим, что теперь $M = 2a_0 + pk$, $N = 2b_0 + ql$, где $k = \lfloor\frac{M}{p}\rfloor$, $k = \lfloor\frac{N}{q}\rfloor$. С учётом этих соотношений, в дискретном случае получаем

$$\frac{F_{a_0,b_0}}{\Phi} = \frac{\sin\left(2\pi a_0\frac{u}{M}\right)}{\sin\left(\pi p\frac{u}{M}\right)}\,\frac{\sin\left(2\pi b_0\frac{v}{N}\right)}{\sin\left(\pi q\frac{v}{N}\right)}\,e^{\pi i\left[p\frac{u}{M}+q\frac{v}{N}\right]}.$$

Как было замечено выше, при практической реализации метода известными можно считать только k и l, поэтому фильтрующую функцию $H_{a,b}$ целе-

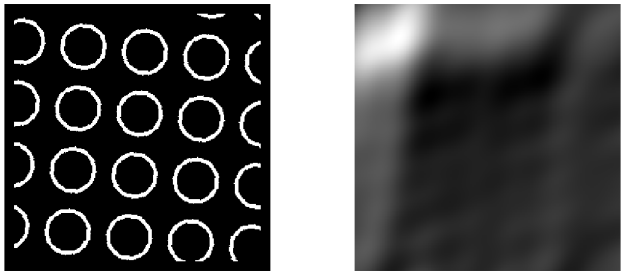


Рис. 4. Выделение Φ с подавлением высоких частот.

сообразно записывать в следующем виде:

$$H_{a,b}(u,v)\cdot e^{\pi i\left(\frac{u}{k}\left[1-\frac{2a}{M}\right]+\frac{v}{l}\left[1-\frac{2b}{N}\right]\right)} = \\ = \frac{\sin\left(\pi\frac{u}{k}\left[1-\frac{2a}{M}\right]\right)}{\sin\left(2\pi a\frac{u}{M}\right)}\frac{\sin\left(\pi\frac{v}{l}\left[1-\frac{2b}{N}\right]\right)}{\sin\left(2\pi b\frac{v}{N}\right)}.$$

Случай $a_0 = b_0 = 0$, когда края изображения совпадают с границами Φ, является особым, поскольку $H_{0,0}$ не определена. Заметим, что

$$F_{0,0}(u,v) = \begin{cases} k\Phi, & \text{если } u \equiv 0 \bmod k; \\ l\Phi, & \text{если } v \equiv 0 \bmod l; \\ kl\Phi, & \text{если } u \equiv 0 \bmod k, v \equiv 0 \bmod l; \\ 0, & \text{в иных случаях.} \end{cases}$$

В этом случае область Φ легко построить, удалив из $F_{0,0}$ нулевые строки и столбцы.

Следует отметить высокую чувствительность метода к поворотам изображения и отклонениям от периодичности. Одним из способов уменьшения этого является использование вместо $H_{a,b}$ функции

$$H^{(s)}_{a,b}(u,v) = H_{a,b}\exp\left[-s(u^2+v^2)\right].$$

Несмотря на искажения, вносимые подавлением высоких частот, фильтрация с $H^{(s)}_{a,b}$ позволяет оценить размеры и найти примерный вид элементарных областей и для «неидеальных» орнаментов. В частности, на рис. 4 показана периодическая структура, повернутая на 5°, и результат выделения элементарной ячейки путём фильтрации с $H^{0.02}_{a,b}$. Светлое пятно в левом верхнем углу является искаженным изображением области Φ. Его наличие указывает на периодичность структуры, а диаметр может служит оценкой размеров элементарной области.

Выводы

Показано, как преобразование Фурье может быть использовано для решения задачи выделения элементарной ячейки изображения периодической структуры на плоскости.

Литература

[1] *Gool L., Moons T., Ungureanu D., Pauwels E.* Symmetry from Shape and Shape from Symmetry // Int. J. Robotics Res., 1995. — Vol. 14, No. 5. — P. 407–424.

[2] *Никулин В. В., Шафаревич И. Р.* Геометрии и группы — Москва: Наука, 1983. — 239 с.

[3] *Gallian J. A.* Contemporary Abstract Algebra — 2002. — 426 p.

[4] *Liu Y., Lin W., Hays J.* Near-regular texture analysis and manipulation // ACM Trans. Graph. — 2008. — Vol. 23, No. 3. — P. 368–376.

[5] *Liu Y., Collins R. T., Tsin Y.* Gait Sequence Analysis Using Frieze Patterns // ECCV, 2002. — Vol. 2. — P. 657–671.

[6] *Liu Y., Collins R. T., Tsin Y.* A computational model for periodic pattern perception based on frieze and wallpaper groups // Trans. Pattern Analysis and Machine Intelligence, 2004. — Vol. 26, No. 3. — P. 354–371.

[7] *Gorban A. S., Karkishchenko A. N.* Detection of symmetry of images based on similarity measures of sets // Proc. 9th Int. Conf. «Pattern Recognition and Image Analysis», Nizhni Novgorod, 2008. — P. 261–264.

[8] *Горбань А. С., Каркищенко А. Н.* Инвариантные характеристики в задачах обнаружения симметрии изображений // САИТ-2007, Москва: Изд-во ЛКИ, 2007. — С. 210–212.

[9] *Горбань А. С., Каркищенко А. Н.* К определению мер сходства полутоновых изображений // Известия ЮФУ. — 2008. — №4. — С. 98–103.

[10] *Каркищенко А. Н., Мнухин В. Б.* Классификация изображений периодических структур на основе непрерывного преобразования симметрии // ИОИ-2010, Москва, МАКС Пресс, 2010. — С. 359–362.

[11] *Каркищенко А. Н., Мнухин В. Б.* Метод непрерывного преобразования симметрии в задачах классификации изображений периодических структур // Труды НИИАС, Москва: Изд-во НИИАС, 2011, (в печати).

[12] *Zhang J., Tan T.* Affine invariant classification and retrieval of texture images // Pattern Recognition, 2003. — Vol. 36, No. 3. — P. 657–664.

[13] *Poularikas A. D.* The Transform and Applications Handbook — CRC Press, 2000. — 1336 p.

[14] *Ben-Arie J., Wang Z. Q.* Pictorial recognition of objects employing affine invariance in the frequency domain // PAMI, 1998. — Vol. 20, No. 6.— P. 604–618.

Согласование изображений пространственно расположенных групповых точечных объектов по угловым координатам

Фурман Я. А., Егошина И. Л., Ерусланов Р. В.
krtmbs@marstu.net
Йошкар-Ола, Марийский государственный технический университет

Рассмотрен подход к устранению углового рассогласования изображений двух групповых точечных объектов. Объекты расположены в трехмерном пространстве и заданы в виде кватернионных сигналов. Задача решена итерационным путем на основе свойств скалярного произведения таких сигналов.

Постановка задачи

В докладе рассмотрен подход к решению важной для распознавания образов проблемы согласования по угловым параметрам двух 3D изображений групповых точечных объектов (ГТО). Под ГТО будем понимать изолированное по значениям тех или иных параметров множество из s упорядоченных (пронумерованных) точек $E = \{\varepsilon(n)\}_{n=0}^{s-1}$. Изображение одного из ГТО является распознаваемым, а другое — эталонным изображениями. Мера схожести двух изображений, по величине которой принимается решение о классе распознаваемого объекта, является функцией нескольких переменных: степени схожести их форм, величины угла между изображениями объектов, сдвига начальных точек и др. Обычно все переменные, кроме первой, играют роль помеховых факторов и перед вынесением решения их влияние должно быть минимизировано. Цель доклада заключается в исследовании подхода к снижению помехового влияния углового рассогласования на величину меры схожести двух 3D изображений ГТО одинаковой размерности (с одинаковым количеством составляющих их точек), основанного на более высокой информативности скалярного (внутреннего) произведения (СП) векторов в кватернионном пространстве H по сравнению с СП этих векторов в линейном действительном пространстве R_3 [1].

Кватернионные сигналы как модели ГТО

Кватернионный сигнал является одной из адекватных математических моделей пространственного ГТО [2]. На базе такой модели ниже будет рассмотрен итерационный подход к решению поставленной задачи.

У. Гамильтон в поисках обобщений комплексных чисел, позволяющих задать точки в 3D и 4D пространствах, ввел, ценой отказа от свойства коммутативности операции умножения, кватернионы [2]. Элементами кватернионного пространства H_1 служат полные кватернионы

$$\mathbf{q}(n) = q_0(n) + q_1(n)\mathbf{i} + q_2(n)\mathbf{j} + q_3(n)\mathbf{k}.$$

Здесь $q_0(n) = \mathrm{Re}\mathbf{q}(n)$ — вещественная часть, а $\mathbf{v}(n) = q_1(n)\mathbf{i} + q_2(n)\mathbf{j} + q_3(n)\mathbf{k}$ — векторная или гиперкомплексная части кватерниона, $q_0(n)$, $q_1(n)$, $q_2(n)$, $q_3(n)$ — произвольные вещественные числа, i, j и k — мнимые единицы. Умножение кватернионов некоммутативно, но именно таким оно должно быть, чтобы описывать вращения трёхмерного пространства, которые не все коммутируют. Исходя из того что билинейная эрмитова форма в H_1 может рассматриваться в качестве СП векторов [3], для одномерного случая будем иметь

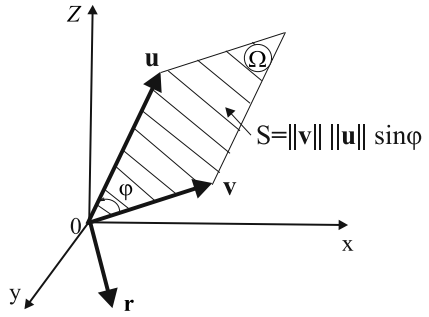


Рис. 1. Геометрическая интерпретация СП векторов **v** и **u** в пространстве H_1.

$$\eta_{H_1} = (\mathbf{v}, \mathbf{u}) = \mathbf{v}\mathbf{u}^* = (\mathbf{v}, \mathbf{u})_{R_3} - [\mathbf{v}, \mathbf{u}].$$

Здесь **u** и **v** — векторные кватернионы, задающие в R_3 векторы $\mathbf{v} = (v_1, v_2, v_3)$ и $\mathbf{u} = (u_1, u_2, u_3)$. Используя известные выражения для СП векторов в R_3 и векторного произведения $[\mathbf{v}, \mathbf{u}]$, получим

$$\eta_{H_1} = (\mathbf{v}, \mathbf{u})_{H_1} = \|\mathbf{v}\|\|\mathbf{u}\| \cos\varphi - \mathbf{r}\|\mathbf{v}\|\|\mathbf{u}\| \sin\varphi, \quad (1)$$

где φ — угол между векторами **v** и **u**, а $\mathbf{r} = r_1(n)\mathbf{i} + r_2(n)\mathbf{j} + r_3(n)\mathbf{k}$ — нормаль к собственной плоскости Ω этих векторов (рис. 1). Отметим, что СП векторов, заданных в H_1, более информативно чем их СП в R_3. Во-первых, оно включает в качестве своей составной части величину $(\mathbf{v}, \mathbf{u})_{R_3}$, и, во-вторых, содержит всю необходимую информацию для вращения вектора **u** до совмещения с вектором **v** (рис. 1).

Пусть в R_3 задан ГТО E. Образуем с полюсом в некоторой точке O пучок из s векторов $\mathbf{V} = \{\mathbf{v}(n)\}_{n=0}^{s-1}$, соединяющих точку O со всеми точками $\varepsilon(n)$ множества E (рис. 2). Каждый из векторов зададим одноимённым векторным кватернионом. Представленный подобным образом пу-

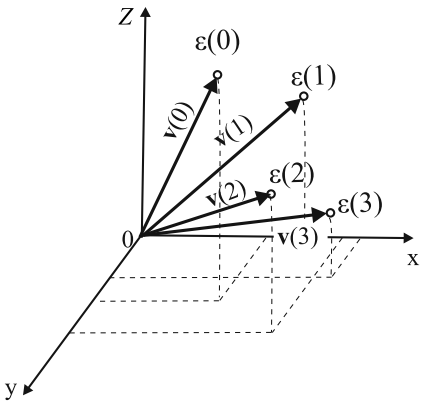


Рис. 2. Задание ГТО $E = \{\varepsilon(n)\}_{n=0}^{3}$ пучком векторов $\{\mathbf{v}n\}_{n=0}^{3}$.

чок $\mathbf{V} = \{\mathbf{v}(n)\}_{n=0}^{s-1}$ назовем кватернионным сигналом (КТС). Скалярное произведение двух КТС $\mathbf{V}$ и $\mathbf{U} = \{\mathbf{u}(n)\}_{n=0}^{s-1}$ на основе выражения (1) имеет вид [1]:

$$(\mathbf{v},\mathbf{u})_{H_s} = \sum_{n=0}^{s-1} \|\mathbf{v}(n)\|\|\mathbf{u}(n)\| \cos\varphi_n - \\ - \sum_{n=0}^{s-1} \mathbf{r}(n)\|\mathbf{v}(n)\|\|\mathbf{u}(n)\| \sin\varphi_n, \quad (2)$$

где φ_n — угол между парциальными векторами $\mathbf{v}(n)$ и $\mathbf{u}(n)$, а $\mathbf{r}(n)$ — нормаль к собственной плоскости Ω_n этих векторов, $n = 0, 1, \ldots, s-1$.

Распознавание ГТО с неизвестным ракурсом

Предполагается, что задан алфавит ГТО из классов $\{A_m\}_{m=0}^{M-1}$, причём каждый класс представлен одним эталонным КТС $\mathbf{U}_m$, $\|\mathbf{U}_m\| = 1$, $m = 0, 1, \ldots, M-1$. Поскольку нас интересует влияние на процесс распознавания неконтролируемых вращений, будем считать, что распознаваемый ГТО был получен только поворотом одного из эталонных ГТО, для конкретности, класса A_l, на угол ψ вокруг оси с направляющим вектором $\boldsymbol{\rho} = \rho_1(n)\mathbf{i} + \rho_2(n)\mathbf{j} + \rho_3(n)\mathbf{k}$. Параметры l, ψ и $\boldsymbol{\rho}$ предполагаются неизвестными. КТС $\mathbf{U}_l$, задающий распознаваемый сигнал, будет иметь вид:

$$\mathbf{V} = \mathbf{b}\mathbf{U}_l\mathbf{b}^{-1}, \quad (3)$$

где $\mathbf{b} = \cos\psi/2 + \boldsymbol{\rho}\sin\psi/2$ — вращающий кватернион. Устройство распознавания принимает решение по критерию минимума расстояния

$$d_m^2 = \|\mathbf{V}\|^2 + \|\mathbf{U}\|^2 - 2\mathrm{Re}(\mathbf{V},\mathbf{U}_m),\ m = 0, \ldots, M-1.$$

Для нормированных КТС $d_m^2 = 2\,(1 - \mathrm{Re}(\mathbf{V},\mathbf{U}_m))$, т. е. в данном случае степень различия (схожести)

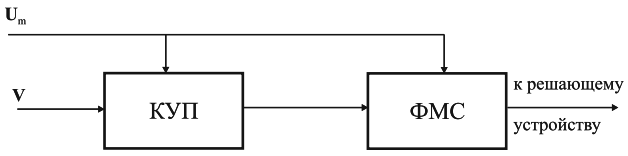


Рис. 3. Структура одного из каналов устройства распознавания ГТО, формирующего сигнал для решающего устройства.

КТС определяется лишь величиной их нормированного СП. С учётом выражения (3), получим

$$\mathrm{Re}(\mathbf{V},\mathbf{U}_n) = \mathrm{Re}(\mathbf{b}\mathbf{U}_l\mathbf{b}^{-1},\mathbf{U}_m), \\ m = 0, 1, \ldots, M-1. \quad (4)$$

Видно, что из-за поворота КТС $\mathbf{U}_l$ реальное значение СП зависит уже не только от степени сходства КТС $\mathbf{V}$ и $\mathbf{U}_l$, но и от параметров вращения ψ и $\boldsymbol{\rho}$. Поскольку эти параметры являются неконтролируемыми, то величина $\mathrm{Re}(\mathbf{V},\mathbf{U}_m)$, независимо от степени сходства этих КТС, может принимать произвольные значения в пределах ± 1. Поэтому устройство распознавания перед вычислением статистики, по величине которой КТС $\mathbf{V}$ будет отнесен к одному из классов алфавита A, для каждого из M классов должно выполнить процедуру согласования КТС $\mathbf{V}$ и $\mathbf{U}_m$, компенсирующую негативные последствия вращения КТС $\mathbf{U}_l$. С учётом этого в каждом канале устройства перед формирователем меры схожести (ФМС) должен находиться корректор угла поворота (КУП) (рис. 3).

Итерационный подход к угловому согласованию ГТО

В рамках принятой модели получения распознаваемого сигнала $\mathbf{V} = \mathbf{b}\mathbf{U}_l\mathbf{b}^{-1}$, мера схожести КТС $\mathbf{V}$ и $\mathbf{U}_l$ достигает максимума, равного 1 при $\psi = 0$. Мера схожести КТС $\mathbf{V}$ и $\mathbf{U}_m$, $m \neq l$, $m = 0, 1, \ldots, M-1$, при их угловом согласовании также достигает максимума, но из-за различия форм ГТО, задаваемыми этими КТС, величина полученного максимума будет меньше единицы. Известные подходы к угловому согласованию двух ГТО базируются на оценке параметров вращения $\boldsymbol{\rho}_m$ и φ_m с последующим компенсирующим поворотом одного из них. Для получения оценок вычисляются элементы матрицы вращения [4] или используются соотношения между кватернионами в составе КТС [5]. Данные подходы будут корректными лишь для случая, когда один из ГТО был получен в результате поворота другого, т. е. когда $\mathbf{V} = \mathbf{b}\mathbf{U}_l\mathbf{b}^{-1}$. Если же $\mathbf{V}$ и $\mathbf{U}_m$, $m \neq l$, не связанные друг с другом КТС, упомянутые подходы не позволяют осуществить их угловое согласование (в смысле достижения максимума их меры схожести), но дают возможность установить отсутствие между ними функциональной зависимости (3). Рассмотрим основанный на свойствах СП

$(\mathbf{V}, \mathbf{U}_m)_{H_s}$ подход к минимизации величины угла между ГТО, заданными КТС $\mathbf{V}$ и $\mathbf{U}_m$, $m = 0, 1, \ldots, M-1$. Из (1) и (2) следует, что вектор $\mathbf{v}(n)$ можно совместить с вектором $\mathbf{u}(n)$, повернув его вокруг нормали $\mathbf{r}(n)$ на угол φ_n, $n = 0, 1, \ldots, s-1$. Но из-за разницы параметров вращения каждого из n парциальных векторов совмещение КТС не произойдет. Угол Φ между КТС можно уменьшить за счёт поворота КТС $\mathbf{V}$ вокруг усреднённой нормали

$$\mathbf{r} = \frac{1}{s} \sum_{n=0}^{s-1} \mathbf{r}(n).$$

Нормированное СП $\cos\Phi = \mathrm{Re}(\mathbf{V}, \mathbf{U})/\|\mathbf{V}\|\|\mathbf{U}\|$ количественно характеризует степень схожести сигналов. При повороте с параметрами $\mathbf{r}$ и Φ будет получен КТС $U^{(1)}$, степень схожести которого с КТС $\mathbf{V}$ равна $\cos\Phi^{(1)}$. При $\Phi^{(1)} < \Phi$ данная операция повторяется, но уже для КТС $U^{(1)}$ и $\mathbf{V}$ и т. д. Критериями остановки могут быть либо условие достижения заданного значения меры схожести $\cos\Phi^{(1)} \geqslant \tau$, либо отсутствие значимых изменений косинуса на нескольких итерациях подряд. С учётом изложенного алгоритм углового согласования изображений ГТО по их КТС в пределах отдельной, c-ой итерации имеет следующий вид.

1. Вычислить СП КТС $\mathbf{U}^{(c)}$ и $\mathbf{V}$:

$$(\mathbf{U}^{(c)}, \mathbf{V}) = \sum_{n=0}^{s-1} \|\mathbf{u}^{(c)}(n)\|\|\mathbf{v}(n)\| \left(\cos\varphi_n^{(c)} - \mathbf{r}_{(n)}^{(c)} \sin\varphi_n^{(c)}\right).$$

2. Найти косинус угла $\Phi^{(c)}$ и направляющий вектор $\mathbf{r}^{(c)}$:

$$\cos\Phi^{(c)} = \frac{1}{\|\mathbf{U}^{(c)}\|\|\mathbf{V}\|} \sum_{n=0}^{s-1} \|\mathbf{U}^{(c)}(n)\|\|\mathbf{V}(n)\| \times \cos\varphi_n^{(c)};$$

$$\mathbf{r}^{(c)} = \left(\sum_{n=0}^{s-1} \mathbf{r}_{(n)}^{(c)} \|\mathbf{U}^{(c)}(n)\|\|\mathbf{V}(n)\| \sin\varphi_n^{(c)} \right)^0,$$

где верхний индекс «0» обозначает операцию нормирования вектора.

3. Повернуть КТС $\mathbf{U}^{(c)}$ на угол $\left(-\Phi^{(c)}\right)$ вокруг оси с направляющим вектором $\mathbf{r}^{(c)}$:

$$\mathbf{U}^{(c+1)} = \left(\cos\Phi^{(c)}/2 - \mathbf{r}^{(c)} \sin\Phi^{(c)}/2\right) \mathbf{U}^{(c)} \times \left(\cos\Phi^{(c)}/2 + \mathbf{r}^{(c)} \sin\Phi^{(c)}/2\right).$$

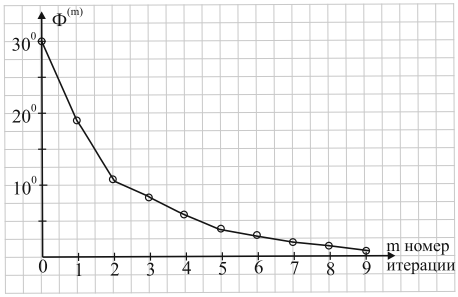


Рис. 4. Зависимость угла между КТС $\mathbf{V}$ и $\mathbf{U}$ от номера итерации (корректная задача)

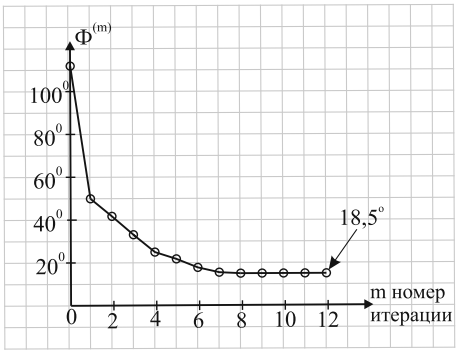


Рис. 5. Зависимость угла между КТС $\mathbf{V}$ и $\mathbf{U}$ от номера итерации (некорректная задача).

4. Найти меру схожести $\cos\Phi^{(c+1)}$ КТС $\mathbf{U}^{(c+1)}$ и $\mathbf{V}$ и сравнить её с величиной $\Phi^{(c)}$ для КТС $\mathbf{U}^{(c)}$ и $\mathbf{V}$.

Если КТС $\mathbf{U}$ и $\mathbf{V}$ связаны преобразованием (3), то они задают один и тот же ГТО. Поэтому с ростом номера итерации величина углового рассогласования $\Phi^{(c)}$ будет стремиться к нулю. Если же эти КТС задают различные ГТО, то величина углового рассогласования между ними зафиксируется на некотором не равном нулю уровне, зависящем лишь от степени различия форм этих ГТО.

Пример углового согласования ГТО для корректной задачи. Распознаваемый ГТО задан КТС $\mathbf{V} = \{2\mathbf{i} + \mathbf{j} + 3\mathbf{k}; -1,667\mathbf{i} + 0,333\mathbf{j} + 0,333\mathbf{k}\}$ и был получен поворотом исходного ГТО на угол $\psi = 60°$ вокруг оси $\boldsymbol{\rho} = \mathbf{i}+\mathbf{j}+\mathbf{k}$. Исходный ГТО был задан КТС $\mathbf{U} = \{\mathbf{i} + 2\mathbf{j} + 3\mathbf{k}; -\mathbf{i} + \mathbf{j} - \mathbf{k}\}$. На рис. 4 приведен график процесса углового согласования этих КТС, проведённого в соответствии с описанным выше алгоритмом.

Пример углового согласования ГТО для некорректной задачи. Распознаваемый и эталонный ГТО заданы КТС $\mathbf{V} = \{2\mathbf{i} - 4\mathbf{j} + \mathbf{k}; -\mathbf{i} + \mathbf{j} - 3\mathbf{k}\}$, $\mathbf{U} = \{\mathbf{i} + 3\mathbf{j} - \mathbf{k}; -4\mathbf{i} - \mathbf{j} + \mathbf{k}\}$. На рис. 5 представлен график процесса углового согласования ГТО по описанному алгоритму.

Как видно из приведённых графиков, для изображений ГТО, отличающихся друг от друга лишь параметрами вращения, достигается полное угло-

вое согласование. Если же ГТО разные, то полученный алгоритм минимизирует величину угла между ними на уровне $\Phi_{min} > 0$. Для рассматриваемого примера $\Phi_{min} \approx 18,5^\circ$.

Заключение

В статье показана возможность минимизации итерационным путём угла между 3D изображениями групповых точечных объектов, заданных в виде КТС. С этой целью выполняются пошаговые повороты одного из КТС вокруг усреднённой нормали к собственным плоскостям парциальных векторов сигнала. Углы поворота выбираются равными значениям текущих углов между КТС.

Литература

[1] *Фурман Я. А., Кревецкий А. В.* Комплекснозначные и кватернионные сигналы и подходы к их обработке — Известия высших учебных заведений. Приборостроение, 2009. — Т. 49,№. 4 — С. 7–18.

[2] *Арнольд В. И.* Геометрия комплексных чисел, кватернионов и спинов — Москва: МЦНМО, 2002. — С. 40.

[3] *Ефимов Н. В., Розендорн Э. Р.* Линейная алгебра и многомерная геометрия — Москва: Наука. 1974. — 528 с.

[4] *Роженцов А. А., Хафизов Р. Г., Егошина И. Л., Хафизов Д. Г.* Оценка трудоемкости определения параметров вращений трехмерных объектов — Вестник Марийского технического университета. Радиотехнические и инфокоммуникационные системы, 2008 — №. 3. — С.51–61.

[5] *Фурман Я. А., Егошина И. Л.* Обратная задача вращения трехмерных векторных сигналов — Автометрия, 2010. — Т. 46., №. 1. — С.46–56.

Обнаружение точек на контурах теней объекта, сопряжённых с точками на его поверхности

Фурман Я. А., Ерусланов Р. В.
krtmbs@marstu.net
Йошкар-Ола, Марийский государственный технический университет

Рассмотрена задача восстановления координат 3D объекта по параллельным проекциям в виде его теней. Показано существование проходящей через эти проекции экстремальной линии, на которой расположены необходимые для восстановления сопряжённые точки. Разработан алгоритм обнаружения находящихся на контурах теней сопряжённых точек.

Постановка задачи

Цель доклада заключается в получении полезной информации из изображений теней объекта для вычисления координат расположенных на его поверхности точек. Предполагается, что 3D объект W находится на плоской поверхности P, освещаемой источником света, расположенным в точке полюса S_1. Он удалён от объекта W настолько, что падающие на объект лучи можно считать параллельными. За объектом на поверхности P образуется его проекция в виде изображения тени W_1, имеющей постоянный уровень яркости. Аналогично, когда источник света оказывается в полюсе S_2, на поверхности P формируется изображение тени W_2 этого же объекта и т. д. Угловые положения полюсов S_n заданы направляющими векторами $\mathbf{d}_n$ линий визирования (проецирования) объекта W, $n = 1, 2, \ldots$, (рис. 1). Необходимо найти координаты точек на его поверхности по заданной серии теней $\{W_n\}$ и направляющих векторов $\{\mathbf{d}_n\}$ линий визирования, $n = 1, 2, \ldots$. Рассматриваемая задача связана с обнаружением и распознаванием антропогенных объектов на подстилающей 3D поверхности. Ключевая проблема при решении подобных задач заключается в нахождении на изображениях W_1 и W_2 сопряжённых точек (СТ) $\mathbf{w}_1(m)$ и $\mathbf{w}_2(m)$, являющих идентичными точке $\mathbf{w}(m) \in W$. Основные известные подходы к обнаружению СТ на изображениях проекций объекта следующие: отождествление точек по уровню яркости, поиск площадок с близкими распределениями яркостей (корреляционные методы), контурные методы (поиск в местах с резкими изменениями яркости) и поиск дополнительных СТ интерполяцией по линии контура [1]. Этим методам присущ ряд недостатков: нет гарантированной однозначности восстановления точки $\mathbf{w}(m) \in W_1$, отсутствие теоретической сходимости процесса поиска СТ, сложность оценки размеров коррелируемых областей операций [2]. Рассмотрим один из простых и достаточно надежных подходов к получению СТ на изображениях теней W_n, $n = 1, 2, \ldots$. Он базируется на геометрических соотношениях, представленных на рис. 1, и не использует корреляционные связи между проекциями.

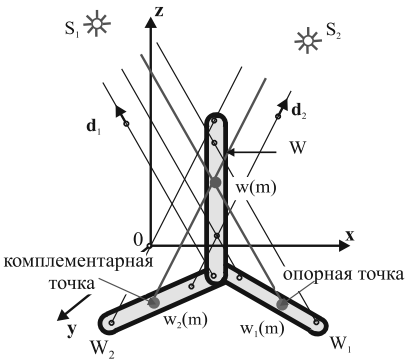


Рис. 1. Формирование теней W_1 и W_2 объекта W.

Экстремальная линия

При поиске СТ на одной из проекций, например, W_1, выбирают из тех или иных соображений точку $\mathbf{a}$, называемую опорной. Вторая СТ $\mathbf{b} \in W_2$ называется комплементарной. Как видно из рис. 2а, визирующие лучи $\mathbf{d}_1$ и $\mathbf{d}_2$, выходящие соответственно из полюсов S_1 и S_2, пересекаются в точке $\mathbf{w}_{ab}$ на поверхности объекта W. Далее они пересекают плоскость XOY в точках $\mathbf{a} \in W_1$ и $\mathbf{b} \in W_2$. Эти точки являются искомыми СТ для точки $\mathbf{w}_{ab} \in W$. Проходящие через эти три точки лучи $\mathbf{d}_1$ и $\mathbf{d}_2$ образуют плоскость H.

Исследуем свойства линии пересечения E_{ab} плоскостей H и XOY, включающей СТ $\mathbf{a}$ и $\mathbf{b}$. Параллельные лучи с направляющими векторами $\mathbf{d}_2$ проецируют на линию E_{ab} каждую точку визирующего луча $\mathbf{d}_1$. Отсюда следует, что часть линии в пределах тени W_1 служит геометрическим местом сопряжённых точек, комплементарных к опорной точке $\mathbf{a}$ для всех внутренних точек тела W по лучу $\mathbf{d}_1$. Остальная часть точек линии E_{ab} вместе с опорной точкой $\mathbf{a}$ задаёт 3D точки за пределами объекта W. Поскольку яркость каждой точки теней W_1 и W_2 одинакова, задача поиска на тени W_2 комплементарной точки $\mathbf{b}$ не имеет однозначного решения. Линию E_{ab} далее будем называть экстремальной [1]. Она образована пересечением двух плоскостей и поэтому является прямой

[1] В стереовидении вводится прямая линия, называемая эпиполярой, содержащая две СТ [3]. Рассматриваемая в докладе экстремальная линия в отличие от эпиполяры содер-

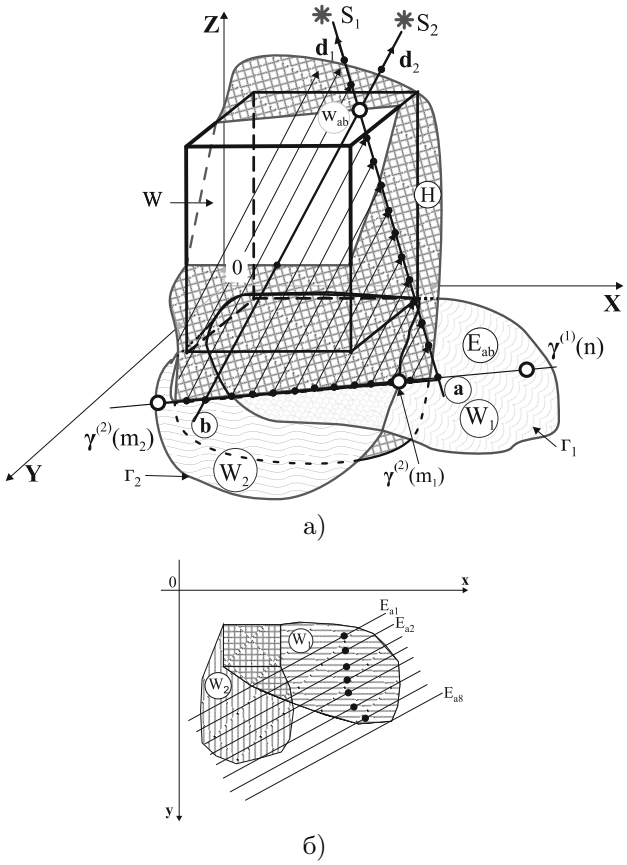


Рис. 2. Экстремальная линия а) векторная диаграмма, б) вид поля экстремальных линий.

линией. Направляющий вектор $\mathbf{r}_E$ этой линии равен

$$\mathbf{r}_E = [\mathbf{r}_H, \mathbf{r}_{XOY}],$$

где $\mathbf{r}_H = [\mathbf{d}_1, \mathbf{d}_2]$ — нормаль к плоскости H, а $\mathbf{r}_{XOY} = \mathbf{k}$ — нормаль к плоскости XOY. Отсюда следует: 1) линия проходит через опорную точку $\mathbf{a} \in W_1$; 2) её угловой коэффициент зависит только от направляющих векторов $\mathbf{d}_1$ и $\mathbf{d}_2$ проецирующих лучей; 3) экстремальные линии на плоскости XOY образуют поле параллельных прямых (рис. 2б).

Анализ контура изображения тени

Переход от тени к освещённому источником S фону сопровождается интенсивным скачком яркости. Поэтому её края хорошо очерчены и рассматриваемые в качестве сигнала характеризуются высоким отношением сигнал/шум. При выделении и прослеживании контура Γ тени приемлемые результаты даёт алгоритм Розенфельда [3, 4]. Он позволяет устойчиво, без зацикливания прослеживать контуры теней сколь угодно сложной формы. Квадратную сетчатку с изображением контура целесообразно рассматривать как комплексную плоскость и кодировать элементарные векторы контура комплексными числами $\gamma(n) = \gamma_1(n) + i\gamma_2(n)$, $n = 0, 1, \ldots, s-1$, где s — размерность контура [3]. Контур $\Gamma = \{\gamma(n)\}_{n=0}^{s-1}$ содержит сведения о форме объекта W, полученные в результате стопроцентной модуляции объектом светового потока. При параллельном проектировании объекта прямая линия на его поверхности отображается прямолинейным участком контура тени, а фрагмент в виде угла, образованного двумя прямыми — аналогичным по форме фрагментом контура. В условиях слабой зашумлённости линии контура подобные фрагменты уверенно выделяются обнаружителями на основе контурных согласованных фильтров [3]. Для обнаружения прямолинейных фрагментов инвариантно к их ориентации таким фильтром будет фильтр скользящего среднего:

$$\eta(m) = \frac{1}{t}\sum_{r=m}^{m+t-1}\gamma(r), \quad m = 0, 1, \ldots, s-1,$$

а для фрагментов в виде углов — фильтр скользящей разности (ФСР):

$$\eta(m) = \frac{1}{t}\left[\sum_{r=m}^{0,5t-1+m}\gamma(r) - \sum_{r=0,5t+m}^{t-1+m}\gamma(r)\right], \quad m = 0, 1, \ldots, s-1.$$

В этих выражениях $\eta(m)$ — выходной сигнал фильтра, t — ширина окна. При обнаружении фрагментов инвариантно к их ориентации решение принимается по результатам сравнения модуля выходного сигнала с порогом.

4. Восстановление координат точек на поверхности объекта W

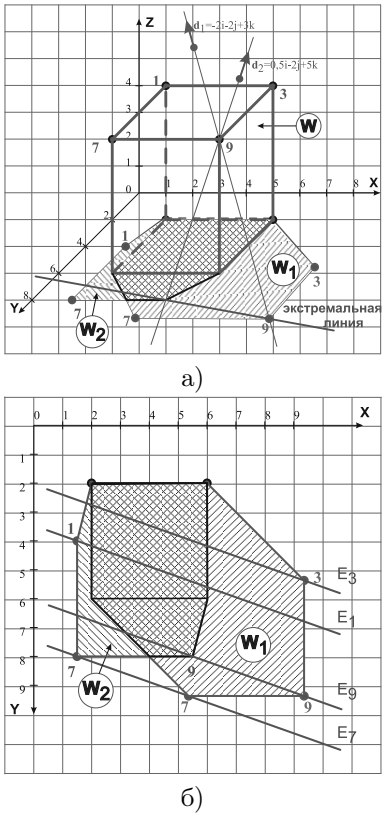
В предыдущем разделе было показано, что если опорная точка **a** расположена внутри изображения тени, то однозначное определение комплементарной точки **b** становится невозможным. В контурах $\Gamma_1 \in W_1$ и $\Gamma_2 \in W_2$ содержится сравнительно небольшое количество СТ, причём часть из них доступна для обнаружения. Такие точки расположены в вершинах углов контура и находятся на одной экстремальной линии E. Алгоритм обнаружения таких СТ реализуется следующим образом:

1) выделяются контуры Γ_1 и Γ_2 проекций W_1 и W_2 (рис. 3);

2) по результатам контурной согласованной фильтрации обнаруживаются фрагменты в форме углов и оцениваются координаты $\gamma^{(1)}(n)$, $n = 0, 1, \ldots$ и $\gamma^{(2)}(m)$, $m = 0, 1, \ldots$ вершин углов;

3) вычисляются параметры экстремальной линии, проходящей через вершину $\gamma^{(1)}(n)$ контура Γ_1 и фиксируется точка $\gamma^{(2)}(m)$ пересечения линия с контуром Γ_2; точки $\gamma^{(1)}(n)$ и $\gamma^{(2)}(m)$ принимаются в качестве сопряжённых при выполнении условия

жит бесконечное количество СТ, что объясняется одинаковой яркостью всех точек теней.



а) б)

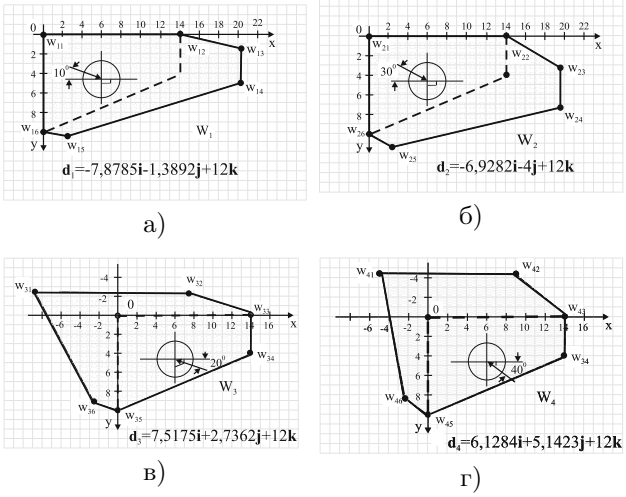
Рис. 3. К обнаружению СТ: а) объект W и его тени, б) поля диспарантности и экстремальных линий объекта W.

$$\Delta_{nm} \leqslant \Delta_0,$$

где $\Delta_{nm} = \left|\gamma_n^{(2)} - \gamma_m^{(2)}\right| = \min\limits_m$ — минимальное расстояние между точкой $\gamma_n^{(2)}$ и вершинами контура Γ_2, Δ_0 — пороговое значение Δ_{mn}, определяемое допустимой ошибкой восстановления координат 3D точки. Как видно из рис. 3б, экстремальная линия E_7 проходит через вершину 7, а линия E_9 — через вершину 9 контура Γ_2. Поэтому пары точек 7-7 и 9-9 будут сопряжёнными точками. Экстремальная линия E_3 не проходит ни через одну из вершин контура Γ_2 и поэтому на этом контуре соответствующая комплементарная точка отсутствует.

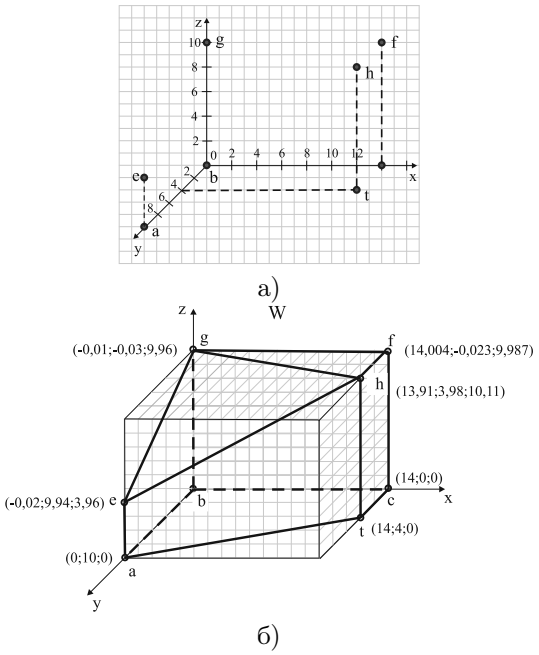
На рис. 4 показаны изображения теней W_1, W_2, W_3 и W_4 некоторого многогранника W и приведены выражения для направляющих векторов $\mathbf{d}_1$, $\mathbf{d}_2$, $\mathbf{d}_3$ и $\mathbf{d}_4$ соответствующих проецирующих лучей. На рис. 5а представлены восстановленные точки вершин многогранника W, а на рис. 5б — реконструированное в соответствии с [5] изображение самого многогранника.

Незначительное количество СТ, определяемых в контурах теней, должно компенсироваться большим количеством теней. Учитывая, что получение теней объекта не требует специального оборудования, а также простоту вычисления СТ, данное требование в ряде случаев не вызывает затруднений.



а) б) в) г)

Рис. 4. Изображения теней объекта W.



а) б)

Рис. 5. Восстановленные координаты точек и реконструкция изображения многогранника W: а) вершины многогранника, б) изображение многогранника.

Дополнительное количество СТ может быть получено путём интерполяции имеющихся СТ. Наиболее просто такая процедура реализуется, если объект W получен пересечением ряда плоскостей, т. е. является многогранником.

Заключение

Данная работа посвящена использованию полезной информации об объекте, содержащейся в изображениях его теней. Показана возможность восстановления координат точек на поверхности объекта по сопряжённым точкам, расположенных на контурах теней и обнаруживаемых при структурном анализе этих контуров на основе согласо-

ванной фильтрации. Метод обладает наибольшей эффективностью при восстановлении изображений объектов многогранной формы и для обнаружения сопряжённых точек не использует корреляционных связей между изображениями теней объекта.

Литература

[1] *Хорн Б. К. П.* Зрение роботов — Москва: Мир, 1989. — 487 с.

[2] *Потапов А. А.* Новейшие методы обработки изображений — Москва: ФИЗМАТЛИТ, 2008. — 496 с.

[3] *Фурман Я. А.* Введение в контурный анализ и его приложения к обработке изображений и сигналов — Москва: ФИЗМАТЛИТ, 2003. — 592 с.

[4] *Розенфельд, А.* Распознавание и обработка изображений — Москва: Мир, 1972. — 232 с.

[5] *Фурман Я. А.* Проволочная модель пространственного группового точечного объекта — Автометрия, 2008. — №. 3. — С. 3—16.

Метод оценки параметров вращения пространственного-группового точечного объекта*

Хафизов Д. Г.
HafizovDG@marstu.net
Йошкар-Ола, Марийский государственный технический университет

Предложен метод оценки параметров вращения изображений пространственных групповых точечных объектов на основе применения метода главных компонент. Применение данного метода позволяет получить оценку параметров вращения пространственного группового точечного объекта при отсутствии информации о порядке нумерации точек в эталонном и распознаваемом объектах, что существенно упрощает решение задач обработки пространственных групповых точечных объектов.

Задачи обработки 3D изображений в виде групповых точечных объектов характерны для ряда систем распознавания образов. При этом надежность результатов обработки таких изображений зависит от выбранной модели описания точечного объекта. В работах [1, 2, 3] показано, что наиболее адекватной моделью представления таких объектов с позиции применяемого математического аппарата и наибольшей информативности меры схожести является представление в виде кватернионных сигналов (КТС). В этих работах было показано, что величина скалярного произведения таких сигналов, как мера схожести и базовая операция, применяемая при их обработке, не является инвариантной к параметрам вращения точечного объекта по отношению к эталонному. В качестве подходов к решению данной проблемы применялись различные методы совмещения изображений пространственных точечных объектов, что, по сути, сводится к оценке параметров вращений. Решение задачи оценки параметров вращения необходимо для возможности дальнейшей обработки пространственных групповых точечных объектов (ПГТО).

Задачам оценки параметров вращения или совмещения подобных объектов посвящён ряд работ и здесь можно выделить следующие методы: метод совмещения с последующим усреднением [1]; метод оценки параметров вращения на основе амплитудно-фазовых моделей [4]; метод оценки параметров вращения на основе сферических гармоник [2, 5]; метод оценки параметров вращения по результатам фильтрации [7] и двухэтапный метод оценки параметров вращения [6]. Предлагаемое в данной работе решение задачи оценки параметров вращения основано на известном методе, используемом в многомерном статистическом анализе данных — методе главных компонент [8, 9].

Пусть имеется пространственный групповой точечный объект, заданный в виде множества точек расположенных в трёхмерном пространстве $\boldsymbol{\Xi} = \{\boldsymbol{\xi}(n)\}_{0,s-1}$, где $\boldsymbol{\xi}(n) = (\xi_1(n), \xi_2(n), \xi_3(n))$ — пространственные координаты n-й точки; а также ПГТО $\boldsymbol{\Omega} = \{\boldsymbol{\omega}(m)\}_{0,s-1}$, где $\boldsymbol{\omega}(m) = (\omega_1(m), \omega_2(m), \omega_3(m))$, который отличается от исходного тем, что повернут относительно некоторой неизвестной оси (вектора) $\boldsymbol{\rho}$ на неизвестный угол 2ψ. Также следует отметить, что порядок нумерации точек ПГТО $\boldsymbol{\Omega}$ не совпадает с порядком нумерации в ПГТО $\boldsymbol{\Xi}$. Требуется оценить параметры вращения ПГТО $\boldsymbol{\Omega}$ относительно $\boldsymbol{\Xi}$.

Применение метода главных компонент для анализа ПГТО

Анализ главных компонент — это метод преобразования одной последовательности наблюдаемых переменных в другую последовательность переменных. Он заключается в получении новых показателей — главных компонент, являющихся линейными комбинациями исходных. Главные компоненты упорядочиваются в порядке убывания той дисперсии, которую они «объясняют» [8, 9]. Метод главных компонент осуществляет переход к новой системе координат $y_1, \ldots, y_p$ в исходном пространстве признаков $x_1, \ldots, x_p$, которая является системой ортонормированных линейных комбинаций. Линейные комбинации выбираются таким образом, что среди всех возможных линейных нормированных комбинаций исходных признаков первая главная компонента $y_1(x)$ обладает наибольшей дисперсией. Геометрически это выглядит как ориентация новой координатной оси y_1 вдоль направления наибольшей вытянутости эллипсоида рассеивания объектов исследуемой выборки в пространстве признаков $x_1, \ldots, x_p$. Вторая главная компонента имеет наибольшую дисперсию среди всех оставшихся линейных преобразований, некоррелированных с первой главной компонентой. Она интерпретируется как направление наибольшей вытянутости эллипсоида рассеивания, перпендикулярное первой главной компоненте и т. д.

Вычисление коэффициентов главных компонент основано на определении собственных векторов $\boldsymbol{w}_1 = (w_{1,1}, \ldots, w_{p,1})^T, \ldots, (w_{1,p}, \ldots, w_{p,1})^T$ ковариационной матрицы $\boldsymbol{S} = \{s_{i,j}\}$ полученной

Работа выполнена при финансовой поддержке РФФИ, проекты № 11-07-00585-а, 10-01-00445-а, и программы «Развитие научного потенциала высшей школы», проекты 2.1.2/2204 и 2.1.2/10218.

на основе координат точек объекта

$$s_{i,j} = \frac{1}{s}\sum_{k=0}^{s-1}(\xi_i(k) - \bar{\xi}_i)(\xi_j(k) - \bar{\xi}_j), \qquad (1)$$

где $i, j = 1, 2, 3$.

Таким образом, для определения главных компонент ПГТО необходимо определить собственные вектора $\boldsymbol{w}_1 = (w_{1,1}, w_{2,1}, w_{3,1})$, $\boldsymbol{w}_2 = (w_{1,2}, w_{2,2}, w_{3,2})$ и $\boldsymbol{w}_3 = (w_{1,3}, w_{2,3}, w_{3,3})$ ковариационной матрицы (1).

Выполнив пересчёт координат исходного ПГТО по формуле [11]

$$\boldsymbol{\varphi}_{\boldsymbol{\Xi},j}(n) = \sum_{i=1}^{3} w_{i,j}(\xi_i(n) - \bar{\xi}_i), j = 1, 2, 3, \qquad (2)$$

получим изображение ПГТО $\boldsymbol{\Phi}_{\boldsymbol{\Xi}} = \{\boldsymbol{\varphi}_{\boldsymbol{\Xi}}(n)\}_{0,s-1}$, где $\boldsymbol{\varphi}_{\boldsymbol{\Xi}}(n) = (\varphi_{\boldsymbol{\Xi},1}(n), \varphi_{\boldsymbol{\Xi},2}(n), \varphi_{\boldsymbol{\Xi},3}(n))$ в новых координатных осях, причём данное изображение по определению не будет зависеть от исходного положения ПГТО.

Выделим основные свойства, связанные с использованием собственных значений и собственных векторов.

1. Собственные значение линейного оператора выражают его свойства и их значения не зависят от используемой системы координат [12]. Это означает, что и для исходного и повернутого ПГТО собственные значения ковариационной матрицы будут одинаковы.
2. Использование собственных векторов, означает использование в качестве координатных базисных векторов, направление которых зависит только от конфигурации группового точечного объекта.
3. Совокупность собственных векторов образует невырожденную модальную матрицу $\boldsymbol{T}$, определитель которой равен 1, и с помощью которой задаётся преобразование координат, т. е. является оператором вращения $p = \boldsymbol{T}q$ [2]; здесь q и p — исходная и повернутая точка соответственно, причём $q = \boldsymbol{T}^{-1}p$.

Решение задачи оценки параметров вращения

Как было показано, метод главных компонент позволяет получить представление ПГТО инвариантное к параметрам вращения, т. е.

$$\boldsymbol{\Phi}_{\boldsymbol{\Xi}} = \left\{ \begin{pmatrix} \boldsymbol{\varphi}_{\boldsymbol{\Xi},1}(n) \\ \boldsymbol{\varphi}_{\boldsymbol{\Xi},2}(n) \\ \boldsymbol{\varphi}_{\boldsymbol{\Xi},3}(n) \end{pmatrix} \right\}_{0,s-1} = \left\{ \boldsymbol{T}_{\boldsymbol{\Xi}} \begin{pmatrix} \boldsymbol{\xi}_1(n) \\ \boldsymbol{\xi}_2(n) \\ \boldsymbol{\xi}_3(n) \end{pmatrix} \right\}_{0,s-1},$$

где $\boldsymbol{T}_{\boldsymbol{\Xi}}$ — оператор вращения, представляющий собой невырожденную модальную матрицу, сконструированную из совокупности собственных векторов ковариационной матрицы $\boldsymbol{S}$.

Для ПГТО Ω аналогично получим

$$\boldsymbol{\Phi}_{\boldsymbol{\Omega}} = \left\{ \boldsymbol{T}_{\boldsymbol{\Omega}} \begin{pmatrix} \boldsymbol{\omega}_1(n) \\ \boldsymbol{\omega}_2(n) \\ \boldsymbol{\omega}_3(n) \end{pmatrix} \right\}_{0,s-1},$$

причём $\boldsymbol{\Phi}_{\boldsymbol{\Xi}} \equiv \boldsymbol{\Phi}_{\boldsymbol{\Omega}}$, здесь тожественное равенство означает, что $\boldsymbol{\Phi}_{\boldsymbol{\Xi}}$ равен $\boldsymbol{\Phi}_{\boldsymbol{\Omega}}$ с точностью до порядка нумерации точек в ПГТО.

Таким образом, для оценки параметров вращения ПГТО $\boldsymbol{\Omega}$ относительно ПГТО $\boldsymbol{\Xi}$ при неизвестном порядке нумерации точечных отсчетов в объекте можно записать:

$$\left\{ \begin{pmatrix} \boldsymbol{\xi}_1(n) \\ \boldsymbol{\xi}_2(n) \\ \boldsymbol{\xi}_3(n) \end{pmatrix} \right\}_{0,s-1} \equiv \left\{ \boldsymbol{T}_{\boldsymbol{\Xi}}{}^{-1}\boldsymbol{T}_{\boldsymbol{\Omega}} \begin{pmatrix} \boldsymbol{\omega}_1(n) \\ \boldsymbol{\omega}_2(n) \\ \boldsymbol{\omega}_3(n) \end{pmatrix} \right\}_{0,s-1};$$

здесь $\boldsymbol{T} = \boldsymbol{T}_{\boldsymbol{\Xi}}{}^{1}\boldsymbol{T}_{\boldsymbol{\Omega}}$ — матрица вращений, являющаяся, по сути, оценкой параметров вращений ПГТО $\boldsymbol{\Omega}$ относительно $\boldsymbol{\Xi}$.

В том случаи, если ПГТО задан кватернионным сигналом, то из матрицы $\boldsymbol{T}$ можно получить оценку параметров вращений в виде вращающего кватерниона, т. е. если

$$\boldsymbol{\Xi} = \{\boldsymbol{\xi}(n)\}_{0,s-1} = \{\xi_1(n)\boldsymbol{i} + \xi_2(n)\boldsymbol{j} + \xi_3(n)\boldsymbol{k}\}_{0,s-1},$$

и

$$\boldsymbol{\Omega} = \{\boldsymbol{\omega}(n)\}_{0,s-1} = \{\omega_1(n)\boldsymbol{i} + \omega_2(n)\boldsymbol{j} + \omega_3(n)\boldsymbol{k}\}_{0,s-1},$$

при условии, что $\boldsymbol{\Xi} \equiv \boldsymbol{b}\boldsymbol{\Omega}\boldsymbol{b}^{-1} = \{\boldsymbol{b}\boldsymbol{\omega}(n)\boldsymbol{b}^{-1}\}_{0,s-1}$, где $\boldsymbol{b} = b_0 + b_1\boldsymbol{i} + b_2\boldsymbol{j} + b_3\boldsymbol{k}$ — вращающий кватернион, т. е. $\boldsymbol{b} = cos(\psi) + sin(\psi)(\rho_1\boldsymbol{i} + \rho_2\boldsymbol{j} + \rho_3\boldsymbol{k})$, то, согласно [2], получим

$$2\psi = \arccos\left(\frac{t_{0,0} + t_{1,1} + t_{2,2} - 1}{2}\right),$$

$$\rho_1 = \frac{t_{1,2} - t_{2,1}}{2\sin(2\psi)}; \rho_2 = \frac{t_{2,0} - t_{0,2}}{2\sin(2\psi)}; \rho_3 = \frac{t_{0,1} - t_{1,0}}{2\sin(2\psi)},$$

здесь $t_{i,j}$ — элементы матрицы $\boldsymbol{T}$.

Заключение

Использование метода главных компонент, основанного на свойствах характеристических векторов ковариационной матрицы, рассчитанной по координатам точек пространственного группового точечного объекта, позволяет получить оценку параметров вращений пространственного группового точечного объекта. При этом достоинством метода является то, что порядок следования точек в исходном и повернутом ПГТО может не совпадать. Данное обстоятельство позволяет упростить решение таких задач обработки пространственных групповых точечных объектов, как обнаружение и распознавание ПГТО, так как не требуется разрабатывать трудоемкую процедуру нумерации точек в объекте, устойчивую к воздействию шумов и к процедуре вращения ПГТО.

Литература

[1] *Фурман Я. А., Кревецкий А. В., Передреев А. К., Роженцов А. А., Хафизов Р. Г., Егошина И. Л., Леухин А. Н.* Введение в контурный анализ и его приложения к обработке изображений и сигналов. — 2-е изд., испр. — Москва: ФИЗМАТЛИТ, 2003. — 592 с.

[2] *Фурман Я. А., Кревецкий А. В., Роженцов А. А., Хафизов Р. Г., Леухин А. Н., Егошина И. Л.* Комплекснозначные и гиперкомплексные системы в задачах обработки многомерных сигналов. — Москва: ФИЗМАТЛИТ, 2004.— 456 с.

[3] *Фурман Я. А., Хафизов Д. Г.* Распознавание групповых точечных объектов в трехмерном пространстве // Автометрия. — 2003. — Т. 39, № 1. — С. 3–18.

[4] *Хафизов Д. Г.* Упорядочение точек пространственного изображения группового точечного объекта на базе амплитудно-фазового представления // Автометрия. — 2007. — Т. 43, № 1. — С. 10–23.

[5] *Леухин А. Н.* Оценка параметров вращений трехмерного группового точечного объекта без предварительной нумерации формирующих точек // Математические методы распознавания образов: Доклады 11-й Всероссийской конфе-ренции. — Москва, 2003. — С. 130–133.

[6] *Фурман Я. А.* Обратная задача вращения трехмерных векторных сигналов // Автометрия.— 2010. — Т. 46, № 1. — С. 46–56.

[7] *Фурман Я. А.* Нахождение параметров вращения пространственного группового точечного объекта по результатам его фильтрации // Радиотехника и электроника. — 2008. — Т. 53, № 1. — С. 86–97.

[8] *Дронов С. В.* Многомерный статистический анализ: учебное пособие. — Барнаул: Изд-во Алт. гос. ун-та, 2003. — 213 с.

[9] *Калинина В. Н.* Введение в многомерный статистический анализ: учебное пособие. — ГУУ. — Москва, 2003. — 66 с.

[10] *Кантор И. Л.* Гиперкомплексные числа. — Москва: Наука, 1973.

[11] *Хафизов Д. Г.* Получение аналитического представления формы пространственного группового точечного объекта // Вестник МарГТУ Радиотехнические и инфокоммуникационные системы. — 2008. — № 2. — С. 35–41.

[12] *Корн Г.* Справочник по математике для научных работников и инженеров. — Москва: Наука, 1968. — 720 с.

Применение ПЛИС в решении задачи распознавания изображений пространственных объектов с неупорядоченными отсчётами*

Баев А. А., Роженцов А. А.
krtmbs@marstu.net
Йошкар-Ола, Марийский государственный технический университет

В данной работе представлен специализированный процессор для решения задачи распознавания изображений пространственных объектов, представленных кватернионными моделями. Для проверки работоспособности устройства использована отладочная плата Nexys II на базе ПЛИС Spartan3E-1200.

Введение

Решаемые современными системами технического зрения задачи требуют перехода от плоских сцен к анализу пространственных изображений ввиду их большей информативности, появлению относительно недорогих систем 3D сканирования, а также росту вычислительной мощности систем обработки. Задачи, решаемые системами технического зрения, зависят от конкретного их применения. Например, в системах автономной навигации большую роль играет выделение плоских поверхностей, а также распознавание плоских и пространственных объектов. В настоящее время разработано множество методов распознавания пространственных объектов, в том числе с неупорядоченными отсчётами, однако, остаётся проблема технической реализации. Создание интегральных схем и конечных автоматов распознавания для этих методов является очень трудоемкой задачей. Одним из актуальных направлений их реализации является разработка процессоров и контроллеров на программируемых логических интегральных схемах (ПЛИС) [1]. В этом случае разработанный процессор представляет собой проект, свойства которого определяются разработчиком. Однако специфика ПЛИС заключается в том, что реализованные с их помощью цифровые устройства имеют меньшую тактовую частоту и более высокую стоимость, чем аналогичные специализированные микросхемы. Поэтому прямой перенос в ПЛИС одного из существующих микропроцессоров не всегда оправдан с технико-экономической точки зрения. При разработке синтезируемых ядер необходимо руководствоваться соображениями повышения функциональных возможностей проектируемого процессора, аппаратного распараллеливания операций и реализации в одном корпусе всей совокупности устройств «процессор–память–периферия».

В данной работе представлен специализированный процессор на ПЛИС для решения задачи распознавания изображений пространственных объектов, позволяющий выполнять математические действия с кватернионами.

Работа выполнена при финансовой поддержке РФФИ, проекты №10-01-00445-а, №11-07-00585-а, и по программе «Развитие научного потенциала высшей школы», проекты 2.1.2/2204 и 2.1.2/10218

Метод распознавания

Как показано в работе [2], для обработки объёмных изображений могут использоваться методы кватернионного анализа. В этом случае векторы, проведённые в пространстве к точкам, задающим поверхность объекта, описываются векторными кватернионами, а их набор представляет собой кватернионный сигнал. В работе [3] показано, что использование аппарата кватернионного анализа позволяет связать поверхность, заданную в пространстве, с функцией кватернионного переменного, например, отображающей её отсчёты на сферу. Для этого применяется полиномиальная функция вида:

$$\sum_{m=0}^{M-1}(q_n^m * a_m) = p_n, \qquad (1)$$

где a_m — коэффициенты полинома, также являющиеся кватернионами, задающие отображение пространственной фигуры на поверхность сферы, q_n — кватернионы, соединяющие точки поверхности объекта с началом координат, p_n — проекции кватернионов q_n на сферу.

Формула (1) позволяет вычислить коэффициенты полинома a, связывающего поверхность исследуемого объекта с поверхностью сферы. При использовании метода наименьших квадратов, для вычисления коэффициентов полинома степени M, следует решить систему линейных кватернионных уравнений (СЛКУ), элементы которой определяется из соотношений (2):

$$\boldsymbol{q_{r,m}} = \sum_{n=0}^{N-1}(\overline{q_n^r} * q_n^m),\ \boldsymbol{p_r} = \sum_{n=0}^{N-1}(\overline{q_n^r} * p_n), \qquad (2)$$

где $r = 0, \ldots, M-1,\ m = 0, \ldots, M-1,\ \ M-1$ — степень полинома, N — количество элементов исходного сигнала.

Решение данной системы уравнений, например, методом Гаусса, позволяет найти значения коэф-

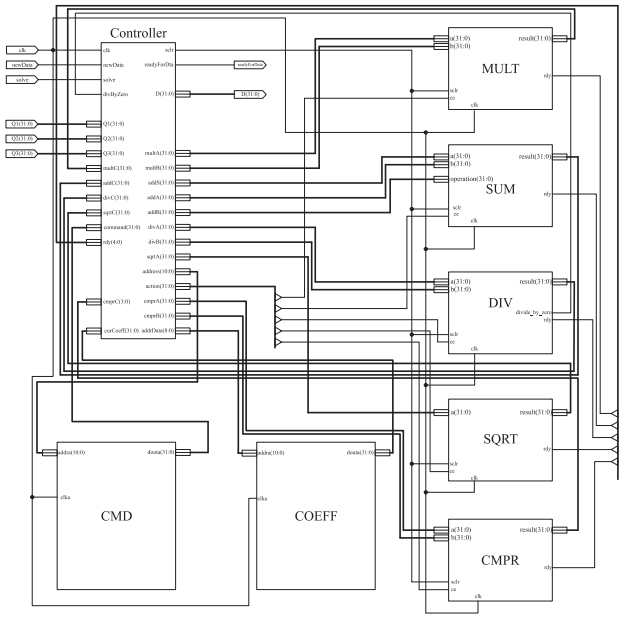


Рис. 1. Функциональная схема процессора.

фициентов a_m полиномиальной функции, выполняющей отображение пространственной фигуры на сферу.

Описание процессора на ПЛИС

Разработанное устройство представляет собой специализированный процессор, содержащий блоки работы с математическими функциями, блоками постоянной и разделяемой памяти.

На рис. 1 изображена функциональная схема процессора. В его состав входят: умножитель (MULT), сумматор (SUM), делитель (DIV), блок извлечения корня (SQRT), компаратор (CMPR), память команд (CMD), банк эталонных коэффициентов (COEFF) и блок управления (Controller). Блоки АЛУ используют встроенный в ПЛИС DSP. Блоки памяти команд и банк эталонов построены на основе распределённой памяти.

Также процессор содержит ряд выводов, где: clk — тактовый синхроимпульс; newData — флаг, используемый при формировании СЛКУ, указывающий на то, что на входе присутствует новый набор данных; solve — флаг начала вычисления СЛКУ, выставляется источником сигнала, когда все данные переданы в процессор; Q1, Q2, Q3 — шины данных, соответствуют компонентам i, j, k входного кватерниона; readyForData — флаг готовности принять новый набор данных; D0 — выходная шина, содержащая индекс наиболее похожего эталона.

Основным является блок управления, алгоритм его работы описан на языке VHDL и включает четыре основных части:

1) вычисление уникальных элементов СЛКУ;
2) решение СЛКУ;
3) сравнение полученных коэффициентов полинома с коэффициентами из банка эталонов;
4) управление блоками математических функций и блоками памяти.

Таблица 1. Список поддерживаемых команд.

Название	Аналитическая запись	Нех -код	A	B	C	Количество тактов
Mult	$C = A * B$	01	+	+	+	10
Sum	$C = A + B$	02	+	+	+	16
Sub	$C = A - B$	03	+	+	+	16
Coef	$C = 1/B$	04	-	+	+	30
Sqrt	$C = \sqrt{A}$	05	+	-	+	30
Inv	$C = -A$	06	+	-	+	2
Inc	$C = A + 1$	07	+	-	+	16
Cmpr	$C = A > B$	08	+	+	+	16

Задачей первых трёх частей является управление выбором адресов памяти команд. Следует отметить, что части 1 и 3 являются многоитерационными, однако, извне управляется только процесс вычисления уникальных компонент СЛКУ. Управление блоками математических функций производится при помощи набора команд. Каждая команда представляет собой 32-х битное слово, где первый байт содержит индекс математической операции, последующие два — индексы ячеек памяти, соответствующие исходным данным, последний — индеек ячейки памяти в которую необходимо поместить результат. Команды, поддерживаемые устройством, сведены в таблицу 1.

Как видно из таблицы 1, некоторые команды не требуют одного из входных параметров; в данном случае этот байт может принимать любое значение, так как при чтении команды не учитывается. Команда Cmpr возвращает наибольшее значение.

Общая трудоемкость для метода проецирования на сферу [4] при реализации на данном процессоре определяется как:

$$T = N * (26 * 10 + 30 * 16 + 30 + 30) +$$
$$+770 * 10 + 664 * 16 + 76 * 30 + 30 +$$
$$+K * (20 * 10 + 19 * 16 + 16) =$$
$$= N * 800 + 18324 + K * 520$$

При тактовой частоте 500 МГц, размерности эталона $N = 10000$ и количестве эталонов $K = 16$, время выполнения составит $t = 16$ мс.

Таким образом, показана возможность реализации метода проецирования на сферу на ПЛИС. Приведённый расчёт трудоемкости показал обоснованность выбора средства технической реализации.

Заключение

Разработанный специализированный процессор для решения задачи распознавания изображений пространственных объектов, позволяет выполнять, математические действия с кватернионами. Процессор выполнен с использованием ПЛИС. Для проверки работоспособности устройства использована отладочная плата Nexys II на базе ПЛИС Spartan3E-1200. Применение более мощных ПЛИС позволит добавить команды, выполняющие математические операции над подряд идущими блоками данных. Так, дополнение схемы тремя умножителями и тремя сумматорами снизит требуемое время вычислений в 2,5 раза.

Литература

[1] *Ивченко В. Г.* Применение языка VHDL при проектировании специализированных СБИС: Учебное пособие. — Таганрог: Издательство ТРТУ, 1999. — 80 с.

[2] *Фурман Я. А., Кревецкий А. В., Роженцов А. А. и др.* Комплекснозначные и гиперкомплексные системы в задачах обработки многомерных сигналов. — Москва: Наука, 2004. — 456 с.

[3] *Роженцов А. А., Баев А. А., Наумов А. С.* Оценка параметров и распознавание изображений трехмерных объектов с неупорядоченными отсчетами // Автометрия. — 2010. — Т. 46, № 1. — С. 57–69.

[4] *Роженцов А. А., Баев А. А., Ерусланов Р. В.* Решение задачи распараллеливания вычислений при обработке кватернионных сигналов // Вестник Марийского государственного технического университета. Радиотехнические и инфокоммуникационные системы. — 2011. — Т. 1, № 1. — С. 34–46.

Криволинейные скелеты трёхмерных форм*

Местецкий Л. М., Хромов Д. В.
l.mest@ru.net, denis.v.khromov@gmail.com
Московский государственный университет им. М. В. Ломоносова

В задачах анализа формы трёхмерных изображений часто используются криволинейные скелеты — пространственные графы, описывающие геометрию рассматриваемой фигуры. К настоящему моменту существует большое количество разнообразных способов для построения криволинейных скелетов, однако все эти подходы являются разнородными эвристиками, не имеющими теоретического обоснования. В данной работе предлагается математическая модель, позволяющая строго определять криволинейные скелеты и численно оценивать их соответствие форме исходного объекта. Подход основан на аппроксимации цилиндрических фрагментов трёхмерного объекта при помощи специальных примитивов — жирных кривых. На основе предлагаемой модели разработан и реализован алгоритм построения криволинейных скелетов, использующий двумерные скелеты плоских проекций.

Введение

Пусть $\tilde{\Omega}$ — открытое ограниченное n-мерное множество в $\mathbb{R}^n$ с границей $\partial\Omega$. Через Ω обозначим замыкание этого множества:

$$\Omega = \tilde{\Omega} \cup \partial\Omega.$$

Такое замыкание также будем называть фигурой.

Определение 1. *Серединной осью (medial axis) открытого множества $\tilde{\Omega}$ называется множество $\mathcal{M} \subset \Omega$ точек, для которых существует не менее двух ближайших точек на границе Ω:*

$$\mathcal{M} = \{\boldsymbol{a} \in \Omega | \exists \boldsymbol{x}, \boldsymbol{y} \in \partial\Omega : \boldsymbol{x} \neq \boldsymbol{y}, \rho(\boldsymbol{a}, \boldsymbol{x}) = \rho(\boldsymbol{a}, \boldsymbol{y}), \\ \forall \boldsymbol{z} \in \partial\Omega \rho(\boldsymbol{a}, \boldsymbol{z}) \geqslant \rho(\boldsymbol{a}, \boldsymbol{x})\}.$$

Определение 2. *Серединной осью фигуры Ω называется замыкание серединной оси открытого множества $\tilde{\Omega}$.*

Понятие серединной оси впервые было введено в [1]. Серединная ось n-мерной фигуры в общем случае содержит в себе многообразия размерности $(n-1)$. Между фигурой и её серединной осью всегда имеет место гомотопическая эквивалентность [3].

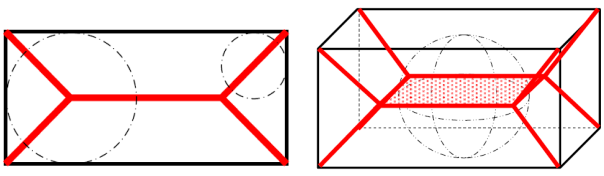


Рис. 1. Серединная ось двумерного прямоугольника и трёхмерного параллелепипеда.

Серединную ось двумерной фигуры можно рассматривать как плоскую укладку некоторого планарного графа. Рёбрам этого графа соответствуют кривые, образованные точками, для каждой из которых существуют ровно две ближайшие точки на границе фигуры; остальные точки серединной оси соответствуют вершинам графа. Такую укладку графа также называют скелетом фигуры. Скелет плоской фигуры хорошо схватывает основные метрические и топологические свойства фигуры. Поэтому плоские скелеты активно используются в задачах компьютерного зрения, т. к. извлекать из них признаковую информацию зачастую проще, чем из граничного описания фигуры.

В отличие от плоского случая, серединная ось трёхмерного объекта не является пространственной укладкой некоторого графа, т. к. может содержать в себе фрагменты двумерных поверхностей. Такой объект в общем случае является достаточно сложным, поэтому его практическое применение ограничено. Тем не менее, в ряде задач, особенно в тех, где анализируется форма объектов с трубчатыми, вытянутыми фрагментами, было бы очень удобно иметь инструмент, аналогичный плоскому скелету — некоторый пространственный граф, описывающий форму объекта. Объекты, соответствующие такому чисто интуитивному понятию, называют криволинейными скелетами (curve-skeletons).

К настоящему моменту существует большое количество публикаций, посвящённых криволинейным скелетам. Основные подходы к построению таких скелетов подробно рассмотрены в работах [4, 5]. Однако до сих пор отсутствует не только строгое определение криволинейного скелета, но и математический критерий, который по крайней мере позволил бы сравнивать между собой различные подходы к решению этой задачи, представляющие собой эвристики самой разнообразной природы.

В настоящей работе предлагается определение криволинейного скелета, в рамках которого можно описать различные способы для его построения; при этом определяется мера соответствия между скелетом и исходной фигурой, позволяющая проводить теоретически обоснованное сравнение различных конкретных способов задания криволинейных

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований, грант №11-01-00783.

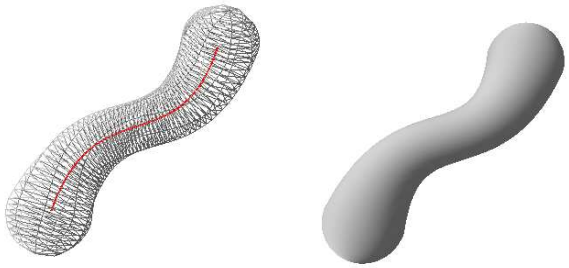


Рис. 2. Жирная кривая.

скелетов. Для демонстрации практической состоятельности предлагаемой модели в статье описан алгоритм, основанный на использовании критерия соответствия.

Жирные линии и циркуляры

Определение 3. *Пусть γ — гладкая кривая в $\mathbb{R}^n$, представляющая собой гомеоморфный образ отрезка $[0;1]$ и не имеющая особых точек. Пусть, кроме того, на ней задана неотрицательная гладкая функция r*

$$r\colon \gamma \to \mathbb{R}_+,$$

которая не обращается в нуль нигде, кроме, может быть, концевых точек γ[1]. Тогда жирной кривой с осью γ и радиальной функцией r называется множество точек

$$\mathcal{F}(\gamma, r) = \{\boldsymbol{a} \in \mathbb{R}^n | \exists \boldsymbol{x} \in \gamma : \rho(\boldsymbol{a}, \boldsymbol{x}) \leqslant r(\boldsymbol{x})\}.$$

Концевые точки кривой γ называются вершинами жирной кривой.

Жирная кривая (термин, предложенный в [2]) представляет собой объект, хорошо аппроксимирующий тела вытянутой, «цилиндрической» формы. При этом ось такой кривой соответствует интуитивному представлению о линии скелета такого тела (см. рис. 2). Поэтому удачной идеей представляется приближённое представление объекта произвольной формы в виде набора жирных кривых.

Определение 4. *Циркуляром $\mathcal{C}$ называется объединение конечного множества жирных кривых.*

Определение 5. *Графом связности циркуляра называется граф, множество вершин V которого совпадает с множеством вершин жирных кривых, образующих циркуляр, а две вершины v_1, v_2 этого графа соединены ребром тогда и только тогда, когда пространственный циркуляр содержит в себе жирную кривую с парой вершин v_1, v_2.*

Для графа смежности циркуляра естественным образом определяется вложение в $\mathbb{R}^n$ в виде объединения осей жирных кривых, составляющих циркуляр. Такое вложение будем называть осями циркуляра.

Криволинейный скелет трёхмерной фигуры

Аппроксимация фигуры циркуляром. Пространственный циркуляр представляет собой фигуру, составленную из набора примитивов (жирных линий). Пусть задана некоторая мера близости между фигурами в $\mathbb{R}^n$ (например, метрика Хаусдорфа; подробнее вопрос о выборе меры близости будет рассмотрен ниже), тогда можно поставить задачу об аппроксимации произвольной фигуры Ω циркуляром с заданной погрешностью.

Для циркуляра можно естественным образом определить его криволинейный скелет как множество его осей. Поэтому задача построения криволинейного скелета произвольной трёхмерной фигуры, рассмотренная во введении, может быть строго сформулирована как задача аппроксимации фигуры некоторым циркуляром. При этом формы циркуляра и исходной фигуры можно считать достаточно близкими (в смысле используемой метрики), и поэтому криволинейный скелет фигуры можно определить как множество осей аппроксимирующего циркуляра. Такая формулировка позволяет применять различные алгоритмы построения криволинейных скелетов, вместе с тем предоставляя возможность для численной оценки качества полученных скелетов — таковой оценкой является величина погрешности аппроксимации.

Определение 6. *Пусть Ω — фигура в $\mathbb{R}^n$, $\mathcal{C}$ — аппроксимирующий её циркуляр. Тогда оси циркуляра $\mathcal{C}$ называются криволинейным скелетом фигуры Ω.*

Мера близости между исходной фигурой и аппроксимирующим её циркуляром может быть определена различными способами и в значительной степени определяется решаемой задачей. Одним из возможных вариантов является использование метрики Хаусдорфа, хорошо зарекомендовавшей себя в задачах анализа и сравнения формы.

В реализации алгоритма, описываемого в настоящей статье, в качестве меры близости используется интегральное расстояние между границами фигуры $\partial\Omega$ и огибающей поверхностью циркуляра $\partial\mathcal{C}$:

$$\varepsilon(\Omega, \mathcal{C}) = \int_{\boldsymbol{x} \in \partial\Omega} \rho^2(\boldsymbol{x}, \partial\mathcal{C}) dS,$$

где

$$\rho(\boldsymbol{x}, \partial\mathcal{C}) = \min_{\boldsymbol{y} \in \partial\mathcal{C}} \rho(\boldsymbol{x}, \boldsymbol{y}).$$

Класс жирных кривых, в котором выполняется поиск решения, имеет большое значение.

[1] Т.е. образов точек 0 и 1.

Класс кривых, наряду с мерой близости, в значительной степени определяет как внешний вид получаемых скелетов, так и методы оптимизации, используемые для поиска оптимального (или локально оптимального) решения. В качестве примера достаточно мощного и полезного на практике класса жирных кривых можно привести кривые Безье порядка d. Каждая такая жирная кривая определяется набором из $(d+1)$ векторов $\boldsymbol{p}_i \in \mathbb{R}^n$ и $(d+1)$ неотрицательных чисел r_i:

$$\gamma(t) = \sum_{i=0}^{d} \boldsymbol{p}_i B_d^i(t),$$

$$r(t) = \sum_{i=0}^{d} r_i B_d^i(t),$$

где B_d^i — полином Бернштейна степени d.

Метод построения трёхмерных скелетов

Поиск аппроксимирующего циркуляра, для которого погрешность аппроксимации достигает своего минимума — сложная задача. В настоящем разделе предложен ряд простых эвристик, позволяющих найти достаточно точное решение.

Общая схема алгоритма построения аппроксимирующего циркуляра включает в себя два важных этапа: выбор некоторого циркуляра в качестве начального приближения и его последующая подгонка численными методами. Минимизация погрешности аппроксимации фигуры циркуляром является достаточно нетривиальной задачей, поэтому важно, чтобы начальное приближение циркуляра уже было качественным. В частности, задача численной оптимизации сильно упрощается, если граф смежности циркуляра полностью определён в начальном приближении и не меняется при итерациях численных методов.

Использование плоских проекций. Рассмотрим плоскую проекцию трёхмерной фигуры. Интуитивно понятно, что скелет этой двумерной проекции визуально схож с проекцией криволинейного скелета исходной трёхмерной модели. Плоская серединная ось является хорошо изученным объектом; для её построения существуют эффективные алгоритмы. Поэтому можно попытаться восстановить криволинейный скелет пространственной фигуры по плоскому скелету проекции (или нескольких проекций, см. [6]). Основная проблема, возникающая при таком подходе, обусловлена тем, что скелеты, восстановленные по различным проекциям, существенно различаются между собой, и достаточно трудно разрешить вопрос о том, какой именно из полученных скелетов является наилучшим. Однако наличие меры близости между циркуляром и фигурой позволяет численно оценить качество полученных скелетов и выбрать среди них тот, для которого погрешность аппроксимации соответствующего пространственного циркуляра является наименьшей. Полученный циркуляр является хорошим начальным приближением, которое можно улучшить, минимизируя погрешность аппроксимации при помощи численных методов.

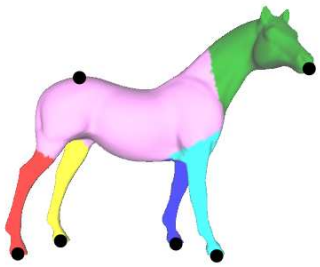


Рис. 3. Сегментация модели; точки множества Q_s показаны чёрными кружками.

Окклюзии. Важно выбрать такую плоскую проекцию, которая не содержала бы самопересечений (окклюзий). Например, если для модели лошади на рис. 3 выбрать такой ракурс, при котором одна из ног закрывает другую, то полученный скелет окажется некорректным. Если плоская проекция, не содержащая окклюзий, существует, то её поиск упрощается тем, что для некорректных скелетов значение погрешности аппроксимации чрезвычайно велико. Однако в общем случае проекции без окклюзий может не существовать. В качестве одного из возможных решений указанной проблемы следует упомянуть предварительную сегментацию модели на простые фрагменты, для каждого из которых легко найти собственную плоскую проекцию, не содержащую окклюзий. После этого достаточно лишь объединить между собой циркуляры, аппроксимирующие отдельные сегменты, в один общий циркуляр.

Сегментация модели задаётся множеством точек

$$Q_s = \{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_s\}, \boldsymbol{q}_i \in \partial\Omega.$$

Пусть ρ_Ω — геодезическое расстояние по поверхности $\partial\Omega$. Тогда каждый сегмент S_i определяется как множество точек поверхности, ближайших к $\mathbf{q_i}$:

$$S_i = \{\boldsymbol{x} \in \partial\Omega | \forall k, 1 \leqslant k \leqslant s, \rho_\Omega(\boldsymbol{x}, \boldsymbol{q}_i) \leqslant \rho_\Omega(\boldsymbol{x}, \boldsymbol{q}_k)\}.$$

Заданную таким образом сегментацию можно рассматривать как диаграмму Вороного на поверхности фигуры в метрике, определённой через длину геодезических линий. Пример сегментации показан на рис. 3.

Восстановление трёхмерного скелета по плоской проекции выполняется так, чтобы оси

получаемого циркуляра проектировались в линии двумерного скелета. Каждая вершина двумерного скелета (т. е. центр некоторого максимального вписанного круга) является проекцией по крайней мере двух точек A, B поверхности фигуры (см. рис. 4). В качестве хорошего начального приближения точки оси циркуляра можно взять середину отрезка AB.

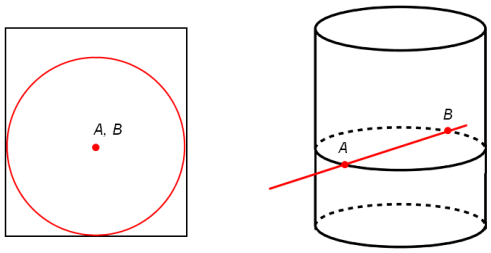


Рис. 4. Двумерная проекция (слева) трёхмерного цилиндра (справа); центр максимального вписанного круга является проекцией точек A, B.

Вычислительный эксперимент

Вышеописанный алгоритм был реализован на практике. На рис. 5 показаны примеры полученных скелетов для модельных изображений, которые обычно демонстрируются в публикациях, посвящённых трёхмерным криволинейным скелетам.

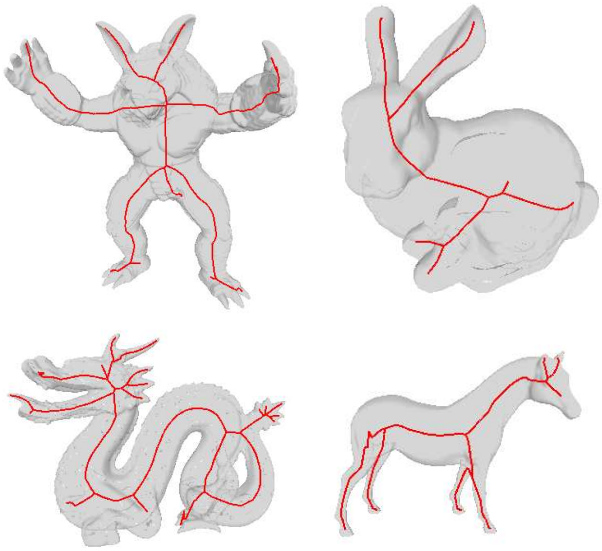
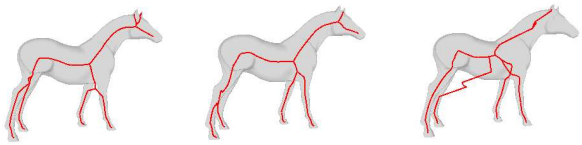


Рис. 5. Примеры трёхмерных криволинейных скелетов.

Одним из основных результатов, полученных в настоящей работе, является критерий соответствия между скелетом и исходной фигурой. Чтобы экспериментально продемонстрировать состоятельность этого критерия, были сгенерированы скелеты из различных плоских проекций без какой-либо последующей численной оптимизации. Для каждого такого скелета была вычислена величина погрешности аппроксимации. Примеры показаны на рис. 6. Хорошо видно, что чем больше величина погрешности, тем менее осмысленным получается скелет.



(а) $\varepsilon = 0.0077$ (б) $\varepsilon = 0.0080$ (в) $\varepsilon = 0.0089$

Рис. 6. Криволинейные скелеты с различной величиной погрешности аппроксимации.

Выводы

В настоящей работе представлен подход к построению трёхмерных криволинейных скелетов, новизна которого связана с использованием численного критерия для оценки соответствия между скелетом и исходной фигурой, что позволяет говорить о теоретической обоснованности, а кроме того, даёт возможность строго сравнивать между собой различные алгоритмы.

Среди направлений для дальнейших исследований можно указать следующие:

— дальнейшая разработка функции критерия качества;
— алгоритмы для борьбы с окклюзиями в плоских проекциях фигуры;
— численные методы оптимизации начального приближения скелета.

Литература

[1] *Blum H.* A Transformation for Extracting New Descriptors of Shape. — Models for the Perception of Speech and Visual Form, 1967.

[2] *Mestetskiy L.* Fat curves and representation of planar figures — Computers & Graphics, vol.24, No. 1, 2000, pp.9-21.

[3] *Lieutier A.* Any open bounded subset of R^n has the same homotopy type than its medial axis. — Proceedings of the eighth ACM symposium on Solid modeling and applications, 2003.

[4] *Cornea N.D., Silver D., Min P.* Curve-Skeleton Properties, Applications, and Algorithms. — IEEE Transactions on Visualization and Computer Graphics, 13(3) 530–548 (2007).

[5] *Siddiqi K., Pizer S. M.* Medial Representations: Mathematics, Algorithms and Applications. — Springer, 2008.

[6] *Цискаридзе А.К.* Математическая модель и метод восстановления позы человека по стереопаре силуэтных изображений. — «Информатика и её применения», том.4, вып.4, 2010.

Параметрический дескриптор формы на основе гранично-скелетной модели*

Жукова К. В., Рейер И. А.
kz@pisem.net,reyer@forecsys.ru
Москва, Вычислительный центр РАН

В работе представлен дескриптор формы объекта — множество вершин выпуклых углов аппроксимирующей объект многоугольной фигуры с оценкой значимости выпуклой особенности границы, соответствующей каждой вершине. Оценки значимости вычисляются на основе анализа параметрического семейства гранично-скелетных моделей формы. Описано применение дескриптора для выделения линии и характерных точек профиля на изображениях лица.

Одной из основных характеристик формы объекта, которую можно использовать для получения признакового описания, является граница объекта. Установлено, что человеческий глаз анализирует форму объекта, опираясь на выпуклые и вогнутые фрагменты границы. На этом соображении основан ряд методов структурного анализа формы, использующих представление контура в виде последовательности особенностей-примитивов (выпуклостей и вогнутостей) [1, 2]. Если при этом рассматривать представления контура с различной точностью приближения, то можно говорить о значимости особенности — чем существенней особенность, тем в более «грубых» представлениях она сохраняется. Поэтому помимо нахождения особенностей нужно получить оценку значимости каждой особенности границы. Таким образом, возникает задача построения такого дескриптора формы, который содержит информацию об особенностях формы контура для разных уровней детализации.

Один из подходов к решению этой задачи представляют методы «обнаружения углов» (corner detection) [3, 4]. Идея этих методов состоит в отборе точек локальных экстремумов границы по пороговому значению. Оценкой значимости особенности здесь является абсолютная величина кривизны фрагмента границы, соответствующего этой особенности.

Другим популярным инструментом является масштабируемая модель кривизны границы (curvature scale space) [5, 6], которая основана на аппроксимации границы кусочно-гладкой кривой, сглаживании этой кривой и выявлении экстремумов или нулей кривизны границы при разных степенях сглаживания.

Описанные методы анализа особенностей границы основываются на оценке кривизны границы, поэтому для их реализации требуется либо адаптировать понятие кривизны для дискретного представления контура, либо аппроксимировать границу кривыми высших порядков.

Работа выполнена при поддержке РФФИ, проекты № 11-07-00462 и № 11-01-00783.

Для решения поставленной задачи мы предлагаем использовать параметрическое семейство гранично-скелетных моделей формы [7], строящихся на основе аппроксимирующей объект многоугольной фигуры и состоящих из базового скелета фигуры и границы объединения множества базовых кругов. Анализируя изменение моделей при росте величины точности аппроксимации, для каждой вершины выпуклого угла исходного многоугольника можно получить оценку значимости — минимальную величину точности, при которой соответствующая вершине особенность границы исключается из граничного описания. Соответственно, набор вершин выпуклых углов многоугольной фигуры с сопоставленными им величинами точности аппроксимации будем использовать в качестве дескриптора формы.

Базовый скелет многоугольной фигуры

Напомним основные моменты концепции базового скелета, представленной в [7].

Пусть P — односвязная многоугольная фигура, ε — некоторое неотрицательное число.

Определение 1. *Круг C называется ε-допустимым кругом для P, если:*

1) расстояние Хаусдорфа между областями P и $P \cup C$ $H(P, P \cup C) \leqslant \varepsilon$;

2) расстояние Хаусдорфа между границами областей $H(\partial P, \partial(P \cup C)) \leqslant \varepsilon$.

Определение 2. *Круг C называется максимальным ε-допустимым кругом для P, если:*

1) C является ε-допустимым кругом для P;

2) C не содержится целиком ни в каком другом ε-допустимом для P круге.

Нетрудно видеть, что множество центров максимальных ε-допустимых кругов для P совпадает со множеством центров максимальных пустых кругов для P.

Определение 3. *Круг C называется базовым кругом для многоугольной фигуры P, если выполнено следующее:*

1) C является максимальным ε-допустимым кругом для P;

2) пусть точки, в которых максимальный пустой круг C', соответствующий кругу C, касается границы фигуры, разбивают границу на фрагменты $P_1, \ldots, P_n$, $n \geqslant 2$, а радиусы круга C, проходящие через эти точки, разбивают окружность круга C на дуги $L_1, \ldots, L_n$; тогда существуют $i\colon 1 \leqslant i \leqslant n$, $j\colon 1 \leqslant j \leqslant n$, $i \neq j$, такие, что $H(P_i, L_i) \geqslant \varepsilon$ и $H(P_j, L_j) \geqslant \varepsilon$ (рис. 1).

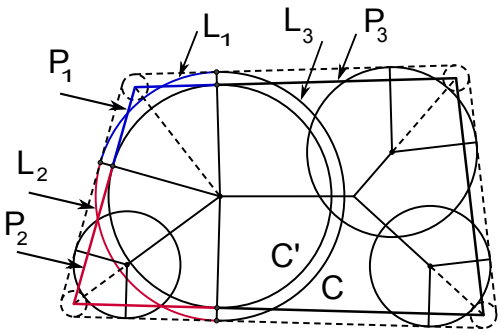


Рис. 1.

Определение 4. *Базовым скелетом многоугольной фигуры P называется множество центров всех базовых кругов области.*

Отметим, что базовый скелет P является подмножеством скелета P.

«Разметка» скелета

В [7] показано, что при росте ε базовый скелет изменяется монотонным и непрерывным в смысле расстояния Хаусдорфа образом. Ребра скелета «стираются» парами кривых — парабол и гипербол (рис. 2). Состав пары зависит от типа ребра: две параболы для ребра-отрезка, порожденного сайтами-сегментами; две гиперболы для ребра-отрезка, порожденного сайтами-точками; парабола и гипербола для ребра — фрагмента параболы, порожденного сайтом-сегментом и сайтом-точкой. Положение стирающих кривых для каждого ребра v определяется положением соответствующих ребру сайтов, величиной ε и положением наиболее удаленной от v точки F_v из некоторого подмножества U'_v вершин границы.

При этом ребро скелета может стираться несколькими парами кривых. В самом деле, точка F_v может быть различной для разных фрагментов ребра v. Диаграмма Вороного дальней точки [8] множества вершин U'_v позволяет определить для каждой точки ребра скелета самую удаленную вершину. Таким образом, в точках пересечения ребер скелета с ребрами диаграммы происходит смена пары стирающих кривых. Кроме того, в таких точках возможно нарушение связности: ребро стирается двумя парами кривых, двигающимися в противоположных направлениях.

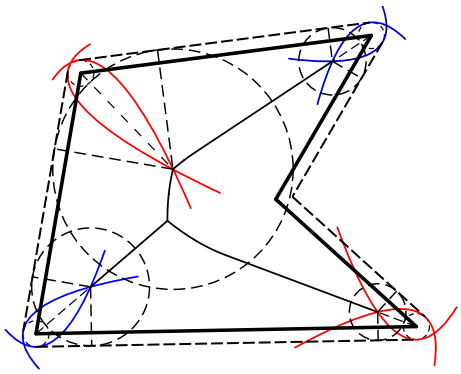


Рис. 2.

Также интерес представляют ситуации, когда одна стирающая кривая из пары касается ребра:

— ребро — фрагмент параболы; одна из парабол в паре стирает ребро с двух сторон и процесс заканчивается в точке касания параболы и ребра;

— ребро — отрезок, порожденный сайтами-точками; одна из пары гипербол стирает ребро с двух сторон и процесс заканчивается в точке касания гиперболы и отрезка;

— ребро — отрезок, порожденный сайтами-сегментами; одна из стирающих парабол сначала касается ребра, а затем пересекает его в двух точках, то есть нарушается связность базового скелета.

Таким образом, в точках касания ребра и кривой заканчивается либо начинается (нарушение связности) стирание ребра скелета.

Очевидно, что при некотором ε из базового скелета исчезнут все ребра. Ребро, которое выпадет из базового скелета последним, назовем центральным ребром. На нем лежит центральная точка, в которой заканчивается стирание ребра и всего скелета.

Будем рассматривать точки трех перечисленных типов как вершины скелета. В результате получим «размеченный» скелет (рис. 3), каждое ребро которого стирается одной парой кривых, в одном направлении. При этом с каждым ребром связаны два значения точности ε, при которых стирающая пара проходит через концевые точки ребра.

Таким образом, размеченный скелет дает возможность получить базовый скелет для любых заданных значений точности.

Параметрический дескриптор формы

Итак, с помощью размеченного скелета многоугольной фигуры можно получить базовый скелет для любого заданного значения ε. При этом границу объединения множества всех базовых кругов можно рассматривать в качестве модели контура фигуры, отражающей те свойства границы, кото-

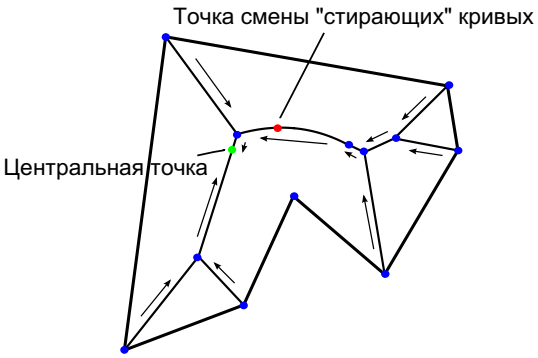


Рис. 3.

рые являются существенными в пределах точности аппроксимации.

При увеличении ε из базового скелета «выпадают» ветви, которые порождены выпуклостями границы, несущественными при данной точности аппроксимации. Таким образом, каждой выпуклости границы соответствует значение ε, при котором соответствующая выпуклости ветвь скелета перестает быть базовой. На этих соображениях основан параметрический дескриптор формы: он представляет собой граничное описание в виде списка вершин выпуклых углов аппроксимирующего многоугольника, где для каждой вершины определена оценка значимости, вычисляемая следующим образом.

Разметим скелет многоугольника и рассмотрим процесс стирания ветвей скелета. При этом будем считать, что среди вершин размеченного скелета нет точек нарушения связности. Стирание начинается с терминальных вершин скелета (они совпадают с вершинами выпуклых углов границы) и распространяется «внутрь» фигуры. Для каждого ребра в размеченном скелете определены значения точности ε в концевых точках и известно направление стирания: от меньшего значения ε к большему. Вершине q степени $n > 2$ инцидентны $n-1$ входящих ребер $v_1, \ldots, v_{n-1}$ и одно исходящее ребро v_n со значениями точности $\varepsilon_1, \ldots, \varepsilon_n$ в точке q; при этом значение в точке q для ребра v_n является максимальным $\varepsilon_n = \max(\varepsilon_1, \ldots, \varepsilon_n)$ и совпадает как минимум с одним из значений $\varepsilon_1, \ldots, \varepsilon_{n-1}$. Это означает, что последним из всех входящих ребер сотрется ребро с максимальным ε. Таким образом, если при стирании ребра концевая точка с большим значением ε является вершиной скелета степени $n > 2$ и значение ε не является максимальным из всех значений ε входящих ребер в этой точке, то стартовой вершине присваивается оценка значимости ε и процесс для этой вершины закончен. В противном случае стирание продолжается. Если достигается концевая точка ребра, которая является центральной точкой скелета, процесс останавливается. Так, для ребра q_1q_3 размеченного скелета фигуры, изображенной на рис. 4, значения ε для точек q_1 и q_3 составляют 0 и $\varepsilon_1 = 23$ соответственно, а для ребра q_2q_3 — 0 и $\varepsilon_2 = 1.5$. Поскольку $\varepsilon_1 > \varepsilon_2$, то ребро q_2q_3 сотрется раньше, чем ребро q_1q_3, и вершине q_2 будет присвоена оценка значимости ε_2. Для вершины же q_1 процесс стирания продолжится, и оценка значимости будет равна значению точности в центральной точке скелета q_c. Описанную процедуру нетрудно обобщить на случай, когда связность базового скелета нарушается.

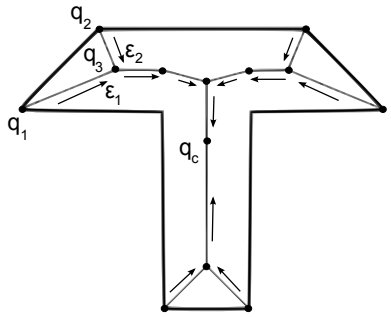


Рис. 4.

На рис. 5 приведен пример формы и показаны вершины границы, являющиеся существенными при указанных значениях точности аппроксимации ε.

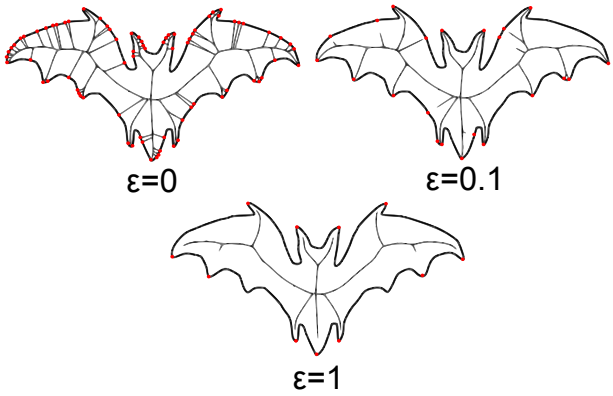


Рис. 5.

Задача выделения линии профиля на изображении лица

В задачах идентификации человека по линии профиля и распознавания движений лицевых мышц [9] требуется получить линию профиля на изображении головы человека и определить на ней набор точек, указывающих на основные части лица (рис. 6). Для решения такой задачи параметрический дескриптор формы был применен следующим образом. Строятся гранично-скелетные модели для головы и фона и вычисляются оценки значимости вершин выпуклых углов. Предполагая, что голова повернута в определенную сторону, выделяется лицевой фрагмент контура. На основе полученных

оценок выбираются 10 точек контура (обозначим их $P_1, \ldots, P_{10}$ в соответствии с [9]), которые определяют особенности профиля. Поиск точек проводится по следующим правилам. В центральной части изображения вершина выпуклого угла контура головы с максимальной оценкой значимости принимается за кончик носа P_4. Точкой P_{10}, соответствующей подбородку, является вершина выпуклого угла с максимальной оценкой значимости среди лежащих «ниже» точки P_4. Точки $P_5, \ldots, P_9$ находятся между точками P_4 и P_{10}: P_6 и P_8 соответствуют вершинам выпуклых углов с максимальными оценками значимости, а P_5, P_7 и P_9 — вершинам вогнутых углов с максимальными оценками значимости. Точка P_3 (переносица) принадлежит группе вершин вогнутых углов, лежащих «выше» точки P_4, с относительно большой значимостью. Далее расположена группа вершин вогнутых углов, одна из которых есть точка P_1. Между P_1 и P_3 находится точка P_2 — как вершина выпуклого угла с максимальной значимостью. Для исключения ошибок при поиске точек проверяется соответствие взаимного расположения точек анатомическим пропорциям.

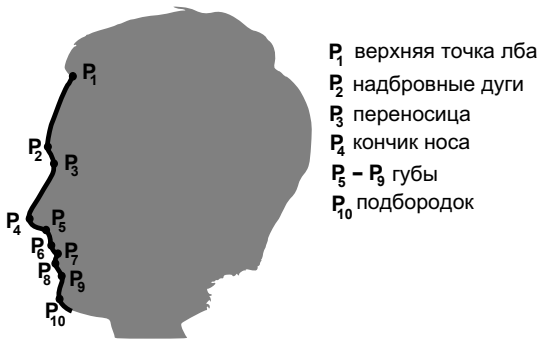


Рис. 6.

Описанный алгоритм поиска характерных точек профиля был применен к сводной базе изображений, в которую вошли профильные изображения из базы Бернского университета [10], базы Color FERET [11, 12] и базы, подготовленной авторами. Набор точек был правильно выделен в 387 случаях из 411.

Выводы

В работе предложен параметрический дескриптор формы объекта, который можно использовать для анализа свойств формы, проявляющихся при различных значениях точности аппроксимации. Дескриптор представляет собой множество вершин выпуклых углов границы аппроксимирующей объект многоугольной фигуры, каждой из которых сопоставлена оценка значимости. Вычисление оценок основано на анализе монотонного и непрерывного изменения гранично-скелетных моделей формы при росте величины точности аппроксимации. Полученный дескриптор дает возможность оценить количество существенных для заданного значения точности аппроксимации особенностей формы и их взаимное расположение. При решении задач анализа и распознавания формы можно также использовать дескриптор вместе с размеченным скелетом для получения граничного представления объекта с нужной степенью детализации. Предложенный способ представления позволяет выделять и анализировать особенности кривизны границы без использования методов оценки кривизны по дискретному представлению контура и аппроксимации границы кривыми высших порядков.

Литература

[1] *Galton A., Meathrel R.* Qualitative Outline Theory // School of Engineering and Computer Science, University of Exeter, UK.

[2] *Rosin P.L.* Multiscale representation and matching of curves using codons // CVGIP: Graphical Models and Image Processing 55(4), 286-310.

[3] *Koplowitz J., Plante S.* Corner detection for chain codes curves // 1994.

[4] *Ray B. K., Pandyan R.* ACORD – an adaptive corner detector for planar curves, Pattern Recognition, Vol.36, 2003, pp. 703-708.

[5] *Abbasi S., Mokhtarian F., Kittler J.* Curvature scale space image in shape similarity retrieval, MultiMedia Systems, Vol.7, 1999, pp. 467–476.

[6] *Dudek G., Tsotsos J. K.* Shape representation and recognition from mutliscale curvature, Computer Vision and Image Understanding, Vol.68, No.2, 1997, pp. 170–189.

[7] *Жукова К. В., Рейер И. А.* Параметрическое семейство гранично-скелетных моделей формы // Доклады всероссийской конференции ММРО-14, 2009.

[8] *Препарата Ф., Шеймос М.* Вычислительная геометрия: введение — Москва: Мир, 1989. — 478 с.

[9] *Pantic M., Rothkrantz L.* Facial Action Recognition for Facial Expression Analysis From Static Face Images // IEEE Transactions on Systems, Man, and Cybernetics - part B: Cybernetics, Vol.34, No 3, 2004

[10] *Achermann B.* University of Bern Face Database. — Copyright 1995, University of Bern, all rights reserved. — `ftp://iamftp.unibe.ch/pub/Images/FaceImages/`.

[11] *Phillips P. J., Wechsler H., Huang J., Rauss P.* The FERET database and evaluation procedure for face recognition algorithms // Image and Vision Computing J, Vol. 16, No. 5, pp. 295-306, 1998.

[12] *Phillips P. J., Moon H., Rizvi S. A., Rauss P. J.* The FERET Evaluation Methodology for Face Recognition Algorithms // IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 22, pp. 1090-1104, 2000.

Классификация лекарственных растений по форме листа на основе скелетного представления

Макарова Е. Ю.
Luar.Soll@gmail.com
Москва, Московский государственный университет им. М. В. Ломоносова

В данной работе предлагается признаковое описание для бинарных изображений листьев, полностью вычисляемое по скелетному представлению этих изображений, и приводится краткий анализ результатов классификации тестовой выборки с использованием предложенного описания.

Возможность автоматического или полуавтоматического распознавания растений является весьма полезной, причём не только для тех, кто занимается ботаникой профессионально, поскольку может оказаться необходимым, например, отличить лекарственное или съедобное растение от похожего на него ядовитого. Но фотографии растения в целом могут слишком сильно различаться между собой, поэтому ограничим нашу задачу только изображениями листьев, причём только их формой, без учёта расцветки и жилкования, то есть бинарными изображениями.

Конечно, существуют различные варианты численного описания различных морфологических признаков листа, используемого потом в качестве признакового описания. Однако в работах по этой теме, например, в [2], в основном использовались характеристики либо границы листа, либо отдельных вписанных в лист и описанных около листа фигур, или, например, в [3], соотношения площадей частей листа и содержащих эти части прямоугольников.

В данной работе предлагается построение признакового описания, вычисляемого на основе скелетного представления, поскольку скелетное представление фигуры содержит всю информацию о форме этой фигуры, причём даёт не локальное, а интегральное представление о ней.

Основные понятия

Будем считать, что заданное бинарное изображение листа аппроксимировано многоугольной фигурой, то есть его граница состоит из одной или нескольких замкнутых ломаных без самопересечений, не имеющих общих точек. Тогда к нему применимы следующие понятия.

Определение 1. *Круг, полностью лежащий внутри фигуры, называется пустым кругом. Максимальный пустой круг — пустой круг, не содержащийся ни в каком другом пустом круге.*

Определение 2. *Скелет фигуры — множество центров максимальных пустых кругов.*

Определение 3. *Радиальная функция точки скелета — величина радиуса максимального пустого круга с центром в этой точке.*

Для скелета существует графовое представление, причём вершинами графа являются центры максимальных кругов, касающихся границы фигуры в трёх и более точках, а ребрами — линии, состоящие из центров пустых кругов, касающихся границы ровно в двух точках. Если граница фигуры состоит только из отрезков прямых и выпуклых наружу дуг окружностей, то ребра её скелета являются отрезками либо прямых, либо парабол. В таком случае форма фигуры однозначно определяется по графовому представлению её скелета и значениям радиальной функции в его вершинах, а вдоль ребер радиальная функция изменяется линейно.

Определение 4. *Объединение всех максимальных пустых кругов с центрами на подграфе скелета называется силуэтом этого подграфа.*

Определение 5. *Вершина скелета, которой в графовом представлении инцидентно ровно одно ребро, называется терминальной.*

Определение 6. *Скелет со стрижкой степени p — скелет размыкания исходного изображения с кругом радиуса p с центром в начале координат в качестве примитива.*

Утверждение 1. *Скелет со стрижкой степени p является подграфом скелета исходного изображения.*

Утверждение 2. *Расстояние Хаусдорфа от силуэта скелета со стрижкой степени p до исходного изображения равно p, причём расстояние от силуэта любого подграфа скелета со стрижкой степени p до исходного изображения больше p.*

Биологические понятия. Необходимо уточнить, что под *листом* далее подразумевается одна листовая пластинка. Кроме того, не будем делать различия между сегментами, долями и лопастями листа — частями листовой пластинки, разделёнными вырезами, более 2/3, от 1/3 до 2/3 и менее 1/3 полуширины листа соответственно ([4]). Для краткости будем называть все такие части *долями*, а все меньшие вырезы относить к *изрезанности края листа*. Лист, вырезы края которого не превышают 1/4 полуширины листа будем называть *цельным*.

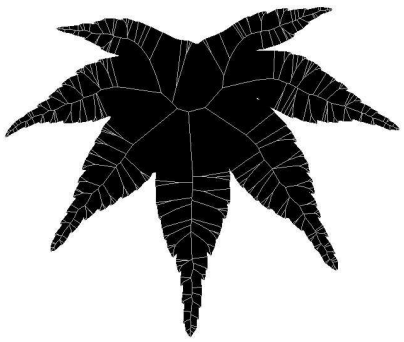


Рис. 1. Скелет листа

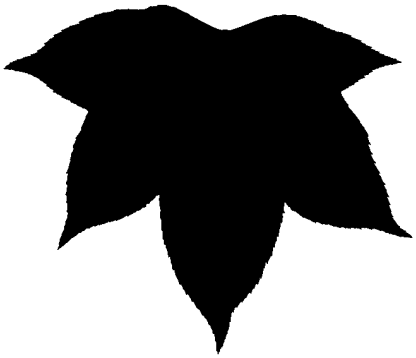


Рис. 3. Лист с не сильно изрезанным краем.

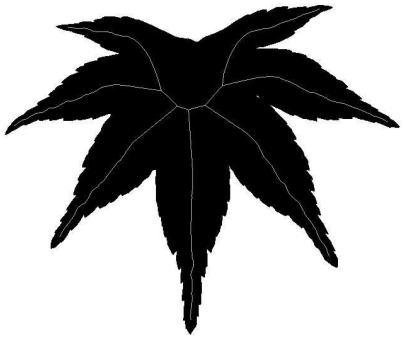


Рис. 2. Скелет листа со стрижкой степени 15.

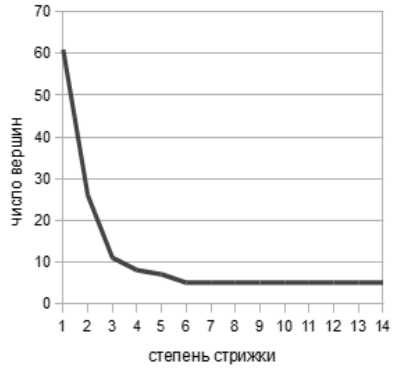


Рис. 4. График зависимости числа терминальных вершин от степени стрижки для листа со слабо изрезанным краем (рис. 3).

Признаковое описание

Терминальные вершины и стрижка скелета. Наиболее важными и заметными характеристикой листовой пластинки являются количество долей и степень изрезанности края листа. Но определить количество долей по скелету листа непосредственно как число терминальных вершин этого скелета невозможно, так как терминальных вершин оказывается намного больше (рис. 1). Поэтому количество долей приходится определять по скелету листа с некоторой стрижкой (рис. 2). Необходимо обратить внимание, что число терминальных вершин и в таком случае является не строго равным количеству долей, а только характеристикой числа долей, поскольку, например, у цельного листа доля одна, а терминальных вершин скелета две.

Однако в таком случае возникает вопрос, каким образом определить такую степень стрижки скелета, когда уже нет тех терминальных вершин, которые происходят от локальной изрезанности края, но ещё имеются все вершины, соответствующие отдельным долям. Конечно, можно выбрать число, примерно равное 1/8 ширины листа, так как при меньших выемках лист считается цельным. Но возможно также поступить иначе.

Построим последовательность скелетов одного и того же листа со степенями стрижки $p = 0, 1, 2, \ldots$. Рассмотрим количества терминальных вершин $t_0, t_1, t_2, \ldots$ этих скелетов.

Утверждение 3. *Для любого интервала стабилизации n существует p_0 такое, что $t_{p_0} = t_{p_0+1} = \cdots = t_{p_0+n-1}$, причём если n не велико, а скелет листа не содержит циклов, то $t_{p_0} \geqslant 2$.*

При удачно подобранном интервале стабилизации p_0 оказывается минимальным радиусом круга-примитива, операция размыкания с использованием которого сглаживает изрезанность краев, но не сглаживает вырезы между долями. Возможно заметить, что чем больше изрезанность края листа при постоянном n и примерно постоянных линейных размерах листа на изображении, тем больше этот радиус, поэтому его можно считать признаком, описывающим изрезанность края листа. В качестве примера можно рассмотреть графики зависимости числа терминальных вершин от степени стрижки скелета: на рис. 4 график для листа с рис. 3 (менее изрезанный край) и на рис. 5 график для листа с рис. 2 (более изрезанный край).

Максимальный и средний круги. Помимо числа долей и изрезанности края, листья могут заметно различаться соотношением длины и ширины, что особенно заметно и важно для цельных

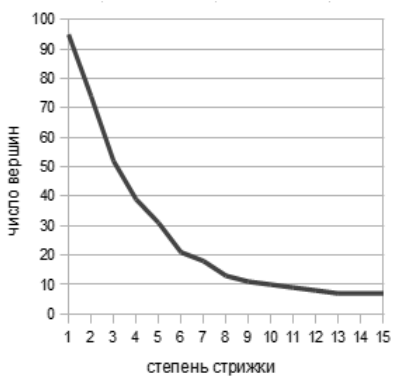


Рис. 5. График зависимости числа терминальных вершин от степени стрижки для листа со слабо изрезанным краем (рис. 2).

листьев. Кроме того, могут существовать как листья, почти равномерно широкие по всей длине, так и с близким к линейному изменением ширины от нуля до максимума. Поэтому необходимо ввести признаки, отвечающие как за максимальную, так и за среднюю ширину листьев.

В качестве характеристики максимальной ширины листа в данной работе используется отношение наибольшего значения радиальной функции к расстоянию между двумя наиболее удалёнными друг от друга вершинами скелета. Необходимо заметить, что, во-первых, максимальное значение радиальная функция всегда принимает в вершине скелета, во-вторых, эта вершина и это значение совпадают для полного скелета и скелета со стрижкой степени p_0, в-третьих, наиболее удалённые друг от друга вершины также возможно искать в скелете со стрижкой степени p_0, поскольку перебор по его вершинам делается гораздо быстрее, а разница в расстояниях не может превысить $2 * p_0$, что в большинстве случаев слабо влияет на оценку.

Среднюю ширину листа нельзя считать как среднее арифметическое значений радиальной функции в вершинах скелета, поскольку вклад в среднее для вершин, инцидентных коротким ребрам и вершин, инцидентных длинным, заметно различается. Поэтому среднюю ширину листа будем вычислять следующим образом:

$$R_{av} = \left(\frac{\sum \left(b_i * \left(r_{i,1} + r_{i,2} \right) / 2 \right)}{\sum b_i} \right) / d_0,$$

где b_i — длина i-го ребра, $r_{i,1}$ и $r_{i,2}$ — значения радиальной функции в вершинах, инцидентных i-му ребру, d_0 — расстояние между наиболее удалёнными друг от друга вершинами скелета, а оба суммирования ведутся по всем ребрам скелета со стрижкой степени p_0.

Положение максимального круга. Ещё одним достаточно заметным признаком, отличающим листья растений разных видов может оказаться положение максимального круга. У одних листьев ширина наибольшая примерно в центре, у других — ближе к одному из концов. К сожалению, определение по листовой пластинке, с какого конца её присоединялся черенок, иногда затруднительно, поэтому положение максимального круга будем определять только по близости к любому из концов.

Для нахождения значения этого признака необходимо найти две наиболее удалённые друг от друга вершины скелета, и вычислить отношение расстояния от вершины, в которой достигается максимум радиальной функции, до ближайшего из них, к расстоянию между ними.

Признаковое пространство. Во-первых, нужно заметить, что, с одной стороны, наиболее важным из приведённых признаков является характеристика числа долей: цельный лист и лист с несколькими долями не могут принадлежать одному и тому же классу, с другой стороны, листья с числом долей, например, 5, 6 и 7 могут быть от растений одного вида. Поэтому в качестве признака предлагается использовать не само число t_{p_0}, а его двоичный логарифм, но использовать этот признак с большим весом.

Во-вторых, различия в признаках, характеризующих ширину листа, в десятые доли, являются более заметными, чем различия в значениях p_0 в единицы. Значит, веса признаков ширины должны быть больше, чем вес признака стрижки.

Классификация тестовой выборки

В качестве метода классификации в описанном выше признаковом пространстве со взвешенными признаками использовался метод ближайшего соседа. Причём, так как итоговый алгоритм должен иметь скорее рекомендательный, чем решающий характер, то есть результаты его работы должны только подсказывать, на принадлежность какому виду прежде всего надо проверить исследуемое растение, то рассматривались три ближайших соседа классифицируемого объекта, и классификация считалась успешной, если хотя бы один из этих соседей был того же класса, что и классифицируемый.

Во время экспериментов использовались бинарные изображения листьев растений 32 видов, всего 1907, каждого вида — от 50 до 70. В качестве обучающей выборки от каждого вида было взято по 10 объектов; в качестве тестовой выборки — все 1907 объектов, включая вошедшие в обучающую. Веса признаков и интервал стабилизации n подобраны вручную, поэтому, возможно, не оптимальны.

В результате из 1907 объектов успешно классифицированы 1560 объектов; среди 32 классов по ко-

личеству правильно определённых объектов можно выделить следующие группы:

— 2 класса, все объекты которых классифицированы успешно;
— 10 классов, на объектах которых ошибок было не более 5;
— 15 классов, на объектах которых число ошибок не более трети от числа объектов класса;
— 5 классов, на объектах которых число ошибок более трети от числа объектов класса.

Выводы

Если вычесть из общего числа объектов тестовой выборки и из числа успешно классифицированных объектов число объектов, вошедших в множество эталонов, то окажется, что удалось добиться доли удачных классификаций около 75%. При таком количестве классов, которое имеется у нас, это является неплохим результатом, однако требует дальнейшей работы.

В процессе проведения экспериментов были обнаружены важные характеристики формы листа, не охваченные имеющимся признаковым описанием, однако могущие помочь различить некоторые, плохо различимые сейчас, классы. Однако теперь уже имеет смысл вычислять дополнительные признаки для объектов не всех классов, а только наиболее часто смешиваемых.

Кроме того, так как, как уже было упомянуто, параметры классификации подбирались вручную, и алгоритм классификации был использован наиболее простой; возможно, в какой-то степени улучшить качество классификации сможет также более точный подбор параметров и смена алгоритма классификации.

Литература

[1] *Местецкий Л. М.* Непрерывная морфология бинарных изображений: фигуры, скелеты, циркуляры. — Москва: ФИЗМАТЛИТ, 2009.

[2] *Knight D., Painter J., Potter M.* Automatic Plant Leaf Classification for a Mobile Field Guide. — 2010. — `http://www.stanford.edu/~jpainter/documents/PlantLeafClassification.pdf`

[3] *Суботэ А. Е.* О методе автоматизированной классификации листьев высших растений (на примере ископаемых буковых) // Вестник ДВО РАН. — 2004. — № 3. — С. 174–177.

[4] *Коровкин О. А.* Анатомия и морфология высших растений: словарь терминов. — Москва: Дрофа, 2007.

Вычисление морфологических спектров плоских фигур с использованием непрерывных скелетных представлений*

Визильтер Ю. В., Сидякин С. В., Рубис А. Ю.
viz@gosniias.ru, sersid@bk.ru
Москва, ФГУП «Государственный научно-исследовательский институт авиационных систем»

Предложен новый подход к вычислению морфологических спектров с дисковыми структурирующими элементами, основанный на использовании непрерывных скелетных представлений плоских фигур. Этот подход позволяет формировать дискретно-непрерывные морфологические спектры плоских фигур в реальном времени, что делает их пригодными для использования в прикладных задачах машинного зрения.

Введение

В работе (Maragos, 1989) [1] по аналогии со спектрами Фурье было предложен способ описания плоских фигур и изображений при помощи форморазмерных спектров, вычисляемых средствами математической морфологии Серра [2]. Было показано, что морфологические спектры являются полезным и устойчивым дескриптором формы изображений. Однако несмотря на значительное время, прошедшее со времени публикации работы [1], морфологические спектры до сих пор не находили широкого практического применения в связи с отсутствием эффективных в вычислительном смысле процедур их построения. В данной работе предлагается новый подход к вычислению морфологических спектров, основанный на использовании непрерывных скелетных представлений плоских фигур, введённых в работах Л. М. Местецкого [3]. Данный подход позволяет формировать морфологические спектры плоских фигур в реальном времени, что обеспечивает возможность использования спектральных морфологических характеристик формы объектов в прикладных задачах машинного зрения.

Морфология Серра и морфологический спектр

Классическое описание операций бинарной математической морфологии (ММ) [2] дано в терминах теории множеств. При этом бинарные изображения (плоские фигуры) рассматриваются как множества ненулевых точек плоскости $P = \mathbb{R}^2$. Определим трансляцию множества $X \subset P$ по вектору $z \in P$ как преобразование

$$X_z = \{y \,|\, x \in X,\, y = x + z\},$$

где точки плоскости суммируются как вектора (координаты покомпонентно складываются). Пусть $X,\ B \in P$. Операция

$$X \oplus B = \{x + b | x \in X, b \in B\} = \bigcup_{x \in X} B_x = \bigcup_{b \in B} X_b \qquad (1)$$

называется сложением Минковского. Операция

$$X \ominus B = \{z \,|\, B_z \subseteq X\} \qquad (2)$$

называется вычитанием Минковского. В рамках ММ операции (1) и (2) называются дилатацией (расширением) и эрозией (сжатием) изображения X со структурирующим элементом B и являются базовыми операциями ММ. Операция

$$X \circ B = (X \ominus B) \oplus B \qquad (3)$$

называется открытием X по B и имеет ясный физический смысл — это объединение всех структурирующих элементов формы B, целиком входящих в фигуру X:

$$X \circ B = \bigcup_{B_z \subseteq X} B_z. \qquad (4)$$

Закрытием X по B называется операция

$$X \bullet B = (X \oplus B) \ominus B. \qquad (5)$$

Операторы (3), (5) являются идемпотентными (проективными) и сохраняют отношение включения, в силу чего они считаются морфологическими фильтрами. Рассмотрим теперь, следуя логике работы [1], структуру преобразования Фурье. Оно содержит два этапа:

1. умножение одномерного сигнала $f(t)$ на комплексную синусоиду $e^{-\omega t}$;
2. измерение площади под этим модифицированным сигналом $f(t)e^{-\omega t}$.

Можно считать некоторым «пробным образом», зависящим от частотного параметра ω и выделяющим некоторую информацию (спектральный состав) из сигнала путём модуляции и измерения преобразованного сигнала. При этом собственной спектральной характеристикой «пробного образа» является импульс на соответствующей частоте ω. Проведем аналогию с заменой:

1) $f(t) \to X$ — двумерный образ;
2) $e^{-\omega t} \to B_n$ — двумерный структурирующий элемент;
3) $\omega \to n$ — размерный (масштабный) параметр;

Работа выполнена при финансовой поддержке РФФИ, проекты № 11-08-01114-а, 11-08-01039-а.

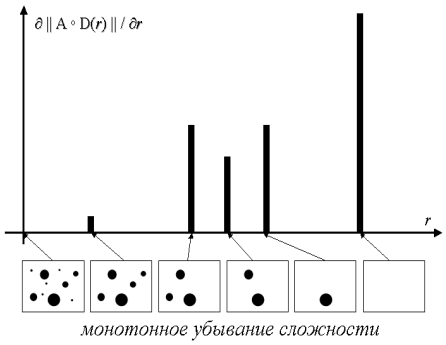


Рис. 1. Морфологический спектр с круглым структурирующим элементом и этапы морфологической обработки изображения при его построении.

4) модуляция → морфологическая фильтрация с использованием структурирующего элемента B_n размера n.

Далее будем рассматривать непрерывный бинарный случай. Введем понятие размера множества B. Пусть на плоскости P дано выпуклое множество , включающее начало координат. Пусть размер B считается единичным. Тогда множество rB, имеющее относительно B размер r, определяется как

$$rB = \{rb \,|\, b \in B\}, r \geqslant 0.$$

Очевидно, форма r повторяет форму B. Рассмотрим компактное (связное) бинарное изображение $X \subseteq P$. Определим морфологический спектр (pattern spectrum, PS) X относительно $B \subseteq P$ как функцию:

$$PS_X(r, B) = -\partial S(X \circ B)/\partial r,\ r \geqslant 0, \qquad (6)$$

$$PS_X(-r, B) = \partial S(X \bullet B)/\partial r,\ r > 0, \qquad (7)$$

где $S(X) = ||X||$ — площадь X, и выражения (6) и (7) задают спектр соответственно для положительной и отрицательной частей оси r. Пусть r есть rD — диск радиуса r. Убедимся в том, что спектральной характеристикой rD является импульс в точке r (как спектром $e^{-\omega t}$ является импульс в точке ω). В самом деле, так как $X = rD$ — компактный диск, то существует максимальное $p > 0$, такое, что $\forall\ r > p$: $X \circ rB = \varnothing$. При $0 < r < p$ имеем $X \circ rB = X$. Следовательно, функция $S(X \circ B)$ является ступенчатой, и её производная имеет один δ-импульс в точке $r = p$. Физический смысл спектра легко понять, если учесть, что согласно (4) $S(X \circ rB)$ есть мера содержания rB в X (рис. 236).

Непрерывная бинарная морфология и скелетные представления фигур

Следуя описанию непрерывной бинарной морфологии, данному Л. М. Местецким [3], примем следующие определения.

Жордановой кривой называется непрерывный инъективный образ окружности при отображении его в евклидову плоскость $P = \mathbb{R}^2$. Здесь $\mathbb{R}$ — множество действительных чисел. Важно, что жорданова кривая не имеет самопересечений. Фигурой называется связная замкнутая область плоскости, ограниченная конечным числом непересекающихся жордановых кривых. Пусть P — евклидова плоскость с соответствующим расстоянием $d(p, q)$, $p, q \in P$. Тогда граница фигуры X определяется как множество точек

$$\partial X = \{p : p \in P, \forall r > 0, D(p, r) \cap X \neq \varnothing,$$

$$D(p, r) \cap X^C \neq \varnothing\},$$

где $X^C = P \setminus X$ — дополнение или фон фигуры X, $D(p, r)$ — открытый круг радиуса r с центром в точке p, определяемый выражением

$$D(p, r) = \{q : q \in P, d(p, q) < r \in \mathbb{R}\}.$$

Пустым или вписанным кругом фигуры X называется круг $D(p, r) \subset X$. Максимальным пустым кругом называется пустой круг, который не содержится целиком ни в одном другом пустом круге данной фигуры. Скелетом $Sk(X)$ фигуры X называется множество центров всех её максимальных пустых кругов. Радиальной или дистанционной функцией точки $p \in P$ для фигуры X называется максимальная величина радиуса пустого круга с центром в данной точке $r_X(p)$.

Скелетным представлением фигуры является совокупность её скелета и радиальной функции, определённой в точках скелета

$$SR(X) = \{\langle p, r_X(p)\rangle : p \in Sk(X)\}. \qquad (8)$$

Полная реконструкция фигуры по скелетному представлению в точности совпадает с самой фигурой:

$$X = \bigcup_{\langle p,r\rangle \in SR(X)} D(p, r).$$

При этом связь с ММ Серра определяется следующим очевидным выражением:

$$X \circ rD = \bigcup_{\langle p,t\rangle \in SR(X)} \{D(p, t) : t \geqslant r\}. \qquad (9)$$

То есть, результат открытия (3) с дисковым структурирующим элементом может быть вычислен с использованием скелетного представления (8) на основе формулы (9). При этом для фигур, ограниченных многоугольниками с конечным числом сторон, скелет оказывается состоящим из конечного числа отрезков аналитических (алгебраических) кривых всего двух видов — прямых и парабол.

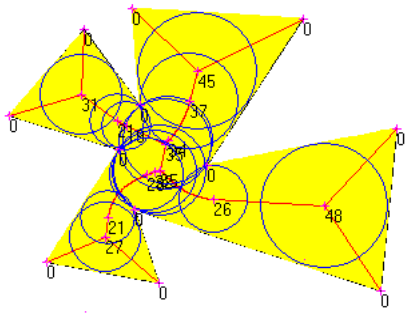


Рис. 2. Непрерывный скелет многоугольной фигуры. Показаны: скелет, точки сочленения отрезков прямых и парабол, пустые круги с центрами в этих точках и значения их радиусов.

Поэтому для построения непрерывных скелетных представлений существуют вычислительно эффективные алгоритмы, основанные на использовании обобщённых диаграмм Вороного [3]. На рис. 237 представлен пример построения непрерывного скелета для простой многоугольной фигуры.

Непрерывно-аналитический подход к вычислению морфологических спектров на основе скелетных представлений

Сравнение непрерывной скелетной морфологии (НСМ) [3] с известными ранее непрерывной и дискретной ММ Серра [2] показывает, что НСМ можно считать специфической реализацией бинарной ММ, в которой за счёт наложения дополнительных ограничений на форму бинарных образов (они должны быть приближённо описаны многоугольной границей) становится возможным вычислять результаты морфологических операций в непрерывно-аналитическом виде. Вычислительная эффективность НСМ по сравнению с ММ определяется тем, что от исходного описания изображения в виде дискретной матрицы пикселей (возможно, очень большого размера — до десятков мегапикселей) осуществляется переход к гораздо более компактному описанию формы в виде набора фиксированных графических примитивов (назовем их анкселями — ANalitcal piCTure ELements), число которых оказывается не только конечным (в отличие от непрерывной ММ Серра), но и на несколько порядков меньше исходного числа пикселей.

В рамках НСМ [3] анксели представляют собой объединение многоугольников и кругов (см. рис. 237). Идея аналитического вычисления спектра (6) показана на рис. 238 на примере одного из базовых анкселей НСМ. Выражение для изменения площади анкселя при малом изменении значения r имеет вид $\partial S(r+\partial r) = \varphi r \partial r$, откуда спектр данного анкселя может быть аналитически вычислен

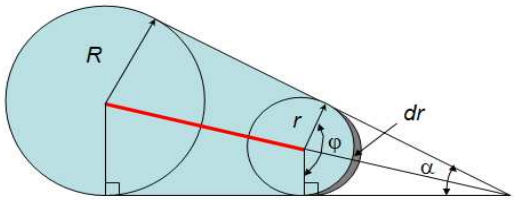


Рис. 3. Один из базовых анкселей НСМ и геометрическая иллюстрация построения спектра.

как

$$-\partial S/\partial r = \begin{cases} (\pi-\alpha)r, 0 \leqslant r < R; \\ \pi R^2, r = R; \\ 0, r > R, \end{cases}$$

где α — угол между опорными сайтами данного анкселя. Аналогичным образом может быть аналитически вычислен спектр и для других базовых анкселей. Однако необходимость учёта всех случаев пересечения анкселей для каждой данной фигуры делает крайне запутанной логику написания подобной вычислительной программы. Более перспективным оказывается дискретно-непрерывный алгоритм, описанный ниже.

Дискретно-непрерывный алгоритм вычисления морфологических спектров на основе скелетных представлений

Алгоритм основан на голосовании точек скелета в двумерный аккумулятор, размер которого совпадает с размером исследуемого бинарного изображения (фигуры).

Шаг 1. Представить бинарное изображение в виде непрерывной многоугольной фигуры F.

Шаг 2. Вычислить скелетное представление $SR(F)$ (8).

Шаг 3. Инициализировать все ячейки аккумулятора нулями.

Шаг 4. Выбрать очередное ребро из списка ребер скелета и перейти к шагу 5. Если список пуст, перейти к шагу 6.

Шаг 5. Текущее ребро скелета представить в дискретном виде при помощи алгоритма Брезенхэма [4], образовав набор точек $\{p_i\}$, характеризуемых координатами (x_i, y_i) центра и радиусом r_i соответствующих пустых кругов. Для каждой точки p_i при помощи алгоритма Брезенхэма построить окружность радиуса r_i в аккумуляторе, поместив в соответствующие ячейки значение радиуса r_i, если r_i больше текущего значения в рассматриваемой ячейке, в противном случае оставить её без изменений. Вернуться к шагу 4.

Шаг 6. Вычислить морфологический спектр как гистограмму значений аккумулятора. Конец алгоритма.

Поскольку в результате проведённых вычислений каждый элемент аккумулятора содержит такое значение r, которое соответствует радиусу проходящего через него максимального пустого круга, гистограмма аккумулятора является достаточно точным дискретным приближением положительной оси морфологического спектра (6). Точность вычисления спектра, естественно, зависит от выбранного дискрета (шага) алгоритма Брезенхема.

Отрицательная часть спектра (7) может быть вычислена при помощи такого же алгоритма, применяемого к дополнению (негативу, фону) исходного изображения.

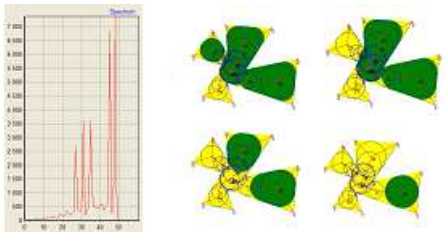


Рис. 4. Дискретно-непрерывный морфологический спектр и пиковые составляющие формы фигуры.

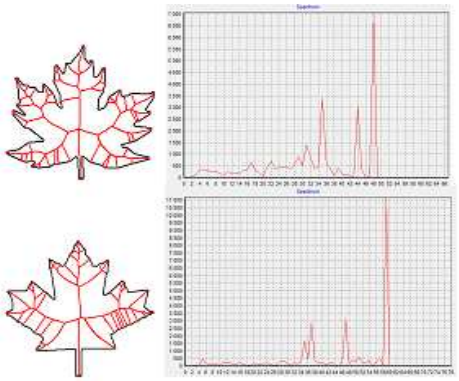


Рис. 5. Скелеты и ДНМС реального и схематического силуэтов кленового листа.

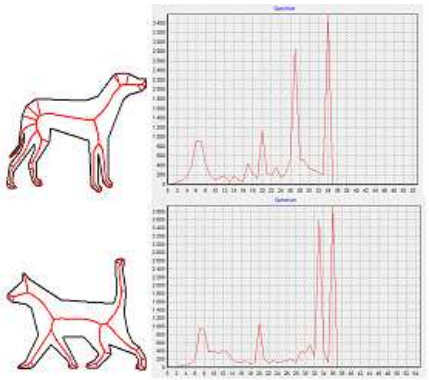


Рис. 6. Скелеты и ДНМС силуэтов животных.

На рис. 239 представлен вычисленный описанным способом дискретно-непрерывный морфологический спектр (ДНМС) многоугольной фигуры (рис. 237). Рядом помещены изображения, иллюстрирующие последовательное изменение площади фигуры при открытии с круглым структурирующим элементом, размер которого соответствует явным пикам построенного спектра.

Заключение

В данной работе предложен новый подход к вычислению морфологических спектров с дисковыми структурирующими элементами, основанный на использовании непрерывных скелетных представлений плоских фигур. Этот подход позволяет формировать дискретно-непрерывные морфологические спектры плоских фигур в реальном времени, что обеспечивает возможность использования их в качестве характеристик формы объектов в прикладных задачах машинного зрения.

На рис. 240, 241 представлены скелеты и ДНМС реальных изображений (силуэтов). Как видно, ДНМС является достаточно информативным описанием формы объектов, устойчивым к существенным её изменениям. Однако метрики сравнения морфологических спектров должны быть нетривиальными и основываться на анализе взаимного расположения пиков спектра. Разработка методов адаптивной обработки и сравнения формы изображений с использованием ДНМС должна стать предметом дальнейшей работы в данном направлении.

Кроме того, в дальнейшем должны быть разработаны аналогичные эффективные процедуры вычисления морфологических спектров для полутоновых изображений, а также ДНМС на основе структурирующих элементов иной формы.

Литература

[1] *Maragos P.* Spectrum, Multiscale Shape Representation. // IEEE Trans.on pattern analysis, machine intelligence, 1989. — Vol. II, No. 7.

[2] *Serra J.* Image Analysis and Mathematical Morphology. // London: Academic Press, 1982.

[3] *Местецкий Л. М.* Непрерывная морфология бинарных изображений. Фигуры. Скелеты. Циркуляры. // Москва: ФИЗМАТЛИТ, 2009.

[4] *Препарата Ф., Шеймос М.* Вычислительная геометрия: Введение. // Москва: Мир, 1989. — 478 с.

Структурное распознавание бинарных изображений с использованием скелетов

Рогов А. А., Быстров М. Ю.
rogov@psu.karelia.ru, maksimkab@yandex.ru
Петрозаводск, Петрозаводский государственный университет

Данная статья посвящена исследованиям по созданию системы поиска в электронных графических коллекциях бинарных изображений. Предлагаемый алгоритм основывается на структурном подходе к распознаванию с использованием скелетов фигур и включает следующие этапы: построение скелета фигуры, регуляризация скелета, аппроксимация скелета прямыми линиями, построение структурного описания скелета (цепочки примитивов) и сравнение цепочек различных фигур между собой. Предлагаются алгоритмы решения каждой из поставленных задач. Рассматриваемый в работе способ построения цепочек примитивов обладает рядом преимуществ и позволяет сравнивать скелеты фигур как целиком, так и частями.

В данной статье рассматривается применение структурного подхода к распознаванию с использованием скелетов при решении задачи поиска в электронных коллекций графических документов (баз данных для хранения изображений). Структурные методы основаны на получении структурно-грамматических признаков, когда в изображении выделяются элементы — признаки, и вводятся правила соединения этих элементов [2]. Анализ и сравнение получаемых последовательностей элементов разных изображений способствует принятию решения. Такой подход обеспечивает высокое быстродействие, т. к. задача распознавания сводится к сравнению символьных структур, а не исходных изображений, что даёт ему неоспоримое преимущество при решении задачи поиска в больших коллекциях графических документов.

Основные определения

Определение 1. *Под бинарным изображением будем понимать черно-белую картинку в растровой решётке, на которой точки объекта являются чёрными (значение цвета равно 1), а точки фона — белыми (значение цвета равно 0).*

В растровой решётке каждая точка имеет целочисленные координаты. Для оценки связности будем использовать 8-смежную структуру соседства [3], где соседними считаются точки, евклидовое расстояние между которыми 1 или $\sqrt{2}$.

Определение 2. *Линии в растровой решетке — связные протяженные фрагменты шириной в один пиксель [3].*

Определение 3. *Множество точек называется связным, если любые две точки множества могут быть соединены последовательностью соседних точек, принадлежащих данному множеству.*

Определение 4. *Под дискретной фигурой (далее фигура) будем понимать максимальное связное множество чёрных точек в растровой решётке, т. е. такое множество, которое не содержится ни в каком другом связном множестве чёрных точек [3].*

Определение 5. *Под «дырой» в фигуре понимается связное множество белых точек, окруженное точками фигуры.*

В описываемом алгоритме рассматриваются только бинарные изображения, содержащие одну фигуру без «дыр» (Рис. 1, а).

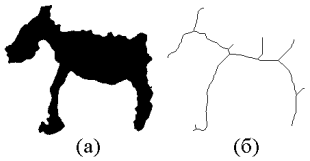


Рис. 1. Бинарное изображение с одной фигурой (а); скелет фигуры (б).

В основе рассматриваемого метода лежит понятие скелета фигуры.

Определение 6. *Скелет фигуры — множество точек, лежащих внутри фигуры и имеющих не менее двух ближайших точек на границе фигуры [4] (Рис. 1, б).*

В нашем случае, в силу дискретности фигуры, получить скелет, соответствующий данному определению, невозможно. Под скелетом дискретной фигуры обычно понимается изображение на растровой решётке, которое приближённо удовлетворяет этому определению. Такой скелет состоит из ребер — линий на растровой решётке. Ребра, которые одним концом соединены с другим ребром, а вторым нет, называются терминальными.

Рассмотрим все этапы предлагаемого алгоритма.

Получение скелета фигуры

Методы получения скелетов фигур можно разделить на 2 категории: использующие непрерывный и дискретный подходы. В рассматриваемом алгоритме структурного поиска для получения скелетов используется известный алгоритм Зонга-Суня [6], который реализует дискретный подход. Алгоритм Зонга-Суня относится к классу методов

топологического утончения фигуры и позволяет получить скелет фигуры на растровой решётке по 8-смежной структуре соседства. Идея метода заключается в последовательном утончении фигуры от границы к её середине путём последовательного перекрашивания чёрных граничных точек в белые. После применения данного алгоритма и некоторой постобработки можно получить искомый скелет дискретной фигуры.

Пример работы данного алгоритма представлен на рис. 2.

Рис. 2. Построение скелета дискретной фигуры с помощью алгоритма Зонга-Суня.

Регуляризация скелета

Обычно скелет содержит множество шумовых ребер, которые не являются существенными при описании общей формы фигуры. Требуется их удаление, так называемая регуляризация скелета [1]. Появление подобных ребер связано с неровностями границы фигуры — в процессе топологического утончения в каждой выпуклости границы возникает новое терминальное ребро скелета, часто шумовое. Процесс регуляризации сводится к последовательному удалению шумовых терминальных ребер.

Предлагается алгоритм регуляризации, основанный на утверждении о том, что по скелету можно восстановить исходную фигуру, если в каждой точке скелета нарисовать круг соответствующего радиуса. Данное утверждение вытекает из определения скелета. В случае дискретной фигуры в качестве радиуса в точке скелета берется минимальное расстояние от этой точки до контура исходной фигуры. При этом фигура восстанавливается приближённо.

Пусть F_0 — восстановленная по скелету фигура, а её площадь — S_0. Под площадью фигуры понимается количество точек, из которых она состоит. Удалим какое-либо терминальное ребро i из скелета и снова восстановим фигуру, используя оставшиеся ребра и старые радиусы (фигура F_i), обозначим её площадь S_i. Форма и площадь фигуры F_i будет отличаться от фигуры F_0. При этом если S_i отличается от S_0 незначительно, то удалённое терминальное ребро является шумовым, т. к. не сильно повлияло (или вообще не повлияло) на площадь. Стоит отметить, что критерий сравнения площадей является достаточным, т. к. F_i всегда будет лежать внутри F_0.

Основываясь на данной идеи, предлагается следующий алгоритм регуляризации скелета.

1. Составляется список всех терминальных ребер скелета T.
2. По скелету восстанавливается исходная фигура и вычисляется её площадь S_0.
3. Выбирается i-е терминальное ребро скелета из списка T.
4. По скелету без ребра i восстанавливается фигура и считается её площадь S_i.
5. Если $(S_0 - S_i)/S_0 < P_R$, где P_R — пороговая величина, то ребро i удаляется из скелета, иначе переход к п. 7.
6. Если после удаления ребра возникли новые терминальные ребра, то поместить их в список T.
7. Если остались нерассмотренные терминальные ребра — переход к п. 3, иначе конец алгоритма.

Пример работы алгоритма регуляризации представлен на рис. 3.

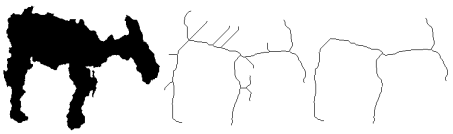


Рис. 3. Работа алгоритма регуляризации скелета.

Аппроксимация скелета

На данном этапе происходит аппроксимация ребер скелета прямыми линиями (Рис. 4, а, б).

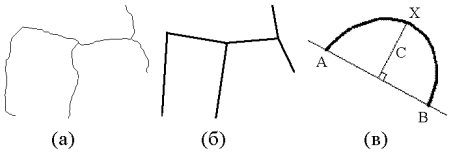


Рис. 4. Аппроксимация скелета.

Для проведения аппроксимации используется метод, идея которого состоит в вычислении кривизны ребер. Кривизна ребра C измеряется максимальным евклидовым расстоянием от точек ребра скелета до прямой, проведённой через его концы A и B (Рис. 4, в): $C = \max(d(X_i, AB))$, $i = 1, \ldots, n$, где X_i — i-я точка ребра, n — количество ребер в скелете, $d(X_i, AB)$ — расстояние от точки X_i до отрезка AB. Далее вводится следующее условие: $C > P_A \times D$, где P_A — пороговая величина, D — диаметр минимальной окружности, описанной вокруг скелета. Если условие выполняется, то ребро делится на два новых ребра в точке, соответствующей значению C, иначе проводится отрезок, соединяющий концы ребра. Новые возникающие в процессе обработки ребра далее аппроксимируются аналогичным способом.

Получение цепочек примитивов

В качестве структурного описания скелета предлагается использовать цепочку примитивов, состоящую из прямых ребер аппроксимированного скелета и углов между ребрами. Для этого делается обхода скелета «против часовой стрелки», как показано на рис. 5.

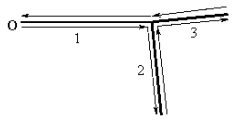


Рис. 5. Получение цепочки примитивов.

Здесь стрелками обозначено направление обхода скелета, O — точка, в которой начинается обход. Следующий алгоритм реализует данную идею и получает цепочку для скелета, состоящего из n ребер:

1. Выбирается терминальная вершина скелета и инцидентное ей терминальное ребро $i = 1$. Вторая, нетерминальная вершина, инцидентная ребру i, объявляется текущей.
2. Вычисляется длина i-го ребра l_i и записывается в цепочку.
3. Среди ребер, инцидентных текущей вершине, выбирается то, которое составляет наименьший угол относительно i-го ребра против часовой стрелки (ребро $i + 1$). Рассматривается также само i-е ребро с углом, равным 360°.
4. Угол α_i между ребрами i и $i + 1$ записывается в цепочку.
5. Если $i + 1 = 2n + 1$, то конец алгоритма. Иначе вершина, инцидентная ребру $i+1$, не являющаяся текущей, объявляется текущей, i увеличивается на 1 и переход на п. 2.

В результате такого обхода каждое ребро скелета рассматривается два раза. Получаемое структурное описание записывается в следующем виде: $\{l_i, \alpha_i\}$, $i = 1, \ldots, 2n$, где l_i — длина i-го элемента, α_i — угол, между i-м и $i + 1$-м элементами, n — количество ребер в скелете.

Длина l_i является относительной величиной в процентах. За 100% берется диаметр минимальной окружности, описанной вокруг скелета.

Цепочка примитивов, описывающая скелет на рис. 5, имеет вид:

$$\{70\,;\,95\}\ \{50\,;\,360\}\ \{50\,;\,90\}$$
$$\{40\,;\,360\}\ \{40\,;\,175\}\ \{70\,;\,360\}.$$

Свойства цепочек примитивов. Предлагаемый способ структурного описания скелетов бинарных изображений в виде цепочки примитивов обладает рядом полезных свойств:

— цепочки устойчивы к масштабированию;
— цепочки устойчивы к повороту фигуры;
— изменение начала обхода соответствует циклическому сдвигу элементов в цепочке;
— цепочка, записанная в обратном порядке, соответствует зеркальному отображению исходной фигуры.

Сравнение цепочек

Конечным этапом алгоритма распознавания является сравнение цепочек примитивов скелетов различных фигур между собой. Если цепочки похожи, то значит и скелеты, и фигуры, из которых они получены, также похожи.

Пусть имеется две цепочки длины $2n$, описывающие разные фигуры: $\{l_i, \alpha_i\}$, $\{k_i, \beta_i\}$, $i = 1, \ldots, 2n$. Тогда введем следующее условие равенства: цепочки равны, если для всех $i = 1, \ldots, 2n$ выполняются условия:

$$|l_i - k_i| \leqslant P_L, \quad |\alpha_i - \beta_i| \leqslant P_C,$$

где P_L и P_C — пороговые величины.

Данное условие имеет один недостаток: для его выполнения необходимо по парное соответствие всех элементов двух цепочек, что возможно только в случае, когда начало обхода скелетов выбрано одинаково. Однако на практике этого добиться очень сложно. Поэтому для сравнения двух цепочек достаточно зафиксировать одну из них и сравнить с $2n$ циклическими сдвигами второй. Если хотя бы одна пара сравниваемых цепочек будет равна, то и исходные цепочки равны между собой.

То есть, если найдется такое $\gamma = 1, \ldots, 2n$, что выполняются условия

$$|l_i - k_j| \leqslant P_L, \quad |\alpha_i - \beta_j| \leqslant P_C,$$

где

$$j = \begin{cases} i + \gamma, & \text{если } i + \gamma \leqslant 2n; \\ i + \gamma - 2n, & \text{если } i + \gamma > 2n. \end{cases}$$

Сравнение цепочек частями. Рассмотренное способ сравнения двух цепочек требует равенства длин цепочек и, соответственно, совпадения форм фигур, из которых они получены, целиком. Однако на практике зачастую существует необходимость поиска общей одинаковой части у двух фигур. Тогда возникает задача нахождения общей части у двух цепочек.

Рассмотрим скелет на Рис. 5. Пусть необходимо получить цепочку скелета, состоящего только из ребер 1 и 3. Для того, чтобы это сделать, необходимо удалить из строки элементы, соответствующие ребру 2 — l_2 и l_3, а также угол, связывающий их между собой α_2: Заметим далее, что угол между ребрами 1 и 3 (снизу) равняется сумме углов между ребрами 1 и 2 и ребрами 2 и 3, а это углы α_1 и α_3 соответственно. Сложив и объединив эти два угла между собой, получим искомую цепочку:

$$\{70\,;\,185\}\ \{40\,;\,360\}\ \{40\,;\,175\}\ \{70\,;\,360\}.$$

В общем случае процесс удаления терминального ребра из цепочки $\{l_i, \alpha_i\},\ i = 1, \ldots, 2n$ можно записать следующей последовательностью действий:

1. Ребро на позиции k в цепочке является терминальным, если $\alpha_k = 360°$, иначе — не терминальным, и удаление невозможно;
2. Для того, чтобы удалить терминальное ребро k, необходимо из цепочки удалить l_k, l_{k+1} и α_k, а α_{k-1} и α_{k+1} сложить между собой.

Рассмотренная операция удаления терминального ребра позволяет для цепочки длины $2n$ получить множество всех подцепочек длины $2r$ $(r \leqslant n)$, описывающих части исходного скелета. Тогда задача поиска общей части длины $2r$ у цепочек длин $2n$ и $2m$ сводится к процессу сравнения всевозможных их подцепочек длины $2r$. Если какая-то пара сравниваемых подцепочек равна, то общая часть найдена.

Апробация алгоритма

Описанный метод распознавания был реализован в виде программного комплекса и протестирован на электронной коллекции изображений петроглифов Карелии [5]. Для тестирования были отобраны 200 черно-белых изображений петроглифов. Результаты работы оценивались субъективно авторами статьи. Для данной коллекции были подобраны оптимальные значения параметров работы алгоритма (таблица 1).

Таблица 1. Параметры алгоритма

Параметр	Обозначение	Значение
Регуляризация	P_R	$0,02$
Аппроксимация	P_A	$0,25$
Сравнение (ребра)	P_L	20
Сравнение (углы)	P_C	30

Точность работы алгоритма была оценена в 70%. В таблице 2 представлена информация о средней скорости работы для изображения 250×250 пикселей.

Таблица 2. Скорость работы

Этап	Время (в мсек)
Получение скелета	15
Регуляризация скелета	100
Аппроксимация скелета	1
Получение цепочки примитивов	1
Сравнение двух цепочек	0,01

Как видно из таблицы, основное время занимают операции получения и регуляризации скелета. Однако для каждого изображения в коллекции можно проделать эту операцию 1 раз, построить цепочки, и в дальнейшем пользоваться ими при поиске. Поэтому в целом можно утверждать, что алгоритм работает достаточно быстро.

Выводы

В данной статье описан алгоритм и его реализация для распознавания бинарных изображений при помощи скелетов с использованием структурного подхода.

Рассмотренный способ построения и сравнения структурных описаний скелетов бинарных изображения в виде цепочек примитивов обладает рядом преимуществ:

1) цепочки устойчивы к масштабированию;
2) цепочки устойчивы к повороту фигуры;
3) цепочка, записанная в обратном порядке, соответствует зеркальному отображению исходной фигуры;
4) существует возможность сравнивать цепочки как целиком, так и отдельные их части;
5) выбор ребра для начала обхода скелета не влияет на результат.

Система успешно прошла апробацию на коллекции изображений петроглифов Карелии [5].

Описанный алгоритм позволяет производить быстрый поиск в электронной коллекции, однако является недостаточно точным. Решением данной проблемы может быть использование двухуровнего поиска — после применения рассмотренного метода поиска, сужающего набор изображений, применяется точный, но медленный метод для окончательного отбора. В дальнейшем планируется усовершенствование алгоритма путём подключения точного алгоритма поиска.

Литература

[1] *Домахина Л. Г.* Регуляризация скелета для задачи сравнения формы // ММРО: доклады XIV всероссийской конференции, Москва: Макс-Пресс, 2009. — С. 342–345.

[2] *Фу К.* Структурные методы в распознавании образов. — Москва: Мир, 1977. — 320 с.

[3] *Местецкий Л. М.* Непрерывная морфология бинарных изображений. Фигуры. Скелеты. Циркуляры. — Москва: ФИЗМАТЛИТ, 2009. — 287 с.

[4] *Местецкий Л. М.* Непрерывный скелет бинарного растрового изображения // Труды международной конференции «Графикон-98», Москва: МГУ, 1998. — С. 71–78.

[5] *Рогов А. А., Рогова К. А., Кириков П. В.* Применение методов распознавания образов в системе управления коллекциями графических документов // Математические методы распознавания образов: доклады XIV всероссийской конференции, Москва: Макс-Пресс, 2009. — С. 429–432.

[6] *Zhang T. Y., Suen C. Y.* A fast parallel algorithm for thinning digital patterns // Commun. ACM. — 1984. — Vol. 27, № 3. — Pp. 236–239.

Идентификация модели ладони по серии её снимков в разных положениях*

Бакина И. Г.
irina_msu@mail.ru
Москва, Московский Государственный Университет имени М. В. Ломоносова

Ранее была предложена параметрическая модель ладони, метод её построения, а также использование этой модели для решения задачи биометрической идентификации личности. В данной работе предлагается метод автоматической идентификации этой модели под каждого человека. А именно, определение кончиков, оснований и точек поворотов пальцев по серии снимков ладони в разных положениях. Работа содержит результаты экспериментов по оценке качества предложенной разметки модели ладони.

Системы распознавания личности, основанные на анализе ладони человека, относятся к числу наиболее старых систем автоматического распознавания личности. Они удобны в использовании, более гигиеничны, менее зависимы от изменений условий окружающей среды. Для измерения признаков ладони обычно не требуется специальных сканеров как, например, при распознавании по отпечаткам пальцев или радужной оболочке глаза. У взрослого человека форма ладони меняется незначительно, за исключением случаев травм и порезов. К недостаткам можно отнести чувствительность таких систем к распознаванию в ситуациях, когда у человека присутствуют кольца, прикрыта часть пястья или какие-то болезни.

Обычно при распознавании предъявляемое (тестовое) изображение сравнивается с эталонными и вычисляется их мера похожести. В условиях регистрации, при которых отсутствуют ограничения на способ позиционирования ладони, выбор такой метрики становится более трудной задачей. Для сравнения ладоней недостаточно простого совмещения изображений, поскольку пальцы в этом случае занимают различные положения при каждом акте регистрации. Для правильного сравнения ладоней нужно по меньшей мере «пошевелить» пальцами таким образом, чтобы они совпали на обоих изображениях.

Ранее в [3] была предложена модель ладони, позволяющая достаточно просто выполнять такие трансформации. Эта модель строится по эталонному изображению ладони, а при сравнении происходит её подгонка под тестовую ладонь. Применение трансформаций к эталону, а не тесту объясняется тем, что тестовое изображение может содержать некоторые дефекты. Например, частично соприкасающиеся пальцы. Также в [3] была предложена функция похожести моделей ладоней и метод её вычисления. Эта функция напрямую зависит от разметки модели ладони — определения положения кончиков, оснований и точек поворотов пальцев. Ранее точки поворотов пальцев просто назначались, исходя из некоторых общих соображений о геометрии ладони. Однако можно получить более высокое качество распознавания ладоней, если эти параметры подбирать для каждого человека индивидуально. Именно такой подход предлагается в данной работе — способ параметризации модели ладони и метод автоматического определения точек поворотов пальцев по серии снимков ладони в разных положениях.

Отметим, что в работе при регистрации человек помещает свою ладонь на горизонтальную однотонную поверхность. Снимки ладоней делаются с помощью недорогой web-камеры, позволяющей уверенно выделять лишь их контура; поэтому для распознавания используются бинаризованные изображения, на которых ладонь представлена чёрными пикселями на белом фоне.

Модель ладони

Модель ладони была предложена ранее в работе [2]. В качестве модели ладони рассматривается её представление в форме гибкого объекта [1]. Напомним, что гибкий объект G есть пара $G = (C, V)$, где C — циркулярный граф (семейство максимальных пустых кругов), а V — множество его допустимых деформаций. Центры кругов из C образуют планарный граф T, который называется осевым графом (скелетом). Силуэтом S гибкого объекта является огибающая семейства кругов C. На рис. 1 приведен пример бинарного изображения ладони и её циркулярного графа. Алгоритм построения циркулярного графа для произвольного бинарного изображения описан в [1].

Трансформации модели. В качестве допустимых трансформаций гибкого объекта рассматриваются следующие:

- сдвиг циркулярного графа (сдвиг ладони);
- поворот циркулярного графа (поворот ладони);
- поворот ветвей циркулярного графа (поворот пальцев).

Первые две трансформации могут быть достаточно просто применены к любому гибкому объекту. Для применения третьей трансформации необходимо иметь дополнительную информацию о структуре анализируемого объекта. В частности, нужно

Работа выполнена при финансовой поддержке РФФИ, проекты № 10-07-00609, 11-01-00783 и 11-07-00462

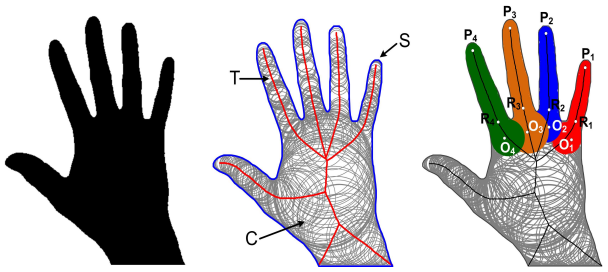


Рис. 1. Бинарное изображение ладони (слева), его циркулярный граф (по центру) и модель ладони (справа).

уметь выделять пальцы и определять положение их точек поворотов.

Параметризация модели. Обозначим через P_i положение кончиков пальцев, через O_i — оснований пальцев, через R_i — точек поворотов пальцев. Здесь мы рассматриваем все пальцы за исключением большого (рис. 1 справа). Считается, что к пальцу относится часть циркулярного графа, содержащая ветвь осевого графа (со всеми кругами) от вершины, ближайшей к точке O_i, до концевой вершины этой ветви. Именно эта часть графа поворачивается относительно точки O_i при «шевелении» пальцев. На рис. 1 справа для каждого пальца ладони закрашена выделенная для него область на циркулярном графе.

Таким образом, предлагается определять модель ладони следующим вектором параметров, составленном из 12 компонент: $\boldsymbol{p} = \{(P_i, O_i, R_i)\}_{i=1}^{4}$. Конкретные значения данных параметров должны быть определены для каждой модели в отдельности. Так, например, разметка модели может быть вручную выполнена экспертом. Однако в случае большого количества пользователей в базе этот процесс может занимать значительное время. Поэтому далее предлагается метод автоматической разметки модели ладони, т. е. её идентификации для отдельно взятого пользователя.

Идентификация модели ладони

Задача идентификации модели ладони ставится следующим образом. Имеется n неразмеченных моделей ладони, построенных для исходных n эталонных изображений ладони одного человека. Для каждой модели необходимо определить её вектор параметров $\boldsymbol{p}$.

Построение разметки модели ладони состоит из двух основных шагов:

1) выделение пальцев;
2) определение точек поворотов пальцев.

Первый шаг может быть выполнен независимо для каждой из n моделей ладони. Для выполнения второго шага предлагаются два метода: один позволяет назначать точки поворотов пальцев отдельно для каждой модели, а второй определяет положение этих точек на основании информации обо всех моделях сразу.

Выделение пальцев. Метод выделения пальцев, а именно алгоритм определения их кончиков и оснований был предложен в [2]. Он основан на последовательном просмотре всех вершин ветви циркулярного графа, относящейся к пальцу, и вычислении нескольких локальных характеристик. Пример поиска кончика и основания пальца приведена на рис. 2 слева. Как и раньше, P_1, P_2, P_3 и P_4 — найденные кончики пальцев, а R_1, R_2, R_3 и R_4 — основания пальцев. Прямая R_iP_i, проходящая через выделенные точку основания R_i и кончик пальца P_i, называется *осью пальца*.

Определение точек поворотов. Если посмотреть на строение ладони, то можно отметить, что для пальца наиболее подвижным является основание проксимальной фаланги (за исключением большого пальца). В работе рассматривается два способа определения точек поворотов пальцев — простое назначение точек (по одной модели ладони) и оптимизационный (на основе нескольких моделей ладони). Рассмотрим каждый из них в отдельности. В обоих подходах подразумевается, что точка поворота пальца O_i находится на его оси R_iP_i.

Назначение точек. В качестве точки поворота пальца назначается точка на оси пальца, отстоящая от его основания на $k\%$ его длины, где k — параметр метода (например, в [2] рассматривается случай $k = 30$). Эксперименты показали, что такое определение точки поворота с параметром $k = 30 - 40$ позволяет получать адекватные результаты при моделировании поворотов пальцев.

Оптимизационный подход. При назначении точек поворотов пальцев достаточно одного изображения ладони. В случае, когда имеется несколько моделей ладони в разных положениях, задача определения точек поворотов пальцев может рассматриваться в другой постановке.

Пусть имеется n неразмеченных моделей ладони одного человека и, соответственно, n гибких объектов: $G_1, G_2, \ldots, G_n$. Считается, что точка поворота пальца находится на его оси. Обозначим через $\boldsymbol{l} = (l_1, l_2, l_3, l_4)$ вектор параметров, определяющих расстояние от кончика каждого пальца до точки поворота этого пальца вдоль его оси (рис. 2 по центру). Пусть также задана функция $\mu^* = \mu^*(G_1, G_2)$, определяющая степень похожести гибкого объекта G_1 на гибкий объект G_2. Эта функция вводится ниже. Значение этой функции неявно зависит от выбора точек поворотов пальцев ладони, т. к. при её вычислении проводятся трансформации циркуляра G_1 с целью получения наилучшего совмещения силуэтов G_1 и G_2. В число таких трансформаций входят также повороты паль-

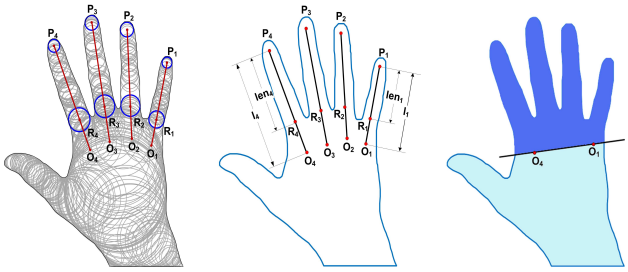


Рис. 2. Разметка ладони (слева), параметризация (по центру) и выделение области для сравнения (справа).

цев. Чтобы указать, что рассматривается гибкий объект G с некоторым выбранным положением точек поворотов пальцев $\boldsymbol{l}$, будем писать $G(\boldsymbol{l})$.

Понятно, что ладони одного человека должны быть максимально похожи. В качестве интегральной меры похожести n ладоней будем рассматривать среднее значение попарных расстояний:

$$\mu_{\text{ср}}(\boldsymbol{l}) = \sum_{\substack{i,j=1,\dots,n \\ i \neq j}} \frac{\mu^*(G_i(\boldsymbol{l}), G_j(\boldsymbol{l}))}{n(n-1)}$$

Здесь учтено, что функция μ^* является несимметричной, т. е. в общем случае $\mu^*(G_1, G_2) \neq \neq \mu^*(G_2, G_1)$, и число попарных расстояний равно $n(n-1)$.

Значение $\mu_{\text{ср}}(\boldsymbol{l})$ напрямую зависит от того, каким образом выбираются точки поворотов пальцев. Тогда можно поставить следующую оптимизационную задачу: найти такой набор параметров $\boldsymbol{l^*}$, при котором $\mu_{\text{ср}}(\boldsymbol{l})$ достигает своего минимума:

$$\mu_{\text{ср}}(\boldsymbol{l^*}) = \min_{\boldsymbol{l} \in \mathbb{R}^4} \mu_{\text{ср}}(\boldsymbol{l}) \qquad (1)$$

При решении этой задачи могут быть использованы методы как условной, так и безусловной оптимизации. Например, область поиска может быть ограничена пространственным параллелепипедом: $l_i \in [0.2\, len_i, 0.5\, len_i], i = 1, \dots, 4$, где $len_i = |P_i R_i|$ — длина i-ого пальца ладони (рис. 2 по центру). Отметим, что данная задача является многоэкстремальной.

Задача (1) была решена двумя способами: полным перебором $\boldsymbol{l}$ с шагом $0.05\, len_i$ в рамках указанного выше пространственного параллелепипеда, а также методом Нелдера-Мида [4]. Начальный симплекс для метода Нелдера-Мида брался из параллелепипеда $l_i \in [0.35\, len_i, 0.45\, len_i], i = 1, \dots, 4$. Параметры метода: коэффициент отражения $\alpha = = 1$, коэффициент сжатия $\beta = 0.5$ и коэффициент растяжения $\gamma = 2$.

Эксперименты, приведённые ниже, показали, что поиск точек поворотов пальцев на основе решения задачи минимизации функции $\mu_{\text{ср}}(\boldsymbol{l})$ приводит к меньшим ошибкам ложного отказа (FRR — *False Reject Rate*) и ложного узнавания (FAR — *False Accept Rate*) при распознавании личности. При этом использование метода Нелдер-Мида позволяет ускорить этот поиск без существенного влияния на качество распознавания.

Таким образом, может быть выполнена полная разметка ладони — определены положения пальцев и их точки поворотов (рис. 2 слева). Предложенная разметка строится offline по заранее собранному множеству бинарных изображений ладони человека. Желательно, чтобы это множество содержало наиболее часто встречающиеся положения ладони человека, т. к. от этого напрямую зависит выбор точек поворотов в оптимизационном подходе.

Сравнение моделей

Рассмотрим функцию μ^*, отражающую степень похожести сравниваемых ладоней. Процесс её построения и вычисления был предложен в [3] и состоит из двух шагов:

1) начальная укладка моделей;
2) подгонка моделей.

Рассмотрим два гибких объекта G_1 и G_2, первый из которых является эталоном, а второй — тестом. Начальная укладка заключается в наложении моделей друг на друга путём совмещения кончиков P_3 и осей P_3R_3 средних пальцев. После чего вычисляется значение функции mu:

$$\mu(G_1, G_2) = \text{Area}(S_1 \backslash S_2) + \text{Area}(S_2 \backslash S_1)$$

Здесь S_1 и S_2 — силуэты сравниваемых гибких объектов, а $Area$ — площадь фигуры, измеряемая в квадратных пикселях. С целью исключения влияния разницы в форме запястья человека и подвижности кожи в области большого пальца при сравнении учитывается лишь область, лежащая выше прямой O_1O_4 эталонной ладони. Эта область выделена тёмным цветом на рис. 2 справа.

На этапе подгонки проводятся трансформации эталонного гибкого объекта G_1 с целью получения наилучшего совмещения силуэтов G_1 и G_2 с точки зрения функции μ. А именно, решается следующая оптимизационная задача:

$$\mu^*(G_1, G_2) = \min_{v_1 \in V_1} \mu(v_1(G_1), G_2)$$

Эксперименты

Для проведения экспериментов была собрана база ладоней группы людей. Использовалась web-камера с невысоким разрешением, закрепленная над столом на фиксированном расстоянии. Стол представлял из себя однотонную поверхность, на которой человек помещал свою ладонь тыльной стороной вверх. Таким образом, было собрано 97 изображений ладоней 22 человек.

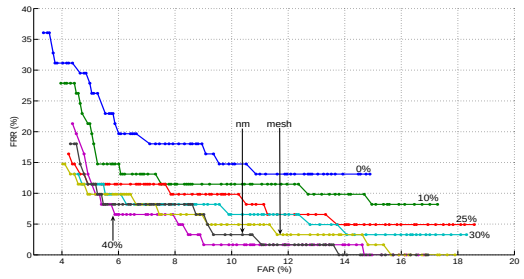


Рис. 3. Сравнение результатов верификации при простом назначении точек поворотов пальцев и в случае оптимизационного подхода.

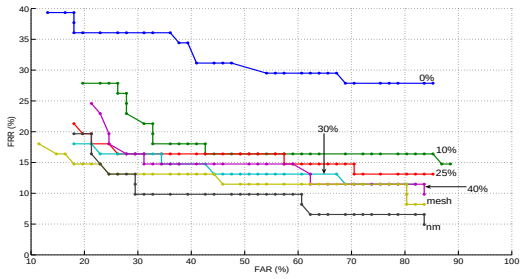


Рис. 4. Сравнение результатов идентификации при простом назначении точек поворотов пальцев и в случае оптимизационного подхода.

Собранные изображения ладоней были вручную разделены на две группы: 45 были отнесены к эталонам (2–3 для каждого человека) и 52 к контролю (2–3 для каждого человека). Качество разметки модели ладони оценивалось на основе экспериментов по идентификации и верификации личности. При распознавании использовалось простое пороговое правило. Так как ладонь каждого пользователя была представлена несколькими моделями, то в качестве меры похожести тестовой ладошки на ладонь этого человека бралось минимальное расстояние между этой ладонью и каждой из ладоней этого человека.

На рис. 3 представлены результаты верификации при простом назначении точек поворотов пальцев и в случае оптимизационного подхода. Условные обозначения такие: пять кривых — назначение точек с параметром k равным соответственно 0, 10, 25, 30 и 40; другие две — оптимизационный подход при полном переборе значений $\boldsymbol{l}$ с шагом в 5% (кривая с меткой "mesh") и минимизации методом Нелдера-Мида (кривая с меткой "nm").

Как видно, в случае верификации простое назначение и оптимизационный методы определения точек поворотов пальцев ведут себя примерно одинаково. Однако в случае идентификации (рис. 4), когда тестовая ладонь сравнивается с ладонями всех пользователей, а не одного человека, как при верификации, оптимизационный подход показывает свое явное преимущество. При этом оптимизация методом Нелдера-Мида позволяет получать лучшее качество распознавания по сравнению с перебором всех значений $\boldsymbol{l}$ на сетке.

Ошибки FRR и FAR остаются высокими в случае идентификации. Но они, например, могут быть снижены путём сведения задачи идентификации к задаче верификации при использовании дополнительной модальности. Так, в [3] предлагается метод комбинирования признаков голоса и ладони человека. Голосовые признаки служат фильтром заведомо непохожих людей, а идентификация проводится уже в небольшой группе (2–3 человека).

Выводы

В работе предложен способ параметризации модели ладони, а также метод идентификации её параметров для отдельно взятого человека. В качестве входной информации используются одно или несколько бинарных изображений ладони человека. Разметка модели включает в себя процедуру выделения пальцев и определения их точек поворотов. Предложены два подхода к определению точек поворотов — простое назначение и оптимизационный подход (по группе ладоней в разных положениях). Представлены эксперименты, демонстрирующие преимущество оптимизационного подхода. Оценка проводилась на основе вычисления FRR и FAR при верификации и идентификации личности с заданной разметкой модели.

В дальнейшем планируется расширить множество допустимых трансформаций ладони, включив в него возможные движения большого пальца; и учитывать его характеристики при построении функции похожести моделей. Также планируется провести эксперименты на большей выборке ладоней.

Литература

[1] *Местецкий Л. М.* Непрерывная морфология бинарных изображений: фигуры, скелеты, циркуляры. — Москва: ФИЗМАТЛИТ, 2009. — 288 с.

[2] *Бакина И. Г., Местецкий Л. М.* Метод сравнения ладоней при наличии артефактов // Доклады 14-ой Всероссийской конференции Математические методы распознавания образов (ММРО-14), Москва: МАКС Пресс, 2009. — С. 301–304.

[3] *Bakina I. G.* Palm Shape Comparison for Person Recognition // Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP 2011), INSTICC, 2011. — Pp. 5–11.

[4] *Nash J. C.* Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation. — International Series on Biometrics. — Adam Hilger, 1990.

[5] *Ross A. A., Nandakumar K., Jain A. K* Handbook of multibiometrics. — International Series on Biometrics.

Распознавание жестов ладони с помощью непрерывного скелета*

Куракин А. В.
alekseyvk@yandex.ru
Москва, Московский физико-технический институт (государственный университет)

В статье рассмотрено использование непрерывного скелета для анализа формы ладони и распознавания жестов на плоскости и в пространстве. Скелет позволяет выполнять сложный анализ формы и измерять широкий диапазон признаков на силуэте ладони. А использование эффективных алгоритмов для построения и анализа скелета позволяет использовать предложенные методы в реальном времени.

В настоящее время задача распознавания жестов активно исследуется в силу её важного практического значения. Потенциальные применения технологий распознавания жестов включают человеко-машинное взаимодействие, приложения виртуальной реальности, распознавание языка жестов и другие.

В литературе фигурирует большое количество подходов к этой задаче [1]. Часть методов требует для работы специального оборудования или роботизированных перчаток. Есть методы, в основе которых лежит максимизация вероятности того, что рука примет ту или иную позу, при условии заданного входного изображения, однако они вычислительно сложны и неприменимы в реальном времени. Также существуют методы, работающие в реальном времени только за счёт анализа изображения ладони. Но в большинстве таких подходов спектр жестов существенно ограничен из-за скудного механизма выделения признаков.

В данной работе рассмотрено использование непрерывного скелета для анализа формы ладони в задаче распознавания жестов. Непрерывный скелет позволяет измерять широкий диапазон признаком на силуэте ладони; более того, алгоритмы построения и анализа скелета работают в реальном времени.

Понятие скелета

Рассмотренные в статье методы анализа формы базируются на понятии непрерывного скелета фигуры. Определение непрерывного скелета вместе со способом получения кратко описано в этом разделе. Более детальную информацию по непрерывным скелетам можно найти в книге [2].

Определение 1. *Многоугольной фигурой будем называть связное множество точек на плоскости, ограниченное конечным числом непересекающихся простых многоугольников.*

Определение 2. *Максимальный пустой круг, или максимальный вписанный круг — это круг B, полностью содержащийся внутри фигуры F, такой, что любой другой круг $B' \subset F, B' \neq B$ не содержит в себе B.*

Работа выполнена при финансовой поддержке РФФИ, проекты № 11-01-00783-а и № 11-07-00462-а.

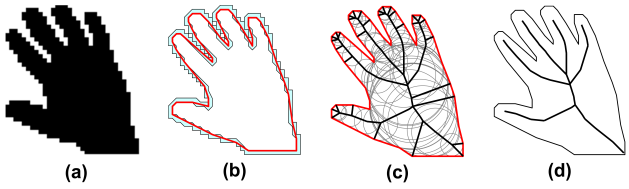


Рис. 1. Процесс построения скелета. (a) исходная бинарная картинка; (b) построение многоугольной аппроксимации; (c) скелет многоугольника; (d) скелет после стрижки.

Определение 3. *Скелетом многоугольной фигуры F является множество центров её максимальных пустых кругов.*

Определение 4. *Радиальная функция ставит в соответствие каждой точке скелета значение радиуса максимального пустого круга с центром в этой точке.*

Можно доказать [3], что скелет многоугольной фигуры состоит из объединения конечного числа отрезков и дуг парабол. Таким образом скелет многоугольной фигуры можно рассматривать как планарный граф, вершины которого — это точки соединения отрезков и дуг парабол, а ребра — собственно отрезки и дуги парабол составляющие скелет. Степень любой вершины в таком графе будет равна 1, 2 или 3.

Для многоугольных фигур существуют эффективные алгоритмы, которые позволяют строить скелет за время $O(N \log N)$, где N — количество вершин. На практике обычно приходится иметь дело с бинарными растровыми изображениями, на которых пиксели одного цвета обозначают фигуру, а второго — фон. В таком случае предварительно строится многоугольная аппроксимация изображенной фигуры, после чего уже выполняется построение скелета. Кроме этого, после того, как скелет построен, выполняется его регуляризация (стрижка) с целью удаления шумовых и малозначимых ветвей. Весь процесс построения скелета проиллюстрирован на рис. 1.

Ветви скелета

Анализ формы ладони производился за счёт анализа ветвей скелета — частей скелета, рассмотренных как непрерывные кривые.

Определение 5. *Рассмотрим непрерывную кусочно-гладкую кривую без самопересечений* $\boldsymbol{s}(\bullet)$: $s(l) = \{x(l), y(l)\}$, $l \in [0, L]$, *и пусть* l *является естественной параметризацией кривой (т. е. длиной дуги кривой). Пусть каждая точка кривой* $\boldsymbol{s}(\bullet)$ *является одновременно и точкой скелета. В таком случае кривую* $\boldsymbol{s}(\bullet)$ *будем называть ветвью скелета, соединяющей точки скелета* $r(0)$ *и* $r(L)$.

Для каждой точки скелета с координатами (x, y) известно значение радиальной функции $R(x, y)$, равное радиусу максимального пустого круга с центром в этой точке. Следующим образом определим радиальную функцию вдоль ветви скелета:

Определение 6. *Для ветви скелета* $\boldsymbol{s}(\bullet)$ *функцию* $R_s(l) = R(\boldsymbol{s}(l))$, $l \in [0, L]$ *будем называть радиальной функцией вдоль ветви* $\boldsymbol{s}(\bullet)$.

Можно доказать, что радиальная функция вдоль ветви будет непрерывной кусочно-гладкой функцией.

Из-за наличия дуг-парабол работать с радиальной функцией вдоль ветви неудобно. Однако у скелетов реальных изображений дуги парабол очень короткие, имеют малую кривизну и приближённо могут быть рассмотрены как отрезки. Соответственно, заменив все параболические дуги на отрезки, получим новый аппроксимированный скелет, радиальные функции вдоль ветвей которого будут кусочно-линейными.

Такую радиальную функцию будем называть *линейно аппроксимированной радиальной функцией*, и для ветви, проходящей через последовательность вершин $V_0, V_1, \ldots, V_{n-1}, V_n$, она может быть вычислена следующим образом:

$$\tilde{R}_{\tilde{s}}(l) = \begin{cases} R_i, & l = L_i, \\ \alpha R_i + (1-\alpha) R_{i+1}, & l = \alpha L_i + (1-\alpha) L_{i+1}, \end{cases}$$

где $\alpha \in (0, 1)$, $R_i = R(V_i)$ — значения функции радиуса в вершинах скелета, а $L_i = \sum_{k=0}^{i-1} |V_k V_{k+1}|$.

Детектирование пальцев

Поиск пальцев на силуэте ладони производится путём анализа ветвей скелета, оканчивающихся в висячих вершинах скелетного графа, а также радиальных функции вдоль них.

Пусть A является произвольной висячей вершиной скелетного графа. Пусть B является ближайшей (в терминах расстояния вдоль графа) к A вершине скелетного графа, имеющей степень 3.

Рассмотрим ветвь скелета $s(\bullet)$, порождённую кратчайшим путём P между вершинами A и B, и значение линейно аппроксимированной радиальной функции $R_{\tilde{s}}(l)$ вдоль этой ветви. Примеры радиальных функций вдоль ветви, соответствующих

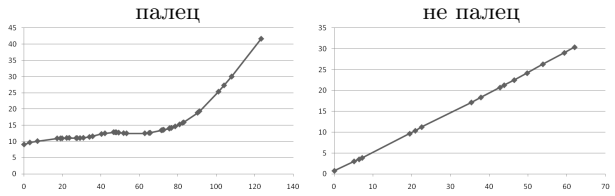


Рис. 2. Радиальная функция для пальца (слева) и для ветви, не являющейся пальцем (справа).

пальцу, и ветви, не являющейся пальцем, приведены на рис. 2. Данный рисунок демонстрирует, что особенности радиальной функции вдоль ветви могут использоваться для определения пальцев.

Собственно алгоритм детектирования пальцев по скелету силуэта ладони состоит в следующем:

1) анализируем все ветви скелета соединяющие висячие вершины скелета с ближайшими вершинами степени 3;
2) для каждой такой ветви выполняем поиск точки C, которая является наиболее вероятным место сочленения пальца и пясти;
3) когда определена точка C, проверяем набор критериев на геометрические размеры (длину, толщину) ветви, чтобы отсечь те ветви, которые не являются пальцами.

Будем использовать обозначения из предыдущего раздела ($V_0 \ldots, V_n$, $R_0 \ldots, R_n$ и $L_0 \ldots, L_n$), и ещё введем дополнительно величину D_i как дискретную частную производную R по L. Положим $D_0 = 0$, $D_n = +\infty$, а в остальных точках будем D_i вычислять по формуле:

$$D_i = \frac{R_{i+1} - R_{i-1}}{L_{i+1} - Li - 1}, \quad i = 1, \ldots, n-1.$$

Поиск точки C выполняется из тех соображений, что в момент, когда заканчивается палец и начинается пясть, происходит выполнение одного из следующих условий:

— R увеличивается более чем в 2–2,5 раза по сравнению с началом пальца;
— радиус начинает резко расти, т. е. частные производные D_i превосходят наперед заданный порог (использовалось значение порога, равное 0,4–0,6).

После того, как для сегмента AB найдена точка C — вероятного сочленения пальца и ладони, выполняется вычисление длин сегментов AC и AB и толщины пальца (радиус максимальной вписанной окружности в определённой точке скелета, либо среднее значение радиуса вдоль AC). Сегмент AB классифицируется как палец, если выполняются все следующие условия:

— $|AC|/|AB| \geqslant 0.35$;
— толщина ветви AC должна быть в заданных пределах;

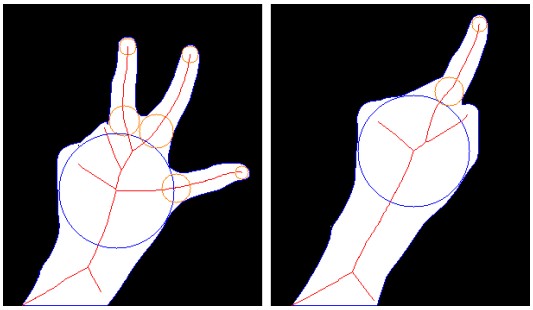


Рис. 3. Пример детектирования пальцев.

— длина $|AB|$ должна быть больше порога (т. е. палец должен быть достаточно длинным).

На рисунке 3 приводится пример результата работы алгоритма детектирования пальцев. Большой круг — это максимальный круг, вписанный в ладонь, его центр полагается центром ладони. А маленькие круги соответствуют кончикам пальцев и местам сочленения пальцев и ладони.

Анализ двухмерных жестов

С помощью алгоритма описанного в предыдущем разделе можно выполнять распознавание простого набора жестов по силуэту ладони. В частности, можно выполнять детектирование жестов, состоящих из 1–5 пальцев, выполнять определение положения кончиков видимых пальцев, положение центра ладони (как максимального вписанного круга, имеющего самый большой радиус), а также определять наличие кольца, образованного двумя пальцами.

В разработанном демонстрационном стенде набор простых жестов использовался для управления объектами на экране компьютера. В частности использовались жесты состоящие из одного, двух и трёх пальцев для перемещения, масштабирования и вращения объектов. Жест-кольцо использовался для захвата объектов, а жест из пяти пальцев — для опускания.

Жесты используемые для перемещения, масштабирования и вращения засчитывались только в случае, когда рука двигалась как целое. Это позволило отличать собственно жесты от ситуаций перехода от одного жеста к другому, например, когда человек сгибает или разгибает палец.

Жест-кольцо и жест из пяти пальцев распознавались как динамические жесты, т. е. измерялось время наблюдения жеста, и он засчитывался только если время было больше заданного порога. Динамическое распознавание жестов возможно благодаря высокой скорости обработки кадров.

Анализ трёхмерных жестов

Достаточно широкий класс объектов (например, тела человека, животных, ладонь человека) могут приближённо считаться локально симметричными. В таком случае объект может быть описан в виде объединения набора т. н. пространственных жирных кривых.

Определение 7. *Рассмотрим гладкое отображение* $C\colon [a,b] \to \mathbb{R}^3 \times [0,\infty)$ *отрезка* $[a,b]$ *в множество шаров. Каждому значению параметра* $t \in [a,b]$ *соответствует шар* C_t *с центром в точке* $P_C(t) = (x_C(t), y_C(t), z_C(t))$ *и радиусом* $r_C(t)$. *Объединение всех шаров* $\bigcup_{t\in[a,b]} C_t$ *будем называть пространственной жирной кривой с осью* $P_C(t)$ *и шириной* $r_C(t)$.

Пусть пространственный объект, являющийся объединением жирных кривых, снимается камерой. Рассмотрим те части объекта, при проецировании которых на плоскость не возникает окклюзий, т. е. прообразом точки плоскости является либо одна точка исходного объекта, либо отрезок, полностью принадлежащий объекту. Если для силуэтов проекций этих частей объектов построить скелет, то он с высокой точностью будет совпадать с проекцией соответствующих пространственных осей исходного объекта [4]. Это наблюдение можно использовать для восстановления положения частей трёхмерного объекта по его проекциям на плоскости. Ниже показано применение этой идеи к восстановлению положения частей ладони.

Рассмотрим систему, состоящую из двух откалиброванных камер. Пусть с помощью этих камер получена пара изображений ладони, и для каждого изображения выделен силуэт ладони. В итоге имеем стереопару силуэтов S_1 и S_2. Используя следующий алгоритм, можно восстановить положение кончиков пальцев и центра ладони в пространстве.

1) Для силуэтов S_1 и S_2 выполняется построение и стрижка скелетов M_1 и M_2.
2) Каждый скелет M_i анализируется с целью выделения положения кончиков пальцев $\{A^i_j = (x^i_j, y^i_j),\ j = 1,\dots,N_i\}$ и центра ладони $A^i_c = (x^i_c, y^i_c)$.
3) Для множеств точек $\{A^1_j\}$ и $\{A^2_j\}$ выполняется определение соответствия между ними. В результате получается набор стереопар кончиков пальцев. При этом для установления соответствия используются ограничения, накладываемые эпиполярной геометрией [5], а также тот факт, что ориентация пальцев относительно центра ладони должна быть одинаковой на обоих кадрах.
4) Для стереопар кончиков пальцев, а также для стереопары центров ладоней выполняется вычисление трёхмерных координат [5].

Следует отметить, что алгоритм будет работать и при наличии частичных окклюзий при проецировании ладони. В такой ситуации будут восста-

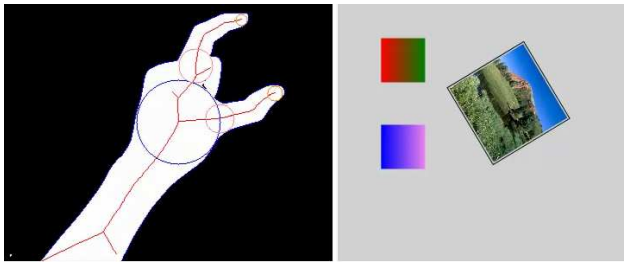


Рис. 4. Пример работы системы для распознавание жестов в двух измерениях. На рисунке изображен жест, используемый для масштабирования.

новлены положения только тех кончиков пальцев, которые проецируются на плоскости камер без окклюзий.

Эксперименты

Для экспериментов и демонстрации вышеописанных методов был разработан программный комплекс, состоящий из двух программных систем. Одна из программных систем выполняла распознавание жестов руки посредством анализа изображения полученного с одной веб-камеры. Вторая выполняла анализ изображения с двух веб-камер и определяла положение руки и кончиков пальцев в трёхмерном пространстве.

Для двухмерного распознавания жестов использовалась следующая схема экспериментальной установки. Обычная веб-камера (Logitech 9000) располагалась над однородной тёмной поверхностью. Пользователь двигал рукой перед поверхностью, и изображение снималось камерой. За счёт попиксельных методов детектирования кожи [6] выделялся силуэт ладони. По силуэту ладони строился скелет и выполнялось распознавание жестов. Обнаруженные жесты использовались для перемещения, вращения и масштабирования объектов на экране компьютера. На рисунке 4 приведен снимок экрана данной программы.

Для восстановления положения ладони и пальцев в трёхмерном пространстве использовалась пара из откалиброванных веб-камер. В остальном экспериментальная установка повторяла двумерный вариант. На паре изображений, полученных с камер, выделялись силуэты руки и производилось детектирование кончиков пальцев и центра ладони, после чего вычислялись их координаты в трёхмерном пространстве. На рисунке 5 приведен пример работы системы.

Эффективные алгоритмы построения и стрижки скелета позволяют использовать системы для слежения за рукой в реальном времени. Для примера, однопоточная реализация описанного тестового стенда обрабатывает один кадр за 22 мс на 2.4 Ghz Intel Core 2 Quad CPU.

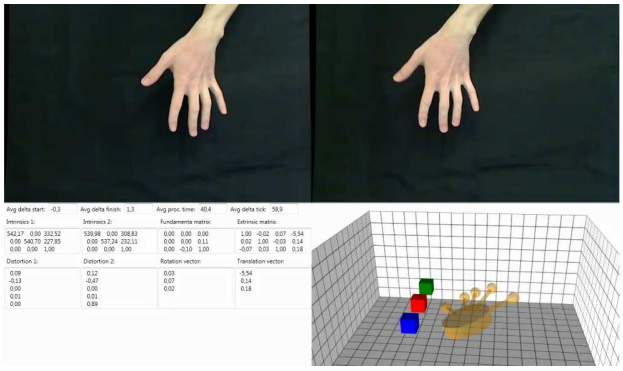


Рис. 5. Пример работы системы для слежения за рукой в трёхмерном пространстве.

Выводы

В работе рассмотрены методы анализа формы ладони и распознавания жестов за счёт применения непрерывного скелета. Непрерывный скелет позволяет выполнять сложный анализ формы и топологии руки, определять количество и положение видимых пальцев. Кроме того, метод применим и для определения положения руки в пространстве по стереопаре силуэтов. Более того, методы работают в реальном времени, что делает их применимыми в практических приложениях.

В дальнейшем предполагается расширение подхода за счёт привлечения модели руки, что позволит восстанавливать «позу» руки при наличии окклюзий.

Литература

[1] *Garg P., Aggarwal N., Sofat S.* Vision Based Hand Gesture Recognition. World Academy of Science Engineering and Technology, 2009. — Pp. 972–977.

[2] *Местецкий Л. М.* Непрерывная морфология бинарных изображений: фигуры, скелеты, циркуляры. — Москва: Физматлит, 2009.

[3] *Mestetskiy L.* Skeleton representation based on compound Bezier curves. VISAPP 2010: Proc. of 5th Int. Conf. on Computer Vision Theory and Applications, 2010. — Vol. 1. — Pp. 44–51.

[4] *Pillow N., Utcke S., Zisserman A.* Viewpoint-invariant representation of generalized cylinders using the symmetry set. 5th British Machine Vision Conference. Image and Vision Computing, June 1995. — Vol. 13, Iss. 5.

[5] *Hartley R. I., Zisserman A.* Multiple View Geometry in Computer Vision. — 2nd edition. — Cambridge University Press, 2004.

[6] *Phung S. L., Bouzerdoum A., Chai D.* Skin Segmentation Using Color Pixel Classification: Analysis and Comparison. IEEE Trans. Pattern Anal. Mach. Intell., 2005. — Vol. 27, No. 1. — Pp. 148–154.

Сегментация с моделью формы на основе циркулярного графа

Янгель Б. К., Ветров Д. П.
hr0nix@acm.org, vetrovd@yandex.ru
Московский государственный университет им. М. В. Ломоносова

В настоящей работе предложен подход к использованию глобальных ограничений на форму объекта в задаче сегментации изображений. Также предложена модель формы, основанная на упрощённом циркулярном графе объекта. Модель позволяет учитывать вариабельность формы объекта, а также задавать глобальные ограничения на его размер и местоположение на изображении.

Сегментация изображений является одной из важных задач компьютерного зрения. Современные методы сегментации, как правило, опираются на локальную низкоуровневую информацию об изображении, такую как цвет пикселей или модуль градиента яркости. Примером может служит сегментация путём минимизации энергии случайного марковского поля (Markov Random Field, MRF) [1]. Подобные методы хорошо работают в случае, когда имеется дискриминативная модель цвета объекта, а сам объект на изображении обладает чёткими границами. В случае размытых границ или высокого уровня шума методы, основанные только на низкоуровневой информации, могут давать сбои. Один из способов решения этой проблемы — дополнительно использовать высокоуровневую информацию, например, знания о предполагаемой форме сегментируемого объекта, тем более что такая информация обычно известна заранее. На вопрос о том, в каком виде представлять и как эффективно учитывать ограничения на форму в оптимизационной задаче сегментации наряду с локальными ограничениями, тем не менее, окончательного ответа пока не получено.

В работах [10, 4] предложено описывать форму объекта бинарной маской, а сегментацию производить итеративно, обновляя предполагаемую позицию объекта на изображении на каждой итерации. Несмотря на то, что такое описание формы позволяет на каждой итерации решать возникающую оптимизационную задачу эффективно, жёсткая маска не пригодна для классов объектов со значительной вариацией формы. В ряде работ предложено ввести весьма общие ограничения на форму объекта, такие как «звездность» [9] или «натянутость» [7]. Такие ограничения, напротив, не позволяют достаточно ограничить форму сегментируемого объекта. В нескольких работах объект описывается как совокупность частей, в которой каждая часть имеет свою специфическую форму, а вместе они образуют гибкую конфигурацию [5, 2]. Предложенная в настоящей работе модель формы, по мнению авторов, имеет большую гибкость за счёт того, что позволяет моделировать неоднородные изменения масштаба частей объекта, а также ряд глобальных ограничений на ориентацию и размер объекта.

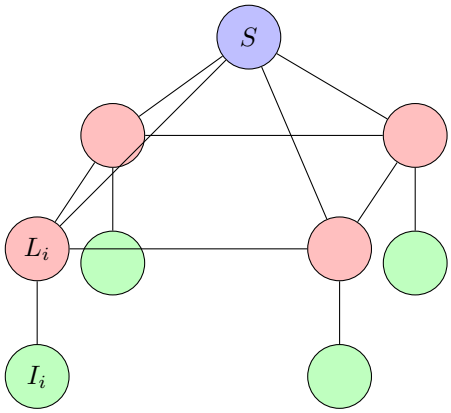


Рис. 1. Графическая модель для сегментации с моделью формы.

Далее мы будем рассматривать задачу сегментации изображения на 2 класса: объект и фон.

Сегментация с моделью формы

Предложенный в настоящей работе подход к сегментации с моделью формы основан на вероятностной модели, изображенной на рисунке 1. В рамках данной модели случайная величина S, соответствующая параметрам формы объекта, порождает бинарные метки пикселей L_i, $L_i = 1$ соответствует объекту, а $L_i = 0$ — фону. Метки, в свою очередь, независимо порождают пиксели изображения I_i согласно цветовой модели объекта и фона. Метки пикселей порождаются формой не независимо, а так, чтобы соседние пиксели имели одинаковую метку с большей вероятностью (за это отвечают ребра между переменными — метками).

Сегментация как покоординатный спуск. Совместное распределение переменных в предложенной модели может быть записано как

$$P(S, L, I) = \frac{1}{Z} f(S) \times$$
$$\times \prod_i h_i(I_i, L_i) \prod_{(i,j)\in\mathcal{N}} \varphi_{ij}(L_i, L_j) \prod_i \varphi_i(L_i, S).$$

где $\mathcal{N}$ — модель соседства пикселей, потенциальные функции h_i отвечают за связи между метками и пикселями, φ_{ij} отвечают за связи между парами меток, φ_i — за связь между меткой и формой, f — потенциальная функция формы. Тут мы предпо-

лагаем, что в рассмотренной вероятностной модели потенциалы, соответствующие 3-кликам, могут быть дополнительно факторизованы.

В рамках предложенной модели можно сформулировать задачу сегментации как задачу нахождения

$$\arg\max_{S,L} P(S,L\,|\,I) = \arg\max_{S,L} P(S,L,I) = \\ = \arg\min_{S,L}\Big[-\log f(S) - \sum_i \big(\log h_i(I_i,L_i) + \\ + \log\varphi_i(L_i,S)\big) - \sum_{(i,j)\in\mathcal{N}} \log\varphi_{ij}(L_i,L_j)\Big].$$

Эту оптимизационную задачу можно решать, например, покоординатным спуском по S и L. Соответствующие формулы будут иметь вид

$$S^{new} = \arg\min_S \Big[-\log f(S) - \sum_i \log\varphi_i(L_i^{old},S)\Big], \quad (1)$$

$$L^{new} = \arg\min_L \Big[-\sum_i \big(\log h_i(I_i,L_i) + \\ + \log\varphi_i(L_i,S^{new})\big) - \sum_{(i,j)\in\mathcal{N}} \log\varphi_{ij}(L_i,L_j)\Big]. \quad (2)$$

Следует отметить, что минимизация по L — обычная задача минимизации MRF-энергии с бинарными метками, которая может быть эффективно решена (например, с помощью разрезов графов [1]) для субмодулярных слагаемых $\log\varphi_{ij}(l_1,l_2)$. Метод минимизации по S следует выбирать в зависимости от конкретного вида модели формы.

Модель формы на основе циркулярного графа

В качестве модели формы в настоящей работе рассмотрен граф, ребра которого соответствуют ветвям скелетного графа объекта, а с каждой вершиной дополнительно связано значение толщины. Такое представление объекта можно рассматривать как упрощённый аналог циркулярного графа. Пример такого представления для фигуры жирафа приведен на рисунке 3.

Априорное распределение параметров формы. Потенциальную функцию формы $f(S)$ в рамках данной модели можно ввести следующим образом:

$$-\log f(S) = \sum_i U_i(e_i) + \sum_{(i,j)\in\mathcal{N}_S} B_{ij}(e_i,e_j), \quad (3)$$

где e_i — i-е ребро графа формы, $\mathcal{N}_S$ — множество неупорядоченных пар индексов соседних ребер. При помощи унарных членов U_i можно задавать глобальные ограничения на положение каждого из ребер, а также на радиусы вершин. Бинарные члены B_{ij} позволяют задать ограничения на взаимное положение ребер. С их помощью можно регулировать гибкость различных частей модели.

Связь формы с сегментацией. Чтобы завершить описание введённой модели формы, необходимо указать вид потенциалов $\varphi_i(l,S)$. Естественно предположить, что пиксели, расположенные недалеко от какого-либо из ребер графа формы, скорее всего будут принадлежать объекту, в то время как пиксели, расположенные в отдалении от ребер графа, будут относиться к фону. Это наблюдение позволяет ввести φ_i следующим образом:

$$\varphi_i(l,S) = l\max_{e\in E_S} Z(e,i) + \\ + (1-l)(1-\max_{e\in E_S} Z(e,i)). \quad (4)$$

Тут E_S — множество ребер графа формы, а $Z(e,i)$ — некоторая функция, монотонно убывающая с увеличением расстояния от ребра e до пикселя i. В данной работе использовалась функция

$$Z(e,i) = \\ = \exp\left[-w\max\left\{0,\left(\frac{D(e,i)-\alpha W(e,i)}{(1-\alpha)W(e,i)}\right)^p\right\}\right], \quad (5)$$

где $D(e,i)$ — расстояние от ребра e до пикселя i, а $W(e,i)$ — ширина ребра вдоль кратчайшего отрезка, соединяющего e с i.

Имитация отжига для обновления параметров формы. Для минимизации (1) в рамках предложенной модели формы в данной работе использовался метод имитации отжига. На каждой итерации отжига к позициям и радиусам вершин графа формы добавлялся гауссовский шум с дисперсией, пропорциональной температуре, равной $\frac{1}{\log k}$ на итерации k. В проведённых экспериментах процесс сходился за 2000–4000 итераций.

Эксперименты

Предложенный в данной работе метод сегментации с моделью формы был протестирован на двух наборах изображений. Первый набор был образован фотографиями жирафов в профиль из [8]. Второй набор изображений, составленный из ряда источников, содержал различные изображения заглавной буквы «E». В изображениях из обоих наборов отсутствовали четкие границы между объектом и фоном, а их цвета зачастую были очень близки.

Для каждого изображения был дополнительно задан приблизительный ограничивающий прямоугольник объекта, который использовался для построения моделей цвета объекта и фона (как в работе [7]), а также при выборе начальной формы для имитации отжига.

Унарные и бинарные члены. Модели цвета объекта и фона представляли из себя смеси Гауссиан с тремя компонентами. Для бинарных членов была использована 4-связная модель близости. Соответствующие члены имели вид

$$\varphi_{ij}(l_1, l_2) = \\ = \exp\left[-\lambda|l_1 - l_2|\left(e^{-c\frac{(B_i - B_j)^2}{\sigma^2}} + d\right)\right],$$

где B_k — интенсивность цвета пикселя k, $c = 1.2$, $d = 0.1$, $\lambda = 10$. Параметр σ был установлен равным среднему модулю разности между интенсивностями соседних пикселей.

Для членов $\varphi_i(l, S)$ использовалось выражение (4) с функцией $Z(e, i)$ вида (5). Параметры имели значения $w = \ln 2$, $\alpha = 0.7$, $p = 2$.

Покоординатный спуск. На первой итерации покоординатного спуска для инициализации имитации отжига использовалась мода распределения $P(S)$, вписанная в известный ограничивающий прямоугольник объекта. На последующих итерациях в качестве начального приближения использовалось решение, найденное на предыдущей итерации.

Для минимизации выражения (2) использовался метод разрезов графов. Минимизируемое на этом шаге выражение имело вид

$$F(L) = -\sum_i \log h_i(I_i, L_i) - \\ -w_s \sum_i \log \varphi_i(L_i, S) - \sum_{i,j \in \mathcal{N}} \log \varphi_{ij}(L_i, L_j). \qquad (6)$$

Мы обнаружили, что плавное увеличение параметра w_s с 0 до 1 в течение первых нескольких итераций может сделать сегментацию более устойчивой. Это может рассматриваться как эвристика для предотвращения попадания в локальный минимум.

Для инициализации покоординатного спуска была использована разметка L, полученная путём минимизации выражения (6) с $w_s = 0$, т. е. путём сегментации без модели формы. Мы считали, что процесс спуска сошёлся, если после очередной итерации доля пикселей с изменившейся меткой не превышала 0.0002. Процесс обычно сходился за 12–15 итераций, что занимало 2–3 минуты на современном компьютере для изображения размерами 320×240 пикселей.

Использованные модели формы. Использованная в экспериментах модель формы для фигуры жирафа приведена на рисунке 3, для заглавной «Е» — на рисунке 2. В обеих моделях члены U_i и B_{ij} из (3) имели следующий вид:

$$U_i(e) = V_{v(e,1)}\big(r(e,1)\big) + V_{v(e,2)}\big(r(e,2)\big) + \\ + I[i = 1]\frac{1}{\sigma^G}(\|e\| - \rho^G l)^2,$$

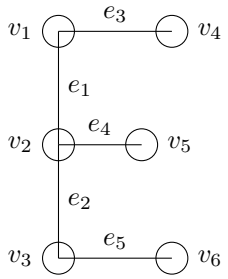


Рис. 2. Модель формы заглавной буквы «Е».

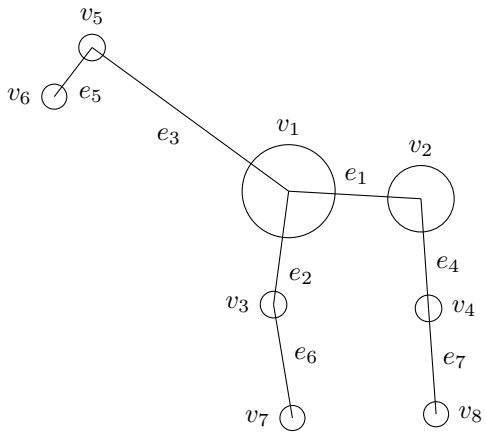


Рис. 3. Модель формы жирафа.

$$V_i(r) = \frac{1}{\sigma_i^r \deg_i}(r - \rho_i l)^2,$$

$$B_{ij}(e_1, e_2) = \frac{1}{\sigma_{ij}^{\alpha}}(\angle(e_1, e_2) - \alpha_{ij})^2 + \\ + \frac{1}{\sigma_{ij}^{l}}(\|e_1\| - \rho_{ij}\|e_2\|)^2.$$

Тут через $v(e, k)$ обозначен индекс k-й вершины ребра e, через $r(e, k)$ — её радиус, $k \in \{1, 2\}$, $\deg_i$ — степень i-й вершины. Параметр ρ_i связывает радиус i-й вершины с l, длиной наибольшей из сторон ограничивающего прямоугольника, заданного пользователем; параметр σ_i^r регулирует жесткость этой связи. Параметр α_{ij} задаёт средний угол между i-м и j-м ребрами, ρ_{ij} связывает их длину между собой, σ_{ij}^{α} и σ_{ij}^{l} служат для задания жесткости соответствующих ограничений. Параметры ρ^G и σ^G отвечают за глобальный масштаб. Значения всех параметров для обеих моделей были откалиброваны вручную по нескольким изображениям.

Результаты применения предложенного в данной работе метода к нескольким изображениям приведены на рисунке 256. Для сравнения рядом приведена сегментация без модели формы, которая использовалась как инициализация для покоординатного спуска. Видно, что использование дополнительного ограничения на форму области позволяет значительно улучшить качество сегментации.

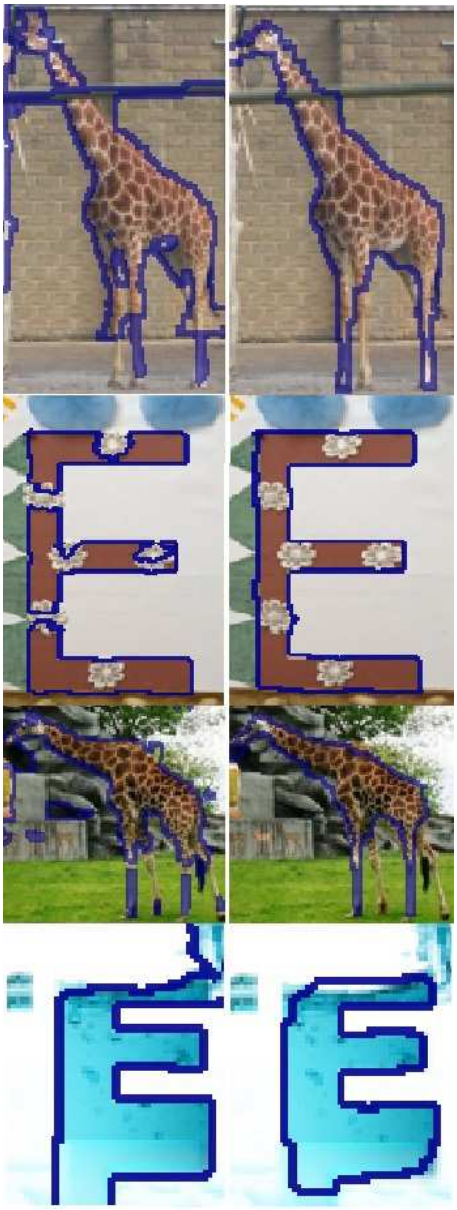


Рис. 4. Некоторые результаты. Слева: сегментация без модели формы. Справа: с моделью.

Заключение

В данной работы был предложен итеративный алгоритм сегментации изображений с использованием модели формы. Каждая итерация алгоритма состоит из двух шагов: сперва выполняется выбор наиболее вероятной формы при помощи имитации отжига, затем производится сегментация изображения с помощью разрезов графов с учётом выбранной формы.

В работе была также предложена модель формы на основе упрощённого циркулярного графа объекта, которая позволяет моделировать широкий класс вариаций формы, а также задавать глобальные ограничения на размер и положение объекта. Эксперименты на двух коллекциях изображений показывают, что использование предложенной модели формы позволяет существенно улучшить качество сегментации.

Одно из возможных направлений дальнейших исследований — отказ от имитации отжига для выбора формы в пользу более эффективных методов оптимизации. Варианты включают в себя динамическое программирование [3], а также метод ветвей и границ [6].

Другое возможное направление исследований — автоматическое построение модели формы по размеченным вручную изображениям. При этом отдельный интерес представляет автоматический выбор структуры графа модели по набору бинарным масок объектов.

Литература

[1] *Boykov Y. Y., Jolly M. P.* Interactive graph cuts for optimal boundary & region segmentation of objects in ND images. In *2001 IEEE 8th International Conference on Computer Vision*, volume 1, pp. 105–112. IEEE, 2001.

[2] *Bray M., Kohli P., Torr P. H. S.* Posecut: Simultaneous segmentation and 3d pose estimation of humans using dynamic graph-cuts. *Proceedings of the 8th European Conference on Computer Vision*, 01:642–655, 2006.

[3] *Felzenszwalb P. F., Huttenlocher D. P.* Pictorial Structures for Object Recognition. *International Journal of Computer Vision*, 61(1):55–79, January 2005.

[4] *Freedman D., Zhang T.* Interactive Graph Cut Based Segmentation with Shape Priors. *2005 IEEE Conference on Computer Vision and Pattern Recognition*, pages 755–762, 2005.

[5] *Kumar M. P., Torr P. H. S., Zisserman A.* Obj Cut. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1. IEEE Computer Society, 2005.

[6] *Lempitsky V., Blake A., Rother A.* Image segmentation by branch-and-mincut. *Proceedings of the 10th European Conference on Computer Vision*, pages 15–29, 2008.

[7] *Lempitsky V., Kohli P., Rother C., Sharp T.* Image segmentation with a bounding box prior. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 277–284. IEEE, September 2009.

[8] *Quack T., Ferrari V., Leibe B., van Gool L.* Efficient Mining of Frequent and Distinctive Feature Configurations. *2007 IEEE 11th International Conference on Computer Vision*, October 2007.

[9] *Veksler O.* Star Shape Prior for Graph-Cut Image Segmentation. In *Proceedings of the 10th European Conference on Computer Vision*, 2008.

[10] *Vu N., Manjunath B. S.* Shape prior segmentation of multiple objects with graph cuts. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, June 2008.

Случайная морфология: алгоритмы обучения и классификации

Зубюк А. В.
zubuk@cmpd2.phys.msu.ru
Москва, физический факультет МГУ им. М. В. Ломоносова

Рассмотрена задача классификации изображений, имеющих случайную форму, как задача проверки статистических гипотез. Предложен метод эмпирического построения решающих правил в задачах проверки гипотез на основе обучающих последовательностей. Исследованы условия применимости предложенного метода.

Математические методы морфологического анализа изображений, рассмотренные в [2], разработаны для решения задач анализа сцен по их изображениям. В качестве примеров можно привести задачи поиска известного объекта на неизвестном фоне, выделения неизвестного объекта на известном фоне, совмещения изображений, полученных в разных спектральных диапазонах. Трудности при решении подобных задач связаны с тем, что всякое изображение сцены содержит информацию не только о регистрируемой сцене, но также и об условиях регистрации, при которых оно получено. Такими условиями являются, например, характер освещения объектов сцены, его спектральный состав и т. п. В основе математических методов морфологического анализа изображений лежит понятие *формы изображения* как максимального инварианта относительно изменения условий регистрации. Форма изображения сцены несёт наиболее полную информацию о форме сцены, доступную по её изображению.

В ряде задач анализа сцен формы предъявляемых изображений можно считать случайными. Так, например, изображение легкового автомобиля не может быть охарактеризовано одной формой. Каждый легковой автомобиль, а значит и его изображение, имеет свою форму, отличающуюся, вообще говоря, от форм других легковых автомобилей. Распределение этих форм несёт информацию о форме легкового автомобиля «в целом» и отличается, например, от распределения форм грузовых автомобилей.

В качестве другого примера рассмотрим формы изображений рукописных букв. На рис. 1 приведены изображения рукописной буквы «А», отличающиеся друг от друга лишь наклоном, размером и положением на координатной плоскости. Будем считать, что наклон, размер и положение являются несущественными параметрами в задаче классификации рукописных букв, т. е. теми параметрами, которые выше мы назвали условиями регистрации. Таким образом, изображения на рис. 1 отличаются лишь условиями их регистрации и, следовательно, имеют одинаковую форму. На рис. 2 приведены изображения рукописной буквы «А», отличающиеся не только наклоном, размером и положением на координатной плоскости, но также и начертанием. Эти изображения имеют разную форму, однако их отличия по форме связаны со случайными искажениями, объясняющимися, например, дрожанием руки пишущего. В такой ситуации можно считать, что изображения рукописной буквы «А» имеют случайную форму.

Напомним формальное определение *случайной формы изображения*, данное в [1].

Случайная форма изображения

Пусть $(X, \mathfrak{A}_X, \mu)$ — измеримое пространство, в котором множество X представляет собой подмножество координатной плоскости, которое в дальнейшем будем называть полем зрения. Математической моделью изображения будем считать любую интегрируемую с квадратом функцию, определённую на X и принимающую значения на числовой прямой $(-\infty, \infty)$. Пространство всех таких функций с естественными линейными операциями и скалярным произведением

$$(g_1, g_2) = \int\limits_X g_1(x) g_2(x) d\mu(x),$$

обозначим $\mathcal{L}^2_\mu(X)$.

Формой изображения в морфологическом анализе принято считать любое подмножество пространства изображений $\mathcal{L}^2_\mu(X)$, содержащее изображения одной и той же сцены, отличающиеся лишь условиями регистрации, при которых они получены (см. [2]).

Определение 1. *Случайной формой изображения назовём вероятностное пространство* $(\Omega, \mathfrak{A}_\Omega, \mathsf{P})$, *где* Ω — *это некоторое множество форм в пространстве* $\mathcal{L}^2_\mu(X)$, $\mathfrak{A}_\Omega$ — σ*-алгебра подмножеств множества* Ω, *а* P — *вероятность, определённая на* $\mathfrak{A}_\Omega$.

Таким образом, случайная форма — это, фактически, случайное множество в пространстве изображений $\mathcal{L}^2_\mu(X)$. Каждая форма $V \in \Omega$ является реализацией случайной формы $(\Omega, \mathfrak{A}_\Omega, \mathsf{P})$ и представляет собой множество изображений, отличающихся лишь условиями их регистрации, и несёт информацию о тех свойствах изображённой сцены, которые инвариантны относительно изменения этих условий. В свою очередь вероятность P моделирует случайные изменения таких инвариантных свойств.

Изображения, имеющие случайную форму $(\Omega, \mathfrak{A}_\Omega, \mathsf{P})$, формируются по следующей схеме.

1. В результате стохастического эксперимента, моделью которого является вероятностное пространство $(\Omega, \mathfrak{A}_\Omega, \mathsf{P})$, определяется форма $V \in \Omega$, $V \subset \mathcal{L}^2_\mu(X)$.
2. Из формы V произвольным образом выбирается предъявляемое изображение g. Выбор конкретного изображения из V отражает условия регистрации, при которых получено предъявляемое изображения.

Схематично это можно изобразить следующим образом:

$$\Omega \xrightarrow{(\Omega,\mathfrak{A}_\Omega,\mathsf{P})} V \xrightarrow{\text{произв.}} g. \qquad (1)$$

Преобразования, моделирующие изменения условий регистрации

Изменение условий регистрации при получении изображений некоторой сцены приводит к изменению изображений этой сцены. В ряде практически важных случаев можно выделить класс $\mathbb{G} = \{\gamma\colon \mathcal{L}^2_\mu(X) \to \mathcal{L}^2_\mu(X)\}$ преобразований пространства изображений $\mathcal{L}^2_\mu(X)$, содержащий тождественное преобразование, и такой, что

$$\gamma_1 \circ \gamma_2 \in \mathbb{G} \quad \forall \gamma_1,\, \gamma_2 \in \mathbb{G},$$

где знаком «$\circ$» обозначена композиция преобразований γ_1 и γ_2, который моделирует всевозможные изменения условий регистрации изображений. То есть изображение[1] $\gamma \circ g \in \mathcal{L}^2_\mu(X)$ является изображением той же сцены, что и изображение g, для любого $g \in \mathcal{L}^2_\mu(X)$. Форму изображения $f \in \mathcal{L}^2_\mu(X)$ определим как множество изображений

$$V_f = \{\gamma \circ f \mid \gamma \in \mathbb{G}\} \subset \mathcal{L}^2_\mu(X).$$

В ряде случаев класс $\mathbb{G}$ оказывается группой с естественной операцией композиции преобразований. Например, группой является класс преобразований изображений рукописных символов, изменяющих их наклон, размер и положение на координатной плоскости X (см. рис. 1). Такие преобразования имеют вид

$$\gamma \circ g(x) = g\left(Ax + b\right),$$
$$x \in X, \quad A = \begin{pmatrix} \alpha & \beta \\ 0 & \alpha \end{pmatrix},$$

где x и b — столбцы размера 2×1, элементы которых являются координатами на плоскости X.

Пусть $j(\cdot)\colon \mathcal{L}^2_\mu(X) \to \mathcal{J}$ — максимальный инвариант группы преобразований $\mathbb{G}$. Для случайной формы $(\Omega, \mathfrak{A}_\Omega, \mathsf{P})$ отображение $j(\cdot)$ индуцирует вероятностное пространство $(\widetilde{\Omega}, \widetilde{\mathfrak{A}}, \widetilde{\mathsf{P}})$, где

$$\widetilde{\Omega} = \{j(V) \mid V \in \Omega\},$$
$$\widetilde{\mathfrak{A}} = \left\{\{j(V) \mid V \in A\} \;\middle|\; A \in \mathfrak{A}_\Omega\right\}, \qquad (2)$$
$$\widetilde{\mathsf{P}}\left(\{j(V) \mid V \in A\}\right) = \mathsf{P}(A),\ A \in \mathfrak{A}_\Omega.$$

Заметим, что в (2) $j(V)$ — это значение максимального инварианта $j(\cdot)$, одинаковое для всех изображений из формы $V \in \Omega$.

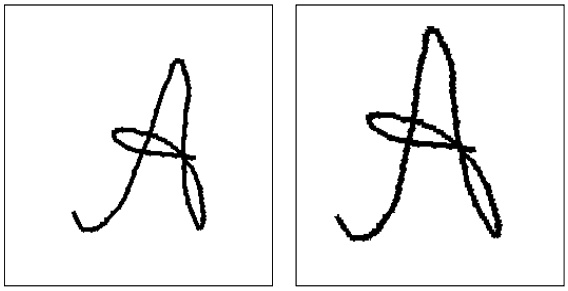


Рис. 1. Изображения рукописной буквы «A», имеющие одинаковую форму.

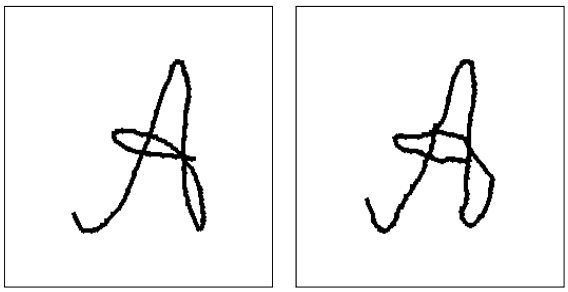


Рис. 2. Изображения рукописной буквы «A», имеющие разную форму.

Задача классификации изображений, имеющих случайную форму

Пусть заданы N случайных форм $(\Omega, \mathfrak{A}_\Omega, \mathsf{P}_i)$, $i = 1, \ldots, N$, и предъявляемое изображение $g \in \mathcal{L}^2_\mu(X)$ имеет одну из этих форм. То есть изображение g получено по схеме (1), и форма V, из которой выбрано изображение g, получена в результате стохастического эксперимента $(\Omega, \mathfrak{A}_\Omega, \mathsf{P})$, где вероятность P — это одна из вероятностей P_i, $i = 1, \ldots, N$. Требуется определить, какую из случайных форм $(\Omega, \mathfrak{A}_\Omega, \mathsf{P}_i)$, $i = 1, \ldots, N$, имеет предъявленное изображение g.

Будем считать, что преобразования изображений, связанные с изменением условий регистрации, образуют группу $\mathbb{G}$, и $j(\cdot)\colon \mathcal{L}^2_\mu(X) \to \mathcal{J}$ — максимальный инвариант этой группы. Тогда описанная задача классификации изображений может быть поставлена как многоальтернативная минимаксная задача проверки статистических гипотез с альтернативами $(\widetilde{\Omega}, \widetilde{\mathfrak{A}}, \widetilde{\mathsf{P}}_i)$, $i = 1, \ldots, N$ (см. предыдущий

[1] Здесь и далее запись вида $\gamma \circ g$ означает результат применения преобразования γ к изображению $g \in \mathcal{L}^2_\mu(X)$.

пункт):

$$\max_{i=1,\dots,N}\alpha_i(\pi)\sim\min_{\pi\in\text{Ш}},$$
$$\alpha_i(\pi)=1-\int\limits_{\widetilde{\Omega}}\pi_i(y)d\widetilde{\mathsf{P}}_i(y),\ i=1,\dots,N, \qquad (3)$$

где Ш — множество всех рандомизированных критериев в N-альтернативной задаче проверки гипотез, т. е. всех векторнозначных функций $\pi(\cdot)=(\pi_1(\cdot),\dots,\pi_N(\cdot))$, определённых на $\widetilde{\Omega}$ и таких, что

$$\pi_i(y)\geqslant 0,\quad \sum_{i=1}^{N}\pi_i(y)=1,\quad y\in\widetilde{\Omega}.$$

Пусть $\pi_*\in\text{Ш}$ — решение задачи (3). Классификация предъявленного изображения g осуществляется путём проведения стохастического эксперимента, моделью которого является вероятностное пространство $(\mathcal{D},\mathfrak{A}_{\mathcal{D}},\Phi)$, где $\mathcal{D}=\{1,\dots,N\}$, $\mathfrak{A}_{\mathcal{D}}$ — σ-алгебра всех подмножеств множества $\mathcal{D}$, а Φ — вероятность, определённая на $\mathfrak{A}_{\mathcal{D}}$ следующим образом:

$$\Phi(D)=\sum_{i\in D}\pi_i(j(g)),\quad D\subset\mathcal{D}.$$

Решение принимается в пользу случайной формы, номер которой является результатом описанного стохастического эксперимента.

Обучение в задаче классификации изображений со случайной формой

В прикладных задачах, требующих классификации изображений, вероятности P_i в случайных формах $(\Omega,\mathfrak{A}_\Omega,\mathsf{P}_i)$, $i=1,\dots,N$, (см. предыдущий пункт) чаще всего не заданы. При этом знание об этих вероятностях может быть получено из обучающей последовательности изображений, набранной в результате серии независимых стохастических экспериментов, моделью которой является вероятностное пространство $(\Omega,\mathfrak{A}_\Omega,\mathsf{P}_1)^L\times\ldots\times(\Omega,\mathfrak{A}_\Omega,\mathsf{P}_N)^L$, где L — натуральное число, которое будем называть длиной обучающей последовательности. Эта последовательность содержит по L изображений-представителей для каждой из N случайных форм. Обозначим их

$$\begin{array}{lll} f_1^{(1)} & \dots & f_L^{(1)} \quad \text{— представители } (\Omega,\mathfrak{A}_\Omega,\mathsf{P}_1),\\ \dots & & \\ f_1^{(N)} & \dots & f_L^{(N)} \quad \text{— представители } (\Omega,\mathfrak{A}_\Omega,\mathsf{P}_N),\end{array}$$
$$f_l^{(i)}\in\mathcal{L}_\mu^2(X),\ i=1,\dots,N,\ l=1,\dots,L.$$

Пусть класс преобразований $\mathbb{G}=\left\{\gamma\colon\mathcal{L}_\mu^2(X)\to\mathcal{L}_\mu^2(X)\right\}$, моделирующих изменения условий регистрации, является группой, $j(\cdot)\colon\mathcal{L}_\mu^2(X)\to\mathcal{J}$ — максимальный инвариант этой группы, и $(\widetilde{\Omega},\widetilde{\mathfrak{A}},\widetilde{\mathsf{P}}_i)$, $i=1,\dots,N$, — вероятностные пространства, индуцированные отображением $j(\cdot)$ (см. (2)).

Рассмотрим $\widetilde{\mathfrak{A}}$-измеримое разбиение $\mathbb{CL}_L$ множества $\widetilde{\Omega}$ на K_L подмножеств:

$$\mathbb{CL}_L=\left\{\mathcal{C}l_1^{(L)},\ \dots,\ \mathcal{C}l_{K_L}^{(L)}\right\}\subset\widetilde{\mathfrak{A}},$$
$$\bigcup_{k=1}^{K_L}\mathcal{C}l_k^{(L)}=\widetilde{\Omega},\quad \mathcal{C}l_{k_1}^{(L)}\cap\mathcal{C}l_{k_2}^{(L)}=\varnothing\text{ при }k_1\neq k_2.$$

Обозначим $\widetilde{\mathfrak{A}}_L$ σ-подалгебру алгебры $\widetilde{\mathfrak{A}}$, порождённую разбиением $\mathbb{CL}_L$, и определим на ней частоты $\nu_i^{(L)}(\cdot)\colon\widetilde{\mathfrak{A}}_L\to[0,1]$, $i=1,\dots,N$, следующим образом:

$$\nu_i^{(L)}(A)=\sum_{k\colon\mathcal{C}l_k^{(L)}\subset A}\nu_i^{(L)}\left(\mathcal{C}l_k^{(L)}\right),\quad A\in\widetilde{\mathfrak{A}},$$
$$\nu_i^{(L)}\left(\mathcal{C}l_k^{(L)}\right)=\frac{1}{L}\sum_{l=1}^{L}\chi_{\mathcal{C}l_k^{(L)}}\left(j\left(f_l^{(i)}\right)\right),$$
$$k=1,\ \dots,\ K_L,\ i=1,\dots,N,$$

где $\chi_{\mathcal{C}l_k^{(L)}}(\cdot)$ — индикаторные функции множеств $\mathcal{C}l_k^{(L)}$, $k=1,\dots,K_L$.

Пусть $\mathcal{C}l^{(L)}:\widetilde{\Omega}\to\widetilde{\mathfrak{A}}_L$ — функция, которая каждому $y\in\widetilde{\Omega}$ ставит в соответствие такое подмножество $\mathcal{C}l_k^{(L)}\in\mathbb{CL}_L$, что $y\in\mathcal{C}l_k^{(L)}$. Обозначим $\widetilde{\pi}^{q,L}$ рандомизированный критерий из множества критериев Ш, удовлетворяющий условиям

$$\begin{array}{ll}\widetilde{\pi}_i^{q,L}(y_1)=\widetilde{\pi}_i^{q,L}(y_2), & \text{если } y_1\in\mathcal{C}l^{(L)}(y_2),\\ \widetilde{\pi}_i^{q,L}(y)=0 & \text{если } i\notin\widetilde{I}_{q,L}(y),\end{array} \qquad (4)$$
$$i=1,\dots,N,\quad y,\ y_1,\ y_2\in\widetilde{\Omega},$$

где

$$\widetilde{I}_{q,L}(y)=\left\{i:q_i\nu_i^{(L)}\left(\mathcal{C}l^{(L)}(y)\right)=\max_{j=1,\dots,N}q_j\nu_j^{(L)}\left(\mathcal{C}l^{(L)}(y)\right)\right\},$$

и $q=(q_1,\dots,q_N)$ — N-мерный вектор неотрицательных чисел, таких, что $\sum_{i=1}^{N}q_i=1$ (множество всех таких векторов обозначим $\mathbb{Q}$).

Для любого $q\in\mathbb{Q}$ критерий $\widetilde{\pi}^{q,L}$ может быть построен на основе обучающей последовательности изображений. При выполнении определённых условий может быть организован процесс обучения, в результате которого по обучающей выборке изображений будет построена последовательность критериев, удовлетворяющих (4), в определённом смысле приближающая критерий π_*, являющийся решением минимаксной задачи (3). Рассмотрим эти условия.

Потребуем, чтобы на множестве $\widetilde{\Omega}$ были определены метрика ρ и мера ν, вероятности $\widetilde{\mathsf{P}}_i, i = 1, \ldots, N$, были абсолютно непрерывны относительно меры ν, и их плотности $\widetilde{\mathrm{pr}}_i(\cdot)$ удовлетворяли условиям Липшица

$$\begin{aligned} &|\widetilde{\mathrm{pr}}_i(y_1) - \widetilde{\mathrm{pr}}_i(y_2)| \leqslant C_{\mathrm{pr}}\rho(y_1, y_2), \\ &y_1, y_2 \in \widetilde{\Omega}, \quad i = 1, \ldots, N, \end{aligned} \tag{5}$$

где C_{pr} — некоторая константа.

Обозначим

$$\mathcal{S}_\delta(q) = \Big\{ y \in \widetilde{\Omega} \colon q_I \widetilde{\mathrm{pr}}_I(y) - \max_{\substack{i=1,\ldots,N \\ i \neq I}} q_i \widetilde{\mathrm{pr}}_i(y) \leqslant \delta \Big\}, \quad q \in \mathbb{Q},$$

где для каждого $y \in \widetilde{\Omega}$ номер I определяется из условия $q_I \widetilde{\mathrm{pr}}_I(y) = \max\limits_{i=1,\ldots,N} q_i \widetilde{\mathrm{pr}}_i(y)$. Потребуем, чтобы $\mathcal{S}_\delta(q) \in \widetilde{\mathfrak{A}}$ для любых $\delta \geqslant 0$ и $q \in \mathbb{Q}$, и

$$\begin{aligned} &\widetilde{\mathsf{P}}_i\left(\mathcal{S}_\delta(q)\right) \leqslant C_{\mathcal{S}}\delta \text{ при } \delta < \overline{\delta}, \\ &q \in \mathbb{Q}, \quad i = 1, \ldots, N, \end{aligned} \tag{6}$$

где $C_{\mathcal{S}} \geqslant 0$ и $\overline{\delta} > 0$ — некоторые константы.

Определим последовательность $\{\mathbb{CL}_1, \mathbb{CL}_2, \ldots\}$ $\widetilde{\mathfrak{A}}$-измеримых разбиений множества $\widetilde{\Omega}$ так, чтобы

$$\begin{aligned} &K_L < K_{L+1}, \quad L = 1, 2, \ldots, \\ &\forall r = 1, \ldots, K_{L+1} \quad \exists s = 1, \ldots, K_L : \\ &\quad \mathcal{C}l_r^{(L+1)} \subset \mathcal{C}l_s^{(L)}. \end{aligned} \tag{7}$$

Для всякого критерия $\widetilde{\pi}^{q,L}$, удовлетворяющего (4), определим эмпирическую ошибку $\widetilde{\alpha}^{q,L}\left(\widetilde{\pi}^{q,L}\right)$ следующим образом:

$$\widetilde{\alpha}^{q,L}\left(\widetilde{\pi}^{q,L}\right) = \sum_{i=1}^{N} q_i \widetilde{\alpha}_i^{q,L}\left(\widetilde{\pi}^{q,L}\right),$$

$$\widetilde{\alpha}_i^{q,L}\left(\widetilde{\pi}^{q,L}\right) = 1 - \sum_{k=1}^{K_L} \widetilde{\pi}_i^{q,L}\left(y_k^{(L)}\right) \nu_i^{(L)}\left(\mathcal{C}l_k^{(L)}\right),$$

где $y_k^{(L)} \in \mathcal{C}l_k^{(L)} \in \mathbb{CL}_L$, $q \in \mathbb{Q}$.

Теорема 1. *Пусть $(\widetilde{\Omega}, \rho, \widetilde{\mathfrak{A}}, \nu)$ — метрическое измеримое пространство, вероятности $\widetilde{\mathsf{P}}_i$ абсолютно непрерывны относительно меры ν, $\mathcal{S}_\delta(q) \in \widetilde{\mathfrak{A}}$ для любых $\delta \geqslant 0$ и $q \in \mathbb{Q}$, и выполнены условия* (5), (6).

Пусть решение π_* задачи (3) *единственно.*

Пусть последовательность $\{\mathbb{CL}_1, \mathbb{CL}_2, \ldots\}$ $\widetilde{\mathfrak{A}}$-измеримых разбиений $\widetilde{\Omega}$ удовлетворяет (7) *и*

$$\max_{k=1,\ldots,K_L} \sup_{y_1, y_2 \in \mathcal{C}l_k^{(L)}} \rho(y_1, y_2) \xrightarrow[L\to\infty]{} 0,$$

$$K_L \leqslant C_K L^a,$$

где C_K и a — неотрицательные константы.

Пусть последовательность $\{q_1, q_2, \ldots\} \subset \mathbb{Q}$ удовлетворяет условию

$$\max_{i=1,\ldots,N} \widetilde{\alpha}_i^{q_L,L}\left(\widetilde{\pi}^{q_L,L}\right) - \widetilde{\alpha}^{q_L,L}\left(\widetilde{\pi}^{q_L,L}\right) \xrightarrow[L\to\infty]{} 0. \tag{8}$$

Тогда почти наверное[2]

$$\begin{aligned} \widetilde{\pi}^{q_L,L}(y) &\xrightarrow[L\to\infty]{} \pi_*(y) \quad \forall y \in \widetilde{\Omega}, \\ \widetilde{\alpha}^{q_L,L}\left(\widetilde{\pi}^{q_L,L}\right) &\xrightarrow[L\to\infty]{} \max_{i=1,\ldots,N} \alpha_i(\pi_*), \end{aligned} \tag{9}$$

где функции $\alpha_i(\cdot)$ определены в (3).

Теорема 1 определяет процесс обучения принятию решений в задаче классификации изображений со случайной формой, позволяющий на основе обучающей последовательности изображений построить последовательность критериев $\left\{\widetilde{\pi}^{q_1,1}, \widetilde{\pi}^{q_2,2}, \ldots\right\}$, приближающую решение π_* задачи (3) в смысле сходимостей (9). При этом на L-ом шаге вектор $q_L \in \{q_1, q_2, \ldots\}$, удовлетворяющий (8), может быть найден с помощью алгоритмов случайного поиска, описанных в [3].

В докладе также будут приведены результаты применения методов случайной морфологии в прикладных задачах классификации изображений.

Выводы

Разработаны адаптивный алгоритм морфологической классификации изображений со случайной формой и его компьютерная реализация. Работа алгоритма исследована на примере классификации изображений рукописных символов.

В заключение хочется выразить благодарность проф. Пытьеву Ю. П. за постановку задачи и помощь в её решении.

Литература

[1] *Зубюк А. В.* Алгоритмы идентификации изображений в случайной и нечёткой морфологии // Математические методы распознавания образов. 13-я Всероссийская конференция: Сборник докладов, Москва: МАКС Пресс, 2007. — С. 30–32.

[2] *Пытьев Ю. П., Чуличков А. И.* Методы морфологического анализа изображений. — Москва: ФИЗМАТЛИТ, 2010. — 336 с.

[3] *Ермольев Ю. М.* Методы стохастического программирования. — Москва: Наука, 1976. — 340 с.

[2] Утверждение «почти наверное» понимается здесь в терминах вероятностного пространства $(\Omega, \mathfrak{A}_\Omega, \mathsf{P}_1)^\infty \times \ldots \times (\Omega, \mathfrak{A}_\Omega, \mathsf{P}_N)^\infty$.

Методы морфологического анализа изображений в задаче интерпретации данных ядернофизического эксперимента.*

Фаломкина О. В., Пытьев Ю. П., Пятков Ю. В., Каманин Д. В., Хербст Б. М., Трзаска В. Х.

yuri.pytyev@gmail.com, olesya.falomkina@gmail.com

Москва, МГУ им. М. В. Ломоносова, физический факультет

Теоретические описания ядерных реакций, таких как деление и квазиделение, позволяют представить эволюцию ядерной системы как траекторию в многомерном деформационном пространстве. Мы предлагаем стратегию выявления изображений таких траекторий в пространстве экспериментально наблюдаемых переменных при заданном уровне надежности. Предлагаемый подход основан на математических методах морфологического анализа изображений и позволяет детально анализировать последовательные стадии ядерных реакций длящихся порядка 10^{-20} секунды.

В работе предложено решение задачи выделения «тонких структур» из известных массово-энергетических распределений продуктов ядерных реакций посредством непосредственной обработки двумерных данных [1] на основе методов морфологического анализа изображений [5].

Типичное $M - E$ (масса-энергия) распределение осколков деления, например, в реакции $^{233}U(n_{th}, f)$, на первый взгляд, выглядит как гладкий холм. Более детальное рассмотрение показывает, что любое сечение $E = const$ этого распределения (см. рис. 1) не является абсолютно гладким, но напротив, демонстрирует локальные нерегулярности (пики), показанные стрелками.

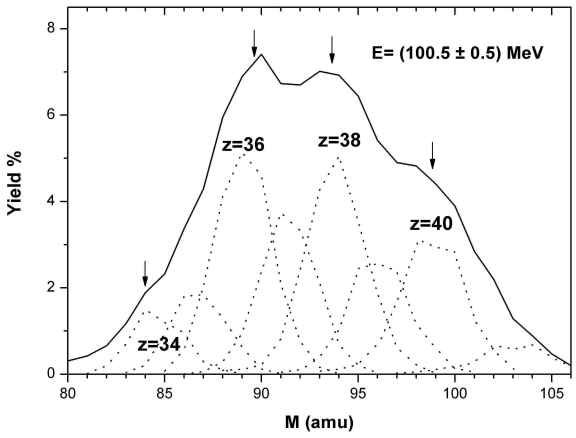


Рис. 1. Сечение $E - M$ распределения для энергии осколка $E = (100.5 \pm 0.5)$ MeV [3]. Парциальные выходы для фиксированных зарядов показаны пунктирными линиями.

Известно, что пики являются проявлением протонного чётно-нечётного эффекта, вследствие которого изотопы с чётным ядерным зарядом имеют повышенный выход. Совокупность обсуждаемых пиков образует на плоскости масса-энергия систему почти периодических хребтов параллельных оси . Эта структура будет называться ниже «вертикальные хребты».

Определим термин «тонкая структура» (ТС). По определению, это локальные области (пики) двумерного распределения с повышенным, по отношению к гладкой подложке, выходом. Мы задались вопросом: есть ли какая-либо структура в массово-энергетическом распределении осколков деления (ОД), отличная от вертикальных хребтов, продуцируемых чётно-нечетной модуляцией массовых выходов и обусловленная, следовательно, другими физическими причинами? Для автоматического подавления вертикальных хребтов при поиске тонкой структуры анализировались сечения $M =$ $=$ const. В наших исследованиях использовались алгоритмы идентификации пиков, известные в гамма-спектроскопии, и методы обработки зашумленных изображений [1]. На рис. 2 показаны примеры ТС, выявленных в распределениях полная кинетическая энергия-масса ОД. Более тёмные точки в тоновой диаграмме (рис. 2a)) соответствуют большей амплитуде эффекта. Показан только лёгкий пик массового распределения ОД.

Симметричность показанных на рис 2b структур обусловлена как методикой измерения масс ОД (метод 2-х скоростей [2]), так и использованным фильтром [4].

Превалирующая ТС представляет собой последовательность змееподобных кривых, иногда имеющих точки бифуркации [1, 4].

Какова цель исследования обсуждаемых тонких структур? По современным представлениям, эволюция распадающейся ядерной системы, например, в делении, определяется в основном потенциальной энергией, зависящей от деформации системы или, в трехмерном представлении, поверхностью потенциальной энергии (ППЭ). Отдельные потенциальные долины на ППЭ [6, 7] являются причиной существования выделенных траекторий системы в деформационном пространстве. Как показано в [8], в любой точке спуска системы по до-

Работа выполнена при финансовой поддержке РФФИ, проекты №11-07-00722-а, 11-07-00338-а.

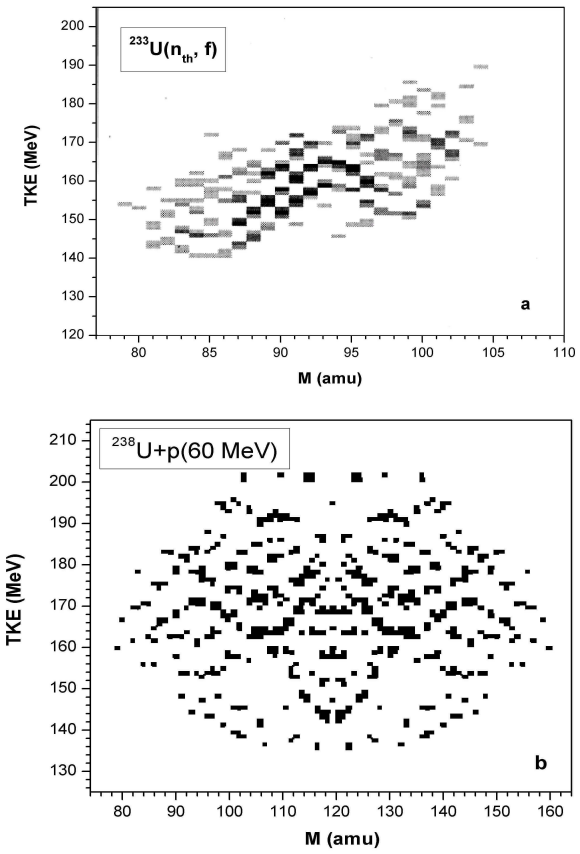


Рис. 2. «Змееподобные» тонкие структуры, найденные в распределениях масса-полная кинетическая энергия ($M-TKE$) ОД из реакций $^{233}U(n_{th},f)$ (a) и $^{238}U+p$ (60 MeV) (b). Детали представлены в тексте.

лине может произойти разрыв, что регистрируется как событие деления в пространстве экспериментально наблюдаемых переменных. Другими словами, дискретные траектории в деформационном пространстве как непрерывная последовательность ядерных состояний в долине деления отображается на непрерывные траектории (гладкие кривые) на плоскости экспериментально наблюдаемых переменных E-M [9]. Таким образом, мы полагаем, что обсуждаемые ТС являются уникальным изображением отдельных путей деления, проходимых системой за времена всего лишь $\sim 10^{-20}$ сек [10].

До настоящего времени слабым местом описанного подхода к анализу данных было отсутствие количественной оценки надежности выделенной тонкой структуры. Эта проблема усугубляется тем, что ищется структура с априори неизвестной формой.

Для решения этой проблемы был разработан подход, основанный на методах морфологического анализа изображений [5].

Пусть $\tilde{f}$ — полученный в эксперименте сигнал ($E-M$ распределение), который может быть представлен в виде

$$\tilde{f} = S + h + \nu, \tag{1}$$

где S — изображение «гладкой подложки», h — сигнал, в котором, возможно, содержатся несколько интересующих нас «тонких структур», ν — аддитивный шум. На первом этапе происходит выделение «гладкой подложки» S из сигнала $\tilde{f}$, например, с помощью сплайн-интерполяции [1, 11]. В результате получается сигнал $f = h + \nu$. На втором этапе из сигнала f извлекается «тонкая структура» с помощью морфологических методов анализа изображений.

Рассмотрим вкратце некоторые понятия методов морфологического анализа изображений. *Изображением* сигнала $f(\cdot)$ называется квадратично интегрируемая на X числовая функция, заданная на подмножестве X евклидовой плоскости $\mathcal{R}^2$. Область X называется *полем зрения*, а значение $f(x)$ функции $f(\cdot)$ в точке $x \in X$ — яркостью точки x поля зрения X. В рассматриваемом случае $X = \{x_1, \ldots, x_n\}$ и, соответственно, изображения $\tilde{f}(\cdot)$ $f(\cdot)$, $S(\cdot)$, $h(\cdot)$ и $\nu(\cdot)$ из (1) определены в одних и тех же точках и являются элементами евклидовой плоскости $\mathcal{R}^n$. О погрешности $\nu \in \mathcal{R}^n$ будем полагать, что это случайное изображение с нулевым математическим ожиданием $\mathbf{E}\nu = 0$ и корреляционным оператором $\sigma^2 I$, $I \in (\mathcal{R}^n \to \mathcal{R}^n)$ — единичный оператор, σ^2 неизвестно.

Изображение «тонкой структуры» обозначим $\omega(\cdot)$ и будем считать его заданным на подвижном, переменного размера подмножестве Ω поля зрения X.

Формой изображения $\omega(\cdot)$ назовем множество изображений

$$V_\omega = \{\omega(\cdot), \omega(x) = c_1\chi_{A_1}(x) + c_2\chi_{A_2}(x), c_1 \geqslant c_2, \\ c_1, c_2 \in \mathcal{R}_1,\ x \in \Omega\}, \tag{2}$$

где

$$\chi_{A_i}(x) = \begin{cases} 1, & x \in A_i, \\ 0, & x \notin A_i, \end{cases} \quad i = 1, 2.$$

V_ω является выпуклым замкнутым конусом в $\mathcal{R}^2$ и в $\mathcal{R}^n$. В этом определении A_1 и A_2 — различные подобласти постоянной яркости в Ω. В соответствии с этим определением, *форма* изображения объекта содержит все изображения этого объекта, отличающиеся яркостями на подобластях постоянной яркости в Ω.

На рис. 3 показаны области $A_1, A_2 \subset \Omega$ постоянной яркости изображения тонкой структуры. На этом рисунке поле зрения разбито на области A_1, A_2. Непосредственно «тонкой структуре» соответствует область A_1, окружающим её точкам — область A_2. Форма (в обычном понимании этого сло-

ва) и размер областей A_1, A_2 заданы (постулируются) априори исследователем. Предлагаемый метод позволяет проверить правильность этого постулата. В нашем случае форма «тонкой структуры» была определена на основе рис. 2. Яркости над областями A_1, A_2 предполагаются постоянными. То, что яркость в точках изображения, принадлежащих «тонким структурам», должна быть выше, чем в окружающих её точках, отражено в условии $c_1 \geqslant c_2$ в выражении (2).

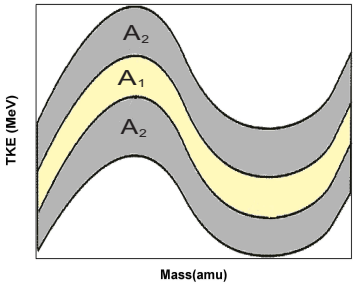


Рис. 3. Пример областей $A_1, A_2 \subset \Omega$ постоянной яркости на модельном изображении «тонкой структуры».

Проекцией (см. ниже) некоторого изображения $g(\cdot)$, определённого на Ω, на форму V_ω называется изображение $(P_{V_\omega} g)(\cdot)$, которое корректно определено равенством

$$(P_{V_\omega} g)(x) = \hat{c}_1 \chi_{A_1}(x) + \hat{c}_2 \chi_{A_2}(x),\ x \in \Omega, \qquad (3)$$

поскольку V_ω является выпуклым замкнутым конусом (см. [5]). В (3) $\hat{c}_1, \hat{c}_2$ — решения следующей задачи на минимум,

$$\int\limits_{\Omega} \left(g(x) - \hat{c}_1 \chi_{A_1}(x) - \hat{c}_2 \chi_{A_2}(x)\right)^2 dx = \\ = \min_{c_1, c_2 \in \mathcal{R}_1 c_1 \geqslant c_2} \int\limits_{\Omega} \left(g(x) - c_1 \chi_{A_1}(x) - c_2 \chi_{A_2}(x)\right)^2 dx. \qquad (4)$$

Далее для краткости знак $(\cdot)$ при обозначении изображений как функций опущен.

Рассмотрим проблему выделения тонкой структуры в рамках сформулированной выше модели (1) регистрации сигнала как задачу проверки статистической гипотезы H о том, что на изображении f существует фрагмент f_ω, который может быть представлен в виде

$$H: \exists f_\omega = g + \nu,\ \exists t \in T,\ g \in t(V_\omega), \nu \in \mathcal{N}(0, \sigma^2 I), \\ \sigma^2 > 0,\ \|\nu\|^2 << \|g\|^2, \qquad (5)$$

где форма g с точностью до сдвига и масштабного преобразования совпадает с (2), $t \in T$ — преобразование сдвига и масштаба, T — класс таких преобразований, $\|\cdot\|$ — символ евклидовой нормы. Альтернатива K заключается в том, что такого фрагмента нет.

Для решения задачи проверки гипотезы используется функционал [5]

$$j(z) = \frac{||(I - P_{V_\omega})z||^2}{||(P_{V_\omega} - P_{V_U})z||^2}. \qquad (6)$$

В выражении (6) z — некоторое изображение, $P_{V_U} z$ — проекция изображения $z(\cdot)$ на форму U ровного поля зрения,

$$U = \{u(\cdot), u(x) = \text{ const} \cdot \chi_\Omega(x),\ x \in \Omega\}. \qquad (7)$$

Функционал (6) обладает следующими свойствами.

1. Пусть нашёлся фрагмент f_ω, удовлетворяющий условию (5), но не представимый в виде
$$f_\omega = g + \nu, \exists t \in T,\ g \in t(U), \nu \in (0, \sigma^2 I),\ \sigma^2 > 0. \qquad (8)$$
Числитель в (6) равен $||(I - P_{V_\omega})\nu||^2$, знаменатель равен $||(P_{V_\omega} - P_{V_U})\nu + (P_{V_\omega} - P_{V_U})g||^2$ и имеет значения порядка $O(\|g\|^2)$. Поэтому значение функционала (6) мало, т.к. $\|\nu\|^2 << \|g\|^2$.
2. Пусть нашёлся фрагмент f_ω, удовлетворяющий условию (8). Числитель в (6) равен $||(I - P_{V_\omega})\nu||^2$ и имеет значения порядка $O(\|\nu\|^2)$, знаменатель равен $||(P_{V_\omega} - P_{V_U})\nu||^2$ и также имеет значения порядка $O(\|\nu\|^2)$. Таким образом, функционал $j(z)$ имеет значения порядка $O(1)$.
3. Пусть не нашёлся фрагмент f_ω, удовлетворяющий условию (5) или (8). Числитель в (6) равен $||(I - P_{V_\omega})\nu + (I - P_{V_\omega})g||^2$ и имеет значения порядка $O(\|g\|^2)$. Знаменатель, равный $||(P_{V_\omega} - P_{V_U})\nu + (P_{V_\omega} - P_{V_U})g||^2$, также имеет значения порядка $O(\|g\|^2)$, т. о. функционал $j(z)$ снова имеет значения порядка $O(1)$.

Таким образом, только в случае 1. значение функционала (6) мало, ибо $\|\nu\|^2 << \|g\|^2$.

Решающее правило имеет следующий вид: гипотеза H принимается, если путём сдвига и масштабного преобразования найдется такой фрагмент f_ω, что $j(f_\omega) \leqslant a$, где a — эмпирически найденная константа (см. ниже), и отвергается, если такого фрагмента нет.

Значение функционала (6) характеризует «близость» между изображением z и изображением формы (2). Следует отметить, что функционал (6) инвариантен относительно преобразований яркости и контраста, т. е. относительно преобразований $z \to \alpha z + \beta$, где α — некоторое число, β — изображение, определённое на Ω.

Для определения значения константы a на реальных данных экспериментально было найдено значение константы, при котором изображение найденной тонкой структуры удовлетворяет исследователя. В нашем случае оказалось, что $a = 40$.

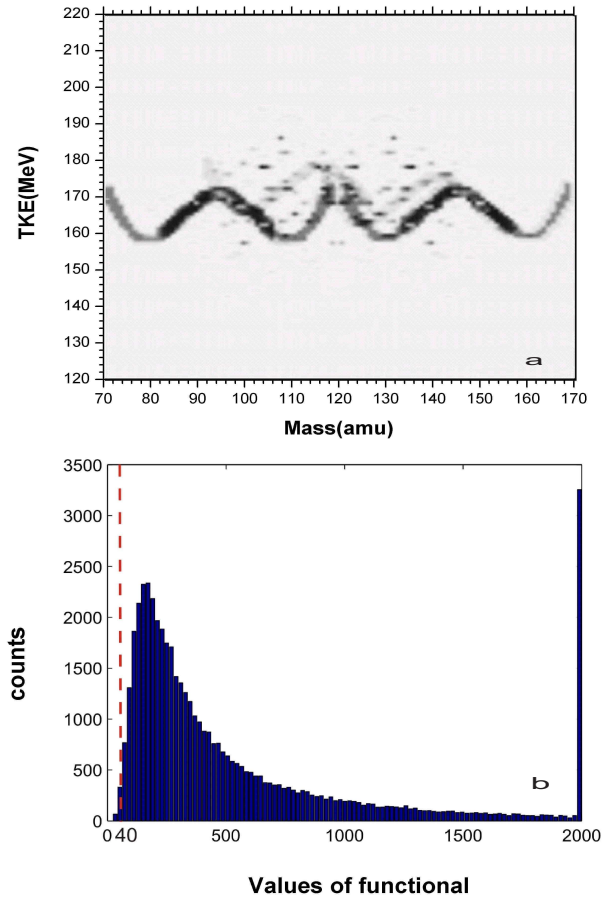


Рис. 4. (a) Тонкая структура, найденная на распределении $TKE - M$, изображенном на рис. 2. (b) Спектр значений функционала (6), полученный на основе модельных данных. Пунктирная линия показывает пороговое значение a, для которого $P(j \leqslant a) = 0.001$.

Затем надежность найденного значения была проверена с помощью модельного эксперимента. Было использовано 10000 модельных изображений гладкой подложки S, зашумленной аддитивным пуассоновским шумом. Параметры шума были подобраны соответствующими реальному эксперименту. С помощью методов, предложенных в [1, 11], шум был «отделен» от гладкой подложки, и в результате было построено эмпирическое распределение значений функционала (6) при заданном уровне шума. На основе этого распределения была найдена вероятность $P(j \leqslant a) = 0.001$, см. рис. 4. Эта вероятность является вероятностью ошибочно принять гипотезу против ближайшей к гипотезе альтернативы «однородное поле зрения». Из свойств 1.-3. функционала (6) следует, что эта вероятность ошибки оценивает сверху вероятность ошибочно принять гипотезу против альтернативы «такого фрагмента нет». Данный критерий — аналог принципу локально равномерно наиболее мощного критерия [12].

В соответствии с проведённым статистическим анализом вероятность того, что «тонкая структура», найденная на реальных данных, порождена шумом, действительно мала.

На рис. 4 приведен результат поиска «тонких структур» на реальных данных с помощью предложенным метода (a) в сравнении с результатами, полученными ранее (b).

Выводы

1. Многодолинная структура поверхности потенциальной энергии ядерной системы, по крайней мере, в таких реакциях, как слияние, деление, квазиделение обуславливает наличие отдельных путей эволюции системы, проходящих вдоль соответствующих долин. Каждый путь проявляется как траектория в пространстве наблюдаемых переменных таких как масс-асимметрия и полная кинетическая энергия, пропорциональная, в первом приближении, предразрывному удлинению системы. Визуализация таких траекторий (выявление тонкой структуры) может дать уникальную физическую информацию, ранее абсолютно недоступную.
2. Для получения количественных оценок надежности выделенных структур мы развили подход, основанный на методах морфологического анализа изображений. В рамках этого подхода оценивается вероятность случайной реализации (из-за наличия шума) структуры или её масштабной копии. Только наличие такой оценки даёт основание для физической трактовки выявленных структур, снимая вопрос, не являются ли они статистическим артефактом.

Литература

[1] Yu.V. Pyatkov et al., NIM **A 488** 381, (2002).

[2] The Nuclear Fission Process, edited by Cyriel Wagemans, CRC Press, 1991.

[3] U.Quade et al., Nucl. Phys. **A 487**, 1, (1988).

[4] W.H. Trzaska et al., Proc. Symposium on Nuclear Clusters, Rauischholzhausen, Germany, 5-9 August 2002, p. 237.

[5] Yu.P. Pytyev. Morphological Image Analysis // Pattern Recognition and Image Analysis, Vol. 3, No.1, pp. 19–28, MAIK "Nauka/Interperiodica Pleiades Publishing, 1993.

[6] V.V. Pashkevich, Nucl. Phys, **A 169**, 275, (1971).

[7] U.Brosa et al., Phys. Rep, **197**, 167, (1990).

[8] J.F.Berger et al., Nucl. Phys. **A 428**, 230, (1984).

[9] Yu.V. Pyatkov et. al., Nucl. Phys. **A 624**, 140, (1997).

[10] D.J. Hinde et al., Nucl. Phys. A452 (1986) 550

[11] O.V.Falomkina et al., Heavy Ion Physics, FLNR JINR Scientific Report 2003 - 2004, Dubna, Russia, 2006, p. 158-159.

[12] C. R. Rao (chief ed.). Handbook of Statistics, Vols 1-18. New York and Amsterdam: North Holland/Elsevier Science Publishers.

Эмпирическое упорядочение яркости пикселей изображения, задающее его форму*

Чуличков А. И., Цыбульская Н. Д.
achulichkov@gmail.com
Москва, Московский Государственный Университет им. М. В. Ломоносова, Физический факультет

По результатам регистрации серии изображений, полученных от фиксированной сцены при неконтролируемых и изменяющихся условиях, построены характеристики сцены, инвариантные к преобразованиям, моделирующим изменения условий регистрации её изображений. Показано, что для случая, когда изменения условий регистрации приводят к монотонным преобразованиям яркости изображений сцены, максимальным инвариантом этих преобразований является функция, упорядочивающая яркость изображения по невозрастанию. Сформулированы условия, при выполнении которых максимальный инвариант может быть безошибочно построен (с вероятностью единица) по конечному числу наблюдаемых изображений.

При экспериментальных исследованиях часто информация об изучаемом объекте содержится в некоторых характерных особенностях поступающих от объекта сигналов. Методы анализа сигналов, инвариантные к указанным искажениям их амплитуды, носят название морфологических [1], изначально они применялись для анализа изображений [2].

Основным используемым в методах морфологического анализа понятием является форма — инвариант преобразований, моделирующих результаты воздействия на регистрируемые сигналы изменяющихся условий регистрации. В терминах формы решается обширный класс задач анализа сцен по их изображениям [1], оценки свойств источника звука по результатам регистрации звукового давления [3] и др.

Однако на практике часто информация о форме сигнала априори недоступна, и её приходится извлекать из результатов регистрации тех же сигналов.

Форма изображения

Под изображением понимается числовая (для полутоновых) или векторнозначная функция (для цветных изображений), заданная на ограниченном подмножестве (поле зрения) X плоскости. В данной работе ограничимся полутоновыми изображениями. Значение изображения в точке x поля зрения X будем называть яркостью изображения. Далее изображения рассматриваются как элементы евклидова пространства $L^2(X)$. В большинстве приложений поле зрения X состоит из конечного числа пикселей; в этом случае множество всех изображений является евклидовым пространством $\mathbb{R}^N$, размерность которого совпадает с числом N пикселей.

Особенности сцены будут присутствовать во всех её изображениях, полученных при различных условиях. Если известен класс преобразований изображения, которым оно подвергается при изменениях условий регистрации, то информация о сцене будет содержаться в инвариантах этих преобразований, носящих название название формы изображения сцены.

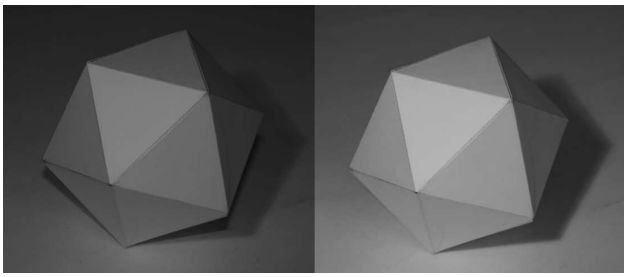


Рис. 1. Изображение многогранника при различных условиях регистрации хорошо аппроксимируется кусочно постоянной функцией яркости; области постоянной яркости не изменяются при вариациях освещения.

В качестве примера рассмотрим форму изображения сцены как разбиение поля зрения на участки постоянной яркости: изменение условий регистрации изображения сцены (освещения, параметров видеокамеры и т. п.) моделируются изменением яркости участков поля зрения, границы же этих участков остаются неизменными, см. рисунок 1.

Для формального построения формы изображения в работе [2] определена операция сравнения изображений по форме: считается, что форма изображения g *не сложнее формы* изображения f, если $g = F * f$ для некоторого $F \in \mathcal{F}$, сигналы f и g называются *сравнимыми по форме*. Преобразование $f \to F * f$ моделирует влияние условий регистрации.

В реальных ситуациях преобразование F неизвестно, однако задан класс $\mathcal{F}$ преобразований, моделирующих влияние всех возможных условий регистрации. Если функция F пробегает весь класс $\mathcal{F}$, то результаты преобразований

$$V_f = \{F * f, F \in \mathcal{F}\}$$

образуют множество всех возможных изображений сцены.

Работа выполнена при финансовой поддержке РФФИ, проект № 11-07-00338-а.

Множество V_f названо формой изображения f. В [1] указаны условия, при которых V_f является выпуклым и замкнутым множеством. Тогда с V_f взаимно однозначно связан оператор проецирования на V_f, обозначаемый P_f и также называемый формой изображения f:

$$\|P_f x - x\|^2 = \inf_{g \in V_f} \|g - x\|^2 \equiv \inf_{g = F*f,\, F \in \mathcal{F}} \|g - x\|^2. \quad (1)$$

Если форма изображения априори известна, то для решения задач узнавания, классификации, оценки параметров сцен по их изображениям и других задач разработаны морфологические методы, основой которых является операция проецирования P_f [1].

Задача описания сцены по серии её изображений, искаженных шумом

Далее будем считать, что изображение $f \in \mathbb{R}^N$ задано своими координатами $(f_{(1)}, \ldots, f_{(N)})$ (яркостями пикселей), а функции из класса $\mathcal{F}$ определены так, что $F * f = (F(f_{(1)}), \ldots, F(f_{(N)})) \in \mathbb{R}^N$ для любой $F \in \mathcal{F}$, определяющая преобразование $F\colon \mathbb{R}^N \to \mathbb{R}^N$ функция $F\colon \mathbb{R}^1 \to \mathbb{R}^1$ задана на числовой прямой и принимает числовые значения.

Пусть при некоторых идеальных условиях изображение сцены имеет вид $(f_{(1)}, \ldots, f_{(N)}) \in \mathbb{R}^N$, причём все координаты вектора f попарно различны. Заметим, что множество векторов, не удовлетворяющих этому требованию, имеет в $\mathbb{R}^N$ меру нуль.

Реальное измерение яркости изображения сцены в j-й точке поля зрения даёт результат

$$\xi_{(j)} = F(f_{(j)}) + \nu_{(j)}, \quad j = 1, \ldots, N, \qquad (2)$$

где $\nu_{(j)}$ — погрешность регистрации яркости изображения $F * f$ в точке поля зрения X с номером j, $j = 1, \ldots, N$. Преобразование $f \to F * f$ приводит к тому, что значения некоторых координат вектора $F * f$ могут совпасть.

Пусть идеальное изображение f сцены не наблюдается и известен лишь результат регистрации этой сцены в серии измерений, проводимых по схеме

$$\xi_i = F_i * f + \nu_i, \quad i = 1, 2, \ldots, \qquad (3)$$

где случайный вектор $\nu_i \in \mathbb{R}^N$ моделирует погрешность измерения изображения $F_i * f$; функции $F_i \in \mathcal{F}$, $i = 1, 2, \ldots$, неизвестны, но задан класс $\mathcal{F}$ преобразований $\mathbb{R}^N$ и при этом множество $V_f = \{g \in \mathbb{R}^N\colon g = F * f,\ F \in \mathcal{F}\}$ выпукло и замкнуто в $\mathbb{R}^N$ при фиксированном $f \in \mathbb{R}^N$, случайные векторы $\nu_i \in \mathbb{R}^N$ независимы в совокупности и имеют нулевое математическое ожидание $\mathbf{E}\nu_i = 0$, координаты вектора $\nu_i \in R^y N$ независимы и с вероятностью единица ограничены по модулю: $|\nu_{i(j)}| \leqslant \delta$, $i = 1, 2, \ldots$, $j = 1, \ldots, N$. Задача состоит в том, чтобы по результатам наблюдений (3) охарактеризовать сцену.

Если класс $\mathcal{F}$ достаточно широк, то оценить значения координат вектора f по результатам измерения (3) при неизвестных и произвольных $F_i \in \mathcal{F}$, $i = 1, 2, \ldots$, невозможно, тем не менее измерения (3) содержат некоторую информацию об изображении f, а значит и об исходной сцене. Охарактеризуем сцену полным инвариантом группы преобразований $\mathcal{F}$ и оценим его по результатам измерения (3).

Рассмотрим функцию $\pi(\cdot)$, действующую из $\mathbb{R}^N$ в $\bigotimes_{k=1}^{N} \{1, 2, \ldots, N\}$, результатом действия которой на вектор f является набор номеров его координат, упорядоченных по невозрастанию:

$$\pi(f) \equiv \pi((f_{(1)}, f_{(2)}, \ldots, f_{(N)})) = (i_1, i_2, \ldots, i_N), \quad (4)$$

при этом

$$f_{(i_1)} > f_{(i_2)} > \cdots > f_{(i_N)}.$$

Зададим группу (класс) монотонных преобразований $\mathcal{F}$ пространства $\mathbb{R}^N$, считая, что для любых $F \in \mathcal{F}$ и $f = (f_{(1)}, \ldots, f_{(N)}) \in \mathbb{R}^N$

$$F * f = (F(f_{(1)}), \ldots, F(f_{(N)})) \in \mathbb{R}^N,$$

где $F(\cdot) \in \mathbb{R}^1 \to \mathbb{R}^1$ — монотонно возрастающая функция.

Теорема 1. *Функция $\pi(\cdot)$ является максимальным инвариантом группы монотонных преобразований $\mathbf{F}$ пространства $\mathbb{R}^N$.*

Следовательно, указанная упорядоченность координат — это «наиболее полное» свойство сигнала f, которое может быть извлечено из наблюдений (3).

Проверка адекватности модели регистрации изображений сцены

Если при анализе данных используется их математическая модель, то необходимым этапом исследования является установления факта непротиворечивости модельных ограничений и результатов наблюдений [4]; отсутствие таких противоречий означает, что нет причин считать модель неадекватной.

Рассмотрим n изображений $\xi_1, \ldots, \xi_n \in \mathbb{R}^N$ и сформулируем ряд условий, которые будем называть условиями $\{H\}$:

— изображения $\xi_1, \ldots, \xi_n \in \mathbb{R}^N$ зарегистрированы согласно схеме (3);
— класс $\mathcal{F}$ является классом монотонных преобразований пространства $\mathbb{R}^N$;

— погрешности $\nu_i \in \mathbb{R}^N$, $i = 1, \ldots, n$, независимы в совокупности, обладают независимыми в совокупности координатами, причём модуль каждой из которых с вероятностью единица не превосходит δ.

Выполнение условий H равносильно тому, что все изображения (3) получены от одной и той же (неизвестной) сцены.

Упорядочение координат вектора f

Пусть выполнены условия $\{H\}$ и требуется определить функцию $\pi(\cdot)$, упорядочивающую координаты вектора $f = (f_1, \ldots, f_N)$ по убыванию. Для этого введем следующие обозначения:

$$\theta_\alpha^{(n)} = \frac{1}{n}\sum_{j=1}^{n} F_j(f_{(\alpha)}); \quad \theta_{\alpha\beta}^{(n)} = \theta_\alpha^{(n)} - \theta_\beta^{(n)}; \qquad (5)$$

$$\mu_\alpha^{(n)} = \frac{1}{n}\sum_{j=1}^{n} (\nu_j)_{(\alpha)}; \quad \mu_{\alpha\beta}^{(n)} = \mu_\alpha^{(n)} - \mu_\beta^{(n)}; \qquad (6)$$

$$\eta_{\alpha\beta}^{(n)} = \theta_{\alpha\beta}^{(n)} + \mu_{\alpha\beta}^{(n)}. \qquad (7)$$

В формулах (5)–(7) числа $\alpha, \beta = 1, \ldots, N$ — номера координат соответствующих векторов, $\alpha < \beta$, индекс n указывает количество измерений, по которым производилось усреднение. В соответствии с введёнными обозначениями задача эмпирического упорядочивания элементов вектора $f = (f_1, \ldots, f_N)$ в силу монотонности преобразований F свелась к задаче определения знаков $\theta_{\alpha\beta}^{(n)}$ на основе наблюдаемых значений $\eta_{\alpha\beta}^{(n)}$, $\alpha, \beta = 1, \ldots, N$, $\alpha < \beta$.

Теорема 2. *Пусть для любого конечного числа измерений* (3) *выполнены условия* $\{H\}$. *Тогда упорядоченность координат вектора* $f \in \mathbb{R}^N$ *определяется с вероятностью единица по конечному числу наблюдений* $\xi_1, \ldots, \xi_n \in \mathbb{R}^N$.

Построим алгоритм, упорядочивающий координаты изображения $f \in \mathbb{R}^N$ по наблюдениям (3). Для этого воспользуемся алгоритмом эмпирического упорядочения вероятностей элементарных событий, изменяющихся в процессе испытаний [5] и сформулируем его в следующем виде.

На n-м шаге алгоритма для каждой пары индексов (α, β), $\alpha, \beta = 1, \ldots, N$, $\alpha < \beta$ принимается одно из решений

— если $\eta_{\alpha\beta}^{(n)} > \Delta^{(n)}$, то принимается решение $\mathbf{S}_1$: считать $\theta_{\alpha\beta}^{(n)} > 0$,

— если $\eta_{\alpha\beta}^{(n)} < -\Delta^{(n)}$, то принимается решение $\mathbf{S}_2$: считать $\theta_{\alpha\beta}^{(n)} < 0$,

— если $|\eta_{\alpha\beta}^{(n)}| \leqslant \Delta^{(n)}$, то принимается решение $\mathbf{R}$: необходимо увеличить количество измерений n.

В этом алгоритме при каждом $n = 1, 2, \ldots$ проверяются условия решений $\mathbf{S}_1$, $\mathbf{S}_2$ и $\mathbf{R}$ для всех $N(N-1)/2$ пар (α, β), $\alpha, \beta = 1, \ldots, N$, $\alpha < \beta$. Если для всех пар приняты только решения $\mathbf{S}_1$ или $\mathbf{S}_2$, то алгоритм завершен, если же хотя бы для одной пары принято решение $\mathbf{R}$, то в (3) добавляется ещё одно измерение и для нового набора изображений вновь проверяются условия решений $\mathbf{S}_1$, $\mathbf{S}_2$ и $\mathbf{R}$ для всех $N(N-1)/2$ пар $(\alpha < \beta)$, $\alpha, \beta = 1, \ldots, N$.

Задача эмпирического упорядочения координат вектора f будет решена, если приняты все $N(N-1)/2$ решений $\mathbf{S}_1$ или $\mathbf{S}_2$ и оценены вероятности сопутствующих ошибок. Это позволит определить упорядоченность координат вектора $f \in \mathbb{R}^N$ с гарантированной вероятностью, совпадающей с их истинной упорядоченностью.

Лемма 3. *Вероятности ошибочных решений* $\mathbf{S}_1$ *и* $\mathbf{S}_2$ *в зависимости от величины* $\Delta^{(n)}$ *для фиксированных* α *и* β *удовлетворяют неравенствам*[1]

$$P(\{\eta_{\alpha\beta}^{(n)} > \Delta^{(n)}\} | \theta_{\alpha\beta}^{(n)} < 0) \leqslant \exp(-n(\Delta^{(n)})^2/8\delta^2) \equiv \lambda,$$

$$P(\{\eta_{\alpha\beta}^{(n)} < \Delta^{(n)}\} | \theta_{\alpha\beta}^{(n)} > 0) \leqslant \exp(-n(\Delta^{(n)})^2/8\delta^2) = \lambda.$$

Заданная априори величина λ оценивает сверху вероятность ошибочных решений $\mathbf{S}_1$ и $\mathbf{S}_2$ и определяет значение

$$\Delta^{(n)} = \left(\frac{8\delta^2}{n}\ln\frac{1}{\lambda}\right)^{1/2}, \quad n = 1, 2, \ldots \qquad (8)$$

Следующая теорема фактически переформулирует свойства алгоритма упорядочения из работы [5], являющегося прототипом алгоритма упорядочения яркости изображения f.

Теорема 4. *Пусть в предложенном алгоритме эмпирического упорядочения координат вектора* f *величина* λ *оценивает сверху вероятность ошибочного решения* $\mathbf{S}_1$ *или* $\mathbf{S}_2$, *и при всех достаточно больших* n *и всех* $\alpha < \beta$, $\alpha, \beta = 1, \ldots, N$, $|\theta_{\alpha\beta}^{(n)}| \geqslant (1 + \varepsilon^{(n)})\Delta^{(n)}$, *где* $\Delta^{(n)}$ *определено в* (8), *а* $\varepsilon^{(n)} > 0$ *и удовлетворяют условиям*

$$\sum_{n=1}^{\infty} \exp(-n(\varepsilon^{(n)}\Delta^{(n)})^2/8\delta^2) < \infty.$$

Тогда все $N(N-1)/2$ *решений* $\mathbf{S}_1$ *или* $\mathbf{S}_2$ *будут приняты на основе почти наверное конечного числа испытаний, и если* n — *число испытаний, при котором первый раз приняты все* $N(N-1)/2$ *решений, то вероятность того, что полученная упорядоченность совпадает с истинной, больше*

$$p = \frac{1 - N\lambda}{1 - \lambda} + \frac{\lambda^2(1-\lambda)^{N-1}}{(1-\lambda)^2} = 1 - (N-1)\lambda + o(\lambda).$$

[1] Под знаком вероятности после вертикальной черты стоит условие, которому удовлетворяет параметр вероятности.

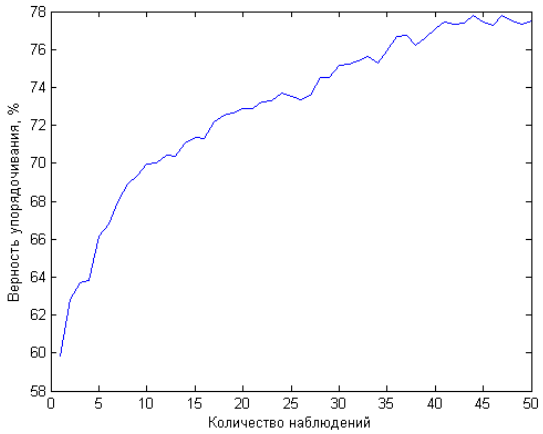


Рис. 4. Зависимость количества наблюдений (числа шагов n) от вероятности ошибки для изображения размера 70×100.

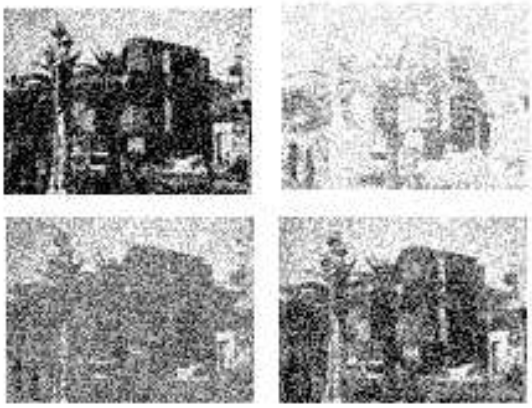


Рис. 2. Примеры измерений изображения сцены, выполненные при различном освещении с погрешностью.

Рис. 3. Результат работы алгоритма упорядочивания за 40 шагов.

Вычислительный эксперимент

В качестве демонстрации практического применения предложенной теории был проведен вычислительный эксперимент. Использовались 50 измерений сцены изображения размера 70×100, выполненные при различном освещении (с сохранение упорядоченности) и при достаточно сильном зашумлении. Примеры измерений приведены на рисунке 2. На рисунке 3 приведен результат работы алгоритма упорядочивания за $n = 40$ шагов.

На рисунке 4 показана экспериментально полученная зависимость количества наблюдений (числа шагов n) от вероятности ошибки. Хорошо видно, что в данном эксперименте для достижения точной с вероятностью единица упорядоченности необходимо более 50 измерений.

Выводы

В статье рассмотрена задача эмпирического построения формы изображения сцены по набору искаженных шумами изображений той же сцены, полученных при неизвестных условиях регистрации. Форме соответствует максимальный инвариант группы преобразований, моделирующих изменения изображений при вариациях условий регистрации. Предложен алгоритм, восстанавливающий точную форму изображения с любой наперед заданной вероятностью за конечное число шагов.

Построенная таким образом форма позволяет применять известные методы морфологического анализа [1], в которых форма считается известной.

Литература

[1] *Пытьев Ю. П., Чуличков А. И.* Методы морфологического анализа изображений — Москва: ФИЗМАТЛИТ, 2010.

[2] *Пытьев Ю. П.* Задачи морфологического анализа изображений. — В сб.: Математические методы исследования природных ресурсов Земли из Космоса. — Москва: Наука, 1984.

[3] *Куличков С. Н., Чуличков А. И., Демин Д. С.* Морфологический анализ инфразвуковых сигналов в акустике. — Москва: Изд-во "Новый Акрополь", 2010.

[4] *Пытьев Ю. П.* Методы математического моделирования измерительно-вычислительных систем. — Москва: ФИЗМАТЛИТ, 2004.

[5] *Пытьев Ю. П.* Математические методы и алгоритмы эмпирического восстановления стохастических и нечетких моделей. // Интеллектуальные системы. — 2007. — Т. 11. Вып. 1–4. — С. 277–327.

Сравнение двух классов функций преобразования яркости в задаче поиска структурных изменений*

Корнилов Ф. А., Костоусов В. Б., Перевалов Д. С.
vkost@imm.uran.ru
Екатеринбург, Институт математики и механики УрО РАН

В работе исследуется задача поиска структурных изменений на двух изображениях с помощью методов локального преобразования яркости. Показано, что для сильно зашумленных изображений использование класса линейных функций даёт более качественный результат, чем класса функций, порождаемых морфологическим проектором Ю. П. Пытьева.

Задача поиска изменений на двух изображениях возникает в разных областях компьютерного зрения, таких как сжатие видеоданных или системы видеонаблюдения. В данной работе рассматривается поиск изменений на космических снимках земной поверхности. Для задачи автоматического анализа космических снимков особый интерес представляет задача обнаружения не всех изменений, а так называемых *структурных изменений*, которые заключаются в существенном изменении наблюдаемой сцены. Таковыми изменениями будем считать появление, исчезновение или изменение формы объектов: домов, дорог, участков леса и т. д. При этом изменения освещённости и цвета объектов структурными изменениями не считаются.

В задаче поиска структурных изменений входными данными являются два разновременных космических снимка одного и того же участка земной поверхности. Предполагается, что они геометрически выровнены и имеют одинаковый размер в пикселах. Геометрическая выравненность означает, что каждый элемент (x, y) на обоих снимках имеет одни и те же координаты на местности. В работе рассматриваются только полутоновые изображения.

Все многообразие методов, разработанных для решения поставленной задачи, нам удобно разделить на три группы: поточечный анализ, поиск и анализ контуров объектов и текстурная классификация. Поточечный анализ работает с яркостями пикселов без понимания структуры изображения. Второй метод заключается в построении контуров вокруг объектов на изображениях и их последующих сравнении и анализе. В третьем методе проводится классификация пикселов изображений на типы объектов и структурным изменением считаются те области, где эти типы отличаются сильно (например, дома и лес). Как видно, второй и третий методы требуют предобработки, чтобы понять структуру изображения. Изучаемый нами метод относится к первой группе.

Работа выполнена при финансовой поддержке программ фундаментальных исследований Президиума РАН №09-П-1-1003, №09-П-1-1013 и гранта РФФИ №09-01-00523.

Алгоритм поиска структурных изменений

Общая схема алгоритма. Исходная пара изображений сканируется окном заданного размера. Как было указано во введении, алгоритм не должен находить объекты, цвет которых изменился. Поэтому нам необходимо провести «выравнивание» яркостей двух сравниваемых изображений с сохранением их структуры. Для этого будет использована функция преобразования яркости. После этого с использованием пороговой обработки определяется результат: наличие или отсутствие в данном окне структурных изменений и, в случае наличия, их локализация. Таким образом, обработка каждого сканирующего окна даёт на результирующем изображении одну точку, яркость которой 255 или 0. Далее из этих точек формируются связные области структурных изменений — конечный результат. На рис. 1 схематично представлена общая схема алгоритма.

Метод локальной корректировки яркости. Структура алгоритма, основанного на локальной корректировки яркости, такова: сначала по двум сравниваемым изображениям строятся две функции преобразования яркости $f_{\mathbf{AB}}$ и $f_{\mathbf{BA}}$. Использование схемы с двумя функциями позволяет добиться симметризации результата, т. е. становится неважно, сравнивать первое изображение со вторым или наоборот.

На основе этих функций строятся преобразованные изображения: $I_{\mathbf{A}'} = f_{\mathbf{AB}}(I_{\mathbf{A}})$, $I_{\mathbf{B}'} = f_{\mathbf{BA}}(I_{\mathbf{B}})$. При этом яркость изображения $I_{\mathbf{A}'}$ «выравнена» по яркости изображения $I_{\mathbf{B}}$ с сохранением структуры изображения $I_{\mathbf{A}}$.

Теперь вычисляются «разностные» изображения $|I_{\mathbf{A}'}(x, y) - I_{\mathbf{B}}(x, y)|$ и $|I_{\mathbf{B}'}(x, y) - I_{\mathbf{A}}(x, y)|$. Для этих изображений яркость точки характеризует величину структурного несоответствия исходных изображений. То есть чем ярче точка, тем более вероятно, что в ней присутствует структурное изменение. Из этих двух изображений строится одно — $I_{\mathbf{R}}$, яркость каждой точки которого есть максимум яркостей точек «разностных» изображений с соответствующими координатами.

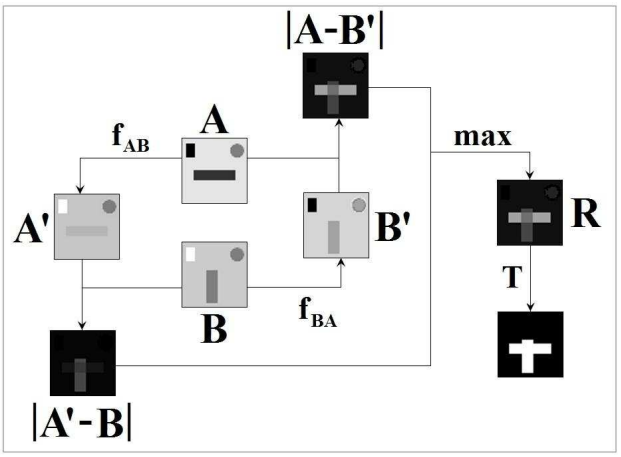


Рис. 1. Общая схема метода локальной корректировки яркости.

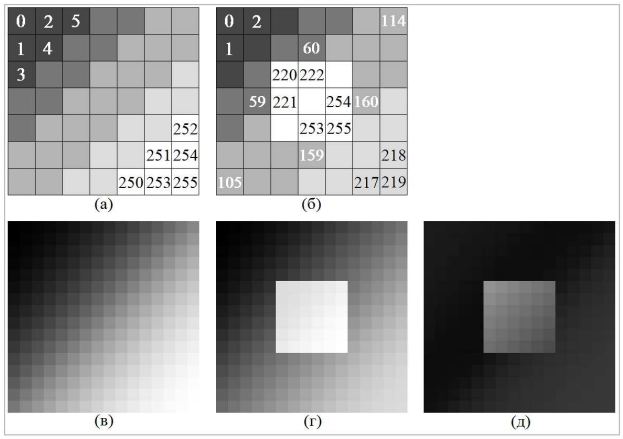


Рис. 2. Примеры изображений, в которых морфологический проектор работает не очень хорошо.

Последний шаг — пороговая обработка изображения $I_{\mathbf{R}}$. Выбор величины этого порога — не совсем простая задача. Можно задать порог один раз для всех сканирующих окон. Можно использовать методы автоматического определения порога, например, метод Оцу [1].

Функции преобразования яркости. Как было указано выше, локальная корректировка яркости основана на использовании функции преобразования яркости. Класс таких функций, в принципе, очень широк. В данной работе мы рассмотрим две такие функции.

Первая из функций была предложена профессором Ю. П. Пытьевым (ниже для краткости называемая морфологическим проектором) в работе [2] и представляет собой оптимальный проектор в некотором пространстве функций относительно специального критерия невязки. Её значение для некоторой яркости c есть среднее значение набора яркостей изображения $\mathbf{B}$ в пикселах, яркость которых на изображении $\mathbf{A}$ равна c. То есть эта функция усредняет второе изображение $\mathbf{B}$ по уровням яркости первого изображения $\mathbf{A}$. Её формула такова:

$$f_{\mathbf{AB}} = E_{\mathbf{A_c}}(\mathbf{B}(x,y))$$
$$\mathbf{A_c} = \{(x,y) : \mathbf{A}(x,y) = c\}$$

Однако у этой функции есть недостаток: в случае, если на первом изображении присутствует слишком много уровней яркости, результат преобразования первого изображения будет очень близок ко второму как по яркости, так и по структуре. Например, для двух произвольных изображений размером 16 на 16 пикселов, принимающих все 256 значений яркости, после преобразования с помощью морфологического проектора на изображениях $\mathbf{A}'$ и $\mathbf{B}'$ значения яркостей пикселов буду совпадать со значениями соответствующих пикселов изображений $\mathbf{B}$ и $\mathbf{A}$ соответственно. Потому никаких изменений на этих изображениях обнаружено не будет.

В то же время, как показано на рис. 2, можно так подобрать эти изображения, что на них будет присутствовать заметное структурное изменение. Здесь (а) и (б) — способ раскраски, (в) и (г) — получившиеся изображения. В реальном случае аналогичная ситуация возможна, если на изображении присутствует сильный шум.

Одним из способов преодоления обнаруженного ограничения является сужение класса функций, из которого выбирается функция преобразования яркости. Простейшим вариантом является использование класса линейных функций:

$$I_{\mathbf{A}'}(x,y) = kI_{\mathbf{A}}(x,y) + b,$$

где коэффициенты k и b находятся методом наименьших квадратов (МНК) из соотношения

$$kI_{\mathbf{A}}(x,y) + b = I_{\mathbf{B}}(x,y)$$

На рисунке 2 изображение (д) — разность преобразованного с помощью этой функции первого изображения и второго изображения. Можно видеть, что структурное изменение будет несложно локализовать с помощью пороговой обработки.

Сравнение функций преобразования яркости

Используем анализ ROC-кривых для сравнения двух описанных функций преобразования яркости. Для этого определим функцию точности (или не-риска) для оценки правильности обнаружения структурного изменения:

$$R_{sum} = S^{+}_{obj}/S_{obj} + S^{-}_{fon}/S_{fon},$$

где S^{+}_{obj} — число пикселов, определённых как структурное изменение среди пикселов объек-

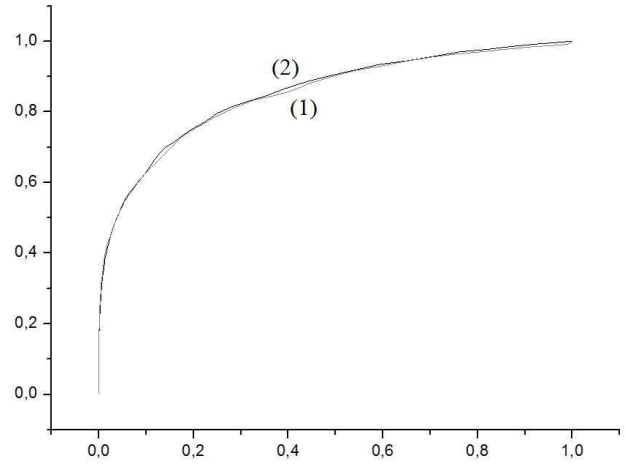


Рис. 3. ROC-кривые по порогу. Случай умеренного шума.

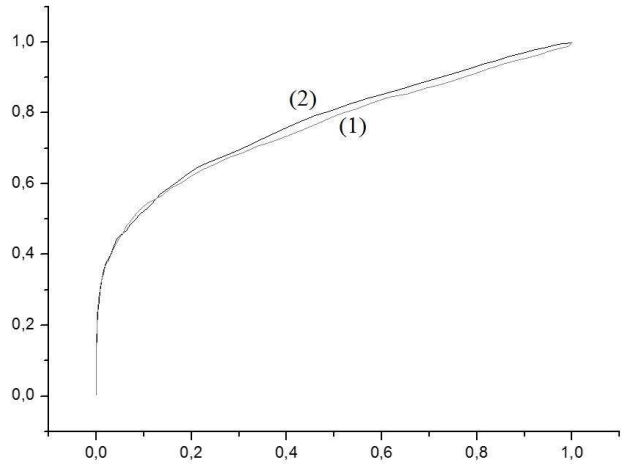


Рис. 4. ROC-кривые по порогу. Случай сильного шума.

та, S_{obj} — площадь объекта, S^-_{fon} — число пикселов, не определённых как структурное изменение среди пикселов фона, S_{fon} — площадь фона. В этой функции: S^+_{obj}/S_{obj} — доля верно обнаруженных структурных изменений (чувствительность), $1 - S^-_{fon}/S_{fon}$ — доля ложных тревог (специфичность).

По реальным, предварительно вручную размеченным и умеренно зашумленным космоснимкам для каждой функции были построены ROC-кривые при изменении значения порога от 0 до 255. На рис. 3 изображены эти ROC-кривые: здесь (1) — ROC-кривая, соответствующая морфологическому проектору, (2) — соответствующая построенной с помощью МНК линейной функции. Как можно видеть, в случае умеренного шума результаты оказались близки.

ROC-кривые, построенные для случая сильнозашумленных изображений, представлены на рис. 4. Как можно видеть, линейная функция (2) превзошла морфологический проектор (1).

Заключение

С теоретической точки зрения сужение класса функций до класса линейных даёт более устойчивые результаты в случае зашумленных изображений. Численный эксперимент показал: такое сужение не привело к значительной потере качества обработки; в то же время, если на изображениях присутствует сильный шум, то основанная на вычислении МНК линейная функция даёт более качественный результат.

Литература

[1] *Otsu N.* A threshold selection method from gray-level histograms // IEEE Trans. Sys., Man., Cyber. 9. Pp. 62–66.

[2] *Пытьев Ю. П.* Морфологический анализ изображений // Доклады академии наук СССР. — 1983 — Т. 269, №. 5. — С. 1061–1064.

[3] *Корнилов Ф. А., Перевалов Д. С.* Обнаружение изменений объектов земной поверхности по спутниковым снимкам // Проблемы теоретической и прикладной математики. Труды 41-й Всероссийской молодежной конференции. — 2010, С. 534–540.

[4] *Корнилов Ф. А., Перевалов Д. С.* Оценка оптимального порога для алгоритма поиска структурных изменений с помощью проектора Пытьева // Современные проблемы математики. Тезисы 42-й Всероссийской молодежной школы-конференции. Екатеринбург: ИММ УрО РАН. — 2011, С. 291–293.

Метод геометризованных гистограмм, дуальное описание сцен и его применение*

Кий К. И.
kikip_46@mail.ru
Москва, Институт Прикладной Математики им. Келдыша РАН

В работе предлагается новый подход к описанию содержания цветных изображений, который объединяет преимущества областного и контурного описаний сцен. Данный подход основан на методе геометризованных гистограмм, разработанном автором, и позволяет работать с «оснащенными» контурами с описанием областей «слева» и «справа» от контура. Рассматриваются возможные приложения разработанного подхода к решению задач анализа и понимания изображений.

В последнее время достигнут большой прогресс в решении задач понимания (категоризации) изображений сцен, типичных для зрительных задач, возникающих из робототехники [1–3]. В значительной степени этот прогресс инициируется расширением круга задач, решаемых автономными роботами. Данные задачи представляются особенно актуальными в свете последних событий в Японии. Становится ясным, что автономные роботы должны быть готовы к выполнению широкого класса работ, невозможных без создания интеллектуальных зрительных систем, способных анализировать изображения в ситуациях, когда модели среды заранее не известны. Работы, цитированные выше (а также те, которые могут найдены в ссылках данных работ) показывают интересные результаты на основе применения различных детекторов точек и множеств интереса. Однако эти методы медленно работают на сложных сценах с многими объектами, типичных для зрительных задач робототехники, плохо отыскивают мелкие объекты, не очень приспособлены к решению задач распознавания в реальном времени и не дают возможности хорошо описывать структурные связи между объектами. В частности, эти методы плохо прикладываются к задаче анализа движения в кадре, когда съемка ведется с движущегося объекта, а в кадре могут двигаться в разных направлениях несколько объектов разного разного размера.

В работах автора [4–8] на основе разработанного метода геометризованных гистограмм предлагается описание изображений с помощь структур окрашенных отрезков, связанных с некоторым разбиением изображения на полосы одинаковой ширины параллельные некоторой оси в плоскости изображения. В настоящей работе даётся приложение разработанной техники к выделению и описании виртуальных цветовых контуров, ограничивающих объекты. При этом получаются «оснащенные» контура, содержащие информацию о «теле» объектов по обе стороны от контура. Такие контура будут полезны при анализе заслонений, движения и анализе стерео пар изображений.

Работа выполнена при финансовой поддержке РФФИ, проект № 10-07-00612а.

Метод геометризованных гистограмм, оснащённые контура

В этом разделе мы приводим развитие данного теоретического метода сжатого описания и анализа цветных изображений, которое позволяет установить связь между подходом, основанным на контурах, с подходом, основанном на областях. В последнее время существенное развитие получили методы описания и анализа изображений, основанные на выделении характерных особенностей (salient features) изображений [9, 1, 2, 3]. Данные методы также применяются к стерео анализу сцен [10] (см. также ссылки в этой работе). Стоит заметить, что для человека анализ с помощью существенных точек выглядит несколько экзотически, несмотря на получающиеся убедительные результаты. По нашему мнению, введённые существенные точки следует дополнить контрастными точками границ областей и самими контрастными частями границ. Вместе с контрастными областями они дают основу анализа сцены людьми и животными при быстром движении, в то время как существенные точки, введённые в вышеупомянутых работах являются результатом внимательного разглядывания сцены. В работе даётся попытка формализации новых классов существенных точек на основе метода геометризованных гистограмм.

Постановка задачи, виртуальные контура, ограничивающие области. Существует много методов выделения контурных препаратов [11, 12]. Однако на сложных сценах с многими объектами и заслонениями задача связывания контуров объектов из отдельных частей без привлечения семантики может не иметь решения и кроме того является сложной вычислительной проблемой. В случае, когда границы объектов содержат прямолинейные сегменты, задача нахождения контуров может эффективно решается с помощью преобразований Хафа [12, 13]. Однако при наличии заслонений могут возникнуть ложные прямолинейные отрезки, которые могут привести к неверным решениям. С помощью только одних контуров так-

же бывает невозможно отследить движение (быстрое) на сложных сценах. Выделение границ объектов с помощью классических методов сегментации может занять много времени и привести к неверным результатам ввиду ошибок классических методов сегментации [3].

Метод геометризованных гистограмм [4–8] позволяет поставить в соответствие каждому цветному изображению CI (пиксельному массиву) некоторую структуру отрезков (структурный граф цветовых сгустков), принадлежащих некоторой оси Ax (горизонтальной или вертикальной), выбранной в плоскости изображения. Обозначим эту структуру $STG(CI)$. Для получения $STG(CI)$ исходное цветное изображение разбивается на узкие полосы одинаковой ширины параллельные выбранной оси Ax. Граф $STG(CI)$ состоит из слоев, каждый слой соответствует некоторой полосе разбиения, вершины $STG(CI)$ представляются окрашенными отрезками на Ax, ребра графа соединяют вершины в соседних полосах, отрезки которых имеют близкие цветовые характеристики и близки как отрезки на Ax. $STG(CI)$ получается из некоторой промежуточной структуры отрезков, которая называется геометризованной гистограммой изображения (полосы). Для выбранной полосы отрезки этой структуры получаются проекцией на Ax множеств уровня характеристической функции полосы CF, которая получается из $G/(G+B)$ введением дополнительных значений, чтобы разделить цветовые диапазоны и отделить малонасыщенные, яркие и тёмные места. Детали могут быть найдены в [5–7]. Каждому множеству уровня соответствует группа отрезков на Ax. Детальное описание проекции может быть найдено в работе [8], доступной в Интернете. Каждому отрезку геометризованной гистограммы ставится в соответствие набор чисел

$$(H_{min}, H_{max}, H_{mean}, S_{min}, S_{max}, S_{mean},$$
$$gr_{min}, gr_{max}, gr_{mean}, card, beg, end), \qquad (1)$$

где $H_{min}, H_{max}, H_{mean}$, $S_{min}, S_{max}, S_{mean}$ и gr_{min}, gr_{max}, gr_{mean} — диапазоны и средние значения оттенка, насыщения и полутоновой компоненты соответственно, $card$ — мощность отрезка (приблизительно число точек в прообразе в полосе, которые имеют значения характеристик, принадлежащие заданным диапазонам), а beg и end — координаты его начала и конца на Ax.

С помощью процедур кластеризации отрезки геометризованной гистограммы объединяются в группы. Каждой группе отрезков присваивается новый набор (1), и полученные интегральные отрезки (группы) называются цветовыми сгустками. При кластеризации в один сгусток объединяются отрезки, которые имеют сильное пересечение с некоторым предварительно выбранным отрезком — зародышем сгустка. Так как данное представление строится без потерь, интервалы цветовых сгустков могут пересекать друг друга и даже совпадать. Многочисленные примеры структур цветовых сгустков, наложенных на изображение, которое определяется полутоновой компонентой исходного изображения, могут быть найдены в [14, 15, 8]. Эти и другие примеры показывают, что графы $STG(CI)$ содержат существенную часть информации исходного изображения. На множестве цветовых сгустков можно определить задачу сегментации (выделения образов объектов изображения на структуре сгустков) [5–7]. В этих статьях описывалась процедура построения глобальных объектов, которая продолжает цветовые сгустки на соседние полосы. Однако в ряде случаев (особенно для изображений при неоднородном освещении и с помехами) таким образом мы получаем только части реальных объектов и иногда достаточно мелкие, с которыми трудно работать при решении задач распознавания. Поэтому необходимо объединить мало различимые цветовые сгустки, убрать те сгустки, которые оцениваются как несущественные и найти интервалы контрастного перехода между сгустками, которые порождают участки полосы. Заметим, что на данной стадии мы будем объединять соседние цветовые сгустки, отрезки которых имеют слабое пересечение, т. е. объединение будет происходить вдоль полос, в то время как при построении цветовых сгустков объединение происходило в основном поперек полос. Каждый цветовой сгусток может быть основным в своей части полосы, быть доминантом (мощность остальных сгустков несущественна по сравнению его мощностью), либо быть одним из нескольких сгустков приблизительно одинаковой мощности. Второй случай возможен в точках, где встречаются границы нескольких областей, окрашенных в разные цвета, либо в точках цветовой текстуры. Как в первом, так и во втором случае небольшие объекты (ориентиры) могут быть расположены на фоне доминирующих сгустков. Расположение цветового сгустка в полосе задаётся интервалом $Int = [beg, end]$ на оси Ax. Для двух цветовых сгустков $bunch_1, bunch_2$ их интервалы $Int_1 = [beg_1, end_1]$ и $Int_2 = [beg_2, end_2]$ могут пересекаться, в том числе один интервал может содержать другой. Интервалы двух соседних доминант могут как не пересекаться, так и иметь некоторое пересечение.

Поясним это на примерах. Пусть выбрано разбиение на горизонтальные полосы. Если две разные области постоянного цвета на модельном изображении разделены вертикальной прямой, то соответствующие цветовые сгустки (их интервалы) не пересекаются. Если разделяющая прямая наклонная, то цветовые сгустки имеют пересечение с длиной отрезка пересечения, пропорциональной

углу наклона прямой и ширине полосы. Разумеется, помехи могут слегка искажать данную картину. Возникает вопрос является ли переход между двумя соседними цветовыми сгустками в полосе контрастным. Этому переходу можно поставить в соответствие либо две граничные точки конца и начала касающихся интервалов, либо интервал пересечения $Int_{12} = Int_1 \bigcap Int_2$. Будем считать, что в общем случае переход между соседними сгустками всегда осуществляется через некоторый интервал. Для соприкасающихся цветовых сгустков положим $Int_{12} = [end_1, beg_2]$. Заметим, что Int_{12} определяет границу между цветовыми сгустками с некоторой неопределённостью, зависящей от направления границы между реальными объектами, соответствующими цветовым сгусткам. В случае, когда Int_{12} вырождается в две или одну точку (когда начало и конец соседних интервалов совпадут), граница полностью определена. В следующем параграфе будут сформулированы правила для разделения границ цветовых сгустков на три класса: контрастная граница, несущественная граница (цветовые сгустки могут быть объединены) и граница со слабым контрастом (дальнейший вывод должен быть сделан с помощью других критериев). Если имеется набор цветовых сгустков в соседних полосах изображения, таких что каждый из них имеет контраст со своим соседним сгустком в полосе и эти сгустки непрерывно продолжают друг друга [5–7] (соответствуют некоторому глобальному объекту на изображении), то интервалы Int_{12} этих цветовых сгустков образуют некоторый виртуальный контур (контур образованный граничными интервалами, а не точками (как в случае классических контуров), определяющий границу с некоторым допуском). Глобальные объекты с контрастными границами дают контрастные объекты на изображении. Заметим, что каждый контрастный интервал имеет цветовой сгусток слева и справа, что даёт дополнительное описание контрастного перехода и может быть использовано при анализе пар изображений. Добавим также, что при разбиении изображения на горизонтальные (вертикальные) полосы более точные границы находятся для вертикальных (горизонтальных) объектов.

Контрастные граничные интервалы на $STG(CI)$. Пусть $bunch_1, bunch_2$ — две соседних доминанты. Надо понять, является ли переход от $bunch_1$ к $bunch_2$ контрастным. Данная задача не является чисто математической и должна решаться с учётом особенности человеческого зрения. Так как роботы должны решать задачи, которые им ставит человек-оператор, они должны видеть подобно окружающую среду и формулировать результаты в терминах, понятных ему. Известно, что если интервалы значений характеристик не пересекаются, но очень близки, человек все равно видит разницу между объектами. Кроме того, в разных диапазонах значений H, S, I различающие способности зрения человека различны. Например, при слабом освещении заметность цветов меняется. В [11] имеются ссылки на работы психологов по определению пределов изменения цвета, незаметного зрению человека. Результаты оказались зависимыми от цветового диапазона. Необходимо отметить, что эксперименты проводились при неизменной полутоновой компоненте.

Нами делается попытка решить поставленную задачу в более общей постановке, но с намерением получить менее тонкие результаты. Решение будет трёх типов: (1) рассмотренные сгустки можно объединить, (2) сгустки являются контрастными, и (3) решение по сгусткам не принято, для принятия решения требуются другие критерии (например, семантика). Нами построена система рассуждений, реализованная продукциями, которая позволяет решать задачу с результатами, представляющимися разумными на широком классе изображений в разных диапазонах цветовых характеристик. Чтобы учесть это, значения H и I разделены на 8 и 6 зон соответственно. Значения S квантованы на 16 позиций $0, 1, \ldots, 15$. Для каждого значения S и каждой зоны H и I вводятся пороги возможных и невозможных значений отклонений H, S, I: $PosThresh_f(x)$, $ImPosThresh_f(x)$, где f принимает значения H, S, I, x — одно из квантованных значений S или зон H, I. Однако эти пороги имеют только «совещательный» голос в правилах и отделяют только очевидные решения. Каждый сгусток $bunch_i$, $i = 1, 2$ характеризуется четырьмя интервалами: $Int_i = [beg_i, end_i]$ на оси Ax (геометрия) и $Int_i^f = [f_{min}, f_{max}]$, $f = H, S, I$ — интервалами изменения H, S, I. Для отбрасывания случайных отклонений Int_i^f симметризованы относительно среднего значения f_{mean} (по меньшему отклонению).

Пусть It любой интервал на некоторой числовой оси и It_1, It_2 — два произвольных интервала. Обозначим длину It как $L(It)$ и две меры близости интервалов [6] как

$$d_{min}(It_1, It_2) = L(It_1 \cap It_2) / \min(L(It_1), L(It_2)),$$

$$d_{max}(It_1, It_2) = L(It_1 \cap It_2) / \max(L(It_1), L(It_2)).$$

Переменные f_{dev}^{12}, $f = H, S, I$, обозначают отклонения средних значений для Int_i^f. Если $Int_1^f \cap Int_2^f = \varnothing$, то переменным $d_{min}(Int_1^f, Int_2^f)$ присваиваются расстояния между Int_i^f, полученные по сходным принципам. Вводятся переменные hue_close, sat_close, $inten\_close$ со значениями в множестве (3,2,1,0,-1,-2,-3). Для различных диапазонов H, S, I, строятся решающие правила, которые на основе значений переменных f_{dev}^{12}, $d_{min}(Int_1^f, Int_2^f)$,

Рис. 1. Контрастные цветовые сгустки границы объектов.

$d_{max}(Int_1^f, Int_2^f)$ присваивают значения hue_close, sat_close, $inten\_close$. Далее строится различающая функция

$$Discr(hue\_close, sat\_close, inten\_close),$$

которая решает, к какому классу отнести границу между $bunch_1, bunch_2$.

Процедура определения контрастности между цветовыми сгустками реализована программно и включена в комплекс программ реализующих метод геометризованных гистограмм [5–7]. Многочисленные эксперименты с изображениями, полученными в различных условиях, показали, что выделяется существенная часть контурных точек реальных объектов. Рисунок приводит пример нахождения контрастных точек цветовых переходов на изображении. Цветовые сгустки нанесены на средние линии полос изображения, задаваемого полутоновой компонентой. Контрастные сгустки снабжены вертикальными отрезками на концах.

Выводы

На основе метода геометризованных гистограмм предложен новый тип характерных точек и множеств на изображении, контрастные точки на структурном графе и виртуальные кривые контрастных границ областей. Разработаны методы для выделения данных особенностей в реальном времени. Предложенные инварианты могут быть основой для разработки системы реального времени для поиска ориентиров на реальных изображениях среды действия автономных мобильных роботов, распознавания контрастных объектов в среде и анализа их движения.

Литература

[1] *Espinace P., Kollar T., Soto A., and Roy N.* Indoor Scene Recognition through Object Detection // IEEE Int'l Conf. on Robotics and Automation (ICRA), 2010.

[2] *Kang H., Efros A., Hebert M., and Kanade T.* Image Composition for Object Pop-out // IEEE Workshop on 3D Representation for Recognition, in association with IEEE International Conference on Computer Vision (ICCV), 2009.

[3] *Mishra A.K., Aloimonos Y.* Active Segmentation // International Journal of Humanoid Robotics. — 2009. — Vol. 6, № 3. — P. 361–386.

[4] *Kiy K. I.* A New Method for Description and Generalized Segmentation of Color Images in Real Time // Int'l Conf. on Pattern Recognition and Image Analysis: New Information Technologies, 2008. — Pp. 297–300.

[5] *Кий К. И.* Геометризованные гистограммы и понимание изображений // XIV всеросс. конф. Математические методы распознавания образов, ММРО-14, Суздаль, Макспресс, 2009. — С. 362–365.

[6] *Kiy K. I.* A New Real-Time Method for Description and Generalized Segmentation of Color Images // Pattern Recognition and Image Analysis. — 2010. — Vol. 6, № 2. — P. 169–178.

[7] *Кий К. И.* Модифицированный метод геометризованных гистограмм и его применение // 8-ая Международная конференция "Интеллектуализация обработки ин-формации ИОИ-2010, Пафос, Кипр, Макспресс, 2010. — С. 367–370.

[8] *Кий К. И.* Автоматическая система реального времени для обнаружения объектов и ориентиров на изображении, основанная на обработке цветных изображений // Техническое зрение в системах управления мобильными объектами-2010, Москва: Университет, книжный дом, 2011. — С. 268–276, http://tvcs2011.technicalvision.ru/.

[9] *Mikolajczyk K., Tuytelaars T., Schmid C., Zisserman A., et. al* A Comparison of Affine Region Detectors // International Journal of Computer Vision. — 2005. — Vol. 65, № 1, 2.

[10] *Blokhinov Yu. B.* Automation of Mutual Orientation of Digital Images Based on Computer Vision Algorithms // Journal of Computer and Systems Sciences International. — 2010. — Vol. 49, № 6. — P. 981–991

[11] *Форсайт Д. А., Понс Ж.* Компьютерное зрение. Современный подход. — Москва: Вильямс, 2004. — 926 с.

[12] *Визильтер Ю. В., Желтов С. Ю., Бондаренко А. В., Ососков М. В., Моржин А. В.* Обработка и анализ изображений в задачах машинного зрения. — Москва: Физматкнига, 2010. — 671 с.

[13] *Barinova O., Lempitsky V., Tretiak E., and Kohli P.* Geometric Image Parsing in Man-Made Environments // European Conference on Computer Vision, 2010.

[14] `www.sites.google.com/site/colorvisionkikiy/` — Color vision — 2010.

[15] `www.tz2010.cosmos.ru` — Cosmos.ru — 2010.

Алгоритм супериерархического подавления шума в видеоряде

Василенко С. И., Прокофьев А. В.
alexandr_prk@mail.ru
Тверь, ТвГУ

В данной работе построена модель и алгоритм супериерархического подавления шума в видеоряде, обладающие рядом преимуществ относительно последовательной компенсации движения. Приведены оценки экспериментальных данных на основе объективного показателя — пикового отношения сигнал-шум в сигнале.

Введение

Эффективность применения алгоритмов и методов выделения, распознавания, сопровождения объектов на основании данных видеоряда в значительной степени зависит от показателя отношения сигнал-шум входной информации [1]. При использовании данных методов в областях технического телевидения, характеризуемых высоким уровнем шума в сигнале, получаемом от приемника (например, рентгена, тепловизора), необходима предварительная обработка входной информации с целью подавления шума с минимальными потерями «полезного» сигнала. В существующих методах подавления шума в видеоряде улучшение требуемого показателя осуществляется с помощью *внутрикадрового* (сглаживающие фильтры для каждого кадра в отдельности) или *межкадрового подавления* шума (усреднение соседних кадров потока с учётом векторов движения областей кадров). Предложенный метод *супериерархической* компенсации можно отнести к третьему классу методов подавления шума — *комбинированному* или *3D-фильтры*. Название вытекает из структуры способа поиска векторов движения и способа получения суммированного кадра видеопоследовательности.

Поиск векторов движения

Поиск векторов движения в предлагаемом методе подавления шума производится для подматриц изображения кадра, имеющих равный размер $m \times m$ пикселей и основывается на минимизации значения суммы абсолютных разностей элементов (пикселей) в подматрицах кадров:

$$\min_{k,l\in[0,w]} \sum_{i=x}^{x+m}\sum_{i=y}^{y+m} \left|T^1_{i,j} - T^2_{i+k,j+l}\right| \to (k,l)\,,$$
$$x \in [1, N-m]\,, y \in [1, M-m]\,, \quad (1)$$

где:
x, y — положение подматрицы в матрице кадра;
N, M — размер матрицы изображения;
w — размер области (окна) поиска переместившейся подматрицы;
$T^1_{i,j}$ и $T^2_{i,j}$ — матрицы изображений кадров;
(k,l) — вектор движения подматрицы.

Выбор метода отыскания минимума зависит от требуемой вычислительной сложности и точности определения вектора движения и может быть различным — от полного перебора w^2 подматриц из области поиска до более сложных алгоритмов, например, описанных в [3] и [7].

Операция поиска вектора движения для любой подматрицы считается выполненной, если найденная минимальная сумма абсолютных разностей её элементов и элементов подматрицы с рассматриваемым вектором движения (k,l) меньше среднего уровня шума в кадрах:

$$\sum_{i=x}^{x+m}\sum_{i=y}^{y+m} \left|T^1_{i,j} - T^2_{i+k,j+l}\right| < \Delta \times m^2,$$
$$\Delta = \frac{1}{M \times N}\sum_{i=1}^{M}\sum_{j=1}^{N}\left|T^1_{i,j} - T^2_{i,j}\right|. \quad (2)$$

Принятие решения об изменении подматрицы и, соответственно, поиск вектора движения осуществляется в том случае, если не выполняется условие (2) с нулевыми значениями k и l.

Найденных таким образом векторов движения достаточно для дальнейшей компенсации перемещения объектов кадра при получении оценки видеоряда на основе усреднения кадров.

Вейвлет-преобразование

С целью уменьшения вычислительной сложности процедура поиска векторов движения в предлагаемом методе проводится не для изображений кадров, а для результатов их n-уровневого двумерного вейвлет-преобразования (ВП) [5], являющегося комбинацией одномерных ВП для строк и столбцов матриц изображения.

$$v_j^{(i+1)} = \sum_{k\in Z} v_{2j+k}^{(i)} h_k,$$
$$w_j^{(i+1)} = \sum_{l\in Z} v_{2j+l}^{(i)} g_l, \quad (3)$$
$$i, j \in Z,$$

где:
$v_j^{(i+1)}$ — сглаженный сигнал на уровне $i+1$,
$w_j^{(i+1)}$ — коэффициенты преобразования на уровне $i+1$,
h_k, g_k — коэффициенты низкочастотного и высокочастотного фильтра применяемого вейвлета.

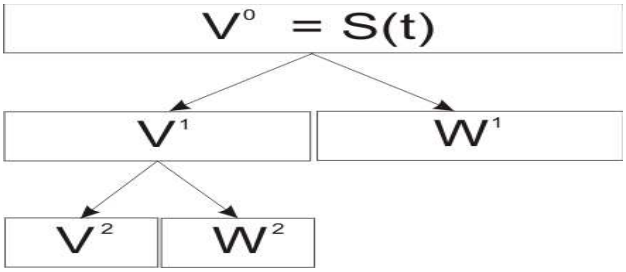


Рис. 1. Схема алгоритма Мала.

Замечание 1. Первый уровень иерархии, давшей название рассматриваемому методу, состоит в вычислении вейвлет-перобразования по схеме Мала, представляющей собой иерархическую структуру Рис. 1.

Одним из результатов такого преобразования на n-м шаге являются сглаженные и уменьшенные в 2^n раз проекции исходных изображений кадров, которые и используются для поиска векторов движения. Из этого следует уменьшение вычислительной сложности процедуры поиска векторов движения. Так, при отыскании минимума для критерия (1) с помощью полного перебора подматриц из окна поиска, необходимо вычислить не w^2 сумм абсолютных разностей пикселей, а $\left(\frac{w}{2^n}\right)^2$.

Другим результатом ВП является возможность подавления шума обработкой вейвлет-коэффициентов перед обратным вейвлет-преобразованием:

$$v_j^{(i)} = \sum_{k \in Z} \left(v_k^{(i+1)} h_{j-2k} + w_k^{(i+1)} g_{j-2k} \right), \quad i = 0 \ldots n, j \in Z. \qquad (4)$$

Она заключается в удалении коэффициентов с малыми амплитудами, то есть с высокой вероятностью относящихся к шумовому сигналу. Проведение данной процедуры является аналогом внутрикадровых методов подавления шума в видеоряде. По результатам многочисленных исследований (например, [2, 6]) отношение сигнал-шум в преобразованных кадрах увеличивается по сравнению с исходными. Что, согласно условию (2), в свою очередь, уменьшает количество подматриц, для которых принимается решение о необходимости поиска вектора движения, и уменьшает ошибки вычисления векторов для изменившихся подматриц.

Иерархический поиск векторов движения

В предлагаемой модели поиск векторов движения организован по иерархической схеме (Рис. 2), являющейся комбинацией методов, описанных выше:

Рис. 2. Иерархическая схема поиска векторов движения.

1) рассчитывается n-уровневое двумерное ВП (количество уровней зависит от разрешения кадра);
2) для самого нижнего n-го уровня ВП кадра по критерию (1) производится поиск значений векторов движения для изменившихся подматриц сглаженных проекций кадров;
3) производится обратное ВП с удалением шумовых вейвлет-коэффициентов;
4) на основе найденных векторов движения на предыдущем уровне вейвлет-преобразования кадра по критерию (1) производится поиск значений векторов для уровня преобразования выше $(n-1, n-2, \ldots, 0)$ со значительно меньшими размерами области поиска (уменьшение окна поиска обусловлено наличием информации о возможных значениях искомых векторов);
5) пункты 3 и 4 повторяются до тех пор, пока не будет восстановлен исходный кадр.

Таким образом, поиск вектора движения осуществляется на самом низком уровне вейвлет-преобразования изображений кадров с последующим иерархическим их уточнением на восстановленных изображениях.

Замечание 2. Иерархическая схема поиска векторов движения создаёт второй уровень иерархии в структуре предлагаемого метода.

Супериерархическая компенсация

Увеличение отношения сигнал-шум в видеоряде производится с помощью получения суммированного кадра, содержащего усреднённые значения сигналов нескольких скомпенсированных кадров. Количество усредняемых кадров зависит от условий решаемой задачи и значительно прореживает видеоряд.

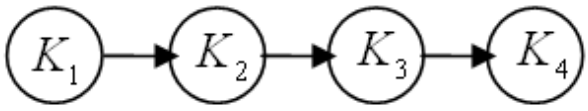


Рис. 3. Схема последовательной компенсации движения.

Стандартный способ формирования суммированного кадра — *последовательная* компенсация движения (Рис. 3), которая осуществляется для каждого последующего соседнего кадра в строгой очередности.

Такой способ компенсации движения обладает рядом характерных недостатков:

1) по мере увеличения числа обработанных кадров в суммированном кадре улучшается отношение сигнал-шум, а в исходном видеоряде оно постоянно, поэтому при обработке каждого нового кадра точность нахождения вектора смещения на основе критерия (1) уменьшается из-за различия значений среднего шума в кадрах;
2) расстояние между движущимися объектами на суммированном и обрабатываемом кадрах увеличивается, что влечёт увеличение размеров области поиска, которые могут превысить заданные или существенно увеличить время поиска.

В предложенном методе для формирования суммированного кадра используется *иерархическая* компенсация (Рис. 4), в которой процесс усреднения сигнала разбивается на несколько уровней, число которых зависит от количества обрабатываемых кадров:

1) исходные кадры разбиваются на пары нулевого уровня, в которых осуществляется поиск векторов движения с помощью иерархического поиска;
2) пиксели пар кадров обрабатываемого уровня усредняются с учётом движения:

$$T_{i,j}^{(L+1)N} = \frac{T_{i,j}^{LN} + T_{i+k_i,j+l_j}^{L(N+1)}}{2} \qquad (5)$$

где:
$T_{i,j}^{(L+1)N}$ — суммированный кадр с номером N нового уровня $L+1$,
$T_{i,j}^{LN}, T_{i+k_i,j+l_j}^{L(N+1)}$ — соседние кадры с номерами N и $N+1$ обрабатываемого уровня L,
k_i, l_j — значения вектора движения для подматрицы, к которой принадлежит усредняемый пиксель;

3) если к новому уровню относятся несколько суммированных кадров, то он принимается за очередной обрабатываемый уровень — кадры разбиваются на пары, производится вычисление векторов движение для пар и повторяются операции с пункта 2. Иначе полученный кадр является искомой оценкой видеоряда.

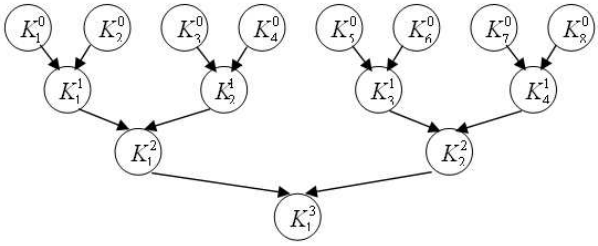


Рис. 4. Схема иерархической компенсации движения.

Замечание 3. Третий уровень иерархии заключается в иерархической компенсации движения, а объединение всех уровней представляет собой многомерную иерархическую структуру метода подавления шума, получившего название *супериерархического*.

У такой схемы компенсации движения отсутствуют недостатки, характерные для последовательной компенсации.

1) В обрабатываемых кадрах равного уровня одинаковое соотношение сигнал-шум, что не влияет на точность определения векторов движения.
2) Расстояние между движущимися объектами увеличивается только с возрастанием уровня обработки суммированных кадров и с меньшей в два раза скоростью. Учитывая равное отношение сигнал-шум в суммированных кадрах одного уровня, можно произвольно выбирать кадр из пары, значения которого будут компенсированы при учёте движения. Это позволяет строить более гибкие схемы иерархической компенсации движения, с меньшими расстояниями между движущимися объектами в сравниваемых кадрах.

Заключение

В работе представлен алгоритм, позволяющий увеличить отношение сигнал-шум в видеоряде с помощью прореживания его во времени, реализованный на основе комбинированного метода подавления шума.

Преимуществами предложенного метода супериерархической компенсации движения являются:

1) более точное вычисление векторов движения при их поиске на проекциях с меньшим уровнем шума;

2) уменьшение количества вычислений в алгоритме иерархического поиска векторов движения, объединяющем процедуру фильтрации и вычисление значений векторов;
3) гибкость схемы иерархической компенсации движения;
4) высокая степень подавления шумов.

В ходе экспериментов описанный алгоритм был апробирован на наборе тестовых видеопоследовательностей, полученных добавлением шума с различными параметрами распределения к «чистому» изображению. Изменение отношения сигнал-шум видеопоследовательности после обработки оценивалось на основе вычисления объективного критерия — пикового отношения сигнал-шум PSNR (peak signal-to-noise ratio) кадров обработанной и необработанной видеопоследовательности. Из результатов проведенных экспериментов сделаны следующие выводы:

1) предложенный метод увеличивает отношение сигнал-шум в видеоряде в не зависимости от закона распределения аддитивного шума в нём;
2) метод эффективно подавляет шумы различного уровня (значение PSNR тестовых видеопоследовательностей до обработки составляло от 14 до 28 дБ);
3) использование различных типов вейвлетов, различных уровней преобразования и порогов приводит к увеличению отношения сигнал-шум в сигнале (например, обработка 8 кадров видеопоследовательности с помощью ВП 2-го уровня и B-сплайна 2-2 увеличила оценку PSNR на 9 дБ).

Литература

[1] *Андреев В. П., Белов Д. А., Вайнштейн Г. Г., Москвина Е. А.* Эксперименты с машинным зрением. — Москва: Наука 1987.

[2] *Батлук А. В., Чобану М. К.* Исследование банков фильтров и применение лифтинг-схемы для декомпозиции изображений // Цифровая обработка сигналов, 2005. — №. 2.

[3] *Гришин С. В., Ватолин Д. С., Стрельников К. Н. и др.* Обзор блочных методов оценки движения в цифровых видеосигналах // Программные системы и инструменты. Тематический сборник. Москва: Изд-во факультета ВМиК МГУ, 2008 г.

[4] *Зубарев Ю. Б., Дворкович В. П., Нечепаев В. В., Соколов А. Ю.* Методы анализа и компенсации движения в динамических изображениях // Электросвязь, 1998. — №. 11.

[5] *Мал С.* Вейвлеты в обработке сигналов. — Москва: Мир, 2005.

[6] *Прокофьев А. В., Василенко С. И.* Применение пакетного вейвлет-преобразования при анализе изображений // Труды 15-ой Всероссийской научно-технической конференции «Современное телевидение». Москва: ФГУП МКБ Электрон, 2007.

[7] *Умняшкин С. В., Стрелков Ф. В., Жуков В. Г.* Трехшаговые алгоритмы поиска перемещенных блоков изображений // Информационные технологии и системы управления. Сб. научн. тр. под ред. В.А.Бархоткина. Москва: МИЭТ, 2000.

Аффинная версия алгоритма Лукаса-Канады*

Хашин С. И.
khash2@mail.ru
Иваново, Ивановский Университет

Предлагается алгоритм поиска межкадрового движения в виде аффинного преобразования общего вида при сжатии видеоинформации. Он является обобщением классического алгоритма Лукаса-Канады [1, 6], применяемого для нахождения междкадрового движения в виде сдвига. Предложен эффективный метод кодирования построенных преобразований.
Все рассмотренные алгоритмы реализованы и протестированы, язык реализиации — $C++$.

Одним из основных инструментов, используемых при сжатии видеоинформации, является приближённое построение области на следующем кадре в виде сдвига соответствующей области с предыдущего кадра. Эту схему используют все работающие в настоящее время алгоритмы — от $MPEG$-2 до $H.264$ [2–5]. Они различаются в этом пункте лишь размером обрабатываемой области (от 16×16 до 2×2), и методом интерполяции яркости точки между пикселами (билинейная, бикубическая и до би-пятой степени в $H.264$). Для нахождения таких сдвигов используется классический алгоритм Лукаса-Канады [1, 6]. Он отличается высокой скоростью работы и большой устойчивостью.

Однако, в разрабатываемых, перспективных алгоритмах видеосжатия (например, [8]) только сдвигов оказывается недостаточно. В основном это связано с тем, что рассматриваемая область на кадре (сегмент) может оказаться достаточно большой и её движение не может быть удовлетворительно описано с помощью сдвига. Поэтому приходится переходить к более сложным движениям на плоскости:

— сдвиг;
— сдвиг и поворот;
— сдвиг, поворот и растяжение;
— общее аффинное преобразование плоскости;
— общее проективное преобразование плоскости.

Алгоритму нахождения таких преобразований и посвящена настоящая статья.

Метод Лукаса-Канады

Опишем вначале, вкратце, классический метод Лукаса-Канады (см. например, [1]).

Пусть яркость первого кадра из пары задаётся в целочисленных точках таблицей $f[x,y]$. Её продолжение с помощью некоторого интерполяционного метода на всю плоскость и будем обозначать её $f(x,y)$. Для второго кадра из пары значения в целочисленных точках — $g[x,y]$, на всей плоскости — $g(x,y)$.

Пусть мы хотим аппроксимировать значения функции $g[x,y]$ в целочисленных точках некоторой области U, обычно это прямоугольник сравнительно небольшого размера, от 2×2 до 16×16. Для этого положим $g[x,y]\approx f(x-v_x,y-v_y)$, где (v_x,v_y) — некоторый вектор (*вектор скорости, вектор сдвига*), не обязательно целочисленный. Для оценки погрешности этого приближения рассмотрим сумму

$$S(v_x,v_y)=\sum_{(x,y)\in U}\left(g[x,y]-f(x-v_x,y-v_y)\right)^2.$$

Определение 1. *Вектором движения для пары кадров (f,g) в области U будем называть вектор $v=(v_x,v_y)$, для которого $S(v_x,v_y)$ минимальна.*

Замечание 1. Значения функции $g[x,y]$ берутся только в целочисленных точках, а функции $f(x,y)$ — в любых действительных. Для этого требуется некоторая интерполяционная формула (см., например, [7, 9]), в алгоритме Лукаса-Канады предполагается билинейная.

Замечание 2. Нахождение векторов скорости требуется не во время просмотра фильма, а во время кодирования. Поэтому временные требования к алгоритму не слишком жесткие — здесь не требуется обрабатывать по 24 кадра в секунду на стандартном процессоре.

В точке минимума должны выполняться соотношения:

$$\begin{aligned}\partial S/\partial v_x &= 2\sum\left(g[x,y]-f(P_v)\right)f'_x(P_v)=0\ ,\\ \partial S/\partial v_y &= 2\sum\left(g[x,y]-f(P_v)\right)f'_y(P_v)=0\ ,\end{aligned}\qquad(1)$$

где через P_v обозначена точка с координатами $(x-v_x,y-v_y)$.

Будем предполагать, что (v_x,v_y) — некоторое достаточно хорошее приближение к точному решению системы уравнений (1), которое будем искать в виде $(v_x+\Delta_x,v_y+\Delta_y)$. *Достаточно хорошее* означает, что для $f(x-(v_x+\Delta_x),y-(v_y+\Delta_y))$ можно использовать линейную аппроксимацию:

$$\begin{aligned}&f(x-(v_x+\Delta_x),y-(v_y+\Delta_y))\approx\\ &\approx f(x-v_x,y-v_y)-\Delta_x f'_x-\Delta_y f'_y\ .\end{aligned}\qquad(2)$$

С учётом этого, система уравнений (1) будет линейной системой размера 2×2 от переменных (Δ_x,Δ_y).

Работа выполнена при финансовой поддержке РФФИ, проект №11-07-00653.

Один шаг метода Лукаса-Канады состоит в решении этой системы уравнений и замене вектора скорости (v_x, v_y) на $(v_x + \Delta_x, v_y + \Delta_y)$.

Так как процедуру приходится выполнять много раз, то следует заранее наряду с матрицами $f[x,y]$, $g[x,y]$ приготовить матрицы частных производных $f'_x[x,y]$ и $f'_y[x,y]$.

Пирамидальная версия. В данном алгоритме важным является нахождение хорошего начального приближения для вектора скорости. Для этого обычно применяют *пирамидальную* версию алгоритма. Её идея заключается в том, что наряду с исходной парой изображений (f,g) рассматривают эти же изображения сжатые в два раза (f_2, g_2), в четыре (f_4, g_4) и т. д. (*пирамида*). Вектора скорости находят сначала на самом верхнем уровне пирамиды, а затем спускаются вниз этаж за этажом. На самом верхнем уровне в качестве начального приближения берут нулевой вектор. На нижних уровнях за начальное приближение берут удвоенную скорость, полученную на предыдущем шаге.

Все это вместе взятое обеспечивает хорошее сочетание скорости, точности и устойчивости алгоритма нахождения межкадрового движения в виде сдвигов.

Обобщение метода Лукаса-Канады

В используемых алгоритмах видеосжатия не используются движения плоскости более сложные, чем сдвиг. Это связано с тем, что обрабатываемые области малы и их движения достаточно хорошо описываются сдвигами. Но если переходить к алгоритмам видеосжатия следующего поколения [8], то придется обрабатывать и довольно большие сегменты. Например, межкадровое движение фона, занимающего большую часть кадра, часто можно описать одной-единственной формулой. В этом случае придется иметь дело с более сложными движениями:

— сдвиг (2 параметра):

$$\begin{aligned} x' &= a_1 + x, \\ y' &= a_2 + y; \end{aligned} \tag{3}$$

— сдвиг и поворот (3 параметра):

$$\begin{aligned} x' &= a_1 + \cos a_3 \cdot x + \sin a_3 \cdot y, \\ y' &= a_2 - \sin a_3 \cdot x + \cos a_3 \cdot y; \end{aligned} \tag{4}$$

— сдвиг, поворот и растяжение (4 параметра):

$$\begin{aligned} x' &= a_1 + a_4 (\cos a_3 \cdot x + \sin a_3 \cdot y), \\ y' &= a_2 + a_4 (-\sin a_3 \cdot x + \cos a_3 \cdot y); \end{aligned} \tag{5}$$

— общее аффинное преобразование плоскости (6 параметров):

$$\begin{aligned} x' &= a_1 + a_2 x + a_3 y, \\ y' &= a_4 + a_5 x + a_6 y; \end{aligned} \tag{6}$$

— общее проективное преобразование плоскости (8 параметров):

$$\begin{aligned} x' &= \frac{a_1 + a_2 x + a_3 y}{1 + a_7 x + a_8 y}, \\ y' &= \frac{a_4 + a_5 x + a_6 y}{1 + a_7 x + a_8 y}. \end{aligned} \tag{7}$$

В общем виде будем это записывать так:

$$\begin{aligned} x' &= A_x (x,y) \ , \\ y' &= A_y (x,y) \ , \end{aligned}$$

где функции A_x, A_y зависят, помимо x, y, ещё от некоторого набора параметров $(a_1, \ldots, a_k)$.

Каждое движение $A = (A_x, A_y)$ можно рассматривать и как более сложное: сдвиг — как сдвиг и поворот на нулевой угол, сдвиг и поворот — как сдвиг, поворот и растяжение с коэффициентом 1 и так далее.

Можно рассмотреть и обратную редукцию — более сложного движения к более простому, при заданной области на плоскости.

Определение 2. *Пусть A_1 — движение плоскости одного из типов от (4) до (7). Преобразование A_2 более простого типа будем называть редукцией A_1 на множестве U, если оно менее всего отклоняется от исходного движения A_1 на точках области U среди всех преобразований данного типа (в среднеквадратичном смысле).*

Задача нахождения редукции решается аналитически и мы будем спускаться от сложных движений к более простым, если это не ведет к потере точности.

По аналогии с обычным алгоритмом Лукаса-Канады, будем искать приближения для функции $g[x,y]$ в виде $g[x,y] \approx f(A_x(x,y), A_y(x,y))$. Рассмотрим сумму по точкам области U:

$$S(a_i) = \sum \left(g[x,y] - f(A_x(x,y), A_y(x,y)) \right)^2 . \tag{8}$$

Определение 3. *Оптимальным движением A для пары кадров (f,g) в области U будем называть движение плоскости одного из перечисленных выше типов, для которого величина $S(A) = S(a_i)$ минимальна среди всех движений данного типа.*

Для того, чтобы преобразование A задавало оптимальное движение области U, должны выполняться соотношения (аналог уравнений (1)):

$$\frac{\partial S}{\partial a_i} = 0 \ , \quad i = 1, \ldots, k, \tag{9}$$

После линеаризации, аналогичной (2), система (9) окажется линейной системой уравнений размера $k \times k$ относительно приращений $(\Delta_1, \ldots, \Delta_k)$ параметров преобразования $(a_1, \ldots, a_k)$.

Определение 4. *Базовым шагом обобщенного метода Лукаса-Канады будем называть решение системы уравнений (9) и замену параметров преобразования* $(a_1, \ldots, a_k)$ *на* $(a_1 + \Delta_1, \ldots, a_k + \Delta_k)$.

Замечание 3. Если движение плоскости ищется в виде (3), то обобщенный метод Лукаса-Канады в точности совпадает с классическим.

Замечание 4. На сегодняшний день движения в виде общего проективного преобразования (7) не реализовано. Возможно, в этом и нет необходимости: переход от общего аффинного преобразования к нему, скорее всего не даст заметного выигрыша. Но это пока лишь предположение, хотя и весьма вероятное.

Замечание 5. Рассмотренные алгоритмы довольно сложны с чисто математической точки зрения, являются итерационными, многошаговыми процессами. Поэтому для отладки были взяты искусственные пары кадров, в которых первый получается из второго путём заранее заданного аффинного преобразования; искомое движение в этом случае заранее точно известно.

Редукция исходной области U

Для решения задачи минимизации мы должны задаться некоторой областью U на втором кадре из пары. На самом деле, для работы алгоритма Лукаса-Канады вовсе не требуется, чтобы U было областью в геометрическом понимании — это просто произвольное множество точек со второго кадра.

Исходная область U может состоять из очень большого количества точек, вплоть до почти всего кадра, а это — до двух миллионов точек в случае `FULL-HD`-кадров (1920×1080). Поэтому для эффективного применения алгоритма количество точек в области надо сократить.

Определение 5. *Подмножество точек* U' *из* U *будем называть его редукцией, если оно содержит все те точки из* U*, координаты которых делятся на* k *для некоторого натурального* k.

Определение 6. *Подмножество точек* U'' *из* U *будем называть его равномерной редукцией, если оно получается следующим образом. Вначале в* U'' *кладем одну, случайно выбранную точку из* U*. Затем последовательно добавляем точку из* U*, находящуюся на расстоянии не менее* k *от всех уже выбранных точек для некоторого* k.

Оба множества, U' и U'', содержат порядка $|U|/k^2$ точек, поэтому, выбрав подходящее k, мы можем получать множества, мощность которых примерно равна любой наперед заданной.

На первый взгляд равномерная редукция кажется более надежным способом, хотя и существенно более сложным вычислительно. Однако на практике оказалось, что простая редукция, гораздо более быстрая, даёт вполне удовлетворительные результаты, мало отличающиеся от равномерной (в случае наших областей и с точки зрения наших целей). Поэтому в дальнейшем мы будем рассматривать только простую редукцию.

В разрабатываемом алгоритме используются две (простых) редукции U_1 и U_2 исходной области U. U_1 — это редукция до примерно 2000 точек. Если мощность исходной области U меньше 5000 точек, то просто полагаем $U_1 = U$. U_2 — это редукция до примерно 200 точек. Если мощность исходной области U меньше 500 точек, то полагаем $U_1 = U_2 = U$.

Область U_2 используется для нахождения начального приближения искомого движения плоскости (A_x, A_y), область U_1 — для последующего уточнения. Если оказывается, что $U_1 = U_2$, то мы получает только один этап, без последующего уточнения.

Общая схема алгоритма

Сначала рассмотрим работу алгоритма для фиксированной области $\tilde{U}$ на втором кадре из пары (f, g) (это может быть как U_1, так и U_2). Пусть A_0 — начальное приближение к оптимальному движению плоскости на этой области. A_0 может быть любого типа — от сдвига до общего аффинного преобразования.

Для каждого преобразования A у нас есть его численная оценка (погрешность): величина $S = S(A)$ из формулы (8), чем она меньше, тем лучше. В идеале она должна равняться нулю.

Рассматриваемый этап алгоритма состоит из следующих шагов.

1. Упрощение начального приближения. Исходное преобразование A_0 редуцируем к более простому типу (в смысле определения 2), до тех пор, пока его погрешность уменьшится не более, чем на C_1 процентов. Полученное преобразование обозначим A_1 (оно может и совпадать с A_0).

2. Базовые шаги. Начиная с преобразования A_1, выполняем базовые шаги (опр. 4) обобщённого метода Лукаса-Канады, пока не исчерпаем количество шагов (не более C_2) и уменьшение погрешности на одном базовом шаге составит не менее $C_3\%$. Найденное преобразование обозначим A_2, оно того же типа, что и A_1.

3. Усложнение типа преобразования. Если тип преобразования A_2 не самый сложный (не полное аффинное преобразование, проективные мы пока не рассматриваем вообще), то повышаем тип (от сдвига — к сдвигу и повороту, и т. д.) и с новым преобразованием повторяем базовые шаги оптими-

зации. Если погрешность уменьшилась менее, чем на C_1 процентов, то от такого усложнения отказываемся и заканчиваем алгоритм.

Замечание 6. В программе константы взяты следующие: $C_1 = 5\%,\ \ C_2 = 10,\ C_3 = 0.1\%$.

На основе большого количества численных экспериментов был сделан вывод, что пирамидальная схема является в данном случае неэффективной. Вместо неё был предложена редукция от заданной области U к паре областей (U_1, U_2), таких, что мощность U_1 порядка 2000 точек, а U_2 — порядка 200 (см. выше).

Сначала, начиная с тождественного преобразования (сдвиг на вектор $(0,0)$), находим оптимальное преобразование для области U_2. Затем, используя полученное преобразование в качестве начального приближения, находим оптимальное преобразование для области U_1.

Кодирование преобразований

Запоминать построенные аффинные преобразования путём хранения всех их коэффициентов, конечно же, неэффективно. Более эффективный способ можно предложить, вспомнив, что преобразование A применяется не на всей плоскости, а лишь в области U. Рассмотрим на плоскости точку P_0 — центр тяжести области U, округленный до целых, и обозначим через r среднеквадратичное расстояние от точек U до P_0, округленное до целых. Построим точку P_1 на расстоянии r от P_0 вправо и P_2 на расстоянии r от P_0 вниз. Любое аффинное преобразование A полностью определяется образами этих трёх точек $A(P_0)$, $A(P_1)$, $A(P_1)$. Введем следующие три пары чисел.

1. Пара (w_0, w_1) — вектор $A(P_0) - P_0$, т. е. вектор сдвига центра тяжести области U.

2. Пара (w_2, w_3) — разность векторов сдвига точек P_1 и P_0.

3. Третья пара. Пусть A' — преобразование типа (5): <<*сдвиг + поворот + растяжение*>>, переводящее P_0 в $A(P_0)$ и P_1 в $A(P_1)$. Такое всегда существует и единственно. В качестве третьей пары чисел (w_4, w_5) возьмём координаты вектора $A(P_2) - A'(P_2)$.

Теорема 1. *Числа $w_0, \ldots, w_5$ однозначно определяют преобразование A.*

Замечание 7. Если преобразование является сдвигом, то $w_2 = w_3 = w_4 = w_5 = 0$. Если преобразование близко к сдвигу, то $w_2, w_3, w_4, w_5 \approx 0$.

Замечание 8. Если преобразование имеет тип «*сдвиг + поворот + растяжение*», то $w_4 = w_5 = 0$. Если преобразование близко к таком типу, то $w_4, w_5 \approx 0$.

Построенное преобразование для области U будем запоминать не через его коэффициенты, а через шесть чисел $(w_0, \ldots, w_5)$, которые можно округлить до заданной точности $(1/16)$.

Полученные результаты

Построено обобщение классического алгоритма Лукаса-Канады для нахождения межкадрового движения не только в виде сдвигов, но и в виде более сложных движений плоскости, вплоть до аффинного преобразования общего вида.

Разработанные алгоритмы реализованы на языке $C{+}{+}$. Разработка велась так, чтобы можно было использовать не менее трёх различных компиляторов, чтобы не быть привязанным в будущем к конкретной системе программирования.

На основе большого количества численных экспериментов построена схема работы, обеспечивающая баланс между точностью, устойчивостью и объёмом вычислений.

Предложен эффективный метод хранения построенных преобразований.

Литература

[1] *Baker S., Gross R., Matthews I.* Lucas-Kanade 20 Years On: A Unifying Framework // Int. J. of Computer Vision, 2002. — Vol. 56. — Pp. 111–122.

[2] *ITU-T and ISO/IEC JTC 1* Generic coding of moving pictures and associated audio information. Part 2: Video // ITU-T Recommendation H.262 – ISO/IEC 13818-2 (MPEG-2), Nov. 1994.

[3] *ITU-T* Video coding for low bit rate communication // ITU-T Recommendation H.263; version 1, Nov. 1995; version 2, Jan. 1998; version 3, Nov. 2000.

[4] *ITU-T Rec. H.264 / ISO/IEC 11496-10.* Advanced Video Coding // Final Committee Draft, Document JVT-E022, September 2002.

[5] Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification // (ITU-TRec.H.264.ISO/IEC14496-10AVC) Joint Video Team (JVT), Mar. 2003, Doc. JVT-G050.

[6] *Lucas B. D., Kanade T.* An iterative image registration technique with an application to stereo vision // Proc. of Imaging Understanding Workshop. 1981. — Pp. 121–130.

[7] *Гонсалес Р., Вудс Р.* Цифровая обработка изображений // Москва: Техносфера, 2006. — 1072 с.

[8] *Хашин С. И.* Применение методов распознавания образов для сжатия видеоинформации // Докл. всеросс. конф. ММРО-13. — Москва: МАКС Пресс, 2007. — С. 420–424.

[9] *Яне Б.* Цифровая обработка изображений // Москва: Техносфера, 2007. — 583 с.

Построение кратнорегрессионных псевдоспектров для выделения и прослеживания объектов в системах видеонаблюдения*

Вишняков Б. В., Визильтер Ю. В., Выголов О. В.
boris.vishnyakov@gmail.com
Москва, ФГУП ГосНИИАС

В работе предлагается подход к выделению и прослеживанию движущихся объектов в системах видеонаблюдения, основанный на модели кратнорегрессионных псевдоспектров. Описывается процедура построения псевдоспектров. Демонстрируется, что методы анализа движения, основанные на использовании псевдоспектров, позволяют единым образом решать задачи как выделения объектов, движущихся с различными скоростями, вплоть до кратковременной остановки, так и исчезнувших/появившихся предметов в поле наблюдения. Показано, что данный метод обеспечивает повышенную помехоустойчивостью по сравнению с методами, основанными на анализе межкадровых разностей и оптических потоков. Представлены результаты работы алгоритма анализа движения на основе псевдоспектров для видеопоследовательностей из общедоступных тестовых баз PETS и ETISEO.

Введение

Задачи автоматического выделения и сопровождения объектов по признаку их движения на изображениях, получаемых от различных видео датчиков, часто возникают при разработке систем видеонаблюдения и систем машинного зрения, предназначенных для мобильных технических средств.

Основным недостатком традиционных систем видеонаблюдения является повышение загруженности оператора при разрастании системы. Это проблема особенно актуальна в случае использования больших систем уровня города. Очевидно, что для системы видеонаблюдения не достаточно лишь передавать изображения со всех камер операторам. Видеосигналы должны быть обработаны таким образом, чтобы при возникновении внештатных или чрезвычайных ситуаций на дисплее оператора появлялось сообщение, привлекающее его внимание.

Анализ движения является хорошо изученной областью машинного зрения. Наиболее распространенными подходами к выделению движущихся объектов на изображениях в настоящее время являются метод оптических потоков [1–4] и метод оценки движения по разностям изображений [2, 5, 6].

Метод оптических потоков обеспечивает достаточно эффективный и гибкий аппарат для анализа видимого движения объектов на цифровых видеопоследовательностях. В то же время, данный метод не лишен недостатков. Одной из наиболее существенных проблем, связанных с использованием техники анализа оптических потоков в практических приложениях, является высокая чувствительность данного метода к помехам. Как известно, «дифференцирующие» процедуры и фильтры всегда являлись наиболее чувствительными к шумам и помехам различного рода. Второй проблемой метода оптических потоков является невозможность прослеживания медленно движущихся или приостановившихся объектов.

Метод оценки движения по разностям изображений показывает высокую эффективность использования предложенного подхода в основном для выделения мелкоразмерных движущихся объектов, площадь которых существенно меньше площади анализируемого кадра [5]. Однако в большинстве систем видеонаблюдения типичной является ситуация, когда размер движущегося объекта заранее неизвестен, и, следовательно, система детектирования движения должна обеспечивать возможность адаптивной настройки алгоритмов на крупно-, средне- и мелкоразмерные объекты, в зависимости от задачи наблюдения, поля зрения и текущих настроек объектива камеры.

Предлагаемый алгоритм выделения и межкадрового прослеживания объектов на видеопоследовательностях с использованием кратнорегрессионных псевдоспектров, основанный на специальным образом модифицированном методе оценки движения по разностям изображений, обеспечивает как помехоустойчивость, возможность настройки алгоритма на выделение не только мелкоразмерных, но и средне-, и крупноразмерных объектов, так и прослеживания медленно движущихся и даже приостановившихся объектов.

Авторегрессионные псевдоспектры

Пусть $I(k)$ — входное изображение ширины w и высоты h на кадре с номером k, $I(k) \in \mathbb{R}^{w\times h}$. Предполагается, что $I(k)$ является полутоновым изображением, то есть $0 \leqslant I(k)_{x,y} \leqslant 255 \quad \forall x = 1,\ldots,w, \quad y = 1,\ldots,h$. Назовем $M_n(k)$ авторегрессионным аккумулятором n кадров с параметром α, вычисленным на кадре с номером k. Аккумулятор $M_n(k)$ является матрицей, $M_n(k) \in \mathbb{R}^{w\times h}$, и может быть вычислен по рекуррентной формуле [6]:

$$M_n(k+1) = \alpha M_n(k) + (1-\alpha)I(k). \qquad (1)$$

Работа выполнена при финансовой поддержке РФФИ, проекты № 11-08-01114 № 11-08-01039

Накопленная сумма в аккумуляторе $M_n(k)$ на кадре k будет иметь вид:

$$M_n(k) = (1-\alpha)\sum_{i=0}^{k-1}\alpha^{k-1-i}I(i). \qquad (2)$$

Пусть $l(k)$ будет произвольным элементом матрицы $I(k)$, а $m_n(k)$ — элементом аккумулятора $M_n(k)$ с такими же координатами ячейки, что и у элемента $l(k)$. Предположим, что на вход исходно нулевого авторегрессионного аккумулятора (1), начиная с некоторого момента k_0 (без потери общности можно положить $k_0 = 0$), в течение достаточно длительного времени N поступает сигнал интенсивности l:

$$m_n(k) = l(1-\alpha)\sum_{i=0}^{k-1}\alpha^{k-1-i} = l(1-\alpha^k). \qquad (3)$$

Определим такой параметр временного усреднения α, чтобы через n кадров сумма в ячейке аккумулятора $m_n(k)$ гарантировано принимала значение, равное доле β сигнала l:

$$m_n(n) = l(1-\alpha^n) = \beta l.$$

Отсюда получаем, что

$$\alpha_n = \sqrt[n]{1-\beta}. \qquad (4)$$

Таким образом, α_n есть такое значение параметра временного усреднения α, при котором через n кадров в аккумуляторе накопится сумма, равная $m_n(n) = \beta l$. В то же время значение n здесь можно назвать β-длиной памяти или просто длиной памяти аккумулятора с соответствующим параметром усреднения $\alpha_n = \sqrt[n]{1-\beta}$.

При значении коэффициента накопления, определённого в формуле (4), значение отклика аккумулятора (3) в текущий момент времени k определяется выражением

$$m_n(k) = l(1-\alpha_n^k) = l\left(1-(1-\beta)^{k/n}\right). \qquad (5)$$

Графики $m_n(k)$ для значений $\beta = 0.5$, $l = 100$, $k_0 = 10$ при различных значениях α_n, $n = 4, 8, 16, 32$ показаны на рисунке 1.

Таким образом, параметр временного усреднения α_n, определённый в формуле (4), фактически играет роль некоторого параметра насыщения функции отклика аккумулятора. Он позволяет судить о том, через какое время n накопленная сумма в аккумуляторе станет равной βl.

Согласно (5), α_n также обладает следующим свойством кратности:

$$\alpha_n = \alpha_{n\cdot s}^s. \qquad (6)$$

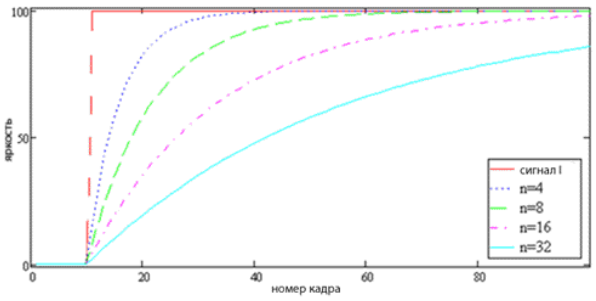


Рис. 1. Значение $m_n(k)$ для различных α_n.

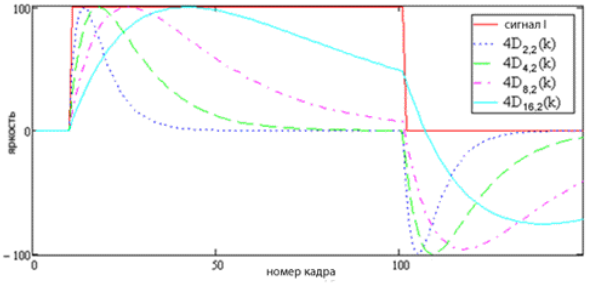


Рис. 2. Псевдоспектр: разности аккумуляторов.

На самом деле, $\alpha_{n\cdot s}^s = \left(\sqrt[n\cdot s]{1-\beta}\right)^s = \sqrt[n]{1-\beta} = \alpha_n$. Назовем

$$D_{n,s}(k) = m_n(k) - m_{n\cdot s}(k)$$

разностью между откликами аккумуляторов с кратными длинами памяти n and $n\cdot s$. Исходя из (6), а также предположения, что некоторый сигнал величины l подаётся с начального времени $k_0 = 0$, данная разность будет обладать следующим интересным свойством:

$$D_{n,s}(n\cdot s) = m_n(n\cdot s) - m_{n\cdot s}(n\cdot s) = \qquad (7)$$
$$= l\left(1-(1-\beta)^{\frac{n\cdot s}{n}}\right) - l\left(1-(1-\beta)^{\frac{n\cdot s}{n\cdot s}}\right) =$$
$$= l(1-\beta)\left(1-(1-\beta)^{s-1}\right) = l(1-\beta)\sum_{i=1}^{s-1}\beta^i.$$

Рассмотрим поведение разности откликов аккумуляторов $D_{n,s}(k)$ с кратной памятью. Пусть $s = 2$, а $\beta = 0.5$. Тогда, согласно формуле (7), разность значений аккумулятора с памятью длиной $2n$ кадров и аккумулятора с памятью длиной n кадров можно вычислить по формуле

$$D_{n,2}(2n) = 0.25l. \qquad (8)$$

На рисунке 2 показано поведение разности аккумуляторов $D_{n,s}(k)$ в зависимости от времени k при параметрах $s = 2$, $l = 100$.

Как видно, учетверённая разность кратных аккумуляторов $4\cdot D_{n,2}(k)$ является выпуклой функцией по k на отрезке действия сигнала. Максимум

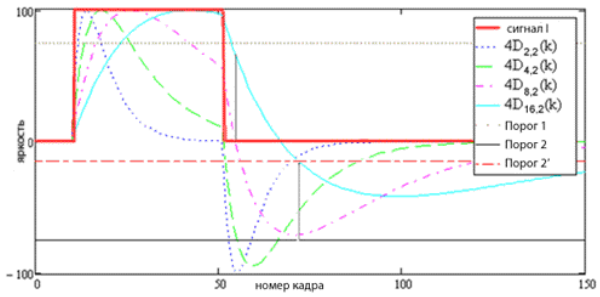


Рис. 3. Динамический порог яркости, определяемый по псевдоспектру.

данной функции, если он достигается, равен l. Соответствующий максимум разности отклика авторегрессионных аккумуляторов с кратной памятью достигается в точке $2n$.

Таким образом, поведение разности авторегрессионных аккумуляторов первого порядка с кратной длиной памяти напоминает спектральное разложение сигнала, точнее — его вейвлет-разложение. Назовем «кратнорегрессионным псевдоспектром» набор разностей откликов авторегрессионных аккумуляторов первого порядка (7) с кратными характеристиками длины памяти, задаваемыми последовательностью степеней двойки: 1, 2, 4, 8, ... (см. рисунок 2). Такой псевдоспектр позволяет качественно и количественно исследовать как продолжительность, так и амплитуду входного временного сигнала типа «меандр».

Если максимум разности откликов был последовательно достигнут для всех аккумуляторов с длиной памяти $n \leqslant N$, а для $n = N + 1$ предсказанный максимум сигнала достигнут не был, это значит, что постоянный входной сигнал имел длину $2N$ кадров, после чего начал убывать или был както иначе резко изменен.

Аналогичным образом можно сделать вывод и о величине сигнала. Поскольку $D_{n,2}(2n) = 0.25l$, то, следовательно, для всех n, для которых был достигнут максимум,

$$l = 4D_{n,2}(2n). \tag{9}$$

Ожидаемое значение максимума $D_{n,2}(k)$ можно легко найти, например, для $n = 1$ и далее сравнивать с ним значения разности аккумуляторов $D_{n,2}(k)$ для других значений n до тех пор, пока максимум в точке $k = 2m$ перестанет совпадать с максимумами в предыдущих точках $n < m$.

Рассмотрим теперь вопрос об определении порога чувствительности алгоритма детектирования яркостных изменений на изображениях. На рисунке 3 показана форма кратнорегрессионного псевдоспектра для случая менее продолжительного по времени изменения интенсивности. Как видно, при меньшей длительности сигнала низкочастотные составляющие псевдоспектра начинают движение в отрицательную сторону с более высоких исходных значений и вследствие этого достигают экстремума при значениях, меньших по модулю, чем порог, основанный на оценке (8). На рисунке 3 это хорошо видно на примере составляющей, представленной линией $D_{16,2}(k)$ (наиболее низкочастотная из представленных составляющих псевдоспектра). Однако эта проблема может быть решена, если совместно рассмотреть пару последовательных составляющих псевдоспектра.

Рассмотрим уровень $D_{8,2}(k)$ по отношению к уровню $D_{16,2}(k)$ псевдоспектра на рисунке 3. Поскольку реакция на изменения входного сигнала является существенно более быстрой, линия $D_{8,2}(k)$ гораздо раньше пересекает нулевую ось, реагируя на приход обратного фронта сигнала — его исчезновение. В этот момент значение текущей составляющей $D_{16,2}(k)$ ещё существенно больше нуля. Значение текущей составляющей псевдоспектра $D_{16,2}(k)$ в момент пересечения нуля предыдущей составляющей $D_{8,2}(k)$ предлагается для каждого пикселя запоминать и затем использовать для внесения динамической поправки в порог регистрации яркостных изменений.

Анализ предложенных кратнорегрессионных псевдоспектров является особенно полезным именно в случае анализа изображений с целью выявления движущихся объектов или вновь появившихся/пропавших предметов. Поскольку, с одной стороны, движение объекта относительно фона за счёт эффекта загораживания (замещения) яркости пикселей изображения порождает в каждом отдельном пикселе временной сигнал типа «меандр», имеющий ярко выраженные передний и задний фронты (яркостные перепады во времени). С другой стороны, возможность анализа сигнала с опорой на разности накопленных сумм с кратной памятью позволяет существенно экономить вычислительные ресурсы системы машинного зрения.

Результаты тестирования

Описанные алгоритмы тестировались на общедоступных базах видеороликов PETS [7] и ETISEO [8]. Пример реальной работы алгоритма можно увидеть на рисунке 4.

Кроме алгоритма было разработано и тестирующее приложение, которое позволяет сравнивать результаты работы алгоритма с эталонными результатами, полученными при разметке видеоролика человеком. Производительность измеряется в кадрах в секунду (FPS — frames per second). Оценка качества детектирования производится с помощью параметров «качества» детектирования (precision) и «надежности» детектирования (recall).

Название видео	Размеры кадра	Тип детектирования	Точность	Надежность	FPS
PETS-2001-SEQ1-CAM1	768x576	Движущиеся объекты	80% (8/10)	100% (8/8)	110
PETS-2001-SEQ1-CAM2	768x576	Движущиеся объекты	88% (8/9)	100% (8/8)	109
PETS-2006-S1-T1-C3	720x576	Движущиеся объекты	74% (26/35)	85% (24/28)	115
ETISEO-VS2-BE-19-C2	768x576	Движущиеся объекты	100% (4/4)	100% (4/4)	110
ETISEO-VS1-AP-5-C5	720x576	Движущиеся объекты	88% (8/9)	100% (6/6)	124
ETISEO-VS1-AP-5-C7	720x576	Движущиеся объекты	100% (9/9)	100% (7/7)	124
ETISEO-VS2-BC-17-C1	640x480	Оставленные предметы	66% (2/3)	100% (2/2)	142
ETISEO-VS1-BC-12-C1	640x480	Оставленные предметы	100% (1/1)	100% (1/1)	144

Таблица 1. Результаты тестирования алгоритма анализа движения.

Рис. 4. Скриншот из записи PETS-2001-SEQ1-CAM1. Объекты, находящиеся на этом кадре: движущийся человек (7), движущаяся машина (5) и паркующаяся машина (3).

Под «качеством» детектирования понимается процентное соотношение числа объектов, найденных алгоритмом из числа размеченных человеком, к общему числу объектов, найденных алгоритмом. Иными словами, «качество» есть 100% минус процент ложных срабатываний алгоритма. Под «надежностью» понимает процентное соотношение числа объектов, найденных алгоритмом, к числу всех объектов, размеченных человеком. Таким образом, «надежность» есть 100% минус процент объектов, которых алгоритм не смог обнаружить.

В таблице 1 содержатся результаты обработки некоторых известных видеороликов из баз PETS и ETISEO. FPS был подсчитан для персонального компьютера низкой производительности: процессор Intel Atom N270 1600 MHz, 1 Gb оперативной памяти.

Заключение

Проблема автоматического анализа видеосигнала с целью обнаружения и прослеживания движущихся объектов и оставленных или унесённых предметов является наиболее крупной в области анализа движения. В работе предложен новый подход к задаче анализа движения, основанный на формировании и обработке кратнорегрессионных псевдоспектров. Исследованы свойства псевдоспектров, позволяющие для входного яркостного сигнала типа «меандр» получать регулярным образом достоверные оценки его амплитуды и длительности. Представлены результаты тестирования алгоритма выделения движущихся объектов и исчезнувших или появившихся предметов, реализующего описанный подход, для видеороликов из общедоступных тестовых баз PETS и ETISEO.

Литература

[1] *Horn B. K. P., Schunck B. G.* Determining optical flow // Artificial Intelligence, 1981. — Vol. 17. — Pp. 185–203.

[2] *Barron J., Fleet D., Beauchemin S.* Performance of optical flow techniques // Internat. Jour. of Computer Vision, 1994. — Vol. 12, No. 1. — Pp. 43–77.

[3] *Anandan P.* A computational framework and an algorithm for the measurement of visual motion // Int. J. Comp. Vision, 1989. — Vol. 2. — Pp. 283–310.

[4] *Singh A.* Optic Flow Computation. A Unified Perspective // IEEE Computer Society Press, 1992. — Pp. 168–177.

[5] *Визильтер Ю. В., Лагутенков А. В., Ососков М. В., Выголов О. В., Блохинов Ю. Б.* Выделение и межкадровое прослеживание движущихся объектов при регистрации изображений сложных пространственных сцен произвольно движущимися двумерными сенсорами. // Вестник компьютерных и информационных технологий. — 2006. — № 3. — С. 34–38.

[6] *Вишняков Б. В., Визильтер Ю. В., Лагутенков А. В.* Использование модифицированного метода оптических потоков в задаче обнаружения и межкадрового прослеживания движущихся объектов // Вестник компьютерных и информационных технологий. — 2007. — № 5. — С. 3–8.

[7] *Pefomance Evaluation of Tracking and Surveillance (PETS) база данных видеозаписей* http://www.cvg.rdg.ac.uk/slides/pets.html

[8] *Video Understanding Evaluation project ETISEO база данных видеозаписей* http://www-sop.inria.fr/orion/ETISEO/

Формирование инвариантных признаков движущегося воздушного объекта*

Емельянов Г. М., Титов И. О.
Gennady.Emelyanov@novsu.ru
Великий Новгород, ГОУ ВПО «Новгородский государственный университет имени Ярослава Мудрого»

Рассматривается проблема создания автоматизированной системы для выделения и идентификации движущегося воздушного объекта. Исследуется вопрос классификации объекта, который чётко отделён от фона, не соприкасается и не перекрывается с другими объектами и представлен контуром. Результатом является набор характеристик, необходимых для реализации отображения «объект — номер класса».

Введение

Одной из серьёзных проблем систем компьютерного зрения является задача наблюдения за объектами при различных трансформациях. Чувствительность этих систем наблюдения к геометрическим искажениям объектов делает эту задачу технически достаточно сложной.

Первый этап в системе распознавания образов — это формирование признаков объекта, которые его уникально идентифицируют. Основным требованием к системе формирования признаков анализируемого объекта является требование их эффективности в процессе распознавания. Требования эффективности распознавания накладывает определённые ограничения на значения признаков, а именно: для объектов различных классов значения должны образовывать компактные области-кластеры в пространстве признаков. Кроме того, желательно чтобы указанные значения были устойчивы к ряду возможных искажений объекта.

В данной работе будут рассмотрены инвариантность системы к следующим изменениям:

— изменению местоположения объекта;
— изменению масштаба объекта;
— изменению ориентации объекта (к повороту объекта в плоскости изображения);
— определённым аффинным преобразованиям.

Для задачи распознавания контурного движущегося объекта нет необходимости рассчитывать характеристики, инвариантные к шумовым и динамическим искажениям, изменению яркости и контрастности. Достаточно остановиться только на инвариантах, причисленных выше.

Известны несколько методов для решения поставленной задачи, которые подробно рассмотрены в [1–3]. Так, метод Фурье-Меллина предназначен для сопоставления изображений в присутствии взаимного геометрического преобразования из группы подобия. Метод не является чувствительным к простым искажениям. Данный метод применим к оценке перспективных искажений, но при этом требует существенных ресурсных затрат.

Работа выполнена при финансовой поддержке РФФИ, проект №10-01-00146.

Другой метод классификации объекта основан на контуре объекта. Контур можно представить как одномерный сигнал, который позволяет проще устранять эффекты геометрических искажений. Метод дескрипторов Фурье реализуют требования к инвариантности, но имеет существенную вычислительную сложность, что усложняет его применение в задачах реального времени.

Метод, основанный на анализе моментов, используется на «поточечных» или растровых изображениях. Он позволяет рассчитывать инвариантные характеристики к любым геометрическим искажениям. Недостатком данного метода также является существенная ресурсоёмкость. Попытаемся применить метод моментов по контурному изображению, тем самым сократив его вычислительную сложность в несколько раз.

Признаки формы на основе анализа моментов

Рассмотрим изображение как функцию двух непрерывных аргументов $f(x, y)$. Момент порядка $p + q$ определяется как

$$m_{pq} = \int\limits_{\infty}^{\infty}\int\limits_{\infty}^{\infty} x^p y^q \, dxdy \qquad (1)$$

для $p, q = 0, 1, 2, \ldots$. Теорема единственности утверждает, что для любой кусочно-непрерывной функции $f(x, y)$, принимающей ненулевые значения только в конечной области плоскости xy, существует момент любого порядка, и последовательность моментов вида (1) однозначно определяется функцией $f(x, y)$. И наоборот, момент m_{pq} однозначно определяет функцию $f(x, y)$. Будем использовать центральные моменты, обладающие инвариантностью к сдвигу:

$$\mu_{pq} = \int\limits_{\infty}^{\infty}\int\limits_{\infty}^{\infty} (x - \bar{x})^p \, (y - \bar{y})^q \; dxdy,$$

где

$$\bar{x} = \frac{m_{10}}{m_{00}}, \; \bar{y} = \frac{m_{01}}{m_{00}}.$$

Здесь $\bar{x}$, $\bar{y}$ — координаты центра тяжести изображения.

Так как функция $f(x,y)$ описывает дискретное изображение, то равенство принимает вид

$$\mu_{pq} = \sum_x \sum_y (x-\bar{x})^p (y-\bar{y})^q f(x,y).$$

Основным достоинством моментных инвариантов

$$\mu_{00} = \sum_x \sum_y (x-\bar{x})^0 (y-\bar{y})^0 f(x,y) = m_{00},$$

$$\mu_{10} = \sum_x \sum_y (x-\bar{x})^1 (y-\bar{y})^0 f(x,y) = 0,$$

$$\mu_{01} = \sum_x \sum_y (x-\bar{x})^0 (y-\bar{y})^1 f(x,y) = 0,$$

$$\mu_{11} = \sum_x \sum_y (x-\bar{x})^1 (y-\bar{y})^1 f(x,y) = m_{11} - \bar{y}_{10},$$

$$\mu_{20} = \sum_x \sum_y (x-\bar{x})^2 (y-\bar{y})^0 f(x,y) = m_{21} - \bar{x}_{10},$$

$$\mu_{02} = \sum_x \sum_y (x-\bar{x})^0 (y-\bar{y})^2 f(x,y) = m_{02} - \bar{y}m_{01},$$

$$\mu_{21} = \sum_x \sum_y (x-\bar{x})^2 (y-\bar{y})^1 f(x,y) = m_{21} - 2\bar{x}m_{11} - \bar{y}m_{20} + 2\bar{x}^2 m_{01},$$

$$\mu_{12} = \sum_x \sum_y (x-\bar{x})^1 (y-\bar{y})^2 f(x,y) = m_{12} - 2\bar{y}m_{11} - \bar{x}m_{02} + 2\bar{y}^2 m_{10},$$

$$\mu_{30} = \sum_x \sum_y (x-\bar{x})^3 (y-\bar{y})^0 f(x,y) = m_{30} - 3\bar{x}m_{20} + 2\bar{x}m_{10},$$

$$\mu_{03} = \sum_x \sum_y (x-\bar{x})^0 (y-\bar{y})^3 f(x,y) = m_{03} - 3\bar{y}m_{02} + 2\bar{y}^2 m_{01}$$

является нечувствительность к поворотам изображения, что делает эффективным их применение в качестве признаков в задаче обнаружения и распознавания на изображении объектов неизвестной ориентации [1, 3, 4]. Данный набор использует момент только до порядка $p+q \leqslant 3$. Он обеспечивает полноту, то есть возможность построения других функционально независимых инвариантов с помощью моментов до заданного порядка. В то же время отсутствует функциональная избыточность, когда есть возможность выразить один из инвариантов как функцию других.

Нормированные центральные моменты определяются как $\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}}$, $\gamma = \frac{p+q}{2} + 1$, где $p+q = 2, 3, \ldots$.

На базе моментных инвариантов формируются признаки, устойчивые к преобразованиям подобия. Использование моментов второго и третьего порядков позволяет получить следующий набор из семи инвариантных моментов [1]:

$$\begin{aligned}
\varphi_1 =& \eta_{20} + \eta_{02},\\
\varphi_2 =& (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2,\\
\varphi_3 =& (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21}^2 - \eta_{03})^2,\\
\varphi_4 =& (\eta_{30} + \eta_{12})^2 + (\eta_{21} - \eta_{03})^2,\\
\varphi_5 =& (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12}) \times \\
& \times \left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] + \\
& + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) \times \\
& \times \left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right],\\
\varphi_6 =& (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + \\
& + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}),\\
\varphi_7 =& (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12}) \times \\
& \times \left[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] + \\
& + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) \times \\
& \times \left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right].
\end{aligned}$$

Указанный набор моментов является инвариантным по отношению к параллельному переносу, повороту и изменению масштаба. Дополнительно можно получить признаки, описывающие силуэт контура изображения за счёт аффинных преобразований [3, 5].

Аффинные преобразования можно рассматривать как декомпозицию следующих трансформаций: преобразование типа сдвига, пропорциональное масштабирование, искажение масштаба вдоль одной из осей координат, поворот и деформацию изображения, не описываемые преобразования подобия. С использованием моментов второго и третьего порядков получается набор из четырёх аффинных моментных инвариантов:

$$\begin{aligned}
i_1 =& \frac{1}{\mu_{00}^4}\left(\mu_{20}\mu_{02} - \mu_{11}^2\right),\\
i_2 =& \frac{1}{\mu_{00}^{10}}\left(\mu_{30}^2\mu_{03}^2 - 6\mu_{30}\mu_{21}\mu_{12}\mu_{03} + 4\mu_{30}\mu_{12}^3 + \right.\\
& \left. + 4\mu_{03}\mu_{21}^3 - 3\mu_{21}^2\mu_{12}^2\right),\\
i_3 =& \frac{1}{\mu_{00}^{10}}\left(\mu_{20}(\mu_{21}\mu_{03} - \mu_{12}^2) - \mu_{11}(\mu_{30}\mu_{03} - \mu_{21}\mu_{12}) + \right.\\
& \left. + \mu_{02}(\mu_{30}\mu_{12} - \mu_{21}^2)\right),\\
i_4 =& \frac{1}{\mu_{00}^7}\left(\mu_{20}^3\mu_{03}^2 - 6\mu_{20}^2\mu_{11}\mu_{12}\mu_{03} - 6\mu_{20}^2\mu_{02}\mu_{21}\mu_{03} + \right.\\
& + 6\mu_{20}\mu_{11}\mu_{02}\mu_{30}\mu_{03} - 18\mu_{20}\mu_{11}\mu_{02}\mu_{21}\mu_{12} + \\
& + 9\mu_{20}^2\mu_{11}\mu_{02}\mu_{30}\mu_{03} - 18\mu_{20}\mu_{11}\mu_{02}\mu_{21}\mu_{12} - \\
& - 8\mu_{11}^3\mu_{30}\mu_{03} - 6\mu_{20}\mu_{02}^2\mu_{30}\mu_{12} + \\
& + 9\mu_{20}\mu_{02}^2\mu_{21}^2 + 12\mu_{11}^2\mu_{02}\mu_{30}\mu_{12} - \\
& \left. - 6\mu_{11}\mu_{02}^2\mu_{30}\mu_{21} + \mu_{02}^3\mu_{30}^2\right).
\end{aligned}$$

Теория центральных моментов служит основой модели инвариантных признаков. На базе указанной модели разработан алгоритм выделения и идентификации движущегося воздушного объекта. Результаты работы алгоритма представлены далее.

Экспериментальная апробация

Исходное изображение, приведённое на рис. 1 слева, было подвергнуто геометрическим искажениям, включая поворот (рис. 2), уменьшение (рис. 1 (справа) и рис. 2), а также перспективное искажение (рис. 3).

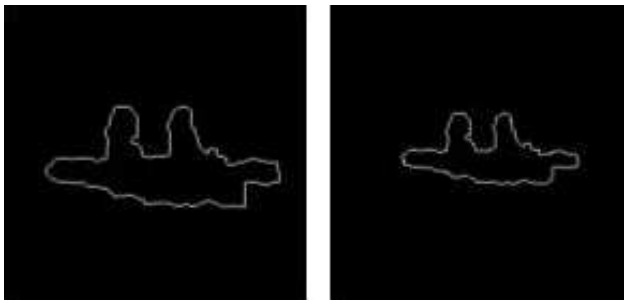


Рис. 1. Слева — исходное изображение (Фиг. 1); справа — уменьшенное в два раза (Фиг. 2).

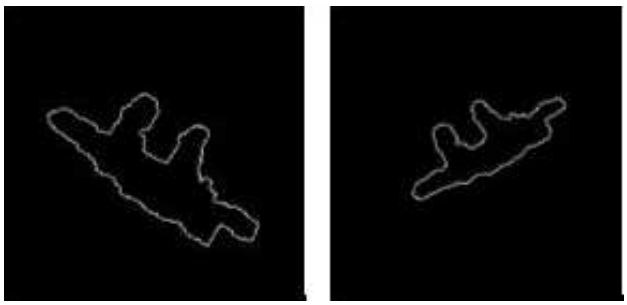


Рис. 2. Слева — повёрнутое на 45^0 (Фиг. 3); справа — уменьшенное в два раза и повёрнутое на 45^0 (Фиг. 4).

Рис. 3. Слева — перспективное искажение (Фиг. 5); справа — перспективное искажение (Фиг. 6).

Для каждого из изображений на рис. 1–4 рассчитывались инвариантные характеристики, представленные в таблицах 1 и 2. Приведённые в указанных таблицах результаты прологарифмированы с целью сужения динамического диапазона.

Рис. 4. Слева — другой контур объекта из той же видеопоследовательности (Фиг. 7); справа — ложный контур (Фиг. 8).

Таблица 1. Значения моментных инвариантных признаков для набора изображений.

	φ_1	φ_2	φ_3	φ_4	φ_5	φ_6	φ_7
Фиг. 1	−1.3	−3.4	−6.9	−11.3	20.5	13.1	−22.2
Фиг. 2	−1.3	−3.2	−6.9	10.8	19.7	12.5	20.7
Фиг. 3	−1.3	−3.4	−6.9	−11.3	20.4	13.0	−27.3
Фиг. 4	−1.3	−3.3	−7.0	−10.8	19.9	12.6	20.6
Фиг. 5	−1.3	−3.1	−7.8	−9.7	−18.6	−11.4	−19.9
Фиг. 6	−1.1	−2.6	−7.1	−8.0	−15.6	−9.3	−17.0
Фиг. 7	−1.5	−4.5	−7.3	−12.3	−22.1	−14.8	26.6
Фиг. 8	−1.7	−8.8	−13.1	−14.8	28.9	−20.5	−30.0

Таблица 2. Значения аффинных инвариантных признаков для набора изображений.

	i_1	i_2	i_3	i_4
Фиг. 1	−4.6	18.0	11.1	−13.6
Фиг. 2	−4.6	18.0	11.2	−13.7
Фиг. 3	−4.6	18.1	11.2	−13.6
Фиг. 4	−4.7	18.3	11.3	−13.9
Фиг. 5	−4.7	20.6	12.3	−14.7
Фиг. 6	−4.7	21.6	12.3	−14.7
Фиг. 7	−4.7	18.8	11.7	−14.3
Фиг. 8	−4.9	−30.7	17.9	−20.5

Выводы

Рассмотренный нами набор изображений включал в себя ложный контур помимо исходного изображения и его геометрически искаженных аналогов. Значения инвариантных признаков, рассчитанных для всего набора изображений, позволяют *сделать вывод*, что применением простейшего линейного классификатора ложный контур определяется достаточно просто и быстро, поскольку степень различия одноименных признаков весьма существенная.

В качестве *направления дальнейших исследований* следует отметить применение более сложных классификаторов для оценки принадлежности выделяемого контура к определённому классу на основе рассмотренных признаках формы.

Литература

[1] *Гонсалес Р., Вудс Р.* Цифровая обработка изображений Москва: Техносфера, 2005. — С. 957–961.

[2] *Яне Б.* Цифровая обработка изображений. Москва: Техносфера, 2007. — С. 515–517.

[3] *Венцель Е. С., Овчаров Л. А.* Теория вероятности и её инженерные приложения Москва: Высш. шк., 2000. — С. 115–128.

[4] *Гашников М. В., Глумов Н. И., Ильясова Н. Ю. и др.* Методы компьютерной обработки изображений Москва: Физматлит, 2001. — С. 601–634.

[5] *Heikkil J.* Pattern matching with affine moment descriptors // Elsevier Science. — 2004. № 3. — Pp. 23–50.

Динамическая модель повышения геометрической разрешающей способности системы регистрации изображений

Григорьева А. М., Пытьев Ю. П.
elf1ka@mail.ru, yuri.pytyev@gmail.com
Москва, физический факультет МГУ им. М. В. Ломоносова

Работа посвящена математическому моделированию системы регистрации изображений с подвижной матрицей сенсоров. Показано, что такая система регистрации позволяет получать разрешение существенно более высокое, чем разрешение, полученное системой регистрации, в которой сетчатка неподвижна.

Как известно, «острота» зрения человека, его геометрическая разрешающая способность, существенно выше, чем это следовало бы из оптической модели глаза и геометрической разрешающей способности сетчатки, обусловленной размером светочувствительных элементов и их плотностью на сетчатке. Этот феномен, по-видимому, обусловлен спецификой формирования зрительного восприятия, в процессе которого глазное яблоко испытывает неконтролируемые человеком движения, классифицируемые как «скачки», «тремор», «дрейф» и др., приводящие к соответствующим движениям изображения, сформированного оптической системой глаза, по его сетчатке. Более того, оптическое изображение, будучи зафиксированным относительно сетчатки глаза, не воспроизводится зрительной системой. Такой «динамический» принцип формирования зрительного восприятия позволяет существенно ослабить влияние гибели отдельных светочувствительных элементов сетчатки на качество изображения. Работа посвящена математическому моделированию «сверхразрешающей» способности зрительной системы человека, её нечувствительности к разбросу характеристик светочувствительных элементов сетчатки.

В докладе представлены результаты исследований следующих динамических моделей регистрации изображений:

Модель с одним подвижным светочувствительным элементом. Обозначим s светочувствительный элемент (сенсор), размер $l \times l$ которого определяет его геометрическую разрешающую способность. Выходной сигнал g сенсора s пропорционален энергии светового потока L с распределением интенсивности I, падающего на s, не содержит информации о вариациях интенсивности I в пределах сенсора s. На рис. 1 показан кусочно-постоянный выходной сигнал неподвижной матрицы сенсоров $m \times k$, состоящей из плотно упакованных сенсоров, таких как s. Эту кусочно-постоянную функцию g для краткости будем называть изображением. В первой модели динамического формирования изображения P — поле зрения размера $m \times k$, в пределах которого сенсор s перемещается (сканирует) с шагом l/n, задерживаясь в каждой точке (i, j), где $i = 1, \ldots, m,\ j = 1, \ldots, k$, на время t_{ij} в каждом акте регистрации его выходного сигнала g_{ij} (см рис. 2). Время нахождения сенсора s на поле зрения P фиксировано.

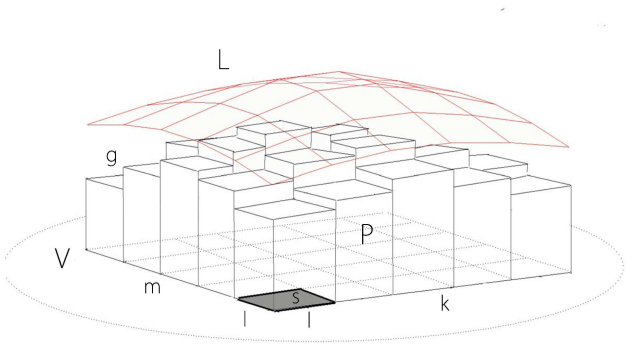


Рис. 1. V — область на плоскости, P — поле зрения, подобласть V, размера $m \times k$, s — сенсор размером $l \times l$, L — световой поток, g — изображение, полученное неподвижной сетчаткой, состоящей из плотно упакованных сенсоров s, покрывающей всё поле зрения P.

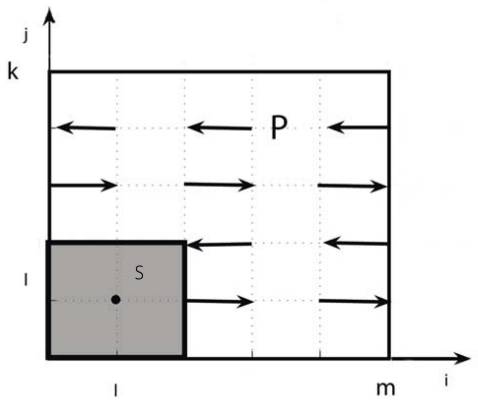


Рис. 2. Сканирование сенсором s поля зрения P с шагом l/n $(n = 2)$.

Так получаемый сигнал g методами математической редукции [3] преобразуем в изображение g', которое было бы получено с помощью неподвижной сетчатки, покрывающей всё поле зрения P и состоящей из плотно упакованных сенсоров s' размера $l/n \times l/n$, то есть сетчатки с более вы-

сокой геометрической разрешающей способностью. Таким образом будет восстановлена вариация интенсивности I потока L в пределах сенсора s до размеров сенсора s'.

Поскольку выходной сигнал g_{ij} сенсора в каждой точке $(i,j),\ i=1,\dots,m,\ j=1,\dots,k$, искажён шумом, следует определить оптимальное распределение времен $t_{ij},\ (i,j)\in\{1,\dots,m\}\times\{1,\dots,k\}$ нахождения сенсора в каждой точке поля зрения P, дающее наиболее точное отображение интенсивности светового потока L, то есть:

$$E\|(\widehat{g'}-g')\|^2 \sim \min_{t_{ij},(i,j)\in\{1,\dots,m\}\times\{1,\dots,k\}},$$

где $\widehat{g'}$ — результат редукции изображения g.

Метод динамической редукции проиллюстрирован в вычислительном эксперименте, выполненном по схеме:

$$\begin{pmatrix}\xi_1\\ \vdots\\ \xi_p\end{pmatrix}=\begin{pmatrix}A_1\\ \vdots\\ A_p\end{pmatrix}g'+\begin{pmatrix}\nu_1\\ \vdots\\ \nu_p\end{pmatrix},$$

где g' — случайная величина с корреляционным оператором F, моделирующая изображение, которое было бы получено с помощью неподвижной сетчатки, состоящей из сенсоров s', покрывающей всё поле зрения;
A_i — оператор, моделирующий отклик сенсора на световой поток L в каждом акте регистрации, $i=1,\dots,p$;
ν — нормально распределённый случайный вектор с $E\nu=0$ и дисперсией, равной 1, моделирующий аддитивный шум, каждая компонента которого соответствует i-ому акту регистрации;
ξ_i моделирует изображение, полученное как выходной сигнал сенсора в результате i-ого измерения, $i=1,\dots,p$.

Результаты получены методом рекуррентной редукции изображения g к изображению g'.

$$\widehat{g'_i}=\widehat{g'_{i-1}}+\frac{F_iA_i(\widehat{g'_i}-(A_i,\widehat{g'_{i-1}}))}{\sigma_i^2+A_i^*F_{i-1}A_i},$$

$$F_i=F_{i-1}-\frac{F_{i-1}A_i(F_{i-1}A_i)^*}{\sigma_i^2+A_i^*F_{i-1}A_i},$$

где $i=1,\dots,p$, $\widehat{g'_0}=0$, $F_0=F$, $\Sigma=\mathrm{diag}(\sigma_1^2,\dots,\sigma_p^2)$ — ковариационный оператор шума [3].

Модель с совокупностью жёстко связанных меду собой светочувствительных элементов. Процесс сканирования одним сенсором занимает достаточно много времени, чтобы сократить время регистрации в работе исследована динамическая сетчатка S, состоящая из жёстко связанных между собой светочувствительных элементов s, причём они не обязательно расположены вплотную к друг другу. Структура сетчатки определяется на основе анализа оптимального распределения времени пребывания сенсора s на поле зрения P модели с одним подвижным светочувствительным элементом, так как один акт регистрации сетчаткой S можно представить как последовательную регистрацию изображения g одним сенсором s.

Модель с областью повышенной точности. Пусть теперь на поле зрения P выделена область Q, на которой требуется восстановить изображение с более высокой точностью, при этом геометрическая разрешающая способность на всем поле зрения фиксирована, так как зафиксирован шаг сканирования l/n. Для этого нужно, чтобы на области Q, то есть области с меньшим шумом, сенсор s находился большее время. Зададим распределение времен пребывания светочувствительного элемента s в каждой точке поле зрения P, учитывающее это условие. Тогда, так как общее время нахождения сенсора s на поле зрения P фиксировано, распределение относительного времени пребывания в каждой точке поля зрения сенсором s можно моделировать распределением вероятности нахождения светочувствительного элемента s в каждой точке поля зрения. Реализовать такую случайную стратегию сканирования можно методом Монте-Карло.

Аналогичную задачу нужно поставить для сетчатки S, причём распределении сенсоров s на ней с учётом перемещения сетчатки должно реализовать это распределение вероятности, рассмотренное выше.

Выводы

Рассмотренные выше задачи были реализованы на примере одномерного поля зрения. Были получены следующие результаты.

1. Получено увеличение разрешения изображения в заданное количество раз.
2. Проиллюстрирована помехоустойчивость динамической системы регистрации изображений к выходу из строя светочувствительных элементов.
3. Разработана оптимальная стратегия сканирования, реализующая повышенную точность восстановления изображения в заданной области поля зрения P.
4. Наряду с вероятностной моделью была построена возможностная модель, максимально согласованная с вероятностной. Проведен их сравнительный анализ качества.

Аналогичные результаты для двумерной модели будут представлены на докладе.

Авторам известна работа [5], в которой рассматривается динамическая регистрация изображений, но используется для решения других проблем.

Литература

[1] *Ярбус А. Л.* Роль движений глаз в процессе зрения, Москва: Наука, 1965. — 167 с.

[2] *Демидов В. Е.* Как мы видим то, что видим, Москва: Знание, 1987. — 240 с.

[3] *Пытьев Ю. П.* Методы математического моделирования измерительно-вычислительных систем, Москва: ФИЗМАТЛИТ, 2011. — 400 с.

[4] *Пытьев Ю. П.* Возможность как альтернатива вероятности. Математические и эмпирические основы, применение. Москва: ФИЗМАТЛИТ, 2007. — 464 с.

[5] *Ben-Ezra M., Zomet A., Nayar S. K.* Video Super-Resolution Using Controlled Subpixel Detector Shifts // IEEE Transactions on Pattern Analysis and Machine Intelligence, Jun, 2005, Vol.27, No.6, pp.977-987.

Выделение радужки методом оптимизации кругового пути*

Матвеев И. А.
matveev@ccas.ru
Москва, Вычислительный центр им. А. А. Дородницына РАН

Для определения границ зрачка и радужки на изображении глаза используется алгоритм построения оптимального кругового пути (ОКП). Исследуются два способа применения алгоритма ОКП: определение границ зрачка и радужки с использованием информации о приближённом положении центра глаза и уточнение границ зрачка на основании известной окружности, приближающей зрачок. В первой задаче ОКП показывает худшую производительность по сравнению с аналогичными имеющимися методами, однако достаточно хорош в задаче уточнения уже выделенных границ. Алгоритм протестирован на базах данных изображений глаз, находящихся в открытом доступе, с общим числом изображений более 80000 для первого варианта применения и 16000 для второго.

Радужка на изображении представляется кольцом, заключённым между двумя округлыми приближённо концентрическими контурами: внутренней границей со зрачком и внешней границей со склерой. Обе границы приближаются окружностями с хорошей точностью, однако существуют приложения, в которых требуется более детальной описание границы [1]. В особенности это касается внутренней (зрачок–радужка) границы. Как правило, зрачок человека близок по форме к кругу, но в большинстве случаев не является идеальным кругом, а имеет нерегулярные отклонения с относительной величиной 5–10%. Задача определения форм, моделируемых кругами, окружностями, эллипсами (т. е. параметрически заданными регулярными фигурами) исследовалась очень подробно. Наиболее интересными в данной задаче представляются парные градиентные векторы [2], восстановление центров окружностей, проходящих через выделенные точки [3]. Эти, а также ряд других методов и модификаций описаны в обзоре [4]. Однако, методы, которые были бы пригодны для уточнения формы зрачка/радужки, прослеживая их округлую, но нерегулярную форму разработаны существенно хуже. Здесь можно выделить метод Оптимального Кругового Пути (ОКП, Circular Shortest Path, CSP) предложенный [5]. Этот метод выделяет округлые контуры и устойчив по отношению к разрывам границы, которые часто возникают на изображении радужки из-за бликов и отражений осветителя.

Рассматриваются для приложения метода ОКП: выделение контуров зрачка и радужки, при некоторой заданной точке, лежащей внутри зрачка (далее — задача детектирования) и уточнение границы зрачка, если есть приближающая её окружность (далее — задача уточнения).

Работа выполнена при финансовой поддержке РФФИ, проект № 09-01-00678.

Алгоритм оптимального кругового пути (ОКП)

Существует множество методов определения оптимального пути на изображении. Специфика ОКП заключается в том, что этот метод обнаруживает замкнутый контур, заключающий внутри себя некоторую заданную точку, которая предполагается его приближённым центром. Таким образом, в задаче детектирования метод ОКП начинает работу располагая координатами некоторой точки. В задаче уточнения начальные данные более полны: указывается окружность, приближающая контур, то есть заданы её центр и радиус. Поскольку контур проходит вокруг заданной точки, имеет смысл произвести полярное преобразование с полюсом в этой точке, что облегчает представление и дальнейшие вычисления. Полярное преобразование переводит кольцо в прямоугольник, его параметры можно подобрать так, чтобы его верхняя сторона соответствовала достаточно большой окружности, заключающей искомый контур, а нижняя сторона соответствовала достаточно малой окружности, находящейся целиком внутри контура. Радиальная координата полярной системы превращается в абсциссу прямоугольника, угловая координата — в ординату. Изображение из системы OXY переводится в систему $O\rho\varphi$, где также представляется дискретным прямоугольным растром размером $W * H$ пикселов. Далее будем называть этот растр *полярным представлением* (см. Рис. 1).

Таким образом, задача поиска ОКП превращается в задачу нахождения оптимального пути между левой и правой сторонами прямоугольника при условии, что терминальные точки пути имеют одинаковую ординату. Поскольку форма искомого контура близка к окружности и полюс преобразования лежит внутри контура, полярное представление контура однозначно относительно угла, и контур можно представить в виде функции $\rho(\varphi)$. Далее, предполагая, что искомый контур не проходит вблизи полюса преобразования, можно утверждать, что производная значения радиуса по углу ограничена: $d\rho/d\varphi < C_1$. Значение C_1 зависит

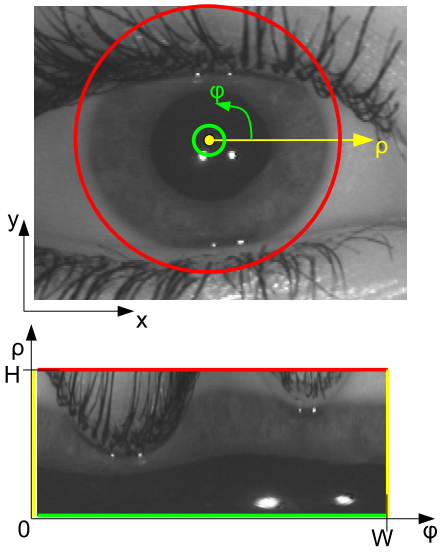


Рис. 1. Пример полярного преобразования радужки.

от того, насколько хорошо полюс преобразования соответствует центру контура и насколько форма самого контура близка к окружности. Задаваясь некоторой точностью определения центра глаза, можно выбрать параметры полярного преобразования так, что $C_1 \leqslant 1$. Итак, контур представляется как последовательность точек прямоугольного растра, в каждом столбце растра содержится ровно одна точка последовательности, точки из соседних столбцов лежат в одной или соседних строках, точки контура из первого и последнего столбцов также лежат в одной или соседних строках.

Введём стоимость перехода между точками (n, ρ') и $(n+1, \rho'')$ из соседних столбцов растра: $C((n, \rho'), (n+1, \rho'')) \equiv C_n(\rho', \rho'')$. Эта стоимость составлена из «внутренней» и «внешней» частей: $C = C^{(I)} + C^{(O)}$. Внутренняя часть определяется формой кривой и минимальна для прямых горизонтальных линий на растре полярного представления (соответствующих окружностям в исходном пространстве OXY):

$$C_n^{(I)}(\rho', \rho'') = \begin{cases} 0\ , & \rho' = \rho'', \\ T_1\ , & |\rho' - \rho''| = 1, \\ \infty\ , & otherwise. \end{cases}$$

Константа $T_1 > 0$ зависит от параметров полярного преобразования. «Внутренняя» часть стоимости определяется лишь формой контура и не зависит от характеристик изображения, «внешняя» часть — наоборот. «Внешняя» часть есть цена прохода через точку (n, ρ') растра, определяемая локальными свойствами изображения: $C_n^{(O)}(\rho', \rho'') = w((n, \rho'))$. Для заданного пути $S = \{\rho_n\}_{n=1}^{W}$ общая стоимость есть $C(S) = C((0, \rho_0), (W, \rho_W)) = \sum_{n=1}^{W} C_n(\rho_n, \rho_{n+1})$. Оптимальный контур — это последовательность, минимизирующая общую стоимость: $S^* = arg \min_S C(S)$.

Эта задача дискретной оптимизации может быть решена, например, жадным алгоритмом, как в [5]. Однако, для малых значений H ($H < 30$) предпочтительнее полный перебор, который совершается рекурсивно, как набор шагов, каждый из которых соответствует очередному столбцу растра (т. е. точкам с одним значением φ). Обозначим стоимость перехода из точки $(0, \rho')$, лежащей в первом столбце растра, в точку (n, ρ''), лежащую в текущем столбце: $C((0, \rho'), (n, \rho'')) \equiv C_{(n)}(\rho', \rho'')$. Поскольку ρ' и ρ'' меняются в пределах $[1, H]$, всего необходимо вычислить H^2 значений $C_{(n)}$. Они определяются рекурсивно, начиная с $C_{(1)}(\rho', \rho'') = 1/\delta(\rho', \rho'')$. Цена достижения точки в следующем столбце есть сумма цены достижения точки ρ''' в предыдущем и цены перехода между точками соседних столбцов:

$$C_{(n+1)}(\rho', \rho'') = \min_{\rho'''}(C_{(n)}(\rho', \rho''') + C_n(\rho''', \rho'')) =$$

$$= \min \left\{ \begin{array}{l} C_{(n)}(\rho'), \rho'') + w(n, \rho'')\ , \\ C_{(n)}(\rho'), \rho'' + 1) + w(n, \rho'' + 1) + T_1\ , \\ C_{(n)}(\rho'), \rho'' - 1) + w(n, \rho'' - 1) + T_1\ . \end{array} \right\}$$

На последнем шаге (под номером $W+1$) получаются H^2 значений $C_{(W+1)}(\rho', \rho'')$. При этом только значения с $\rho' = \rho''$ соответствуют замкнутым контурам. Поэтому стоимость оптимального замкнутого контура равна $\min_\rho C_{(W+1)}(\rho, \rho)$, он заканчивается (и начинается) в точке $\rho^*_{W+1} \equiv \rho^*_0 = arg \min_\rho C_{(W+1)}(\rho, \rho)$.

Рассмотрим «внешнюю» стоимость прохода через точку $C^{(O)}(\varphi, \rho) = w(\varphi, \rho)$. Из постановки задачи ясно, что функция $w(\varphi, \rho)$ должна быть построена так, чтобы быть малой в точках, соответствующих контуру и большой в других. В точках контура большое значение градиента яркости, поэтому точки с малым градиентом яркости должны отвергаться. Это делается условием $\|\mathbf{g}\| > T_2$, где $\mathbf{g}$ — вектор градиента яркости, а T_2 — пороговое значение, выбираемое таким образом, чтобы подавить ложные градиенты, возникающие из-за шума. Для фильтра Собеля можно принять $T_2 = 6\sqrt{2}\max\{\sigma, 2\}$, где σ — среднеквадратичная амплитуда шума.

Следующее специфическое свойство решаемой задачи состоит в том, что и зрачок, и радужка являются тёмными областями на светлом фоне, причём угол между градиентом яркости в точке контура и линией, соединяющей эту точку с полюсом преобразования, достаточно мал. Это условие можно записать как $\arccos(\frac{\mathbf{x}\cdot\mathbf{g}}{\|\mathbf{x}\|\,\|\mathbf{g}\|}) < T_3$. Значение порога T_3 зависит от качества определения приближённого центра, которое можно оценить как отно-

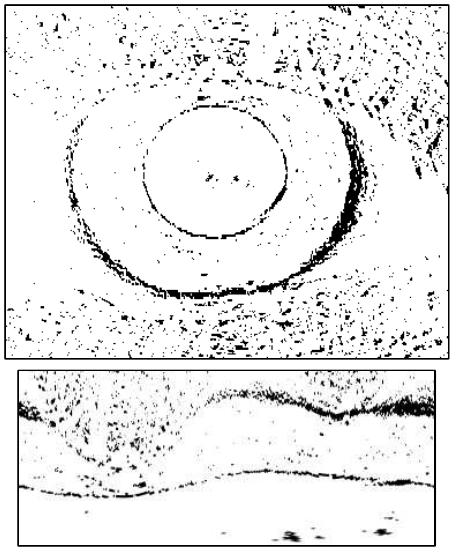


Рис. 2. Пример карты градиентов и полярное преобразование.

шение среднего расстояния D между приближённым и истинным центрами к среднему радиусу R контура: $T_3 = \arcsin(D/R)$. На Рис. 2 точки изображения с Рис. 1, со значениями градиента, удовлетворяющими обеим условиям, показаны чёрным цветом.

Для точек, удовлетворяющих обеим описанным условиям, цена перехода устанавливается в ноль, для всех остальных она равна T_1.

Примение ОКП в задаче детектирования

Алгоритм ОКП был использован для детектирования округлого контура при условии, что задана точка, лежащая внутри этого контура [5]. Для проверки работоспособности были использованы следующие открытые базы изображений радужки:

— UBIRIS.v1 [6];
— CASIA-IrisV3 [7];
— ND-IRIS [8].

Изображения были просмотрены человеком-экспертом, который выделил на каждом из них окружности зрачка и радужки. Параметры этих окружностей далее считались истинными и использовались для проверки метода. Далее обработка происходила автоматически. Приближённый центр глаза задаётся как случайная точка круга с центром в истинном центре глаза и радиусом, равным половине истинного радиуса зрачка, что моделирует ошибку в определении центра глаза (существующие методы определения центра имеют более высокую точность). Контуры (предположительно зрачки) определяются методом ОКП. Для сравнения с «истинным» зрачком для контура строился *эквивалентный круг*, такой, что его площадь (масса) и положение центра масс совпадали с центром масс области, окружённой контуром. Результаты сравнения «истинного» и эквивалентного кругов разбиты на пять непересекающихся категорий.

— «Хороший зрачок». Все параметры эквивалентного круга (x, y, r) не отличаются от соответствующих параметров «истинного» круга более чем на 5% (относительно радиуса «истинного» круга).
— «Плохой зрачок». Некоторые параметры нарушают ограничение в 5%, но все они удовлетворяют ограничению в 10%.
— «Хорошая радужка». Аналогично «хорошему зрачку», но для радужки.
— «Плохая радужка». Аналогично «плохому зрачку», но для радужки.
— «Не определено». Ни одно из перечисленных условий не выполнено. Найденный контур не соответствует истинным зрачку или радужке.

В Таблице 1 даны количества изображений по категории и базе данных. Следует отметить,

Таблица 1. Результаты определения контура по приближённому центру.

БД	UBI	ND-IRIS	CASIA
Общее число изобр.	1207	64980	16213
Хороший зрачок	139	61617	13494
Плохой зрачок	103	1421	922
Хорошая радужка	680	532	786
Плохая радужка	128	67	163
Не определено	157	1343	848

что при тестировании всегда детектировался контур, независимо от его качества. Возможно, наложение ограничения на минимально допустимое качество контура (и введение результата работы метода «контур не найден») уменьшило бы количество ошибок. Однако такие тесты не проводились.

Легко видеть, что результаты применения ОКП для непосредственного определения контуров (с использованием лишь приближённых координат центра глаза) хуже получаемых другими методами, например, перечисленными в [4], многие из которых имеют точность выше 99%. Также при использовании ОКП остаётся вопрос о том, как различать случай обнаружения контура зрачка и контура радужки.

Применение алгоритма ОКП в задаче уточнения

Вторая задача, для решения которой использовался метод ОКП, — уточнение границ контура. Это означает, что некоторый метод нашёл приближённую границу в виде окружности, и теперь необходимо найти точный вид контура, который не является идеальной окружностью. Следует от-

метить, что такая постановка имеет смысл только для контура зрачка. Радужка как правило, имеет регулярную эллиптическую форму, и в том случае, если она не затенена веками и/или ресницами, можно уточнить параметры эллипса методами, предназначенными для поиска эллиптических контуров. С другой стороны, как правило, радужка затенена и видимая её часть по форме сильно отличается и от эллипса, и от окружности. В этом случае при применении метода ОКП кратчайший путь будет прослеживается по богатой текстуре ресниц и век, что приводит к обнаружению ложных и бесполезных контуров. Поэтому здесь приводятся данные тестов лишь для уточнения границы зрачка. К сожалению, не существует простого способа непосредственно и объективно определить качество работы метода уточнения границ, поскольку данных, которые могли бы считаться истинными, для таких контуров не существует. Человек-оператор может вручную разметить лишь небольшое количество изображений, такая разметка достаточно сложна и в значительно большей степени подвержена подвержена ошибкам, чем разметка приближёнными окружностями. Для тестов с большими базами данных могут применяться лишь непрямые методы проверки качества

В качестве такого непрямого метода было избрано сравнение результатов распознавания. Тест выполнялся как набор следующих шагов. Из набора изображений были сформированы эталоны с использованием параметров зрачка, определённых сравниваемыми методами. Набор эталонов был сравнён сам с собой и по результатам вычислена EER. Полученная EER использовалась как характеристика качества метода определения параметров зрачка. Для тестов были использованы 16213 изображений 819 глаз 411 субъектов базы CASIA Iris-Lamp DB [7]. Три следующих метода определения параметров зрачка были сравнены.

— Окружности, размеченные человеком-оператором. Итоговое EER=0.752%;
— Контуры, уточнённые методом ОКП на основании размеченных оператором окружностей. Итоговое EER=0,981%;
— Окружности, полученные как эквивалентные уточнённым контурам. Хотя это тоже окружности, как и в первом пункте, они не всегда совпадают с размеченными человеком-оператором. Итоговое EER=0,390%.

Относительно высокое число ошибок для уточнённых контуров объясняется нестабильностью тонких деталей уточнённого контура в наборе изображений глаза одного человека, что приводит к локальным искажениям эталонов. Построение эквивалентной окружности усредняет эти вариации и приводит к построению более стабильных эталонов.

Заключение

Метод ОКП оказывается в целом неприменим для детектирования контуров зрачка и радужки на изображениях глаза, хотя на некоторых базах данных он даёт сравнительно хорошие результаты. Этот метод недостаточно надёжен даже в случае, если известна точная координата центра зрачка, поскольку на изображениях с низким контрастом границы зрачка или на изображениях с сильно деформированным зрачком он часто обнаруживает радужку вместо зрачка. С другой стороны, метод ОКП полезен в задаче уточнения формы и положения зрачка, уже приближенного окружностью. В этом случае метод работает в более узком кольце, содержащем лишь границу зрачка, в которое не попадают другие контуры, как то граница радужки или контактная линза. В этой задаче метод корректно обрабатывает большинство помех, а именно: блики от осветителя и тени от ресниц, попадающие на границу радужки, шум камеры, размытость изображения, частичное затенение веками, частичный выход за границы изображения. Более того, вычисления в узком кольце уменьшает перебор, что позволяет использовать этот относительно медленный метод в реальном времени.

Литература

[1] ISO, *ISO/IEC 19794-6:2005 Information technology – Biometric data interchange formats – Part 6: Iris image data*, 2005.

[2] *Rad A. A., Faez K., Qaragozlou N.* Fast circle detection using gradient pair vectors // Proc. VIIth Digital Image Computing: Techniques and Applications. Sydney, 2003.

[3] *Chen T.-C., Chung K.-L.* An efficient randomized algorithm for detecting circles // Computer Vision and Image Understanding, 2001. — Vol. 83. — Pp. 172–191.

[4] *Bowyer K. W., Hollingsworth K., Flynn P. J.* Image understanding for iris biometrics: a survey // Computer Vision and Image Understanding, 2008. — Vol. 110, No. 2. — Pp. 281–307.

[5] *Sun C., Pallottino S.* Circular shortest path in images // Pattern Recognition, 2003. — Vol. 36, No. 3. — Pp. 709–719.

[6] *Proenca H., Alexandre L. A.* UBIRIS: A noisy iris image database // 13th Int. Conf. on Image Analysis and Processing, Cagliari, Italy, 2005. — Vol. LNCS 3617. — Pp. 970–977.

[7] CASIA, Chinese academy of sciences institute of automation, 2005. — `http://www.cbsr.ia.ac.cn/IrisDatabase.htm`

[8] *Phillips P. J., Scruggs W. T., O'Toole A. J., et. al.* Frvt2006 and ice2006 large-scale experimental results // IEEE PAMI, 2010. — Vol. 32, No. 5. — Pp. 831–846.

Биометрическая система идентификации личности по радужной оболочке глаза*

Харитонов А. В., Потехин Е. Н., Леухин А. Н.
code@marstu.net
г. Йошкар-Ола, Марийский государственный технический университет

В работе рассмотрены известные алгоритмы распознавания радужной оболочки глаза, а также предложен метод распознавания, основанный на обнаружении характерных точек.

Один из механизмов обеспечения информационной безопасности — аутентификация личности. Для этих целей всё чаще используются биометрические параметры. В качестве отличительного, характерного лишь для одного человека, признака используют отпечатки пальцев, форму лица, сетчатку и радужную оболочку глаза, ДНК (самый надёжный и наиболее дорогой метод). При распознавании выделяются параметры объекта, уникальные в классе ему подобных. Они должны быть инварианты относительно условий регистрации и изменчивости самого объекта.

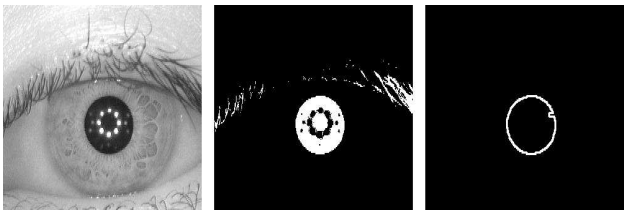
Устойчивым, высокоинформативным и выраженным биометрическим признаком является радужная оболочка глаза, благодаря тому, что имеет сложный рисунок и состоит из множества деталей. Рисунок радужки в большой степени случаен, а чем больше степень случайности, тем больше вероятность того, что конкретный рисунок будет уникальным. Математически данная случайность описывается степенью свободы. Исследования показали, что текстура радужки имеет степень свободы, равную 250, что гораздо больше степени свободы отпечатков пальцев (для сравнения — 35) и изображений лиц (20). Данные показатели наряду с практически неизменной структурой узора радужной оболочки глаза делают её надёжным и простым механизмом идентификации личности.

Алгоритмы

Весь процесс распознавания можно разделить на следующие этапы:

1. изображение, полученное с камеры, проходит процедуру сегментации, то есть определяются радиусы зрачка и радужной оболочки;
2. большинство алгоритмов нормализует изображение, приводя его из полярных координат к декартовым;
3. определяются признаки, по которым происходит дальнейшее сравнение с эталонами в базе.

На первом этапе могут возникнуть сложности, связанные с локализацией зрачка в связи с наложением ресниц и верхнего века. Решается данная задача путём приведения изображения глаза в бинарный вид с заданным порогом. Затем происходит выделение восьмисвязной границы и нахождение контуров.



а б в

Рис. 1. Этапы выделения зрачка.

Для формирования контура используется алгоритм Розенфельда, который состоит из нескольких шагов. На первом шаге необходимо просканировать сцену с целью отыскания верхнего левого граничного пикселя. На втором шаге необходимо проследить линию контура, согласно граничным точкам. На третьем шаге нужно сформировать код, которым будет описываться контур. В качестве такого кода используется код Фримена.

Для всех обнаруженных контуров определяется коэффициент формы, по которому определяется тот факт, является ли контур кругом или нет. Шаги повторяются до момента, пока коэффициент формы не прекратит улучшаться. Коэффициент формы для круга определяется отношением площади к длине; отсюда также можно найти приблизительный радиус зрачка. Затем используется преобразование Хафа для нахождения точного значения координат центра зрачка и его радиуса. На рисунке 1б приведён результат бинаризации с порогом 81; на рисунке 1в результат обнаружения контура зрачка. Поиск внешней границы радужной оболочки основан на обнаружении участка резкого перепада яркости.

Часто для дальнейшей работы производится перевод изображения радужки из полярных координат в декартовы. Однако есть методы, не требующие такого перевода.

После этого приведения к полученному изображению можно применить гауссов фильтр или медианную фильтрацию для устранения высокочастот-

Работа выполнена при финансовой поддержке гранта Президента РФ № МД-5418.2010.9, в рамках ФЦП «Научные и научно-педагогические кадры инновационной России» на 2009-2013 гг., ГК № 02.740.11.0838 и ГК № П 783, гранта РФФИ № 09-07-00072-а.

ного шума. После этого изображение все ещё слабоконтрастно, и для повышения надёжности производят выравнивание гистограммы. Следующий шаг может использовать различные методы: выделение признаков радужной оболочки на основе анализа локальной фазовой информации с помощью фильтров Габора; использование корреляционного анализа; выделение характерных точек.

В первом методе в качестве признаков используются грубо фильтрованные значения фаз фильтров Габора в определённых точках изображения. Корреляционный анализ является, с точки зрения теории, математически обоснованным, однако на практике приходится подбирать порог, при превышении которого можно считать, что два образа схожи.

Последний метод — выделение точек — основан на обнаружении характерных данному изображению точек, которые имеет некие признаки, существенно отличающие их от основной массы точек. Например, это могут быть резкие перепады освещённости. Предполагая, что ключевые точки присутствуют на образце всегда, можно поиск схожего изображения свести к поиску на сцене ключевых точек образца. А поскольку ключевые точки сильно отличаются от основной массы точек, то их число будет существенно меньше, чем общее число точек образца.

В целом, принцип выбора ключевых точек не важен. Главное, чтобы их было не слишком много и они присутствовали на изображении образца всегда.

В данной работе осуществляется поиск ключевых точек на основе вычисления детерминанта матрицы Гессе (гессиана), то есть определении перепадов яркости. Матрица Гессе для двумерной функции и её детерминант определяется следующим образом:

$$H(f(x,y)) = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix},$$

$$\det(H) = \frac{\partial^2 f}{\partial x^2}\frac{\partial^2 f}{\partial y^2} - \left(\frac{\partial^2 f}{\partial x \partial y}\right)^2.$$

Значение гессиана используется для нахождения локального минимума или максимума яркости изображения. В этих точках значение гессиана достигает экстремума.

Рассматриваются 100 точек с максимальным перепадом яркости.

Результат работы выделения ключевых точек представлен на рисунке 2.

Следует отметить, что используется бинаризованная аппроксимация лапласиана гауссиана, поскольку данный фильтр можно эффективно вычислить с помощью интегрального представления изображения.

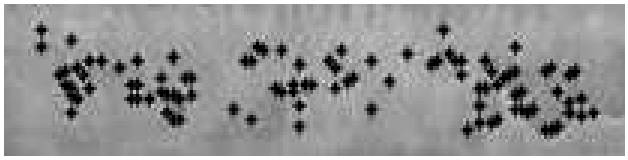


Рис. 2. Ключевые точки выделены чёрным цветом.

Таблица 1. Результаты распознавания

База данных	FAR (%)	FRR (%)
CASIA	0,26	0,86

Далее для каждой точки строится множество, в которое входят ближайшие соседи, расстояние не должно превышать заданного порога. Таким образом, отсеиваются некоторые точки и остаются множества, образующие группы точек. Характеристикой группы являются расстояния между каждым из её элементов.

Сравнение групп происходит как сравнение множеств; если порог совпадения превышен, считается, что идентификация пройдена. Метод инвариантен к линейному смещению и изменению яркости. Чтобы скомпенсировать возможные трансформации масштаба, длины между всеми элементами в группе нормируются.

Результаты распознавания представлены в таблице 1. Здесь FAR — ошибка второго рода. Это вероятность ложного доступа, когда система ошибочно опознает чужого как своего. Для многих систем данный параметр является наиболее критичным, поскольку область применения биометрических считывателей — объекты с повышенными требованиями по безопасности. А FRR (ошибка первого рода) — это вероятность ложного отказа в доступе клиенту, имеющему право доступа. Данная ошибка может появляться при повышении порога чувствительности (в системах, где данный параметр регулируемый) или при сильном повреждении идентификатора.

Выводы

Предложенный метод показывает хорошие результаты на базе данных CASIA. Сильное влияние оказывают включённые в область рассмотрения веки и ресницы, поэтому для устранения нежелательного эффекта, необходимо уточнить область век, ресниц и исключить их.

Методы распознавания во многом ориентированы на уменьшение размерности исходных данных. Сокращение пространства признаков улучшает кластеризацию образов. Распознавание в сокращённом пространстве признаков позволяет значительно уменьшить размер эталона, оставляя только

те признаки, которые имеют принципиальное значение для конкретного образа.

Литература

[1] *Daugman J.* New Methods in Iris Recognition // IEEE Transaction on Systems, Man, Cybernetics - part B, 2007. — Vol. 37, No. 5, Pp. 1167-1175.

[2] `http://www.cbsr.ia.ac.cn/IrisDatabase.htm` — База данных CASIA-IrisV3 — 2011.

[3] *Фурман Я. А., Кревецкий А. В., Передреев А. К., Роженцов А. А., Хафизов Р. Г., Егошина И. Л., Леухин А. Н.* Введение в контурный анализ; приложения к обработке изображений и сигналов. — Москва: ФИЗМАТЛИТ, 2003. — 588 с.

Локальная нормировка меры сходства и её влияние на характеристики биометрического распознавания лиц*

Визильтер Ю. В., Горбацевич В. С.
viz@gosniias.ru, gvs@gosniias.ru
Москва, ФГУП «Государственный научно-исследовательский институт авиационных систем»

Предложен новый класс локальных процедур персональной нормировки биометрических мер сходства, основанный на оценке свойств исходной меры сходства в малой окрестности выбранного эталона. Проведено экспериментальное исследование влияния локальной нормировки на основные биометрические характеристики верификации и идентификации. Показано, что локальная нормировка демонстрирует в целом значительно лучшие эффекты изменения целевых характеристик биометрического распознавания при существенно меньших вычислительных затратах по сравнению с глобальной Z-нормировкой.

Введение

Согласно международным и российским стандартам [1], процесс биометрической идентификации состоит из комбинации вызовов двух базисных функций: создания биометрического шаблона из биометрического образца и сравнения биометрических шаблонов. В силу этого биометрический классификатор представляет собой «черный ящик», на вход которого подаются хранимый *эталон* и предъявляемый *тест*, а на выходе формируется единственное число: *мера сходства*.

Казалось бы, свойства биометрического классификатора как «распознающего чёрного ящика» полностью определены производителями, и повлиять на них уже нельзя. Однако на конкретной базе эталонов поведение классификатора может быть улучшено за счёт построения *персональных мер сходства* (*client-specific, user-specific* или *model-specific score*) для каждого эталона. Говорят также о процедурах *персональной нормировки меры сходства* (*client-specific score normalization procedure*). В работе [2] содержится обзор известных методов персональной нормировки мер сходства, основанных на рассмотрении распределений меры сходства в своих и чужих сравнениях. Такую нормировку мы будем далее называть *глобальной*, так как она учитывает свойства всей базы в целом.

В данной работе предложен новый класс *локальных процедур персональной нормировки*, основанных на оценке свойств исходной меры сходства в малой окрестности выбранного эталона.

Основные характеристики качества биометрического распознавания

В задаче биометрической *верификации* (сравнения $1:1$) решение о соответствии теста A эталону принимается, если мера сходства $\lambda(A, B)$ превышает порог t. При этом качество характеризуется двумя распределениями: $f_{\text{gen}}(\lambda)$ в «своих» сравнениях, и $f_{\text{imp}}(\lambda)$ в «чужих» сравнениях (сравнениях с образами других людей). Ошибкой 1-го рода (FRR) является неправильная идентификация в «своих» сравнениях, 2-го рода (FAR) — ложная идентификация в «чужих» сравнениях. При фиксированном t вероятности ошибок имеют вид:

$$\text{FRR}(t) = \int\limits_{-\infty}^{t} f_{\text{gen}}(x)dx; \qquad (1)$$

$$\text{FAR}(t) = \int\limits_{-\infty}^{t} f_{\text{imp}}(x)dx$$

При изменении t пара значений $\langle \text{FAR}(t), \text{FRR}(t) \rangle$ пробегает в плоскости FAR-FRR так называемую *ROC* или *DET кривую*, которая полностью описывает поведение биометрического классификатора в режиме верификации.

В задаче идентификации (сравнения $1:N$) тесту A ищется соответствие в базе эталонов $\mathbf{B} = B_i$, $i = 1, \ldots, N$. Решение об идентификации принимается по правилу *ближайшего соседа*:

$$c_\lambda(A, \mathbf{B}) = \arg\max_{i=1,\ldots,N} \lambda(A, B_i). \qquad (2)$$

Часто рассматриваются также задачи «нахождения n наиболее похожих» или «построения списка n лучших кандидатов» (сравнение $1:n:N$). Будем называть эту задачу обобщённой или *n-идентификацией*. Она решается путём нахождения n ближайших соседей для A в базе $\mathbf{B}$, причём $n \ll N$. Качество решения определяется вероятностью попадания «своего» кандидата в первые n кандидатов. Основной характеристикой качества идентификации является вероятность попадания «своего» кандидата в первые n кандидатов по убыванию λ. Поскольку качество идентификации явным образом зависит от размера базы N, ось n иногда нормируют как n/N и говорят о попадании «в первые x% базы».

Персонализация мер сходства путём глобальной и локальной нормировки

Пространство биометрических шаблонов неоднородно, поэтому, как было показано в [4], одни

Работа выполнена при финансовой поддержке РФФИ, проекты № 11-08-01114-а, 11-08-01039-а.

эталоны имеют больше шансов быть верно распознанными, чем другие (т. н. «Doddington's zoo effect» [2]). Используется два основных подхода для учёта этого эффекта: выбор *персональных порогов* и использование *персональных мер сходства*. Первый подход развивается в работах [5–7]. При этом персональный порог может быть функцией от глобального порога [8]. Примерами реализации второго подхода являются такие процедуры, как Z-нормировка [3], EER-нормировка [9] и F-нормировка [10]. В [2] показано, что в большинстве случаев методы персонального выбора порога могут быть сведены к методам персональной нормировки, поэтому рассмотрим лишь методы нормировки.

Пусть имеется база эталонов $\mathbf{B} = B_i, \quad i = 1, \ldots, N$. Под персональной нормировкой меры $\lambda(A, \mathbf{B})$ на базе $\mathbf{B}$ понимается переход к набору мер

$$\mu_i(A,\, B_i) = \mu(\lambda(A,\, B_i),\, \{\lambda(A\,, B_j)\}_{j\neq i}). \qquad (3)$$

Здесь μ — закон нормировки. В рамках метода максимального правдоподобия если для $\lambda_i(A,\, B_i$ известны распределения в «своих» и «чужих» сравнениях, то статистика

$$\mu_i(A,\, B_i) = \ln f_{\text{gen}}(\lambda_i) - \ln f_{\text{imp}}(\lambda_i), \qquad (4)$$

согласно лемме Неймана-Пирсона, обеспечивает минимальные ошибки 1-го и 2-го рода. Предположим, что распределения нормальные: $f_{\text{gen}}(\lambda_i) = Norm(\lambda_i^G,\, \sigma_i^G), \quad f_{\text{imp}}(\lambda_i) = Norm(\lambda_i^I,\, \sigma_i^I)$ Тогда из (4) после простой нормировки следует оценка

$$\mu_i^{MS-LLR}(A, B_i) = \frac{\lambda_i - \lambda_i^I}{\sigma_i^I} - \frac{\lambda_i - \lambda_i^G}{2\sigma_i^I}. \qquad (5)$$

Способ *MS-LLR-нормировки* (5) был предложен в [2], и он является наиболее предпочтительным в случае, когда известны распределения «своих» сравнений. К сожалению, обычно в базе имеется лишь по одному «своему» эталону, поэтому $f_{\text{gen}}(\lambda_i)$ неизвестно. Таким образом, в (5) остаётся лишь первая составляющая:

$$\mu_i^Z(A, B_i) = \frac{\lambda_i - \lambda_i^I}{\sigma_i^I} \qquad (6)$$

Способ глобальной нормировки (6) называется *Z-нормировкой*.

В данной работе предлагается механизм локальной нормировки мер сходства, основанный на иных принципах. Исходным толчком к разработке этого подхода послужило следующее наблюдение. Пусть имеется база $\mathbf{B}_{N-1}$, состоящая из $N-1$ эталона, и пара изображений лица одной персоны, ещё не занесённой в базу: B_N^1 и B_N^2. Образуем два варианта базы из N эталонов: $\mathbf{B}_N^1 = \{\mathbf{B}_{N-1},\, B_N^1\}$ и $\mathbf{B}_N^2 = \{\mathbf{B}_{N-1},\, B_N^2\}$. Предъявим оставшееся изображение той же персоны в качестве теста. Поскольку процедура идентификации (2) неидеальна, наряду с правильными ответами, могут встретиться такие случаи, когда

$$c_\lambda(B_N^2,\, \mathbf{B}_N^1) \neq N, \quad c_\lambda(B_N^1,\, \mathbf{B}_N^2) = N. \qquad (7)$$

Выражение (7) означает, что с точки зрения меры сходства λ для образа B_N^2 в базе $\mathbf{B}_{N-1}$ *существует такой ближайший сосед* $\text{Neighbor}(B_N^2, \mathbf{B}_{N-1}) \in \mathbf{B}_{N-1}$, для которого

$$\lambda(\text{Neighbor}(B_N^2,\, \mathbf{B}_{N-1}),\, B_N^2) > \lambda(B_N^1,\, B_N^2), \qquad (8)$$

но при этом для образа B_N^1 в базе $\mathbf{B}_{N-1}$ нет более близкого образа, чем B_N^2:

$$\lambda(\text{Neighbor}(B_N^1,\, \mathbf{B}_{N-1}),\, B_N^1) > \lambda(B_N^1,\, B_N^2). \qquad (9)$$

Значит, согласно (8) и (9), N-ю персону в базе $\mathbf{B}_N^2$ распознать *в принципе легче*, поскольку ближайший «чужой» сосед к N-му эталону в ней расположен дальше, чем ближайший «чужой» сосед в базе $\mathbf{B}_N^1$. Этот факт можно оценить заранее, до предъявления теста: *чем ближе к данному эталону находится ближайший «чужой» сосед, тем больше вероятность его перепутывания со «своим» тестом*. Этот локальный принцип «конкуренции с ближайшим соседом» приводит к следующему простейшему правилу $L(1)$-нормировки:

$$\mu_i^{L(1)}(A, B_i) = \frac{\lambda(A,\, B_i)}{\lambda(\text{Neighbor}(B_i,\, \mathbf{B}),\, B_i)}. \qquad (10)$$

Заметим, что способ локальной нормировки (10) можно рассматривать как разновидность «конкурентной меры сходства» [11], основанной на сравнении расстояния (монотонно убывающего с ростом меры сходства) от эталона до теста и от него же до ближайшего «чужого» соседа в базе.

Общая идея локальной перенормировки меры сходства заключается в том, чтобы *изменять масштаб нормировки меры сходства в зависимости от плотности «чужих» эталонов в небольшой локальной окрестности вокруг каждого эталона в биометрической базе*.

Понятие локальной окрестности эталона в базе определим в терминах соседства: k-окрестностью эталона B_i в базе $\mathbf{B}$ назовем множество $\mathbf{B}^k(B_i, \mathbf{B}) \subseteq \mathbf{B}$, включающее k ближайших соседей эталона B_i в базе $\mathbf{B}$. В качестве оценки средней плотности «чужих» эталонов в k-окрестности эталона B_i используем *среднее значение меры сходства* эталона B_i с его соседями из $\mathbf{B}^k(B_i,\, \mathbf{B})$:

$$\lambda^{(k)}(B_i,\, \mathbf{B}) = \sum_{B\in\mathbf{B}^k(B_i,\, \mathbf{B})} \lambda(B,\, B_i)/k. \qquad (11)$$

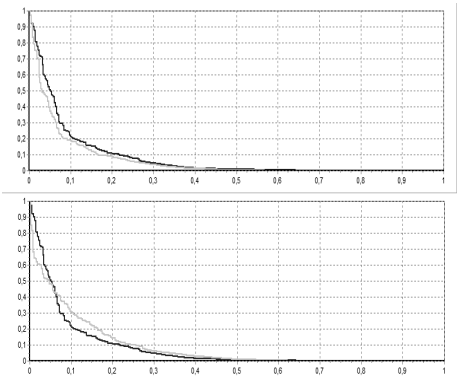


Рис. 1. *DET*-кривые для $L(20)$- и Z-нормировки *CGN*.

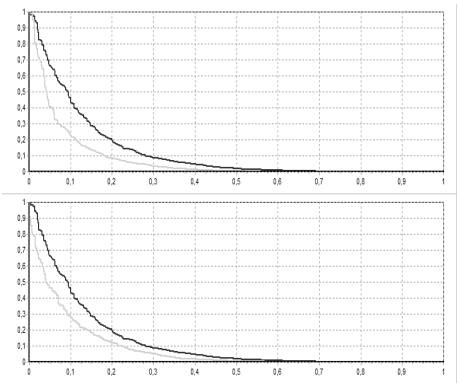


Рис. 2. *DET*-кривые для $L(20)$- и Z-нормировки *IIT-1*.

При помощи (11) можно определить общее правило $L(k)$-нормировки:

$$\mu_i^{L(k)}(A, B_i) = \lambda(A,\, B_i)/\lambda^{(k)}(B_i,\, \mathbf{B}). \qquad (12)$$

Заметим, что (10) — частный случай (12) при $k = 1$.

Экспериментальное исследование эффектов персональной нормировки

Экспериментальное исследование влияния $L(k)$-нормировки (12) на основные биометрические характеристики было проведено для трёх готовых систем биометрического распознавания лиц: *FaceVACS-DBScan, Ver.4.4, 2010* (далее *CGN*) фирмы *Cognitec* (Германия), *IIT Face Indexing* ЗАО «Институт Информационных Технологий» (Россия) версий 2009 г. (далее *IIT-1*) и 2011 г. (далее *IIT-2*).

В качестве тестовой базы использовалась база фронтальных изображений лиц низкого качества (низкое разрешение, наличие различных артефактов и т. п.). Тестовая база включала 1505 разномоментных изображений лиц 699 персон.

Результаты предложенной $L(k)$-нормировки (12) сравнивались с результатами Z-нормировки (6).

Исследование влияния нормировки меры сходства на характеристики верификации. Рис. 294, 295, 296 демонстрируют влияние локальной и глобальной нормировок на DET-кривые трёх различных биометрических классификаторов (тёмным цветом указаны характеристики для исходной меры λ, светлым — для нормированной меры μ). В тестах данной серии использовалось значение $k = 20$. $L(20)$-нормировка демонстрирует небольшое монотонное улучшение характеристик FAR-FRR для алгоритма *CGN* и существенное монотонное улучшение для *IIT-1*. Для классификатора *IIT-2* $L(20)$-нормировка практически не изменяет соотношения FAR-FRR, то есть не даёт ни положительного, ни отрицательного эффекта. В то же время, Z-нормировка демонстрирует монотонное улучшение характеристик FAR-FRR лишь для наиболее слабого алгоритма *IIT-1*. Для алгоритма *CGN* Z-нормировка ведет себя немонотонно — улучшает характеристики в левой части графика и ухудшает в правой. Для алгоритма *IIT-2* Z-нормировка монотонно и существенно ухудшает характеристики верификации по сравнению с исходной мерой сходства.

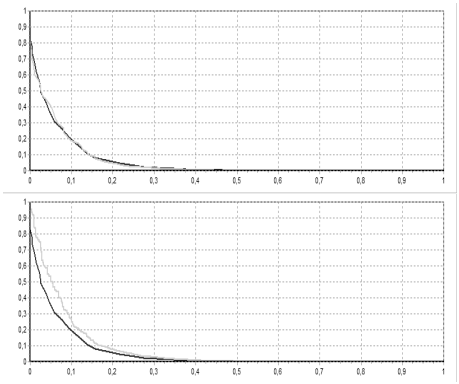


Рис. 3. *DET*-кривые для $L(20)$- и Z-нормировки *IIT-2*.

Исследование влияния нормировки меры сходства на характеристики идентификации. Таблицы 1, 2, 3 демонстрируют влияние $L(k)$-нормировки и Z-нормировки мер сходства *CGN*, *IIT-1* и *IIT-2* на вероятности попадания «своего» кандидата в первые $n = 1, 5, 10, 20$ кандидатов. Кроме того, исследуется влияние изменения параметра k. На основании экспериментов можно сделать следующие основные выводы:

1) $L(k)$-нормировка во всех случаях приводит к статистически значимому улучшению характеристик n-идентификации, особенно это касается 1-идентификации. Важно отметить, что значимое улучшение идентификации наблюдается даже для случая алгоритма *IIT-2*, для которого улучшение верификации не наблюдалось (рис. 296).

2) Лучшие результаты для данной базы изображений лиц были получены при $L(k)$-нормировке с параметром $k = 20$. Однако даже $L(1)$-нормировка демонстрирует определённый эффект улучшения характеристик идентификации (для *IIT-1* этот эффект является весьма значимым).

Таблица 1.

	P1	P5	P10	P20
CGN	**0,3**	**0,45**	**0,52**	**0,6**
Z	**0,32**	**0,47**	**0,52**	**0,63**
$L(1)$	0,33	0,47	0,55	0,61
$L(10)$	0,33	0,47	0,55	0,63
L(20)	**0,34**	**0,47**	**0,54**	**0,65**
$L(40)$	0,33	0,47	0,53	0,65
$L(200)$	0,31	0,47	0,51	0,63
$L(500)$	0,3	0,46	0,52	0,63

Таблица 2.

	P1	P5	P10	P20
IIT-1	**0,4**	**0,52**	**0,57**	**0,62**
Z	**0,49**	**0,63**	**0,7**	**0,77**
$L(1)$	0,5	0,62	0,69	0,75
$L(10)$	0,51	0,66	0,71	0,78
L(20)	**0,52**	**0,66**	**0,72**	**0,78**
$L(40)$	0,52	0,65	0,71	0,77
$L(200)$	0,51	0,64	0,71	0,78
$L(500)$	0,5	0,64	0,71	0,78

Таблица 3.

	P1	P5	P10	P20
IIT-2	**0,43**	**0,58**	**0,65**	**0,7**
Z	**0,48**	**0,6**	**0,65**	**0,71**
$L(1)$	0,45	0,6	0,65	0,71
$L(10)$	0,47	0,6	0,66	0,71
L(20)	**0,48**	**0,6**	**0,66**	**0,71**
$L(40)$	0,48	0,6	0,65	0,72
$L(200)$	0,47	0,6	0,65	0,71
$L(500)$	0,47	0,6	0,65	0,71

3) Z-нормировка также демонстрирует монотонное улучшение характеристик идентификации (хотя верификация при Z-нормировке может даже ухудшаться). Эффект Z-нормировки, как правило, превышает эффект $L(1)$-нормировки, но не превышает эффекта $L(20)$-нормировки.

Заключение

Предложен новый класс локальных процедур персональной нормировки биометрических мер сходства. По результатам проведённых экспериментов локальная нормировка демонстрирует в целом значительно лучшие эффекты изменения целевых характеристик при существенно меньших вычислительных затратах.

Различие наблюдаемых эффектов глобальной Z- и локальной $L(k)$-нормировки, по-видимому, свидетельствует об их принципиально различных механизмах. Эффект глобальной нормировки возникает при рассмотрении «чужих» образов как единого кластера с гауссовым распределением. Локальная нормировка использует локальные вариации плотности образов, возникающих из-за сходства типов лиц, либо сходства условий регистрации их изображений.

Литература

[1] *ГОСТ Р ИСО/МЭК 19784-1-2007* Автоматическая идентификация. Идентификация биометрическая. Биометрический программный интерфейс. Часть 1. Спецификация биометрического программного интерфейса

[2] *Poh N., Kittler J.* Incorporating Variation of Model-specific Score Distribution in Speaker Verification Systems // Speech and Language Processing, 2008. — Vol. 19, No. 3. — Pp. 594–606.

[3] *Auckenthaler R., Carey M., Lloyd-Thomas H.* Score Normalization for Text-Independant Speaker Verification Systems // Digital Signal Processing, 2000. — Vol. 10, No. 1-3. — Pp. 42–54.

[4] *Doddington G., Liggett W., Martin A., Przybocki M., Reynolds D.* Sheep, Goats, Lambs and Woves: A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker ReCGNnition Evaluation // Int'l Conf. Spoken Language Processing (ICSLP), Sydney, 1998.

[5] *Furui S.* Cepstral Analysis for Automatic Speaker Verification // IEEE Trans. Acoustic, Speech and Audio Processing / IEEE Trans. on Signal Processing, 1981. — Vol. 29, No. 2. — Pp. 254–272.

[6] *Chen K.* Towards Better Making a Decision in Speaker Verification // Pattern Recognition, 2003. — Vol. 36, No. 2. — Pp. 329–346.

[7] *Saeta J. R., Hernando J.* On the Use of Score Pruning in Speaker Verification for Speaker Dependent Threshold Estimation // The Speaker and Language Recognition Workshop (Odyssey), Toledo, 2004. Pp. 215–218.

[8] *Lindberg J., Koolwaaij J. W., Hutter H.-P., Genoud D., Blomberg M., Pierrot J.-B., Bimbot F.* Techniques for a priori Decision Threshold Estimation in Speaker Verification // Proc. of the Workshop Reconnaissance du Locuteur et ses Applications Commerciales et Criminalistiques (RLA2C), Avignon, 1998. Pp. 89–92.

[9] *Fierrez-Aguilar J., Ortega-Garcia J., Gonzalez-Rodriguez J.* Techniques for a priori Decision Threshold Estimation in Speaker Verification // LNCS 3072, Int'l Conf. on Biometric Authentication (ICBA), Hong Kong, 2004. Pp. 498–504.

[10] *Poh N., Bengio S.* F-ratio Client-Dependent Normalisation on Biometric Authentication Tasks // IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, 2005. Pp. 721–724.

[11] *Борисова И. А., Дюбанов В. В., Загоруйко Н. Г., Кутненко О. А* Сходство и компактность // ММРО-14, г. Суздаль. Москва: МАКС Пресс, 2009. — С. 89–92.

Алгоритм извлечения бинарного вектора из изображений отпечатков пальцев*

Ушмаев О. С.[1]***, Гудков В. Ю.***[2]***, Кузнецов В. В.***[1]
oushmaev@ipiran.ru, diana@sonda.ru, k.v.net@rambler.ru
[1] Москва, Институт проблем информатики РАН, [2] Миасс, Челябинский государственный университет, филиал

В статье рассмотрена задача, относящаяся к проблеме совмещения биометрической идентификации по отпечаткам пальцев и криптографических конструкций. Основой такого совмещения является алгоритм извлечения достаточно длинной устойчивой бинарной строки из изображения отпечатка пальца. Традиционно в распознавании отпечатков пальцев используются сложные графовые представления. Развертывание таких графов в строку является достаточно сложной задачей. В статье предлагается алгоритм извлечения бинарной строки из отпечатка пальцев на основе топологического представления папиллярных линий. При прослеживании папиллярной линии встречаются «события»: разветвления и окончания этой и соседних папиллярных линий, которые вводят частичные бинарные отношения между различными точками на изображении отпечатка пальцев. Индексация этих отношений позволяет описывать окрестность любой точки изображения бинарным векторов длиной 40–100 бит. Выбирая несколько окрестностей, мы получаем более длинные вектора. Эксперименты показывают, что в топологических векторах для различных предъявлений одного отпечатка пальца в среднем наблюдается 20% ошибок. Такое количество ошибок позволяет использовать корректирующие коды для извлечения побитно устойчивых строк, что может служить основой для реализации схем защищённой биометрической идентификации и биометрической криптографии на основе отпечатков пальцев.

Большинство систем биометрической идентификации личности (т. е. идентификации по физическим или поведенческим характеристикам человека) базируются на стандартизованной схеме реализации, которая изложена в стандарте ISO 19784 (bioAPI) и в гармонизированном с ним российским стандарте ГОСТ 19784. Согласно этим стандартам, процесс биометрической идентификации личности не зависит от используемой биометрической характеристики человека. Идентификация состоит из комбинации вызовов двух базисных функций: создания биометрического шаблона из биометрического образца (изображения или сигнала) и сравнения биометрических шаблонов. Функция создания шаблона по сути является извлечением информативных признаков. Функция сравнения биометрических шаблонов возвращает одно число: меру сходства. На основе сходства двух биометрических шаблонов принимается решение о принадлежности соответствующих биометрических образцов одному человеку.

На практике биометрический шаблон полностью раскрывает личность человека, что сильно затрудняет применение биометрической идентификации в информационных системах. Даже если шаблон хранится отдельно от фамильно-именной группы и прочей информации, нарушитель может выяснить личность путём использования этого шаблона для обращения к другим биометрическим системам (например, криминалистическим учётам) или даже для восстановления исходной биометрии как изображения или сигнала. Эта проблема становится действительно уязвимым местом биометрических систем по мере распространении технологий биометрической идентификации. При утере или компрометации биометрического шаблона полностью отсутствуют возможности его перевыдачи (подобно смене пароля). Стандартные методы защиты, например, хэш-функции, также не работают, так как биометрический шаблон является побитно неустойчивым. Как следствие, прямое применение шифрования и хэш-функций к биометрическим данным делает собственно идентификацию невозможной.

Для обеспечения возможности внедрения биометрической идентификации в информационные системы требуется разработка специализированных конструкций, совмещающих биометрию и криптографию. Как показывают научные результаты, основой для таких конструкций является преобразование биометрических данных в бинарный вектор. Есть успешные примеры преобразования в бинарный вектор биометрических данных, которые традиционно представляются вектором признаков: радужной оболочки глаза, голоса, рукописного почерка. Актуальным является эффективное решение этой задачи для биометрии отпечатка пальцев. Обычно отпечаток пальцев представляется достаточно сложным графом с внутренней структурой, целостность которой легко может быть нарушена. Поэтому, несмотря на достаточно высокую энтропию и индивидуальность отпечатков пальцев, пока не создано конструкций, которые позволяли бы извлекать из него достаточно длинный побитно устойчивый бинарный вектор с приемлемой ошибкой первого рода.

Работа выполнена при поддержке гранта МД-72.2011.9 Президента Российской Федерации для государственной поддержки молодых российских учёных-докторов наук.

В статье предложен алгоритм, который позволяет существенно продвинуться в задаче извлечения побитно точного вектора из отпечатка пальцев. Предложен алгоритм извлечения бинарного вектора (с возможными ошибками в битах), предложена схема реализации корректирующих кодов, исследована статистика ошибок в битах. Проведённые эксперименты показывают, что при наблюдаемом характере ошибок можно эффективно использовать корректирующие коды.

Топологическая модель изображения отпечатка пальца

Большинство систем распознавания отпечатков пальцев используют в качестве информативных признаков контрольные точки: разветвления и окончания папиллярных линий (рис. 1). При сравнении отпечатков анализируется взаимное расположение контрольных точек и их атрибуты (направление потока папиллярных линий, тип). Однако, с точки зрения задачи извлечения бинарной строки такие признаки не являются удобными, так как они неинварианты к естественным искажениям изображений отпечатков пальцев: движению, деформациям, отсутствию фрагментов изображения.

Мы предлагаем использовать не сами контрольные точки, а их отношения. Для этой цели введем понятия топологического события и топологической связи [6]. Для произвольно выбранной точки скелета отпечатка пальца проведем линию, перпендикулярную в каждой точке потоку папиллярных линий, на глубину d линий. Эта линия разделяет каждую пересекаемую папиллярную линию на два луча. Выберем луч (далее *топологическая связь*) и будем прослеживать его до встречи с контрольной точкой или проекцией контрольной точки с соседней папиллярной линии. Всего возможно 14 типов топологических событий [4, 1], их схематичное изображение дано на рис. 2. Пример прослеживания топологии и пример схематического представления локальной топологии представлены на рис. 3 и рис. 4. Как видно из схемы, топологические связи могут быть пронумерованы с точностью до выбора изначального паправления (вправо или влево от исходной точки скелета). Если рассматривать топологическую схему только в точках ветвлений или окончаний, то изначальное направление может быть выбрано естественным образом как совпадающее с направлением ветвления или окончания. Таким образом, для каждой контрольной точки мы получаем бинарное описание длины $16d + 4$. Несмотря на то, что событий всего 14, при том, что четыре бита дают возможность кодирования 16 событий, приведённая на рис. 2 индексация топологических событий даёт строки с равновероятным распределением нулей и единиц.

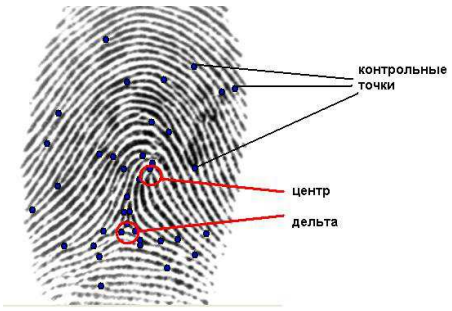


Рис. 1. Пример схожих кластеров на различных отпечатках пальцев.

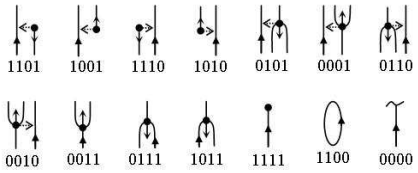


Рис. 2. Индексы топологических событий.

Рис. 3. Пример топологической окрестности.

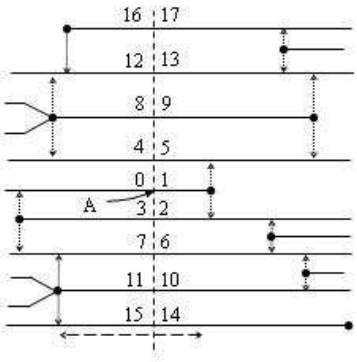


Рис. 4. Пример топологической схемы окрестности.

Алгоритм извлечения бинарного вектора из отпечатка пальцев

Представленное в предыдущем разделе описание контрольных точек потенциально позволяет извлекать достаточно длинный бинарный вектор (при разумных глубинах прослеживания от 3 до 6 линий длина описания контрольной точки варьируется от 52 бит до 100 бит). На изображениях отпечатках пальцах можно в среднем найти от 30 до

60 контрольных точек [3]. Основной проблемой при построении бинарного вектора является упорядочивание контрольных точек.

Для решения этой проблемы предлагается следующий алгоритм. На этапе регистрации отпечатка пальцев выберем l контрольных точек, заданных своими координатами $\{p_i\}_{i=1}^{l}$, и запишем их в хранимый шаблон пользователя открытым образом. При верификации личности или инициализации криптографического протокола мы извлекаем из предъявленного отпечатка пальца контрольные точки $\{q_i\}_{i=1}^{k}$. Далее находим среди них подмножество, наиболее близкое к $\{p_i\}_{i=1}^{l}$, например, по среднеквадратической ошибке:

$$\sum_{i=1}^{l} (q_{j_i} - p_i)^2 .$$

Таким образом, мы получили порядок на некотором подмножестве контрольных точек. Далее мы извлекаем для каждой из найденных точек топологический вектор. Итоговый вектор для предъявленного отпечатка вычисляется как конкатенация топологических векторов для точек q_{j_i}.

Шаблон применения

В дальнейшем планируется использовать полученный алгоритм преобразования отпечатка пальца в бинарную строку длины s для реализации защищённой биометрической идентификации. Для этого предполагается использовать следующую схему, близкую к [2].

На этапе регистрации пользователь получает случайный ключ k. Ключ преобразуется корректирующими кодами в строку $K = Code(k)$ длины s. Далее из предъявленного при регистрации отпечатка случайным образом выбирается подмножество контрольных точек $\{p_i\}_{i=1}^{l}$, из топологических векторов которых строится бинарный вектор $A = D(p_1)||\ldots||D(p_l)$, где $D(p_i)$ — топологический вектор точки p_i. Далее вычисляется защищённый шаблон

$$T = A \oplus K.$$

В качестве шаблона в системе хранятся T, координаты контрольных точек $\{p_i\}_{i=1}^{l}$ и сигнатура $h(k)$ ключа, где h — хэш-функция.

При верификации личности пользователь предъявляет отпечаток пальца. На изображении находятся контрольные точки $\{q_i\}_{i=1}^{l}$, которые наиболее близко соответствуют конфигурации $\{p_i\}_{i=1}^{l}$. Для точек $\{q_i\}_{i=1}^{l}$ вычисляется топологический вектор $B = D(q_1)||\ldots||D(q_l)$, который используется для восстановления ключевой последовательности

$$K' = B \oplus T.$$

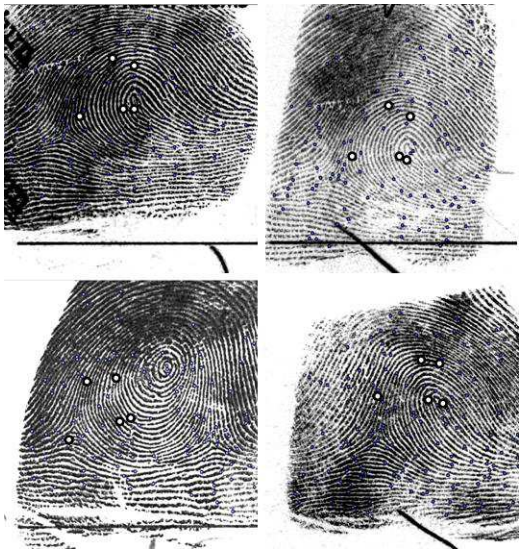


Рис. 5. Пример схожих кластеров на различных отпечатках пальцев.

Если представить разницу A и B как строку ошибочных битов E, т. е. $B = A \oplus E$, тогда

$$K' = B \oplus T = A \oplus E \oplus A \oplus K = K \oplus E.$$

Далее восстанавливается $k' = Decode(K \oplus E)$. Если вектор ошибок таков, что используемые корректирующие коды их исправляют, тогда мы получим побитно точное восстановление $h(k') = h(k)$.

Основной целью защищённой идентификации является возможность идентификации по данным, которые не раскрывают личность («биохэш»). В предложенной схеме $T = A \oplus K$ и $h(k)$ не дают информации об используемой биометрии. Для вычисления исходной биометрии требуется обращение хэш-функции, что является вычислительно сложной задачей для стандартных хэш-функций (например, SHA-256). Объём разглашения в результате опубликования $\{p_i\}_{i=1}^{l}$ уже более существенный. Он зависит от количества l контрольных точек. При увеличении l наступит момент, когда этой информации о координатах контрольных точек $\{p_i\}_{i=1}^{l}$ достаточно для определения личности пользователя. При l от 3 до 10 точек опубликование $\{p_i\}_{i=1}^{l}$ разглашение ещё не очень значительно. Схожие кластеры могут быть найдены практически на любом отпечатке. На рис. 5 приведен пример пятиточечных кластеров на различных отпечатках.

Эксперименты

При реализации технологий защищённой биометрической идентификации требуется оценить две характеристики: точность идентификации и энтропию ключа k. Мы ещё не реализовали корректирующие коды, поэтому эти характеристики оценены следующим приближённым способом. Как вид-

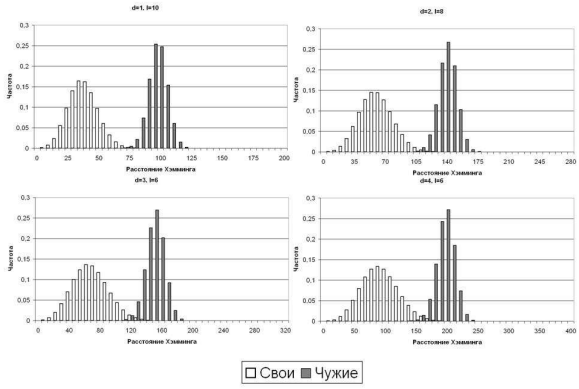


Рис. 6. Гистограммы расстояний Хэмминга для бинарного вектора.

но, из шаблона применения определяющими параметрами являются длина ключевой последовательности K, которая равна $16ld+4l$, и структура ошибок E. Предположим, что мы имеем в распоряжении идеальные коды, которые исправляют не более e ошибок. Если расстояние Хэмминга для двух предъявлений одного отпечатка пальцев («свои») будет больше e, тогда мы наблюдаем ошибку идентификации первого рода. Соответственно, если расстояние для предъявлений различных отпечатков («чужие») будет меньше e, то мы наблюдаем ошибку второго рода. На рис. 6 представлены гистограммы распределений расстояния Хэмминга для различных значений d и l для «своих» и «чужих» сравнений. Изменяя количество корректируемых битов e, мы управляем соотношением ошибок идентификации.

Энтропия ключа в предположении об идеальности кодов может быть оценена на основе Шеннонвского предела [5]

$$len(k) \leqslant s\left(1 + p\log_2 p + (1-p)\log_2(1-p)\right),$$

где $s = 16ld + 4l$ — длина шаблона, $p = e/s$ — доля корректируемых битов.

На рис. 7 приведены DET-кривые соотношения ошибок первого (FRR) и второго (FAR) рода с указанием верхней оценки энтропии ключа в каждой точке. Энтропия ключа сильно зависит от числа корректируемых битов. Мы проводили эксперименты на публично доступной базе FVC2002 DB1, которая содержит изображения среднего качества. Как видно из рис. 6, в среднем наблюдается 20% ошибок в топологических векторах в «своих» сравнениях. Для достижения приемлемой ошибки FRR идентификации требуется корректировать 30–35% битов. На базах лучшего качества (например, по нашей предварительной оценке, на базе NIST SD14 побитных ошибок на четверть меньше) энтропия может увеличиться примерно в два раза.

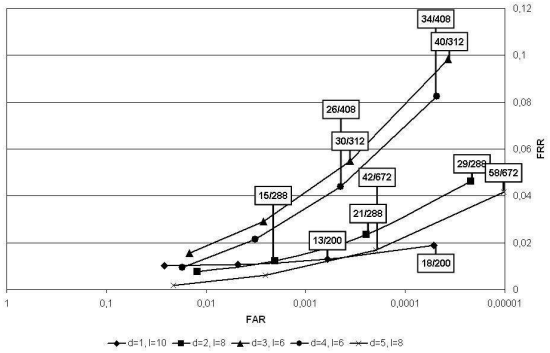


Рис. 7. DET-кривые (на подписях указаны длина ключа/длина бинарного вектора.

Заключение

В статье изложен алгоритм извлечения из изображения отпечатков пальцев бинарную строку, которая может быть использована в задачах защищённой биометрической идентификации и биометрической криптографии.

Направлением дальнейших исследований является выбор корректирующих кодов и сокращение дополнительной информации по позиционированию контрольных точек. Также мы планируем рассмотреть вопрос увеличения длины описания топологического события. Источниками дополнительных битов могут быть расстояния от исходной точки до топологического события. Такая информация может дать дополнительные два или три бита в каждое топологическое событие. Как следствие, длина итогового бинарного вектора может быть увеличена примерно на 50%.

Литература

[1] *Gudkov V. Y., Ushmaev O. S.* Topological Approach to User-Dependent Key Generation from Fingerprints // 20-th Int'l. Conf. on Pattern Recognition, 2010. — Vol. 11, No. 1. — Pp. 111–122.

[2] *Hao F., Anderson R., Daugman J.* Combining crypto with biometrics effectively // IEEE Trans. Comp, 2006. — Vol. 5. — Pp. 1081–1088.

[3] *Pankanti S., Prabhakar S., Jain A. K.* On the individuality of fingerprints // IEEE Trans. PAMI, 2002. — Vol. 24, No. 8. — Pp. 1010–1025.

[4] *Gudkov V. Y.* Mathematical Models of Fingerprint Image On the Basis of Lines Description // 19th Int'l. Conf. on Computer Graphics and Vision (GraphiCon'2009), 2009. — Pp. 223–227.

[5] *Vanstone S.A., van Oorshot P.C.* An introduction to error correcting codes with applications. — Kluwer Academic Publishers, 1989.

[6] *Sparrow M.K.* Pat. 5631971 USA, Int.Cl. G 06 K 9/00. Vector based topological fingerprint matching / M.K. Sparrow (Winchester). — Field: Jul. 15, 1994; Date of patent: May. 20, 1997; U.S.Cl. 382/125. — 17 p.

Анализ эхокардиографических данных на основе вычисления оптического потока

Марьяскин Е. Л., Ивановский С. А., Немирко А. П.
e.maryaskin@gmail.com
г. Санкт-Петербург, СПбГЭТУ «ЛЭТИ»

Данная статья рассматривает подход к анализу эхокардиографических данных с помощью сегментации видеоряда, основанной на вычислении оптического потока. Предлагаются как модификации самого метода вычисления потока, так и алгоритмы последующей сегментации, основанные на кластеризации векторов поля потока. Рассматривается работа реализущего анализ программного средства как на модельных, так и на реальных данных. В конце делается вывод об успешности применения такого подхода в диагностике и перспективах развития методов сегментации на основе вычисления оптического потока

Эхокардиография на протяжении последних 15–20 лет является одним из основных методов визуализации сердца. Недостатком эхокардиографии является выраженная зависимость от квалификации исследователя. У исследователя резко возрастает процент ошибок диагностики. Существующие в настоящее время методы исследования результатов эхокардиографии не подразумевают использования вспомогательных автоматизированных средств [5]. В этой связи возникает естественное стремление протестировать применимость разработанной методики к данной области задач.

Анализ движения

Анализ движения долгое время являлся специализированной тематикой, которая не имела особого значения в общей теории обработки изображений. Существующие методики сегментации динамических сцен оказываются очень простыми, не работающими эффективно для многих актуальных случаев и не учитывающими множество известных факторов, которые могли бы помочь в решении задач, поставленных перед исследователем [1].

Оптический поток. Понятие оптического потока взято из гидрогазодинамики. Оптический поток определяется как «поток» уровней яркости на плоскости изображений. Оптический поток и поле движений равны, если объекты не изменяют энергетическую освещенность на плоскости изображений в процессе движения в сцене. Хотя это кажется разумным на первый взгляд, более полный анализ показывает, что это точно выполняется только в очень ограниченных случаях [1, 3]. Для вычисления оптического потока на основе уравнений потока разработано и реализовано несколько алгоритмов.

Алгоритмы обработки видеопоследовательности

Любая обработка видеопоследовательностей с выделением информации по движению состоит из двух частей: предобработка и вычисление оптического потока и постобработка полученных потоковых данных с попыткой извлечения из них информации. В данной статье описываются модификации, предложенные к обеим частям процедуры обработки.

Вычисление потока. Исследования показали, что уже на этапе вычисления оптического потока возможно внести алгоритмические модификации, которые помогут ускорить процесс вычисления без существенной потери информации. Было разработано специальное программное средство, предназначенное для быстрого вычисления потока и демонстрации его структуры. Вычисление векторного поля потока производится на основе метода Лукаса–Канаде [2, 4] и содержит следующие особенности:

1. Вектора вычисляются отдельно в каждой точке, причем ни одно вычисление не производится дважды.
2. Направление, задаваемое вектором, вычисляется не абсолютно, а с точностью до $\pi/4$.
3. Применяется мультимасштабная схема Лапласа–Гаусса [1].

Обработка поля потока. Представленный метод в значительной степени отличается от принятых в обработке изображений методик, во-первых, поскольку не ограничивается только отсеиванием небольших по длине векторов потока, а учитывает обе компоненты векторов, а во-вторых, поскольку не заканчивается получением векторного поля потока, а работает с ним, как с новым, подлежащим сегментации трехмерным сигналом. Ключевые этапы метода:

1. *Кластеризация.* Результат кластеризации представляет собой векторное поле, разделенное на кластеры близких (в Евклидовом смысле) вектором, причем каждый кластер представляет собой либо совокупность векторов шума, либо часть реального объекта.
2. *Фильтрация.* Разработанные фильтры в целом основаны на существующих методах статической обработки изображений [3]
3. *Вычисление характеристик.* Определение характеристик, позволяющих определить наилучший момент прекращения итеративного процес-

са. Характеристики отражают основные свойства кластеров: форма, размер, расположение, заполненность точками, распределение внутри кластеров и пр.

4. *Постобработка.* На данном этапе происходит окончательное формирование объектов из кластеров, с постфильтрацией и подчеркиванием формы.
5. *Переход к новой итерации.* Итерации повторяются до тех пор, пока не будут достигнуты требуемые характеристики.

Вспомогательные алгоритмы

В рамках формирования процесса обработки были разработаны и применены некоторые дополнительные алгоритмические средства, способствующие эффективной сегментации поля потока. Такие алгоритмы можно разделить на несколько типов.

Алгоритмы анализа кластеров предназначены для анализа текущего состояния данных, возможного перераспределения некоторых логических сущностей, но не направлены на изменение текущего актуального набора данных. В частности, используются определение центра масс каждого кластера, определение радиуса кластеров, общее направление движения кластера, траектория движения, вычисление выпуклой оболочки кластеров и оценки ее геометрических характеристик, эллиптическая аппроксимация формы кластеров.

Основными же в этом разделе являются алгоритм вычисления количества объектов, составляемых кластерами, алгоритм сопоставления кластеров этим объектом и алгоритм вычисления характеристик невидимых в этом кадре объектов.

Алгоритмы модификации кластеров представлены, по сути, различными фильтрами, модифицирующими текущие данные в различных целях, связанных с выделением целевых объектов.

В частности, реализованы фильтры обособленных точек, фильтры удаленных от центров кластеров точек, фильтры малых плотных групп точек, фильтры сжатия выпуклых оболочек кластеров, фильтры разреженных кластеров [1, 3]. Все фильтры могут применяться как на всем изображении, так и отдельно при работе с конкретным кластером.

Алгоритмы оценки кластеризации позволяют количественно оценить эффективность сегментации изображения и судить о том, насколько реально результаты обработки изображения соответствуют исходной сцене.

На уровне анализа кластеров используются реализации алгоритмов оценки статистических характеристик компонент каждого кластера. Применяются:

1) характеристика отношения дисперсий кластеров;
2) характеристика наименьшей доли площади;
3) комбинированная характеристика, учитывающая разреженность кластера, его размер и количество компонент.

Описание процесса обработки

Этот раздел рассматривает более подробно особенности этапов обрабоки, описанные выше. Приводятся примеры, полученные при обработки модельного видеоряда, близкого по характеристикам к реальным данным эхокардиографии.

Первичное выделение кластеров использует идею итерационной кластеризации и фильтрации изображения до тех пор пока не будут достигнуты заданные пороговые значения характеристик. Целью этапа является получение кластеров, отвечающих нескольким свойствам:

1. Каждый кластер полностью является частью какого-либо объекта.
2. Ни один кластер не соответствует шуму.
3. Ни один кластер не является частью двух и более объектов.

В самом общем виде предлагаемый алгоритм предварительной обработки выглядит следующим образом.

Выполнять

1) очередной фильтр;
2) перекластеризацию на большое число кластеров;
3) удаление самого разреженного кластера

пока не будет достигнуто пороговое значение характеристик.

Предлагаемый алгоритм обработки имеет следующие особенности:

1. Фильтры, описанные выше, на каждом шаге выполняются парами, причем наборы различны на четных и нечетных шагах. Успешность применения именно этих фильтров и именно в таком сочетании установлена на основании изучения стандартных средств обработки изображений, исследования предметной области обработки потоковых кадров и серии экспериментов, проведенных на целевых видеопоследовательностях.
2. Кластеризация на 8 кластеров. Цель кластеризации — породить как можно большее число кластеров, соответствующих накладываемым на них условиям. Опытно установлено, что 8 кластеров достаточно для успешной сегментации при случае 1–4 объектов.
3. Вычисление численной характеристики, позволяющей принять решение о продолжении или завершении процесса в зависимости от достиг-

нутых значений. Правило вычисляется по общим характеристикам кластеров, описанным выше, и с помощью алгоритма определения реального количества кластеров

В качестве примера работы первого этапа можно привести пару изображений до и после обработки (рис. 1).

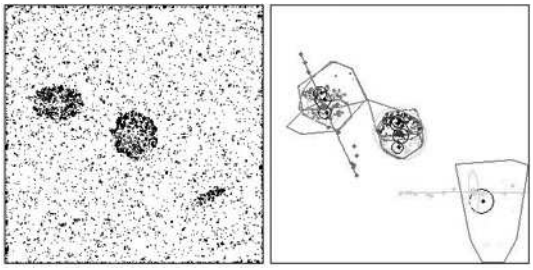


Рис. 1. Пример работы этапа первичной кластеризации

Подготовка к окончательной сегментации особенно актуальна для тех случаев, когда используется информации о предыдущих сессиях, поскольку именно на этом этапе происходит сопоставление объектов кластерам и, в любом случае, начиная с этого этапа кластеры в понимании алгоритма соответствуют объектам, и именно в таком виде будут потом записаны в случае сохранения результатов сегментации.

Кроме того, на этом этапе осуществляется фильтрация шумовых векторов, близких к векторам от объектов, что позволяет более точно задать границы объектов. Это оказывается важно как само по себе, так и при использовании этой информации для сегментации других кадров.

Алгоритм, реализующий этап подготовки, представляет собой последовательное выполнение фильтров индивидуальной очистки удаленных точек и индивидуального сжатия границ, до тех пор пока наименьшая по кластерам характеристика доли занимаемой площади не достигнет порогового значения.

На рис. 2 показаны изображения сразу после первичного выделения кластеров и после выполнения этапа подготовки к окончательной сегментации.

Постообработка

На данном этапе происходит окончательное формирование объектов из кластеров, с постфильтрацией и подчеркиванием формы. Алгоритм представляет собой итерационное выполнение фильтров индивидуального исключения удаленных точек и усечения границ выпуклой оболочки.

На рис. 3 показаны изображения сразу после выполнения этапа подготовки к окончательной сегментации и после этапа постобработки.

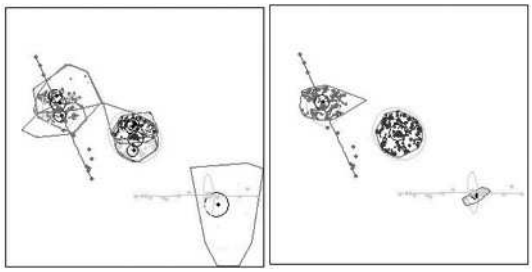


Рис. 2. Пример работы этапа подготовки к окончательной кластеризации

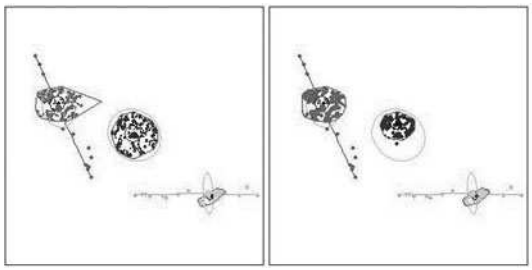


Рис. 3. Пример работы этапа постобработки

Анализ реальных данных

Для тестирования на реальных сюжетах были взяты результаты эхокардиографии, сделанные с помощью соответствующего медицинского прибора. Результаты эхокардиографии имеют большую общность с рассмотренными моделями, поэтому можно предположить достаточно успешную сегментацию. Основные особенности данных последовательностей:

1) сильное зашумление;
2) отсутствие выраженных траекторий движений;
3) постоянное изменение размеров;
4) Отсутствие изменения собственной яркости.

В качестве примера была использована реальная запись последовательности изображений, некоторые кадры из которой приведены на рис. 4. Сегментация в каждом таком кадре очень затруднена и только опытный врач может делать правильные выводы об объекте наблюдения. Попробуем теперь провести предложенный метод на одном из этих кадров. На рис. 5 показаны этапы обработки входной информации. Слева направо:

1) Входной поток, полученный непосредственно проведением алгоритма Лукаса–Канаде.
2) Результат первого этапа работа алгоритма, этапа итеративной кластеризации-фильтрации.
3) Окончательный результат, достигнутый на этапе постобработки, с уточнением форм и границ кластеров.

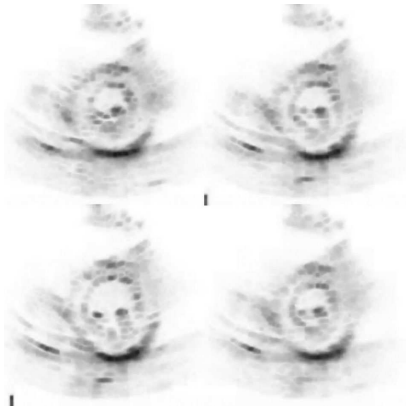


Рис. 4. Некоторые кадры эхокардиографического видеоряда

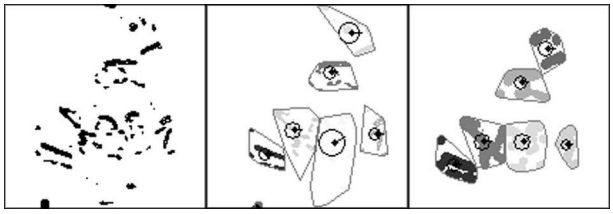


Рис. 5. Этапы обработки данных

Рис. 6. Сегментированное изображение

Для оценки результата достаточно совместить выявленные алгоритмом границы кластеров и исходное изображение, чтобы установить продуктивность работы метода. Изображение теперь представляется сегментированным по принципу совместно движущихся участков и оказывается гораздо более удобным для понимания неспециалистом в области ультразвуковой диагностики.

Выводы

В статье описан подход к анализу эхокардиографических данных на основе вычисления и последующей сегментации поля оптического подхода. Полученные результаты позволяют судить об успешности применения данного подхода к задачам эхокардиографической диагностики, поскольку данный подход, с одной стороны, упрощает понимание сцены неспециалистом, а с другой — предоставляет дополнительную информацию о происходящем в кадре.

В дальнейшем представляется необходимым рассмотреть применение сегментации на основе оптического потока к узким диагностическим задачам по выявлению особенностей или отклонений, плохо диагностируемых невооруженным взглядом, что позволит снизить зависимость успешности диагностики от квалификации специалиста.

Литература

[1] *Яне Б.* Цифровая обработка изображений. — М.: Техносфера, 2007. — 583 с.

[2] *Baker S., Matthews I.* Lucas-Kanade 20 Years On: A Unifying Framework // The Robotics Institute Carnegie Mellon University. — 2009. — 30 p.

[3] *Ballard D.* Computer vision. — Rochester, New-York: Brown Department of computer science. University of Rochester, 2006. — 573 p.

[4] *Verri A., Poggio T.* Motion field and optical flow. Qualitative properties . — London, Washington: IEEE Trans, 2009.

[5] *Рыбакова М. А.* `www.medison.ru/si/art237.htm` — Возможности современной эхокардиографии.

Оценка эффективности распознавания стадий анестезии по энтропийным характеристикам ЭЭГ*

Немирко А. П., Манило Л. А., Калиниченко А. Н., Волкова С. С.
apn-bs@yandex.ru
Санкт-Петербург, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ»

Рассмотрены методы анализа глубины наркоза по ЭЭГ, основанные на применении энтропийных характеристик. Проведен анализ эффективности распознавания состояний бодрствования и наркоза по параметрам аппроксимированной энтропии, вычисляемой во временной области, а также спектральной энтропии. Проведено сравнение предлагаемых методов с известным алгоритмом анализа ЭЭГ, использующим параметры биспектра ЭЭГ-сигнала.

Автоматическое распознавание стадий анестезии в ходе проведения операций представляется важнейшей задачей современной анестезиологии. Непрерывный контроль глубины наркоза способствует поддержанию требуемого уровня воздействия анестетиков, а при определенных условиях обуславливает проведение коррекции программы управления функциональным состоянием пациента. Использование соответствующих технических средств направлено как на оптимизацию условий проведения операции, так и на обеспечение безопасности самого пациента.

В системах наблюдения обычно анализируют ЭЭГ-сигнал, наиболее полно отражающий уровень функционирования головного мозга человека. При этом используют методы обработки, основанные на вычислении показателей биспектра сигнала, параметров спектральной энтропии и ряда других показателей [1]. Несмотря на то, что наиболее надежным на сегодняшний день считается биспектральный метод анализа ЭЭГ, актуальной остается задача исследования других характеристик сигнала, например, основанных на вычислении показателей энтропии во временной и частотной областях. Интерес к энтропийным характеристикам вызван возможностью на их основе распознавать сложные сигналы, отличающиеся степенью выраженности хаотической и детерминированной составляющих. Именно эти различия в свойствах ЭЭГ-сигнала проявляются в процессе изменения глубины наркоза, и особенно выражены для двух состояний: бодрствование и наркоз. Это и является основанием для разработки и исследования энтропийных методов анализа ЭЭГ, направленных на повышение эффективности распознавания стадий анестезии.

Методы вычисления показателей энтропии

В качестве энтропийных характеристик в данном исследовании были выбраны аппроксимированная энтропия (АЭ) и спектральная энтропия (СЭ). Первая характеристика известна как модификация энтропии Колмогорова, вычисляемая во временной области. Как показали исследования, она обеспечивает получение надежных оценок степени хаотичности сигнала на коротких выборках данных [2]. Алгоритм расчета значений АЭ и обоснование использования скорректированной ее оценки подробно описаны в работах [2–4]. В данном исследовании в качестве основного параметра этой характеристики использован относительный минимум скорректированной функции энтропии, определяемый в следующем виде:

$$ME = ApEn(0) - \min_{m=1\ldots6}\{ApEn_{cor}(m)\},$$

где m — длина анализируемых последовательностей отсчетов сигнала, $ApEn(0)$ — энтропия одиночных событий.

Этот показатель аппроксимирует нижнюю границу К-энтропии и может быть использован для оценки выраженности регулярных и хаотических компонент, содержащихся в ЭЭГ-сигнале.

В частотной области может быть рассчитана СЭ, важным преимуществом которой является то, что вклад в энтропию составляющих, лежащих в любом заданном диапазоне частот, может быть выделен отдельно [5]. Основой для расчета СЭ является оценка спектральной плотности мощности сигнала (СПМ), которая может быть получена с использованием быстрого преобразования Фурье. Этот метод позволяет получить из последовательности отсчётов равномерно дискретизованного с частотой f_d сигнала $x(n) = x(nT)$, где $T = 1/f_d$ — интервал дискретизации, такое же количество комплексных величин $X(k)$:

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-i2\pi k f_0 nT}, k = 0, 1, 2, \ldots, N-1,$$

где N — число отсчётов в анализируемом фрагменте сигнала, а каждый из элементов преобразования $X(k)$ соответствует частоте $f_k = kf_d/N$.

Вычисление СЭ основано на такой мере информации, как энтропия Шеннона [6]. Если применить эту характеристику не к самому сигналу, а

Работа выполнена при финансовой поддержке РФФИ, гранты № 09-01-00501, № 10-01-00604; Министерства образования и науки РФ, госконтракт № 02.522.11.2020 от 10.03.2009 г.

к его спектру, то получится спектральная энтропия [5]. Спектральная энтропия в некотором диапазоне частот $[f_1, f_2]$ для заданного фрагмента сигнала может быть вычислена с помощью следующей последовательности шагов.

Путём возведения в квадрат амплитуды каждого из элементов $X(k)$ преобразования Фурье от сигнала $x(n)$ рассчитывается соответствующее значение спектра мощности $P(k)$:

$$P(k) = X(k)X^*(k)\,,$$

где $X^*(k_i)$ представляет собой комплексно сопряжённое значение элемента разложения Фурье $X(k)$.

Далее спектр мощности нормализуется, для чего рассчитывается такая константа нормализации C_n, что сумма нормализованного спектра мощности в пределах заданного диапазона частот $[f_1, f_2]$, равняется единице. Тогда значения нормализованного спектра мощности $P_n(k)$ рассчитываются как

$$\sum_{f_i=f_1}^{f_2} P_n(f_i) = C_n \sum_{f_i=f_1}^{f_2} P(f_i) = 1\,.$$

Спектральная энтропия, соответствующая диапазону частот $[f_1, f_2]$, вычисляется как сумма:

$$E[f_1, f_2] = \sum_{f_i=f_1}^{f_2} P_n(f_i) log\left(\frac{1}{P_n(f_i)}\right).$$

Оба показателя нормализуют до представления каждого в шкале значений от 0 (полная регулярность) до 100 (максимальная нерегулярность). При этом показатель аппроксимированной энтропии определяется в виде:

$$ME_n = \frac{ME}{ApEn(0)}100\,,$$

где $ApEn(0)$ принимает максимальное значение в ряду точечных оценок условной энтропии.

Нормализованный показатель спектральной энтропии $E_n[f_1, f_2]$ рассчитывается как:

$$E_n[f_1, f_2] = \frac{E[f_1, f_2]}{log(N[f_1, f_2])}100\,,$$

где $N[f_1, f_2]$ равно общему числу частотных компонентов в диапазоне $[f_1, f_2]$.

Исследование свойств рассмотренных параметров энтропии на модельных сигналах показало, что для гармонических сигналов соответствующие значения равны нулю, а с ростом степени хаотичности сигнала они приближаются к максимальному значению.

Выбор параметров алгоритмов обработки ЭЭГ

В ходе экспериментов было установлено, что значения рассматриваемых показателей энтропии зависят от ряда параметров, изменение которых влияет на результат распознавания стадий бодрствования и глубокого наркоза. Вопросы оптимизации параметров алгоритмов распознавания стадий анестезии на основе АЭ и СЭ обсуждаются в работах [7, 8].

При вычислении АЭ важным является выбор порога, определяющего зону нечувствительности к изменениям амплитуды сигнала. Анализ зависимости критерия Фишера J от величины порога r позволил выбрать значение $r = 0,3SD$, где SD — стандартное отклонение исходной выборки отсчетов. Кроме того, были проведены исследования, связанные с выбором частоты отсчетов ЭЭГ-сигнала и длины анализируемых фрагментов.

Для получения достоверных значений АЭ, оценивающей вероятность появления новых цепочек символов в исследуемой последовательности, требуется анализ достаточно длительных фрагментов сигнала. С учетом ограничения на время анализа, что связано с требованием непрерывного контроля состояния пациента, длина анализируемых фрагментов ЭЭГ выбрана, равной 5 с.

Выбор частоты поступающих на обработку отсчетов ЭЭГ, определяется двумя факторами. Первый связан с ограничением на скорость обработки и анализа данных. Вычисление АЭ остается чрезвычайно ресурсоемкой задачей даже для современных быстродействующих процессоров. Второй связан с влиянием частоты отсчетов на эффективность распознавания анализируемых состояний по показателю энтропии.

При вычислении параметров АЭ необходимо выполнить число операций N_t, определяемое по формуле:

$$N_t = N^2 + \sum_{i=1}^{M} i\frac{N!}{(N-i)!}\,,$$

где N — длина исходной выборки отсчетов сигнала, M — длина анализируемых последовательностей.

Анализ 5 с фрагмента ЭЭГ-сигнала, дискретизованного с частотой 500 Гц, требует выполнения примерно (5×10^{18}) операций.

С использованием процессора, работающего на тактовой частоте 1,2 ГГц, проведено исследование скорости обработки данных. В табл. 1 приведены усредненные по множеству реализаций значения времени, затраченного на обработку 5 с фрагментов ЭЭГ-сигнала при частоте дискретизации $f_d = 500, 250, 150$ Гц. Здесь указаны нижний и верхний пределы временных затрат, необходимых для вычисления интегрального показателя ME_n по цепочкам отсчетов длиной от 3 до 10 элементов. В таблице также приведены результаты сравнения критерия J для тех же частот f_d. Как показали результаты экспериментов, качество распозна-

вания состояний бодрствования и глубокого наркоза не ухудшается при переходе от частоты $f_d = 500$ Гц к частоте $f_d = 250$ Гц.

Таблица 1. Оценки временных затрат и критерия Фишера для показателя ME_n при разных частотах дискретизации ЭЭГ-сигнала f_d

Оценка	Частота дискретизации f_d, Гц		
	500	250	150
Время, с	1,63–2,71	0,42–0,68	0,16–0,26
Критерий J	3,13	4,49	2,13

Исходя из анализа полученных данных, в алгоритме вычисления энтропии во временной области выбрана частота $f_d = 250$ Гц. При этом время обработки считается допустимым, а объем выборки позволяет получать достоверные оценки АЭ.

Для метода анализа СЭ также проведено исследование влияния на эффективность распознавания стадий анестезии частоты дискретизации f_d, диапазона анализируемых частот ЭЭГ $[f_1, f_2]$, а также способа вычисления СПМ [8]. По результатам экспериментов выбраны следующие параметры алгоритма обработки: $f_d = 500$ Гц, частотный диапазон для расчета СЭ от 8 Гц до 47 Гц, вычисление СПМ по методу Уэлча.

Методика и результаты экспериментальных исследований

Разработанные алгоритмы СЭ и АЭ были включены в состав программного обеспечения специального прибора — монитора анестезии. Последний прошел клиническую апробацию совместно с известным монитором глубины анестезии BIS фирмы «Aspect Medical Systems», рассчитывающим оценку глубины анестезии на основе анализа биспектра ЭЭГ. Испытания прибора проводились во время анестезиологического обеспечения хирургических операций у 30 пациентов. Длительность операций составляла от 1 до 6 часов.

Каждому пациенту на всё время проведения операции на лобную часть головы накладывались два комплекта электродов (испытуемого прибора и импортного аналога). Полученные данные заносились в таблицы, а также использовались для графического представления результатов.

На рис. 1 приведён график, демонстрирующий пример данных по СЭ, полученных в ходе одной из операций. Кружками отмечены показания BIS-монитора, а крестиками — прибора, реализующего рассматриваемый алгоритм.

На данном графике можно отметить следующие основные участки:

— бодрствование перед подачей анестетика (от 0 до 300 с);
— интенсивное применение начальной дозы анестетика (от 300 до 700 с);
— период стабильной анестезии (от 700 до 8 500 с);
— этап пробуждения (после 8500 с).

Для каждой операции рассчитывались следующие статистические показатели: коэффициент взаимной корреляции (C_{xy}), среднеквадратическое отклонение (σ_{xy}), максимальное отклонение (D_{xy}).

Если обозначить последовательности значений, полученных от испытуемого прибора и импортного аналога соответственно $x(n)$ и $y(n)$, где $n = 1, 2, \ldots, N$ (N — число значений в каждой из последовательностей), то расчёт указанных параметров будет выполняться в соответствии со следующими формулами:

$$C_{xy} = \frac{1}{N\sigma_x\sigma_y}\sum_{n=1}^{N}[x(n) - \bar{x}][y(n) - \bar{y}]\,;$$

$$\sigma_{xy} = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}[x(n) - y(n)]^2}\,;$$

$$D_{xy} = max|x(n) - y(n)|\,,$$

где $\bar{x} = 1/N\sum_{n=1}^{N} x(n)$, а $\bar{y} = 1/N\sum_{n=1}^{N} y(n)$.

Далее для показателей, рассчитанных по всем операциям, вычислялись их максимальные, минимальные и средние значения. Полученные результаты для метода СЭ приведены в табл. 2.

Таблица 2. Результаты сравнения индексов глубины наркоза для метода СЭ и BIS-монитора

Показатель	Максимальное значение	Минимальное значение	Среднее значение
C_{xy}	0,86	0,55	0,79
σ_{xy}	12,2	7,1	10,6
D_{xy}	18,7	16,1	19,9

В методе АЭ получены большие отклонения измеряемых величин от показаний BIS-монитора.

Для каждой из полученных последовательностей были выбраны участки, соответствующие длительному устойчивому состоянию глубокого наркоза. По этим участкам оценивались значения стандартного отклонения. В табл. 3 содержатся величины минимальных, максимальных и средних значений указанных стандартных отклонений, рассчитанных по отдельности для данных с испытуемого прибора и импортного аналога.

Анализ полученных данных свидетельствует о хорошей согласованности результатов анализа на участках глубокого наркоза. В то же время существует достаточно устойчивое отклонение параметров СЭ на этапах выхода из наркоза.

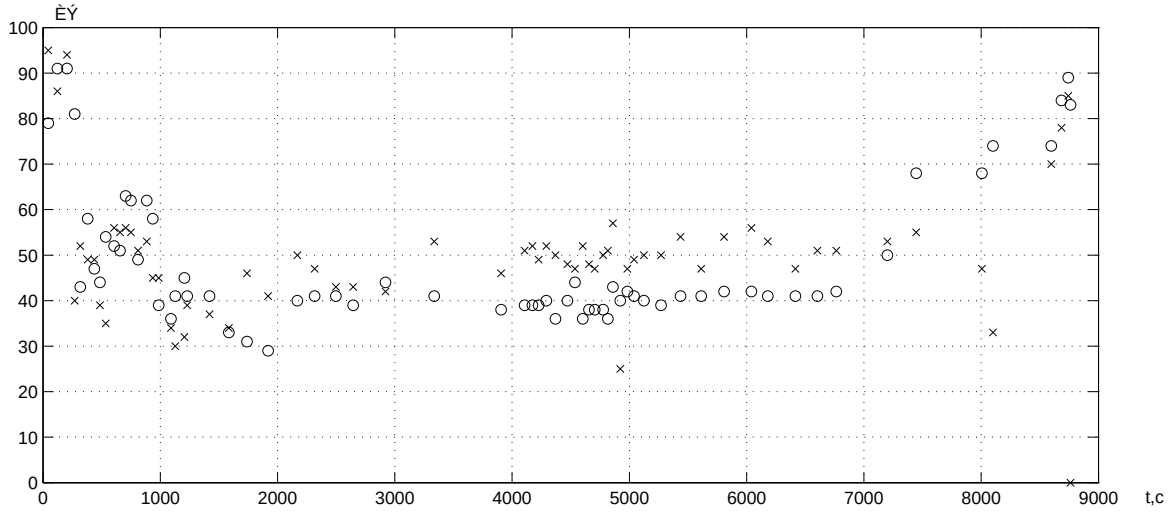


Рис. 1. Пример анализа ЭЭГ-сигнала в ходе одной из операций по показателю СЭ

Таблица 3. Оценка отклонений индексов глубины наркоза по методу СЭ от BIS-монитора

Прибор	Стандартное отклонение на стабильном участке		
	Максимальное значение	Минимальное значение	Среднее значение
СЭ-монитор	9,15	3,34	5,89
BIS-монитор	9,97	3,20	6,62

Выводы

Результаты проведенных экспериментальных исследований показали возможности применения энтропийных характеристик в клинических условиях при распознавании стадий анестезии по ЭЭГ. Параметры энтропии, вычисляемые как во временной, так и в частотной областях достаточно надежно распознают стадии бодрствования и глубокого наркоза. Однако требуются дальнейшие исследования, возможно в области многопараметрического описания, направленные на повышение эффективности распознавания промежуточных состояний в ходе проведения хирургических операций.

Литература

[1] *Schwilden H.* Concepts of EEG processing: from power spectrum to bispectrum, fractals, entropies and all that // Best Practice & Research Clinical Anaesthesiology, 2006. — Vol. 20, No. 1. — Pp. 31–48.

[2] *Манило Л. А., Немирко А. П.* Аппроксимация энтропии Колмогорова при анализе хаотических процессов на конечных выборках // Математические методы распознавания образов ММРО-14: 14-я Всероссийская конференция, Владимирская обл., г. Суздаль. — Сборник докладов. — М.: МАКС Пресс, 2009. — С. 405–407.

[3] Nonlinear Biomedical Signal Processing / Edited by Metin Akay // Dynamic Analysis and Modelling. — N.-Y.: IEEE. 2001. — Vol. 2. — 341 p.

[4] *Немирко А. П., Манило Л. А., Калиниченко А. Н., Волкова С. С.* Энтропийные методы оценки уровня анестезии по ЭЭГ-сигналу // Информационно-управляющие системы. — 2010. — № 3. — С. 69–74.

[5] *Viertio-Oja H., et al.* Description of the Entropy^{TM} algorithm as applied in the Datex-Ohmeda S/5^{TM} Entropy Module // Acta Anaesthesiol. Scand. — 2004. — Vol. 48. — Pp. 154–161.

[6] *Bruhn J., et al.* Shannon entropy applied to the measurement of the electroencephalographic effects of desflurane // Anesthesiology. — 2001. — Vol. 95. — Pp. 30–35.

[7] *Немирко А. П., Манило Л. А., Калиниченко А. Н., Волкова С. С.* Сравнительный анализ применения различных оценок энтропии ЭЭГ-сигнала для распознавания стадий наркоза // Биотехносфера. — 2010. — №3(9).

[8] *Калиниченко А. Н., Манило Л. А., Немирко А. П., Волкова С. С.* Оценка глубины анестезии по ЭЭГ на основе спектральной энтропии // Биотехносфера. — 2010. — №3(9).

Анализ пространственно-временных характеристик данных магнитной энцефалографии*

Лыжко Е. В.[1,2]***, Махортых С. А.***[1]
lyzko@yandex.ru
Пущино, [1]Институт математических проблем биологии РАН, [2]Пущинский государственный университет

В данной работе были изучены особенности пространственно-временных характеристик данных магнитной энцефалографии (МЭГ) при заболевании Паркинсона. Измерение сигнала осуществлялось в 148 точках на поверхности головы с частотой дискретизации 500 Гц. Для изучения пространственных характеристики данных МЭГ измеренные данные для каждого момента времени были разложены по полиномам Лежандра. К коэффициентам разложения было применено преобразования Фурье для получения временных характеристик сигнала.

Человеческий мозг является чрезвычайно сложной анатомической структурой, состоящей примерно из 100 млрд нейронов, которые объединяясь на различных уровнях, образуют структуры. Для каждой такой структуры в состоянии нормы характерна определенная биологическая активность. Однако чтобы осуществлять диагностику, необходимо знать допустимый диапазон изменения сигнала, и какая именно структура является источником патологической активности.

В данной работе для изучения биологической активности мозга были использованы данные магнитной энцефалографии (МЭГ). Магнитная энцефалография — быстро развивающаяся область экспериментального изучения высшей нервной деятельности человека, функциональных областей мозга и диагностики различных патологий. Измерения биомагнитных полей мозга у испытуемых были проведены в Медицинской школе Нью-Йоркского университета (New-York Medical School) на магнитном энфефалографе Magnes 2500 WH производства компании 4-D Neuroimaging (США). Поле, возникшее на поверхности головы, измеряется с помощью набора датчиков, находящихся в шлеме измерительного устройства МЭГ. Одной из главных составляющих установки МЭГ является сверхпроводящий квантовый интерференционный датчик-СКВИД. Это замкнутое кольцо из полупроводника, которое в одном или двух местах имеет джозефсоновский контакт. Измеряемый сигнал представляет собой пространственно-временную структуру: 148-мерный вектор измерений в 148 точках на поверхности головы, развернутый во временной ряд с частотой опроса датчиков 500 Гц.

На основе проведенного анализа динамики сигнала решается задача определения пространственно-временных характеристик по данным МЭГ.

Работа выполнена при финансовой поддержке РФФИ, проекты № 11-07-00519-а, № 10-01-00609-а, № 11-07-00716-а и № 10-07-00300.

Спектрально-аналитический метод описания данных

Аппроксимацию данных МЭГ осуществлялась с использованием сферических функций. В [5] приводятся результаты аппроксимации пространственного распределения биомагнитного поля сферическими гармониками. Если некоторая функция $B(\theta,\varphi)$ интегрируема с квадратом на сфере, она может быть разложена в ряд по сферическим гармоникам [4, 6]:

$$B(\theta,\varphi) \approx \sum_{n=0}^{N}\sum_{k=0}^{n}\left(a_{nk}p_n^k\cos k\varphi + b_{nk}p_n^k\sin k\varphi\right);$$

$$a_{n0}\int_0^{\pi}\int_{-\pi}^{\pi}(p_n)^2\sin\theta d\theta d\varphi = \int_0^{\pi}\int_{-\pi}^{\pi}B(\theta,\varphi)p_n\sin\theta\ d\theta d\varphi;$$

$$a_{nk}\int_0^{\pi}\int_{-\pi}^{\pi}(p_n^k)^2\cos^2(k\varphi)\sin\theta\,d\theta d\varphi = \int_0^{\pi}\int_{-\pi}^{\pi}B(\theta,\varphi)p_n^k\cos(k\varphi)\sin\theta\,d\theta d\varphi;$$

$$b_{nk}\int_0^{\pi}\int_{-\pi}^{\pi}(p_n^k)^2\sin^2(k\varphi)\sin\theta\,d\theta d\varphi = \int_0^{\pi}\int_{-\pi}^{\pi}B(\theta,\varphi)p_n^k\sin(k\varphi)\sin\theta\,d\theta d\varphi.$$

Значения нормировочных множителей вычисляются по формулам:

$$\int_0^{\pi}\int_{-\pi}^{\pi}(p_n)^2\sin\theta d\theta d\varphi = \frac{4\pi}{2n+1};$$

$$\int_0^{\pi}\int_{-\pi}^{\pi} \left(p_n^k\right)^2 \cos^2(k\varphi)\sin\theta d\theta d\varphi = \frac{2\pi(n+k)!}{(2n+1)(n-k)!};$$

$$\int_0^{\pi}\int_{-\pi}^{\pi} \left(p_n^k\right)^2 \sin^2(k\varphi)\sin\theta d\theta d\varphi = \frac{2\pi(n+k)!}{(2n+1)(n-k)!},$$

где p_n — полином Лежандра, p_n^k — присоединенный полином Лежандра. Для вычисления сферических гармоник функции магнитного поля на поверхности головы необходимы значения функции в узлах интегрирования. С этой целью была выполнена триангуляция Делоне (рис. 1), в качестве вершин треугольников были взяты координаты датчиков. Переход к сферическим координатам осуществлялся путем проекции треугольников на сферу единичного радиуса. Интегрирование производилось на равномерной сетке 200×100 соответственно ($\theta \in \in [-\pi,\pi]$, $\varphi \in [0,\pi]$). Значения в узлах равномерной сетки были получены путем линейной интерполяции данных МЭГ, отложенных относительно плоскости треугольника.

Так как данные МЭГ имеются не на всей поверхности сферы, то отсутствующие значения были заменены нулями. Однако количество отсутствующих данных можно минимизировать путем оптимального выбора направления осей для θ и φ. При изменении θ и φ с постоянным шагом плотность точек увеличивается на полюсах для $\varphi = 0$ и $\varphi = \pi$. Если один из полюсов находится в области с отсутствующими данными, то соотношение между отсутствующими и данными, полученными путем интерполяции составляет 1 : 1.

Временные характеристики вычисленных коэффициентов разложения были исследованы с помощью преобразования Фурье. Коэффициенты a_{n0}, a_{nk} и b_{nk} были вычислены для глубины разложения от 0 до 10, частота опроса датчиков 500 Гц, число дискретных моментов 10 172. На рис. 2 представлены зависимости амплитуды спектра для a_{00}, a_{10}, a_{20} и a_{30}. Максимальная амплитуда спектра была получена для коэффициентов a_{10} и a_{30} на частоте 11,7 Гц. Оценка качества аппроксимации осуществлялась по формуле среднеквадратичного отклонения:

$$\sigma = \frac{\sqrt{\sum\limits_{i=1}^{N}(B_i(\theta,\varphi) - \sum{}_i)^2}}{N}.$$

Зависимость качества аппроксимации по формуле среднеквадратичного отклонения от глубины разложения n представлена на рис. 3 для произвольно выбранного момента времени.

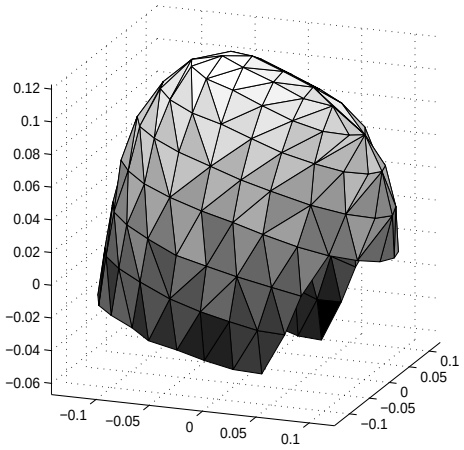


Рис. 1. Триангуляция Делоне, выполненная на поверхности шлема

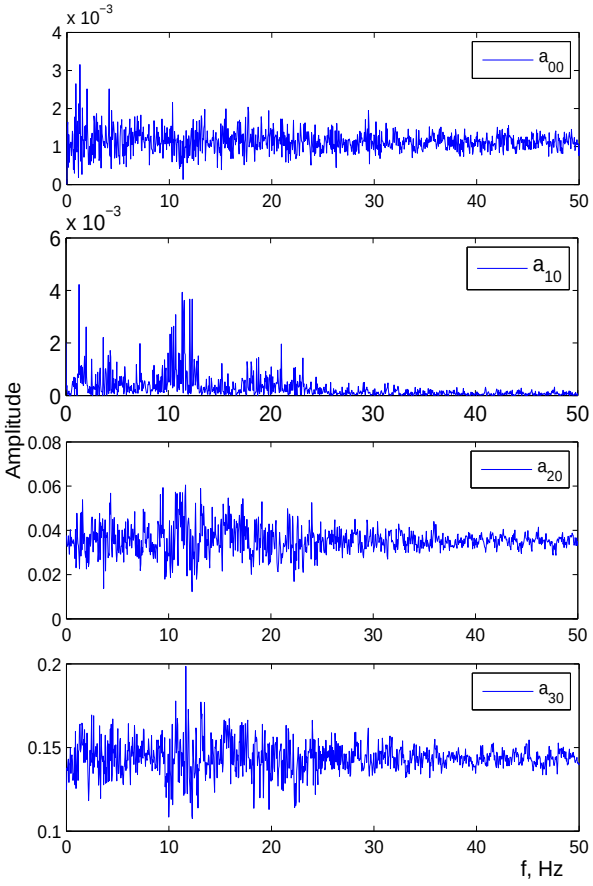


Рис. 2. Зависимость амплитуды спектра от частоты для a_{00}, a_{10}, a_{20}, a_{30}

Стандартные процедуры интерполяции не обеспечивают достаточной точности в этой области [3]. Для решения данной проблемы разработана процедура экстраполяции данных. Построен итерацион-

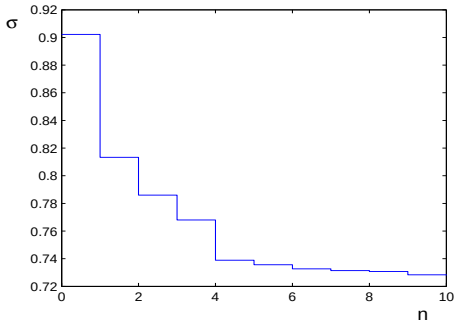


Рис. 3. Зависимость среднеквадратичного отклонения σ от глубины разложения n

ный алгоритм аппроксимации на сфере $A(i)$, где A — вектор коэффициентов разложения, i — номер итерации, приводящий при росте i к близкому к реальному распределению искомой функции на сфере.

Предобработка данных и методы локализации источников биомагнитной активности

Передобработка данных и методы локализации источников биомагнитной активности может быть разбита на следующие этапы:

1. Выделение полезного сигнала, связанного с конкретным видом деятельности мозга (например, сигнала, вызванного подачей периодического стимула — слухового, визуального, осязательного и т. д.; сигнала, связанного с генерацией тремора или слуховых галлюцинаций при паркинсонизме).
2. Выбор моментов времени для решения обратной задачи локализации токовых источников по пространственной картине поля на поверхности головы в эти моменты времени (с использованием результатов [1, 3, 7]).
3. Решение обратной задачи локализации источников как при наличии патологии, так и в случае нормальной активности. Прямой учет физиологических ограничений, получаемых с помощью ЯМР-томографии.

Спектральный подход пространственной локализации источников биомагнитной активности записей МЭГ протестирован на клинических записях, содержащих аномальную активности, связанную с болезнью Паркинсона и ее разновидностями (синдромы Tinnitus, Arkinson), генерация тремора, дистония. Исходные экспериментальные данные получены при обследовании пациента с болезнью Паркинсона для явно выраженного случая наличия одного источника патологии представлены на рис. 4.

На предварительном этапе на основе визуального анализа выделялись временные участки в записи МЭГ с наличием патологии (по амплитуде сигнала) и сигналом в норме. Для выбранных записей соответствующим различным типам активности решалась обратная задача локализация участков биомагнитной активности согласно методу минимизации функционала невязки [8].

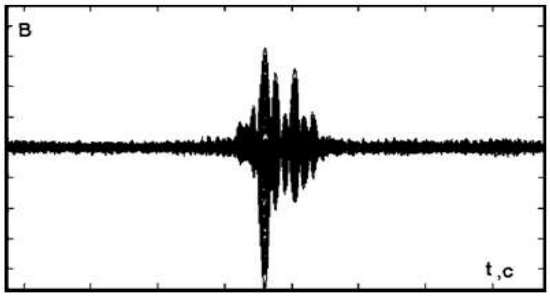


Рис. 4. Экспериментальные данные для случая с болезнью Паркинсона

Решение обратной задачи выполнено для каждого момента времени в виде одного токового диполя с переменным моментом, основную часть времени находящегося центральных, стволовых структурах головы.

По данным настоящего исследования у здоровых людей в состоянии спокойного бодрствования не обнаружены постоянные источники повышенной магнитной активности. Однако при функциональной нагрузке (в аудиторных экспериментах) появляются источники повышенной магнитной активности в височной коре, где находятся корковые проекции слуховой сенсорной системы. При фильтрации магнитного сигнала на разных частотах источники локализуются в разных полушариях — в левом на частоте 10 Гц и правом на частоте 20 Гц.

На рис. 5 представлены два взаимно перпендикулярных сечения ЯМР-томограммы, проходящие через источник (токовый диполь). Кружком отмечено местоположение диполя, а цветом — величина момента, меняющаяся от максимального (черный кружок) до нулевого (белый кружок). Отрезок, выходящий из кружка, указывает направление момента диполя.

На рис. 6 приведена локализация дипольных источников найденных записей на основании предварительной спектральной классификации [1]. Для обучающей выборки в 67 моментах времени было найдено 180 точек с хорошо локализующимися в одной области источниками.

Выводы

Таким образом, комбинирование пространственно-временных спектральных характеристик МЭГ позволяет повысить точность и детальность описания как временных, так и пространственных распределений биомагнитного поля, что, в част-

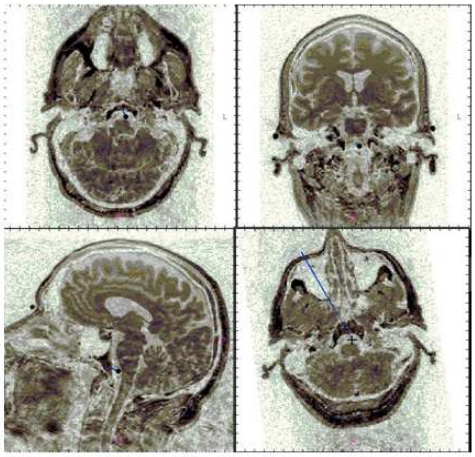


Рис. 5. Сечения ЯМР-томограммы головного мозга с локализованным источником магнитной активности

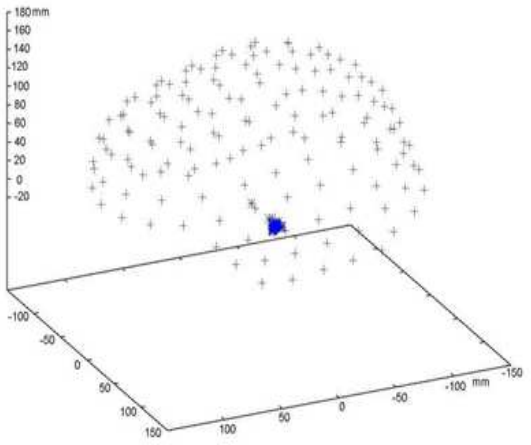


Рис. 6. Локализация дипольных источников для найденных записей МЭГ с патологией на основе спектральных разложений, + — положения диполей эталонных записей, ж — положения диполей найденных патологических записей МЭГ, при этом также локализация патологии дополнительно найденных записей осуществляется в точке (-10, 9, -10)

ности, повышает точность локализации его источников [7].

Визуальное выделение участков в записях МЭГ с патологической активностью при паркинсонизме (например, для случая тремора и синдрома Tinnitus считалось, что при переключении между различного типа сигналами происходит резкое изменение амплитуды и частоты) не всегда эффективно при решении задач локализации источников патологии. Применение дополнительных ограничений, накладываемых определенностью картины пространственного распределения поля, существенно повышает точность локализации [2].

Предварительная спектральная классификация позволяет уточнить моменты времени переключения типа активности. Найденные моменты времени, в которых присутствует аномальная компонента, становятся исходными данными в задаче локализации участков мозга, связанных с рассматриваемой патологией. Представленные выше результаты локализации источников активности и основанный на них метод классификации сигнала находятся в хорошем соответствии с результатами спектрального анализа, описанного в [1, 3, 7].

Обратная задача решалась в предположении наличия одного или двух источников. При этом решение с одним токовым диполем демонстрировало существенно более высокую устойчивость (источник локализовался в компактной области мозга). Причиной плохой локализации для двух токовых диполей является либо некорректность исходного предположения о наличии двух разнесенных источников, отвечающих рассматриваемой патологии, либо вычислительная ошибка оптимизационной процедуры. С большой вероятностью можно считать, что число дипольных источников больше одного в рассматриваемых случаях патологии является артефактом.

Литература

[1] *Derguzov A. V., Makhortykh S. A.* Spectral Analysis and Data Classification in Magnetoencephalography // Pattern Recognition and Image Analysis. — 2006. — Vol. 16, № 3. — 497–505 Pp.

[2] *Дергузов А. В., Махортых С. А.* Комплексная диагностика паркинсонизма по данным магнитной энцефалографии // Электронный журнал Исследовано в России. — 2006. — Т. 65. — С. 646–659.

[3] *Дергузов А. В., Махортых С. А.* Распознавание патологической активности в записях магнитных энцефалограмм при болезни Паркинсона // Электронный журнал Исследовано в России. — 2005. — Т. 149. — С. 1562–1573.

[4] *Джеффрис Г., Свирлс Б.* Методы математической физики. — М.: Мир, 1970. — Вып.3. Т.3. — 344 с.

[5] *Куликова Л. И., Махортых С. А.* Математические операции над двумерными сигналами в базисах сферических гармоник // Электронный журнал Исследовано в России. — 2006. — Т. 60. — С. 598–608.

[6] Курант Р., Гильберт Д. Методы математической физики. — М.-Л.: Гостехиздат, 1951. — Т. 1. — 525 с.

[7] *Махортых С. А., Семечкин С. А.* Локализация источников биомагнитной активности мозга при слуховой стимуляции испытуемого // Бюллетень экспериментальной биологии и медицины. — 2009. — Т. 147, № 4. — С. 477–480.

[8] Устинин М. Н., Махортых С. А., Молчанов А. М. Задачи анализа данных магнитной энцефалографии // Компьютеры и суперкомпьютеры в биологии. — М.: Институт компьютерных технологий, 2002. — 327–349 с.

Нахождение опорных точек в данных магнитной энцефалографии*

Устинин М. Н., Панкратова Н. М.
pnm@impb.ru
Пущино, ИМПБ РАН

Предложена информационная технология интеллектуального анализа данных для обнаружения полезного сигнала на фоне шумов большой амплитуды без использования внешней информации о моментах его возникновения. Для характеристики пространственных паттернов магнитного поля используются базисные функции Карунена-Лоэва. Технология проверена на данных, полученных в контрольном эксперименте, и позволяет выделять патологическую активность на фоне спонтанной активности головного мозга человека.

Задача неинвазивной энцефалографии состоит в том, чтобы узнать, как работает головной мозг человека по магнитному или электрическому полю, регистрируемому на поверхности головы. При этом на экспериментальных данных отражается все многообразие задач, которые головной мозг решает одновременно. Поэтому весьма актуально выделение различных компонент из суммарной электрической активности.

При анализе данных МЭГ производится выделение полезного сигнала на фоне общей спонтанной активности мозга, а затем решается обратная задача: по магнитному полю определяется расположение электрических источников на магниторезонансной томограмме головного мозга испытуемого и делаются научные или диагностические выводы.

Как правило, для выделения сигналов малой амплитуды используются либо внешние проявления патологической активности, например, запись миограммы (электрической активности мышц) при паркинсоническом треморе, либо запись стимула в экспериментах с вызванной активностью. На записи миограммы или стимула определяются опорные точки, по которым полезный сигнал из энцефалограммы выделяется усреднением. Однако остается актуальной задача выделения полезного сигнала при отсутствии такой дополнительной информации. Особенно остро эта проблема возникает при обработке данных со спонтанной активностью, снятых у пациентов с патологией, не имеющей внешних проявлений, которые можно было бы зарегистрировать подобно тремору. Целью методики, изложенной в данной работе, является обнаружение моментов возникновения признаков искомых сигналов.

Методика

Экспериментальные данные представляют собой не просто набор временных рядов, но и имеют четкое распределение в пространстве, поскольку датчики прибора распределены по поверхности головы в виде шлема.

Для эффективного анализа информации о пространственном поведении магнитного поля над поверхностью головы было применено дискретное преобразование Карунена–Лоэва (КЛ-преобразование) [1, 2]. Каждую из функций полученного базиса мы можем использовать в качестве пространственного признака того или иного сигнала. В рамках данной работы мы не пытаемся делать выводы об активности мозга в целом на основании относительных весов тех или иных собственных функций Карунена–Лоэва. Нашей целью является получение оптимального ортогонального базиса для описания пространственной компоненты магнитного поля.

Функции базиса Карунена-Лоэва конкретного эксперимента несут информацию обо всей активности, которая присутствует в данных. Каждая из них может быть использована в качестве пространственного признака определенной активности или полезного сигнала. Выполняя проецирование всего экспериментального массива на выбранную базисную функцию, мы получим двумерную функцию:

$$x_i(l) = \sum_{k=1}^{K} h_k^l v_k^{(i)} ,$$

где $x_i(l)$ — проекция экспериментального массива на $v^{(i)}$ — i-ю базисную функцию КЛ-преобразования. Максимумы этой функции-проекции совпадают с моментами проявления в данных выбранного признака, т. е. определенной структуры магнитного поля. Именно моменты локальных максимумов этой функции мы будем брать за опорные точки для дальнейшего усреднения экспериментального массива. Данный подход был применен в работе [3] к экспериментам с яркой сменой режимов функционирования головного мозга с физиологически нормального на патологический. Тогда в качестве признака искомого сигнала было взято пространственное распределение магнитного поля в момент выраженной патологической активности. Для лучшего очищения сигнала можно использовать дополнительную информацию. С помощью

Работа выполнена при финансовой поддержке РФФИ, проекты № 10-07-00300, № 11-07-00716, № 11-01-00765, № 11-07-00577.

дополнительного признака можно очистить уже полученный, с помощью основного признака, набор опорных точек. Тогда, руководствуясь этим дополнительным знанием о признаках возникновения или отсутствия искомого сигнала, необходимо выбрать соответствующую базисную функцию КЛ-преобразования эксперимента и еще раз найти соответствующую проекцию данных и набор моментов ее локальных максимумов. Выбрав из множества опорных точек, найденных по первой проекции, те, которые присутствуют и во втором наборе, мы очистим это множество от точек, попавших в него случайно. Можно также работать в пространстве нескольких признаков. Для проверки метода мы использовали данные аудиторного эксперимента [4, 5]. По моментам подачи стимула был выделен аудиторный отклик, а решение обратной задачи дало локализацию источников вызванной активности в слуховой зоне коры головного мозга. Таким образом, была получена подробная информация о структуре магнитного поля, возникающего в ответ на аудиторный стимул. Теперь мы можем протестировать предложенный метод и сравнить полученный результат с тем, что получен усреднением по моментам подачи стимула. В качестве признака искомого сигнала было выбрано пространственное распределение магнитного поля, соответствующее максимуму аудиторного отклика. При подробном рассмотрении первых базисных функций Карунена–Лоэва данного эксперимента оказалось, что одна из них качественно соответствует паттерну магнитного поля аудиторного отклика. Обозначим эту функцию как $v^{(audit)}$. Проводя проецирование всего экспериментального массива, т. е. K каналов за все время регистрации, на выбранную пространственную моду КЛ-преобразования, получаем функцию

$$x_{audit}(l) = \sum_{k=1}^{K} h_k^l v_k^{(audit)},$$

локальные максимумы которой совпадают с моментами возникновения аудиторного отклика. Опорные точки определялись уровнем отсечки на функции $x_{(audit)}$. После усреднения сигнала было выполнено решение обратной задачи, которое показало, что источники активности (в нашем случае это отклик на стимул) лежат в слуховой зоне коры головного мозга.

Выводы

Предложенная технология обработки данных магнитной энцефалографии без использования внешней информации об опорных точках, проверена на контрольном эксперименте с известными свойствами отклика на акустический стимул. Для этого был построен базис Карунена–Лоэва экспериментального ряда, и в качестве пространственного признака выбрана мода с распределением магнитного поля, качественно соответствующим паттерну магнитного поля аудиторного отклика. По максимумам проекции экспериментального массива на пространственный признак были найдены опорные точки и по ним проведено усреднение данных. Решение обратной задачи показало совпадение координат источника с найденными по внешнему стимулу [6]. Данный подход был применен к экспериментам, в которых искомая патологическая активность головного мозга, с одной стороны, является спонтанной, а с другой стороны, не порождает внешних физиологических признаков. В этом случае, единственным способом получения опорных точек для выделения такой активности является обнаружение ее пространственно-временных паттернов на фоне общей спонтанной активности. Предлагаемая информационная технология позволяет успешно решать эту задачу. Отметим, что в случае использования только одного пространственного признака, очистка сигнала не является идеальной, и вместе с полезным сигналом определяются шумовые компоненты, впрочем, разделенные во времени с хорошо локализуемой активностью.

Литература

[1] *Karhunen K.* Uber lineare Methoden in der Wahrscheinlichkeitsrechnung // Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys., — 1947. — No. 37.— 1–79 Pp.

[2] *Loeve M.* Probability theory // Vol. II, 4th ed., Graduate Texts in Mathematics, Vol. 46, Springer-Verlag, 1978.

[3] *Панкратова Н. М., Устинин М. Н., Молчанов А. М., Линас Р.* Математическая интерпретация переключений между режимами в сигналах электрической активности головного мозга // Биофизика. 2009, Т. 54, № 5. С. 916–920.

[4] *Устинин М. Н. и др.* Задачи анализа данных магнитной энцефалографии // Компьютеры и суперкомпьютеры в биологии. Под ред. В. Д. Лахно и М. Н. Устинина. Москва–Ижевск: Институт компьютерных исследований, 2002 — С. 327–348.

[5] *Llinas R., Ribary U., Jeanmonod D., Kronberg E., Mitra P.* Thalamocortical dysrhythmia: A neurological and neuropsychiatric syndrome characterized by magnetoencephalography // Proc. of the National Academy of Sciences of the USA. 1999; 96: 15222.

[6] *Устинин М. Н.* Спектрально-аналитические методы обработки данных вычислительного и натурного эксперимента // Дисс. д. ф.-м. н., Пущино, 2004.

Применение комплекснозначных нейронных сетей в задачах распознавания заболеваний органа зрения*

Наумов А. С., Роженцов А. А., Смирнов А. С.
krtmbs@marstu.net
Йошкар-Ола, Марийский государственный технический университет

Предложен подход к распознаванию заболеваний органа зрения, основанный на его диагностике с использованием компьютерной периметрии. Точечные объекты, расположенные на полусфере, для последующей обработки представляются в виде проекций точек полусферы на плоскость. Последующее распознавание группового точечного объекта, представленного в плоскости, осуществляется с применением модели комплекснозначной нейронной сети.

Одной из важных проблем современной офтальмологии является увеличение количества пациентов с патологией органа зрения. Одним из существующих методов диагностики органа зрения является компьютерная периметрия. В рамках данной работы предложен алгоритм распознавания заболеваний органа зрения по данным его периметрических исследований на базе комплекснозначной нейронной сети.

Введение

Современная компьютерная периметрия обладает рядом преимуществ по сравнению с проекционным периметром, основными из которых являются следующие:

— отсутствует субъективный фактор влияния врача на результаты обследования;
— нет необходимости в присутствии врача во время исследования, т. к. врач получает уже готовый результат;
— при каждом повторном обследовании условия тестирования абсолютно одинаковы, позволяя абсолютно точно сравнивать последующие тесты;
— современная компьютерная периметрия дает возможность ранней диагностики и точного мониторинга течения заболевания.

Представление сигнала в комплексной плоскости

Компьютерная периметрия органа зрения представляет собой один из методов исследования периферического зрения, в основе которого лежит проекция сферической поверхности сетчатки на сферическую же и концентрическую с ней внешнюю поверхность, на которой расположены точечные объекты (светодиоды). На рис. 1 представлена схема расположения головы пациента при проведение периметрических исследований.

Полученное в результате проведения периметрических исследований множество видимых пациентом точек сферы можно представить в виде их проекций на плоскость. На рис. 2 представлен пример периметрических данных, представляющих из себя спроецированные на окружность точки сферы для диагностики поля зрения.

Как видно из рис. 2, периметрические данные представляют из себя групповой точечный объект (ГТО) на плоскости. Его можно представить в виде сигнала, отсчетами которого являются комплекс-

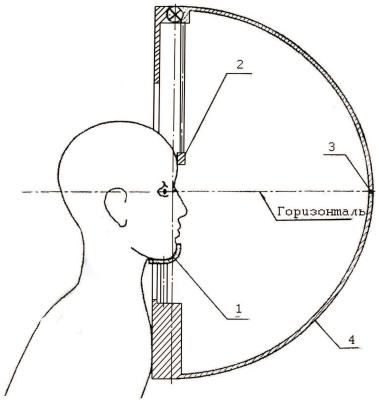


Рис. 1. Схема расположения головы пациента при исследовании поля зрения: *1* — упор для подбородника; *2* — лобный упор; *3* — фиксационная точка; *4* — полусфера

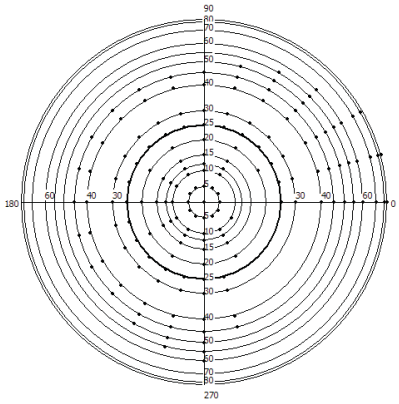


Рис. 2. Квадрантный дефект поля зрения, доходящий до слепого пятна при окклюзии верхней носовой ветви центральной артерии сетчатки

Работа выполнена при финансовой поддержке РФФИ, проект № 10-01-00445-а; и по программе «Развитие научного потенциала высшей школы», проекты 2.1.2/2204 и 2.1.2/10218.

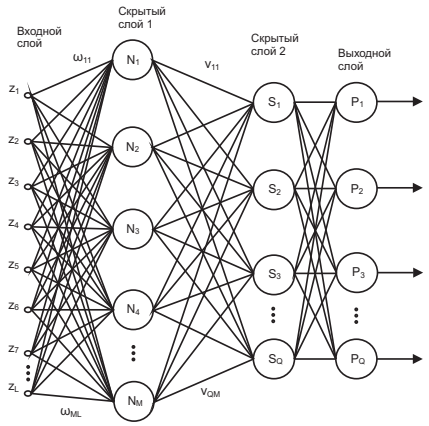


Рис. 3. Комплекснозначная нейронная сеть со структурой L–M–Q–Q

ные числа. Реальная и мнимая части отсчетов сигнала $z_l = x_l + i \times y_l$ соответствуют координатам по оси абсцисс и оси ординат соответственно с точкой начала координат в центре окружности. Ниже представлен комплекснозначный сигнал, соответствующий ГТО на рис. 2, заданный матрицей $\boldsymbol{Z}$, в котором невидимые пациентом точки имеют нулевое значение реальной и мнимой части:

$$\boldsymbol{Z} = \begin{pmatrix} 0,011 + i \times 0 \\ 0 + 0 \times i \\ \dots \\ 0,121 - i \times 0,032 \\ 0,123 - i \times 0,033 \end{pmatrix}.$$

Данный сигнал нормирован по его энергии, что делает его инвариантным к масштабу, в котором представлен ГТО.

Модель комплекснозначной нейронной сети

В рамках данной работы для распознавания ГТО, отсчетами которого являются комплексные числа, построена модель комплекснозначной нейронной сети (КНС), состоящая из входного, двух скрытых и выходного слоев. Схема модели КНС представлена на рис. 3.

На входы сети поступают отсчеты $z_l = x_l + i \times \times y_l$ распознаваемого сигнала. Далее со входного слоя полученный сигнал поступает на каждый из нейронов первого скрытого слоя. Состоянием m-го нейрона первого скрытого слоя N_m является взвешенная сумма всех отсчетов входного сигнала $\boldsymbol{Z}$, которая вычисляется по (1):

$$N_m = N_m^R + i \times N_m^I = \sum_{l=1}^{L} (w_{jl}^R \times x_l - w_{jl}^I \times y_l) + \\ + i \times \sum_{l=1}^{L} (w_{jl}^I \times x_l + w_{jl}^R \times y_l), \quad (1)$$

где $\boldsymbol{w} = \boldsymbol{w^R} + i \times \boldsymbol{w^I} = (w_0, w_1, \dots, w_{L-1}) \in \mathbb{C}^L$ — весовой вектор комплекснозначного нейрона, в котором $w_l = w_l^R + i \times w_l^I$, w_l^R и $w_l^I \in \mathbb{R}^1$, $i = \sqrt{-1}$ и $l = 0, \dots, L-1$; N_j — состояние j-го нейрона первого скрытого слоя.

Выходное значение H_j j-го нейрона первого скрытого слоя определяется по (2). В данном случае использовалась функция активации гиперболического тангенса. Для нейронов первого скрытого слоя использовался реально мнимый тип функции активации, при котором реальная и мнимая части выходного значения нейрона есть отдельные функции от реальной и мнимой частей его состояния, т. е. $H_j = f(N_j^R) + i \times f(N_j^I)$:

$$H_m = f(N_j^R) + i \times f(N_j^I) = \\ = \frac{e^{N_j^R} - e^{-N_j^R}}{e^{N_j^R} + e^{-N_j^R}} + i \times \frac{e^{N_j^I} - e^{-N_j^I}}{e^{N_j^I} + e^{-N_j^I}}, \quad (2)$$

где N_j^R и N_j^I — соответственно реальная и мнимая компоненты состояния j-го нейрона, H_j — выходное значение j-го нейрона.

Полученный на выходе первого скрытого слоя сигнал поступает на нейроны второго скрытого слоя. Аналогично первому скрытому слою состояние q-го нейрона второго скрытого слоя S_q является взвешенной суммой отсчетов сигнала на выходе первого скрытого слоя и вычисляется по (3):

$$S_q = S_q^R + i \times S_q^I = \sum_{j=1}^{J} (v_{qj}^R \times H_j^R - v_{qj}^I \times H_j^I) + \\ + i \times \sum_{j=1}^{J} (v_{qj}^I \times H_j^R - v_{qj}^R \times H_j^I), \quad (3)$$

где $v_{qj} = v_{qj}^R + i \times v_{qj}^I$ — весовые коэффициенты q-го комплекснозначного нейрона второго скрытого слоя, $i = \sqrt{-1}$, $q = 1, \dots, Q$, Q — количество нейронов второго скрытого слоя, $m = j, \dots, J$, J — количество нейронов первого скрытого слоя.

Количество нейронов второго скрытого слоя Q равно количеству эталонов, которым инициализирована нейронная сеть. Функцией активации нейронов второго скрытого слоя является Гауссова функция активации, которая вычисляется по (4):

$$G_q = f(S_q) = \frac{1}{\sqrt{2 * \pi} * \sigma} \times e^{-((S_q^{R\,2} + S_q^{I\,2} - M)^2)/(2 \times \sigma^2)}, \quad (4)$$

где S_q^R и S_q^I — реальная и мнимая компоненты состояния q-го нейрона второго скрытого слоя, σ — среднеквадратическое отклонение, M — математическое ожидание функции. При равенстве значений энергии состояния нейрона $||E|| = S_q^{R\,2} + S_q^{I\,2}$ и математического ожидания M на выходе данного

нейрона будет пиковое значение. Каждому нейрону второго скрытого слоя соответствует свой эталон из базы эталонов для инициализации сети, т. е. нейроны данного слоя имеют разные значения математических ожиданий M, которые присваиваются им при инициализации нейронной сети.

Инициализация нейронной сети заключается в поочередной подаче на ее входной слой всех Q эталонных сигналов, вычислении энергии состояния q-го нейрона второго скрытого слоя, соответствующего q-му эталонному сигналу и присвоении полученного значения энергии значению математического ожидания данного нейрона.

Полученный на выходе 2-го слоя сигнал поступает на нейроны выходного слоя. Активационной функцией для данных нейронов является функция (5) единичной суммы:

$$P_q = \frac{G_q}{\sum_{q=1}^{Q}(G_q)}\,. \qquad (5)$$

Данная функция активации используется в вероятностных нейронных сетях. Как видно из формулы, значение на выходе каждого из нейронов показывает вероятность принадлежности поданного на вход сигнала к одному из сигналов, которым ранее была инициализирована нейронная сеть. Значением на выходе q-го нейрона выходного слоя является вероятность сигнала на выходе одного из нейронов предыдущего слоя по отношению к суммарному значению вероятностей сигналов всех нейронов предыдущего слоя.

Выводы

В рамках данной работы рассмотрена вероятностная комплекснозначная нейронная сеть для распознавания заболеваний органа зрения по данным периметрических обследований. Рассмотрена возможность представления периметрических данных в комплексной форме с последующей обработкой полученного сигнала с применением комплекснозначной нейронной сети.

Преимуществами представленного метода распознавания заболеваний органа зрения являются:

— возможность обучения системы новым эталонным сигналам, которые соответствуют заболеваниям, неизвестным системе на момент распознавания данного сигнала;
— устойчивость к выпадению (в случае, если пациент случайно не зафиксировал видимую точку) или появлению внешних (при случайной фиксации точки, находящейся за полем зрения) точек.

Литература

[1] *Zhang H., Zhang C., Wu W.* Convergence of Batch Split-Complex Backpropagation Algorithm for Complex-Valued Neural Networks // Discrete Dynamics in Nature and Society. — 2009. — Vol. 2009.

[2] *Фурман Я. А., Кревецкий А. В., Роженцов А. А. и др.* Введение в контурный анализ и его приложения к обработке изображений и сигналов.. — М.: Физматлит, 2002. — 592 с.

[3] *Naumov A., Rozhentsov A.* Contour signals recognition by probabilistic neural network // PRIA-10, 2010. — Vol. 2, No. 1.

[4] *Наумов А. С., Роженцов А. А.* Обработка контурных изображений с применением комплекснозначных нейронных сетей // Оптико-электронные приборы и устройства в системах распознавания образов, обработки изображений и символьной информации. Сборник материалов IX Международной конференции «Распознавание — 2010», Курск, 2010. — С. 61–63.

Вероятностный подход к поиску поведенческих паттернов*

Вишневский В. В., Ветров Д. П.
valera.vishnevskiy@yandex.ru, vetrovd@yandex.ru
Москва, Москвоский Государственный Университет им. М. В. Ломносова

В данной работе предложен новый метод для поиска скрытых закономерностей (паттернов) в последовательностях событий, основанный на вероятностном представлении P-Паттернов (probabilistic pattern) в дискретных последовательностях событий. Поиск производится снизу вверх: сначала находятся простые закономерности, потом, путем их соединения, образуются более сложные паттерны. Рассматривается применение данного алгоритма для анализа поведения мышей. Найденные паттерны используются для классификации животных. Проведено сравнение реализованного алгоритма с существующими аналогами, показавшее, что предложенный метод более устойчив к шуму в исходных данных.

Задача поиска закономерностей (стереотипов, паттернов, шаблонов — здесь синонимы) в поведении животных и людей крайне важна в современной нейробиологии и когнитивных науках. Выделив характерные паттерны, можно, например, делать выводы о сложности поведения различных особей, определять изменения в поведении наблюдаемых процессов, другими словами, решив задачу поиска паттернов, мы можем определенным образом *измерять* поведение особи, или группы особей.

В данной работе мы будем рассматривать временны́е паттерны в «структурном», или «эффектном» описании [3, с. 57]. Исходными данными будет размеченное поведение особи, т. е. последовательность пар «момент времени», «поведенческий акт». Неформально можно сказать, что интересующие нас паттерны — это упорядоченная последовательность поведенческих актов, следующие один за другим через *относительно инвариантные* временны́е интервалы. Причем этот паттерн должен повторяться в исходных данных *достаточно* часто.

Несмотря на то, что описанные выше паттерны широко распространены в описании поведения, стандартные статистические методы не подходят для их поиска: эти методы либо не учитывают всю сложность паттернов (например, периодические орбиты [4]), либо оперируют такими понятиями как циклы, волны, тренды, что невозможно напрямую использовать для поиска интересующих нас паттернов.

На сегодняшний день, для анализа таких поведенческих закономерностей наиболее широкое распространение получил метод поиска T-Паттернов (temporal patterns), предложенный в 2000 г. Магнусом Магнуссоном в [2].

Основные определения. Понятие T-Паттерна

Пусть время наблюдения разбито на N_t интервалов. В каждый момент *периода наблюдения* $[1, N_t]$ может произойти некоторое событие $\mathbf{e}$ (*действие, поведенческий акт, event*) [1] из множества допустимых событий $\mathcal{E}$. Соответственно, каждому типу события сопоставляется множество моментов времени $TS(\mathbf{e})$, когда это событие имело место:

$$TS(\mathbf{e}) = \{t_1^{\mathbf{e}}, \ldots, t_{N_{\mathbf{e}}}^{\mathbf{e}}\},$$
$$\mathbf{e} \in \mathcal{E}, \quad 0 \leqslant t_i^{\mathbf{e}} \leqslant N_t \quad (i = 1, \ldots, N_t),$$

здесь $N_{\mathbf{e}}$ — количество появлений события $\mathbf{e}$ в данных.

Понятие T-Паттерна включает в себя определение модели связи последовательных событий и способа их появления в данных (т. е. определения, что данная структура не случайна, а является *закономерностью*). Каждое событие паттерна определяется фиксированным временны́м интервалом, в течение которого это событие должно присутствовать после предыдущего события. Другими словами, расстояния между событиями моделируются равномерным распределением.

Основной недостатк метода поиска T-Паттернов заключается, во-первых, в том, что само определение T-Паттерна не позволяет ему иметь пропуски событий. По этой причине метод становится крайне чувствителен к шуму в исходных данных, из-за чего можно пропустить информативные длинные и сложные паттерны. Во-вторых, полученные T-Паттерны сильно специфичны особи, в поведении которой они были найдены.

Вероятностная модель паттерна. P-Паттерны

Определение 1. *Нечетким паттерном, или P-Паттерном* $\mathbf{P}$ *длины* $N_{\mathbf{P}}$ *назовем упорядоченную*

[1] Чаще всего понимается, что в этот момент времени имеет место *начало* действия

Работа выполнена при финансовой поддержке РФФИ, проект №08-01-00405; гранта Президента РФ, МК3827.2010.9; и федеральной целевой программы «Научные и научно-педагогические кадры инновационной России» на 2009–2013 годы, контракт №П1265.
Авторы статьи выражают благодарность членам «Лаборатории Нейробиологии Памяти» при Институте Нормальной Физиологии П. К. Анохина, возглавляемой членом-корреспондентом РАН К.В. Анохиным. Отдельно благодарим Ирину Зарайскую, предоставившую нам экспериментальные данные по поведению.

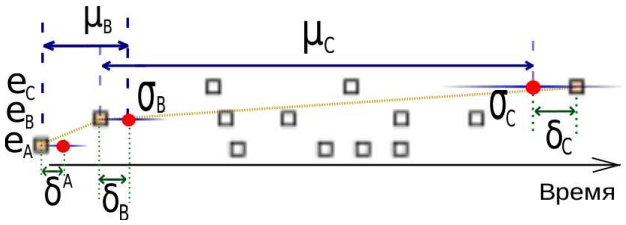


Рис. 1. Вхождение нечеткого паттерна $\mathbf{P} = \mathbf{e}_A[\mu_A, \sigma_A]\mathbf{e}_B[\mu_B, \sigma_B]\mathbf{e}_C[\mu_C, \sigma_C]$ в данные. Маркеры-кружки соответствуют ожидаемым позициям событиям, закрашенные маркеры-квадраты соответствуют позиции реальной позиции соответствующего события в данных

последовательность событий $\mathbf{e}_i, (i = 1, \ldots, N_\mathbf{P})$, *где каждое событие паттерна характеризуется смещением и разбросом от предыдущего события. Будем записывать паттерн* $\mathbf{P}$ *в следующем виде:*

$$\mathbf{P} = [\,\mu_1, \sigma_1]\mathbf{e}_1[\,\mu_2, \sigma_2]\mathbf{e}_2, \ldots, [\,\mu_{N_\mathbf{P}}, \sigma_{N_\mathbf{P}}]\mathbf{e}_{N_\mathbf{P}}, \quad \mu_1 = 0.$$

Здесь μ_i *и* σ_i *— метематическое ожидание и корень из дисперсии нормального распределения, моделирующего величину времени, прошедшего между событиями.*

Представление Р-Паттерна иллюстрировано на рис. 1.

Далее чтобы иметь возможность обрабатывать пропуски в Р-Паттернах, введем понятие *функции потерь*, которая определяет «штраф» за пропуск m событий в паттерне длины $N_\mathbf{P}$ следующим образом:

$$f_{LOSS}(m, N) = \begin{cases} \exp\left(-\frac{\lambda m}{N_\mathbf{P}}\right), & m < N, \\ 0, & m = N. \end{cases}$$

Здесь λ является структурным параметром, определяющим уровень «нечеткости» паттернов. Если этот параметр велик, то мы, по сути, запрещаем реализациям паттерна иметь пропуски. Если выставить этот параметр слишком малым, то будут обнаруживаться паттерны, не разу полностью не встречающиеся в данных, т. е. закономерности могут быть найдены даже в случайных данных.

Определение 2. *Правдоподобие паттерна* $\mathbf{P}$ *— это функция, определенная в каждый момент времени наблюдения* ε $(\varepsilon = 1, \ldots, N_t)$ *следующим образом:*

$$L_\mathbf{P}(\varepsilon) = f_{LOSS}(N_-, N_\mathbf{P}) \prod_{i=1}^{N_\mathbf{P}} \left(\frac{1}{\sqrt{2\pi}\,\sigma_i}\right) \times \times \prod_{i \in \mathcal{N}_+} \exp\left(-\frac{\delta_i^2}{2\sigma_i^2}\right), \quad (1)$$

где δ_i *— расстояния от ожидаемой позиции события в Р-Паттерне до ближайшего события в данных (более наглядно см. рис. 1). То есть:*

$$\delta_i = \min_{x \,\in\, TS(\mathbf{e}_i)} \Bigg| \underbrace{\varepsilon + \sum_{j=1}^{i-1}(\mu_j + \delta_j) + \mu_i}_{\text{ожидаемая позиция события}} - x \Bigg|,$$

здесь, если событие было пропущено, то соответствующее $\delta_i = 0$. *Далее,* N_- *— количество пропущенных событий в паттерне, а* $\mathcal{N}_+$ *— множество индексов присутствующих в паттерне событий. Событие считается пропущенным, если* $\exp\left((-\delta_i^2)/(2\sigma_i^2)\right) < \exp\left((-\lambda)/(N_\mathbf{P})\right)$.

По сути, правдоподобие показывает насколько можно быть уверенным, что данный Р-Паттерн начинается в определенный момент времени ε.

Заметим, что правдоподобие Р-Паттерна может быть отсчитано с конца или с m-го события паттерна.

Утверждение 1. *Математическое ожидание функции правдоподобия* (1) *в момент времени* ε, *при условии, что в данный момент времени имеет место начало модельного Р-Паттерна* $\mathbf{P}$ *вычисляется следующим образом:*

$$\mathsf{E}\,[L_\mathbf{P}(\varepsilon)] = \frac{1}{(2\sqrt{\pi})^{N_\mathbf{P}}\,\sigma_1, \ldots, \sigma_{N_\mathbf{P}}}.$$

Доказательство приведено в [1].

Здесь используется тот факт, что межточечные расстояния между событиями в модельном Р-Паттерне распределены по нормальному закону:

$$\delta_i \sim \mathcal{N}(0, \sigma_i), \quad (i = 1, \ldots, N_\mathbf{P}).$$

Теперь мы можем считать, что Р-Паттерн $\mathbf{P}$ имеет место быть только в следующие моменты времени:

$$t\colon L_\mathbf{P}(t) \geqslant \gamma \mathsf{E}\,[L_\mathbf{P}(\varepsilon)], \quad (2)$$

где γ — заданная константа.

Конструирование Р-Паттернов. Рассмотрим пару Р-Паттернов $\mathbf{P}_L$ (левый) и $\mathbf{P}_R$ (правый). Пусть $\{\alpha_i\}_{i=1,\ldots,N_L}$ и $\{\beta_j\}_{j=1,\ldots,N_R}$ — значения правдоподобия соответствующих Р-Паттернов в моменты времени, соответствующие паттерны имели вхождения (2). Важно, что правдоподобие левого Р-Паттерна отсчитывается с конца, т. к. мы ищем связь между концом левого паттерна и началом правого. Также пусть, $\{t_{L,i}\}_{i=1,\ldots,N_L}$ и $\{t_{R,j}\}_{j=1,\ldots,N_R}$ — моменты времени, когда эти Р-Паттерны имели место в смысле (2). N_L и N_R — количество вхождений паттернов $\mathbf{P}_L$ и $\mathbf{P}_R$, соответственно. Определим множество межточечных расстояний:

$$\rho = \{t_{R,j} - t_{L,i} \mid t_{R,j} \geqslant t_{L,i}\}\,.$$

Для каждого расстояния из этого множества введем соответствующий вес $w_l = \ln(1 + \alpha_i\beta_j)$, $l = 1, \ldots, M$, где $M = |\rho|$.

Рассмотрим гипотезу H_0, что моменты времени и веса вхождения P-Паттернов распределены независимо и равномерно на всем наблюдаемом промежутке. Тогда плотность распределение введенных выше межточечных расстояний $\{t\}$ имеет следующий вид [1]:

$$p_{LR}(t) = \begin{cases} (N_t - t)\frac{2}{N_t^2}\,, & t \in [0, N_t]\,; \\ 0, & x \notin [0, N_t]\,. \end{cases} \qquad (3)$$

Введем статистическую модель связи между паттернами(проверяемые параметры связи μ и σ фиксированы):

$$g_{\mu,\sigma}(t_i) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(t_i - \mu)^2}{2\sigma^2}\right).$$

Рассмотрим следующую сумму:

$$k = \sum_{i=1}^{M} w_i g_{\mu,\sigma}(t_i)\,, \qquad (4)$$

где $t_i \sim p_{LR}$.

Теорема 1. *При выполнении H_0 статистика k распределена по нормальному закону:*

$$k \sim \mathcal{N}\left(\mu_*, \sigma_*^2\right),$$

где

$$\mu_* \approx M\,\mathsf{E}[w]\,\frac{2}{N_t}\left(1 - \frac{\mu}{N_t}\right);$$

$$\sigma_*^2 \approx \frac{M}{N_t}\left(1 - \frac{\mu}{N_t}\right)\left[\left(\frac{1}{\sqrt{\pi}\,\sigma} - \frac{\mu}{N_t}\left(1 - \frac{\mu}{N_t}\right)\right) \times \right.$$
$$\left. \times \left((\mathsf{E}[w])^2 + \mathsf{D}[w]\right) + \frac{4\mathsf{D}[w]}{N_t}\left(1 - \frac{\mu}{N_t}\right)\right], \qquad (5)$$

здесь $\mathsf{E}[w]$ и $\mathsf{D}[w]$ — выборочное среднее и дисперсия весов, соответственно.

Доказательство приведено в [1].

Замечание 1. Приближенные формулы (5) дают удовлетворительный результат для задачи поиска паттернов, однако точные формулы для значений μ_* и σ_* можно найти в [1].

Теперь для различных пар μ и σ можно вычислить статистику (4), сравнить ее с α-квантилью распределения (5). Если гипотеза H_0 о «случайности» данных будет отвергнута односторонним критерием, то считается, что соответствующие P-Паттерны образуют новый паттерн с параметрами μ и σ. Если существует несколько пар μ и σ, для которых отвергается гипотеза H_0, то для конструирования P-Паттернов берутся непересекающиеся[2] параметры соответствующие максимальным значениям k.

Более подробно о процессе формирования нового P-Паттерна можно найти в [1].

Редукция P-Паттернов. Для удаления [1] паттернов-дубликатов и неполных копий анализируется коэффициент корреляции функций правдоподобия. Пусть $\overrightarrow{L_{\mathbf{P},i}}$ — вектор-столбец значений функции правдоподобия, отсчитанной от i-го события во всех моментах времени наблюдения.

$$\operatorname{corr}\left(\overrightarrow{L_1}, \overrightarrow{L_2}\right) = \frac{\overrightarrow{L_1}^{\top}\overrightarrow{L_2}}{\sqrt{\overrightarrow{L_1}^{\top}\overrightarrow{L_1}}\sqrt{\overrightarrow{L_2}^{\top}\overrightarrow{L_2}}} \in [0, 1]$$

— коэффициент корреляции между двумя P-Паттернами.

Проверяются все пары паттернов, если все поведенческие акты, присутствующие в паттерне $\mathbf{P}_L$, также присутствуют в $\mathbf{P}_R$ с учетом порядка, и

$$\exists m\colon\ \operatorname{corr}\left(\overrightarrow{L_{P_L,1}}, \overrightarrow{L_{P_R,m}}\right) > \nu\,,$$

тогда паттерн $\mathbf{P}_L$ удаляется из множества найденных паттернов.

Алгоритм поиска P-Паттернов

1. Инициализировать текущее множество P-Паттернов событиями (паттерны длины 1).
2. Для всевозможных пар P-Паттернов из текущего множества провести процедуру *конструирования*.
3. Провести процедуру редукции паттернов.
4. Если текущее множество паттернов изменилось, перейти к п. 2.

Параметры алгоритма и способы их настройки описаны в [1].

Сложность предложенного алгоритма — $O(n^3)$, где n — общее количество событий во временном ряде. В [1] представлена параллельная реализация данного метода на GPU, что позволило применять алгоритм поиска P-Паттернов на реальных данных.

Эксперименты на реальных данных

Описанный ниже эксперимент демонстрирует способ применения предложенного метода на реальных поведенческих данных. Целью эксперимента является анализ того, как влияет отсутствие гиппокампа на поведение. Гиппокамп — один из древнейших отделов головного мозга млекопитающих, его функции связывают с механизмами работы памяти, обучением, пространственной навигацией.

Особи были разделены на 5 групп:

[2] $[\mu' - 3\sigma', \mu' + 3\sigma'] \cap [\mu'' - 3\sigma'', \mu'' + 3\sigma''] = \varnothing$

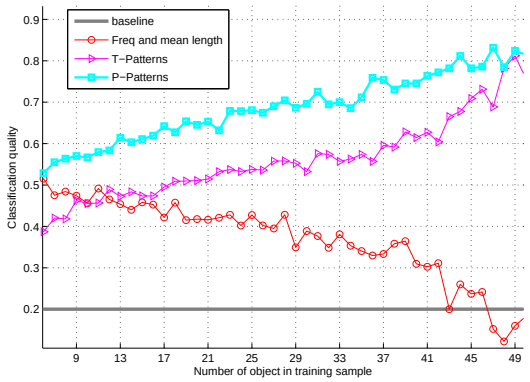


Рис. 2. Качество — средняя доля правильных классификаций. По горизонтали откладывалось количество объектов в обучении (качество усреднено по ста случайным разбиениям). Классификации по P- и T-паттернам производилась с помощью SVM. Классификации на основе частот и длин актов производилась с помощью решающих лесов

1) Контрольная группа, содержащая разметку поведения здоровых мышей — 12 особей.
2) Гиппокампальная группа. Гиппокамп этих животных разрушали путем введения в эту структуру лидокаина, растворенного в искусственной спинномозговой жидкости (2 мкл. 4% раствора) — 12 особей.
3) Шумовая группа, с параметрами частоты и продолжительности актов первой (контрольной) группы — 12 «особей».
4) Шумовая группа, с параметрами частоты и продолжительности актов второй (гиппокампальной) группы — 12 «особей».
5) Искусственные данные, содержащие один модельный P-Паттерн — 7 «особей».

Поведение каждой особи было представлено временным рядом длины ~ 12 минут, всего было 24 детектируемых поведенческих актов, более подробно условия эксперимента описаны в [1]. Требовалось решить задачу классификации особи по поведению. Далее данные разбивались на обучение (N_l объектов) и контроль (N_c объектов). $N_l + N_c = 55$. Каждая особь описывалась вектором длины N_l, i-е значение которого равняется количеству паттернов данной особи, также найденное в поведении i-й особи из обучающей выборки.

На рис. 2 видно, что предложенный метод поиска P-Паттернов дает заметно лучшее качество классификации, чем метод поиска T-Паттернов. Наивная классификация на основе описания частот и средней продолжительности актов неприменима в данном эксперименте из-за двух шумовых классов.

Характерные паттерны присущие определенному классу также могут быть выделены в экспериментальных данных. Неформально, паттерн является характерным для заданного класса, если он присутствует в поведении многих особей этого класса и редко выявляется в поведении животных из других классов. Пример характерного для контрольной группы P-паттерна (в квадратных скобках указаны смещения и дисперсии в секундах между событиями):

— «вылизывание гениталий» [2,5; 6,2];
— «вылизывание ладоней» [1,2; 7,9];
— «умывание головы с ушами» [0,4; 5,7];
— «вылизывание задних конечностей» [2,6; 7,9];
— «умывание носа».

Выводы

Представленный метод решает поставленные перед ним задачи и производит качественный поиск закономерностей как в синтетических временны́х рядах, так и в реальных поведенческих данных. В открытом доступе свободная, документированная, параллельная реализация представленного метода.

Главным преимуществом метода поиска P-Паттернов является их вариабельность: если на то есть предпосылки, то P-Паттерны, найденные в поведении одной особи, будут также найдены в поведении другой особи. Данный факт позволяет описывать поведение на основе найденных паттернов и использовать стандартные алгоритмы машинного обучения для решения, например, задач классификации кластеризации, или восстановления регрессии.

Литература

[1] *Вишневский В. В.* Параллельная реализация метода поиска закономерностей в последовательностях событий. — Дипломная работа. ВМиК МГУ, 2011.

[2] *Magnusson M. S.* Discovering hidden time patterns in behavior: T-patterns and their detection. — Behavior Research Methods, Instruments, Computers, 2000.

[3] *Martin P., Bateson P.* Measuring Behaviour: An Introductory Guide. — Cambridge University Press, second edition, 1993.

[4] *Stoop R., Arthur B.* Periodic orbit analysis demonstrates genetic constraints, variability, and switching in Drosophila courtship behavior. — Chaos, 2008. — Vol.18/2.

Использование эпитомного подхода в задаче автоматической сегментации гистологических изображений срезов мозга мыши*

Елшин Д. А.[1]***, Кропотов Д. А.***[2]
dennis179@rambler.ru, dmitry.kropotov@gmail.com
[1] Москва, МГУ им. М. В. Ломоносова; [2] Москва, ВЦ РАН

Рассматривается задача автоматической сегментации изображений на основе подхода минимизации энергии марковского случайного поля. Функционал энергии здесь состоит из слагаемого, отвечающего за текстурные особенности каждого класса, а также набора стандартных слагаемых, включающих цветовые статистики классов, потенциалы положения и потенциалы Поттса. В работе предлагается вводить текстурный потенциал, основанный на предложенном ранее т.н. эпитомном подходе к описанию изображений. Эффективность предложенного текстурного потенциала демонстрируется на модельной задаче сегментации текстур, а также на сложной практической задаче сегментации гистологических изображений срезов головного мозга мыши на анатомические структуры.

Одним из перспективных направлений в современной нейробиологии для понимания когнитивных процессов, происходящих в головном мозге животных, является анализ экспрессии генов в мозге. Стандартной процедурой здесь является декапитация головного мозга животного, замораживание, нарезка на тонкие слои и последующая окраска на гистологию (общую структуру) и экспрессию определенных генов. Проведение статистического анализа уровня экспрессии генов в различных анатомических структурах мозга требует, в частности, процедуру автоматической сегментации гистологических изображений срезов мозга на анатомические структуры.

В данной работе рассматривается подход к подобной сегментации для данных из Алленовского атласа мозга мыши [4]. Этот атлас содержит 132 коронарных среза мозга, для каждого из которых доступно его гистологическое изображение (см. рис. 2, a, e) и ручная экспертная разметка на анатомические структуры (см. рис. 2, b, f). Особенностью данных гистологических изображений является отсутствие видимых границ между многими парами соседних анатомических структур. Поэтому основной задачей при построении процедуры сегментации таких изображений является выделение текстурных особенностей каждой анатомической структуры.

В данной работе для текстурного описания изображения предлагается адаптировать т.н. эпитомный подход [1]. В этом подходе для обучающего изображения строится его уменьшенная версия (эпитом), которая содержит текстурные особенности исходного изображения. Далее предлагается ввести вероятностное распределение на классы каждого пикселя эпитома, полученное из наблюдаемой разметки на классы исходного обучающего изображения. Затем тестовое изображение проектируется на известный эпитом, а вероятностное распределение на классы отдельных пикселей тестового изображения вычисляется путем обратного преобразования вероятностного распределения для эпитома. Итоговая сегментация тестового изображения получается путем минимизации энергии соответствующего марковского случайного поля, в энергии которого к текстурному потенциалу, полученному на основе эпитома, добавляются стандартные компоненты, включающие потенциалы цвета, положения и Поттса [3].

Процедура сегментации

Данная процедура сегментации заимствована из работы [3]. Задача состоит в следующем: данному изображению $X = \{x_i\}_{i=1}^R$ поставить в соответствие маску $M = \{m_i\}_{i=1}^R$:

$$x_i \to m_i,\ m_i \in \{1, \ldots, K\},$$

где K — число классов. Для её решения используется модель марковского случайного поля. Изображение рассматривается как марковская сеть $(V, \mathcal{E})$, где каждому пикселю x_i изображения ставится в соответствие вершина сети $i \in V$, а ребра $(i, j) \in \mathcal{E}$ определяют соседние пикселы. Здесь и далее используется четырехсвязная система соседства. Переменная в вершине i принимает значения $m_i \in \in \{1, \ldots, K\}$. Тогда энергия разметки M при известном изображении X определяется следующим образом:

$$E(M \mid X) = \sum_{i \in V} \psi_i(m_i, x_i) + \\ + \sum_{(i,j) \in \mathcal{E}} \varphi_{ij}(m_i, m_j, x_i, x_j),$$

где $\psi_i(m_i, x_i)$ — унарные потенциалы, отвечающие за степень принадлежности пиксела i классу m_i, а $\varphi_{ij}(m_i, m_j, x_i, x_j)$ — бинарные потенциалы, штрафующие ситуацию разных классов у соседних пикселей.

Оптимальная разметка $\hat{M}$ получается как

$$\hat{M} = arg \min_M E(M \mid X).$$

Работа выполнена при финансовой поддержке РФФИ, проект № 09-01-00409.

Нетрудно видеть, что минимум энергии марковской сети будет однозначно соответствовать разбиению на классы пикселей исходного изображения.

В данной энергии в качестве унарных потенциалов ψ_i использовалась линейная комбинация текстурных потенциалов на основе эпитомов, потенциалов цвета и потенциалов положения. В качестве бинарных потенциалов φ_{ij} использовалась модель Поттса. Энергия $E(M \mid X)$ минимизировалась алгоритмом α-расширения [2]. Опишем подробнее каждый из потенциалов.

Бинарные потенциалы Здесь используются простейшие потенциалы, штрафующие длину границы между классами:

$$\varphi(m_i, m_j, x_i, x_j) = \begin{cases} 1, & \text{если } m_i \neq m_j\,; \\ 0, & \text{если } m_i = m_j\,. \end{cases}$$

Потенциалы положения. Каждое обучающее изображение разбивается на P непересекающихся областей $\{L_l\}_{l=1}^{P}$, в каждой области вычисляется N_l^m, равное числу пикселей класса m в области l, и $N_l = \sum_m N_l^m$. Тогда, обозначив за L^i область, содержащую пиксель i, имеем:

$$\mathsf{P}(m \mid i) = \frac{N_{m,L^i}}{N_{L^i}}\,.$$

Непосредственно значение $\psi_i(m_i, x_i)$ вычисляется как $-\log \mathsf{P}(m_i \mid i)$.

Потенциалы цвета. Цветовая модель представляет собой гауссовскую смесь для каждого класса. Восстановление смеси осуществляется стандартным EM-алгоритмом. Условное распределение на цвет пикселя задаётся как

$$\mathsf{P}(x \mid m) = \sum_{k=1}^{n(m)} w_k^m \mathcal{N}(x \mid \mu_k^m, \sigma_k^m)\,, \qquad \sum_k w_k^m = 1, \quad w_k^m \geqslant 0,$$

где $n(m)$ — число компонент смеси, а $\mathcal{N}(x \mid \mu_k^m, \sigma_k^m)$ — значение плотности нормального распределения с мат.ожиданием μ_k^m и дисперсией σ_k^m. Значение $\psi_i(m_i, x_i)$ вычисляется как $-\log \mathsf{P}(x_i \mid m_i)$.

Эпитомный подход

Обучение эпитома по изображению. Эпитомом [1] для изображения размера $N \times M$ является изображение меньшего размера $N_e \times M_e$, в каждом пикселе которого помимо цветовой характеристики μ_t хранится её дисперсия φ_t. Примеры эпитомов (компонента μ) для изображений на рис. 1, *a* и 2, *a* показаны, соответственно, на рис. 1, *c* и 2, *c*. Исходное изображение разбивается на P блоков $\{Z_k\}_{k=1}^{P}$. Обозначим через S_k множество индексов пикселей блока Z_k, через $z_{k,i}$ — пиксели блока Z_k, а через R — множество индексов элементов эпитома e. Рассмотрим взаимно однозначное преобразование T_k для блока Z_k между S_k и некоторым подмножеством R. В дальнейшем будем рассматривать только квадратные блоки Z_k размера $D \times D$ и преобразования, сопоставляющие блоку Z_k блок такого же размера в эпитоме. Тогда при известном эпитоме $e = (\mu, \varphi)$ и преобразовании T_k блок Z_k генерируется путём копирования соответствующего блока эпитома $\{\mu\}_k$ с добавлением гауссовского шума с дисперсией $\{\varphi\}_k$:

$$\mathsf{P}(Z_k \mid T_k, e) = \prod_{i \in S_k} \mathcal{N}(z_{k,i} \mid \mu_{T_k(i)}, \varphi_{T_k(i)})\,.$$

Полагая, что блоки генерируются независимо, получаем совместное распределение:

$$\mathsf{P}(\{Z_k, T_k\}_{k=1}^{P}, e) = p(e) \prod_{k=1}^{P} p(T_k) p(Z_k \mid T_k, e)\,.$$

Априорные распределения $p(e)$ и $p(T_k)$ в дальнейшем будем полагать равномерными.

Поиск наилучшего эпитома e для набора блоков изображения $\{Z_k\}_{k=1}^{P}$ осуществляется с помощью максимизации неполного правдоподобия:

$$P(\{Z_k\}_{k=1}^{P} \mid e) \to \max_e\,.$$

Параметры T_k здесь неизвестны и выступают в данной модели в качестве скрытых переменных. Рассмотрим семейство преобразований $\{T_k^j\}_{j=1}^{F}$, ставящих Z_k в соответствие все возможные блоки такого же размера в e. Тогда неполное правдоподобие представляется в виде:

$$P(\{Z_k\}_{k=1}^{P} \mid e) = \prod_{k=1}^{P} \sum_{j=1}^{F} P(Z_k, T_k^j \mid e) \to \max_e\,.$$

Эта задача может быть решена с помощью EM-алгоритма. E-шаг соответствует вычислению распределения на T_k^j при фиксированном эпитоме e:

$$P(T_k^j \mid Z_k, e) \propto P(Z_k \mid T_k^j, e) = \prod_{i \in S_k} \mathcal{N}(z_{k,i} \mid \mu_{T_k^j(i)}, \varphi_{T_k^j(i)})\,. \tag{1}$$

В дальнейшем для краткости будем обозначать распределение $P(T_k^j \mid Z_k, e)$ через $q(T_k^j)$. На M-шаге происходит пересчёт параметров эпитома e:

$$\mu_t = \frac{\sum_{k=1}^{P} \sum_{i \in S_k} \sum_{j, T_k^j(i)=t} q(T_k^j) z_{k,i}}{\sum_{k=1}^{P} \sum_{i \in S_k} \sum_{j, T_k^j(i)=t} q(T_k^j)}\,; \tag{2}$$

$$\varphi_t = \frac{\sum_{k=1}^{P} \sum_{i \in S_k} \sum_{j, T_k^j(i)=t} q(T_k^j)(z_{k,i} - \mu_t)^2}{\sum_{k=1}^{P} \sum_{i \in S_k} \sum_{j, T_k^j(i)=t} q(T_k^j)}\,. \tag{3}$$

Вероятностное распределение на классы для пикселов эпитома. Помимо элементов (μ, φ) введём дополнительный параметр ψ, который будет отвечать за вероятность принадлежности данного элемента эпитома к каждому из классов $\{1, \ldots, K\}$.

Обозначим через M_k разметку на классы блока Z_k и зафиксируем класс l. Тогда $\psi_{t,l}$ определяется как сумма вероятностей преобразования класса l в элемент эпитома t:

$$\psi_{t,l} = \frac{\sum_{k=1}^{P} \sum_{i \in S_k, M_k(i)=l} \sum_{j, T_k^j(i)=t} q(T_k^j)}{\sum_{k=1}^{P} \sum_{i \in S_k} \sum_{j, T_k^j(i)=t} q(T_k^j)}.$$

Заметим, что подсчёт характеристики ψ производится после окончания обучения эпитома e.

Текстурные потенциалы. Тестовое изображение произвольным образом разбивается на блоки $\{Z_k\}_{k=1}^P$ и проектируется на эпитом $e = (\mu, \varphi)$ путём подсчёта вероятностей $q(T_k^j)$. Проекция изображения может быть восстановлена как решение задачи $P(\{Z_k\}_{k=1}^P \,|\, e) \to \max\limits_Z$:

$$x_i = \frac{\sum_{k=1, i \in S_k}^{P} \sum_j q(T_k^j) \mu_{T_k^j} / \varphi_{T_k^j}}{\sum_{k=1, i \in S_k}^{P} \sum_j q(T_k^j) / \varphi_{T_k^j}},$$

где x_i — пиксель изображения, а суммирование ведётся по всем блокам Z_k, перекрывающимся по данному пикселю. При работе алгоритма с группой эпитомов данные проекции позволяют выбрать наилучший для данного изображения эпитом путем минимизации расстояния между исходным изображением и его проекцией. Вероятность принадлежности пикселя изображения i к классу $m_i \in \in \{1, \ldots, K\}$ вычисляется как

$$\mathsf{P}(x_i \,|\, m_i) = \frac{\sum_{k=1, i \in S_k}^{P} \sum_j \psi_{T_k^j(i), m_i} q(T_k^j)}{\sum_{k=1, i \in S_k}^{P} \sum_j q(T_k^j)}, \qquad (4)$$

а непосредственно текстурный потенциал $\varphi_i(m_i, x_i)$ вычисляется как $-\log \mathsf{P}(x_i \,|\, m_i)$.

Ускорение вычислений

Реализация EM-алгоритма обучения эпитома e по формулам (1)–(3) приводит к алгоритму со значительным временем работы. Например, для изображений размера 300×200, эпитома размера 60×60 и блоков размера 5×5 время работы может достигать нескольких часов. Поэтому для практического использования эпитомного подхода необходима эффективная реализация EM-алгоритма (1)–(3).

Рассмотрим ускорение вычислений на E-шаге (1). Для этого рассмотрим один блок исходного изображения Z и эпитом e. На E-шаге требуется вычислить матрицу C размера $(N_e - D)\times(M_e - D)$ с элементами

$$C_{ts} = \sum_{p,q=1}^{D} \log \mathcal{N}(z_{pq} | \mu_{t+p,s+q}, \varphi_{t+p,s+q}).$$

Здесь (t, s) – позиция в эпитоме, а z_{pq} – интенсивность цвета в блоке Z в позиции (p, q). Заметим, что в отличие от предыдущих разделов, здесь используется двухмерная индексация пикселей изображений. Вычисление матрицы C можно представить через операцию свертки:

$$C_{ts} = -\frac{D^2}{2} \log 2\pi - \frac{1}{2} \sum_{p,q=1}^{D} \left[\frac{(z_{pq} - \mu_{t+p,s+q})^2}{\varphi_{t+p,s+q}} + \log \varphi_{t+p,s+q} \right] =$$

$$= -\frac{D^2}{2} \log 2\pi - \frac{1}{2} \sum_{p,q=1}^{D} \left[z_{pq}^2 \frac{1}{\varphi_{t+p,s+q}^2} - 2 z_{pq} \frac{\mu_{t+p,s+q}}{\varphi_{t+p,s+q}^2} \frac{\mu_{t+p,s+q}^2}{\varphi_{t+p,s+q}^2} + \log \varphi_{t+p,s+q} \right].$$

Отсюда

$$C = -\frac{D^2}{2} \log 2\pi - \frac{1}{2} \Big[H_1 * \mathrm{rot180}(Z^2) - 2 H_2 * \mathrm{rot180}(Z) + H_3 * I \Big].$$

Здесь через $*$ обозначена операция двухмерной свертки, $\mathrm{rot180}(Z)$ – операция поворота против часовой стрелки матрицы Z на 180°, $(Z^2)_{pq} = z_{pq}^2$, $(H_1)_{ts} = 1/\varphi_{ts}$, $(H_2)_{ts} = \mu_{ts}/\varphi_{ts}$, $(H_3)_{ts} = \mu_{ts}^2/\varphi_{ts} + \log \varphi_{ts}$, I – матрица из единиц.

Идея быстрого вычисления операции вида $H * Z$ состоит в использовании быстрого преобразования Фурье. Известно, что $F(H * Z) = F(H)F(Z)$, где $F(\cdot)$ — преобразование Фурье. Тогда $H * Z = F^{-1}(F(H)F(Z))$, где F^{-1} — обратное преобразование Фурье. Вычисление прямого и обратного дискретного преобразования Фурье имеет эффективную реализацию во многих программных системах. В частности, в MatLab для этого используется специальная библиотека IPP для процессоров Intel.

Если рассматривать задачу вычисления матриц C для всех блоков исходного изображения, то здесь можно также сэкономить в вычислениях, если заранее рассчитать $F(H_1), F(H_2)$ и $H_3 * I$.

Эксперименты

Эксперименты с предложенным текстурным потенциалом (4) проводились сначала на модельной задаче сегментации текстур. Исходные данные представляли собой аппликации из нескольких текстур размера 100×100 (см. рис. 1, *a*, *e*). Для обучающего изображения строился эпитом размера 40×40 с блоками размера 5×5. Данные и результаты эксперимента показаны на рис. 1. Как видно из этих результатов, использование эпитомного текстурного потенциала позволяет значительно

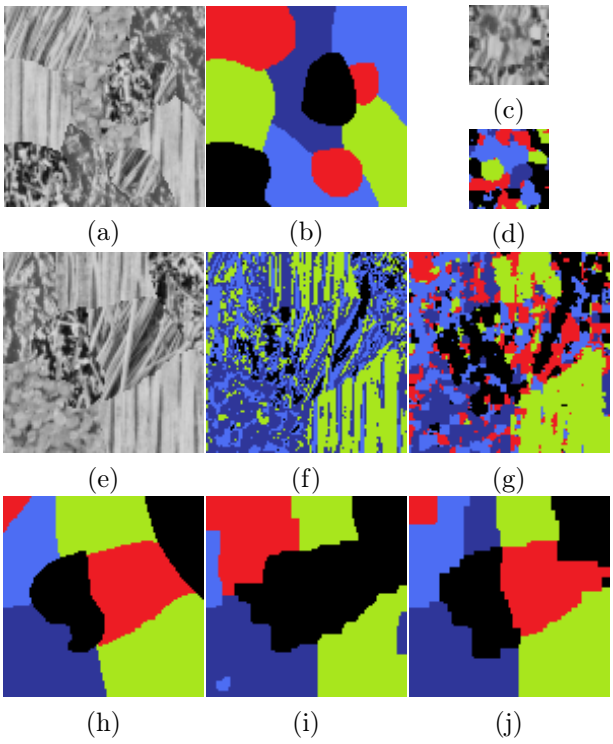


Рис. 1. Данные и результаты в модельном эксперименте: (a) — обучающее изображение; (b) — его истинная сегментация; (c) — эпитом для изображения (a); (d) — наиболее вероятные классы для эпитома (c); (e) — тестовое изображение; (f) — сегментация (e) с помощью цветовых статистик; (g) — сегментация (e) с помощью эпитома (c)–(d); (h) — истинная сегментация (e); (i) — сегментация (e) с помощью цветовых статистик и потенциалов Поттса (ошибка 35,07%), (j) — сегментация (e) с помощью эпитома (c)–(d) и потенциалов Поттса (ошибка 13,12%)

повысить качество сегментации по сравнению с использованием цветовых статистик.

Эксперименты по сегментации изображений срезов мозга из Алленовского атласа проводились по следующей схеме. Изображения разбивались на пары соседних срезов, в каждой паре выбиралось обучающее и тестовое изображение. По обучающему изображению настраивался эпитом, который затем использовался для вычисления текстурного потенциала для тестового изображения. Потенциалы положения вычислялись по четырем соседним изображениями слева и справа от тестового при разбиении на блоки размера 10×10. Потенциалы цвета настраивались по всем имеющимся изображениям. Результаты эксперимента для 10-го и 11-го срезов показаны на рис. 2. Как видно из этого рисунка, текстурный потенциал действительно позволяет снизить ошибку сегментации и выделить некоторые структуры более четко (например, черная и светло-желтая структура на рисунке). В целом, по всем изображениям Алленовского атласа удается снизить ошибку сегментации на $5\% - 10\%$.

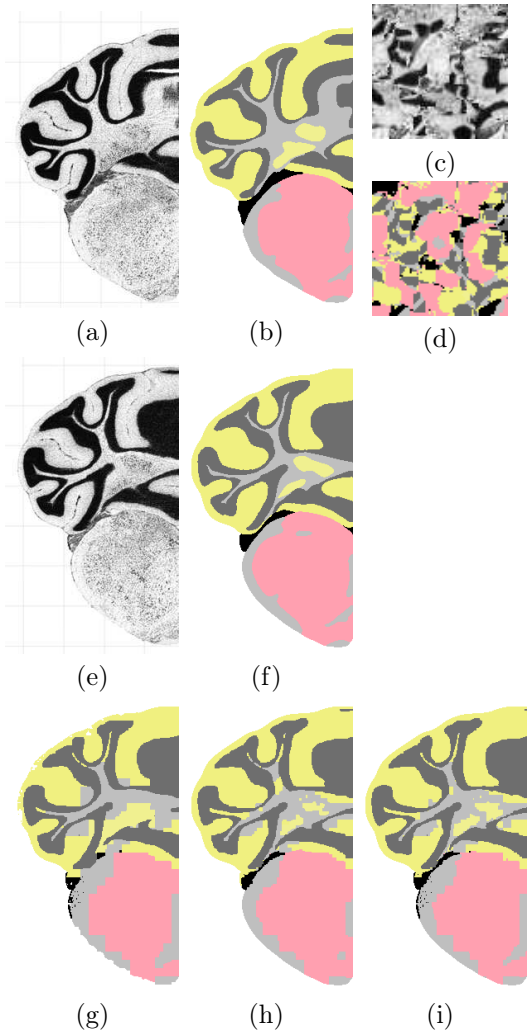


Рис. 2. Данные и результаты сегментации изображения среза: (a) — обучающее изображение; (b) — его истинная сегментация; (c) — эпитом для изображения (a); (d) — наиболее вероятные классы для эпитома (c); (e) — тестовое изображение; (f) — его истинная сегментация; (g) — сегментация (e) с помощью цветовых статистик, потенциалов положения и Поттса (ошибка 28,47%); (h) — сегментация (e) с помощью эпитома (c)–(d), потенциалов положения и Поттса (ошибка 22,41%); (i) — сегментация (e) с помощью цветовых статистик, эпитома (c)–(d), потенциалов положения и Поттса (ошибка 21,32%)

Литература

[1] *Jojic N., Frey B., Kannan A.* Epitomic analysis of appearance and shape // ICCV, 2003.

[2] *Boykov Y., Veksler O., Zabih R.* Fast Approximate Energy Minimization via Graph Cuts // TPAMI, 2001.

[3] *Shotton J., Winn J., Rother C., Criminisi A.* TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context // IJCV, 2009.

[4] `http://mouse.brain-map.org/atlas/index.html` — Allen Brain Atlas.

Распознавание спектров флуоресценции бактерий и полиароматических углеводородов

Суханов А. Я., Креков Г. М.
say@iao.ru
Томск, Институт оптики атмосферы им. В. Е. Зуева СО РАН

Данная работа посвящена решению задачи классификации спектров флуоресценции патогенных, органических аэрозолей на основе метода нейронных сетей. Спектры флуоресценции, полученные лидарной системой, подвержены помехам различного рода, что требует привлечения соответствующих помехоустойчивых алгоритмов классификации. Для решения поставленной задачи была проведена модификация алгоритма обучения нейронной сети, что позволило повысить вероятность распознавания зашумленных спектров.

Введение

Методы распознавания образов или образов сигналов получили широкое распространение во многих направлениях человеческой деятельности. В частности, в спектрометрии при распознавании спектров различных веществ, а также при дистанционном зондировании атмосферы с помощью лидаров. Одной из важнейших задач является обнаружение патогенных аэрозолей, болезнетворных бактерий в окружающей среде, что необходимо для сохранения должного качества жизни людей и сохранения их здоровья. Обнаружение подобных веществ желательно осуществлять с помощью бесконтактных методов без участия человека, это, в свою очередь, требует применения и развития методологий, техники и алгоритмов для автоматизированного дистанционного зондирования и идентификации данных веществ. Одной из возможностей является обнаружение подобных вредоносных аэрозольных образований по их спектрам флуоресценции. Данная задача не является тривиальной, т. к. многие спектры флуоресценции похожи друг на друга, при этом сигнал, проходя толщу атмосферы, может искажаться, также различные приборные шумы накладываются на сигнал, а при наличии нескольких патогенных флуорофоров спектры могут перекрываться, что необходимо учитывать при идентификации.

Для распознавания и идентификации спектров флуорофоров предлагается использовать нейронные сети. Достоинство применения нейронных сетей в возможности параллельной обработки информации, технической и аппаратной реализации, а также простом и понятном принципе решения прикладных задач.

Решение задачи классификации с помощью нейронной сети

При реализации метода нейронных сетей важно выбрать метод обучения, число слоев, тип нейронов, а также создать обучающую выборку. Создание обучающей выборки — отдельная и сложная задача, необходимо чтобы она была полной, наиболее информативной и не сильно большой по объему.

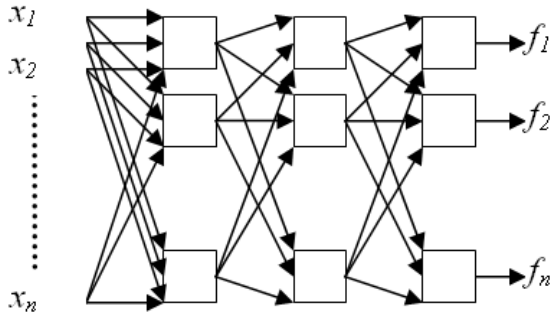


Рис. 1. Трехслойная нейронная сеть

Для решения задачи распознавания флуорофоров по их спектрам флуоресценции была выбрана сеть типа многослойный персептрон (см. рис. 1). Сеть Хопфилда и Хэмминга обладают тем недостатком, что их необходимо обучать на каких-то конкретных эталонных векторах или спектрах, что не всегда удобно, т. к. не позволяет учитывать искажение таких спектров под влиянием различных атмосферных условий, шумов или вклад в полученный спектр различных флуорофоров.

Математически функционирование трехслойной нейронной сети можно описать следующим выражением:

$$f_i = \varphi_{3,i}\Big(\sum_{j1=0}^{L-1} w^3_{i,j1}\varphi_{2,j1}\Big(\sum_{j2=0}^{M-1} w^2_{j1,j2}\times \times \varphi_{1,j2}\Big(\sum_{j3=0}^{N-1} w^1_{j2,j3}x_{j3}\Big)\Big)\Big),$$

где $\varphi_{i,j}$ — активационная функция j-го нейрона i-го слоя; $w^n_{i,j}$ — j-й весовой коэффициент i-го нейрона n-го слоя; f_i — выходные значения нейронной сети, соответствующие выходу i-го нейрона последнего слоя; x_i — входные значения нейронной сети, их число равно числу входов каждого входного нейрона нейронной сети.

На вход такой сети подается некоторый нормированный спектр:

$$I(\lambda_i) = \frac{2I_\gamma(\lambda_i)}{\Delta\lambda \sum_{i=1}^{N-1} I_\gamma(\lambda_i) + I_\gamma(\lambda_{i+1})},$$

где $2I_\gamma(\lambda_i)$ — дискретные значения полученного спектра. Сеть имеет число выходов в соответствии с числом распознаваемых видов флуорофоров, при этом значение близкое к единице на соответствующем выходе указывает на один из флуорофоров, значение близкое к нулю указывает на отсутствие флуорофора. Сеть имеет три слоя, первый и второй слой имеют количество нейронов, совпадающих с размерностью входного вектора, каждый такой нейрон имеет такое же число входов. На основе имеющихся спектров проводится Монте–Карло моделирование переноса излучения для различных атмосферных условий, при этом считается, что вклад в спектр других флуорофоров по сравнениию с основным, не превосходит 20%. Таким образом, мы получаем различные искаженные спектры для различных ситуаций, при этом при создании примеров учитывается возможные случайные шумы различной амплитуды. Каждому такому спектру ставится в соответствие выходной вектор, с единицей на выходе, соответствующий основному флуорофору и со значениями менее 0,2 для остальных флуорофоров в соответствии с их вкладами в общий спектр. Создается до тысячи обучающих примеров и далее проводится обучение такой сети с применением алгоритма обратного распространения ошибки. Для случая, когда распознается четыре бактерии и пять опасных аэрозолей при проверке на тестовой выборке, получены следующие результаты (см. рис. 2). Считается, что конкретный аэрозольный компонент идентифицирован успешно, если значение выхода максимально среди остальных значений выходов и превосходит значение 0,5, номер выхода сравнивается с номером в тестовом примере. Ясно, что уже при величине шума более 15% и, это в лучшем случае, сеть не распознает нужного флуорофора, что неприемлемо для реальной ситуации и накладывает жесткие условия на технику зондирования.

Если в обучающей выборке не учитывать шум в обучающих примерах, это приводит еще к более плачевной ситуации, например, вероятность распознавания Флуорантена при 25% зашумлении падает от 0,3 до 0,1 (см. рис. 3).

Исследования показали, что при уменьшении числа распознаваемых спектров вероятность распознавания при зашумлении растет, но не столь значительно (см. рис. 4).

В качестве метода обучения было решено использовать комбинированный алгоритм линейной регрессии и обратного распространения ошибки.

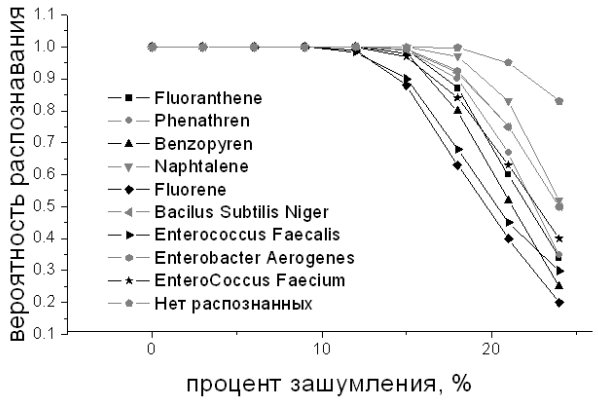


Рис. 2. Вероятность распознавания спектра флуоресценции при зашумлении

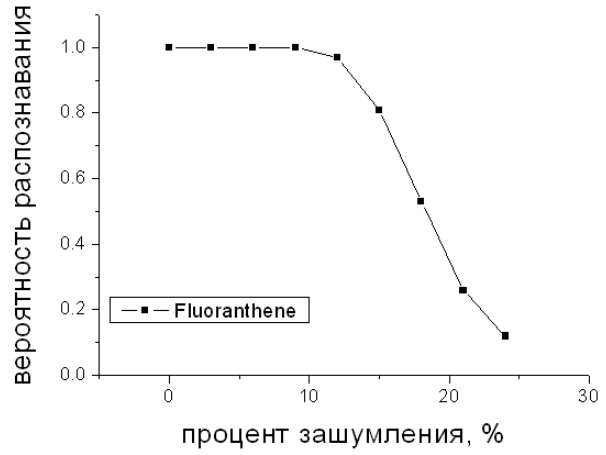


Рис. 3. Вероятность распознавания Флуорантена при зашумлении

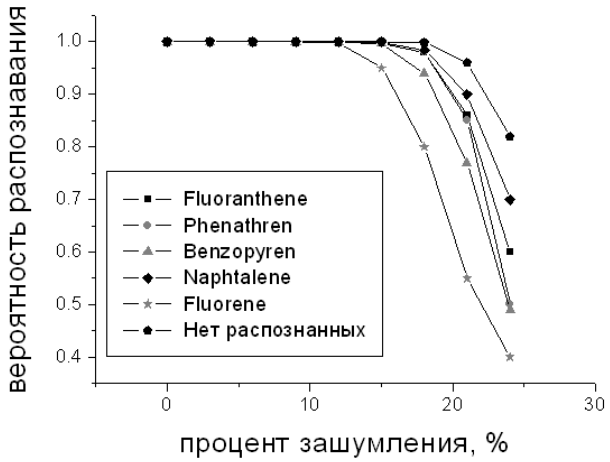


Рис. 4. Процент распознавания пяти флуорофоров в зависимости от зашумления

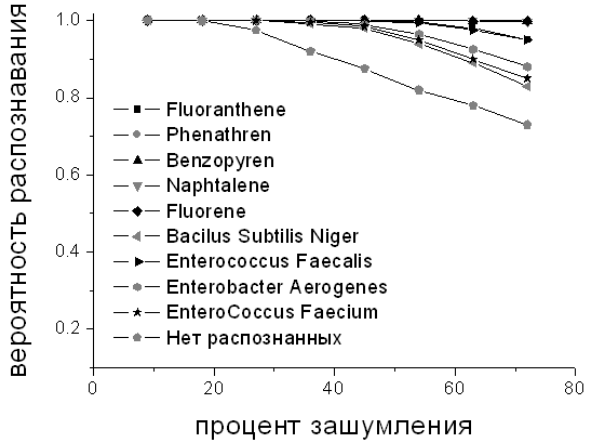


Рис. 5. Распознавание флуорофоров с помощью нейронной сети с комбинированных обучением

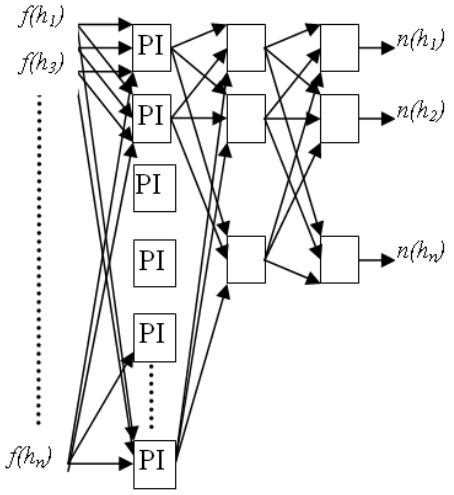


Рис. 6. Пример комбинированной сети с нейронами первого слоя обученными методом псевдо-обратных матриц

При этом первый слой нейронной сети обучался с помощью метода линейной регрессии, а далее нейронная сеть дообучался методом обратного распространения ошибки. Результаты распознавания, полученные с помощью такой сети, представлены на рис. 5.

Нейронная сеть в данном случае представляет собой трехслойную нейронную сеть (см. рис. 6), с предварительно обученными нейронами первого слоя.

Алгоритм линейной регрессии основан на матрично-векторных преобразованиях [1]. Если вход и выход нейронной сети представить в виде векторов, а весовые коэффициенты в виде матрицы, то связь входа нейронной сети с ее выходом можно представить в виде следующего выражения:

$$Y = F^k(W^k F^{k-1}(...F^2(W^2 F^1(W^1 X)))),$$

где Y — матрица выходных значений, F^k — вектор активационных функций k-го слоя, W^k — матрица весовых коэффициентов k-го слоя, X — матрица входных значений.

Во входных X и выходных матрицах D обучающие примеры расположены в столбцах матрицы, каждая строка матрицы весовых коэффициентов принадлежит одному нейрону. Алгоритм заключается в выражении матриц коэффициентов через входную, выходную матрицы и активационные функции. Для однослойной сети данные выражения можно получить следующим образом: $D = F(WX)$, $F^{-1}(D) = WX$, где F^{-1} — функция, обратная к активационной функции.

Например, для активационной функции $\varphi(g) = \arctan(g)$ обратная активационная функция это $g(\varphi) = \tan(\varphi)$, а для $\varphi(g) = 1/(1 + e^{-g})$ это $g(\varphi) = -\log(1/(\varphi) - 1)$.

Далее умножаем справа на транспонированную матрицу входных примеров,

$$F^{-1}(D)X^T = WXX^T,$$
$$F^{-1}(D)X^T(XX^T)^{-1} = WXX^T(XX^T)^{-1},$$
$$F^{-1}X^T(XX^T)^{-1} = W.$$

Таким образом, матрица весовых коэффициентов определяется по формуле

$$W = F^{-1}(D)X^T(XX^T)^{-1}. \qquad (1)$$

Так как размерность входных и выходных векторов для нашего случая распознавания спектров различается, а первый слой нейронной сети имеет одинаковое число входов и одинаковое число выходов, искусственно размерность выходного вектора увеличивается до размерности входного вектора, при этом новым введенным элементам выходного вектора присваиваются случайные значения от -1 до 1. В качестве активационной функции для нейронов первого слоя взята функция $\varphi(g) = \sin(g)$, для второго и третьего слоев $\varphi(g) = 1/(1 + e^{-g})$.

Положим элементы матрицы X входных элементов равными $x_{i,j} = I_x^{(j)}(\lambda_i)$, $i = 1, \ldots, N$, $j = 1, \ldots, L$, вектор выходных элементов $c_{i,k}$, $i = 1, \ldots, N$, $k = 1, \ldots, M$, где N — число спектральных входов, M — число выходов сети соответствующих числу распознаваемых спектров, L — число обучающих векторов, $I_x^{(j)}$ — j-й обучающий вектор. Элементы матрицы выходных элементов D для обучения методом линейной регрессии положим равными $d_{i,k} = c_{i,k}$ при $k \leqslant M$, $d_{i,k} = r - 0{,}5$ при $M < k \leqslant N$, r — случайное значение, распределенное в соответствии с равномерной плотностью вероятности.

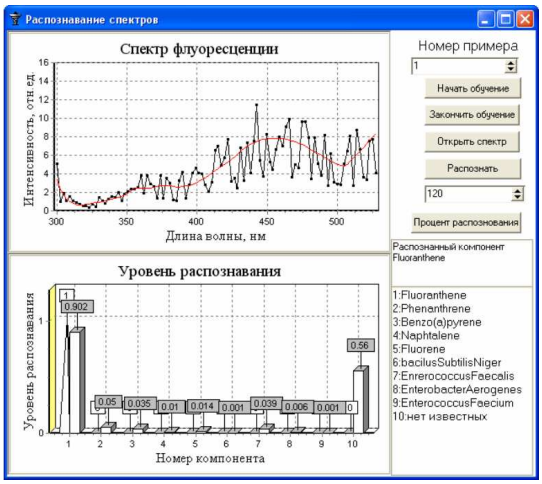


Рис. 7. Пример комбинированной сети с нейронами первого слоя обученными методом псевдо-обратных матриц

Используя алгоритм линейной регрессии, определяются коэффициенты нейронов первого слоя по формуле (4). После установки коэффициентов первого слоя начинается обучение нейронной сети по методу обратного распространения ошибки [2–4].

Таким образом, применение комбинированного алгоритма позволяет повысить помехоустойчивость нейронной сети по сравнению с обучением методом обратного распространения ошибки, также обеспечивается обучение многослойной сети с различным числом входов, чего не обеспечивает алгоритм линейной регрессии, т. к. при его применении возникает необходимость получать обратные матрицы для плохо-обусловленных матриц, что влечет расхождение алгоритма обучения.

Реализованная программа для распознавания и пример спектра флуоресценции приведены на рис. 7.

Выводы

Предложенный алгоритм позволяет повысить эффективность классификации спектров флуоресценции в условиях помех при применении нейронных сетей в сравнении со стандартными алгоритмами обучения, например, обратным распространением ошибки, это существенно продвигает возможности использовании лидаров для идентификации веществ в среде, в том числе и при достаточно высоких уровнях случайных помех.

Литература

[1] *Суханов Д. Я., Суханов А. Я.* Метод итерационной настройки многослойной нейронной сети на основе метода наименьших квадратов // Доклады Томского государственного университета систем управления и радиоэлектроники. — 2004. — Т. 10, № 2. — С. 111–116.

[2] *Rumelhart D. E., Hinton G. E. and Williams R. J.* Learning internal representations by error propagation // Parallel Distributed Processing: Explorations in the Microstructure of Cognition, 1986. Vol. 1. — Pp. 318–362.

[3] *Werbos P. J.* The Roots of Backpropagation // NY: John Wiley Sons. Includes Werbos's 1974 Harvard Ph.D. thesis, Beyond Regression, 1974/1994.

[4] *Розеблатт Ф.* Принципы нейродинамики. — М.: Мир, 1964.

Распознавание скрытой периодичности в кодирующих последовательностях ДНК

Чалей М. Б.[1]***, Кутыркин В. А.***[2]
maramaria@yandex.ru, vkutyrkin@yandex.ru
[1]Пущино, Учреждение Российской Академии наук Институт математических проблем биологии РАН
[2]Москва, Московский Государственный Технический Университет им. Н. Э. Баумана

Методы распознавания нового типа скрытой периодичности в ДНК — скрытой профильности — используются для исследования кодирующих последовательностей ДНК человека из базы данных KEGG (Kyoto Encyclopedia of Genes and Genomes). Эти методы позволили выделить двухуровневую организацию кодирования генетической информации. Результаты исследования могут быть использованы для создания математических методов выявления кодирующих районов в ДНК.

Носителем наследственной информации любого живого организма является биополимерная молекула дезоксирибонуклеиновой кислоты (ДНК), образованная мономерными звеньями — нуклеотидами четырёх типов, называемых аденин, гуанин, цитозин и тимин (часто обозначаемых буквами a, g, t, c). Таким образом, молекула ДНК может быть представима в виде уникальной последовательности букв (нуклеотидов) исходного четырёхбуквенного алфавита.

В нуклеотидных последовательностях ДНК существуют различные уровни регулярной структурной организации: длина шага двойной спирали ДНК в 10–11 пар нуклеотидов (нукл.), характерная длина $\backsim$ 200 пар нукл. для фрагмента ДНК в нуклеосоме, характерные длины $\backsim 2 \times 10^4 - 10^5$ нукл., выявляемые на более высоких уровнях квазирегулярной упаковки молекулы ДНК [1]. Такие особенности, могут быть обусловлены некоторой закономерностью чередования нуклеотидов в исходной последовательности ДНК. Поэтому исследования корреляций в последовательностях ДНК имеют большое значение для понимания основы известных и выявления новых структурных особенностей ДНК. В графиках различных функций, используемых для представления корреляций в кодирующих районах ДНК, наблюдаются регулярно повторяющиеся пики с шагом в три, в соответствии с триплетной природой генетического кода. Отсюда возникло представление о триплетной периодичности кодирующих районов.

До настоящего времени наиболее распространённое описание скрытой периодичности опиралось на понятие размытых тандемных повторов [2]. Результатом распознавания такой периодичности является текстовой «консенсус-паттерн», который служит оценкой исходного паттерна повтора. Если повреждение копий паттерна заменами и вставками/делециями нуклеотидов составляет не более 30%, то консенсус-паттерн признаётся достоверным. Несмотря на то, что в кодирующих районах встречаются короткие тандемные повторы три- и гекса-нуклеотидов, вывести достоверный консенсус-паттерн тандемного повтора на всей длине кодирующего района, как правило, не удаётся.

Слабое предпочтение нуклеотида конкретного типа в фиксированной позиции триплетов кодирующего района способствует появлению в спектре Фурье доминирующего пика спектральной плотности на частоте 0,33, соответствующей периоду в 3 нукл. [3], но оно не является фундаментальной причиной такой картины спектра. Оказалось что, чем больше дисперсия распределения конкретного нуклеотида, даже не доминирующего по позициям триплета, тем больший вклад он вносит в амплитуду спектральной плотности на частоте, соответствующей периоду 3 нукл. [4]. Поэтому появление в Фурье-спектре пика на частоте 0,33 может быть обусловлено всего лишь неоднородностью распределения нуклеотидов по позициям триплетов.

Применение методов Фурье-анализа для оценки длины периода скрытой периодичности стало традиционным [5]. Для этой цели использовались и другие статистические методы [6–8], в основе которых лежит вычисление меры неоднородности в распределении нуклеотидов по позициям периода. На практике в последовательности, не являющейся размытым тандемным повтором, может наблюдаться высокий показатель неоднородности и Фурье-спектр с доминирующим пиком на некоторой частоте. В этом случае использование термина «скрытая периодичность» некорректно, пока не выявлен паттерн периодичности какого-либо нового типа.

В работе [9] для распознавания нового типа скрытой периодичности, расширяющего понятие размытого тандемного повтора и названного профильной периодичностью (профильностью), был предложен оригинальный спектрально-статистический подходе (2C подход). Методы, распознающие размытые тандемные повторы, не могут быть использованы для выявления этого нового типа скрытой периодичности.

Ранее [9] было показано, что предложенный 2C подход позволяет выявить два уровня организации кодирования генетической информации: регу-

лярную неоднородность в распределении нуклеотидов по позициям кодонов и скрытую профильность. Второй уровень кодирования может коррелировать со структурной организацией кодируемых белков [9]. Непосредственное выявление такой организации является достаточно сложной задачей, поскольку цель поиска априори неизвестна.

Как будет показано далее, Фурье-анализ не позволяет выделить второй уровень организации кодирования (скрытую профильность). Фурье-спектры кодирующих районов ДНК были получены с помощью программ Фурье-анализа [10, 11] на WEB-сервере [12].

В настоящей работе с помощью 2С подхода проведён количественный структурный анализ кодирующих последовательностей ДНК человека из базы данных KEGG-54.1 (Kyoto Encyclopedia of Genes and Genomes) [13] в сравнении с последовательностями интронов (некодирующих фрагментов генов) человека из базы данных EID (The Exon-Intron Database) [14, 15]. Результаты сравнительного анализа могут быть использованы для создания математических методов выявления кодирующих районов в ДНК.

Спектрально-статистический подход к выявлению скрытой профильной периодичности в ДНК

Модель профильной периодичности (профильности) описывается схемой независимых испытаний, в которой в каждых L последовательных испытаниях реализуется заданный список распределений вероятностей для четырёх букв алфавита ДНК. Такую схему испытаний можно рассматривать как последовательное сцепление L полиномиальных схем с $K = 4$ элементарными исходами. В этой схеме каждое распределение из заданного списка отождествляется со случайной буквой. Следовательно, паттерн скрытой профильности может быть описан с помощью конечной случайной строки, состоящей из независимых случайных букв с соответствующими распределениями вероятностей для текстовых букв из алфавита ДНК [9].

Пусть $Chr(\mathrm{p})$ — случайная буква со столбцом частот $\mathrm{p} = (p^1, \ldots, p^K)^{\mathsf{T}}$, где p^i — вероятность буквы a_i алфавита $A = \langle a_1, \ldots, a_K \rangle$. Если $p^i = 1$, то $Chr(\mathrm{p}) = a_i$ — буква алфавита A. Введём случайную строку $Str = Str_n(\pi) = Chr(\mathrm{p}_1), \ldots, Chr(\mathrm{p}_n)$ из n независимых случайных букв, где $\pi = (\mathrm{p}_1, \ldots, \mathrm{p}_n) = (\pi^i_j)^K_n$.

Число L из диапазона $1, \ldots, L_{\max}$, где $L_{\max} \backsim n/(5K)$, называется тест-периодом строки Str. Если $L = nm$, то $Str = Str_n(\pi) = Str_L(\pi_1), \ldots, Str_L(\pi_m))$ — разложение строки Str на подстроки длины L. Тогда Π_{Str} — профильно-матричный спектр строки Str, значение которого $\Pi_{Str}(L) = 1/m \sum_{i=1}^{m} \pi_i$ — профильная матрица строки Str. Если $\pi_1 = \cdots = \pi_m = \pi$, то строка Str называется L-профильной строкой со случайным паттерном периодичности $Str_L(\pi)$. В этом случае для строки Str используется обозначение $Tdm_L(\pi, n)$.

Пусть str — текстовая строка длины n в алфавите A и $Tdm_L = Tdm_L(\Pi_{str}(L), n)$. Если $\Pi_{str}(\lambda) = (\pi^{*i}_j)^K_\lambda$ и $\Pi_{Tdm_\Lambda}(\lambda) = (\pi^i_j)^K_\lambda$, то для строки str и её тест-периода λ вводятся спектры:

$$\psi(\Pi_{str}(\lambda), \Pi_{Tdm_\Lambda}(\lambda), n) = \\ = \frac{n}{\lambda} \sum_{j=1}^{\lambda} \sum_{i=1}^{K} \frac{(\pi^{*i}_j - \pi^i_j)^2}{\pi^i_j(1 - \pi^i_j)} \backsim \chi^2_{(K-1)\lambda}; \quad (1)$$

$$H(\lambda) = \psi\big(\Pi_{str}(\lambda), \Pi_{Tdm_1}(\lambda), n\big) - M\big(\chi^2_{(K-1)\lambda}\big), \quad (2)$$

где $M(\chi^2_N)$ — математическое ожидание χ^2-распределения с N степенями свободы. Первый тест-период с ярко выраженным максимальным значением характеристического спектра $\mathbf{H}$ (см. (2)) служит оценкой скрытого периода в строке str (см. рис. $1, a$). Такой тест-период L рассматривается в качестве периода скрытой профильности, если строка str статистически неотличима от L-профильной строки Tdm_L. Для проверки неотличимости используется спектр $\mathbf{D_L}$ отклонения строки str от L-профильности:

$$D_L(\lambda) = \frac{\psi(\Pi_{str}(\lambda), \Pi_{Tdm_L}(\lambda), n)}{\chi^2_{crit}((K-1)\lambda, \alpha)}, \quad (3)$$

где $\chi^2_{crit}(N, \alpha)$ — критическое значение χ^2_N-распределения на уровне значимости $\alpha = 0{,}05$. Гипотеза о неотличимости принимается, если $\mathbf{D_L} \leqslant 1$ на более, чем 95% тест-периодов. В частности, если $L = 1$, неотличимость означает статистическую однородность строки str. В этом случае поиск структурной регулярности в последовательностях ДНК не проводился.

Результаты и обсуждение

В характеристических спектрах (см. (2)) кодирующих районов наблюдается регулярность пиков в 3 нукл. (рис. $1, a$, $2, a$). Так проявляется первый уровень организации кодирования, обусловленный генетическим триплетным кодом. Часто, в спектрах Фурье этому уровню соответствует доминирующий пик на частоте 0,33 (см. рис. $2\,в$). При наличии 3-регулярности, скрытая профильность, отличная от 3-профильности, выявляет второй уровень организации кодирования. На этот уровень организации указывает ярко выраженный максимум характеристического спектра (см. рис. $1, a$). Как правило, его нельзя выделить с помощью спектра Фурье, в котором даже нет доминирующего пика на частоте 0,33 (см. рис. $1, в$). На тест-периоде, соответствующему максимуму характеристического спектра, проверяется наличие скрытой

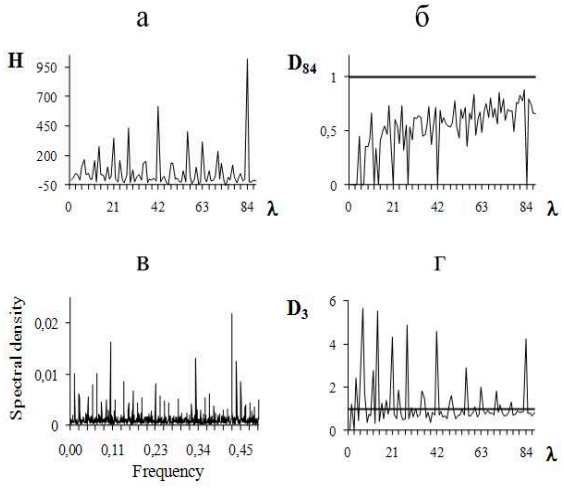


Рис. 1. Спектры 2С подхода (см. (2), (3)) и спектр Фурье кодирующей последовательности ДНК для белка 285A из группы «цинковых пальцев» (KEGG, hsa:26974, zinc finger protein 285A, 1773 нукл.): *а* — характеристический спектр (см. (2)); *б* — спектр отклонения от 84-профильности (см. (3)); *в* — Фурье-спектр; *г* — спектр отклонения от 3-профильности (см. (3))

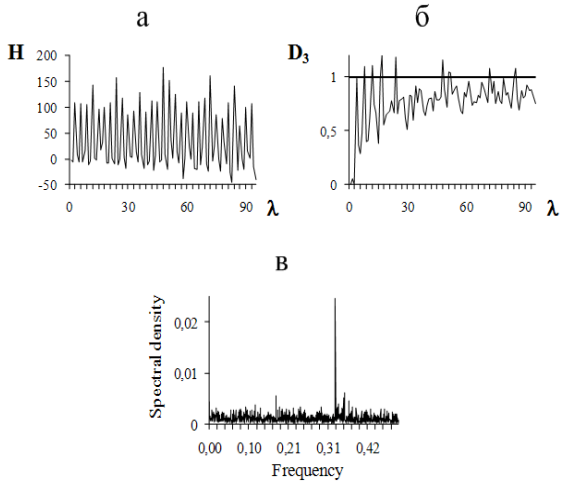


Рис. 2. Спектры 2С подхода (см. (2), (3)) и спектр Фурье кодирующей последовательности ДНК для белка-рецептора, принимающего сигнал (KEGG, hsa:6734, signal recognition particle receptor (docking protein), 1917 нукл.): *а* — характеристический спектр (см. (2)); *б* — спектр отклонения от 3-профильности (см. (3)); *в* — Фурье-спектр

профильной периодичности. Наличие скрытой профильности в 84 нукл. (см. рис. 1, *а*) в кодирующей последовательности ДНК соответствует в белке повторяющемуся домену «цинкового пальца», содержащему альфа-спираль и две антипараллельные бета-структуры. Как правило, «цинковый палец» включает около 20 аминокислот и стабилизируется одним или двумя ионами цинка. Основной группой белков с цинковыми пальцами являются ДНК-связывающие факторы транскрипции.

На рис. 2, *а* показан характеристический спектр кодирующей последовательности ДНК для белка-рецептора, связывающего передающие сигнал частицы. На фоне очевидной 3-регулярности пиков этого спектра невозможно выделить несомненно доминирующий пик. Однако, как видно из спектра на рис. 2, *б*, в анализируемой кодирующей последовательности белка нельзя признать и существование скрытой 3-профильности. Следовательно, характеристический спектр (см. рис. 2, *а*) последовательности отражает лишь её неоднородность вследствие триплетного кодирования аминокислот. Именно неоднородностью этой кодирующей последовательности в 3 нукл. обусловлен доминирующий пик на частоте 0,33 в спектре Фурье на рис. 2, *в*.

С помощью предложенных методов 2С подхода было проанализировано наличие 3-регулярной и скрытой профильной структур в 18 140 кодирующих последовательностях ДНК человека из базы данных KEGG [13], получивших экспериментальное подтверждение. С учётом погрешности статистических методов кодирующие последовательности являются неоднородными и 3-регулярными. Кроме того, в 74% кодирующих последовательностей ДНК наблюдается скрытая профильная периодичность. Второй уровень кодирования (отличный от 3-профильности) проявляется в 11% проанализированных кодирующих последовательностей.

В предложенных методах не учитывались повреждения последовательности ДНК вставками и делециями. Возможно, этим объясняется отсутствие профильного уровня организации в 3-регулярных последовательностях, составляющих 21% от исходных кодирующих последовательностей ДНК.

Аналогичный анализ был выполнен и для интронов. Многие гены человека имеют «мозаичную структуру», в которой кодирующая последовательность ДНК прерывается «лишними» фрагментами нуклеотидов. Чтобы информация о белке могла быть прочитана на рибосоме, с помощью специальных генетических механизмов такие фрагменты удаляются из последовательности гена. Длина отдельных интронов может составлять десятки тысяч нуклеотидов. Рассматривались 277 477 последовательностей интронов (некодирующих фрагментов генов) человека из базы данных EID (The Exon-Intron Database) [14, 15]. Неоднородность была выявлена для 24% всех интронов. Доля 3-регулярных последовательностей среди них составила 3% от исходного числа интронов. То есть в рамках погрешности статистических методов можно считать, что для интронов характерно отсут-

ствие 3-регулярности. Скрытая профильность в неоднородных интронах была выявлена для 13% их исходного числа. Если в кодирующих последовательностях ДНК скрытая профильная периодичность часто служит генетической основой для формирования структурных доменов соответствующих белков, то вопрос о том, какую роль может играть скрытая профильность в интронах, требует будущих исследований. Одной из причин выявления профильности в интронах могут являться размытые тандемные повторы [2].

Выводы

С помощью предложенных методов спектрально-статистического подхода к распознаванию скрытой профильной периодичности (профильности) в последовательностях ДНК был проведён сравнительный анализ структурных свойств кодирующих последовательностей ДНК и интронов человека. Показано, что фундаментальным свойством кодирующих последовательностей ДНК является 3-регулярность, обусловленная неоднородностью последовательности на длине в 3 нукл. вследствие триплетного кодирования аминокислот. Строго говоря, феномен 3-регулярности не гарантирует наличия скрытой периодичности в кодирующей последовательности ДНК, на что ранее не обращали внимание. Такая 3-регулярность сама по себе или совместно с возникающей на её основе скрытой 3-профильной периодичностью (3-профильностью) составляют первый уровень в организации кодирования генетической информации. Второй уровень в организации кодирования представляет скрытая профильная периодичность, отличная от 3-профильности. Этот уровень может коррелировать с особенностями структуры кодируемых белков. Существующие на сегодняшний день статистические методы анализа последовательностей ДНК не выделяют уровни организации кодирования в последовательностях ДНК. На основе полученных в работе результатов возможно создание новых математических методов для предсказания кодирующих районов в геноме.

Литература

[1] *Lobzin V. V., Chechetkin V. R.* Order and correlations in genomic DNA sequences.The spectral approach // Physics–Uspekhi, 2000. — Vol. 43, No. 1. — Pp. 55–78.

[2] *Benson G.* Tandem repeats finder: a program to analyze DNA sequences // Nucl. Acids Res., 1999. — Vol. 27, No. 2. — Pp. 573–580.

[3] *Tsonis A. A., Elsner J. B., Tsonis P. A.* Periodicity in DNA coding sequences: Implications in gene evolutions // J. Theor. Biol., 1991. — Vol. 151, No. 3. — Pp. 323–331.

[4] *Gutierrez G., Oliver J. L., Marin A.* On the origin of the periodicity of three in protein coding DNA sequences // J. Theor. Biol., 1994. — Vol. 167, No. 4. — Pp. 413–414.

[5] *Fickett J. W., Tung C.-S.* Assessment of protein coding measures // Nucleic Acid Res., 1992. — Vol. 20, No. 24. — Pp. 6441–6450.

[6] *Chaley M. B., Kutyrkin V. A.* Model of perfect tandem repeat with random pattern and empirical homogeneity testing poly-criteria for latent periodicity revelation in biological sequences // Math. Biosci., 2008. — Vol. 211. — Pp. 186–204.

[7] *Korotkov E. V., Korotkova M. A., Kudryashov A. A.* Information decomposition method to analyze symbolical sequences // Phys. Lett. A, 2003. — Vol. 312. — Pp. 198–210.

[8] *Gatherer D., McEvan N. R.* Analysis of sequence periodicity in E. coli proteins: empirical investigation of the «duplication and divergence» theory of protein evolution // J. Mol. Evol., 2003. — Vol. 57, No. 2. — Pp. 149–158.

[9] *Chaley M. B., Kutyrkin V. A.* Structure of proteins and latent periodicity in their genes // Moscow Univ. Biol. Sci. Bull., 2010. — Vol. 65, No. 4. — Pp. 133–135.

[10] *Tiwari S., Ramachandran S., Bhattacharya A., Bhattacharya S., Ramaswamy R.* Prediction of probable genes by Fourier analysis of genomic sequences // Comput. Appl. Biosci., 1997. — Vol. 13, No. 3. — Pp. 263–270.

[11] *Issac B., Singh H., Kaur H., Raghava G. P. S.* Locating probable genes using Fourier transform approach // Bioinformatics, 2002. — Vol. 18, No. 1. — Pp. 196–197.

[12] `www.imtech.res.in/raghava/ftg/` — FTG::Fast Fourier Transform based GENE Prediction Server. — 2011.

[13] `www.genome.jp/kegg/` — KEGG: Kyoto Encyclopedia of Genes and Genomes. — 2011.

[14] `www.utoledo.edu/med/depts/bioinfo/database.html` — The University of Toledo: BPG Databases. — 2011.

[15] *Shepelev V., Fedorov A.* Advances in the Exon-Intron Database // Brief. Bioinform., 2006. — Vol. 7, No. 2. — Pp. 178–185.

О методе оценки качества поиска повторов в генетических последовательностях

Горчаков М. А., Панкратов А. Н.
skiffcmc@gmail.com
г. Пущино, Институт математических проблем биологии РАН

В процессе анализа генетических последовательностей на наличие повторяющихся участков, в том числе с учетом возможных мутаций, необходимо обладать критерием оценки качества полученных результатов. В данной статье предложен способ такой оценки на основе расчета вероятности появления повторов с заданными характеристиками (длина и степень мутации) в последовательности заданной длины, а также описан основанный на такой оценке способ выбора оптимальных параметров метода поиска повторов.

В последнее время одно из самых бурно развивающихся научных направлений — изучение и анализ геномов, в том числе и человека. Один из подходов к анализу генетических последовательностей — перевод генетической последовательности в текстовую форму (как известно, такая последовательность представляет собой цепь из 4-х возможных аминокислот — аденин, гуанин, тимин и цитозин, так что ее цифровая форма представляет собой «текст» из 4-х допустимых «букв» — A, G, T, C). Анализ подобного «текста» можно проводить с использованием различных методов, один из них — поиск повторяющихся участков достаточно большой длины, в том числе с использованием обобщенного спектрально-аналитического метода (ОСАМ). Вполне естественной задачей, возникающей при таком анализе, является создание критерия качества работы алгоритма, который мог бы быть вычислен без участия человека и в дальнейшем использован для модификации имеющихся методов. Основной проблемой при построении такого критерия является слабая формализация понятия качественной работы алгоритма — с одной стороны, он должен находить повторы в большем количестве по сравнению с неавтоматизированным поиском, но с другой — нахождение слишком большого числа повторов, скорее всего, означает, что многие из них являются случайно возникшими, так называемом «шумом». Найденное и описанное в данной статье решение предлагает использование подхода теории вероятностей в оценке качества распознавания, а именно — оценку вероятности возникновения повтора(ов) заданной длины в случайной последовательности, не являющейся реально существующим геномом.

Построение вероятностной модели

Так как из-за сходства пар A–T и G–C исходный «текст» для анализа легко переводится в двоичную последовательность, то будем рассматривать материал для анализа как последовательность из двух возможных символов «0» и «1».

Определение 1. *Назовем бинарной последовательностью (БП) длины N строку длины N, каждый символ которой равен либо «0», либо «1»:*

$$B(N) = \{b_i\}, i \in [1;N], b_i \in \{0,1\}\,. \qquad (1)$$

Определение 2. *Бинарной последовательностью со строгим повтором длины k назовем такую БП (1), в которой:*

$$\exists m,n \in [1;N], m \neq n : \forall i \in [0;k-1] b_{i+m} = \\ = b_{i+n}\,. \qquad (2)$$

Определение 3. *Бинарной последовательностью с повтором длины k и l заменами назовем БП, для которой после замены l символов на противоположные в повторе будет выполнено (2).*

Определение 4. *Бинарной последовательностью с повтором длины k и s удалениями назовем БП, для которой после удаления s символов в повторе будет выполнено (2).*

Задача 1. Исходя из случайности и совместной независимости элементов БП, необходимо определить вероятность возникновения в такой БП повтора длины k ($P(N,k)$), с l заменами ($P(N,k,l)$) и s удалениями ($P(N,k,l,s)$).

Для определения такой вероятности разумно воспользоваться методом математической индукции, т. е. определять $P(N+1,k)$ через $P(N,k)$.

Теорема 1. *Значение вероятности $P(N+1,k+1)$ равно:*

$$P(N+1,k+1) = P(N,k+1) + \\ + (1 - P(N-k-1,k+1))(2^{-k-1} - 2^{-N})\,, \qquad (3)$$

при этом

$$P(2k,k) = 2^{1-k} - 2^{1-2k}\,.$$

Из того, что $P(N,k) < (N+1)\cdot 2^{-k}$, можно сделать вывод, что возникновение повтора без мутаций, даже, на первый взгляд, относительно небольшой (несколько десятков символов из нескольких миллионов в БП), не может быть отнесено к случайностям.

Теорема 2.

$$P(N,k,l) = C_k^l P(N,k)\,;$$
$$P(N,k,l,s) = 2^s C_k^s P(N,k,l)\,.$$

Приближенные вычисления вероятностей повтора на ЭВМ

Очевидно, что функция $P(N,k)$ не может быть посчитана обычными методами в силы очень малых ее значений при больших k (необходимая точность не может быть достигнута обычными числовыми типами). Таким образом, необходимо придумать алгоритм ее приближенного подсчета, каким-то образом обходящий это ограничение.

Теорема 3.

$$P(N+1,k) = P(N,k) + 2^{-k-1} + \bar{o}(2^{-3k/2})$$

при $k \ll N \ll 2^k \ll 2^N$.

Далее:

Теорема 4.

$$P(n,k) = (n+3)2^{-k-1} + \bar{o}(2^{-3k/2})$$

при $n \ll N$.

И наконец:

Теорема 5.

$$P(N+1,k) = P(N,k) + \\ + (1 - P(N-k-1,k+1))2^{-k-1} + \bar{O}(2^{-N}) \quad (4)$$

при $k \ll N \ll 2^k \ll 2^N$.

Кроме того, для корректных вычислений нам понадобится приближенная оценка величины C_k^l, которая может быть выражена как

$$C_k^l \approx \sqrt{\frac{1}{k\alpha(1-\alpha)}}\alpha^{-k\alpha}(1-\alpha)^{-(1-\alpha)k},$$

где $\alpha = l/k$. Исходя из полученных результатов, становится ясно, как можно реализовать подсчет вероятности возникновения повторов с мутациями. Для такого подсчета целесообразно считать сразу $P(N,k,l,s) = 2^s C_k^s C_k^l P(N,k)$, так как умножение вероятности возникновения повтора без мутаций на «большое» число дает возможность получить более точные результаты. Расчеты должны вестись, исходя из приближенной формулы, изложенной в теореме 5, а также того, что при минимально допустимой точности δ первое значение n, для которого $P(n,k) > \delta$, равно

$$n = \left[\frac{\delta \cdot 2^{k+1-s}}{C_l^k C_s^k}\right] - 3.$$

Общий алгоритм подсчета функционала качества

Теперь опишем общий принцип подсчета оценки качества работы функционала. Эта оценка в дальнейшем может быть использована как критерий качества для любого алгоритма оптимизации:

— в качестве входных данных для подсчета оценки качества дается набор найденных повторов в БП. Каждый повтор характеризуется длиной и количеством мутаций в нем;
— для каждого повтора считается вероятность его возникновения в случайной БП. Алгоритм такого подсчета был дан выше;
— для повторов, чья приближенная вероятность возникновения не равна 0, считается вероятность возникновения именно такого сочетания повторов как произведение вероятностей возникновения каждого повтора и вероятности того, что в оставшейся части последовательности нет ни одного повтора длины k или больше;
— полученное значение вычитается из количества повторов с вероятностью возникновения, равной 0. Полученная разность и есть критерий качества.

Полученный критерий, будучи применен в алгоритме оптимизации (цель — максимизировать значение этого критерия) будет заставлять работать поиск повторов в соответствии со следующими правилами:

— любой найденный повтор с нулевой вероятностью возникновения «ценится выше» любого числа повторов с ненулевой вероятностью возникновения, т. е. первичная оптимизация будет вестись с целью найти как можно больше таких повторов;
— когда все подобные повторы будут найдены, поиск будет минимизировать совместную вероятность возникновения повторов с ненулевой вероятностью возникновения.

Альтернативный подход учитывает только повторы с нулевой вероятностью возникновения. Он позволяет существенно ускорить работу алгоритма вычисления качества функционала, так как для вычисления вероятности возникновения повтора выше заданной необходимо провести N итераций подсчета, а для сравнения с заданной величиной — всего 1 действие. Однако при этом могут быть упущены некоторые повторы, не представляющие интереса по одиночке, но свидетельствующие о своей неслучайности в совокупности.

Оптимизация

Теперь, зная способ оценить качество поиска найденных повторов и то, какие данные необходимы ему для корректной работы, можно полностью формализовать подход к созданию самообучающегося алгоритма поиска повторов на основе обобщенного спектрально-аналитического метода.

1. В качестве входных данных модулю поиска повторов с помощью OCAM передается текстовый файл с генетической последовательностью. После оцифровки последовательности с

помощью OCAM производится поиск повторов при установленной минимально необходимой степени схожести.

2. Следующим шагом должна стать процедура объединения пересекающихся повторов. Это необходимо сделать по соображениям корректности подсчета общей вероятности как произведения вероятностей — такой подход можно считать верным только при соблюдения условия независимости событий появления повторов, а это возможно только при отсутствии взаимопересечений.
3. Так как оцифровка в п. 1, возможно, происходила с шагом, отличным от 1, то любой найденный повтор не локализован окончательно — его положение в реальной генетической последовательности может отличаться от найденного на длину шага оцифровки. Для установления точного местонахождения повтора необходимо воспользоваться алгоритмом сравнения двух строк по расстоянию Левенштейна и, используя это расстояние в качестве оптимизационного критерия, найти пару минимально отличающихся строк при условии сохранения длины повтора, найденной в п. 2. Для такой пары нужно также получить количество вставок, удалений и замен, которое необходимо для полного совпадения этих строк.
4. После того как в точности определено местоположение повтора и количество мутаций (вставок, удалений, замен) в нем, можно посчитать вероятность его возникновения. После того как посчитаны вероятности для всех найденных повторов, можно оценить качество найденной алгоритмом картины.
5. После того как с помощью алгоритма подсчета вероятностей получено значение показателя качества найденной картины повторов, необходимо использовать данный результат для определения необходимости проведения новой итерации алгоритма (а также выбора новых параметров в случае, если итерацию нужно провести). Для оптимизации функционала качества допустимо использовать любой из созданных многомерных оптимизационных методов.

Выводы

В работе были проведены строгие вычисления вероятности возникновения повторов в случайной бинарной последовательности. На основе этих формул также были получены способы приближенного вычисления этих вероятностей на ЭВМ. Был описан алгоритм вычисления критерия качества найденных повторов, основывающийся на вероятностном подходе, а также полностью описана процедура автоматического выбора оптимальных параметров метода поиска повторов с использованием данного критерия качества.

Литература

[1] *Дедус Ф. Ф, Куликова Л. И, Панкратов А. Н, Тетуев Р. К.* Классические ортогональные базисы в задачах аналитического описания и обработки информационных сигналов. — 2004.

[2] *Марков А. А.* Избранные труды. — Л.: Изд-во Академии наук СССР, 1951.

[3] *Ибрагимов И. А., Линник Ю. В.* Независимые и стационарно связанные величины. — М.: Изд-во Наука, Главная редакция физико-математической литературы, 1965.

[4] Алгоритм вычисления расстояния Левенштейна. `ru.wikibooks.org/wiki/Расстояние_Левенштейна`, 2011.

[5] Сайт библиотеки вычислений с плавающей точкой высокой точности. `www.apfloat.org`, 2006.

[6] *Ширяев Н. А.* Вероятность. — М.: Изд-во Наука, 1980.

О спектральном алгоритме распознавания протяженных тандемных повторов в геномах*

Панкратов А. Н., Пятков М. И.
pan@impb.ru
г. Пущино, Институт математических проблем биологии РАН

Предложен алгоритм для автоматического распознавания тандемных повторов, основанный на спектральном анализе статистических профилей последовательности. Найдены условия оптимальной аппроксимации повторов одного типа, встречающихся в геномах, обеспечивающие изображение повтора на спектральной матрице гомологии в виде квадрата.

Совершенствование и автоматизация методов секвенирования — точного определения последовательности оснований ДНК — предоставило возможность изучения человека и других организмов на биоинформационном уровне. Одной из хорошо формализованных задач бионформатики является исследование повторяющихся элементов (повторов) и изучение их структуры в генетических текстах человека и животных. Выделяют множество видов повторов. В данной статье рассмотрены прямые, инвертированные и отдельный акцент сделан на поиске тандемных повторов большой протяженности, мотив которых более 1000 нуклеотидов. Повторы играют важную роль в функционировании и перестройке геномов: они могут быть вирусами, мигрирующими в течение жизни по геному. Кроме того, исследование повторяющихся структур геномов используется при решении филогенетических задач — определении родства групп организмов на генном уровне.

Описание алгоритма

Постановка задачи и существующие подходы к решению. Рассмотрим задачу сравнения двух последовательностей (строк) и выделения повторов (подстрок). Сложность в обнаружении повторов обусловлена возникновением мутаций в последовательностях. Мутации — это одиночные замены, вставки и делеции (удаления) букв. Для решения задачи поиска неточных подстрок создано множество алгоритмов, большинство из которых основано на методе динамического программирования.

Относительно простой метод анализа гомологичных участков двух последовательностей с помощью точечных матриц гомологии [1] заключается в нахождении и отображении на прямоугольной матрице общих для двух последовательностей слов, т. е. подпоследовательностей длиной W, в которых совпадает не менее M букв. Параметры W и M называются параметрами фильтрации, а величина W называется размером окна. Такое построение обладает большой наглядностью, т. к. гомологичные участки выявляются в виде диагональных линий. Основная проблема использования подобных матрицы возникла после того, как были полностью отсеквенированы полные геномы различных организмов и размеры матриц и время сравнения значительно выросли.

Сравнительно недавно был предложен и стал развиваться спектральный алгоритм поиска повторяющихся последовательностей в геномах [2–4], основным отличием которого является то, что с его помощью можно быстро обрабатывать очень длинные последовательности, сопоставимые с размерами хромосом (10^8 нуклеотидов), и искать протяженные ($\geqslant 1000$ нуклеотидов) повторы.

Для поиска протяженных повторов изначально был предложен алгоритм [5], основанный на оценке периодичности статистического профиля последовательности. В данной работе для автоматизации поиска тандемного повтора применяется более общий алгоритм, основанных на построении и анализе матрицы гомологии. Исходный алгоритм является по сложности линейным в зависимости от длины обрабатываемой последовательности, рассматриваемый в этой работе алгоритм является более сложным с вычислительной точки зрения, хотя он также линеен, если построение и анализ матрицы проводить только вдоль диагонали. После того, как с помощью этого метода визуально были найдены уникальные протяженные повторы [6, 7], возникла задача полной автоматизации поиска таких повторов, решение которой было найдено и описано в этой работе.

Получение ДНК-профиля. В первую очередь, для того чтобы воспользоваться аппроксимативными возможностями полиномиальных ортогональных базисов, необходимо преобразовать генетическую последовательность $s_i \in \{A, T, G, C\}$ в числовую функцию, которая в дальнейшем будет служить для анализа. Для этого разбиваем набор символов на две группы следующим образом:

$$f(s_i) = \begin{cases} 1, & \text{если } s_i \in \{G, C\}; \\ 0, & \text{если } s_i \in \{A, T\}. \end{cases}$$

Работа выполнена при финансовой поддержке РФФИ, проекты №11-07-00716, 11-07-00577, 11-07-00519, 11-01-00765, 10-01-00609, 10-07-00300, а также компаний Интел и Т-платформы.

Тогда, если зафиксировать длину подпоследовательности W и двигать это «окно» с единичным шагом, то всю последовательность можно описать так:

$$f_{GC}(j) = \sum_{i=j}^{j+W_1} f(s_i)\,. \qquad (1)$$

Таким образом, f_{GC} — некоторая дискретная функция, которая на следующем шаге преобразуется в спектры разложения. На рис. 1 изображен пример GC профиля.

Для того чтобы повысить качество распознавания, одновременно используются два профиля f_{GC} и f_{GA}, что соответствует полному описанию последовательности, при котором последовательность может быть восстановлена по профилям. Заметим, что для полного описания последовательности в алфавите из n букв потребуется $\log_2 n$ профилей.

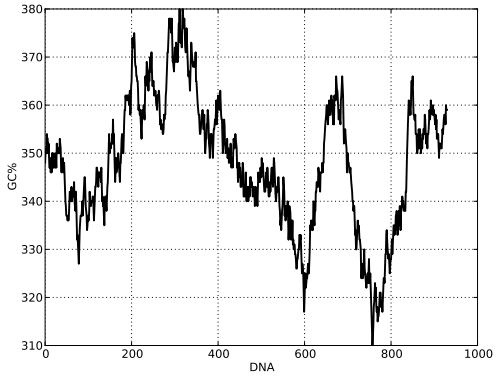


Рис. 1. Пример GC профиля ($W_1 = 700$)

Стоит отметить, что перед переводом в статистическое представление с помощью профилей может потребоваться предобработка последовательности нуклеотидов. Данная предобработка включает в себя удаление из последовательности букв N, которые возникают на этапе секвенирования, когда нуклеотид не определен. Обычно N-нуклеотиды сгруппированы в протяженные регионы, и при распознавании данным методом это является помехой, т. к. все N-регионы нуклеотидов будут похожи друг на друга. Координаты N-регионов на этапе предобработки запоминаются и используются далее при определении координат найденных повторяющихся структур.

Получение спектров разложения. На этом этапе, как и на предыдущем, используется метод «скользящего окна». Часть ДНК-профиля, попавшая в окно W_2, преобразуется в коэффициенты разложения, после чего окно перемещается на шаг S и процедура повторяется. Вектора коэффициентов разложения, полученные на каждом шаге, запоминаются в матрице для дальнейшей оценки близости между ними.

Для представления ДНК-профиля в спектральном виде использовалось разложение в ряды Фурье по тригонометрическим функциям: $\{1/\sqrt{2}, \cos kx, \sin kx, \ldots\}$, $k = 1, 2, \ldots, n$, где n — количество гармоник разложения. Норма всех базисных функций равна $\sqrt{\pi}$.

Для представления профиля в виде коэффициентов разложения используются также базисы полиномов Лежандра непрерывного аргумента и Чебышева дискретного аргумента. Каждый из этих базисов имеет недостатки в рамках алгоритма решения данной задачи. Базис Лежандра относится к базисам непрерывного аргумента, поэтому требуется интерполяция аппроксимируемой функции с постоянной сетки на сетку узлов квадратурной формулы Гаусса. Базис Чебышева дискретного аргумента является аналогом базиса Лежандра и состоит из функций, ортогональных на постоянной сетке с единичным весом. Однако известно, что вычисление полиномов Чебышева дискретного аргумента по рекуррентным формулам обладает неустойчивостью, что связано с отказом от интерполяции на сетку Гаусса. Кроме того, рекуррентное соотношение для полиномов Чебышева дискретного аргумента является более сложным по количеству вычислительных операций. В данной задаче неустойчивость рекуррентного алгоритма несильно влияет на результат, поскольку окно аппроксимации сильно превосходит количество коэффициентов разложения. Базис тригонометрических полиномов Фурье состоит из функций, ортогональных на постоянной сетке. Недостатком базиса Фурье является то, что он предназначен для аппроксимации периодических сигналов. Но в данной задаче этот недостаток также несильно влияет на результат, поскольку в задаче используется оценка интегрального среднеквадратичного отклонения сглаженных профилей. Эффект Гиббса, который возникает при аппроксимации разрывных сигналов имеет локальный характер (в данном случае, на концах интервала) и практически не влияет на оценку интеграла отклонения сигналов.

Сравнение спектров разложения. Для оценки близости векторов коэффициентов разложения используется интегральное среднеквадратичное отклонение. Выбор критерия обусловлен, в основном, выбором базиса, его параметров и глубины разложения. В настоящий момент используется критерий, который основан на оценке квадрата нормы:

$$(f, f) = \sum_{i=0}^{n} c_k^2 \parallel \varphi_k \parallel^2 = \int_{-\pi}^{\pi} f^2 dt < 2\pi W_1^2\,,$$

где f — разница сравниваемых профилей, а c_k — разница коэффициентов разложения сравниваемых профилей. Оценка интеграла основана на том, что максимальное значение функции-профиля ограничено величиной W_1. Таким образом, решающее правило основано на проверке следующего неравенства:

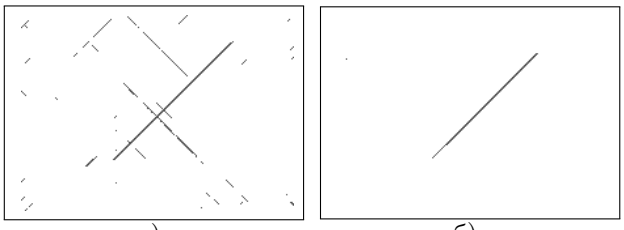
$$\sum_{i=0}^{n} c_k^2 < 2W_1^2\varepsilon\,,$$

где ε — пороговое значение, один из основных параметров алгоритма.

Спектральное представление позволяет получать спектры инвертированного образца непосредственно из спектра прямого образца, если используемый базис состоит из четных и нечетных базисных функций. Таким образом, поиск инвертированных повторов практически не добавляет вычислительной сложности алгоритму.

Среднеквадратичное отклонение является монотонным по числу коэффициентов разложения, данное свойство позволяет по первым нескольким коэффициентам оценить, насколько схожи профили, если пороговое значение ε превышено, дальнейшее сравнение можно не проводить.

Анализ матрицы гомологии. На последнем этапе анализируется спектральная матрица гомологии. Спектральная матрица гомологии создана по аналогии с точечной матрицей гомологии, основным отличием является то, что в спектральной матрице гомологии одна точка соответствует сравнению двух спектров разложения, а не отдельных букв, как в точечной матрице. Прямые повторы отображаются как отрезки линий, параллельных диагонали, инвертированные повторы — перпендикулярно. При сравнении последовательности самой с собой возникает главная диагональ.



а) б)

Рис. 2. Зашумленный инвертированный повтор: а) только GC-профиль; б) GC и GA профили

На рис. 2 изображена фильтрация матрицы гомологии за счет добавления дополнительного профиля, построенного по нуклеотидам «G» и «A». Оценка близости между спектрами коэффициентов разложения происходит по двум профилям раздельно, после чего повтор отображается на матрице, если обе пары профилей близки. Из рисунка видно, что использование двух профилей позволяет существенно повысить качество распознавания повторов. Анализ матриц гомологии позволяет получить координаты всех повторов последовательности; тем не менее, подобный анализ предполагает этап верификации найденных повторов.

Поиск протяженных тандемных повторов

В настоящее время идет активное изучение различных видов тандемных повторов. Основной акцент сделан на три вида повторов: это микросателлиты, повторяющиеся фрагменты ДНК длиной от 1 до 6 пар оснований (п.о.), минисателлиты от 7 до 100 п.о. и сателлиты от 100 до 200 п.о. Изучение повторяющихся структур подобной длины имеет исторические корни, когда ученые не могли себе позволить просканировать хромосому или весь геном по причинам низкого быстродействия компьютеров; соответственно по этой же причине алгоритмы поиска были направлены на небольшие по протяженности последовательности и, соответственно, на поиск минисателлитов. Таким образом, до последнего времени не было ясно, существуют ли более протяженные повторы (более 2000 п.о.). Как оказалось, в геноме подобные структуры существуют и имеют важное значение для функционирования организма в целом [5]. В частности, такие протяженные повторы далее мы будем называть как мегасателлиты, предлагается использовать в качестве маркеров в задаче поиска районов синтении при сравнении различных групп организмов на генном уровне.

Настройка алгоритма. Поиск мегасателлитных тандемных повторов с помощью спектрального алгоритма является задачей распознавания с обучением по образцу. В качестве образца выбраны структуры ранее найденных данным методом повторов IMPB_01 [6] и IMPB_02 [7] и предложено выполнить поиск мегасателлитов со схожей структурой. Отличительной особенностью данного повтора является то, что его период составляет свыше 2400 п.о., что значительно превышает периоды всех ранее найденных неточных тандемных повторов в геномах млекопитающих.

На первом этапе проведен анализ эталонного повтора средствами программы Spectral Revisor, позволяющей строить матрицу гомологии участков геномов. Найдены наиболее оптимальные параметры и опции запуска программы Spectral Revisor, при которых шаблонные структуры IMPB_01 и IMPB_02 выделяются на спектральной матрице гомологии с минимальным шумом, что позволило искать эти повторы в виде красного квадрата в автоматическом режиме (рис. 3).

В результате полного сканирования генома мыши было выявлено три мегасателлитных тандемных повтора помимо шаблонного IMPB_01 и

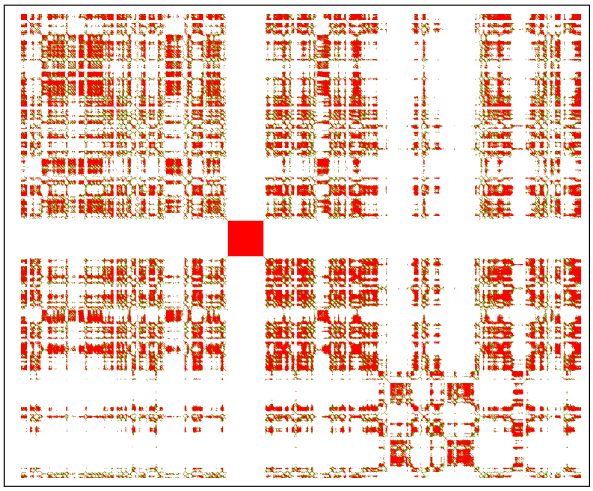


Рис. 3. Тандемный повтор IMPB_01 в виде красного квадрата: ($W_1 = 2300$, $W_2 = 10\,000$, $S = 2300$, $\varepsilon = 0{,}0011$, количество коэффициентов 50)

IMPB_02. Наибольший из ранее неизвестных повторов имеет период 1850 п.о. и количество повторений более 100 раз.

Масштабируемость алгоритма. Особенности математического аппарата и структура алгоритма позволяют добиться максимальной эффективности на многопроцессорных машинах. Создана версия программы, поддерживающая многоуровневый параллелизм MPI/OpenMP/Intel IPP, а также версия для гетерогенных систем с использованием графических процессоров. Анализ масштабируемости как на машинах с общей памятью, так и на машинах, поддерживающих массивно-параллельные вычисления, показал практически линейную масштабируемость алгоритма поиска [8, 9].

Выводы

В настоящей работе предложен алгоритм поиска протяженных тандемных повторов на основе обобщенного спекрально-аналитического метода. Отличием от классических алгоритмов поиска повторов является то, что непосредственная работа с символьной последовательностью ведется только на этапах предобработки, создания профиля и верификации, а основную часть алгоритма составляют спектральные методы. Алгоритм позволяет быстро обрабатывать протяженные последовательности, сопоставимые с размерами хромосом (порядка 10^8), за счет того, что полностью построен на вычислениях с плавающей точкой и хорошо векторизуется и распараллеливается на машинах с различной архитектурой вычислений.

Сканирование генома мыши (Mus Musculus) на предмет поиска протяженных тандемных повторов выявило ранее неизвестный повтор, период которого составляет 1850 п.о. Найдены оптимальные параметры для поиска такого рода повторов в любых организмах.

Литература

[1] *Gibbs A. J., McIntyre G. A.* The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. // Eur. J. Biochem, 1970. No. 16. P. 1–11

[2] *Дедус Ф. Ф., Куликова Л. И., Махортых С. А., Назипова Н. Н., Панкратов А. Н., Тетуев Р. К.* Аналитические методы распознавания повторяющихся структур в геномах // Доклады Академии наук, 2006. Т. 411. № 5. С. 599–602.

[3] *Tetuev R. K., Dedus F. F., Kulikova L. I., Makhortykh S. A., Nazipova N. N., Pankratov A. N.,* Recognition of the structural-functional organization of genetic sequences // Moscow University Computational Mathematics and Cybernetics. — 2007. Vol. 31, No. 2. P. 49–53.

[4] *Pankratov A. N., Gorchakov M. A., Dedus F. F., Dolotova N. S., Kulikova L. I., Makhortykh S. A., Nazipova N. N., Novikova D. A., Olshevets M. M., Pyatkov M. I., Rudnev V. R., Tetuev R. K., Filippov V. V.* Spectral analysis for identification and visualization of repeats in genetic sequences // Pattern Recognition and Image Analysis, 2009. Vol. 19. No. 4. P. 687–692.

[5] *Тетуев Р. К., Назипова Н. Н., Панкратов А. Н., Дедус Ф. Ф.* Поиск мегасателлитных тандемных повторов в геномах эукариот по оценке осцилляций кривых GC-содержания // Математическая биология и биоинформатика, 2010. Т. 5, № 1. С. 30–42.

[6] *Tetuev R. K., Nazipova N. N.* Consensus of repeated region of mouse chromosome 6 containing 60 tandem copies of a complex pattern // Repbase Reports, 2010. Vol. 10, No. 5. P. 776.

[7] *Tetuev R. K., Nazipova N. N., Dedus F. F.* Consensus of repeated region of rat chromosome 4 similar to mouse chromosome 6 repeated region, enclosed in the intergenic region between genes Hrh1 and Atg7 // Repbase Reports, 2010. Vol. 10, No. 8. P. 1185.

[8] *Pankratov A. N., Tetuev R. K., Pyatkov M. I.* `http://software.intel.com/en-us/articles/fast-spectral-estimation-of-genetic-homology.` Fast Spectral Estimation of Genetic Homology, 2010.

[9] *Панкратов А. Н., Тетуев Р. К., Пятков М. И.* `http://software.intel.com/ru-ru/articles/fast-spectral-estimation-of-genetic-homology.` Быстрое спектральное оценивание генетической гомологии, 2010.

Преимущество оценок подобия фрагментов ДНК с помощью спектрально-аналитического метода*

Дедус Ф. Ф., Тетуев Р. К., Назипова Н. Н., Ольшевец М. М., Панкратов А. Н., Пятков М. И.

ruslan.tetuev@gmail.com

Институт математических проблем биологии РАН, Пущино

В работе обсуждаются наиболее вероятные причины успешного применения приближённого спектрально-аналитического представления в таких задачах, как сравнение генетических последовательностей. Основное внимание уделено построению устойчивых сигналов, вопросам компактности спектрального представления и задачам сверхбыстрого поиска схожих спектральных описаний.

Спектрально-аналитический метод

Измерения реального мира неизбежно производятся с погрешностями, поэтому для снижения влияния ошибок часто приходится увеличивать количество измерений, затем по определенному правилу производить усреднение полученных величин. Классическая задача оценки неизвестных параметров приводит к методу наименьших квадратов (МНК), сводящемуся к минимизации следующего функционала

$$f(x) = \int_a^b \Big[y(x) - \sum_{n=1}^{N} a_n x^n \Big]^2 dx,$$

где $y(x)$ — измеряемая зависимость, $\sum_{n=1}^{N} a_n x^n$ — полиномиальное приближение, a_n — искомое значение коэффициентов. Определение неизвестных коэффициентов a_n в функционале приводит к матричным преобразованиям с промежуточными матрицами размером $n \times n$. В случаях, когда полученный приближающий многочлен недостаточно хорошо описывает измеренные значения $y(x)$ следует повышать степень многочлена. При этом всю процедуру определения коэффициентов a_n приходится повторять заново. В середине 19-го столетия П. Л. Чебышев при решении задач по интерполяции предложил искать минимум другого функционала:

$$f(x) = \int_a^b \Big[y(x) - \sum_{n=1}^{N} A_n \varphi^n(x) \Big]^2 dx.$$

По сути приближающий многочлен теперь представляет собой сумму других многочленов повышающих степеней, заранее определённых и «специальным образом организованных». Для определённых систем многочленов коэффициенты A_n могли быть определены по-прежнему по методу наименьших квадратов, но также и намного проще по формуле:

$$A_n = \frac{\int_a^b y(x)\varphi_n(x)dx}{\int_a^b \varphi_n(x)^2 dx}.$$

В дальнейшем такие функции $\varphi_n(x)$ он определил как взаимно ортогональные, принадлежащие ортогональному базису Чебышева.

Можно показать, что вариационная задача минимизации последнего функционала приводит к системе уравнений, решение которой определяет значения коэффициентов A_n искомого приближающего многочлена в виде:

$$y(x) = \sum_{n=0}^{N} A_n \varphi^n(x),$$

где $\varphi^n(x)$ — ортогональные многочлены; A_n — коэффициенты разложения функции $y(x)$ по ортогональным полиномам $\{\varphi^n(x)\}$.

Преимущество последнего представления в том, что в случае, когда нам необходимо повысить точность аналитического описания $y(x)$ теперь достаточно вычислить только новые коэффициенты разложения A_{n+1}, A_{n+2} и т. д., так как все ранее вычисленные коэффициенты разложения остаются неизменными. Этим предложением Чебышев фактически ввел в решение задач по методу наименьших квадратов применение всех классических ортогональных полиномов, которые обладали такими основными свойствами [3] и обеспечил получение несмещенных оценок.

В настоящее время все классические ортогональные полиномы и функции непрерывного и дискретного аргументов представлены в виде таблиц и подробно описаны в работах [4, 5]. Их разнообразные свойства хорошо изучены и позволяют с помощью разработанных адаптивных процедур [4] получать аналитические описания различных измеренных характеристик с квазиоптимальной точностью.

За счет разнообразных весовых функций, определенных однозначно для каждого конкретного ортогонального базиса, можно практически всегда удовлетворять различные требования к особенностям описания поступающих сигналов. Кроме того, у нас разработаны алгоритмы и программы сверхглубокого разложения сигналов для ряда ортогональных базисов из числа классических [7].

Работа выполнена при поддержке РФФИ, гранты № 10-01-00609-а, 11-07-00519-а, 11-07-00716-а.

При этом число членов разложения в указанных случаях достигает 10 000 и выше.

Важная особенность аналитической аппроксимации данных на основе классических ортогональных базисов состоит в том, что отрезки ортогональных рядов линейно независимы и вся информация об описываемых сигналах сосредоточена в коэффициентах разложения. Благодаря этому свойству удалось разработать алгебру коэффициентов разложения [8], которая позволила реализовать весь цикл обработки сигналов исключительно в самом пространстве коэффициентов разложения. В работе [8] показаны пути оценки надежности получаемых результатов в виде вычисления вероятности ошибок, превосходящих определенные пределы.

Практическая ценность МНК заключается главным образом в том, что он дает в руки исследователям аналитические зависимости, позволяющие адекватно оценивать скрытые закономерности в наблюдаемых явлениях. А также проверять надежность получаемых оценок путем вычисления вероятностных характеристик с учетом соответствующих законов распределения.

Применяя данный подход к задачам биоинформатики, выявлена перспективность методов на пути ускорения алгоритмов, к примеру, поиска тандемных повторов в ДНК последовательностях, нахождения особых паттернов в белковых последовательностях и т. д. и т. п. Особая привлекательность заключается в том, что использование спектрально-аналитических методов обуславливает интегральность и целостность проводимых оценок, гладкость исследуемых функций и устойчивость получаемых результатов. Наряду с этим компактность спектрального представления функций благоприятно сказывается на объеме всех расчетов и общей скорости разрабатываемых алгоритмов.

Спектрально-аналитическое описание генетических последовательностей

Широко известно, что генетические последовательности наиболее часто представимы в виде последовательностей букв. В случае ДНК последовательностей в качестве букв просто берутся заглавные буквы в названиях нуклеиновых кислот, радикалы которых составляют данный участок ДНК молекулы.

Переход от текстового вида непосредственно к спектрально-аналитическому описанию ДНК представляется авторам данной работы не столь эффективным, хотя следует отметить, что варианты такого подхода продолжают регулярно публиковаться в соответствующих изданиях. Как показала практика, разрабатываемый подход намного эффективнее при введении промежуточного представления ДНК в виде некоторого сигнала. Но для этого следует определить способ *количественной* оценки *качественного* состава фрагментов ДНК.

Существует масса разнообразных способов получения сигнала, как количественного описания генетической последовательности, и каждый из них обладает рядом преимуществ и недостатков, но применительно к данной задаче все они не имеют почти никаких принципиальных различий. Ввиду этого разумнее выбрать в качестве сигнала тот, получение которого проще с вычислительной точки зрения. Таким сигналом оказался так называемый GC%-профиль, значения которого линейно зависят от простой, легко вычисляемой физической величины – количества водородных связей соответствующего участка ДНК молекулы. Выпишем простую формулу количества водородных связей F на участке ДНК $[x, x+W-1]$:

$$F_W^{(G,C)}(x) = 2\left(\sum_{i=x}^{x+W-1} A + \sum_{i=x}^{x+W-1} T\right) + 3\left(\sum_{i=x}^{x+W-1} G + \sum_{i=x}^{x+W-1} C\right).$$

где ΣA, ΣT, ΣG, ΣC — условные обозначения для количества букв A, T, G, C на заданном интервале. Как видим, тем самым мы предложили функцию, определяющую $L - W + 1$ значений при рассмотрении ДНК строки длины L. При вычислениях на ЭВМ принято использовать схожую, но более простую функцию, определяемую в следующем виде:

$$f_W^{(G,C)}(x) = \frac{\sum_{i=x}^{x+W+1} G + \sum_{i=x}^{x+W+1} C}{W} \times 100,$$

котора и называется GC%-профилем или функцией GC-содержания данного участка ДНК последовательности. Выпишем замечательные свойства этой функции:

— устойчивость к изменению аргумента;
— устойчивость к изменению параметра;
— устойчивость к замене букв в ДНК;
— устойчивость к вставке букв в ДНК;
— устойчивость к изъятию букв из ДНК.

Таким образом, в качестве аналитического описания произвольных фрагментов последовательности ДНК в данной работе предлагается рассматривать несколько первых коэффициентов разложения: $A_0, A_1, \ldots, A_{n-1}$, полученных при разложении GC%-профиля, определённого на заданном фрагменте. Причём, как обычно, меньшему количеству коэффициентов разложения соответствует более гладкое и устойчивое к случайным искажениям приближённое представление сигнала, в то

время как при больших n мы получаем более детальное описание участка, а при некоторых условиях даже можем восстановить ДНК последовательность в её прежнем, буквенном представлении.

Преимущества спектрально-аналитического описания ДНК

Устойчивость спектрально-аналитического приближённого описания сигнала к, пускай даже сильным, но локальным его искажениям — хорошо известное явление, которое весьма благотворно сказалось на получаемых сегодня результатах и в целом на разрабатываемом подходе. Причиной локальных искажений в случае GC%-профиля, очевидно, является присутствие многочисленных точечных мутаций ДНК последовательности. И если бы не это свойство устойчивости, характерное для приближённых спектральных описаний, то едва ли в рамках представляемой работы удалось бы получить уникальные результаты, «ускользнувшие» от любого другого альтернативного подхода, как было показано в работах [9], [10].

Дополнительным важным преимуществом спектрально-аналитического подхода к задачам исследования генетических последовательностей является компактность получаемых описаний при сопутствующей простоте количественных оценок. Например, в задаче поиска схожих участков ДНК достаточно искать пары описаний, имеющих малое среднеквадратичное (или абсолютное) отклонение, что возможно организовать как высокоскоростной вычислительный процесс. Идея подобных быстрых вычислений и основные принципы спектрального анализа ДНК последовательностей впервые были изложены в рамках конференции BGRS-2006 [11], а сегодня уже существует и активно используется ряд программных продуктов, рассчитанных на реализацию быстрых вычислений на платформах как с линейной, так и с параллельной архитектурами, к примеру, SpectralRevisor. Успешному применению технологий параллельных вычислений дана высокая оценка в ходе целого ряда различных конкурсов и проектов, проводимых такими компаниями как Intel Co., T-Platforms, РФФИ и др.

Выводы

Как показано выше, предлагаемый спектрально-аналитический подход позволяет получить весьма устойчивое приближённое описание последовательности ДНК, а также результаты, устойчивые к естественным искажениям «исходного материала», а именно, ко всем точечным мутациям: заменам, вставкам и изъятиям.

К общим недостаткам подхода можно отнести необходимость подбора некоторых параметров, таких, например, как оптимальное количество используемых коэффициентов разложения. Однако по отношению к данным параметрам результаты обработки данных также оказываются устойчивыми, что позволяет успешно применять подход даже при квазиоптимальных параметрах.

Литература

[1] *Марков А. А.* Исчисление вероятностей. — Л.: Госиздательство, 1924.

[2] *Колмогоров А. Н.* Теория вероятностей и математическая статистика. — М.: Наука, 2005. — С. 582.

[3] *Чебышев П. Л.* Вопросы о наименьших величинах, связанные с приближенным представлением функций. — Л: Госиздательство, 1947. — С. 580.

[4] *Дедус Ф. Ф., Куликова Л. И., Панкратов А. Н., Тетуев Р. К.* Классические ортогональные базисы в задачах аналитического описания и обработки информационных сигналов. — М: Издательский отдел ВМиК МГУ, 2004. — С. 147.

[5] *Дедус Ф. Ф., Махортых С. А., Устинин М. Н., Дедус А. Ф.* Обобщенный спектрально-аналитический метод обработки информационных массивов. Задачи анализа изображений и распознавания образов. — М: Машиностроение, 1999. — С. 357.

[6] *Худсон Д.* Статистика для физиков. — М: Мир, 1970. — С. 243.

[7] *Панкратов А. Н., Бритенков А. К.* Обобщенный спектрально-аналитический метод: проблемы описания цифровых данных семейства ортогональных полиномов. // Вестник Нижегородского государственного университета им. Н. И. Лобачевского. Серия Радиофизика. — Н. Новгород: Издательство ННГУ, 2004. — С. 5-14.

[8] *Тетуев Р. К., Дедус Ф. Ф.* Классические ортогональные полиномы. Применение в задачах обработки данных. — М: 11-формат, 2007. — С. 60.

[9] *Tetuev R. K., Nazipova N. N.* — Repbase Reports. — 2010. — V. 10, № 5. — P. 776.

[10] *Tetuev R. K., Dedus F. F., Nazipova N. N.* — Repbase Reports. — 2010. — V. 10, № 8. — P. 1185.

[11] *Tetuev R. K., Dedus F. F., Kulikova L. I, Makhortikh S. A., Pankratov A. N., Nazipova N. N.* — Analytical methods in problems of recognition the structural and functional organization of genetic sequences // The 2006 Bioinformatics of The Genome Regulation And Structure International Summer School for young scientists “Evolution, Systems Biology and High Performance Computing Bioinformatics”, July 12–15, Novosibirsk, Russia, 2006.

Критерии локальной разрешимости и регулярности как инструмент исследования морфологии аминокислотных последовательностей*

Торшин И. Ю.
tiy135@yahoo.com
Москва, Московский физико-технический институт, Центр систем прогнозирования и распознавания

В рамках алгебраического подхода поиск корректных алгоритмов распознавания ограничен фундаментальными критериями разрешимости и регулярности исследуемой задачи. В настоящей работе проведен анализ критериев локальной разрешимости и регулярности одной из задач биоинформатики — задачи распознавания вторичной структуры белка. Показано, что регулярность (и, следовательно, разрешимость) локальной формы задачи гарантирована тупиковыми множествами наиболее информативных мотивов заданной размерности и протяженности. Приведены результаты экспериментов, проведенных на выборке всех известных на сегодняшний день аминокислотных последовательностей. Установлены тупиковые множества мотивов, обеспечивающие регулярность локальной формы задачи при произвольном множестве прецедентов.

Введение

В биоинформатике имеется отдельный класс задач распознавания, связанных с обработкой символьных последовательностей [1, 2]. Распознавание вторичной структуры белка на основе его аминокислотной последовательности представляет особый интерес, так как является одним из важных шагов к установлению взаимосвязи между химической и пространственной уровнями структуры белка. Задача рассматривается как перевод последовательности символов из одного алфавита в другой, а накопленный материал о третичном и вторичном уровнях структуры белка — как основа для построения непротиворечивых множеств прецедентов [3].

В [3–5] предложен формализм для анализа разрешимости и локальности данной задачи распознавания. Введение ключевых понятий для анализа локальной разрешимости задачи (окрестность, система масок, объект, множество мотивов, монотонность и тупиковость по системам масок и множествам мотивов) позволило провести эксперименты по установлению тупиковых множеств аминокислотных мотивов с наибольшей информативностью по отношению ко вторичной структуре белка.

При практическом применении рассматриваемого формализма следует принимать во внимание, что объем данных по первичной структуре белка (миллионы аминокислотных последовательностей) в сотни раз превышает массив имеющихся прецедентов «первичная структура–вторичная структура». Поэтому при переносе закономерностей, установленных в ходе анализа множеств прецедентов, возникает вопрос о возможности обобщения установленных закономерностей на все имеющиеся аминокислотные последовательности. В настоящей работе, данный вопрос исследуется на основе критериев разрешимости и регулярности локальной формы рассматриваемой задачи.

Работа выполнена при поддержке грантов РФФИ 09-07-12098, 09-07-00212-а и 09-07-00211-а.

Критерии локальной разрешимости и регулярности на множествах мотивов

Одним из основных результатов работ [3–5] является формулировка критериев разрешимости исследуемой задачи распознавания. Используются два алфавита: алфавит A для описания первичной структуры белка («верхнего слова») и алфавит B для описания вторичной структуры («нижнего слова»). Алфавит A (однобуквенные обозначения аминокислот) обычно определяется как $A = \{\texttt{A, C, D, E, F, G, H, I, K, L, M, N, P, R, S, T, V, W, Y}\}$. Алфавит B может быть определён существенно различными способами [3]; для целей настоящей работы вполне приемлем трехбуквенный алфавит $B = \{\text{S, H, L}\}$, описывающий три принципиально различных вида вторичной структуры: «стрэнды» (S, англ. strand), «спирали» (H, helix), и «петли» (L, loop).

Критерий локальной разрешимости с использованием отдельных масок (выражение $(6'')$ в работе [3]) был сформулирован для системы масок $M = \{\hat{m}_1, \dots \hat{m}_{|M|}\}$, где $\hat{m}_k = \{\mu_1^k, \dots, \mu_{m(k)}^k\}$ — k-я маска, $\mu_i^k \in \mathbb{Z},\ \mu_1^k < \dots < \mu_{m(k)}^k,\ k = 1, \dots, |M|$, $m(k) = |\hat{m}_k|$ — размерность маски. Каждая маска описывает конкретный способ выбора подпоследовательности в заданной последовательности символов и является признаком верхнего слова. В работах [4, 5] был осуществлен переход от анализа разрешимости на множествах признаков (масок) к анализу разрешимости на множествах значений признаков (т. н. «мотивов») и от множеств прецедентов — к множествам объектов.

emphЭлементарными объектами q (далее просто «объектами») назовем элементы множества $Q = A^{\hat{m}_\Sigma(M)}$, где $\hat{m}_\Sigma(M) = \bigcup\limits_k \hat{m}_k$ — *объединённая маска. Элементарные мотивы* k (далее просто «мотивы») — элементы множества

$$K = \{(\hat{m}, V) \mid \hat{m} \in M,\ V \in A^{|\hat{m}|}\}.$$

Мотив k назовем *отличающим* для произвольной пары объектов q_1 и q_2, если k присутствует в одном из объектов и отсутствует во втором. Тогда критерий локальной разрешимости (Теорема 1 в [5]) записывается следующим образом:

$$\underset{Q}{\forall}(i,j)\colon w_i \neq w_j \;\Rightarrow\; \underset{K}{\exists} k\colon (k \in V_i) \neq (k \in V_j). \quad (1)$$

Пусть $r_1 = N_{(1)}/N(N-1)$, где $N = |Q|$, а $N_{(1)}$ — множество пар объектов, на котором выполнено условие (1). В разрешимой задаче $Z(Q,K)$ $r_1 \equiv 1$. Утверждение (1) соответствует переходу от задачи $Z(Pr,M)$ [3] к эквивалентной задаче $Z(Q,K)$ [5].

Наряду с разрешимостью в современной теории распознавания [6–8] изучается также *регулярность* задач — разрешимость задачи, сопровождающаяся разрешимостью задач из некоторой её окрестности в изучаемом множестве задач. Определим окрестность задачи $Z(Q,K)$ со множеством объектов $Q = \{(V_1,w_1),\dots,(V_i,w_i),\dots\}$ как множество задач Z' со множеством объектов $Q' = \{(V_1,w'_1),\dots,(V_i,w'_i),\dots\}$ при произвольных $w_1,\dots,w'_i,\dots$, т. е. задача Z будет регулярной на множестве объектов Q тогда и только тогда, когда выполняется *условие регулярности на множестве мотивов*:

$$\underset{Q}{\forall} q_i, q_j, i \neq w \;\Rightarrow\; \underset{K}{\exists} k\colon (k \in^* V_i) \neq (k \in^* V_j). \quad (2)$$

Если задача $Z(Q,K)$ регулярна, то будем называть Q *регулярным множеством объектов*, а K — *регулярным множеством мотивов*. Пусть $r_0 = N_{(2)}/N(N-1)$, где $N = |Q|$, а $N_{(2)}$ — множество пар объектов, на котором выполнено условие (2). В регулярной задаче $Z(Q,K)$ $r_0 \equiv 1$.

Для исследования вопроса о возможности обобщения установленных закономерностей на все имеющиеся аминокислотные последовательности, представляет практический интерес нахождение некоторых минимальных множеств мотивов, гарантирующих регулярность на произвольном множестве объектов. Регулярное множество мотивов назовём *тупиковым*, если условие (2) выполнено для K, но не выполнено для любого $K' \subset K$.

Теорема 1. *В задаче $Z(Q,K)$ тупиковое множество мотивов K_1, обеспечивающее разрешимость, является подмножеством тупикового множества мотивов K_0, обеспечивающего регулярность.*

Следствие 1. *Пусть $\Delta_{1,0} = 1 - |K_1 \cap K_0|/|K_1|$ — параметр, описывающий соответствие множества K_1 множеству K_0. При фиксированном множестве объектов $\Delta_{1,0} \equiv 0$.*

Следствие 2. *Пусть $r_{1,0} = r_1(K_1 \cap K_0)$. При фиксированном множестве объектов $r_{1,0} \equiv 1$.*

Таким образом, множество мотивов, удовлетворяющее критерию регулярности (2), содержит в себе множество мотивов, обеспечивающее разрешимость (1) задачи распознавания. Важно, что тестирование регулярности может проводится без какой-либо информации о вторичной структуре белка, т. е. на таких множествах объектов как $Q' = \{(V_1,\Delta),\dots,(V_i,\Delta),\dots\}$, где Δ — неопределённость. Для сокращения полного перебора в ходе вычисления тупиковых множеств мотивов вводятся эвристические оценки информативности мотивов.

Оценки информативности мотивов и критерий регулярности

В духе теории классификации значений признаков можно сказать, что следует оставлять мотивы с «высокой информативностью» и удалять мотивы с «достаточно низкой» информативностью так, что регулярность задачи (2) не нарушена. Оценка информативности мотивов $D\colon K \to \mathbb{R}_+$ может быть введена различными способами так, чтобы бóльшая «информативность» мотива соответствовала бóльшим значениям D.

Отметим принципиальное различие между оценками информативности, используемыми при тестировании условий разрешимости и регулярности. В случае разрешимости «более информативными» являются мотивы, которые (а) характеризуются наибольшей частотой встречаемости и (б) выделяют «достаточно много» объектов l-го класса и «достаточно мало» объектов всех остальных классов. При отборе мотивов следует учитывать оба эти фактора, так что построение функции D, адекватной для тестирования разрешимости, представляет собой нетривиальную задачу. Практически полезным является использование функций

$$D_1(\alpha) = \sum_{l=1}^{m} D_l^\alpha, \quad D_2(\alpha) = N_\Sigma^\alpha \sum_{l=1}^{m} D_l^\alpha,$$

$m = |B|$; D_l^α определяется в соответствии с (3):

$$D_l^\alpha = \begin{cases} 1 - \frac{v_l^\alpha}{v_l^0}, & v_l^\alpha \leqslant v_l^0; \\ \frac{v_l^\alpha - v_l^0}{1 - v_l^0}, & v_l^\alpha > v_l^0. \end{cases} \quad (3)$$

где $v_l^\alpha = N_l^\alpha / N_\Sigma^\alpha$ — частота встречаемости значения $b_l \in B$, а v_l^0 — частоты встречаемости литеры $b_l \in B$ во всем множестве объектов Q. Заметим, что функция $D_1(\alpha)$ также может быть использована для оценки степени непротиворечивости нерегулярных множеств объектов [5], при этом *долей непротиворечивых объектов* во множестве Q назовём отношение $|Q_{\text{нп}}|/|Q|$,

$$Q_{\text{нп}} = \{q_\alpha \subset Q \mid D_1(\alpha) = |B|\}.$$

В случае условия регулярности, которое не включает информации о нижних словах объектов,

необходимым условием наибольшей информативности мотива является только частота его встречаемости. Поэтому оценка информативности мотива $k_\alpha \in K(Q,M)$, который входит в состав N_Σ^α объектов из Q, может быть определена просто как $D_{\mathrm{reg}}(\alpha) = N_\Sigma^\alpha/|Q|$.

Эвристические оценки информативности мотивов необходимы, прежде всего, для нахождения тупиковых множеств мотивов на основе критериев разрешимости или регулярности. Функция $D\colon K \to \mathbb{R}_+$ порождает *линейный порядок* $I(K)$ на множестве мотивов K. Принцип отбора мотивов состоит в том, что для каждой пары объектов из Q находится различающий мотив с наивысшей информативностью. Отобранные таким образом мотивы образуют тупиковое множество мотивов K_0, которое определяется *характеристической функцией* $T(\alpha, Q)$ [5]:

$$T(\alpha,Q) = \begin{cases} 1, & \exists\limits_{Q}(i,j)\colon \big(K_f(i,j) = \alpha\big); \\ 0, & \text{в противном случае;} \end{cases} \qquad (4)$$

где $K_f(i,j) = \min\limits_{1,\ldots,|K|} \alpha\colon (k_\alpha \in V_i) \neq (k_\alpha \in V_j)$.

Теорема 2. *При вычислении K_1 $K(Q,M)$ на основе функционала D, а K_0 — по $K(Q',M')$ с использованием D' выполнение условия $K_1 \subseteq K_0$ гарантировано при $Q = Q'$, $M = M'$ и $D = D'$.*

Следствие 3. *Вложение порядков $I(K_1) \subseteq I(K_0)$ является необходимым и достаточным условием выполнения $K_1 \subset K_0$.*

Следствие 4. *Вложение $I(K) \subseteq I(K')$ — необходимое условие выполнения $K_1 \subset K_0$.*

Следствие 5. *Параметр $\Delta_{1,0}$ отражает длину наибольшей общей под-последовательности линейных порядков $I(K_1)$ и $I(K_0)$.*

При экспериментальном тестировании регулярности и разрешимости следует проводить усреднение вычисляемых K_1 и K_0 по различным выборкам объектов одного размера, контролируя при этом значение определённого ранее параметра $\Delta_{1,0}$. Определим z_α — заполненность элементарного мотива k_α при тестировании n выборок объектов как $z_\alpha = \sum\limits_{i=1}^{n} T(\alpha, Q_i)/n$. Во множестве мотивов K (это может быть множество K_1 и K_0), усреднённом по n выборкам Q, информативным назовём мотив k_α с заполненностью $z_\alpha \geqslant z_{\min}$. Очевидно, что при заданной D наиболее информативны мотивы с $z_{\min} = 1$. Так как при снижении значения параметра $z_{\min}$ ($z_{\min} = 0{,}9$, $0{,}8$ и т. д.), в K войдёт большее число различающих мотивов, то параметры разрешимости (r_1) и регулярности (r_0) увеличатся. Размер выборки объектов $|Q|$ является важным параметром, определяющим значения z_α конкретных мотивов при данной системе масок.

Экспериментальное тестирование условий разрешимости и регулярности

Выражения (1–4) позволяют вычислять тупиковые множества мотивов для данных Q и M. Эксперименты по тестированию разрешимости (1) и регулярности (2) проводились на общедоступных экспериментальных данных по первичной, вторичной и третичной структурам белков (PDB, Protein Data Bank), суммарно включающих более 150 000 последовательностей и структур [9]. Использовались различные протяжённости объектов с $n = 4, \ldots, 16$; доля непротиворечивых объектов при $n \geqslant 6$ составила не менее 0,87.

Регулярность также тестировалась на множествах объектов, полученных на основе всех известных аминоксилотных последовательностей в базе данных UNIPROT [10], в которой присутствует 15 млн. попарно различных последовательностей общей длиной в $5 \cdot 10^9$ литер. Из данных БД UNIPROT были сформированы выборки объектов длиной в 6, 8 и 10 литер, при этом объекты с частотой встречаемости менее 10^{-7} были исключены.

В качестве оценок информативности мотивов использовались $D_{\mathrm{reg}}(\alpha)$ и $D_2(\alpha)$. Были исследованы системы масок с размерностью всех масок равной $m = 2$ (системы M_n^2) и $m = 3$ (M_n^3), полученные полным перебором по m позиций из $n = 6, 8, 10, 12, 16$.

Далее рассматриваются результаты тестирования разрешимости и регулярности $Z(Q,K)$ с использованием $D_{\mathrm{reg}}(\alpha)$ и $D_2(\alpha)$, результаты исследования выполнения условия регулярности на K_0 (UNIPROT) при множествах мотивов $\{z_\alpha = 1\}$ и, наконец, морфология аминокислотных последовательностей с учётом данных K_0 (UNIPROT) при $\{z_\alpha = 1\}$.

Тестирование разрешимости и регулярности $Z(Q,K)$ проводилось на непротиворечивых выборках объектов из PDB [5] с использованием $D_{\mathrm{reg}}(\alpha)$ и $D_2(\alpha)$ в качестве эвристических оценок информативности. Были исследованы выборки размером $2 \cdot 10^4$, $5 \cdot 10^4$, 10^5 и $2 \cdot 10^5$ объектов, сформированные путем случайного отбора объектов без возвращения, по 10 выборок для каждого из приведенных выше значений $|Q|$. Для каждого размера выборки вычислялись множества мотивов K_1 (подмножество $\{z_\alpha = 1\}$) и K_0 ($\{z_\alpha = 1\}$); рассчитывались значения показателей r_1, r_0, $r_{1,0}$ и $\Delta_{1,0}$. Затем сравнивались множества мотивов, полученные с использованием различных оценок информативности.

Зависимость разрешимости (r_1) от размера выборки объектов, исследованная на объектах PDB (подмножество $\big\{z_\alpha \in K_1 \;\big|\; z_\alpha = 1\big\}$) посредством оценки информативности $D_{\mathrm{reg}}(\alpha)$ показала,

что наибольшие различия в значениях r_1 для масок M_n^3 наблюдались при малых размерах выборок ($2 \cdot 10^4$ объектов); наилучший результат показала система масок M_6^3 ($r_1 = 0{,}94 \pm 0{,}01$; $|Q| = 2 \cdot 10^5$). Для данной системы масок $r_1 \geqslant 0{,}99$ достигалось при $z_{\min} = 0{,}7$. Для системы масок M_n^2 практически полная разрешимость ($r_1 \geqslant 0{,}99$) достигалась при любых значениях $n = 6, \ldots, 16$ даже на малых выборках объектов (20000–50000), причём пересечение множеств $\{k_\alpha \in K_1(D_{\text{reg}}(\alpha)) \mid z_\alpha = 1\}$ и $\{k_\alpha \in K_1(D_2(\alpha)) \mid z_\alpha = 1\}$ обеспечивало $r_1 \geqslant 0{,}98$. Сравнение результатов для $D_{\text{reg}}(\alpha)$ и $D_2(\alpha)$ в системе масок M_6^3 показало, что множества мотивов $\{k_\alpha \in K_1(D_{\text{reg}}(\alpha)) \mid z_\alpha = z_{\min}\}$ и $\{k_\alpha \in K_1(D_2(\alpha)) \mid z_\alpha = z_{\min}\}$ содержат общее подмножество, обеспечивающее $r_1 \geqslant 0{,}99$ при $z_{\min} = 0{,}7$ и $|Q| = 2 \cdot 10^5$.

Таким образом, при тестировании разрешимости на произвольном множестве объектов достаточного размера, схожие тупиковые множества мотивов получаются при использовании различных оценок информативности.

Результаты тестирования *выполнимости условия регулярности* на выборках из PDB и UNIPROT с использованием $D_{\text{reg}}(\alpha)$ показало, что параметр $\Delta_{1,0}$ имел наименьшее значение ($\Delta_{1,0} = 0{,}005 \pm 0{,}003$) в системах масок M_6^3 и M_6^2. При этом множества вида $\{k_\alpha \in K_1 \cap K_0 \mid z_\alpha = 1\}$ обеспечивали различение 0,9995 пар объектов по критерию регулярности. Иначе говоря, при таких параметрах системы масок, практически идентичные тупиковые множества мотивов могут быть получены на произвольном множестве объектов.

Морфология аминокислотных последовательностей. Множества мотивов, получаемые в результате тестирования $(1, 2)$, характеризуют морфологию или, в некотором смысле, «структуру» аминокислотных последовательностей. В соответствии с (4), для каждой пары из i-го и j-го объектов множества Q функция $K_f(i, j)$ находит наиболее информативный различающий мотив. Для всех таких мотивов $T(\alpha) = 1$, т. е. эти мотивы образуют K_0. После вычисления $T(\alpha)$ для всех пар объектов из Q, каждому i-му объекту из Q соответствует n_i^{rm} различающих мотивов из K_0, $n_i^{rm} = |\{T(\alpha) = 1\}_i|$. Зачастую (в 90% случаев) эти мотивы покрывают не все позиции объекта, выделяя тем самым некоторые «информативные» позиции аминокислотной последовательности, соответствующие «информативным» мотивам, см. таблицу 1.

RKSGnS L	SEKfRE L	QAgfHd L	EmVnDA H
ARVhEy H	VATTGE H	DPVhKA H	hrSySc S

Таблица 1. Примеры структур объектов с учётом позиций, выбранных по тупиковому множеству мотивов K_0. Указаны литеры вторичной структуры объектов; информативные позиции выделены заглавными A-литерами.

Заключение

Показано, что регулярность (и, следовательно, разрешимость) локальной формы задачи гарантирована тупиковыми множествами наиболее информативных мотивов заданной размерности и протяженности. Приведены результаты экспериментов, проведенных на выборке всех известных на сегодняшний день аминокислотных последовательностей. Установлены тупиковые множества мотивов, обеспечивающие регулярность локальной формы задачи при произвольном множестве прецедентов. Установление тупиковых множеств мотивов и информативных позиций в первичной структуре необходимо не только для синтеза корректных алгоритмов распознавания вторичной структуры белка, но и для решения ряда других задач биоинформатики.

Литература

[1] *Torshin I. Y.* Bioinformatics in the Post-Genomic Era: The Role of Biophysics. — NY: Nova Biomedical Books, 2006. ISBN: 1-60021-048.

[2] *Torshin I. Y.* Sensing the change from molecular genetics to personalized medicine. Nova Biomedical Books, NY, USA, 2009, In "Bioinformatics in the Post-Genomic Era" series, ISBN 1-60692-217-0.

[3] *Рудаков К. В., Торшин И. Ю.* Вопросы разрешимости задачи распознавания вторичной структуры белка. Информатика и её применения, 2010. — Т. 4., № 2. — С. 25–35.

[4] *Рудаков К. В., Торшин И. Ю.* О разрешимости формальной задачи распознавания вторичной структуры белка. ММРО-14, Суздаль, 21–25 сентября, 2009, С. 596–597.

[5] *Рудаков К. В., Торшин И. Ю.* Анализ информативности мотивов на основе критерия разрешимости в задаче распознавания вторичной структуры белка. Информатика и её применения, 2011.

[6] *Журавлев Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики, М.: Наука, 1978. — Вып. 33. — С. 5–68.

[7] *Журавлев Ю. И., Рудаков К. В.* Об алгебраической коррекции процедур обработки (преобразования) информации // Проблемы прикладной математики и информатики, М.: Наука, 1987. — С. 187–198.

[8] *Рудаков К. В.* Универсальные и локальные ограничения в проблеме коррекции эвристических алгоритмов // Кибернетика, 1987. — № 2. — С. 30–35.

[9] *Berman H. M., Henrick K., Nakamura H.* Announcing the worldwide Protein Data Bank // Nature Structural Biology, 2003. — Vol. 10 No. 12. — Pp. 980–982.

[10] The UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res. 39: D214-D219 (2011).

Математическая модель данных микрочипов ДНК, учитывающая эффекты кросс-гибридизации и насыщения*

Когадеева М. С., Рябенко Е. А.
m.kogadeeva@gmail.com, riabenko.e@gmail.com
Москва, Московский Государственный Университет им. М. В. Ломоносова

Анализ данных микрочипов ДНК является одной из перспективных задач современной биоинформатики, однако он осложняется биологическими и техническими вариациями, возникающими на разных стадиях лабораторных экспериментов. В данной работе предлагается модель данных микрочипов ДНК, учитывающая эффекты неспецифических биохимических взаимодействий и насыщения. Модель построена в рамках подхода к определению абсолютных концентраций молекул в образце, что существенно отличает её от распространённых на сегодняшний день методов, ориентированных на оценку относительных концентраций. Таким образом, предложенная модель может быть применима для более широкого спектра задач.

Технология микрочипов ДНК используется в современной молекулярной биологии для определения экспрессии десятков тысяч генов в образце одновременно. Гены — это участки молекулы ДНК, которые заключают в себе информацию о том, какой белок, где и в каком количестве будет синтезирован. *Экспрессия генов* — это процесс, в ходе которого происходит транскрипция и трансляция информации, закодированной в молекуле ДНК (то есть, происходит синтез молекул мРНК и белков). Ген считается экспрессированным, если кодирующая его ДНК транскрибируется в комплементарную мРНК, и оценкой экспрессии генов является концентрация молекул мРНК в образце.

С помощью микрочипов можно изучать мРНК, полученные из разных тканей при разных состояниях клеток. Можно сравнивать экспрессию генов в больных и здоровых или обработанных лекарством клетках и определять, какие гены и при каких экспериментальных условиях активируются и могут влиять на процессы, происходящие в клетке. Анализ данных микрочипов ДНК осложняется различными техническими и биологическими вариациями, что побуждает исследовать и улучшать методы обработки результатов микрочиповых экспериментов.

Технология микрочипов ДНК

Устройство микрочипа ДНК. Микрочип представляет собой небольшую пластину, на поверхности которой закреплены на известных позициях определённые участки одинарных цепочек ДНК, называемые *пробами*. Исследуемый образец готовят таким образом, чтобы в нём находились одинарные цепочки мРНК. Согласно принципу комплементарности одинарные цепочки в образце вступают в реакцию с пробами на микрочипе. К образцу добавляют флюоресцентные метки, после чего микрочип сканируют. Основная задача — по интенсивности свечения проб определить, какие именно цепочки мРНК вступили в реакцию, и оценить их концентрацию в образце.

При решении этой задачи может возникнуть ряд проблем. Во-первых, на этапе сканирования появляется фоновый шум и засветка, обусловленные погрешностью сканера. Во-вторых, пробы существенно отличаются по характеристикам взаимодействия со свободными цепочками мРНК в образце. Возникает биологический шум, вызванный побочными реакциями, влияющими на интенсивность свечения проб. Основным источником биологического шума является реакция *кросс-гибридизации*, то есть связывания молекул мРНК в образце с пробами на микрочипе, комплементарными им лишь частично. Другой эффект, наблюдаемый на поверхности микрочипа, это *эффект насыщения*, возникающий при больших концентрациях определённых молекул мРНК в образце. В этом случае с увеличением концентрации мРНК интенсивность свечения проб перестаёт изменяться линейно.

Анализ данных микрочипов ДНК. Распространённые подходы к оценке экспрессии генов, как правило, состоят из трёх этапов: вычитания фонового шума, нормализации данных и суммаризации интенсивностей соответствующих генам проб.

Вычитание фонового шума направлено на устранение технических погрешностей сканера. Разработчики чипов Affymetrix предложили метод MAS 5.0 (MicroArray Suite 5.0) [1], в котором из значений интенсивности свечения проб вычитаются фоновые значения, усреднённые по нескольким секторам на микрочипе. Популярный на сегодняшний день метод RMA (Robust Multi Average) [10] вычитает фоновый шум, опираясь на предположение о нормальности распределения интенсивности свечения проб. В методе DFCM (Distribution Free Convolution Model) [5] не делается предположений о распределении интенсивностей, но вычитается определённая квантиль значений интенсивностей свечения на микрочипе.

Работа выполнена при финансовой поддержке Минобрнауки РФ, государственный контракт № 16.512.11.2222.

Нормализация данных необходима для приведения интенсивностей на нескольких микрочипах к одному распределению для сравнения данных на разных стадиях эксперимента. Наиболее часто используемым методом является квантильная нормализация, применяемая в комплексах RMA, DFCM и FARMS (Factor Analysis For Robust Multiarray Summarization) [8].

Метод суммаризации FARMS основан на предположении о нормальности распределения логарифма интенсивности свечения проб. Каждому гену соответствует несколько проб на микрочипе, и *суммаризацией* называется определение значения экспрессии гена по интенсивности свечения всех соответствующих ему проб. Другими методами суммаризации является взвешенное Тьюки (Tukey biweight), используемое в MAS 5.0 и DFCM, и MedianPolish, используемый в RMA.

Недостатком вышеназванных методов является то, что они не принимают во внимание физические процессы, протекающие на этапе гибридизации. Эти методы ориентированы на определение изменения уровня экспрессии одного гена в пределах одного эксперимента. Они оказываются неприменимы, если требуется сравнить уровни экспрессии разных генов или одного и того же гена в разных экспериментах.

Во многих работах отмечалось, что при анализе данных микрочипов ДНК необходимо учитывать неизбежно возникающие эффекты кросс-гибридизации [4, 6] и насыщения [7, 9, 11]. В данной работе мы попытались построить гибкую модель, позволяющую учесть неспецифические взаимодействия (кросс-гибридизацию) и как линейный, так и нелинейный характер зависимости интенсивности от концентрации.

Предлагаемые модели данных

Модель кросс-гибридизации. Пробы на поверхности чипа могут вступать в реакцию с молекулами специфических генов, которым они полностью комплементарны, и с молекулами некоторых других генов, которым они комплементарны лишь частично. Чтобы описать эти взаимодействия, введём матрицу $A = (A_{ij})$, которую назовём матрицей взаимодействия проб и генов. Каждый коэффициент $A_{ij} \geqslant 0$ матрицы A выражает склонность i-й пробы к взаимодействию с j-м геном. Матрица A имеет размерность $P \times N$, где P — количество проб на микрочипе, N — число генов, состав которых известен и взаимодействие с которыми предусмотрено разработчиками микрочипа. Микрочипы Affymetrix ориентированы на распознавание более 22 000 генов и содержат от 500 000 до 900 000 проб. Каждая проба соответствует только одному специфическому гену и, возможно, может вступать в неспецифические взаимодействия с несколькими другими генами. Таким образом, матрица A является сильно разреженной матрицей большой размерности.

В предлагаемой модели интенсивность свечения пробы выражается через концентрации всех генов, которые могут быть в образце:

$$I_i^t = d^t \sum_j A_{ij} C_j^t + b_i^t,$$

где t — номер чипа, i — номер пробы; суммирование идёт по всем генам, которые могут присутствовать в образце, $C_j^t \geqslant 0$ — концентрация j-го гена на t-м чипе, d^t — параметр нормализации, b_i^t — фоновая поправка.

Параметры d^t предлагается вычислять с помощью алгоритма квантильной нормализации, рекомендуемого в [3]. Фоновую поправку b_i^t предлагается вычислять алгоритмом MedianPolish [10], учитывающим информацию обо всех микрочипах в эксперименте, тем самым позволяющим учесть индивидуальные характеристики проб.

Построение матрицы взаимодействий A проб с генами опирается на предположение о том, что некомплементарные молекулы не могут вступать в реакцию. Для оценки возможных взаимодействий проб и генов предлагается восстановить матрицу A^{blast} с помощью алгоритма выравнивания символьных последовательностей BLASTN [2]. Для построения матрицы A^{blast} на вход алгоритму BLASTN были поданы последовательности проб, предложенных Affymetrix [1], и последовательности генов из сборки человеческого генома версии 18, использованной при создании данных чипов. Отметим, что последовательности проб и генов зависят от модели микрочипа. На выходе алгоритма была построена матрица A^{blast}.

Элементы матрицы A_{ij}^{blast} определяются как число совпадающих нуклеотидов в последовательности i-й пробы и j-го гена. Если коэффициент матрицы A_{ij}^{blast} равен нулю, то проба и ген некомплементарны. После построения матрицы A^{blast} в ходе минимизации квадратичной невязки восстанавливаются только те коэффициенты матрицы A, для которых $A_{ij}^{\text{blast}} > 0$, остальные элементы матрицы A_{ij} полагаются равными нулю.

Комбинированная модель. Предложенная выше модель кросс-гибридизации восстанавливает линейную зависимость между интенсивностью свечения проб и концентрацией генов. Wu et al. [12] тоже предлагали линейную модель кросс-гибридизации, так как при неспецифических взаимодействиях насыщения не возникает. Однако Cambon et al. [4] показали, что построить модель интенсивности, учитывая только кросс-гибридизацию, с достаточной точностью не представляется возможным.

Было предложено скомбинировать модель кросс-гибридизации с моделью адсорбции Ленгмюра [7, 11], описывающей эффект насыщения проб. Зависимость интенсивности от концентрации генов с учётом насыщения выглядит следующим образом:

$$I_i^t = d^t\left(\frac{\alpha_i C_{j_i}^t}{1+\beta_i C_{j_i}^t} + \sum_{j,j\neq j_i} A_{ij} C_j^t + \gamma_i\right) + b_i^t,$$

α_i, β_i — параметры насыщения, γ_i — параметр неучтённой кросс-гибридизации, j_i — номер специфического гена для пробы i.

Нелинейная составляющая специфических взаимодействий немного видоизменена относительно базовой модели Ленгмюра. При $\beta_i = 0$ модель описывает линейную зависимость интенсивности от концентрации. Дополнительно вводится параметр неучтённых взаимодействий γ_i. Они могут возникнуть в том случае, если в образце присутствуют молекулы генов, не участвовавших при построении матрицы взаимодействий A, либо если по каким-то причинам произошла реакция пробы с некомплементарным геном.

Параметры модели $\alpha_i, \beta_i, \gamma_i, A = (A_{ij})$ настраиваются одновременно в ходе минимизации функционала квадратичной невязки

$$Q = \sum_t \left(\sum_i (I_i^t - Y_i^t)^2 + \sum_{ij} |A_{ij}|\lambda_i\right) \to \min_{\alpha,\beta,\gamma,A},$$

где Y_i^t — наблюдаемая интенсивность i-й пробы на t-м чипе, λ_i — параметры регуляризации.

Коэффициенты матрицы A можно искать тем же способом, который был использован в модели кросс-гибридизации. Предлагается восстанавливать только те коэффициенты A_{ij}, для которых $A_{ij}^{\text{blast}} > 0$. Для этого соответствующие значения λ_i полагаются равными 0 или 1.

Поиск исходных концентраций генов осуществляется путём минимизации функционала

$$Q = \sum_t \sum_i \left(I_i^t(C) - Y_i^t\right)^2 \to \min_C, \qquad C \geqslant 0.$$

Далее предоставлены результаты идентификации параметров модели по экспериментальным данным и сравнения её качества с существующими методами оценки экспрессии генов.

Эксперименты

Исходные данные «Латинский Квадрат». Эксперимент «Латинский квадрат» был проведён на платформе Affymetrix специально для исследования и тестирования методов обработки данных микрочипов ДНК. В ходе эксперимента в образец с РНК клеточной линии HeLa (ATCC CCL-13) были добавлены РНК 42 генов, изначально не содержащихся в данном образце. Все гены были одинаковым образом выделены и подготовлены, а затем в известных концентрациях нанесены на микрочипы. Концентрация принимала значения {0, 0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512} пМ. Гены были сгруппированы по три в одной пробирке. Эксперимент состоял из 14 стадий, на каждой из которых на микрочип были нанесены смеси генов из разных пробирок в разных концентрациях. На каждой стадии одинаковые смеси были нанесены на три микрочипа — так называемые *технические репликаты*, служащие для контроля за воспроизводимостью эксперимента.

Таким образом, в нашем распоряжении оказались 42 микрочипа (14×3) с известными интенсивностями свечения проб и концентрациями 14-ти образцов (по три гена в каждом).

Восстановление концентраций генов. При восстановлении исходных концентраций генов была использована комбинированная модель с матрицей взаимодействий A, полученной с ограничениями, накладываемыми матрицей A^{blast}. Было проведено сравнение предложенной комбинированной модели с методами MAS 5.0, RMA, DFCM и FARMS.

На Рис. 1 изображены восстановленные изменения уровня экспрессии относительно концентрации в 1 пМ. Прямая чёрная линия обозначает истинные изменения уровня экспрессии. Как можно заметить, комбинированная модель предсказывает изменение уровня экспрессии не хуже, чем распространённые методы. Более того, в отличие от остальных методов, предложенная модель не имеет тенденции завышать оценку для маленьких концентраций и занижать её для больших.

На Рис. 2 изображены восстановленные логарифмированные концентрации, прямая чёрная линия — истинные значения. Распространённые методы, не ориентированные на точное восстановление концентраций, сильно завышают значения концентраций. Предложенная модель гораздо лучше приближает исходные концентрации.

Таким образом, можно сделать вывод, что предложенная комбинированная модель, учитывающая кросс-гибридизацию и эффект насыщения, более предпочтительна как для определения исходных концентраций, так и для оценки изменения уровня экспрессии.

Выводы

В данной работе была исследована актуальная проблема анализа данных микрочипов ДНК и рассмотрены различные подходы к её решению. Предложена гибкая модель данных, учитывающая технические вариации, эффект кросс-гибридиза-

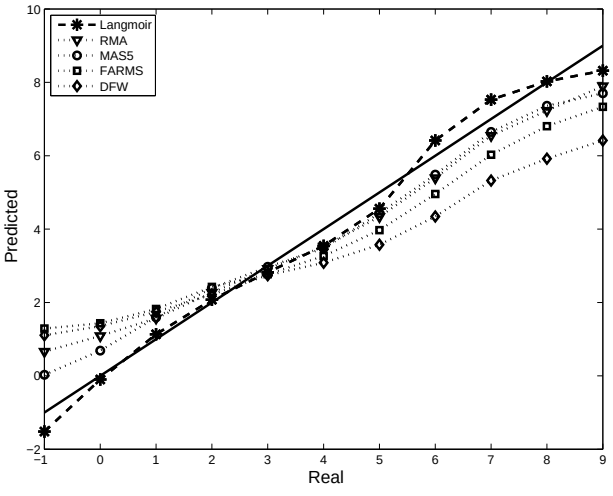


Рис. 1. Восстановленные изменения концентрации (Fold Change). Комбинированная модель обозначена длинным пунктиром, истинные значения — сплошной линией.

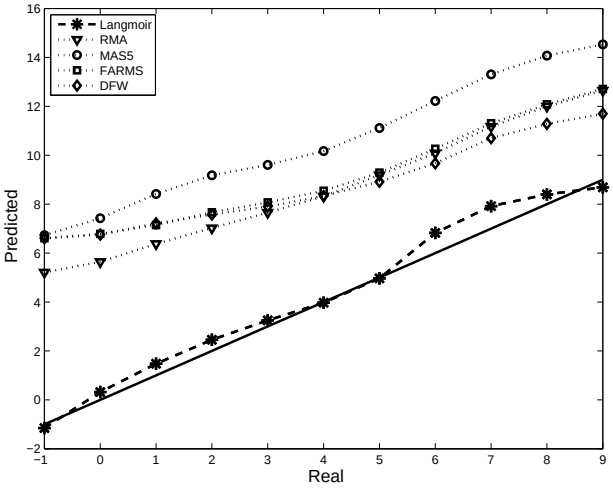


Рис. 2. Восстановленные концентрации ($\log_2 C$). Комбинированная модель обозначена длинным пунктиром, истинные значения — сплошной линией.

ции и допускающая как линейный, так и нелинейный характер зависимости интенсивности от концентрации. Проведённые вычислительные эксперименты показали, что модель даёт более точные результаты, чем многие существующие методы. Однако применение модели для определения экспрессии генов на микрочипах других моделей остаётся предметом исследования.

Литература

[1] *Affymetrix Inc.* Statistical Algorithms Description Document `http://www.affymetrix.com/support/technical/whitepapers.affx`

[2] *Altschul S. F., Gish W., et al.* Basic local alignment search tool // Journal of Molecular Biology, 1990. — Vol. 215, No. 3. — Pp. 403–410.

[3] *Bolstad B. M., Irizarry R. A., et al.* A comparison of normalization methods for high density oligonucleotide array data based on variance and bias // Bioinformatics, 2003. — Vol. 19, No. 2. — Pp. 185–193.

[4] *Cambon A. C., Khalyfa A., et al.* Analysis of probe level patterns in Affymetrix microarray data // BMC bioinformatics, 2007. — Vol. 8, No. 1. — Pp. 146–156.

[5] *Chen Z., McGee M., Liu Q.* A Distribution-Free Convolution Model for background correction of oligonucleotide microarray data // BMC genomics, 2009. — Vol. 10. — Art. S19.

[6] *Furusawa C., Ono N., et al.* Model-based analysis of non-specific binding for background correction of high-density oligonucleotide microarrays // Bioinformatics, 2009. — Vol. 25, No. 1. — Pp. 36–41.

[7] *Hekstra D., Taussig A. R., et al.* Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays // Nucleic Acids Research, 2003. — Vol. 31, No. 7. — Pp. 1962–1968.

[8] *Hochreiter S., Clevert D. A., Obermayer K.* A new summarization method for Affymetrix probe level data // Bioinformatics, 2006. — Vol. 22, No. 8. — Pp. 943–949.

[9] *Held G. A., Grinstein G., Tu Y.* Modeling of DNA microarray data by using physical properties of hybridization // Proceedings of the National Academy of Sciences of the United States of America, 2003. — Vol. 100, No. 13. — Pp. 7575–7580.

[10] *Irizarry R. A., Hobbs B., et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data // Biostatistics, 2003. — Vol. 4, No. 2. — Pp. 249–264.

[11] *Mulders G. C., Barkema G. T., Carlon E.* Inverse Langmuir method for oligonucleotide microarray analysis // BMC Bioinformatics, 2009. — Vol. 10. — Pp. 64–72.

[12] *Wu C., Carta R., Zhang L.* Sequence dependence of cross-hybridization on short oligo microarrays // Nucleic acids research, 2005. — Vol. 33, No. 9. — Art. e84.

Нижняя граница числа комплементарных нуклеотидов при моделировании кросс-гибридизации

Рябенко Е. А., Когадеева М. С.
riabenko.e@gmail.com, m.kogadeeva@gmail.com
Москва, Московский Государственный Университет им. М. В. Ломоносова

Рассматривается проблема моделирования кросс-гибридизации в рамках задачи анализа данных ДНК-микрочипов. В качестве модельных данных изучается интенсивность флуоресценции проб, соответствующих генам, экспрессия которых не характерна для образцов на имеющейся выборке микрочипов по данным исследований тканеспецифичности. Показано, что использование информации о числе комплементарных нуклеотидов в паре «проба–ген» позволяет добиться значительного сокращения числа определяемых коэффициентов, снижая вычислительную сложность задачи.

ДНК-микрочипы — это технология, позволяющая проводить полномасштабный анализ экспрессии генов, что способствует выявлению и изучению новых молекулярно-генетических механизмов в клетке. В основе технологии лежит принцип комплементарности — свойство одинарных последовательностей нуклеотидов соединяться в двухцепочечную молекулу (аденин соединяется с тимином, цитозин — с гуанином). Микрочип представляет собой небольшую пластину, на которой в определённых позициях закреплены короткие нуклеотидные последовательности (пробы), состав каждой из которых известен. В ходе эксперимента на микрочип наносят выделенную из исследуемого образца смесь, содержащую помеченные флуоресцентными метками одноцепочечные участки молекул кДНК. В результате реакции гибридизации цепочки с комплементарными последовательностями соединяются. Далее производят сканирование микрочипа и по интенсивности флуоресценции дают оценку количества молекул РНК конкретных генов, содержавшихся в исходном образце. Более подробное описание устройства и принципов работы ДНК-микрочипов приведено в статье [1] настоящего сборника.

Несмотря на широкую распространённость микрочипов, методы анализа получаемых с их помощью данных всё ещё далеки от совершенства, поскольку учесть все влияющие на результат эксперимента шумовые факторы практически невозможно. Одним из основных неучтённых источников шума остаётся вариация эффективности гибридизации проб. На реакцию гибридизации, то есть связывания нуклеотидных цепочек в образце с пробами на поверхности микрочипа, влияет множество факторов, важнейшие из которых связаны с нуклеотидным составом молекул [2]. Даже если молекула кДНК и соответствующая проба полностью комплементарны, эффективности гибридизации разных проб могут сильно отличается в зависимости от состава пробы. Это делает практически невозможным количественное сравнение уровня экспрессии различных генов на одном ДНК-микрочипе. Дополнительная сложность заключается в том, что гибридизация может проходить и между цепочками, комплементарными лишь частично — так называемая кросс-гибридизация. В этом случае интенсивность флуоресценции пробы будет зависеть не только от экспрессии её целевого гена, но и от экспрессии генов, в которых встречаются участки с частично комплементарной этой пробе последовательностью. Учёт этой зависимости является важным этапом, необходимым для улучшения методов анализа микрочиповых данных [3, 4].

Поскольку явной модели процесса гибридизации, учитывающей все влияющие на неё факторы, на данный момент не построено [5], в [1] было предложено неявно учитывать её эффект за счёт введения матрицы A взаимодействий «проба–ген», определять которую предполагается путём решения оптимизационной задачи. Простейшая версия модели имеет следующий вид:

$$I_i^t = \sum_j A_{ij} C_j^t,$$

Здесь I_i^t — предобработанная (см. [1]) интенсивность флуоресценции i-й пробы на t-м чипе, $C_j^t \geqslant 0$ — уровень экспрессии j-го гена в t-м образце, суммирование идёт по всем генам, которые могут присутствовать в образце.

Матрица A_{ij} имеет размер, равный числу генов, умноженному на число проб, что, например, на микрочипах Affymetrix последних поколений составляет порядка 30000×1000000. Ясно, что определение такого количества коэффициентов сопряжено как с вычислительными трудностями, так и с переобучением. Чтобы исключить из списка настраиваемых параметров как можно больше коэффициентов A_{ij}, предлагается использовать имеющуюся информацию о пробах и соответствующих генам молекулах РНК. Это можно сделать, опираясь на предположение о том, что некомплементарные молекулы не могут вступать в реакцию. Для оценки возможных взаимодействий проб и генов предлагается построить матрицу A^{blast}с помощью алгоритма выравнивания символьных последовательностей BLASTN [6]. Коэффициенты матрицы взаимодействий A_{ij}^{blast} содержат информацию о числе

комплементарных нуклеотидов в последовательностях пробы j и гена i.

Целью данной работы было определение порогового числа совпадений в последовательностях нуклеотидов, ниже которого кросс-гибридизация пары «проба–ген» пренебрежимо мала и не оказывает значимого воздействия на результат микрочипового анализа.

Анализируемые данные

Для исследования были использованы данные эксперимента с чипами Affymetrix Human Gene 1.0 ST, содержащими 861493 проб к 31273 генам человека. В ходе эксперимента изучалась экспрессия генов в клетках крови испытуемых. В распоряжении имелось 70 микрочипов с образцами. Перед дальнейшим анализом была проведена предобработка данных (фоновая поправка и квантильная нормализация) с использованием стандартного метода RMA [7].

Для построения матрицы A^{blast} на вход алгоритму BLASTN были поданы последовательности проб, предложенных Affymetrix, и последовательности генов из сборки человеческого генома версии 18, использованной при создании данных чипов.

Для изучения кросс-гибридизации было предложено использовать информацию о тканеспецифичных генах. Известно, что в различных типах клеток профили экспрессии могут значительно отличаться, что объясняется различием их функций. Так как исследуемые образцы были получены из клеток крови, для анализа шумовых факторов в микрочиповых данных интерес представляли гены, не экспрессируемые в клетках крови. Предполагалось, что флуоресценция проб, соответствующих таким генам, объясняется исключительно шумовыми факторами.

В качестве источника информации о тканеспецифичных генах были использованы базы данных Gene Expression Barcode [8] и TiGER [9]. Первая построена при помощи мета-анализа более чем 30 тыс. микрочипов Affymetrix. Вторая сформирована на основе базы NCBI EST, содержащей информацию о транскриптах, соответствующих каждому гену, и их наличии в различных типах тканей; ген считается специфичным, если число соответствующих ему транскриптов в данной ткани значимо выше случайного. В обеих базах были выбраны все гены, специфичные для всех типов тканей, кроме крови. В итоговый список попал 1271 ген, экспрессия каждого из которых нехарактерна для клеток крови по данным обеих баз. К этим генам на изучаемых микрочипах имеется 41286 проб.

На рис. 1 приведены эмпирические плотности распределений интенсивностей флуоресценции проб к отобранным генам и ко всем остальным

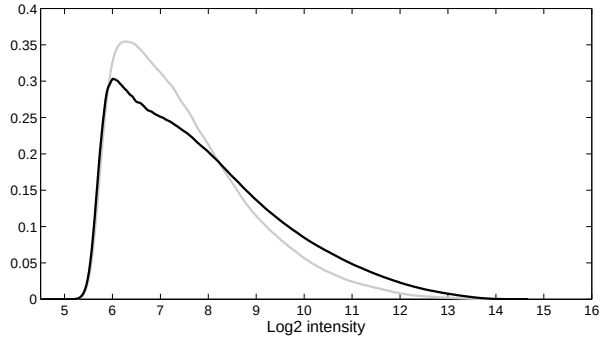


Рис. 1. Распределения интенсивностей флуоресценции проб, серый — к генам, экспрессия которых нехарактерна для крови, чёрный — ко всем остальным.

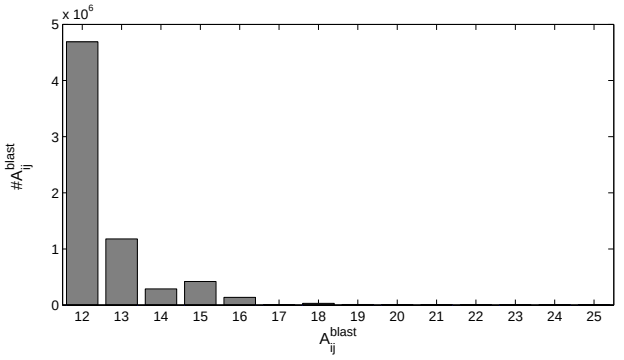


Рис. 2. Распределение коэффициентов A_{ij}^{blast} для 41286 проб к генам, неспецифичным для крови.

на имеющихся 70 микрочипах. Двухвыборочный критерий Дарлинга отклонил гипотезу об одинаковой распределённости интенсивностей в двух группах ($p < 10^{-16}$).

На рис. 2 показано распределение ненулевых значений матрицы A^{blast} для проб к отобранным генам. Поскольку длина проб на рассматриваемых микрочипах составляет 25 нуклеотидов, гибридизация к ним последовательностей, комплементарных менее, чем на 12 символов (меньше половины длины пробы), практически невозможна, поэтому все элементы A^{blast}, меньшие 12, были обнулены.

Анализ эффекта кросс-гибридизации

Для каждой из 41286 проб к отобранным генам была поставлена следующая задача:

$$I_i^t = \sum_{j=1}^{n_j} A_{ij} C_j^t,$$
$$A_{ij} \geqslant 0,$$
$$A_{ij} = 0, \text{ если } A_{ij}^{\text{blast}} \leqslant k.$$

Здесь $t = 1, \ldots, 70,\ \ i = 1, \ldots, 41286,\ \ k = 12, \ldots, 25$; в качестве C_j^t взяты оценки экспрессии, построенные стандартным методом (RMA, [7]). Задача решалась методом наименьших квадратов с ограничением на неотрицательность коэффициентов. Вы-

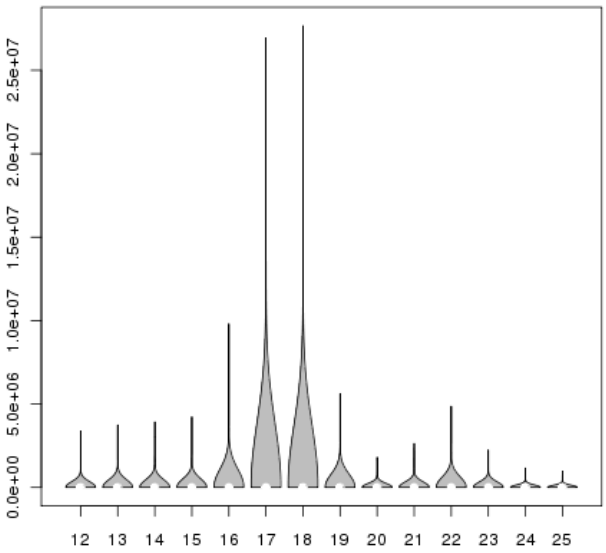


Рис. 3. Распределения среднеквадратичных невязок при различных значениях порога k.

борка из 70 микрочипов была многократно разбита на обучающую и контрольную в отношении 6:1, для каждого разбиения проводилась настройка коэффициентов a_{ij} на обучающей выборке и вычисление среднеквадратичной невязки на контроле.

На рис. 3 показаны распределения значений ошибки на контроле для различных значений порога k. По ним можно сделать следующие выводы:

— повышение порога k от 12 до 15 практически не приводит к увеличению среднеквадратичной ошибки модели на контроле, в то время как число ненулевых коэффициентов A_{ij} значительно уменьшается (см. рис. 2);
— максимальные значения среднеквадратичной ошибки возникают при установлении значениях порога k равным 17–18.

Последний факт объясняется тем, что около половины (18748 из 41286) рассматриваемых проб не комплементарны никаким генам кроме тех, к которым они сделаны, более чем на 17 нуклеотидов. Влияние кросс-гибридизации на интенсивность флуоресценции некоторого подмножества таких проб, по всей видимости, невелико, в связи с чем модель и демонстрирует высокий уровень ошибки на контроле.

Дальнейший анализ показал наличие положительной корреляции между среднеквадратичной ошибкой модели на контроле и интенсивностью пробы, для которой проводилась настройка модели. Скорее всего, высокие интенсивности флуоресценции проб объясняются тем, что соответствующие им гены на самом деле присутствуют в образце, и в сигнале есть не только компонента, соответствующая кросс-гибридизации. Такой результат может объясняться как тем, что базы, взятые в качестве источника информации о тканеспецифичности генов, могут содержать ошибки, так и тем, что строгая тканеспецифичность в природе недостижима.

Выводы

В работе изучалось влияние кросс-гибридизации на интенсивность флуоресценции проб ДНК-микрочипа путём отбора генов, экспрессия которых нехарактерна для изучаемого типа ткани (кровь). Было показано, что при моделировании эффекта кросс-гибридизации разумно рассматривать только те пары «проба–ген», которые комплементарны не менее, чем на 15 из 25 нуклеотидов. Это позволяет значительно снизить число настраиваемых параметров ценой не слишком сильного увеличения ошибки модели.

Литература

[1] *Когадеева М. С., Рябенко Е. А.* Математическая модель данных микрочипов ДНК, учитывающая эффекты кросс-гибридизации и насыщения // ММРО-15: доклады XV Всеросс. конф. Сборник докладов, Москва: МАКС Пресс, 2011. — С. 536–539.

[2] *Wu C., Carta R., Zhang L.* Sequence dependence of cross-hybridization on short oligo microarrays // Nucleic acids research, 2005. — Vol. 33, No. 9. — Art. e84.

[3] *Cambon A. C., Khalyfa A., et al.* Analysis of probe level patterns in Affymetrix microarray data // BMC bioinformatics, 2007. — Vol. 8, No. 1. — Pp. 146–156.

[4] *Furusawa C., Ono N., et al.* Model-based analysis of non-specific binding for background correction of high-density oligonucleotide microarrays // Bioinformatics, 2009. — Vol. 25, No. 1. — Pp. 36–41.

[5] *Koltai H., Weingarten-Baror C.* Specificity of DNA microarray hybridization: characterization, effectors and approaches for data correction // (ucleic acids research, 2008. — Vol. 36, No. 7. — Pp. 2395–2405.

[6] *Altschul S. F., Gish W., et al.* Basic local alignment search tool // Journal of Molecular Biology, 1990. — Vol. 215, No. 3. — Pp. 403–410.

[7] *Irizarry R. A., Hobbs B., et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data // Biostatistics, 2003. — Vol. 4, No. 2. — Pp. 249–264.

[8] *McCall M.,N., Uppal K., et al.* The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes // Nucleic acids research, 2011. — Vol. 39. — Pp. 1011–1015.

[9] *Liu X., Yu X., et al.* TiGER: a database for tissue-specific gene expression and regulation // BMC bioinformatics, 2008. — Vol. 9. — Art. 271.

Об одном алгоритме поиска плотных областей и его геофизических приложениях

Агаян С. М., Богоутдинов Ш. Р., Добровольский М. Н.
m.dobrovolsky@gcras.ru
Москва, Геофизический центр РАН

В работе предложен алгоритм выделения плотных областей в конечных множествах точек на основе формальной конструкции плотности. Даны результаты работы алгоритма на искусственных примерах и на реальных данных, полученных по результатам гравиметрических измерений региона Молуккского моря (Индонезия).

В многомерном конечном массиве X любой природы особую роль играют плотные области. Они важны для содержательного анализа X. Так, например, в геолого-геофизических приложениях такие области часто бывают кластерами или трассами в X.

В работе предлагается алгоритм выделения областей повышенной плотности в X на основе формальной конструкции плотности. Если пространство X удовлетворяет условиям стандартного кластерного анализа, т. е. допускает естественное расслоение на однородные части относительно некоторой меры близости [1], то предлагаемый алгоритм при соответствующем выборе параметров даёт требуемое разбиение и поэтому может рассматриваться как алгоритм кластерного анализа.

Построение множества $A(\alpha)$

Пусть X конечное множество, а A, B, … и x, y, … соответственно подмножества и точки в нём.

Определение 1. *Назовём плотностью P на множестве X отображение из $2^X \times X$ в отрезок $[0,1]$, возрастающее по первому аргументу:*

$$P(A,x) = P_A(x)$$
$$\forall x \in X\colon A \subseteq B \Rightarrow P_A(x) \leqslant P_B(x).$$

$P_A(x)$ есть плотность подмножества A в точке x.

Для плотности P, заданной на X, подмножества A и уровня $\alpha \in [0,1]$ построим последовательность α-n-оболочек подмножества A во множестве X по плотности P:

$$A^1 = \{x \in X\colon P_A(x) \geqslant \alpha\},$$
$$\dots$$
$$A^n = \{x \in X\colon P_{A\cup A^{n-1}}(x) \geqslant \alpha\},$$
$$\dots$$

Утверждение 1. $A^1 \subseteq \dots \subseteq A^n \subseteq \dots$.

Замечание 1. Так как плотность P есть возрастающая функция множества, то все α-n-оболочки подмножества A содержатся в α-1-оболочке всего множества X:

$$\forall n\ A^n \subseteq X^1 = \{x \in X\colon P_X(x) \geqslant \alpha\}.$$

В силу конечности множества X в неубывающей и ограниченной последовательности α-n-оболочек, начиная с некоторого номера n^*, наступит стабилизация:

$$A^1 \subset \dots \subset A^{n^*} = A^{n^*+1} = \dots .$$

Определение 2. *Назовём множество A^{n^*} α-∞-оболочкой подмножества A и обозначим через A^∞.*

Утверждение 2. *A^∞ содержит свою α-оболочку по плотности:*

$$(A^\infty)^1 \subseteq A^\infty.$$

Следствие 1. *Последовательность α-n-оболочек для A^∞ является постоянной:*

$$\forall n\ (A^\infty)^n = (A^\infty)^1.$$

Следствие 2. *Для A^∞ α-∞-оболочка совпадает с α-1-оболочкой:*

$$(A^\infty)^\infty = (A^\infty)^1.$$

Обозначим α-∞-оболочку для A^∞ через $A^{2\infty}$. Имеем:

$$A^{2\infty} = (A^\infty)^\infty = (A^\infty)^1 \subseteq A^\infty.$$

Последовательно строя α-∞-оболочки по плотности P, получим следующую схему:

$$A \to A^1 \subseteq \dots \subseteq A^\infty$$
$$A^\infty \supseteq (A^\infty)^1 = A^{2\infty}$$
$$\dots$$
$$A^{m\infty} \supseteq (A^{m\infty})^1 = A^{(m+1)\infty}$$
$$\dots$$

В силу конечности X в невозрастающей последовательности

$$A^\infty \supseteq \dots \supseteq A^{m\infty} \supseteq \dots ,$$

начиная с некоторого номера m^* наступит стабилизация:

$$A^\infty \supset \dots \supset A^{m^*\infty} = A^{(m^*+1)\infty} = \dots .$$

Определение 3. *Множество $A^{m^*\infty}$ будем обозначать через $A(\alpha)$.*

Замечание 2. Процесс построения $A(\alpha)$ имеет стадию возрастания от A^1 до A^∞ и стадию убывания от A^∞ до $A(\alpha)$:

$$A \to A^1 \subset \cdots \subset A^{n^*} = A^\infty \supseteq \cdots \supseteq A^{m^*\infty} = A(\alpha).$$

Утверждение 3. *$A(\alpha)$ совпадает со своей α-оболочкой:*

$$(A(\alpha))^1 = A(\alpha).$$

Следствие 3. *Результатом повторного применения алгоритма к $A(\alpha)$ будет само $A(\alpha)$:*

$$(A(\alpha))(\alpha) = A(\alpha).$$

Замечание 3. Оператор построения $A(\alpha)$ по A в пространстве X является идемпотентным.

Замечание 4. Множество $A(\alpha)$ состоит ровно из тех точек, где его плотность больше или равна α:

$$A(\alpha) = \{x \in X \colon P_{A(\alpha)}(x) \geqslant \alpha\}.$$

Во всех точках дополнения плотность $A(\alpha)$ меньше α:

$$\overline{A(\alpha)} = \{x \in X \colon P_{A(\alpha)}(x) < \alpha\}.$$

Неформальная интерпретация $A(\alpha)$**.** Будем понимать плотность $P_A(x)$ как меру предельности точки x для множества A. Точки x с достаточно большой плотностью

$$P_A(x) \geqslant \alpha$$

считаем предельными для A. Множество A^∞ содержит все свои α-предельные точки из X и является в этом смысле замкнутым. Множество $A(\alpha)$, содержащееся в A^∞, совпадает со множеством своих α-предельных точек из X и является в этом смысле совершенным.

Зависимость $A(\alpha)$ от параметров

Результат работы алгоритма построения $A(\alpha)$ зависит от трёх составляющих: самого множества A, плотности P и уровня α:

$$A(\alpha) = A_P(\alpha).$$

Зависимость по A и P является возрастающей, а по α убывающей.

Утверждение 4. *Если $A \subseteq B$, то $A(\alpha) \subseteq B(\alpha)$.*

Утверждение 5. *Если P и Q плотности на X и для любого подмножества $B \subseteq X$ во всех точках $x \in X$ выполнено*

$$P_B(x) \leqslant Q_B(x),$$

то

$$A_P(\alpha) \subseteq A_Q(\alpha).$$

Утверждение 6. *Если $\alpha < \beta$, то $A(\alpha) \supseteq A(\beta)$.*

Свойства и примеры плотности

Плотность $P_A(x)$ при фиксированном A является нечёткой структурой на X. Поэтому при помощи операций нечёткой логики, а также некоторых других, можно получать новые плотности из существующих. Это расширяет возможности конструкции $A(\alpha)$ в изучении пространства X.

Утверждение 7. *1) Если P и Q плотности на X и $R = R(y_1, y_2)\colon [0,1] \times [0,1] \to [0,1]$ неубывающее отображение, то*

$$R(P,Q)_A(x) = R(P_A(x), Q_A(x))$$

будет плотностью на X.

2) Если $\neg$ нечёткое отрицание, то

$$\neg P_A(x) = \neg(P_{\overline{A}}(x))$$

будет плотностью на X.

Следствие 4. *Плотностью будет R-соединение P и $\neg P$:*

$$R(P, \neg P)_A(x) = R(P_A(x), \neg P_A(x)).$$

Следствие 5. *1) Если $\mathrm{T}(\bot, \mathrm{M_w})$ — t-норма (t-конорма, обобщённый оператор осреднения) [2], то*

$$\mathrm{T}(P_A(x), Q_A(x)),$$
$$\bot(P_A(x), Q_A(x)),$$
$$\mathrm{M_w}(P_A(x), Q_A(x)))$$

будут плотностями на X.

2) Если $\lambda \in [0,1]$, то

$$\lambda P_A(x) + (1-\lambda) Q_A(x)$$

и

$$\lambda P_A(x) + (1-\lambda)(1 - P_{\overline{A}}(x))$$

плотности на X.

Конкретные примеры плотностей рассматриваются, когда на множестве X задано расстояние d. Остановимся на двух вариантах определения плотности.

1) **«Количество точек».** Плотность зависит от радиуса близости r и параметра $q \geqslant 0$. Для каждой точки x из множества X рассматривается шар с центром в x радиуса r:

$$D(x,r) = \{y \in X \colon d(x,y) \leqslant r\}.$$

Для каждого шара вычисляется сумма точек множества X в нём с учётом расстояния от точек до центра шара:

$$N_X(x,r) = \sum_{y \in D(x,r)} \left(1 - \frac{d(x,y)}{r}\right)^q.$$

Максимум таких сумм по всем точкам $x \in X$ обозначим через $C(X,r)$:

$$C(X,r) = \max_{x \in X} N_X(x,r).$$

Также для каждого шара вычисляется сумма точек с учётом их удалённости от центра шара только по точкам подмножества $A \subseteq X$:

$$N_A(x,r) = \sum_{y \in D_A(x,r)} \left(1 - \frac{d(x,y)}{r}\right)^q.$$

Здесь $D_A(x,r)$ есть пересечение шара и подмножества A:

$$D_A(x,r) = D(x,r) \cap A.$$

Плотность подмножества $A \subseteq X$ в точке $x \in X$ определяется как отношение суммы точек по шару в A с учётом их удалённости от центра к максимальной сумме по шарам в X:

$$P_A(x) = \frac{N_A(x,r)}{C(X,r)}.$$

2) **«Монолитность»** [3]. Рассматривается проколотая окрестность точки x радиуса r

$$D'_A(x,r) = \{y \in A\colon 0 < d(x,y) \leqslant r\}.$$

Она разбивается на m непересекающихся колец:

$$D'_A(x,r) = \bigcup_{n=1}^{m} S_n,$$

где

$$S_n = \{y \in A\colon r_{n-1} < d(x,y) \leqslant r_n\},$$
$$0 = r_0 < \cdots < r_m = r.$$

Каждому кольцу ставится в соответствие вес

$$1 \geqslant \psi_1 \geqslant \cdots \geqslant \psi_m > 0.$$

Плотность подмножества $A \subseteq X$ в точке $x \in X$ определяется как отношение суммы весов непустых колец к сумме весов всех колец:

$$P_A(x) = \frac{\sum\limits_{\substack{n=1 \\ S_n \neq \varnothing}}^{m} \psi_n}{\sum\limits_{n=1}^{m} \psi_n}.$$

Для проверки работы алгоритма использовались веса, линейно убывающие при удалении от центра:

$$\psi_n = 1 - \frac{n}{m+1}.$$

А также веса, получающиеся из них возведением в положительную степень $k > 0$:

$$\psi_n^k = \left(1 - \frac{n}{m+1}\right)^k.$$

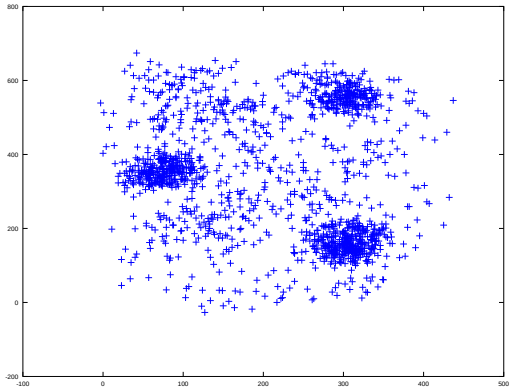


Рис. 1. Исходное множество точек

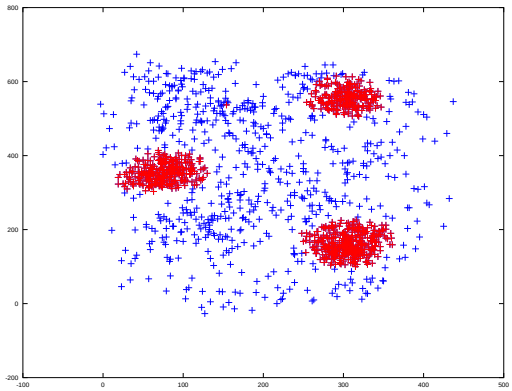


Рис. 2. Результат работы алгоритма при $\alpha = 0{,}1$

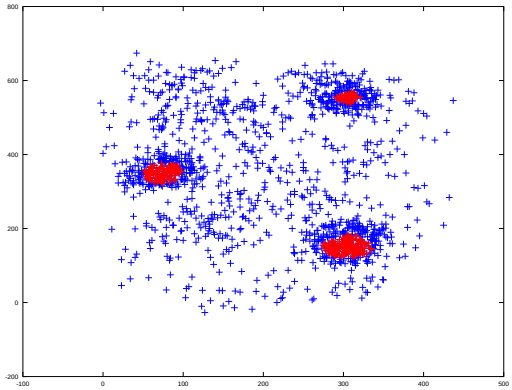


Рис. 3. Результат работы алгоритма при $\alpha = 0{,}35$

Результаты численного эксперимента

Множества $A(\alpha)$ на основе плотностей «Количество точек» и «Монолитность» строились для 14 искусственных примеров для значений α от 0,1 до 0,6 с шагом 0,05. Для расчёта радиуса близости использовалось обобщённое среднее $\mathrm{M_w}$:

$$r = r_w = \mathrm{M_w}(d(x,y)\colon x,y \in X, x \neq y)$$

при $w = -4$; -3,5; -3; -2,5; -1.

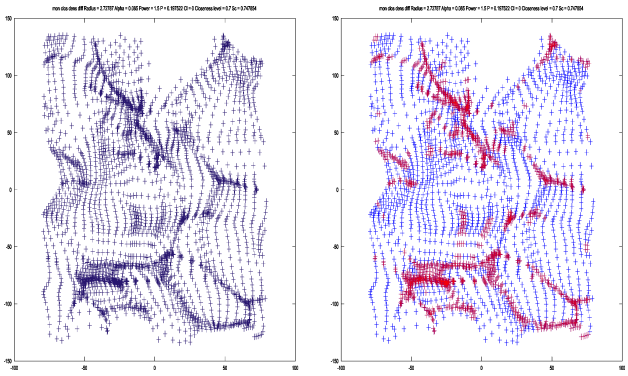


Рис. 4. Результаты работы алгоритма на реальном примере

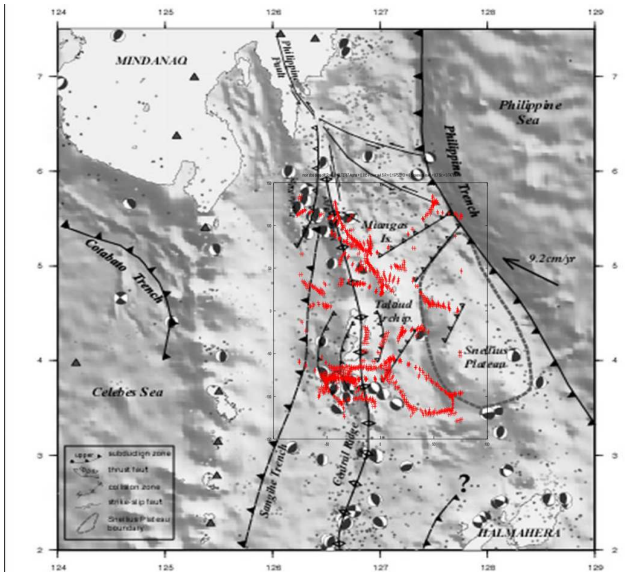


Рис. 5. Результаты работы алгоритма на карте региона

Результаты работы алгоритма при разных значениях параметра α для тестового примера на рис. 1 показаны на рис. 2, 3.

Геофизические приложения

В качестве примера геолого-геофизического приложения алгоритма рассмотрим задачу анализа гравиметрических данных по региону Молуккского моря (Индонезия) [4]. Задача интерпретации гравиметрических данных в полной редукции Буге заключалась в решении обратной задачи методом эйлеровой деконволюции. Однако в регионах со сложной тектонической структурой, где аномальные геофизические поля создаются сложными сочетаниями возмущающих геологических тел, методика эйлеровой деконволюции не всегда позволяет получить хорошо интерпретируемые данные: образуются облака точек, затрудняющие однозначный анализ данных. Поэтому применяются методы поиска участков повышенной плотности и удаления лишних эйлеровых решений. Это позволяет уточнить тектонический план исследуемого участка. Плотные области отражают положение главных возмущающих тел региона. Результаты работы алгоритма на множестве эйлеровых решений показаны на рис. 4. Вид результатов на карте региона показан на рис. 5.

Выводы

В работе построена формальная конструкция плотности для случая конечных множеств. На её основе предложен алгоритм выделения плотных областей в заданном множестве. Работа алгоритма проверена на искусственных примерах для двух вариантов плотности. А также на примере, полученном по реальным геофизическим данным. Остаётся открытым вопрос полной автоматизации выбора параметров алгоритма.

Литература

[1] *Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика: Классификация и снижение размерности: Справочное издание. — М.: Финансы и статистика, 1989. — 607 с.

[2] *Аверкин А. Н., Батыршин И. З., Блишун А. Ф., Силов В. Б., Тарасов В. Б.* Нечёткие множества в моделях управления и искусственного интеллекта. — М.: Наука. Главная редакция физико-математической литературы, 1986. — 312 с.

[3] *Гвишиани А. Д., Агаян С. М., Богоутдинов Ш. Р., Соловьев А. А.* Дискретный математический анализ и геолого-геофизические приложения // Вестник КРАУНЦ. Науки о Земле. — 2010. — № 2. — С. 109–125.

[4] *Widiwijayanti C., Mikhailov V., Diament M., Deplus C., Louat R., Tikhotsky S., Gvishiani A.* Structure and evolution of the Molucca Sea area: constraints based on interpretation of a combined sea-surface and satellite gravity dataset // Earth and Planetary Science Letters, 2003. — Vol. 215. — Pp. 135–150.

Исследование математической модели экологической системы на основе синдромальных представлений распознавания образов*

Котельников И. В., Неймарк Ю. И.
neymark@pmk.unn.ru
г. Нижний Новгород, Научно-исследовательский институт прикладной математики и кибернетики Нижегородского государственного университета им. Н. И. Лобачевского

Приведены результаты исследования математической модели экологической системы, состоящей из двух пар «хищник–жертва», с помощью метода исследования на основе распознавания образов.

На конференции «ММРО-14» в докладе [1] был представлен новый метод исследования многомерных динамических систем (ДС) со многими параметрами, который благодаря применению представлений распознавания образов помог обойти непреодолимые в динамике систем трудности, связанные с многомерностью пространства переменных (фазового пространства) и пространства параметров ДС. В настоящей статье приводятся результаты исследования этим методом математической модели экологической системы [2], состоящей из двух пар «хищник–жертва». Модель достаточно сложная: размерность $n = 4$, число параметров $p = 13$. Она интересна тем, что содержит устойчивые состояния равновесия и их многообразия, предельные циклы и хаотические движения, т. е. полный набор возможных движений, исследуемых рассматриваемым методом.

Некоторые понятия, определения, свойства

Базовыми объектами численного исследования ДС являются их траектории, представляющие собой кривые в пространстве переменных, или фазовом пространстве, которые являются решением системы дифференциальных уравнений.

Траектории различаются по типу. Под типом траектории понимается характер её движения, т. е. стремится ли она к устойчивому состоянию равновесия (точке в пространстве переменных), к устойчивому предельному циклу (замкнутой траектории, по которой отображающая траекторию точка совершает периодические движения) или совершает хаотические движения в ограниченной области пространства переменных, из которой никогда не выходит. В целях упрощения исследования в методе рассматриваются только эти три типа траекторий. Но именно они и определяют прикладное значение конкретной ДС.

Области предельных (при больших значениях времени) состояний ДС в фазовом пространстве (ε–окрестности точек устойчивых состояний равновесия, области ε-окрестностей траекторий предельного цикла, ограниченные области с хаотическими движениями) в теории ДС называются *аттракторами ДС*.

Отдельные фазовые переменные во всей области аттрактора могут удовлетворять условию $x_i = const$. Аттракторы с различными наборами таких фазовых переменных в методе исследования рассматриваются как различные.

Один и тот же аттрактор при заданных параметрах может быть получен на целом ряде различных начальных условий. Это означает, что в фазовом пространстве существует целая область, все начальные условия из которой ведут к данному аттрактору. Такая область называется *областью притяжения аттрактора*. Каждый аттрактор имеет свою область притяжения. Области притяжения различных аттракторов не пересекаются.

При одном и том же наборе параметров ДС может иметь целый набор различных аттракторов в зависимости от различных начальных условий. Полный набор различных аттракторов ДС для конкретного набора параметров в динамике систем называется *фазовым портретом ДС* для данного набора параметров.

Понятия аттрактора, области его притяжения, фазового портрета являются фундаментальными понятиями динамики систем.

Каждому фазовому портрету в пространстве параметров соответствует своя область параметров, в каждой точке которой он существует. Полный набор таких областей для множества фазовых портретов составляет *параметрический портрет ДС*.

Аттракторы кодируются n-разрядным числом, разряды которого в порядке слева направо соответствуют последовательным фазовым переменным. Устойчивое состояние равновесия кодируется цифрой 1. Для $n = 4$ код аттрактора устойчивого состояния равновесия принимает значение 1111. Аттрактор предельного цикла кодируется цифрой 2 с кодом аттрактора 2222, аттрактор с хаотическим движением кодируется цифрой 3 с кодом аттрактора 3333. Если на всех точках траектории в области аттрактора какие-то переменные принимают постоянное нулевое значение, то цифры кода аттрактора в разрядах, соответствующих данным переменным, заменяются цифрой 0. Например, 1011

Работа выполнена при финансовой поддержке РФФИ, проекты № 08-01-00248, № 11-01-00379.

для устойчивого состояния равновесия, 2200 для предельного цикла, 3033 для хаотического аттрактора.

Встречаются случаи, когда на заданном наборе параметров получается множество устойчивых состояний равновесия, в значениях переменных которых наблюдается определённая закономерность. Например, x_2 у всех состояний равновесия принимает значение 0, переменная x_3 у всех состояний равновесия имеет одно и то же ненулевое значение, кодируемое цифрой 1, а переменные x_1 и x_4 принимают каждая некоторое множество различных значений. Переменные, принимающие множественные значения, кодируются цифрой 9. В результате получим код аттрактора 9019. Он будет соответствовать многообразию устойчивых состояний равновесия на плоскости (x_1, x_4). Возможны и другие многообразия. Например, коду аттрактора 9000 соответствует многообразие устойчивых состояний равновесия на оси x_1.

Под синдромами в геометрическом представлении понимаются q-мерные прямоугольные параллелепипеды, обладающие свойством содержать в своём внутреннем объёме, включая поверхность, объекты какого-то одного и только этого класса.

Построение решающего правила распознавания на основе оптимальных тупиковых нечётких тестов и синдромов [3] приводит к покрытию объектов каждого класса своим набором синдромов. Синдромальные покрытия различных классов объектов не пересекаются в пространстве признаков.

Поскольку каждый синдром имеет аналитическое представление в виде двусторонних неравенств вида

$$a_i \leqslant y_i \leqslant b_i, \quad i = 1, \ldots, d\,,$$

где a_i и b_i — некоторые константы, y_i — переменная рассматриваемого пространства, а d — его размерность, то можно говорить о формальном аналитическом представлении полученных в решающем правиле покрытий.

Динамическая система

Экологическая система [2] состоит из двух пар «хищник–жертва». Растения M_1 и M_1' (жертвы) в процессе жизнедеятельности путём фотосинтеза потребляют какое-то вещество M_0 (биоген), запасы которого в окружающей среде ограничены. Расход вещества M_0 компенсируется за счет распада растений M_1 и M_1' и животных M_2 и M_2' после их гибели. Животные M_2 и M_2' (хищники) поддерживают свое существование, поедая растения. Уравнения модели имеют вид:

$$\left.\begin{aligned}
\dot{x_1} &= -\varepsilon_1 x_1 - \gamma_1 \frac{x_1 x_2}{1 + a x_1} + \beta_1 \frac{x_1 x_0}{1 + b x_0}\,; \\
\dot{x_2} &= -\varepsilon_2 x_2 + \gamma_2 \frac{x_1 x_2}{1 + a x_1}\,; \\
\dot{x_3} &= -\varepsilon_3 x_3 - \gamma_3 \frac{x_3 x_4}{1 + a x_3} + \beta_3 \frac{x_3 x_0}{1 + b x_0}\,; \\
\dot{x_4} &= -\varepsilon_4 x_4 + \gamma_4 \frac{x_3 x_4}{1 + a x_3}\,,
\end{aligned}\right\} \qquad (1)$$

где ε_i — коэффициенты смертности, γ_i — коэффициенты потребления, β_i — коэффициенты фотосинтеза, a и b — коэффициенты насыщения, x_0 — количество биогенного элемента M_0, x_1 и x_3 — содержание биогена в жертвах M_1 и M_1', x_2 и x_4 — содержание биогена в хищниках M_2 и M_2'. Полное количество биогенного элемента остаётся постоянным, т. е.

$$\sum\nolimits_0^n x_i = M = const\,.$$

Величина M представляет собой интеграл системы (1), в силу чего система (1) представляет собой дифференциально-алгебраическую систему. Итак, исследуется ДС размерности $n = 4$ с числом параметров $p = 13$.

Результаты исследования

Пожалуй, самым основным результатом исследования является успешное применение нового метода к исследованию новой довольно сложной ДС с полным набором рассматриваемых методом возможных аттракторов: устойчивых состояний равновесия и их многообразий, предельных циклов и хаотических движений.

Начальная выборка наборов параметров состояла из 10 000 наборов. Однако дифференциально-алгебраические уравнения при некоторых наборах параметров и начальных условий могут не иметь решений, что обусловлено выходом соответствующих этим случаям траекторий за пределы положительного квадранта фазового пространства. Таких случаев оказалось 3463, так что все результаты получены на выборке из 6537 наборов параметров.

В результате исследования получен 271 фазовый портрет. Для каждого из них получена p-мерная область параметров, на которой существует фазовый портрет. Число наборов параметров, на которых построены синдромальные покрытия областей параметров фазовых портретов, различно для различных фазовых портретов и колеблется в пределах от 1 до 933. Малым числам наборов параметров соответствуют редкие фазовые портреты, представленные на малых по размеру областях параметрического пространства. Число аттракторов в составе фазовых портретов колеблется в пределах от 1 до 12. Большое число аттракторов в фазовом портрете часто получается за счет повторения в фазовом портрете одних и тех же кодов аттракторов, соответствующих различным устойчивым состояниям равновесия, предель-

ным циклам и хаотическим движениям. Например, фазовый портрет с 12 аттракторами состоит из 12 аттракторов с кодом 1111, соответствующих 12 различным устойчивым состояниям равновесия на заданном наборе параметров. С другой стороны, существует, например, фазовый портрет (1010 0099 1011 2022 3033), включающий два различных аттрактора устойчивых состояний равновесия 1010 и 1011, аттрактор многообразия устойчивых состояний равновесия 0099, аттрактор предельного цикла 2022 и аттрактор хаотического движения 3033 на одном наборе параметров, т. е. полный набор возможных типов аттракторов.

Важной характеристикой фазовых портретов является наличие в них только таких аттракторов, которые соответствуют полному набору видов экологической среды, т. е. не содержат нулей в своих кодовых обозначениях. Из 271 фазового портрета таких фазовых портретов 40.

При построении фазовых портретов зафиксирован 71 аттрактор, 12 из которых не содержат в своих кодах нулевых фазовых переменных, т. е. соответствуют полному набору обитателей экологической среды. Ниже приводятся коды этих 12 аттракторов.

1111 2222 3333 1911 9111 1191 1119 1999 1199

9919 9911 1991.

Число возможных аттракторов ограничено, поскольку существует конечное число n-разрядных кодов на конечном наборе символов. Для 4-мерной ДС число возможных кодов равно 105 с учётом того, что не может быть предельных циклов и хаотических аттракторов с 3 нулевыми фазовыми переменными. Реальное число аттракторов зависит от размеров фиксированной области параметров, из которой выбираются наборы параметров для проведения исследования. Может быть, например, область параметров, на которой существует единственный фазовый портрет с единственным аттрактором в нём. Число 71 аттракторов, полученных в проведённом исследовании, близко к насыщению, поскольку около 90% аттракторов получено на первой половине выборки из 10 000 наборов параметров. Число кодов фазовых портретов на фиксированном числе аттракторов и фиксированном числе применяемых для получения фазовых портретов начальных условий тоже конечно, но может быть довольно большим. На 10 000 наборов параметров число фазовых портретов для исследуемой ДС не проявило тенденции к насыщению. Для больших областей параметров, применяемых для исследования ДС, нахождение полного набора фазовых портретов может оказаться трудоёмкой задачей, но нахождение всех аттракторов представляется необходимым условием для проведения наиболее полного исследования.

Для исследуемой экологической ДС наиболее общими являются задачи нахождения области параметров, на которой существует конкретный аттрактор, и области параметров, на которой существуют фазовые портреты и аттракторы, соответствующие полным наборам животных и растений изучаемой экологической среды. Областью параметров для первой задачи является объединение областей параметров всех фазовых портретов, содержащих выбранный конкретный аттрактор. Для второй задачи областью параметров является объединение областей параметров фазовых портретов с отмеченным в задаче свойством. Исходной информацией для решения обеих задач является обучающая выборка для получения областей параметров фазовых портретов. Каждый набор параметров этой обучающей выборки имеет номер класса в виде порядкового номера полученного на данном наборе фазового портрета в общем списке полученных фазовых портретов. Исходя из этой информации, легко разбить обучающую выборку фазовых портретов на два класса, первый из которых включает наборы параметров, соответствующие постановке задачи, а во второй класс относятся все остальные наборы параметров. На полученной обучающей выборке из двух классов строится синдромальное решающее правило [3]. Синдромальное покрытие объектов первого класса будет соответствовать области пространства параметров, составляющей решение задачи. Полученные синдромальные покрытия могут отличаться не очень высокой достоверностью. Но в предложенном новом методе есть процедура получения из синдромального покрытия области одного синдрома с достоверностью, приближающейся к 1. На основе этой процедуры для первой задачи получен, например, синдром области параметров

$[8,39621 \leqslant M \leqslant 12,9864$
$1,1564 \leqslant \varepsilon_1 \leqslant 1,71512$
$0,14727 \leqslant \varepsilon_2 \leqslant 0,891747$
$0,144218 \leqslant \varepsilon_3 \leqslant 1,39315$
$1,37191 \leqslant \varepsilon_4 \leqslant 1,94888$
$0,470728 \leqslant \gamma_1 \leqslant 1,1209$
$0,730228 \leqslant \gamma_2 \leqslant 1,65749$
$0,615162 \leqslant \gamma_3 \leqslant 2,04867$
$1,57601 \leqslant \gamma_4 \leqslant 2,4821$
$2,1343 \leqslant \beta_1 \leqslant 3,8152$
$1,34745 \leqslant \beta_3 \leqslant 3,04291$
$0,126383 \leqslant a \leqslant 0,403501$
$0,0160175 \leqslant b \leqslant 0,177039]$,

на наборах параметров из которого с достоверностью 0,99 получаются фазовые портреты, содержащие аттрактор предельного цикла с кодом 2222. Для проверки на независимой контрольной выборке из области параметров синдрома была сформирована выборка из 400 наборов параметров случайным выбором на основе равномерного распреде-

ления. Исследование на этой выборке параметров дало следующие результаты. Получено 2 аттрактора с кодами 2222 — предельный цикл и 1111 — устойчивое состояние равновесия. На этих аттракторах получено 3 фазовых портрета: (2222) — один предельный цикл, (2222 2222) — два различных предельных цикла, (1111) — устойчивое состояние равновесия. Последний фазовый портрет не удовлетворяет условию задачи, т. к. не содержит в своём составе аттрактора с кодом 2222. Но такие фазовые портреты получены только на 4 наборах параметров из 400, т. е. достоверность 0,99 для полученного синдрома подтверждается.

Для второй задачи получен синдром области параметров

$[3,90251 \leqslant M \leqslant 16,5098$
$0,0881906 \leqslant \varepsilon_1 \leqslant 1,73521$
$0,141644 \leqslant \varepsilon_2 \leqslant 1,17311$
$0,0258228 \leqslant \varepsilon_3 \leqslant 1,50934$
$0,714768 \leqslant \varepsilon_4 \leqslant 1,95091$
$1,72106 \leqslant \gamma_1 \leqslant 2,85285$
$0,847679 \leqslant \gamma_2 \leqslant 2,26177$
$2,09844 \leqslant \gamma_3 \leqslant 2,94639$
$1,18646 \leqslant \gamma_4 \leqslant 2,78336$
$2,3265 \leqslant \beta_1 \leqslant 3,82336$
$2,14515 \leqslant \beta_3 \leqslant 3,82937$
$0,00760083 \leqslant a \leqslant 0,367437$
$0,0099212 \leqslant b \leqslant 0,559999]$,

на наборах параметров из которого с достоверностью 0,99 получаются фазовые портреты, все аттракторы которых соответствуют полному набору обитателей экологической среды. Аналогично первой задаче сформирована независимая контрольная выборка из 400 наборов параметров, принадлежащих полученному синдрому. В результате исследования на этой выборке наборов параметров получено 4 аттрактора с кодами 3333 — хаотическое движение, 2222 — предельный цикл, 1111, 1110 — два аттрактора устойчивых состояний равновесия и 7 фазовых портретов на этих аттракторах: (3333), (2222), (3333 2222), (1111), (2222 2222), (1111 2222), (1110). Последний фазовый портрет содержит 0 в своём коде, и поэтому является ошибочным, но он получается лишь на двух наборах параметров из 400, что соответствует достоверности 0,995. Что же даёт новый метод исследования ДС с применением идей и методов распознавания образов? Классические методы исследования ДС практически полностью решили проблему исследования ДС размерности $n \leqslant 2$ с таким же соотношением и для числа параметров. Но как только размерность ДС и число параметров превышает 2, так в общем случае сразу возникают проблемы, часто неразрешимые. Возникают они, в частности, потому, что в динамике систем не существует формализованного метода выделения многомерных областей фазового и параметрического пространства, объединяющих объекты с каким-то определённым свойством, например со свойством в каждой точке области параметрического пространства иметь один и тот же фазовый портрет. При этом важно, чтобы эти области выделялись при условии изменения всех фазовых переменных, всех параметров без фиксации какой-либо их части. Это важно, поскольку многие вопросы динамики систем могут быть решены только при этом условии, например вопросы робастной (по параметрам) устойчивости ДС и устойчивости в фазовом пространстве при наличии внешних помех. Выделение таких областей с указанным свойством, как мы видели на примере исследования экологической ДС, успешно реализуется с применением методов распознавания образов.

В настоящее время ведутся работы по синтезу систем квазиинвариантного управления [4]. Это тоже ДС на основе обыкновенных дифференциальных уравнений высокого порядка с большим числом параметров. Рассмотренный выше метод исследования не содержит элементов синтеза ДС, но уже сейчас ясно, что отдельные аспекты метода могут быть применены к выполнению этой важной и интересной темы.

Выводы

Ранее [5] новый метод исследования динамических систем с применением представлений распознавания образов был успешно применён к исследованию динамической системы иммунного ответа организма на вторжение инфекции. Приведённые результаты исследования экологической модели в очередной раз показали, что новый метод может успешно применяться для исследования многомерных динамических систем со многими параметрами при условии изменения всех параметров ДС.

Литература

[1] *Котельников И. В.* Построение параметрического портрета динамической системы на основе синдромальных представлений // ММРО–14. — М.: МАКС Пресс, 2009. — С. 372–375.

[2] *Алексеев В. В., Корниловский А. И.* Автостохастические процессы в биофизических системах // Биофизика, 1982. Вып. 5. С. 890–894.

[3] *Kotel'nikov I. V.* A Sindrome Recognition Method Based on Optimal Irreducible Fuzzy Tests // Pattern Recognition and Image Analysis, 2001. Vol. 11, No. 3. Pp. 553–559.

[4] *Неймарк Ю. И.* Синтез и функциональные возможности квазиинвариантного управления // Автоматика и телемеханика, 2008. № 10. С. 48–56.

[5] *Neimark Yu. I., Kotel'nikov I. V., Teklina L. G.* Study of the Mathematical Model of an Organism's Response to the Intrusion of a Infection Using Methods of Pattern Recognition // Pattern Recognition and Image Analysis, 2009. Vol. 19. No. 1. Pp. 181–185.

Распознавание природно-техногенных объектов по данным гиперспектральных систем аэрокосмического зондирования*

Кондранин Т. В., Козодеров В. В., Дмитриев Е. В.
kondr@kondr.rector.mipt.ru
Московская обл., г. Долгопрудный, Московский физико-технический институт (государственный университет)

Представлены прикладные аспекты тематической обработки данных отечественной гиперспектральной аппаратуры, разработанной в НПО Лептон. Описываются подходы, применяемые на этапах первичной обработки данных и классификации природно-техногенных объектов в сложных атмосферных условиях. Полученные результаты демонстрируются на основе данных самолетных измерительных кампаний лета 2010 г., в районах лесных пожаров.

Разработка методов, алгоритмов и расчетных программ распознавания природно-техногенных объектов по данным гиперспектрального аэрокосмического зондирования находится на переднем крае инновационной деятельности при создании информационно-вычислительных ресурсов решения региональных прикладных задач. Системы гиперспектрального зондирования (высокое пространственное разрешение, сотни спектральных каналов в видимой и ближней инфракрасной области с разрешением в единицы нанометра) открывают новые перспективы мониторинга указанных объектов.

Традиционные подходы к обработке данных дистанционного зондирования — использование вычислительных средств, поставляемых зарубежными фирмами, для интеллектуального анализа получаемой информационной продукции с помощью готового программного обеспечения. В отличие от традиционных подходов к анализу многоспектральных данных (6–7 спектральных каналов) обработка гиперспектральных изображений при таком большом числе спектральных каналов не может быть реализована с помощью стандартных процедур визуально-инструментального дешифрирования снимков, их упрощенного представления в виде «вегетационных индексов» (некоторых комбинаций ограниченного числа измерительных каналов). Требуется обоснование информационного содержания данных гиперспектрального зондирования при реализации новых подходов, в которых дешифрирование снимков уступает место автоматизации распознавания объектов на цифровых изображениях. Необходима оптимизация конкретных спектральных каналов при решении прикладных задач в заданной предметной области. Среди таких задач — оценка состояния лесных и других экосистем в разные периоды их вегетации, рассмотрение влияния разных типов задымления атмосферы от лесных и торфяных пожаров, объединение получаемой новой продукции обработки гиперспектральных изображений с данными наземных лесотаксационных и других обследований выбранных территорий.

В данной публикации иллюстрируются некоторые примеры отображения природно-техногенных объектов (разные типы задымления водоемов от лесных пожаров; разные типы почвенно-растительного покрова) на гиперспектральных аэроизображениях. Результаты начального этапа исследований показаны в работах [1–3].

Исходные данные

Специфика летно-полевых компаний 2010 г. — наличие лесных и торфяных пожаров на тестовой территории летных испытаний аппаратуры в Тверской области. Комплекс приборов, используемых в полете, состоял из: гиперспектрометра, аэрофотоаппарата, системы крепления и подвески аппаратуры на самолете, GPS-навигатора, ноутбука, на который записывались данные съемок в процессе реализации маршрутных полетных заданий. Для каждого из выбранных полетных треков данные гиперспектральной съемки записываются на борту на DVD-диски в специальном формате, предлагаемом разработчиками аппаратуры. В процессе наземной обработки этих дисков проводилась распаковка данных и их последующая обработка.

При практической реализации предлагаемых подходов требуется визуализировать всё множество зарегистрированных спектров для обрабатываемой сцены, в реальном времени провести оконтуривание выделенных объектов с расчетом средних спектров и их изменчивости в пределах этих контуров и осуществить обучение используемого классификатора по соответствующей тестовой выборке. Соответственно, при разработке алгоритмов распознавания образов природно-техногенных объектов по данным гиперспектрального аэрокосмического зондирования должны учитываться два аспекта. Требуется обрабатывать данные очень большого объема, но пространственная и межканальная корреляции соседних элементов разрешения (пикселей) могут быть значительными. Высокое пространственное разрешение, которое может содержать тонкие детали интерпре-

Работа выполнена при финансовой поддержке РФФИ, проекты №11-07-00382-а, №09-05-00171-а и ФЦП НК-568П (г/к П349).

тации конкретной сцены, в некоторых приложениях может не улучшать результаты обработки такой сцены, например вследствие сложной структуры чередования освещенных и затененных элементов для объектов, относящихся к классу «растительность».

Методы обработки данных

Данные гиперспектрального аэрокосмического зондирования позволяют использовать тонкую структуру регистрируемых спектров для повышения информационного содержания обрабатываемых изображений в сравнении с данными многоспектрального зондирования. При этом данные гиперспектрального зондирования содержат информацию о линиях и полосах поглощения излучения в оптической области спектра различными соединениями атмосферы и земной поверхности; эта тонкая структура спектров осредняется аппаратурой многоспектрального зондирования. Вместе с тем, большое число спектральных каналов усложняет проблему классификации природно-техногенных объектов по данным гиперспектрального зондирования, т. к. данные близких каналов гиперспектрального зондирования могут быть в значительной степени линейно зависимы. Следствие взаимной зависимости каналов — неустойчивость решаемых систем алгебраических уравнений, относящихся к разным каналам и обучающим пикселям, которые в совокупности характеризуют выбранные классы объектов в процессе распознавания всего множества объектов на обрабатываемом изображении. Необходимость решения таких систем уравнений связана с требованием обучения классификатора (вычислительной процедуры) по тестовой выборке для известных априори классов с последующей экстраполяцией найденных при обучении зависимостей на все пиксели изображения. Не существует стандартных подходов к осреднению данных гиперспектрального зондирования в отдельных каналах с целью уменьшения размерности признакового пространства, которое определяется числом измерительных каналов.

Каждый спектр гиперспектрального зондирования отображается в виде вектора в многомерном признаковом пространстве с размерностью, равной числу каналов, а множество текущих спектров отображается в виде отдельных точек кластеризации в этом пространстве. Приложения методов аэрокосмической гиперспектрометрии основаны на использовании следующих вычислительных процедур: оконтуривание заданных классов объектов на изображении; отбор обучающих спектров внутри контуров; понижение размерности признакового пространства; распознавание объектов на основе результатов обучения. Новизна предлагаемого подхода характеризуется следующими позициями: из всей совокупности спектральных каналов производится RGB отображение выбранной территории для оконтуривания на ней «объектов интереса»; определяются средние спектры и их мера изменчивости в пределах выделенных контуров; изучаются возможности разных классификаторов для повышения точности распознавания выделенных объектов по их спектрам, используемым для обучения.

Примеры выделения спектров и распознавания объектов

Ниже приводятся некоторые примеры летной кампании на дату съемки 31 июля 2010 г. в период интенсивных лесных и торфяных пожаров над исследуемой территорией.

На рис. 1 демонстрируются примеры исходных (рис. 1, *а*) и нормированных на интегральную яркость (рис. 1, *б*) средних спектров для выделенных контуров водной поверхности при отсутствии задымления над ней и при ее слабом и среднем задымлении, соответственно. Единицы измерений исходных спектров — $\text{Вт}/(\text{см}^2 \cdot \text{мкм} \cdot \text{стер})$, нормированных спектров — нм^{-1}. На рис. 1 демонстрируются также стандартные отклонения этих спектров (более тонкие линии) для каждого из трех иллюстрируемых объектов.

Данные рис. 1 относятся к разным водоемам на тестовой территории и разной интенсивности задымления от источников лесных и торфяных пожаров. Необходимость нормирования спектров связана с требованиями создания универсальной базы данных спектральных образов соответствующих объектов независимо от условий их съемки. Для сравнения показаны исходные спектры и их стандартные отклонения, относящиеся к контурам разных типов почвенно-растительного покрова (рис. 1, *в*), а также нормированные спектры и их стандартные отклонения (рис. 1, *д*) для этих же объектов (зеленая и пожелтевшая растительность, песчаная почва).

Из данных рис. 1, *а* можно видеть, что спектры всех трех типов объектов (водоемы при присутствии над ними дыма от пожаров, водоемы при наличии над ними слабого и среднего задымления) имеют характерный максимум в области 460–480 нм, амплитуда которого уменьшается по мере увеличения длины волны. Для всех трех типов иллюстрируемых кривых заметно влияние линий и полос поглощения излучения водной поверхностью и атмосферой, которое придает этим спектрам характерную изрезанность. Нормализация спектров приводит к уменьшению влияния задымления атмосферы (см. рис. 1, *б*) при сохранении общих особенностей спектров, как бы приближая спектры всех трех объектов к спектру водной поверхности при отсутствии задымления над ней.

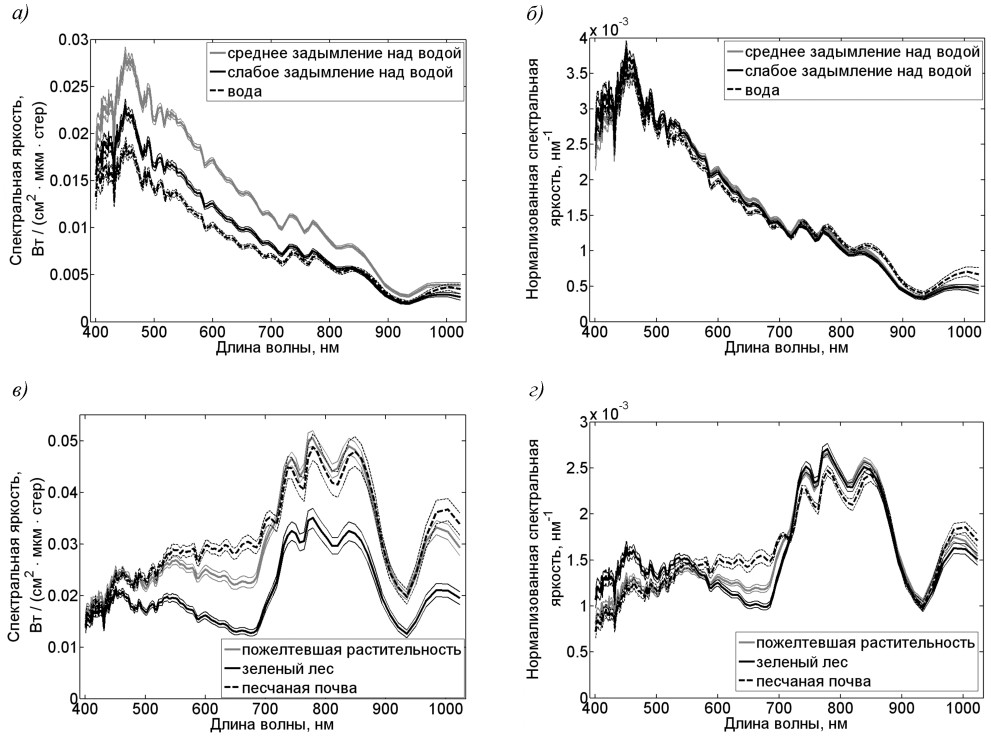


Рис. 1. Средние спектры объектов, полученные на основе обучающих ансамблей

Иной характер носят кривые на рис. 1, *в*, характеризующие спектры вегетирующей лесной растительности, пожелтевшей травянисто-кустарниковой растительности и песчаной почвы, соответственно. Для иллюстрируемого примера зеленой растительности характерно наличие двух небольших максимумов (длины волн 460 и 550 нм), минимума в области 670–680 нм, соответствующего максимуму полосы поглощения хлорофилла (основного пигмента фотосинтезирующей растительности), а также основного максимума для длин волн более 720 нм. Спектры, соответствующие песчаной почве, имеют вид монотонного возрастания их амплитуды при переходе от длин волн 400 нм к 720 нм при наличии упомянутой изрезанности вследствие наличия линий и полос поглощения излучения атмосферной средой. Можно заметить влияние полос поглощения кислорода (718 и 762 нм), водяного пара (820 и 940 нм) и другие особенности спектров в длинноволновой области. Спектры пожелтевшей растительности близки к спектрам открытой почвы во всей области длин волн, за исключением 550–710 нм.

При нормализации спектров (рис. 1, *г*) их амплитуда для зеленой растительности в области коротких волн (400–540 нм) оказывается выше, чем для двух остальных объектов. При этом все три объекта особенно существенно различаются в области длин волн 580–700 нм, слабо различаясь для более длинных волн.

Для оценки степени задымления объектов на рассматриваемой сцене использовались измерения яркости в каналах фиолетовой области. Выбор данного диапазона обусловлен характерной формой спектра дыма, ему соответствует максимум отражательной способности. В то же самое время для растительных объектов фиолетовой области характерно достаточно сильное поглощение. Открытые водные поверхности имеют схожий по форме спектр также с максимумом в фиолетовой области. Тем не менее, в данном диапазоне интенсивность излучения, отраженного от воды, существенно меньше, чем от дыма. Более сильное искажающее влияние могут оказать некоторые виды открытых почвогрунтов. Очень существенным фактором может оказаться наличие антропогенных объектов либо снежного покрова, однако данные объекты отсутствуют на рассматриваемой сцене. Таким образом, если отсутствует дополнительная информация о задымлении, полученная по данным наземных измерений, то изучение градаций яркости в этих каналах является единственным способом получения оценки прозрачности атмосферы в видимом в ближнем инфракрасном диапазоне.

Поскольку в исходных гиперспектральных данных каналы фиолетовой области сильно зашумлены, то для уменьшения данного эффекта использовался «объединенный» канал с центральной длиной волны 414 нм и разрешением 30 нм. По этим

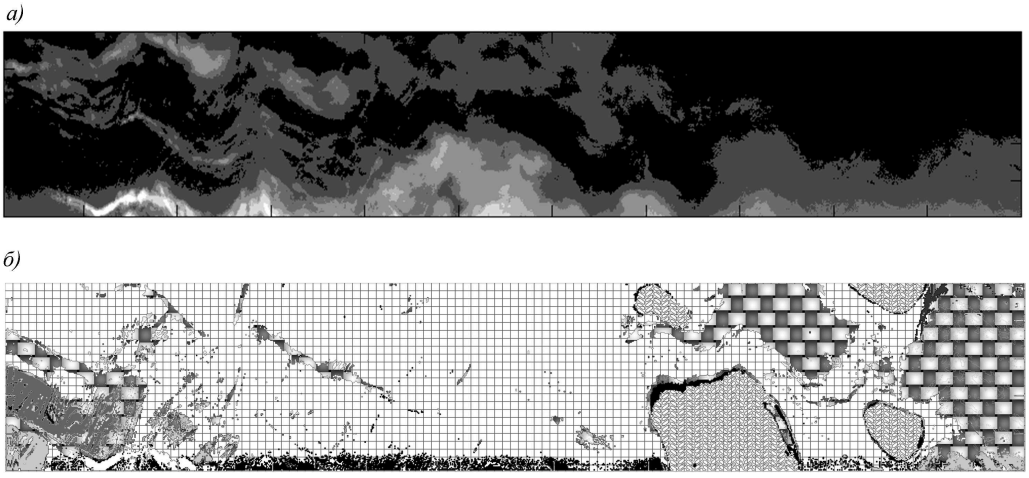


Рис. 2. Результаты тематической обработки: (*а*) — градации задымления; (*б*) — распознавание объектов

данным строилась гистограмма распределения и были выделены наиболее заметные моды, которые на качественном уровне соответствуют градациям плотности дымового шлейфа.

На рис. 2, *а* приведено изображение распределения интенсивности задымления для рассматриваемой сцены. Пороговые значения для 7 выделенных градаций составляют 0,019, 0,023, 0,025, 0,029, 0,031, 0,035 и 0,045. Контуры озер, торфоразработок и лесных массивов не прослеживаются. Однако в центральной части рисунка можно видеть линии, которые, скорее всего, соответствуют наземным объектам. Тем не менее, таких линий немного и в целом метод определения степени задымления по объединенным каналам фиолетовой области работоспособен. Данные, представленные на рис. 1, и аналогичные данные для других оконтуренных объектов использовались для обучения классификатора, основанного на применении квадратичного дискриминантного анализа. На рис. 2, *б* приведен пример распознавания объектов на выбранной сцене гиперспектрального аэрозондирования. В целом результаты распознавания соответствуют наземной информации. Наибольшую площадь занимают выгоревшие участки лесов и торфоразработок, а также леса, поврежденные пожарами (сетчатая штриховка). Достаточно точно выделяются озера (паркетная штриховка), вне зависимости от степени задымления. Белым цветом обозначены шлейфы дыма, которые имеют плотность, исключающую возможность распознавания объектов в данном диапазоне. В правой части изображения обозначены лесные массивы (кирпичная штриховка), которые в действительности подразделялись на здоровые и пораженные засушливыми погодными условиями. В левой части изображения выделены заболоченные участки (темно-серый цвет). Свело-серым цветом обозначены открытые участки торфоразработок, не подвергшиеся воздействию пожаров. Черным цветом обозначены открытые песчаные почвогрунты и неклассифицированные пиксели.

Выводы

Рассмотрены прикладные аспекты тематической обработки данных гиперспектрального аэрозондирования, полученных в условиях пониженной прозрачности атмосферы. Рассмотрены характерные особенности спектров наземных объектов при различной степени задымления. Представленные примеры определения градаций прозрачности атмосферы и распознавания объектов показывают качественное соответствие наземным данным.

Литература

[1] *Дмитриев Е. В., Козодеров В. В., Кондранин Т. В.* Распознавание объектов для территорий, охваченных лесными пожарами, по данным авиационной гиперспектрометрии // Труды МФТИ «Аэрокосмические исследования, прикладная механика», 2010. Т. 2. № 3. С. 133–140.

[2] *Козодеров В. В., Кондранин Т. В.* Методы оценки состояния почвенно-растительного покрова по данным оптических систем дистанционного аэрокосмического зондирования. — М.: Изд-во МФТИ, 2008. — 222 с.

[3] *Козодеров В. В., Кондранин Т. В., Райкунов Г. Г., Казанцев О. Ю., Белоцерковский А. В., Асташкин А. А., Бобылев В. И., Дмитриев Е. В., Каменцев В. П., Борзяк В. В., Щербаков М. В., Лесуновский А. А.* Аэрокосмическая гиперспектрометрия: летные испытания аппаратуры, программно-алгоритмическое обеспечение обработки данных // Исследование Земли из космоса, 2010. № 5. С. 59–68.

Вейвлет-метод выделения геомагнитных возмущений и анализа магнитных данных*

Мандрикова О. В., Соловьев И. С.
oksanam1@mail.kamchatka.ru
г. Петропавловск-Камчатский, Камчатский государственный технический университет,
Институт космофизических исследований и распространения радиоволн ДВО РАН

Работа посвящена созданию методов по анализу геомагнитных данных, выделению и идентификации локальных особенностей, обусловленных солнечной активностью. Предложены алгоритмы, основанные на вейвлет-преобразовании, позволяющие в автоматическом режиме выделить геомагнитные возмущения, выполнить их анализ и идентификацию. Апробация методов выполнена на данных магнитного поля Земли, полученных на обсерватории «Паратунка» (с. Паратунка, Камчатский край).

Из-за сложной внутренней структуры, сильной изменчивости, нерегулярности регистрируемых геомагнитных данных задача их обработки и анализа является весьма сложной. Традиционные методы анализа временны́х рядов, основанные на процедурах сглаживания, позволяют изучить низкочастотные вариации параметров магнитного поля, но не дают информации о локальных изменениях, протекающих в физическом процессе, и их масштабных характеристиках [1, 2]. При обработке данных наблюдается потеря важной информации.

В данной работе для представления геомагнитных данных используются аппроксимирующие вейвлет-схемы [3, 4]. Эта математическая платформа имеет обширный словарь базисов различной формы и длительности, что позволяет исследовать тонкие особенности структуры магнитных данных [4, 5]. Эквивалентность теории непрерывных временны́х вейвлетов и дискретных наборов фильтров дает важную информацию о локальных особенностях исследуемой функции. Для отображения вариаций магнитного поля в работе используются непрерывные вейвлет-преобразования и вейвлет-пакеты [3]. Непрерывные преобразования дают возможность исследовать тонкие особенности протекания процесса накануне и в периоды бурь. На основе дискретных разложений в вейвлет-пространстве геомагнитные данные представляются в виде комбинаций составляющих двух видов — детализирующих и аппроксимирующих [2]. Аппроксимирующие составляющие описывают регулярные изменения поля, детализирующие составляющие определяют различные типы частотно-временны́х структур, формирующих регистрируемый процесс в локальные моменты времени, и характеризуют возмущенность магнитного поля. Процедура анализа геомагнитных данных может быть представлена в виде следующих взаимозависимых этапов:

1. Выделение информативных частотно-временны́х интервалов в данных.
2. Идентификация возмущенных составляющих.
3. Анализ локальных структур, формирующих процесс в периоды повышенной геомагнитной активности и оценка изменений характеристик поля.

Работа поддержана грантом Президента Российской Федерации, МД-2199.2011.9.

Описание метода

Вейвлет-преобразование раскладывает сигналы по растянутым и сдвинутым вейвлетам Ψ. Так как вейвлет Ψ имеет нулевое среднее значение, то вейвлет-преобразование [3]

$$Wf(s,u) = \int f(t)\frac{1}{\sqrt{s}}\Psi\left(\frac{t-u}{s}\right)dt$$

измеряет изменение f в окрестности точки u, размер которой пропорционален s. При стремлении масштаба s к нулю вейвлет-коэффициенты характеризуют свойства функции f в окрестности u. В работе [3] показано, что если f ограничена и $Wf(s,u)$ удовлетворяет для нецелого числа α следующему условию:

$$\forall(s,u) \in R^+ \times [a,b] \quad \mid Wf(s,u) \mid \leqslant As^{\alpha+1/2}, \quad (1)$$

где R^+ — множество действительных чисел, больших нуля, A — некоторая константа, то f удовлетворяет равномерному условию Липшица α на $[a+\varepsilon, b-\varepsilon]$ при любом $\varepsilon > 0$.

Неравенство (1) является условием асимптотического убывания $|Wf(s,u)|$ при $s \to 0$. Будем считать, что функция f в окрестности точки v имеет локальную особенность, если $|Wf(s,u)|$ не удовлетворяет условию (1) в окрестности точки v. В этом случае операция выделения локальных особенностей функции может быть реализована на основе проверки условия

$$\mid Wf(s,u) \mid > T, \quad (2)$$

где T — пороговое значение.

Из существования в вейвлет-преобразовании аналога равенства Парсеваля следует, что через значения вейвлет-коэффициентов может быть определена энергия функции:

$$E_f = \int f^2(t)\,dt = C_\Psi^{-1}\iint W^2(s,u)\frac{dsdu}{s^2}. \quad (3)$$

Плотность энергии функции $E_{f_\Psi} = |Wf(s,u)|$. Мерой магнитной возмущенности, согласно методике Бартельса [6], является амплитуда возмущения. Эта амплитуда, определяемая на трехчасовом временно́м интервале, равна разности между наибольшим и наименьшим отклонениями реальной магнитограммы от невозмущенной вариации поля. Тогда из соотношений (1)–(3) получаем, что характеристикой интенсивности возмущений в вейвлет-пространстве будет являться амплитуда вейвлет-коэффициентов. Интенсивность возмущений в момент времени $t = u$ на анализируемом масштабе s определим как $E_{f(s,u)} = |Wf(s,u)|$.

Это позволяет исследовать изменения энергетических характеристик исследуемого процесса во времени, оценить вклад различных масштабов. Из соотношения (2) получаем, что выделение частотно-временны́х интервалов, содержащих возмущения, может быть реализовано на основе применения пороговой функции

$$P_{T_s}(x) = \begin{cases} x, & \text{если } \mid x \mid \geqslant T_s\,; \\ 0, & \text{если } \mid x \mid < T_s\,, \end{cases} \qquad (4)$$

где T_s — порог на масштабе s. Реализовав операцию (4) и введя в рассмотрение временное окно

$$E_{f(s,\Delta\mu_0)} = \int\limits_{u=u_0-\varepsilon}^{u=u_0+\varepsilon} E_{f(s,u)}du\,,$$

мы можем проанализировать временную динамику интенсивности возмущений поля по масштабам.

Оценку распределения выделенных возмущений поля по масштабам даст величина $E_{f_s} = \int E_{f(s,u)}du$. Максимумы функции E_{f_u} по аналогии с характерными модами Фурье [7] позволят выделить масштабы процесса, вносящие основной вклад в полную энергию E_f. Масштабный уровень u, на котором наблюдается максимум E_{f_u}, в данном случае характеризует среднюю продолжительность локального возмущения, вносящего основной вклад в энергию анализируемого процесса на данном интервале времени на фоне развивающейся бури.

Изменения интенсивности возмущений поля во времени можно проанализировать на основе величины

$$E_{f_u} = \int E_{f(s,u)}ds\,. \qquad (5)$$

Введя в рассмотрение временное окно

$$E_{f(s,\Delta\mu_0)} = \int\limits_{u=u_0-\varepsilon}^{u=u_0+\varepsilon} E_{f_u}du\,, \qquad (6)$$

мы можем проанализировать эти изменения в различных временных диапазонах.

Результаты обработки и анализа данных магнитного поля Земли

Для выделения периодов повышенной геомагнитной активности было выполнено непрерывное вейвлет-разложение и применена операция (4), что позволило выделить в магнитных данных частотно-временны́е интервалы, содержащие возмущения. Идентификация параметров T_s пороговой функции была выполнена статистически по данным K-индекса, характеризующего магнитную возмущенность в месте регистрации (станция «Паратунка»). Путем обработки «спокойных» дней (суммарные суточные значения K-индекса для которых не превышают значения 8) для каждого масштабного уровня s были идентифицированы параметры $T_{s,1}$, позволяющие выделить частотно-временны́е интервалы, содержащие как слабые, так и сильные геомагнитные возмущения. На основе обработки «спокойных» и «слабовозмущенных» дней (суммарные суточные значения K-индекса для которых не превышают значения 18) для каждого масштабного уровня s были идентифицированы параметры $T_{s,2}$, позволяющие выделить частотно-временны́е интервалы, содержащие сильные геомагнитные возмущения. Полученные картины распределения возмущений геомагнитного поля накануне и в периоды сильных магнитных бурь в пространстве (a,b) = (масштаб, время) показаны на рис. 1, *б*, *в* и 2, *б*, *в*. Над магнитограммами рис. 1, *а* и 2, *а* приведены трехчасовые значения K-индекса. Для анализа изменений интенсивности возмущений поля во времени в периоды бурь также были выполнены операции (5) и (6). Операция (6) была реализована в скользящем временно́м окне, равном трем часам. Полученные результаты представлены на рис. 1, *г*, *д* и 2, *г*, *д*. Анализ полученных результатов показывает, что во время развития бури интенсивность возмущений значительно увеличивается, в вариациях магнитного поля возникают локальные разномасштабные периодичности. Анализ данных за разные периоды времени подтверждает нестационарность этого процесса и их неравномерное распределение и по времени, и по масштабам. Накануне бурь и после их протекания в данных наблюдаются слабые возмущения, интенсивность которых не превышает значений $T_{s,1}$ (см. рис. 1, *б* и 2, *б*). За несколько часов до бури и во время ее протекания интенсивность возмущений значительно увеличивается и достигает значений $T_{s,2}$ (см. рис. 1, *в* и 2, *в*), что позволяет в автоматическом режиме фиксировать момент предстоящей бури.

Выводы

Рассчитанные из вейвлет-анализа интенсивности геомагнитных возмущений позволили выполнить оценку энергетических параметров магнитно-

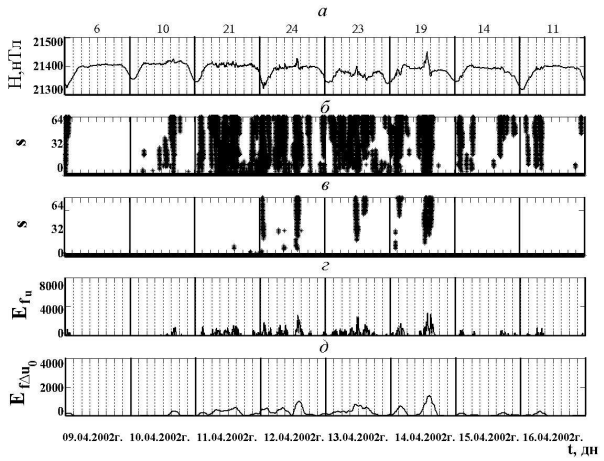


Рис. 1. Результат обработки геомагнитных данных за период 09.04.2002–16.04.2002 (станция регистрации «Паратунка», Камчатский край): (*а*) данные регистрации; (*б*) результат обработки данных на основе алгоритма 2 (пороговое значение $T_{s,1}$); (*в*) результат обработки данных на основе алгоритма 2 (пороговое значение $T_{s,2}$); (*г*) результат выполнения операции (5); (*д*) результат выполнения операции (6) в скользящем временно́м окне

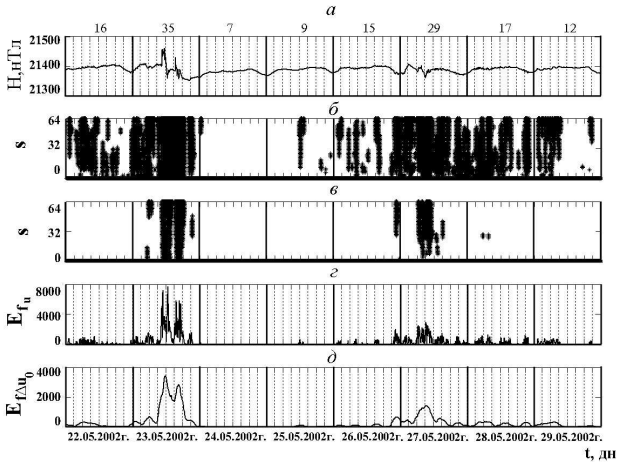


Рис. 2. Результат обработки геомагнитных данных за период 22.05.2002–29.05.2002 (станция регистрации «Паратунка», Камчатский край): (*а*) данные регистрации; (*б*) результат обработки данных на основе алгоритма 2 (пороговое значение $T_{s,1}$); (*в*) результат обработки данных на основе алгоритма 2 (пороговое значение $T_{s,2}$); (*г*) результат выполнения операции (5); (*д*) результат выполнения операции (6) в скользящем временно́м окне

го поля на анализируемом временно́м интервале и проследить динамику их изменения в периоды бурь.

Из результатов непрерывного вейвлет-преобразования магнитных данных выделены частотно-временны́е интервалы, содержащие возмущения, показавшие, что в периоды бурь в вариациях магнитного поля возникают локальные разномасштабные периодичности, имеющие неравномерное распределение и по времени, и по масштабам. За несколько часов до бури и во время ее протекания интенсивность выделенных возмущений значительно увеличивается, что позволяет фиксировать момент предстоящей бури.

Литература

[1] *Будько Н., Зайцев А., Карпачев А., Козлов А., Филиппов Б.* Космическая среда вокруг нас. — Троицк: ТРОВАНТ, 2006. — 232 с.

[2] *Мандрикова О. В., Соловьев И. С.* Вейвлет-технология обработки и анализа вариаций магнитного поля Земли // Интеллектуализация обработки информации: 8-я международная конференция, М.: МАКС Пресс, 2010. — С. 430–433.

[3] *Mallat S.* A Wavelet tour of signal processing / Пер. с английского. — М.: Мир, 2005. — 672 с.

[4] *Мандрикова О. В., Соловьев И. С.* Вейвлет-технология обработки и анализа вариаций магнитного поля Земли // Информационные технологии, 2011. № 1. С. 34–38.

[5] *Козлов В. И., Марков В. В.* Вейвлет-образ гелиосферной бури в космических лучах // Геомагнетизм и аэрономия, 2007. Т. 47. № 1. С. 56–65.

[6] *Bartels J.* Potsdamer erdmagnetische Kennziffern, 1 Mitteilung // Zeitschrift fur Geophysik, 1938. Vol. 14. P. 68–78.

[7] *Ротанова Н. М., Бондарь Т. Н., Иванов В. В.* Вейвлет-анализ вековых геомагнитных вариаций // Геомагнетизм и аэрономия, 2004. Т. 44. № 2. С. 276–282.

Система иерархического распознавания акустических изображений подводных объектов на основе техники SVD*

Макшанов А. В., Гальяно Ф. Р.
makshanov@oogis.ru, galiano@oogis.ru
Санкт-Петербург, СПИИРАН

Описана разработанная система анализа данных, получаемых с помощью гидролокатора бокового обзора (ГБО), установленного на автономном необитаемом подводном аппарате. Для этапов сегментации и описания используется система распознавания изображений посредством представлений в различном числе градаций. Улучшение результатов достигается путем классификации выделенных сегментов с помощью алгоритма, основанного на сингулярном разложении матрицы обучающих векторов.

Гидролокаторы бокового обзора — одно из наиболее известных и эффективных средств для подводных исследований, позволяющее быстро исследовать большие площади дна независимо от прозрачности воды. Они работают одинаково хорошо как в пресной, так и в соленой воде, могут использоваться на озерах, реках, заливах и в открытом океане. Особенность современных гидролокаторов [1–3] состоит в использовании наряду с тональным зондирующим сигналом сложного сигнала с линейной частотной модуляцией (ЛЧМ), что предполагает использование цифровой техники их синтеза и обработки. При использовании частот порядка 500 кГц и максимальной наклонной дальности до 60 м такие локаторы обеспечивают получение изображений с разрешением до 1 см с пиковой мощностью всего в сотни ватт. Все другие виды колебаний имеют существенно большее затухание в воде и позволяют производить исследования на расстояниях только в единицы метров. Акустический сигнал хорошо отражается от границ вода–воздух, вода–камень, вода–металл и плохо отражается, например, от границы вода–ил, что обеспечивает варьирование яркости акустического изображения. Острый «угол зрения» ГБО создает условия формирования акустической тени, образуемой возвышающимися над дном объектами, что помогает их выделять и интерпретировать. Такие особенности гидроакустических изображений в условиях их формирования в реальном масштабе времени требуют создания специальной технологии их распознавания и классификации.

Особенности входных данных

Значительный объем анализируемых данных и ограничения на доступные вычислительные ресурсы осложняют использование высокоэффективных алгоритмов анализа и требуют разработки специального математического и алгоритмического обеспечения, позволяющего обеспечить удовлетворительное качество анализа при соблюдении налагаемых ограничений. Первым этапом анализа данных, поступивших с борта автономного необитаемого подводного аппарата, является построение единого изображения, хранящего информацию, полученную ГБО за время миссии. Форма полученного изображения, как правило, далека от прямоугольной, требование связности также не всегда выполняется; таким образом, встает задача обобщения многих алгоритмов обработки изображений (поворота, масштабирования, сегментации и др.) на случай изображения, представляющего собой множество, в общем случае, несвязных пикселей с произвольными координатами. Во многих случаях оптимальным способом представления подобных изображений является ассоциативный массив.

Система выделения объектов заданного класса

Для решения задач классификации акустических изображений рассматривается вариант техники ABO [4], в котором мера близости сегментов изображения при их динамическом укрупнении основана на поэтапном применении процедуры SVD (Singular Value Decomposition) к матрицам, описывающим сегменты более низкого уровня.

Формирование пространства признаков. Процесс анализа данных ГБО, с целью выделения объектов заданного класса, состоит из трех этапов: сегментации, описания и классификации. На первых двух этапах сокращение требуемых вычислительных ресурсов достигается с помощью представления изображения в заданном числе градаций и использования аппарата признаковых представлений [6]. Полученное в результате маркированное изображение [7] и совокупность значений свойств, вычисленных для каждого сегмента, далее используются для классификации. Единообразное представление набора свойств сегмента в виде вектора признаков используется для их классификации, в частности на основе разновидности метода ближайшего соседа. Векторы состоят из признаков трех типов: яркостные (средняя яркость, дисперсия, инварианты Ху [5]), геометрические (площадь сегмента, периметр, топологические свойства, компактность [9]) и контекстные (свойства сегментов, смежных с текущим). Редукция размерности осно-

Работа выполнена при финансовой поддержке РФФИ, проект № 11-07-00685-А.

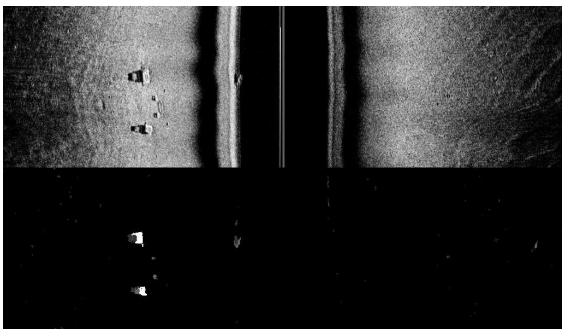


Рис. 1. Пример выделения объектов на основе анализа данных ГБО. В верхней части изображения приведены исходные данные, в нижней — результат анализа. Яркость пропорциональна степени близости найденных сегментов и элементов обучающей выборки

вана на использовании сингулярного разложения матриц (SVD).

Алгоритм классикации на основе SVD. Из m векторов признаков по n элементов формируется матрица A, состоящая из m строк и n столбцов. Известно, что любую матрицу, используя SVD, можно представить в виде:

$$A = \sum_{v=1}^{h} s_v L_v R_v^T ,$$

где s_v — сингулярные числа, L_v — левые сингулярные вектора, R_v — правые сингулярные вектора, а $h = \mathrm{rank}(A)$. При этом слагаемые ряда упорядочены по убыванию сингулярных чисел и учтены дополнительные ограничения на элементы ряда. Чем длиннее ряд, тем выше точность разложения.

Сингулярное разложение матриц позволяет уменьшить размерности входных данных и упорядочить полученные редуцированные признаки по их вкладу в дисперсию. Нормирование данных с заданными коэффициентами позволяет учесть влияние вклада отдельного признака в результат классификации (его вес), а введение весовых коэффициентов при идентификации ближайшего соседа в новом признаковом пространстве позволяет изменять при классификации соотношение ошибок первого и второго рода [8]. Повышение устойчивости классификации достигается путем усреднения результатов по локальной окрестности заданного размера.

Процесс распознавания любого кадра может быть прерван на одном из промежуточных этапов. Процедура в полном объеме выполняется только в тех случаях, когда распознавание на ранних этапах не достигает цели. Такие кадры отмечаются для более детального анализа человеком-оператором. В результате покадрового анализа больших полотен акустических изображений каждому кадру присваивается метка, например «протяженный металлический объект с развитой акустической тенью», что позволяет оператору выделять только кадры с заданными характеристиками. Пример работы приведен на рис. 1. В данном случае процесс анализа был остановлен на этапе сегментации и выделения признаков. Установлено, что наилучшие результаты классификации достигаются при использовании в качестве обучающей выборки векторов свойств сегментов, составляющих акустическую тень, которые выделены на полученном изображении белым цветом.

Выводы

Разработанная система позволяет выполнять выделение объектов заданного класса на основе данных ГБО. К недостаткам следует отнести сложность учета структуры искомых объектов. Сформированный набор алгоритмов позволяет учитывать при классификации широкий набор признаков и обеспечивать приемлемое качество классификации при соблюдении ограничений на доступные вычислительные ресурсы.

Литература

[1] *Fish J. P., Carr H. A.* Sound Underwater Images: A guide to the generation and interpretation of side scan sonar data. — Orleans: Lower Cape Publishing, 2009. — 190 p.

[2] Product Survey on Side-Scan Sonar. — Hydro International, 2004. Vol. 8. No. 3. P. 36–39.

[3] *McFadzean A., Ceri. R.* An automated side scan sonar pipeline inspection system.— UnderWater Magazine, 2000. Vol. 8. No. 6.

[4] *Журавлев Ю. И., Зенкин А. А., Зенкин А. И. и др.* Задачи распознавания и классификации со стандартной обучающей информацией // Журнал вычислительной математики и математической физики, 1980. Т. 20. № 5. С. 1294–1309.

[5] *Wood J.* Invariant pattern recognition: A review. Pattern recognition, 1996. Vol. 29. No 1. P. 1–17.

[6] *Харинов М. В., Гальяно Ф. Р.* Распознавание изображений посредством представлений в различном числе градаций // Математические методы распознавания образов (ММРО–14) / Сб. докл. 14–й Всерос. конф. Владимирская обл., г. Суздаль. — М.: МАКС Пресс. — С. 465-–468. (ISBN 978–5–317–02947–0).

[7] *Шапиро Л., Стокман Дж.* Компьютерное зрение. — М.: Бином, 2006. — 752 с.

[8] *Гальяно Ф. Р.* Алгоритм классификации участков поверхности Земли на основе сингулярного разложения матриц // Информационные технологии, 2010. №12. С. 35–37.

[9] *Гонсалес Р., Вудс Р.* Цифровая обработка изображений. — М.: Техносфера, 2006. — 1072 с.

Методы выявления пространственного группирования землетрясений в сейсмогеодинамическом исследовании районов Центральной Азии*

Дядьков П. Г., Михеева А. В.
anna@omzg.sscc.ru
Новосибирск, Институт нефтегазовой геологии и геофизики им. А. А. Трофимука Сибирского отделения РАН

В работе описаны методы выявления связанных событий, реализованные в геоинформационной системе GIS-EEDB. Первая группа методов связана с распознаванием линейных структур по распределению событий на площади с целью обнаружения приуроченности сейсмичности к зонам активных разломов (например, границ плит или блоков). Вторая группа методов имеет целью выявление афтершоков и роев, а после их исключения — кластеров взаимосвязанных землетрясений, плотность распределения которых по времени не соответствует закону распределения Пуассона для случайных величин.

Изучение закономерностей пространственно-временного развития сейсмического процесса в зоне влияния Индо-Евразийской коллизии в контексте приуроченности к основным тектоническим структурным элементам требует разработки методов кластеризации событий в пространстве и в пространстве–времени. Этими методами был дополнен комплекс геоинформационно-экспертных методов, реализованный в вычислительной интерактивной системе GIS-EEDB (Geo Information System Expert Earthquake Data Base), которая была разработана в ИНГГ СО РАН для проведения работ в области исследования сейсмогеодинамического режима.

Описание прикладной системы анализа данных GIS-EEDB

Логическая структура геоинформационной системы GIS-EEDB представляет собой совокупность взаимодействующих между собой программных блоков: сейсмологической базы данных, географической подсистемы и подсистемы анализа данных. Поскольку развитие процессов в локальной области или в регионе должно рассматриваться с учетом процессов, происходящих на региональном, мегарегиональном и, в ряде случаев, на глобальном уровне, в сейсмологическую базу данных системы GIS-EEDB были включены как глобальные, так и региональные каталоги землетрясений. В частности, для пространственно-временного исследования районов Центральной Азии используются региональные каталоги Алтае-Саянской области, Байкальского региона, Республики Казахстан и Китая, которые постоянно пополняются новым материалом.

Исследование, проводимое в системе GIS-EEDB на различных масштабных уровнях: плит, микроплит, блоков или вблизи их границ со складчатыми областями, — представляет собой пространственно-временной анализ ряда параметров сейсмического режима с использованием алгоритмов математической статистики и анализа временны́х рядов. При этом обычно изучаются следующие параметры сейсмического режима: сейсмические затишья, наклон графика повторяемости, форшоковая активизация, кластеризация событий, плотность сейсмогенных разрывов и т. п. Для построения графиков и карт распределения параметров сейсмического режима используются характеристики очагов землетрясений выбранного каталога: сейсмическая энергия, магнитуда, геометрические размеры очага, момент возникновения землетрясения, координаты эпицентра, глубина очага. При этом особое внимание обращается на представительность и достоверность выбранных каталогов в рассматриваемой области.

Выявление блоковых структур

Для выявления границ блоков и других линейных структур по данным сейсмичности в системе GIS-EEDB реализован простой алгоритм распознавания линейных образов по множеству точек, распределенных в пространстве. В основе данного метода лежит задание максимального шага и максимального угла отклонения для поиска следующей точки (эпицентра события). Угол отклонения рассматривается по отношению к направлению, создаваемому предыдущей парой соединяемых в пространстве точек. Использование этого алгоритма помогает выявлять как прямолинейные, так и искривленные (например, дугообразные) структуры (рис. 1).

Методы идентификации афтершоков

Большую роль при проведении исследований, связанных с вероятностными оценками выборок событий, играет предварительная обработка исходных наборов данных, в частности очищение выбранной части каталога от афтершоков. Эта задача предполагает создание алгоритмов выявления связанных событий, обладающих особыми свойствами пространственно-временного распределения. В нашей системе реализовано три алгоритма этой операции по выбору пользователя. Первый метод,

Работа выполнена при финансовой поддержке РФФИ, проект № 10-05-01042-а.

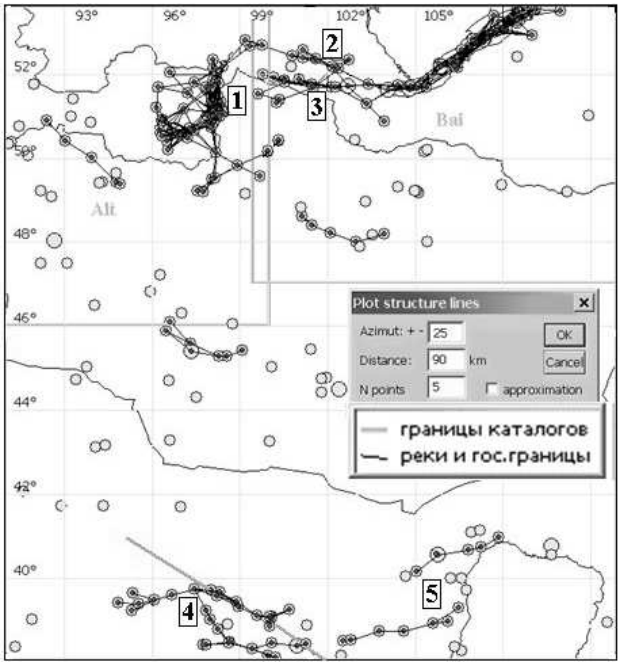


Рис. 1. Пример работы алгоритма распознавания линейных структур в районе юга Сибири, Монголии и северного Тибета для событий с $M \geqslant 4$ (1990–2010): 1 — Бусингольская впадина; 2 — Главный Саянский разлом; 3 — Тункинский разлом; 4 — Килианская складчато-надвиговая зона; 5 — Кайдамский блок [1]. Использован каталог Китая (CSN), Байкала (БФ ГС СО РАН) и Алтая (Алтае-Саянской экспедиции ГС СО РАН)

условно названный статистическим, основан на параметрах разности времён и расстояний афтершокового события и главного толчка, полученных нами из статистики накопленных данных об афтершоковых процессах и зависящих от магнитуды M_s (или энергетического класса K) главного толчка:

$$dT = (M_{s\,\text{главн}} - 4)162 \text{ дней}; \quad dS = 3L \text{ км},$$

где L — длина очага, определяемая по формуле $\lg L_j = aK_j + c$ [2], где $a = 0{,}244$, $c = -2{,}266$.

Но наиболее часто используемым в GIS-EEDB методом фильтрации афтершоков является второй метод, названный эллиптическим, основанный на подходе Прозорова [3] и включающий в себя следующие этапы:

1. Первый проход каталога с целью нахождения плотности неафтершоковых событий.
2. Второй проход с выделением предварительных афтершоков в прямоугольнике.
3. Построение по выделенной группе афтершоков эллипса рассеяния по среднеквадратичному отклонению от центра множества («классический» вариант) или методом наибольшей вероятности («измененный» вариант) по выбору.
4. Последующие проходы каталога с целью послойного выделения афтершоков в 4-кратной эллиптической метрике (при выборе «классического» варианта).

Время афтершокового процесса определяется как отношение числа афтершоков к суммарной плотности в прямоугольнике или эллипсе. «Классический» вариант нахождения параметров эллипса на этапе 3 соответствует предлагаемому в [3] методу расчета метрик по среднеквадратичному пространственному отклонению точек от арифметического центра множества. При этом в системе возможен вариант, учитывающий веса точек, которые определяются по числу событий в ячейке попадания афтершока (рис. 2). Учёт веса имеет смысл в случаях большого разброса афтершокового облака на площади. Возможен также «измененный» вариант метода, когда метрики определяются с помощью эллипса равной вероятности. Опишем его подробнее.

Расчёт эллипса в этом варианте производится в соответствии с представлением о нормальном распределении случайных величин x и y относительно центра множества, которое графически представляется с помощью эллипсов равной вероятности [4]:

$$\varphi(x,y) = \frac{1}{1-\rho_{12}^2}\Big(\frac{x^2}{\sigma_1^2} - 2\rho_{12}^2\frac{xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\Big) = \lambda^2\,,$$

где $\sigma_1^2 = DX$, $\sigma_2^2 = DY$ — дисперсии x и y, а ρ_{12} — коэффициент корреляции между x и y. В качестве наилучшей при рассмотрении вероятностей P, близких к 1, предлагается [4] аппроксимация квантилей для распределения с $m = 2$ степенями свободы, принимающая при $P = 0{,}9995$ следующий вид:

$$\lambda^2 = m\left(1 - \frac{2}{9m} + u_p\sqrt{\frac{2}{9m}}\right)^3 = \\ = 2\left(1 - \frac{1}{9} + 3{,}29\frac{1}{3}\right)^3.$$

Результат применения данной оценки для определения эллиптических метрик мы видим на рис. 2 (эллипс 3). Поиск афтершоков при этом осуществляется уже в единственной кратности (этап 4 отсутствует).

Практика показала, что в ряде случаев преимущество в выявлении афтершоков имеет метод их идентификации с помощью эллипса равной вероятности. Так, при выявлении афтершоков Южнобайкальского землетрясения метод равной вероятности превзошел «классический» как по числу выделеннных событий (861 и 711 соответственно — см. рис. 2), так и по длительности афтершоковой последовательности (5,1 и 1,1 года, соответственно — рис. 3). Преимуществом «измененного» метода является также практическая независимость

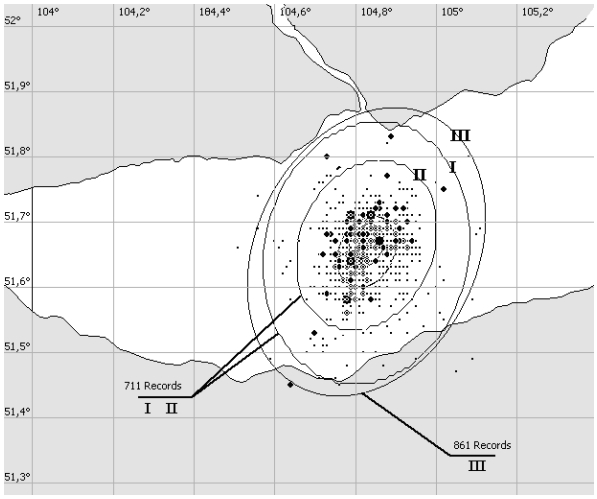


Рис. 2. Результаты трех вариантов расчёта эллиптических метрик в алгоритме выделения афтершоков на примере землетрясения 25.02.1999 ($M = 5{,}9$): 1 — по среднеквадратичному отклонению без веса; 2 — то же с весом; 3 — эллипс равной вероятности. Цифрами 1 и 2 отмечен внешний эллипс (4-кратное увеличение метрик расчетного эллипса согласно методу [3])

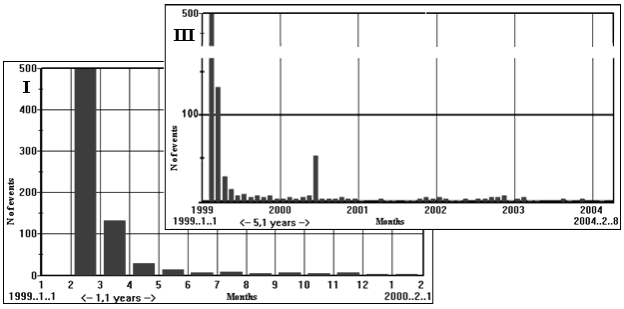


Рис. 3. Распределение по времени афтершоков, выявленных эллипсами 1 и 3; $R_{s/n} = 20$

его результатов от порогового соотношения сигнал/шум $R_{s/n}$ [3].

На примере Чуйского (27.9.2003) землетрясения оба метода выделения афтершоков («классический» и «измененный») показывают одинаковый результат как в определении интервала времени (до конца детальной части каталога, т. е. 4,2 года), так и по числу афтершоков (2009 событий). Это говорит о сближении качества «классического» и «измененного» методов в условиях статистической достаточности по числу событий.

После процедуры очищения от афтершоков в выборке остаётся значительное количество связанных событий, относящихся к роевым последовательностям. Технология очищения каталогов от роев аналогична удалению афтершоков за исключением условия о соотношении магнитуд главного и зависимых событий — в случае выделения роев зависимые события могут иметь как меньшую, так и большую магнитуду по сравнению с начальным

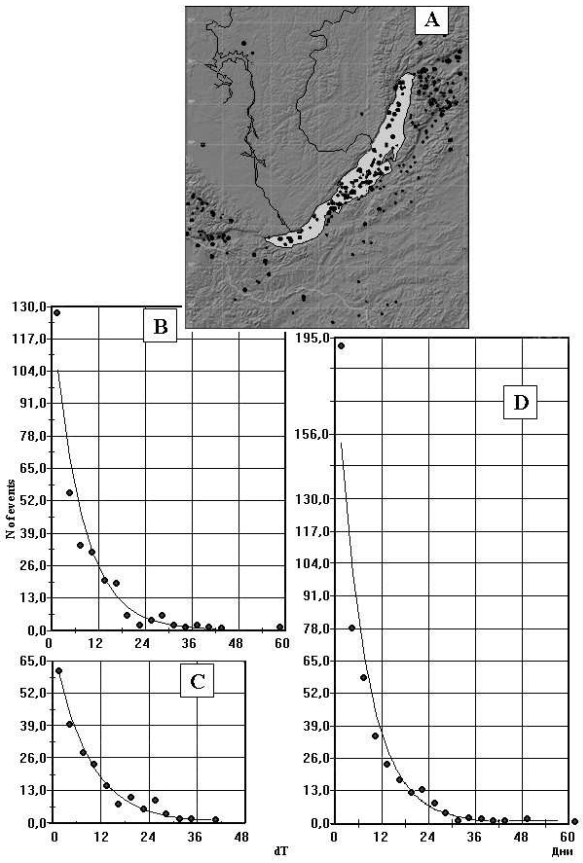


Рис. 4. Гистограммы зависимости числа пар соседних по времени событий с $M \geqslant 3$ (афтершоки и рои удалены) в Байкальской рифтовой зоне (A) от dT для периодов: (B) — с июля 1987 по июнь 1993; (C) — с июля 1993 по июнь 1998; (D) — с июля 1998 по июнь 2007. Шаг суммирования — 3 дня

событием процесса. Кроме того, при выявлении роевых последовательностей время процесса не рассчитывается (как в [3] — по количеству выделенных афтершоков), а задается пользователем интерактивно, поскольку временно́е распределение событий роя не обладает свойствами, характерными для афтершоковых последовательностей.

Алгоритмы выявления кластеров

После удаления афтершоков и роев кластеризация или группирование оставшихся землетрясений обусловлена существующей в природе локализацией сейсмичности в зонах активных разломов, например границ плит или блоков. Для поиска кластеров задаются условия на разность времени и расстояния в каждой паре событий (dT и dS), а также тип кластеризации (временно́й или пространственный). В системе GIS-EEDB заложено два метода нахождения кластеров:

1) метод задания пространственно-временных интервалов (dT и dS);

2) метод автоматического расчета dT и dS исходя из физических процессов разрушения среды [5, 6].

Важным моментом в первом методе является выбор параметров кластеризации (dT и dS). Значения задаваемых параметров можно определить по графикам зависимости числа пар соседних по времени событий от этих параметров, выявляя интервалы заметного превышения числа пар относительно графика экспоненциального пуассоновского распределения (для dT) или максимумов числа событий (для dS).

Превышение наблюденных значений относительно кривой распределения Пуассона на графиках (B) и (D) (рис. 4) в первый 3-дневный интервал может свидетельствовать о наличии эффекта взаимовлияния сейсмических событий [7] в эти периоды. Или, например, о таком аномальном состоянии среды при ее рассмотрении как динамической системы, при котором наблюдаются признаки коллективного поведения ее элементов.

Заключение

Предложены алгоритмы и подходы, позволяющие осуществлять группирование гипоцентров землетрясений в пространственном и пространственно-временном диапазоне. Выявление групп связанных событий необходимо как для построения детальных моделей земной коры: выделения сейсмоактивных границ блоков или отдельных разломов, — так и для изучения сейсмического режима территорий.

Литература

[1] *Лысак С. В.* Термальная эволюция, геодинамика и современная геотермальная активность литосферы Китая // Геология и геофизика, 2009, Т. 50, № 9. — С. 1058–1071.

[2] *Ризниченко Ю. В.* Проблемы сейсмологии. — М.: Наука, 1985. — 408 с.

[3] *Прозоров А. Г.* Динамический алгоритм выделения афтершоков для мирового каталога землетрясений. Математические методы в сейсмологии и геодинамике // Вычислительная сейсмология. — М.: Наука, 1986. — Вып. 19.

[4] *Корн Г., Корн Т.* Справочник по математике. — М., 1977. — 832 с.

[5] *Куксенко В. С.* Модель перехода от микро- к макроразрушению твердых тел // Физика прочности и пластичности. — Л.: Наука, 1986. — С. 38–41.

[6] *Соболев Г. А., Пономарев А. В.* Физика землетрясений и предвестники. — М.: Наука, 2003. — 270 с.

[7] *Ebel J. E., Kafka A. L.* A non-Poissonian element in the seismicity of the Northeastern United States // Bull. Seism. Soc. Amer. — 2002. — Vol. 92. No 5. — P. 2040–2046.

Разработка и реализация алгоритмов анализа подстилающей поверхности по мультиспектральным спутниковым снимкам среднего разрешения*

Потехин Е. Н., Харитонов А. В., Рахманов Х. Э., Леухин А. Н.
code@marstu.net
г. Йошкар-Ола, Марийский государственный технический университет

Статья посвящена методам, применяемым для распознавания объектов природных классов подстилающей поверхности Земли по мультиспектральным спутниковым снимкам среднего разрешения.

Введение

В последнее время возрастает актуальность автоматизации обработки космических снимков подстилающей поверхности ландшафтов поверхности Земли в геоинформационных системах. В частности, при региональной оценке лесных ресурсов важным направлением является распознавание типов растительного покрова и породного состава лесных насаждений по снимкам среднего разрешения. Среди исследователей при работе над картированием растительного покрова широкое распространение получили снимки американского спутника Landsat-7 ETM+ [1–3].

Исходные данные. Для решения задачи классификации типа надземного покрытия и выделения классифицируемых признаков были использованы космические мультиспектральные снимки пяти спектральных диапазонов сцены (p172r021_7dt20010510 от 05.10.2001) спутника Landsat-7 ETM+.

В табл. 1 приведены характеристики используемых спектральных каналов спутника. Каналы 01, 02 и 03 соответствуют видимому излучению (каналам B, G, R, соответственно), а каналы 04 и 05 — ближнему и среднему инфракрасным спектральным диапазонам.

Пример снимка в 03 спектральном диапазоне представлен на рис. 1.

Цель. Составление тематической карты 8 природных классов: открытых участков (ОУ), сельхозугодий (СУ), участков лесовозобновления (УЛ), лиственных лесов (ЛЛ), хвойных лесов (ХЛ), смешанных лесов (СЛ), водных объектов (В) и болот (Б).

Методы анализа

Спектральный анализ. Наиболее информативным методом анализа изображений земной поверхности является спектральный анализ. Это связано с тем, что сами классы природных объектов имеют различную окраску и уровни ультрафиолетового и инфракрасного излучения.

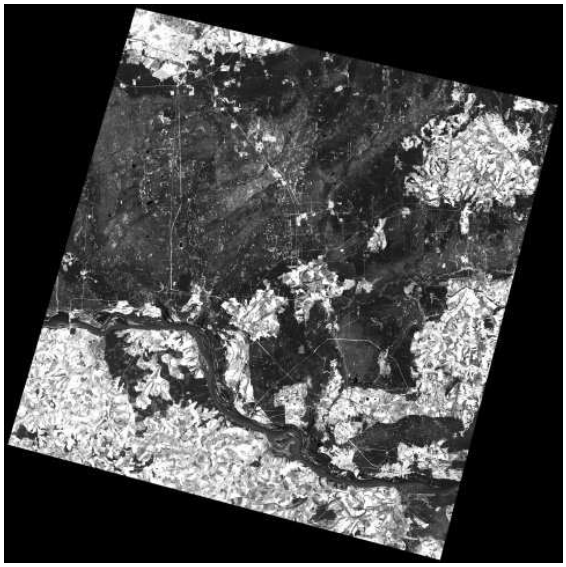


Рис. 1. Спутниковый снимок в 03 спектральном диапазоне

Таблица 1. Характеристики использования спектральных каналов спутника

№ канала	Название спектрального диапазона	Спектральный диапазон (мкм)	Пространственное разрешение (м)
01	B	0,45–0,515	30
02	G	0,525–0,605	30
03	*R*	0,63–0,69	30
04	ближнее ИК	0,75–0,9	30
05	ИК	1,55–1,75	30

Идеальная палитра. Если рассматривать отдельно взятый пиксель изображения, то в любом спектральном диапазоне он характеризуется своим значением цветности или яркости, которое лежит в пределах [0; 255]. Для исследуемого нами изображения земной поверхности использовалось 6 различных снимков (значения яркостей в 3 и R диапазонах не совпадают) в различных частотных диапазонах. Вследствие этого, можно составить некоторую характеристику отдельно взятого пикселя, которая будет описываться 6 значения-

Работа выполнена при финансовой поддержке гранта Президента РФ № МД-5418.2010.9, в рамках ФЦП «Научные и научно-педагогические кадры инновационной России» на 2009–2013 гг., ГК № 02.740.11.0838 и ГК № П 783, гранта РФФИ № 09-07-00072-а.

Таблица 2. Результаты распознавания, полученные по идеальной палитре

		α							
		ОУ	СУ	УЛ	ЛЛ	ХЛ	СЛ	В	Б
β	ОУ	97,3	0	0,52	0	0,18	0	0,07	0,5
	СУ	0	95,3	5,04	0	0	0	0	0,81
	УЛ	1,55	4,64	92,5	0,96	0,01	3,25	0	0,6
	ЛЛ	0	0	0,53	96,8	0	6,6	0	0,09
	ХЛ	0,81	0	0	0	99,2	0,51	0	3,16
	СЛ	0	0	1,38	2,2	0,13	89,1	0	1,38
	В	0,02	0	0	0	0	0	99,9	0
	Б	0,33	0,09	0,07	0,02	0,44	0,5	0	93,5

ми цветности или яркости, что в общей сложности составляет 48-битную характеристику. Несмотря на то, что характеристика имеет достаточно большой размер, это позволяет максимально точно учесть все признаки данного пикселя. В исследуемом снимке земной поверхности с количеством пикселей, составляющим 10498478, удалось выделить 2047792 отдельные характеристики.

В результате составления шестимерной палитры изображения на основе полученных характеристик были получены результаты распознавания классов природных объектов, представленные в табл. 2. Здесь и далее α — это ошибки первого рода, которые несут информативное значение пропуска цели, а β — ошибки второго рода, несут информацию о ложной тревоге.

Данный метод не учитывает тот факт, что изменение характеристики хотя бы на единицу в одном из спектральных снимков приведет к ошибочному отнесению пикселя природному классу. Может оказаться и так, что такая характеристика вовсе не будет содержаться в палитре, тогда этот пиксель будет потерян.

Модель по палитре. Для уменьшения влияния случайных факторов можно создать обобщенную палитру, которая будет обладать более высокой устойчивостью. Для этого необходимо минимизировать размер палитры, заменив её элементы диапазонами характеристик. Так, если представить характеристики в виде 48-битных чисел, то можно произвести их упорядочивание (скажем, по возрастанию) с последующим объединением в группы смежных характеристик, принадлежащих одному природному классу. При объединении появляются выбросы, когда на достаточно большом диапазоне характеристик, принадлежащих одному природному классу, появляются отдельные характеристики с небольшим количеством пикселей других классов. Статистически отнесение таких характеристик к доминирующему классу не принесет значительного ухудшения результатов. В итоге после 5 итерационных проходов с целью группировки характеристик палитры по классам удалось

Таблица 3. Результаты распознавания, полученные по модели

		α							
		ОУ	СУ	УЛ	ЛЛ	ХЛ	СЛ	В	Б
β	ОУ	89,7	0	2,48	0	0,5	0	0,29	1,5
	СУ	0	89,8	7,33	0	0	0	0	1,21
	УЛ	2,61	9,83	87,6	1,1	0,03	4,53	0	1,25
	ЛЛ	0	0	0,65	96,7	0	8,4	0	0,16
	ХЛ	4,12	0	0,18	0	98,5	1,21	0	4,27
	СЛ	0	0	1,2	2,18	0,25	84,6	0	2,63
	В	0,27	0	0	0	0	0	99,7	0
	Б	3,28	0,37	0,55	0,06	0,68	1,28	0	89

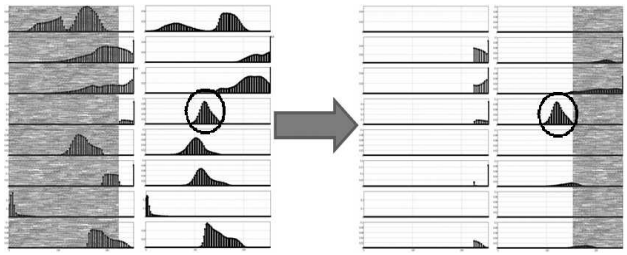


Рис. 2. Пример анализа и усечения гистограмм для выделения природного класса лиственных лесов на спутниковых снимках в 04 и 05 диапазонах

сократить размер палитры до 5682 характеристик (диапазонов), получив следующие результаты распознавания, приведенные в табл. 3.

Анализ и взаимное усечение гистограмм. Кроме построения палитры путем полного перебора характеристик, эффективным методом является анализ гистограмм различных спектральных снимков и их взаимное последовательное усечение. Суть заключается в том, что усечение гистограммы на заданном интервале позволяет не рассматривать пиксели, которые явно не принадлежат искомому классу. Но пиксели с этими же пространственными координатами также не принадлежат исследуемому классу и на снимках в других спектральных диапазонах. Поэтому их гистограммы также перестроятся. Однако, если до перестроения гистограммы природных классов пересекались, то после перестроения ситуация может измениться, и искомый класс может оказаться ортогональным по яркости с другими классами в каком-либо спектральном диапазоне. Подобным методом удалось выделить класс лиственных лесов, используя лишь спектральные снимки 04 и 05 диапазонов (рис. 2). Результаты распознавания приведены в табл. 4.

Кроме спектрального анализа мультиспектральных снимков, который дал хорошие результаты распознавания для всех исследуемых объектов природных классов, стоит упомянуть и о других методах, которые позволяют выделить с достаточной точностью лишь отдельные природные классы, однако эти методы более устойчивы к статистиче-

Таблица 4. Результаты распознавания природного класса лиственных лесов методом анализа и усечения гистограмм

		α							
		ОУ	СУ	УЛ	ЛЛ	ХЛ	СЛ	В	Б
β	ЛЛ	0	0	0,65	96,7	0	8,4	0	0,16

Таблица 5. Результаты контурного анализа снимка

Природный класс	Доля распознавания, %
ОУ	32,2
СУ	92,8
УЛ	80,7
ЛЛ	4,88
ХЛ	2,83
СЛ	5,52
В	4,21
Б	28,46

ской неоднородности снимков и влиянию различных внешних факторов. К таким методам относятся контурный и морфологический анализ.

Контурный анализ. Для контурного анализа спутниковых снимков подстилающей поверхности необходима предварительная обработка изображений. Для этого на первом этапе использовалось квантование 24 битного RGB изображения со сглаживанием по методу Байера [4], в результате чего формировалось квантованное 6-битовое изображение. Для фильтрации шумов квантования использовалась медианная фильтрация. Контрастные детали изображения лучше сохраняются при малых значениях апертуры, поэтому размер окна медианного фильтра был выбран 3×3. Следующим этапом является процедура обнаружения точек классов. Для этих целей использовалась процедура двухальтернативного обнаружения точек объектов, в результате чего формируется бинарное черно-белое изображение необходимого природного класса. Выбор порогового значения в обнаружителе определялся на основании критерия Неймана–Пирсона.

На третьем этапе производится поиск граничных точек сегментированного изображения, удовлетворяющих условию восьмисвязной границы. Пиксель объекта во внутреннем элементе окна размером 3×3 признается граничным, если ходя бы одна из смежных точек в окне принадлежит фону.

На заключительном этапе выполняется процедура выделения контура по алгоритму Розенфельда [5] с последующей фильтрацией контуров по их длине. Для хранения самих контуров использовался код Фримена.

Методом контурного анализа удалось выделить два природных класса — сельхозугодия и участки лесовозобновления — с достаточно высокой степенью вероятности (табл. 5).

Морфологический анализ. Метод морфологического анализа заключается в пространственном анализе пикселей изображения. Он применим в тех случаях, когда на карте остаются пиксели, которые по каким-либо причинам нельзя отнести ни к одному из природных классов. В этом случае можно считать, что этот пиксель будет принадлежать тому классу, пиксели которого граничат с ним.

Выводы

При анализе мультиспектральных спутниковых снимков подстилающей поверхности для выделения природных классов недостаточно учитывать лишь яркостные характеристики пикселей изображения. Для полного и достоверного анализа, который будет устойчив к воздействию внешних факторов, необходим комплексный подход, заключающийся в использовании различных методов для анализа изображений.

Литература

[1] *Курбанов Э. А., Воробьев О. Н., Губарев А. В., Лежнин С. А., Незамаев С. А.* Использование спутниковых снимков высокого и среднего разрешений для изучения естественного возобновления сосны на землях запаса // Международное сотрудничество в лесном секторе: баланс образования, науки и производства: Материалы международной конференции. — 2009. — С. 249–253.

[2] *Hall R. J., Skakun R. S., Arsenault E. J., Case B. S.* Modeling forest stand structure attributes using Landsat ETM+ data: Application to mapping of aboveground biomass and stand volume // Forest Ecology and Management Journal, 2006. — No. 225. — Pp. 378–390.

[3] *Labrecque S., Foumier R. A., Luther J. E., Piercey D.* A comparison of four methods to map biomass from Landsat-TM and inventory data in western Newfoundland // Forest Ecology and Management Journal, 2006. — No. 226. — P. 129–144.

[4] Intel Image Procesing Library. Reference Manual // Intel Corm., V.2: Image and Video Processing, 2009. No. A70805-025US.

[5] *Фурман Я. А., Кревецкий А. В., Передреев А. К., Роженцов А. А., Хафизов Р. Г., Егошина И. Л., Леухин А. Н.* Введение в контурный анализ; приложения к обработке изображений и сигналов // М.: ФИЗМАТЛИТ, 2003. — С. 588.

Подход к измерению активности выброса радиоактивных веществ по данным мониторинга радиационной обстановки

Арутюнян Р. В.[1]***, Огарь К. В.***[1]***, Ушмаев О. С.***[2]
arut@ibrae.ac.ru, kvo@ibrae.ac.ru, oushmaev@ipiran.ru
[1] г. Москва, Институт проблем безопасного развития атомной энергетики РАН
[2] г. Москва, Институт проблем информатики РАН

Предложен подход к оценке суммарной активности аварийного выброса радиоактивных веществ в атмосферу по данным автоматизированного инструментального контроля радиационной обстановки. Перенос радионкулидов в атмосфере зависит от множества внешних факторов, многие их которых сложно или невозможно оперативно определить. Рассмотрены результаты измерения мощности дозы гамма-излучения на удалении от источника выброса как классификаторы, которые свидетельствуют в пользу различных соотношений активности выброса и внешних факторов. На основе множества измерений строится оценка суммарной активности.

В соответствии с федеральным законодательством Российской Федерации, нормами и правилами использования атомной энергии федеральные и областные органы исполнительной власти, объекты использования атомной энергии осуществляют контроль радиационной обстановки на подведомственной территории.

В настоящее время все большее распространение получают автоматизированные системы непрерывного контроля основных параметров радиационной обстановки как на ядерно и радиационно опасных объектов (ЯРОО), так и на прилегающей территории и на территории субъектов Российской Федерации в целом [1]. Назначением автоматизированных систем контроля радиационной обстановки (АСКРО) является инструментальный контроль радиационной обстановки (подтверждение нормальной радиационной обстановки в местах расположения постов контроля при повседневной деятельности, раннее предупреждение об изменении радиационной обстановки, обеспечение данными по радиационной обстановке в местах размещения постов контроля в режиме чрезвычайной ситуации) и информационная поддержка деятельности территориальных и федеральных органов исполнительной власти по обеспечению радиационной безопасности.

В статье рассмотрены возможности использования АСКРО для решения задачи определения параметров аварийного выброса радиоактивных веществ в атмосферу (в первую очередь, общей активности выброса и пропорционального состава выброса). Такая информация может значительно повысить эффективность мер по защите населения за счет своевременного и адекватного реагирования.

Подход к оценке выброса в атмосферу

При определении параметров выброса радиоактивных веществ в атмосферу по данным датчиков АСКРО мы заимствуем идеи из радиолокации и теории объединения свидетельств. Например, в радиолокации один и тот же уровень отраженного сигнала соответствует целой прямой (или конусу при погрешности измерений) равнозначных положений объекта. При наличии нескольких датчиков (приемных станций) положение объекта может быть уточнено за счет наблюдения объекта с различных направлений. В задаче определения параметров выброса радиоактивных веществ мы применяем схожую идею. Проиллюстрируем ее на следующем примере.

Пусть наблюдение за обстановкой осуществляется с одного поста АСКРО, расположенного в 10 км к югу от источника выброса. Известен состав выброса: при выбросе в атмосферу единовременно попадает изотоп $^{137}_{55}$Cs. По данным датчиков мощности дозы АСКРО требуется определить суммарную активность.

На рис. 1 представлено предполагаемое пространственное распределение мощности дозы при различных направлениях ветра через 30 мин после выброса (по данным модели «Нострадамус» [2–4]). На рис. 2 представлены графики мощности дозы гаммы-излучения на посту АСКРО при различных направлениях ветра. Как видно из графика, при одной и той же суммарной активности выброса, но различных направлениях ветра, мощность дозы на посте АСКРО может различаться на порядки. В условия погрешности или полного отсутствия информации, например, о направлении ветра, по данным АСКРО можно принять два принципиально различных решения: активность выброса велика, но пост находится на периферии области распространения выброса или активность невелика, но пост находится на оси области выпадения. Кривая возможных соотношений суммарной активности и направлений ветра при одной наблюдаемой максимальной мощности дозы представлена на рис. 3. Из рисунка очевидно, что по данным одного датчика практически невозможно сделать заключение о величине суммарной активности.

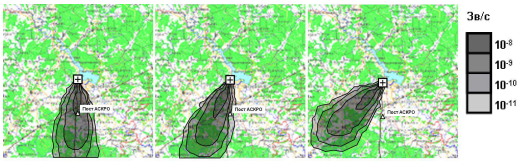


Рис. 1. Мощность дозы при выбросе в атмосферу изотопа $^{137}_{55}$Cs, суммарная активность 10^5 Ки, скорость ветра 5 м/с, 30 мин после выброса, направление ветра (слева направо): Ю, ЮЮВ, ЮВ

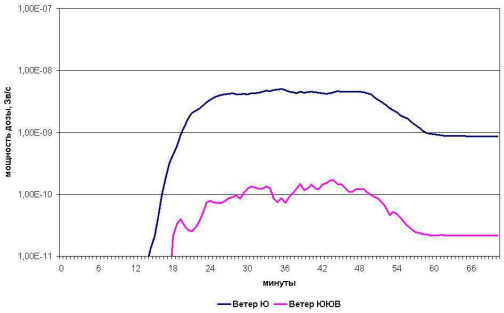


Рис. 2. Динамика мощности дозы на посте АСКРО в зависимости от направления ветра

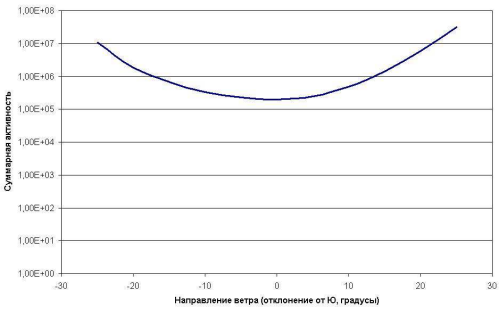


Рис. 3. Кривая возможных соотношений суммарной активности и направления ветра при максимальной мощности дозы 10^{-8} Зв/с на посте АСКРО

Если постов наблюдения два и более (рис. 4), то динамики мощности дозы на различных постах коррелированы (рис. 5) и кривые гипотез могут пересекаться (рис. 6). Точка, в которой измерения всех постов максимально согласованы, может рассматриваться как искомая величина суммарной активности.

Алгоритм оценки суммарной активности при известном радионуклидном составе

Обозначим внешние условия через $c = (c_1, \ldots, c_n)$. Мощности дозы на постах АСКРО при заданных внешних условиях c при суммарной активности выброса 1 Ки обозначим через $u^c(t) = (u^c_1(t), \ldots, u^c_k(t))$. При суммарной активности Q

Рис. 4. Модельное расположение постов АСКРО

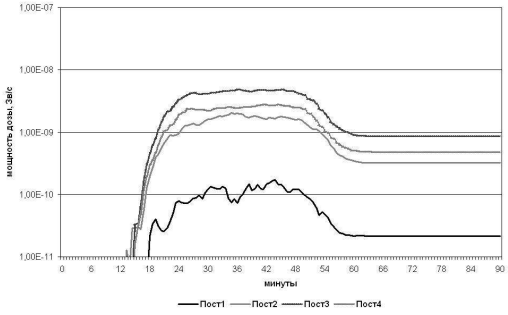


Рис. 5. Модельная динамика мощности дозы на постах АСКРО, ветер ЮЮВ, скорость ветра 5 м/с

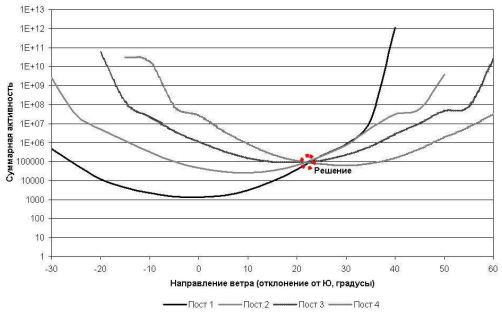


Рис. 6. Кривые гипотез о суммарной активности и направлении ветра на модельном примере (истинное значение суммарной активности изотопа $^{137}_{55}$Cs 10^5 Ки, направление ветра ЮЮВ), максимальные мощности дозы: $1{,}7 \cdot 10^{-10}$ Зв/с (пост 1), $2{,}0 \cdot 10^{-9}$ Зв/с (пост 2), $5{,}0 \cdot 10^{-9}$ Зв/с (пост 3), $2{,}8 \cdot 10^{-9}$ Зв/с (пост 4)

показания $w^c(t) = (w^c_1(t), \ldots, w^c_k(t))$ датчиков постов АСКРО равны

$$w^c_i = Q u^c_i(t) + \varepsilon_i(t)\,,$$

где $\varepsilon_i(t)$ — шум, включающий погрешность измерений и естественный фон.

Пусть известны показания датчиков мощности дозы $W_i(t)$. Требуется, во-первых, определить, произошел ли выброс с активностью, превышающей некоторый порог Q^*; во-вторых, определить суммарную активность Q и вектор внешних условий c. Внешние условия включают такие факторы, как направление и скорость ветра, пропорциональный состав выброса и др. При решении этих задач

мы сталкиваемся с такой проблемой, как неизвестные значения u^c. В силу объективных причин мы не можем провести натурные испытания процесса распространения радионуклидов в атмосфере. Поэтому приходится опираться на модели переноса схожих по физическим свойствам примесей в атмосфере. Пусть при заданных внешних условиях c у нас есть модельные значения $m^c(t)$ мощности дозы на постах АСКРО. Результаты верификации модели переноса [3] позволяют считать, что модели $m^c(t)$ и $u^c(t)$ связаны мультипликативно. Погрешность модели характеризуется распределением $m^c(t)/u^c(t)$. Показания датчиков АСКРО можно представить в следующем виде:

$$w_i^c = Qe_i m_i^c(t) + \varepsilon_i(t)\,, \qquad (1)$$

где $e_i \sim \log N(0,\sigma^2)$ — мультипликативный шум, который характеризует несоответствие модели и процесса переноса радионуклидов.

Для определения активности и внешних условий предлагается следующий алгоритм, который состоит из двух компонентов: обучение по модели и распознавание. При обучении моделируется распространение изотопов при всевозможных значениях внешних условий. Для каждого поста АСКРО формируется калибровочная таблица: модельные мощности дозы $m_i^c(t)$.

При распознавании на каждом посту АСКРО на основе показаний мощности дозы $W_i(t)$ для каждого значения внешних условий c формируются гипотезы об активности

$$Q_i^c(t) = \frac{W_i(t)}{m_i^c(t)}\,.$$

В качестве решения для внешних условий c и наилучшего момента наблюдения выбирается наиболее согласованное значение

$$(C,T) = \arg\min_{c,t} \frac{\max\limits_i Q_i^c(t)}{\min\limits_i Q_i^c(t)}\,. \qquad (2)$$

Оценка суммарной активности $\widehat{Q}$ определяется как среднее геометрическое гипотез датчиков с ненулевыми модельными значениями мошности дозы:

$$\widehat{Q} = \sqrt[l]{Q_{i_1}^C(T)\ldots Q_{i_l}^C(T)}\,. \qquad (3)$$

Ошибки модели

Как видно из (1), итоговый результат оценки мощности дозы зависит от характеристик шума e_i и $\varepsilon_i(t)$. На практике аддитивный шум крайне незначителен в сравнении с представляющими интерес значениями мощности дозы. Поэтому в данном разделе основное внимание уделено мультипликативному шуму e_i, который характеризует несоответствие модели и процесса переноса радионуклидов. Для моделирования шума целесообразно использовать логнормальное распределение, т. е. $u^c/m^c \sim \log N(\mu,\sigma^2)$. Из (3) следует, что распределение отношения оценки к истинному значению $\widehat{Q}/Q$ можно грубо считать логнормальным с параметрами μ/l и σ^2/l. Это является заниженной оценкой ошибки, так как большую погрешность вносит неправильная оценка внешних условий (2) из-за рассогласования датчиков. Иллюстрация рассогласования приведена на рис. 7. Для оценки этого фактора мы провели статистическое моделирование.

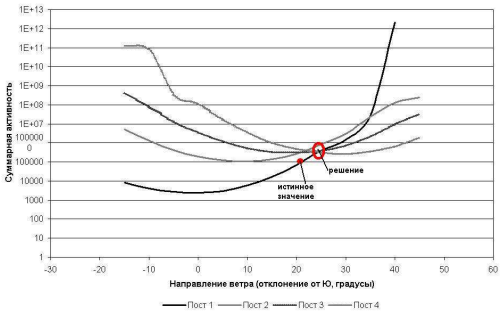


Рис. 7. Пример рассогласования показаний датчиков

Результаты моделирования показали, во-первых, что $\widehat{Q}/Q$ распределено по логнормальному закону. Во-вторых, параметры распредений отличаются от изначальной грубой оценки (рис. 8). При умеренных ошибках модели (среднеквадатическое отклонение менее чем в два раза) мы получаем несмещенную оценку с приемлемой дисперсией. Согласно экспериментальным данным верификационного отчета [3], ошибка используемой нами модели переноса находится в зоне умеренных ошибок ($\sigma \sim 0{,}2$). При больших ошибках модели мы наблюдаем сильное смещение оценки из-за рассогласования датчиков. Поэтому, даже если мы будем использовать очень приближенную модель переноса в атмосфере, предлагаемый алгоритм будет давать завышенную оценку активности выброса, что, с точки зрения обеспечения защиты населения, лучше заниженной оценки.

Эксперименты

Численные эксперименты проводились на примере двух гипотетических АСКРО (схема расположения постов контроля в 20-километровой зоне представлена на рис. 9). Основной целью экспериментов являлось исследование возможности определения активности выброса с учетом действующей конфигурации постов АСКРО. Ставились две задачи:

1. Установить возможность детекции факта выброса.
2. Установить возможность оценки активности выброса и внешних условий.

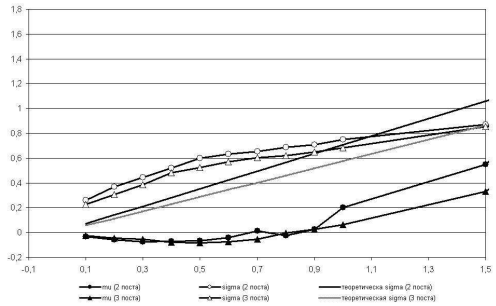


Рис. 8. Параметры ошибки измерения активности по данным статистического моделирования

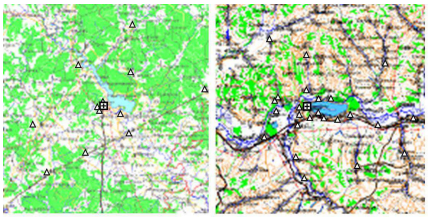


Рис. 9. Расположение постов АСКРО

Эти задачи решались путем анализа калибровочных таблиц датчиков АСКРО по модели распространения радионуклидов в атмосфере. В качестве меры возможности определения активности выброса и внешних условий рассматривалось количество датчиков, фиксирующих повышение уровня мощности дозы гамма-излучения при прохождении радиоактивного облака. Области внешних условий (сила и направление ветра, высота выброса и т. д.), при которых ни на одном из постов АСКРО не фиксируется повышение уровня мощности дозы гамма-излучения, рассматривались как условия, при которых невозможно детектировать факт выброса. Если более двух датчиков фиксируют прохождение облака, то принималось предположение о возможности определения активности выбросов. Визуализация результатов приведена на рис. 10. Изображена диаграмма чувствительности АСКРО при различных направлениях и силе ветра. АСКРО № 1 имеет слепые зоны. Поэтому ее целесообразно расширить.

Заключение

В статье предложен подход к определению суммарной активности выброса радиоактивных веществ по данным датчиков мощности дозы. Предложенный подход позволяет оценивать возможность обнаружения/необнаружения выброса радионуклидов в атмосферу для выбранной пространственной конфигурации датчиков.

Представленное исследование целесообразно развивать в следующих направлениях. Во-первых, известные на сегодняшний день модели переноса примесей в атмосфере имеют погрешность. Поэтому предложенные гипотезы о соотношении активности и мощности дозы следует дополнить вероятностными характеристиками. Во-вторых, многие параметры, которые могут быть приняты как известные (например, скорость ветра), также могут иметь стохастический характер. В-третьих, представляет интерес более детальный анализ показаний мощности дозы во времени. Различные изотопы имеют различный период полураспада. «Расслоение» мощности дозы позволяет сделать заключение о пропорциональном вкладе отдельных изотопов в выброс, что очень важно, так как различные изотопы представляют различную опасность для населения. После определения вклада изотопов в мощность дозы можно решать обратную задачу определения пропорционального состава выброса. Также перспективным является разработка методов уточнения оценки активности за счет дополнительного инструментального контроля. Для этого следует разработать методы определения таких точек, измерения в которых дадут наибольший эффект с точки зрения уточнения оценки выброса.

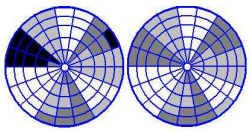


Рис. 10. Диаграмма чувствительности АСКРО № 1 (слева) и АСКРО № 2 (справа) по различным направлениям и силе ветра (радиус — сила ветра; угол — направление ветра; черный сектор: АСКРО не в состоянии детектировать факт выброса; белый сектор: АСКРО позволяет определить внешние условия и активность выброса)

Литература

[1] *Арутюнян Р. В., Осипьянц И. А., Киселев В. П., Гаврилов С. Л., Огарь К. В.* Совершенствование региональных систем радиационного мониторинга и аварийного реагирования // Тезисы докладов 6-й международной конференции МНТК-2008, Безопасность, эффективность и экономика атомной энергетики, Москва, 21–23 мая 2008 г. — С. 183–185.

[2] Паспорт аттестации ПС № 158 от 28.03.2003, свидетельство об официальной регистрации программы для ЭВМ № 2003612220.

[3] Верификация компьютерной системы «НОСТРАДАМУС» для прогнозирования радиационной обстановки на ранней стадии аварии на АЭС. Верификационный отчет. ИБРАЭ РАН, инв. N 3431. — М., 2001.

[4] *Арутюнян Р. В., Беликов В. В., Беликова Г. В., Сороковикова О. С. и др.* Компьютерная система НОСТРАДАМУС для поддержки принятия решений при аварийных выбросах на радиационно опасных объектах // Известия Академии наук, серия Энергетика, 2005. — № 4.

Методы и алгоритмы распознавания объектов сельских поселений на цифровой карте

Шлей М. Д., Рогов А. А., Борисов А. Ю.
shlei@psu.karelia.ru rogov@psu.karelia.ru borisoff@psu.karelia.ru
Петрозаводск, Петрозаводский государственный университет

В докладе представлены результаты разработки алгоритмов для распознавания объектов на цифровом плане местности сельского поселения. В результате работы предложенных алгоритмов на плане местности автоматически находятся жилые постройки, вспомогательные хозяйственные постройки и заборы. Информация о найденных объектах используется в качестве входных данных для модели расчета продолжительности инсоляции жилых построек. Предложенные алгоритмы и методы были реализованы в программной системе, предназначенной для проведения историко-архитектурного анализа объемно-планировочной структуры традиционных сельских поселений Русского Севера.

При проведении историко-архитектурных исследований традиционных сельских поселений основным материалом для исследования являются генеральные планы анализируемых поселений [2]. В классическом виде проведения анализа планы местности обрабатываются исследователем вручную, что замедляет процесс исследования и снижает точность получаемых данных. Для каждого поселения необходимо выделить имеющиеся жилые постройки, определить главные фасады и азимуты их ориентации. Использование цифровых генеральных планов поселений представленных в векторных форматах позволяет автоматизировать процесс их обработки. В докладе представлены результаты исследования по разработке алгоритмов и методов, предназначенных для обработки цифровых карт, с целью выделения построек и определения их пространственных характеристик.

Постановка задачи и объект исследования

Объектом исследования выступает цифровой генеральный план местности, представленный в векторном виде и подготовленный исследователями в соответствии с установленными требованиями (требования указаны в [4]) в системе AutoCAD. На рис. 1 представлен пример фрагмента подготовленного плана сельского поселения.

Для проведения анализа плана поселения требуется найти главные фасады у построек, предоставляющих историко-архитектурный интерес. Перпендикуляр, проведенный к главному фасаду, будет определять азимут его направления (см. рис. 2).

План поселения представлен в векторном формате и представляет собой массив отрезков $[A_i, B_i]$, где $i \in D$ — номер отрезка, точка — начало отрезка, B — соответственно, конец. Исходная задача обработки карты заключается в выделении из общего массива $[A_i, B_i]$ наборов отрезков $[A_e^k, B_e^k]$, таких, что каждый набор будет представлять собой обозначение одной постройки, на карте. Для решения данной задачи предлагается в первую очередь выделить главные фасады, а затем найти вектора, относящиеся к каждой постройке (близкие к главным фасадам).

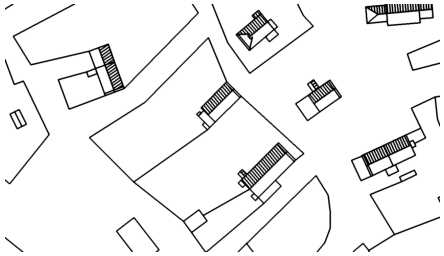


Рис. 1. Пример фрагмента плана сельского поселения

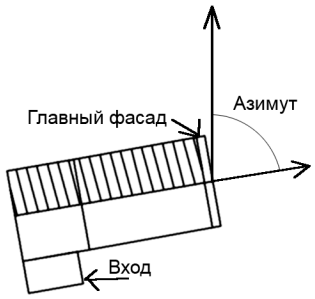


Рис. 2. Обозначение постройки на карте

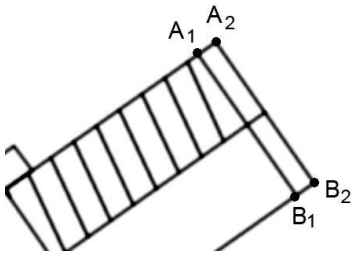


Рис. 3. Выделение отрезков, задающих главный фасад у здания

Распознавание главных фасадов

Пусть существует набор отрезков $[A_i, B_i]$. Необходимо выделить пары векторов, обозначающие главные фасады у построек. На рис. 3 два отрез-

ка $[A_1, B_1]$ и $[A_2, B_2]$ обозначают главный фасад. Они удовлетворяют следующим условиям:

1. $[A_1, B_1]$ параллелен $[A_2, B_2]$ и $[A_1, A_2]$ перпендикулярен $[A_1, B_1]$ и $[B_1, B_2]$ перпендикулярен $[A_1, B_1]$. То есть отрезки являются противоположными сторонами прямоугольника.
2. $[A_1, B_1]/[A_1, A_2] < P$, где P — параметр указывающий отношение расстояния между отрезков, обозначающих главный фасад к их длине. Данный параметр указывается исследователем или определяется эвристическим алгоритмом для заданного исследователем диапазона количества построек на карте.

Опишем алгоритм поиска главных фасадов. На вход исследователем подается минимальное и максимальное количество построек на карте N_1 и N_2 соответственно. Параметр P имеет начальное значение 1. Введем переменные $A = 0$ и $B = 1$ для поиска P.

Шаг 1. Определяем количество главных фасадов. Перебираем все вектора и проверяем их на условие обозначения главного фасада для параметра $P = (A + B)/2$.

Шаг 2. Если полученное количество главных фасадов попадает в диапазон $N_1, \ldots, N_2$, то завершаем работу алгоритма. Иначе на Шаг 3.

Шаг 3. Если полученное количество главных фасадов больше чем $N2$, то $B = P$. Иначе $A = P$. Переходим на Шаг 1.

Замечание 1. При определении параллельности и перпендикулярности линий, проходящих через исследуемые отрезки, необходимо учитывать возможную погрешность инструментальной среды, в которой подготавливалась карта. То есть отрезки, обозначающие главный фасад, могут быть не «идеально» параллельными и составлять не «идеальный прямоугольник». Данная погрешность выведена эмпирически на основе исследования цифровых планов и составляет 5%.

Перебрав все возможные пары отрезков из $[A_i, B_i]$, можно выделить те, которые обозначают главные фасады построек на плане местности (удовлетворяют вышеуказанным условиям). После этого необходимо найти все вектора, относящиеся к данным постройкам.

Выделение жилых построек

Пусть из исходного набора отрезков выделено K пар, обозначающих главные фасады. Поместим данные пары в K групп. Все другие оставшиеся отрезки (свободные) помещаем в группу $K+1$, которую обозначим через H. Пусть в группе H таких отрезков будет m. Далее каждый отрезок из группы H, будем проверять на близость для групп от 1 до K. Для этого определим функцию близости отрезка к группе.

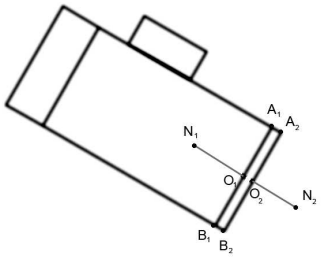


Рис. 4. Определение внешней стены главного фасада

Пусть имеется отрезок $[A_0, B_0]$ и группа M, состоящая из n отрезков ($[C_i, D_i]$, где $i = 1, \ldots, n$). Расстояние между группой и отрезком будет равным 0, если выполняются следующие условия:

1. Отрезок $[A_0, B_0]$ пересекается хотя бы с одним из отрезков группы M и при этом $[A_0, B_0]$ параллелен или перпендикулярен хотя бы с одним из отрезков группы M.
2. Точки A_0 и B_0 должны лежать от концов отрезков главного фасада на расстоянии не больше D (задается исследователем).

Если данные условия не выполняются, то расстояние будет очень большим.

Опишем алгоритм выделения жилых построек.

Шаг 1. Все отрезки, обозначающие главные фасады, разбиваем на K групп. Оставшиеся отрезки переносим в группу H.

Шаг 2. При помощи функции близости определяем все расстояния от отрезков из группы H до каждой группы j (j от 1 до K).

Шаг 3. Проверяем расстояние между каждым свободным отрезком и каждой j-й группой (j от 1 до K), если расстояние равно нулю, то переносим отрезок в соответствующую группу. Если нулевых расстояний не было, то завершаем алгоритм, т. к. дальше не будет происходить изменений в группах от 1 до K. Иначе необходимо пересчитать все расстояния и продолжить проверку, поэтому идем на Шаг 2.

В итоге работы алгоритма получим что в группах от 1 до K, будут сгруппированы отрезки, относящиеся к выделенным постройкам.

Определение азимута у главного фасада

Пусть в результате работы алгоритма распознавания жилых построек, выделено K отрезков относящихся к определенной постройке. Из них отрезки $[A_1, B_1]$ и $[A_2, B_2]$ обозначают главный фасад. Чтобы определить азимут направления главного фасада, необходимо знать какой из этих двух отрезков является обозначением внешней стены у постройки. Для этого построим к отрезкам $[A_1, B_1]$

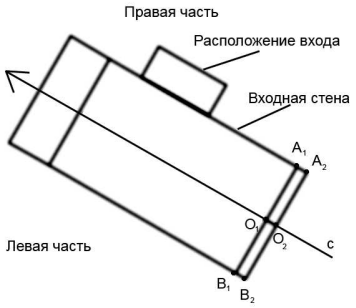


Рис. 5. Определение положения входа

и $[A_2, B_2]$ перпендикулярные отрезки $[O_1, N_1]$ и $[O_2, N_2]$, направленные в разные стороны и с одинаковой длиной (см. рис. 4). Подсчитаем сумму расстояний от точек N_1 и N_2 до всех других точек концов отрезков, обозначающих постройку. Обозначим суммы как S_1 и S_2 соответственно. Если $S_1 > S_2$, то отрезок $[A_1, B_1]$ будет являться внешней стеной. Иначе внешняя стена — отрезок $[A_2, B_2]$. Далее для входной стены оставляем перпендикуляр и подсчитываем угол между осью OY в декартовой системе координат и данным перпендикуляром. Полученный угол будет азимутом направления главного фасада.

Определение положения входной стены

В типизации построек характерных для традиционных сельских поселений Русского Севера, условно выделены два типа относительно положения входной стены: правостороннее и левостороннее расположение.

Пусть имеется К отрезков, обозначающих определенную постройку, где отрезки $[A_1, B_1]$ и $[A_2, B_2]$ являются обозначением главного фасада. Для определения указанной характеристики будем использовать следующий метод. Проведем прямую c, проходящую через середины отрезков $[A_1, B_1]$ и $[A_2, B_2]$. Данная прямая условно разбивает плоскость на левую и правую часть, если смотреть по направлению от главного фасада внутрь дома (см. рис. 5). Подсчитаем расстояние от каждой точки концов отрезка, принадлежащего к постройке, до прямой c. Среди подсчитанных расстояний найдем максимальное, соответствующее наиболее отдаленной точки принадлежащей постройки от прямой c. Если данная точка лежит в левой части, то можно сказать, что постройка имеет левостороннее расположение входа. Иначе правостороннее.

Модель расчета продолжительности инсоляции

В результате работы представлено алгоритма на цифровой карте выделяются постройки. Для каждой выделенной постройки определяется азимут направленности главного фасада и расположение входной стены. Полученные данные используются в качестве входных параметров для других моделей по расчету характеристик постройки. Например, моделью расчета продолжительности инсоляции жилых построек, описанной в [3]. На вход данной модели передается информация об азимуте направления главного фасада. На основании полученных данных модель рассчитывает продолжительность освещения различных частей постройки: главного фасада, входной стены, красного угла и общей продолжительности освещения внутреннего пространства.

Таблица 1. Результаты проверки алгоритма поиска главных фасадов

	План местности 1	План местности 2	План местности 3
Время (сек.)	9,14	0,8	0,2
Полнота	0,98	0,90	1,00
Точность	0,98	0,90	1,00

Поиск дополнительных объектов на плане местности

Кроме жилых построек на плане также присутствуют обозначения дополнительных объектов: сараи, хозяйственные постройки, линии дорог и огородов, которые тоже необходимо найти. Для этого применяется алгоритмом «Форель-1» (параллельная кластер-процедура) [1].

Пусть $[A_i, B_i]$, $i = 1, \ldots, k$ набор отрезков, оставшихся на плане местности после выделения жилых построек. На первом этапе проведем поиск вспомогательных построек. Рассмотрим совокупность точек концов отрезков $\{A_1, \ldots, A_i, B_1, \ldots, B_i\}$. Выберем произвольную точку и найдем все точки, лежащие рядом с ней на расстоянии не больше чем D (задается исследователем или используется максимальная длина жилой постройки). После этого для данной группы производим поиск точки центра масс C, координаты которой будут равны среднеарифметическому координат всех точек группы. Затем находим все точки, ближайшие к C (расстояние $< D$), и рассчитываем новый центр масс. Продолжаем до тех пор, пока точки, попавшие в группу, не перестанут меняться. Если отрезки, к которым принадлежат данные точки, можно включить в окружность с радиусом D, то переносим данные отрезки в класс S_1, обозначающий дополнительную постройку, иначе переносим в класс L, содержащий отрезки, обозначающие линии заборов. После этого из всех оставшихся точек опять выбирается произвольная и повторяется вышеуказанная процедура для поиска дополнительного объекта S_2 и т. д.

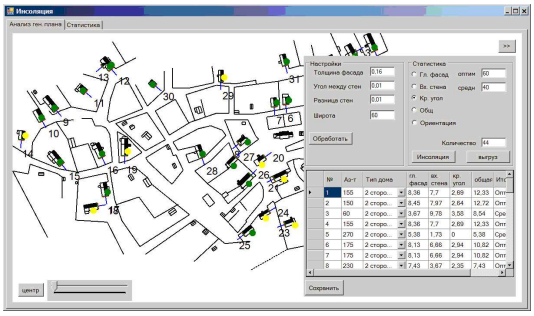


Рис. 6. Программная система

После работы данной процедуры в классах $S_1, \ldots, S_k$ будут находиться объекты, обозначающие сараи и хозяйственные постройки, а в классе L линии заборов на плане местности. На основе информации о линиях заборов с помощью эвристических алгоритмов строится осевые линии дорог, которые располагаются между близко лежащими параллельными линиями заборов. Далее выделенные части дорог соединяются кривыми линиями, так, чтобы они не пересекали найденные постройки и заборы.

Компьютерная реализация

Предложенный метод поиска главных фасадов был реализован в программной системе [3]. При помощи реализованной программной системы была проведена проверка работы алгоритма выделения главных фасадов на трех цифровых планах местности. Полученные результаты представлены в табл. 1.

Для планов местности под номерами 1 и 2 алгоритм не находит все главные фасады у построек, это связано с тем, что в данных планах используются короткие отрезки для обозначения заборов. Для корректировки данных о поселении в системе реализована возможность ручного добавления и удаления информации о жилых постройках. Данные о выделенных постройках используются математической моделью расчета продолжительности освещения жилых построек. Для каждого найденного дома рассчитывается продолжительность освещения различных частей. Также система предоставляет средства для получения описательной статистики по поселению. Подсчитываются следующие показатели: наиболее характерное пространственное положение построек (мода по ориентации) и наиболее характерная продолжительность освещения для данного поселения (мода по продолжительности освещенности). На рис. 6 приведен пример работы данной программной системы.

Выводы

Реализация данного метода в программной системе и проверка ее работы на реальных планах поселений показала, что метод можно использовать для реализации инструментов, позволяющих проводить историко-архитектурный анализ поселения. Информация о выделенных объектах используется в качестве входных данных для математической модели расчета продолжительности инсоляции жилых построек [3]. Дальнейшее направление по данной работе — это создание модели для определения оценки влияния водоема на структуру застройки. В данной модели будет использоваться информация о жилых постройках, рельефе местности и расположении водоема относительно поселения. Также планируется дополнить алгоритм распознавания возможностью классификации постройки с использованием справочника, подготовленным исследователями и содержащий информацию об основных схемах построек для традиционных сельских поселений. Чтобы определить тип найденной постройки, ее необходимо сравнить со схемами из справочника.

Литература

[1] *Айвазян С. А., Мхитарян В. С.* Прикладная статистика и основы эконометрики. — М.: Юнити, 1998.

[2] *Гуляев В. Ф.* Объемно-пространственная структура сельских поселений середины XIX – начала XX вв. и методика ее количественной оценки (на примере Российского севера). Дис. на соиск. уч. ст. кандидата архитектуры. Петрозаводск, 1990. т. I: 148 с., приложения 46 с., т. II: 88 с.

[3] *Шлей М. Д., Борисов А. Ю.* Информационная система комплексного историко-архитектурного анализа объемно-планировочной структуры традиционных сельских поселений Русского Севера. Университеты в образовательном пространстве региона: опыт, традиции и инновации: Материалы научно-методической конференции, посвященной 70-летию Петрозаводского государственного университета (16–17 февраля 2010 г.). Ч. II (Л-Я) / ПетрГУ. — Петрозаводск, 2010. — С. 312–316.

[4] *Шлей М. Д., Борисов А. Ю.* Методы оценки пространственных характеристик сельских поселений Карелии [Электронный ресурс] / М. Д. Шлей, А. Ю. Борисов // Режим доступа: http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2010/part2/SB.

Автоматизированная классификация сцен наземной лесной таксации с использованием статистического анализа текстур*

Ипатов Ю. А.
ipatovya@marstu.net
г. Йошкар-Ола, Марийский государственный технический университет

Рассмотрен один из подходов автоматизированной классификации сцен наземной лесной таксации на основе статистического анализа. Определены минимально достаточные характеристики для решения поставленной задачи с использованием аппарата нейросетевых структур карт самоорганизации без учителя. Исследованы результаты работы на реальных цифровых изображениях.

Введение

Современные подходы в области лесной таксации, экологического мониторинга, оценки и прогнозирования биоресурсов можно условно разделить на два научно-технических направления: первое связано с анализом крупномасштабных (данные дистанционного зондирования земли, аэрофотосъемка) и второе мелкомасштабных (снимки наземной лесной таксации) изображений.

Для первого направления созданы и широко применяются на практике автоматизированные решения: аппаратно-программные средства, реализующие методы цифровой обработки изображений и распознавания образов [1].

В то же время важные массовые наземные измерения, в частности, для решения задач дендрохронологии, динамики плодородности почв, анализа освещенности нижнего яруса лесов, измерения удельного объема биомассы растительного покрова, удельного объема технологичной древесины, сегодня выполняются неавтоматизированными трудоемкими визуальными методами, имеют высокую стоимость и большие временные затраты. Так, для некоторых задачи в данном направлении удалось создать автоматизированные программно-аппаратные комплексы, которые повышают общую эффективность работ примерно на два порядка [2, 3].

В данной работе предлагается подход автоматизированной классификации сцен наземной лесной таксации. Результаты такого анализа позволяют минимизировать работу оператора для этапов: первичной регистрации в единой базе данных, интеллектуальной обработке и анализе результатов, тем самым выигрывая по скорости проведения исследований. Также получены характеристики реализованного метода на реальных цифровых изображениях.

Постановка задачи

Схематически место для пространственной регистрации заданных классов изображений представлено на рис. 1, где E — точка регистрации

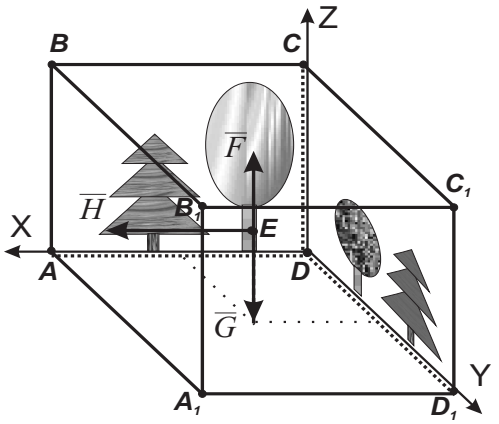


Рис. 1. Место пространственной регистрации изображений для задач наземной лесной таксации

снимка вдоль направления указанного вектора. Первый класс изображений (стволы деревьев на перспективных видах лесных массивов) регистрируется вдоль указанного направления $\overline{H} \parallel ADD_1A_1$, второй класс изображений (лиственный покров растений определенного типа) — $\overline{G} \perp ADD_1A_1$ и третий класс (кроны деревьев верхнего яруса) — $\overline{F} \perp BCC_1B_1$.

Детальный анализ и экспериментальные исследования статистических, корреляционных и спектральных характеристик трех классов изображений был проведен ранее, а также найдены ключевые параметры и закономерности, которые позволили построить математические модели и алгоритмы их анализа [2–5].

Таким образом, проведя сравнительный анализ характеристик для цифровых изображений трех классов, были выбраны доминирующие статистические признаки, которые можно будет использовать для их уникальной идентификации. Вопрос выбора минимально достаточного количества таких признаков требует детального изучения и проведения практических испытаний на существующем классе сцен.

Метод решения

В качестве статистических характеристик для создания «цифрового портрета» сцены были определены: $\bar{x}$ — среднее арифметическое яркости изоб-

Работа выполнена при поддержке гранта Президента РФ МК-2636.2011.9.

ражения $I(x,y)$, а также её дисперсии — σ^2 и среднеквадратического отклонения — σ. При этом в качестве производной рассчитывается величина коэффициента вариации, которая представляет собой относительную меру рассеивания, выраженную в процентах:

$$\nu = \frac{\sigma}{\bar{x}} \cdot 100\% .$$

При исследовании формы распределения яркостей отдельного изображений оценивается степень асимметрии As и эксцесс Ex распределения. Симметричным считается распределение, в котором частоты двух равностоящих от центра значений признака равны между собой. Степень асимметрии распределения в центральной его части определяется с помощью коэффициента асимметрии Пирсона:

$$As = \left(\sum_{i=1}^{m}(x_i - \overline{x})^3 \cdot n_i\right)\left(\sigma^3 \sum_{i=1}^{m} n_i\right)^{-1} .$$

Показатель эксцесса рассчитывается по формуле:

$$Ex = \left(\sum_{i=1}^{m}(x_i - \overline{x})^4 * n_i\right)\left(\sigma^4 \sum_{i=1}^{m} n_i\right)^{-1} - 3 ,$$

если $Ex > 0$, то распределение относится к островершинным, а при $Ex < 0$ распределение относится к плосковершинным.

Статистическая модель цифрового изображения $I(x,y)$ в нашем случае, может быть описана выражением:

$$\overline{S} = \left\{\overline{x}, \sigma^2, \sigma, \nu, As, Ex\right\} . \qquad (1)$$

Далее необходимо определить эмпирическим путем минимально достаточное количество элементов модели $\overline{S}$ для решения задачи наилучшей автоматической классификации таких изображений. При этом критерием качества отбора будет оценка:

$$\hat{\eta} = \arg\min_{i}\left(\overline{S} = \{i, i+1, \ldots, i+j\}\right) ,$$

где j — количество элементов $\overline{S}$.

Модель классификации

Для разделения независимых классов изображений будем использовать обучение без управления для кластеризации образов на основе принципов самоорганизующейся сети Кохонена [6]. Архитектура такой сети представлена на рис. 2.

Входные элементы вектора $\overline{S}$ предназначены для распределения данных между входными элементами сети. Кластерные элементы (выходные) представлены в виде двухмерного массива $M \times N$. В ходе обучения определяется элемент побудитель,

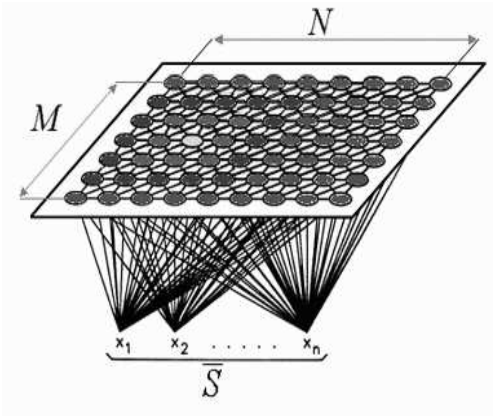


Рис. 2. Базовая архитектура сети самоорганизующейся карты признаков

для которого рассчитывается квадрат евклидового расстояния до учебного вектора:

$$d_{pq} = \sum_{i}^{n}\left(x_{pi} - x_{qi}\right)^2 ,$$

где d_{pq} обозначает квадрат евклидового расстояния между точкой p и x_{pi} — i-я координата образа p (аналогично для образа q), а n — значение размерности.

Кластером будет являться группа векторов, расстояние между которыми внутри этой группы меньше, чем расстояние до соседних групп. Структура кластеров при использовании алгоритма SOM может быть отображена путем визуализации расстояния между опорными векторами (весовыми коэффициентами нейронов) [7].

Результаты работы

Апробируя предложенный метод классификации, была создана выборка Ω_1 из 60 цифровых изображений 800×600, где каждая из исследуемых сцен была представлена в равной мере. Далее проводим оценку (1) модели $\overline{S}$ для всевозможных комбинаций элементов всего — 15. Из них 3 модели по 3 параметра, 7 моделей по 4 параметра, 4 модели по 5 параметров и 1 модель, содержащая все 6 параметров $\left(\overline{x}, \sigma^2, \sigma, \nu, As, Ex\right)$.

Из всех моделей отбираются только те, у которых минимальная ошибка первого и второго рода [8]. Таким образом, получаем модели $\overline{S_1} = \left\{\sigma^2, \nu, As, Ex\right\}$, $\overline{S_2} = \left\{\sigma, \sigma^2, \nu, As\right\}$ и $\overline{S_3} = \left\{\sigma, \sigma^2, \nu, As, Ex\right\}$, для которых средняя ошибка первого рода $F_1 = 0{,}11$ и ошибка второго рода $D_1 = 0{,}15$.

Для испытания полученных моделей $\overline{S_1}$, $\overline{S_2}$, $\overline{S_3}$ была создана тестовая выборка Ω_2 из 95 изображений, где доля первого класса изображений $\overline{H}$ составила 25%, второго класса $\overline{G}$ равняется 20% и третьего класса $\overline{F}$ составляет 55%. На рис. 3 пред-

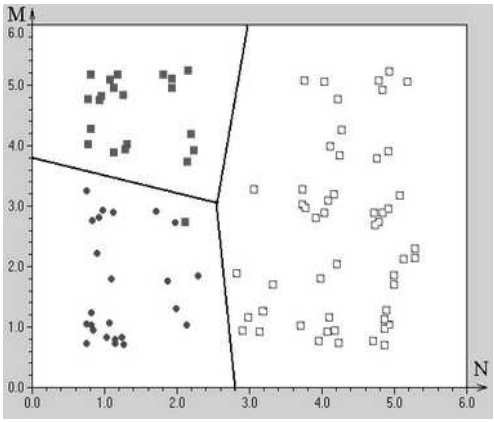


Рис. 3. Результаты разделения трех типов изображений на карте самоорганизации без учителя

ставлен наилучший результат по критерию минимальной ошибки первого и второго рода для модели $\overline{S_3} = \{\sigma, \sigma^2, \nu, As, Ex\}$.

Работа самоорганизующейся сети выполнялась при следующих параметрах: кластерные элементы размерностью $M \times N = 5 \times 5$, скорость обучения $\lambda = 1$, начальный диапазон окрестности $k = 11$ и количество эпох $n = 10\,000$.

Заключение

Таким образом, в результате статистических испытаний были определены минимально достаточные параметры для задачи кластеризации заданных типов таксационных сцен. Предложенный подход апробирован на реальных цифровых изображениях и обладает минимальной ошибкой классификации, что позволяет разделить в автоматическом режиме серию таких изображений для решения задач наземной лесной таксации. Работа выполнена при поддержке гранта Президента РФ МК-2636.2011.9.

Литература

[1] *John A. Richards, Xiuping Jia.* Remote Sensing Digital Image Analysis. — Berlin: Springer, 2006. — 454 p.

[2] *Ипатов Ю. А., Кревецкий А. В.* Алгоритм локализации границ текстурных участков древесины на их цифровых изображениях // Изв. вузов. Приборостроение. — 2009. — Т. 52, № 7. — С. 12–17.

[3] *Krevetsky A. V., Ipatov Y. A.* High Technologies in Measuring Problems of Forestry Complex on The Basis of Scene Analysis and Image Recognition Method // 8-th International Conference «PRIA: New Information Technologies»: Conference Proceedings. Yoshkar-Ola — 2007. — Vol. 2. — P. 287–289.

[4] *Ipatov Y. A.* Process automation of an estimation for relative density of wood plantings the instrumentality of methods of digital image processing // 8-th International Conference «PRIA: New Information Technologies»: Conference Proceedings. Yoshkar-Ola — 2007. — Vol. 1. — P. 307–309.

[5] *Кревецкий А. В., Ипатов Ю. А.* Выделение объектов на сложном неоднородном фоне при анализе цветных изображений в биологических исследованиях // Вестник РГРТУ. — 2008. — Т. 26, № 4. — С. 18–24.

[6] *Хайкин С.* Нейронные сети: полный курс, второе издание. — М.: Вильямс, 2006. — 1104 с.

[7] *Kohonen Teuvo* Self-Organizing Maps. — Springer-Verlag, Heidelberg, 1995.

[8] *Корн Г., Корн Т.* Справочник по математике для научных работников и инженеров. — М.: Наука, 1984. — 832 с.

Анализ методов распознавания и подсчета животных на аэрофотоснимках

Михайлов В. В.[1]***, Харин Я. В.***[2]
mwwcari@gmail.com[1], aferook@yandex.ru[2]
[1]Санкт-Петербургский институт информатики и автоматизации РАН;
[2]Санкт-Петербургский государственный университет аэрокосмического приборостроения

Данная работа посвящена проблеме автоматического распознавания и подсчета животных на фотоснимках. В качестве конкретного объекта при построении прототипа системы выбраны дикие северные олени — важнейший компонент экосистем Арктики, основа благосостояния и этнической самобытности коренных народов Севера России. Основные группировки оленей фотографируются во время их скоплений на летних пастбищах, и количество животных в них подсчитывается. Ручная обработка снимков при численности популяции в 500–600 тыс. особей занимает около 3 мес. Для повышения эффективности управления популяцией время обработки снимков должно быть снижено до 10–15 дней. Автоматизация процесса обработки фотоснимков позволит, таким образом, решить две задачи — освободить специалистов от выполнения рутинной работы и повысить качество функционирования системы «дикий северный олень – человек».

При съемке олени находятся на различном удалении от камеры, поэтому их изображения будут видны под различными углами, изображения будут иметь разные размеры и могут перекрывать друг друга. Помехи: камни, земляные бугры, впадины и т. п. — легко идентифицируются при ручной обработке снимков, но могут создать трудности при работе автоматической системы распознавания.

Система распознавания и подсчета животных должна решать следующие задачи:

1. Распознавать и подсчитывать общее числа животных на снимках. При этом животные могут быть представлены как локальными объектами, так и неразделимыми группами. Неоднородный по цвету и фактуре природный фон может содержать помехи — камни, овраги и т. п.
2. Распознавать и подсчитывать количество животных, имеющих визуально различимые признаки. Для северных оленей это телята и взрослые самцы. Условия распознавания по фону и помехам соответствуют п. 1.

В общем случае в решении задачи подсчета объектов можно выделить следующие этапы [1]: предобработка снимков, сегментация; шумоподавление и фильтрация, отнесение сегментированных областей к классам объектов, подсчет количества найденных объектов.

Первый этап необходим для подготовки изображения к распознаванию. На этом этапе производится отчистка изображений от помех и шумов. Под помехами и шумами понимаются сторонние возмущения, неселективные в отношении объектов и фона, действующие в системах создания, передачи и воспроизведения фотоснимков. В качестве фильтров для удаления помех и шумов служат различного рода усредняющие, частотные и пространственные фильтры [1, 2]. На данном этапе в связи с необходимостью поиска признаков объектов был выбран медианный фильтр [2]. Выбранный фильтр показал лучшие результаты удаления помех по сравнению с линейными сглаживающими фильтрами, сохранив при этом четкость изображения. При использовании медианного фильтра важно определить размер окна фильтра. Практическим путем было установлено, что для удаления помех на представленных снимках размер окна должен составлять 0,1–0,5 от среднего размера объекта.

Второй этап — сегментация — процесс проверки каждого отдельного пикселя для того, чтобы выяснить, принадлежит ли он к интересующим объектам или нет. Результатом сегментирования изображения является бинарное изображение, где выделена область, которая в соответствии с критериями сегментации обладает признаками объектов, и область, которая обладает признаками фона. Поскольку объекты распознавания находятся на неоднородном фоне и имеют разные оттенки цвета, то методы выделения границ не смогут дать хороший результат. Распознаваемые объекты, как правило, контрастируют на зеленом фоне. В связи с этим был выбран пороговый метод сегментации. В качестве порога используется отношение спектральной яркости одной составляющей цвета к другой. Для отбора спектральной пары был проделан эксперимент, в котором использовались участки изображений объектов и фона, полученных из аэрофотоснимков. При проведении эксперимента были взяты фрагменты всех имеющихся в наличии типов фотографий. Другими словами, из множества снимков были перенесены объекты на одно изображение. На другое изображение были помещены фоновые цвета снимков. После этого оба изображения подверглись анализу: каждому пикселю изображения объектов и фона была поставлена в соответствие точка на координатной плоскости. Координатами точки являются значения яркости составляющих цвета. В результате получены три

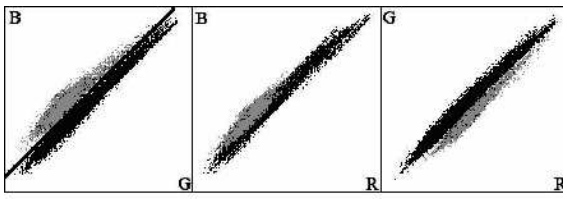


Рис. 1. Цветовые зависимости фона (черные точки) и объектов распознавания (серые точки). Слева — для зеленой и голубой компонентов цвета; по центру — для красной и голубой, справа — для красной и зеленой

графика скоплений точек, соответствующих различным спектральным парам (рис. 1).

Черные точки на графиках соответствуют пикселям фона, серые — пикселям объектов. Как видно из рисунка, скопления точек на левом графике (сочетание зеленого и голубого цвета) не перекрывают друг друга в отличие от двух других. Именно эта спектральная пара была использована нами для отделения объектов от фона. Коэффициенты уравнения прямой были рассчитаны из условия минимума суммы точек, попадающих в чужую область. При этом вероятность ошибочного отнесения области изображения фона к объекту по эталонным изображениям составило около 2%, тогда как вероятность ошибочного отнесения области изображения объекта к фону — 0,001%. Результат сегментации изображения пороговым методом представлен на рис. 2. Выбранный метод устойчив к сложным формам объектов, которые возникают в результате наложения изображений единичных объектов друг на друга.

Были опробован также метод сегментации SWA (Segmentation by Weighted Aggregation), основанный на теории графов [3]. В качестве параметра веса ребра графа была взята разница пикселей в цветовом пространстве. Результаты работы метода представлены на рис. 3.

Для оценки погрешности системы сегментации был проведен эксперимент на автоматически сгенерированных изображениях стад [4]. В результате проведения этого теста на 91 изображениях стад было установлено, что погрешность подсчета с применением описанной системы составляет около 8%. При вычитании из данной погрешности доли ошибок, связанных с наложением объектов друг на друга, погрешность составляет 3%.

Третий этап необходим для удаления помех, возникающих при сегментации. Сглаживание выполняется с помощью набора операций математической морфологии — операции эрозии и масштабного преобразования [5]. Такой подход помимо сглаживания удаляет мелкие помехи, которые, как правило, присутствуют в большом количестве после проведения сегментации пороговыми методами.

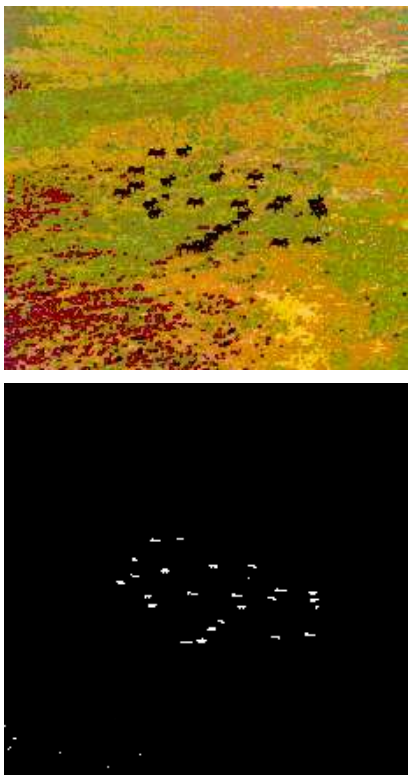


Рис. 2. Результаты сегментации изображения пороговым методом. Вверху исходное изображение, внизу — сегментированное

Входными данными для распознавания объектов являются изображения, полученные в результате процессов сегментации и шумоподавления. Для распознавания северных оленей на аэрофотоснимках были выбраны признаковые методы. Эти методы позволяют решить поставленную задачу в условиях, когда животные на снимках находятся в разных позах, имеют разный размер, цвет, находятся под разным освещением. Другими словами, объекты имеют множество эталонов, определить каждый из которых не представляется возможным, что является причиной отказа от корреляционных методов.

Было выделено 3 класса объектов: одиночные животные, животные, перекрывающие друг друга, и прочие объекты. В качестве признаков были выбраны следующие параметры: форма сегментированной области, ее площадь, вытянутость. Вытянутость области определяется двумя параметрами: протяженность области по оси X, и протяженность по оси Y. О форме области можно судить по такому параметру, как округлость, которая определяется следующим соотношением:

$$c = \frac{p^2}{S},$$

где p — периметр области, S — площадь области, c — округлость области.

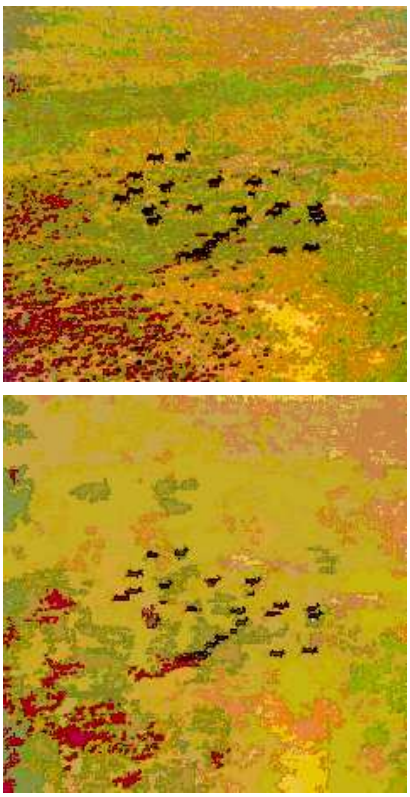


Рис. 3. Результаты применения алгоритма SWA. Вверху — до обработки, внизу — после обработки

Затем были определены пороговые параметры для каждого класса статистическим методом. Некоторые параметры, такие как округлость сегментированной области, имеют постоянное пороговое значение на всех изображениях, другие, такие как площадь, требуют адаптивного подхода на каждом изображении в связи с различным масштабом объектов. Для объекта, находящего в отдалении от других, параметр округлости лежит в диапазоне от 14 до 35. Округлость области сегментированного изображения, соответствующая скоплению животных, лежит в диапазоне от 35 до 300. Области, имеющие значение округлости более 300, как правило, являются крупными помехами.

Анализ методов сегментации и распознавания объектов подтвердил принципиальную возможность автоматического распознавания и подсчета диких северных оленей на фоне летней тундры по реальным аэрофотоснимкам. Однако выявились и недостатки методов, требующие доработки. При смене цветового баланса снимка результаты сегментации могут оказаться неудовлетворительными. Кроме того, темные участки на снимке (овраги, ущелья и т. п.) при выбранном методе сегментации идентифицируются как объекты, их надо выявлять на этапе шумоподавления и фильтрации.

Литература

[1] *Ерош И.Л., Сергеев М.Б., Соловьев Н.В.* Обработка и распознавание изображений в системах превентивной безопасности: Учеб. пособие. — СПб.: СПбГУАП, 2005. 154 с.

[2] *Гонсалес Р., Вудс Р.* Цифровая обработка изображений. — М: Техносфера, 2005. 1072 с.

[3] http://cgm.computergraphics.ru/content/view/147.

[4] *Михайлов В.В., Карташев Н.К* DEER COUNTER — программа–тренажер для выработки навыка визуальной оценки количества животных в группировке //Биологические ресурсы Крайнего Севера; перспективы охраны и рационального использования. — СПб.: РИЦ ГУАП, 2010. С. 205–212.

[5] *Яне Б.* Цифровая обработка изображений. — М.: Техносфера, 2007. 584 с.

Анализ формальных понятий и сжатие текстовой информации в задаче автоматизированного контроля знаний*

Емельянов Г. М., Михайлов Д. В.
Dmitry.Mikhaylov@novsu.ru
г. Великий Новгород, ГОУ ВПО «Новгородский государственный университет имени Ярослава Мудрого»

В данной статье показывается, как можно применять методы анализа формальных понятий для оптимальной организации тестов открытой формы в системах контроля знаний. Рассматривается минимизация базы знаний, описываемых на естественном языке разработчиком теста, путем выделения смысловых эталонов.

Тестовое задание открытой формы [5] в системе контроля знаний предполагает ответ обучаемого в виде одного или нескольких предложений Естественного Языка (ЕЯ).

Как было показано нами в [4], оценка близости ответа обучаемого заданному «правильному» ответу предполагает привлечение тезауруса, формируемого на основе множеств вариантов правильных ответов по совокупности тестов заданной тематики. При этом *актуальна задача* отбора самих форм ЕЯ-описаний каждого отдельного факта предметной области для представления в тезаурусе. В настоящей работе рассматривается решение указанной задачи расширением введенной нами ранее модели Ситуации Языкового Употребления (СЯУ).

Интерпретация ответа обучаемого

Разработчик теста описывает отдельный факт некоторой предметной области множеством Семантически Эквивалентных (СЭ) ЕЯ-фраз, которые определяют СЯУ. В [4] нами использовалась модель СЯУ в виде Формального Контекста (ФК, [2]):

$$K = (G, M, I)\,, \tag{1}$$

рассматриваемого в качестве информационной единицы тезауруса. Здесь множество объектов G составляют основы слов, синтаксически подчиненных другим словам из СЭ-фраз, задающих СЯУ. Множество признаков M включает в себя подмножества, обозначаемые далее посредством M с соответствующим нижним индексом и содержащие:

- указания на основу синтаксически главного слова (M_1);
- указания на флексию главного слова (M_2);
- связи «основа–флексия» для синтаксически главного слова (M_3);
- сочетания флексий зависимого и главного слова (M_4). При этом после флексии главного слова через двоеточие указывается предлог (если такой имеется) для связи главного слова с зависимым;
- указания на флексию зависимого слова (M_5).

Рассматривая в [4] совокупность СЯУ для известных фактов заданной предметной области как основу формирования тезауруса, авторами не накладывались какие-либо ограничения на исходное множество ЕЯ-фраз. Тем не менее, при использовании модели вида (1) в качестве единицы тезауруса ЕЯ-фразы, составляющие её основу, должны максимально точно описывать ситуацию (выражать смысл «на одном дыхании»). Ставится *задача* разделения знаний о сходных языковых формах описания различных ситуаций действительности (с одной стороны) и о внешне различающихся формах наиболее «компактного» описания каждой из ситуаций в тезаурусе (с другой стороны).

Для решения данной задачи рассмотрим единицу знаний, представляемую моделью (1) и сформированную на основе ЕЯ-фраз, отвечающих вышеуказанному требованию, в качестве смыслового эталона СЯУ. При этом введенная ранее модель СЯУ трансформируется к виду:

$$S = (T, K)\,, \tag{2}$$

где K есть ФК вида (1) для эталона, а множество T состоит из последовательностей пар (b_i, f_i), в которых b_i соответствует основе отдельного слова в составе ЕЯ-фразы, f_i — флексии этого слова. Введем обозначения для используемых далее символьных констант: p_{fl} — для «флексия:», p_{bs} — для «главное–основа:», p_{bf} — для «главное–флексия:», а для операции конкатенации — символ $\odot$.

Множество T в составе структуры (2) представляет возможные формы языкового описания заданного факта действительности. В число этих форм входят как ЕЯ-фразы, определяющие эталон СЯУ, так и не являющиеся таковыми. Для связи последних с эталоном поставим в соответствие некоторую переменную x_i каждой основе b_i, для которой существует либо признак $m \in M\colon m = p_{bs} \odot b_i$, либо объект $g \in G\colon g = b_i$. При этом на базе модели (2) строится шаблон СЯУ (верхний индекс P от англ. Pattern — шаблон):

$$S^P = (d^P, T^P, K^P)\,, \tag{3}$$

в котором все обозначения основ в составе имен объектов и признаков формального контекста эталона конкретной СЯУ заменяются переменными и отдельно задается список конкретизирующих четверок вида

$$\left(d^P, d^S, x_i, b_i\right), \tag{4}$$

Работа выполнена при финансовой поддержке РФФИ, проект № 10-01-00146.

где d^S — идентификационный номер СЯУ, d^P — номер её шаблона.

Заметим, что в значительном числе случаев тестирования интерпретация ответа обучаемого состоит в попытке применить шаблон (3) «правильного» ответа, сформулированного разработчиком теста. При этом не требуется производить разбор ЕЯ-ответа, обучаемого с привлечением внешних программ синтаксического анализа, а сама интерпретация происходит за линейное время, пропорциональное $\left|T^P\right|$.

Формирование смыслового эталона

Компоненты K^P в составе шаблонов (3) могут быть использованы для синтаксического разбора ЕЯ-фраз из ответа обучаемого. В ходе разбора строится формальный контекст вида (1) относительно некоторой фиксированной и смежных с ней предметных областей. Наличие структур-конкретизаций (4) по каждой анализируемой ЕЯ-фразе при этом не является обязательным.

Рассмотрим теперь задачу построения формального контекста самого смыслового эталона как основы моделей (2) и (3) по совокупности ФК отдельных СЭ-фраз, задающих СЯУ. Положим, что ФК указанной совокупности, далее упоминаемой как список K^{SE}, строятся по результатам синтаксического анализа этих фраз программой «Cognitive Dwarf» (ООО «Когнитивные технологии», [1]), которая использовалась нами в [4].

Для решения поставленной задачи введем коэффициенты сжатия информации относительно формальных контекстов вида (1).

Коэффициент сжатия информации по основам для формального контекста указанного вида равен:

$$k^S = \frac{\sum\limits_{i=1}^{n^{BS}} k_i^S}{n^{BS}}, \tag{5}$$

где

$$k_i^S = \frac{\sum\limits_{j=1}^{n_i^{BS}} \sum\limits_{k=1}^{n^{MF}} n_{ijk}^{AS}}{n_i^{BS}}, \quad n^{BS} = \left|M_1\right|,\ n^{MF} = \left|M_2\right|;$$

$$n_i^{BS} = \left|\left\{\ g \in G\colon I(g,m) = \text{true},\ m \in M_1,\ m = p_{bs} \odot b_i\ \right\}\right|;$$

$$n_{ijk}^{AS} = \left|\left\{\ m_k \in M_3\colon I(g,m_k) = \text{true},\ \exists\, m_{bf} \in M_2\colon m_{bf} = p_{bf} \odot f_k,\ m_k = b_i \odot f_k\ \right\}\right|.$$

Аналогично определяется коэффициент сжатия информации по флексиям:

$$k^F = \frac{\sum\limits_{i=1}^{n^{FS}} k_i^F}{n^{FS}}, \tag{6}$$

где

$$k_i^F = \frac{\sum\limits_{j=1}^{n_i^{FS}} \sum\limits_{k=1}^{n^{MF}} n_{ijk}^{AF}}{n_i^{FS}}, \qquad n^{FS} = \left|M_5\right|;$$

$$n_i^{FS} = \left|\left\{\ g \in G\colon I(g,m) = \text{true},\ m \in M_5,\ m = p_{fl} \odot f_i\ \right\}\right|;$$

$$n_{ijk}^{AF} = \left|\left\{\ m \in M_4\colon I(g_j,m) = \text{true},\ \exists\, m_{bf} \in M_2\colon m_{bf} = p_{bf} \odot f_k,\ m = f_i \odot f_k\ \right\}\right|.$$

Пусть смысловые эталоны для предметно-языковых знаний эксперта фиксируются в тезаурусе, представляемом формальным контекстом:

$$K^H = (G^H, M^H, I^H); \tag{7}$$

где множество G^H состоит из символьных пометок отдельных СЯУ. Множество M^H содержит элементы множеств признаков ФК вида (1) всех $g^H \in G^H$. Кроме того, в составе M^H выделяются:

- множество указания на основы слов, синтаксически подчиненных другим словам в ЕЯ-описаниях ситуаций $g^H \in G^H$;
- множество связей «основа–флексия» для синтаксически зависимого слова;
- множество сочетаний основ зависимого и главного слова.

Отношение $I^H \subseteq G^H \times M^H$, как и $I \subseteq G \times M$ для формального контекста (1), ставит в соответствие объектам их признаки.

Положим список K^{SE} отсортированным в порядке убывания мощностей множеств объектов для входящих в него ФК. Тогда построение ФК K^E вида (1) для смыслового эталона задаётся двумя нижеприведенными алгоритмами.

Алгоритм 1. Выделение потенциальных эталонов

Вход: K^{SE};
Выход: $P^E = \left\{K^{PE}\colon K^{PE} - \text{ФК вида (1)}\right\}$;

1: взять очередной $K = (G, M, I)$ из K^{SE};
2: $N_{max}^G := |G|$;
3: $P^E := \varnothing$;
4: **для всех** $K \in K^{SE}$ таких, что $|G| = N_{max}^G$
5: $\quad K_{cur}^{SE} := K^{SE} \setminus K$;
6: $\quad K^{PE} := K$;
7: $\quad k_{max}^S := k^S\left(K^{PE}\right)$ согласно формуле (5);
8: $\quad k_{max}^F := k^F\left(K^{PE}\right)$ согласно формуле (6);
9: $\quad$ **цикл**
10: $\quad\quad$ взять очередной $K = (G, M, I)$ из K_{cur}^{SE};
11: $\quad\quad K_{cur}^{PE} := K^{PE} \cup K$;
12: $\quad\quad k_{cur}^S := k^S\left(K_{cur}^{PE}\right)$;
13: $\quad\quad k_{cur}^F := k^F\left(K_{cur}^{PE}\right)$;
14: $\quad\quad Flag := \left(\left(k_{cur}^S > k_{max}^S\right) \wedge \left(k_{cur}^F > k_{max}^F\right)\right)$;
15: $\quad\quad$ при $Flag =$ false **выход**;
16: $\quad\quad k_{max}^S := k_{cur}^S$;
17: $\quad\quad k_{max}^F := k_{cur}^F$;
18: $\quad\quad K^{PE} := K_{cur}^{PE}$;
19: $\quad P^E := P^E \cup \left\{K^{PE}\right\}$;

Замечание 1. Применительно к паре произвольных формальных контекстов $K^X = \left(G^X, M^X, I^X\right)$ и $K^Y = \left(G^Y, M^Y, I^Y\right)$ теоретико-множественная операция $K^X \cup K^Y$ понимается как построение ФК $K^U = \left(G^X \cup G^Y, M^X \cup M^Y, I^X \cup I^Y\right)$.

Для описания следующего алгоритма необходимо ввести ряд дополнительных обозначений и соглашений. Пусть *CheckAndDel* есть функция удаления из состава множества объектов каждого формального контекста в списке P^E тех объектов, которые встречаются не во всех ФК данного списка. Те признаки, которые при этом становятся не принадлежащими ни одному объекту, удаляютя из множества признаков отдельного ФК функцией, обозначаемой далее Pck.

Признак будет включен в множество признаков формального контекста эталона, если он входит в состав пятерки признаков $\{m_1, m_2, m_3, m_4, m_5\}$, в которой $m_1 = p_{bs} \odot b$, $m_2 = p_{bf} \odot f_1$, $m_3 = b \odot f_1$, $m_4 = p_{fl} \odot f_2$, $m_5 = f_2 \odot f_1$, а b — основа некоторого слова. При этом основе b не должен соответствовать объект ФК, если есть другой объект этого же ФК, который обладает одновременно признаком m_1 и некоторым другим признаком $m = p_{bs} \odot b_1$, где $b_1 \neq b$, а основе b_1 не соответствует ни одного объекта этого ФК при том, что признак m относится более чем к одному объекту.

Функции, которая удаляет из признакового набора каждого объекта формального контекста признаки, не отвечающие данному условию, дадим имя *Closure*. Содержательно данная функция удаляет признаки главных слов-причастий в составе оборотов. Кроме того, указанная функция проверяет принадлежность каждого признака формируемого ФК $K^E = \left(G^E, M^E, I^E\right)$ множеству признаков, которые задают последовательности соподчиненных слов по следующему принципу:

$$\begin{cases} \exists\, m_1 \in M_1^E : \left((m_1 = p_{bs} \odot b) \wedge I^E\left(g, m_1\right)\right) = \text{true}; \\ \exists\, m_2 \in M_2^E : \left((m_2 = p_{bf} \odot f) \wedge I^E\left(g, m_2\right)\right) = \text{true}; \\ \exists\, m_3 \in M_3^E : \left((m_3 = b \odot f) \wedge I^E\left(g, m_3\right)\right) = \text{true}; \\ \exists\, m_5 \in M_5^E : \left((m_5 = p_{fl} \odot f) \wedge I^E\left(b, m_5\right)\right) = \text{true} \end{cases}$$

при максимально возможной длине каждой из последовательностей.

Замечание 2. Последовательности из трех и более соподчиненных слов, встречающиеся в 50 и более процентах СЭ-фраз из определяющих заданную СЯУ, выделяются предварительно на этапе синтаксического разбора и не представлены объектами и признаками формальных контекстов из списка K^{SE} на входе *Алгоритма 1*. Для каждой такой последовательности строится свой формальный контекст вида (1), который будет объединен с формальным контекстом эталона (множество таких ФК обозначим далее как P^{SQ}). Данный шаг предпринят в целях предотвращения нежелательного занижения коэффициентов (5) и (6) при выполнении указанного алгоритма.

Будем использовать символ $Null$ для обозначения формального контекста с пустыми множествами объектов и признаков. Тогда окончательный алгоритм формирования смыслового эталона будет выглядеть следующим образом.

Алгоритм 2. Формирование смыслового эталона

Вход: P^E, P^{SQ};
Выход: K^E — ФК вида (1) для эталона;
1: $P_1^E := CheckAndDel\left(P^E\right)$;
2: $P_2^E := \left\{K_2^{PE} \colon K_2^{PE} = Pck\left(K_1^{PE}\right), K_1^{PE} \in P_1^E\right\}$;
3: $K_{tmp}^E := Null$;
4: **пока** $P_2^E \neq \varnothing$
5: взять очередной K_2^{PE} из P_2^E;
6: $K_{tmp}^E := K_{tmp}^E \cup K_2^{PE}$;
7: $P_2^E := P_2^E \setminus \left\{K_2^{PE}\right\}$;
8: $K^E := Closure\left(K_{tmp}^E\right)$;
9: **пока** $P^{SQ} \neq \varnothing$
10: взять очередной K^{SQ} из P^{SQ};
11: $K^E := K^E \cup K^{SQ}$;
12: $P^{SQ} := P^{SQ} \setminus \left\{K^{SQ}\right\}$;

Формируемый *Алгоритмом 2* смысловой эталон соответствует подмножеству максимально проективных ЕЯ-фраз исходного СЭ-множества, представляющих лучшие способы описания заданного факта действительности. Напомним, что ЕЯ-фразу следует считать проективной в содержательном смысле, если все стрелки выявленных синтаксических связей могут быть проведены без пересечений по одну сторону прямой, на которой записана эта фраза. Кроме того, если из позиции некоторого слова выходят несколько стрелок, то эту позицию не должны накрывать стрелки, выходящие из позиций других слов. Говоря о максимальной проективности, здесь мы подразумеваем минимальную суммарную длину синтаксических связей внутри ЕЯ-фразы, не превышающую длины её самой [3].

Экспериментальная апробация

Предложенные методы формирования смыслового эталона и интерпретации ответа обучаемого были апробированы на материале ЕЯ-описаний фактов предметной области «Математические методы обучения по прецедентам». Часть указанного материала использовалась нами в [3] и [4]. При этом число СЭ-фраз, задающих СЯУ, выбиралось экспериментально с целью максиального приближения к реальной ситуации разработки теста. Данный показатель представлен в табл. 1 параметром N_1, его значение варьировалось в пределах от 2 до 54 в зависимости от описываемого факта. Для сравнения

Таблица 1. Смысловые эталоны

i	1	2	3	4	5	6
$N_1(i)$	54	53	26	26	2	3
$N_2(i)$	14	15	5	11	2	3
$N_3(i)$	13	15	13	12	8	11
$N_4(i)$	160	153	135	102	46	68
$N_5(i)$	9	12	12	12	8	11
$N_6(i)$	75	78	65	71	46	68

в этой же таблице приведены значения числа фраз, представляющих эталон (N_2), исходного числа объектов (N_3) и признаков СЯУ (N_4), числа объектов (N_5) и признаков эталона (N_6). Индекс i здесь есть порядковый номер СЯУ, краткое описание самих СЯУ даётся в табл. 2.

Таблица 2. Ситуации языкового употребления

i	Что описывает СЯУ
1	Связь переобучения с эмпирическим риском
2	Связь переусложнения модели с заниженностью средней ошибки на тренировочной выборке
3	Влияние переподгонки на частоту ошибок дерева принятия решений
4	Причина заниженности оценки обобщающей способности алгоритма
5	Зависимость оценки ошибки распознавания от выбора решающего правила
6	Зависимость обобщающей способности логического алгоритма классификации от числа закономерностей алгоритмической композиции

Качественной характеристикой процесса формирования смысловых эталонов в целом может послужить показанное на рис. 1 соотношение размеров тезауруса (7) при формировании его на основе формальных контекстов вида (1) всех СЭ-фраз каждой СЯУ (V_1) и на основе смысловых эталонов с применением предложенных в работе алгоритмов (V_2) при заданном числе СЯУ (N), $N = \left|G^H\right|$.

Модель (7) позволяет при вычислении функции *Closure* в составе *Алгоритма 2* дополненять формируемый эталон информацией слов-синонимов по сходству лексической и флективной сочетаемости на основе ранее сформированных эталонов, представленных в тезаурусе. Именно так были выделены синонимы *«переобучение»–«переподгонка»* для СЯУ с номерами 1, 2 и 4 в табл. 2.

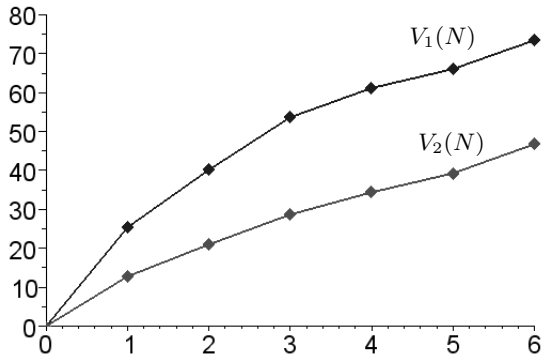


Рис. 1. Размер тезауруса для разного числа СЯУ, Кб

Заключение

Основной *результат* настоящей работы — *метод минимизации базы знаний* для вычисления рассмотренной в [4] количественной оценки схожести СЯУ при их независимом порождении. Применение предложенного метода позволяет уменьшить размер используемой базы примерно на 44%.

Особенностью представленного метода является построение модели смыслового эталона по результатам разбора исходных СЭ-фраз внешней программой синтаксического анализа. Значимыми моментами здесь являются высокая точность разбора (менее 2% ошибок) для случаев существенных смысловых ограничений на перифразирование, а также свободная распространяемость таких программ (включая исходные коды), что немаловажно при построении системы тестирования знаний.

Точность описанного метода может быть оценена средним числом невыделенных (опущенных) признаков на один объект формального контекста сформированного эталона. При этом за основу оценки может быть взят аналогичный ФК, но построенный с привлечением модели процесса выявления закономерностей сосуществования словоформ в линейном ряду, предложенной нами в [3].

Тема отдельного обсуждения — формирование единого смыслового эталона для нескольких СЯУ. Практически это означает доказать возможность их включения в один класс с формированием прецедента в виде модели (1) по наличию одного из случаев синонимии, рассмотренных в [4] и составляющих основу схожести между СЯУ. Само доказательство здесь ведется относительно эталонов отдельных СЯУ, предварительно сформированных описанными в настоящей работе методами.

Литература

[1] `http://cs.isa.ru:10000/dwarf` — 2011.

[2] *Ganter B., Wille B.* Formal Concept Analysis — Mathematical Foundations. — Berlin: Springer-Verlag, 1999. — 284 с.

[3] *Михайлов Д. В., Емельянов Г. М.* Морфология и синтаксис в задаче семантической кластеризации // Всеросс. конф. ММРО-14, М.: Макс Пресс, 2009. — С. 563–566.

[4] *Михайлов Д. В., Емельянов Г. М.* Семантическая схожесть текстов в задаче автоматизированного контроля знаний // Межд. конф. ИОИ-2010, М.: Макс Пресс, 2010. — С. 516–519.

[5] *Челышкова М. Б.* Теория и практика конструирования педагогических тестов. Учебное пособие. — М.: Логос, 2002. — 431 с.

Дискретный подход при извлечении информации из текста с автоматическим построением правил (текстовых запросов)

Прокофьев П. А.
p_prok@mail.ru
г. Москва, Компания «ЛАН-Проект»

Методы извлечения информации из текстов, как правило, не дают экспертам четкого понимания того, какие факторы влияют на принятие решений при обработке информации. В данной работе предложен подход, основанный на построении дискретных (логических) процедур распознавания, правила которых описываются автоматически строящимися текстовыми запросами, понятными экспертам.

Задача извлечения информации из текста относится к разряду тех, для которых практически невозможно построить математическую модель в общепринятом смысле. Частным случаем задачи извлечения информации из текста является отнесение интервалов в тексте к одному или нескольким заранее определенным классам.

В работе [3] извлечение информации использовалось для разрешения неоднозначностей выделения географических объектов в текстах. Слова и словосочетания в текстах распределялись по классам: страна, область, район, город и др. Используемые в работе [3] методы строили модели, анализ которых для экспертов достаточно сложен.

Другой подход, основанный на написании правил экспертами вручную, показывает, что правила быстро разрастаются и становятся плохо понятными самим экспертам.

В работах [1, 4] используются методы автоматического построения логических правил на этапе настройки классификатора. Правила описываются на определенном языке и позволяют отвечать на вопрос об отнесении целого текста или его части к определенной тематике. Последнее обстоятельство не позволяет использовать эти методы в задаче извлечения информации в текстах.

В работе [2] предлагается метод формального описания правил извлечения информации. По правилам строятся наборы интервалов текстов. Подобный подход разрабатывается в данной работе.

В рамках данной работы предлагается формальное описание языка правил (текстовых запросов) и метод автоматического построения правил при настройке алгоритмов классификации, используемых при извлечении информации в тексте. Алгоритмы классификации строятся с помощью дискреционных процедур распознавания по прецедентам, описанных в [5, 6].

Модель текстов

Дадим ряд определений, которыми будем пользоваться в рамках нашей модели.

Определение 1. *Текстами будем называть конечные последовательности слов:* $\tau = (\tau_1, \ldots, \tau_L)$, *где* τ_i *— слова,* $\tau_i \in W$, $i \in \{1, \ldots, L\}$, W *— множество допустимых слов,* $L = L(\tau)$ *— длина текста. Обозначим множество всех текстов* T.

Определение 2. *Интервалом текста* τ *называется совокупность трех объектов:* τ, i, j, *для которых выполняется условие* $1 \leqslant i \leqslant j \leqslant L(\tau)$. *Интервал текста будем обозначать* $\tau[i, j]$, *где* i *— начало интервала,* j *— конец интервала. Обозначим* $I(\tau)$ *— множество всех интервалов текста* τ, I *— множество всех интервалов текстов.*

Текстовые запросы

Определение 3. *Текстовым запросом будем называть отображение* q, *ставящее в соответствие любому тексту* τ *конечный набор интервалов этого текста:*

$$q(\tau) = \{\tau[i_1, j_1], \ldots, \tau[i_c, j_c]\} \subset I(\tau).$$

Обозначим множество текстовых запросов как Q.

Пример 1. Приведем примеры текстовых запросов:

1) запрос $q^*(\tau) = I(\tau)$ возвращает все интервалы текста;
2) запрос $q^{\omega}(\tau) = \big\{\tau[i, i] \big| \tau_i = \omega\big\}$ возвращает интервалы из одного слова, равного $\omega \in W$;
3) запрос $q^{(1)}(\tau) = \big\{\tau[i, i] \big| 1 \leqslant i \leqslant L(\tau)\big\}$ возвращает все интервалы из одного слова;
4) запрос $q^{(g,\alpha)}(\tau) = \{\tau[i, j] \mid g(\tau_i, \ldots, \tau_j) = \alpha\}$ возвращает интервалы, последовательность слов в которых отображается с помощью функции $g : T \to A$ в фиксированное значение $\alpha \in A$.

Чтобы задание запросов сделать конструктивным, рассмотрим ряд операций, позволяющих получать новые запросы из существующих:

1) конъюнкция:

$$(q_1 \wedge q_2)(\tau) = q_1(\tau) \cap q_2(\tau);$$

2) дизъюнкция:

$$(q_1 \vee q_2)(\tau) = q_1(\tau) \cup q_2(\tau);$$

3) отрицание:

$$(\neg q)(\tau) = q^*(\tau) \backslash q(\tau);$$

4) последовательность:

$$(q_1 \to q_2)(\tau) = \\ = \left\{\tau[i,j] \middle| \exists k:\ \tau[i,k] \in q_1(\tau), \tau[k+1,j] \in q_2(\tau)\right\},$$

возвращает интервалы, начальная часть которых принадлежит результату первого запроса, а конечная — результату второго;

5) внешнее включение:

$$(\uparrow q)(\tau) = \left\{\tau[i,j] \middle| \exists \tau[i_1,j_1] \in q(\tau), i \leqslant i_1, j_1 \leqslant j\right\},$$

возвращает интервалы, каждый из которых окружает некоторый интервал из результата запроса q;

6) внутреннее содержание:

$$(\downarrow q)(\tau) = \left\{\tau[i,j] \middle| \exists \tau[i_1,j_1] \in q(\tau), i_1 \leqslant i, j \leqslant j_1\right\},$$

возвращает интервалы, каждый из которых лежит внутри некоторого интервала из результата запроса q;

7) внешняя дизъюнкция:

$$q_1 \vee_\uparrow q_2 = (\uparrow q_1) \vee (\uparrow q_2);$$

8) внешняя конъюнкция:

$$q_1 \wedge_\uparrow q_2 = (\uparrow q_1) \wedge (\uparrow q_2).$$

Замечание 1. Набор этих операторов вместе с базовыми запросами, приведенными выше в качестве примеров, является избыточным. Например, $q_1 \wedge q_2 = \neg((\neg q_1) \vee (\neg q_2))$, $\uparrow q = q \vee (q^* \to \to q) \vee (q \to q^*) \vee (q^* \to q \to q^*)$. Однако конструировать и представлять запросы, пользуясь этим или, возможно, более расширенным набором операторов, значительно удобнее экспертам.

Признаки интервалов текстов

Определение 4. *Признаком объекта называется любое отображение, ставящее в соответствие объекту S определенное значение $\alpha = f(S)$. Множество всех возможных значений признака f будем называть доменом признака и обозначать $D(f)$.*

Если на множестве I задана система признаков $\bar{f} = (f_1, \ldots, f_n)$, то любому S из I соответствует вектор значений признаков $\bar{f}(S) = \\ = (f_1(S), \ldots, f_n(S))$, принадлежащий $D(f_1) \times \ldots \\ \ldots \times D(f_n)$.

Рассмотрим два способа конструирования признаков интервалов.

1. Признаки *словарного* типа:

$$f'_{(g,u,v)}(\tau[i,j]) = \begin{cases} g(\tau_r, \ldots, \tau_s), & \text{если } r \leqslant s; \\ g(\Lambda), & \text{иначе,} \end{cases}$$

где $r = \max\{1, i{+}u\}, s = \min\{i{+}v, L(\tau)\}$, Λ — специально зарезервированный символ для обозначения подпоследовательности текста нулевой длины. Признак возвращает значение отображения g последовательности слов, индексы которых заданы относительно начала интервала: от $i{+}u$ до $i{+}v$.

2. Признаки *запросного* типа:

$$f''_{(q,u,v)}(\tau[i,j]) = \begin{cases} 1, & \text{если } \tau[r,s] \in q(\tau); \\ 0, & \text{иначе,} \end{cases}$$

где $r = \max\{1, i{+}u\}, s = \min\{i{+}v, L(\tau)\}$. Признак возвращает значение 1, если интервал $\tau[r,s]$, заданный относительно начала контекста, удовлетворяет запросу q, и возвращает 0 в противном случае.

Дискретные (логические) процедуры распознавания

Пусть известно, что множество I представимо в виде объединения непересекающихся классов $K_1, K_2, \ldots, K_m$. Имеется конечный набор интервалов $\mathfrak{S} = \{S_1, S_2, \ldots, S_t\} \subset I$, для которых известна их принадлежность к классам (обучающая выборка или набор размеченных текстов). Требуется для произвольного интервала определить класс, к которому он принадлежит.

Подход в данной работе может быть использован на базе нескольких схем построения дискретных процедур распознавания по элементарным классификаторам: голосование по представительным наборам, голосование по покрытиям класса и голосование по антипредставительным наборам. Эти схемы подробно описаны в работах [5, 6]. В каждой схеме алгоритм распознавания описывается набором элементарных классификаторов и функцией вычисления оценки.

Наибольшую сложность представляет нахождение требуемых элементарных классификаторов для классов. В работе [5] описаны способы нахождения элементарных классификаторов как импликантов частично определенной логической функции, переменные которой заданы на множестве значений признаков объектов. Импликанты получаются при нахождении сокращенной ДНФ доопределенной функции. Сокращенная ДНФ находится методами, приведенными в книге [7].

Рассмотрим процедуру распознавания на примере схемы голосования по представительным наборам. Пусть для класса K частично определена на $\mathfrak{S}$ функция

$$u^{(K,\mathfrak{S})}(S) = \begin{cases} 1, & \text{если } S \in \mathfrak{S} \cap K; \\ 0, & \text{если } S \in \mathfrak{S} \backslash K. \end{cases}$$

Доопределяя функцию $u^{(K,\mathfrak{S})}$ на множество I, получаем функцию

$$U^{(K,\mathfrak{S})}(S) = \begin{cases} 0, & \text{если } S \in \mathfrak{S} \backslash K; \\ 1, & \text{в остальных случаях.} \end{cases}$$

Если $\mathfrak{S} \backslash K = \{S_{i_1}, \ldots, S_{i_m}\}$, то

$$U^{(K,\mathfrak{S})}(S) = \delta_1(S) \wedge \ldots \wedge \delta_m(S),$$

где $\delta_r(S) = (f_1(S) \neq \alpha_{r1}) \vee \ldots \vee (f_n(S) \neq \alpha_{rn})$ — конъюнкции, соответствующие интервалу S_{i_r}, и $\alpha_{rj} = f_j(S_{i_r})$ — значения признаков, $r = 1, \ldots, m, j = 1, \ldots, n$.

Прежде чем перейти к поиску импликант функции $U^{(K,\mathfrak{S})}(S)$, исследуем свойства предикатов на множестве I.

Определение 5. *Предикаты p_1 и p_2, значения которых совпадают на $M \subset I$, будем называть равносильными на M и обозначать $p_1 \overset{M}{\equiv} p_2$. Равносильность предикатов на I будем обозначать $p_1 \equiv p_2$*

Утверждение 1. *Если f — признак интервалов текста, тогда $(f(S) \neq \alpha) \equiv (f(S) = \beta_1) \vee \ldots \vee (f(S) = \beta_d)$, где $\{\beta_1, \ldots, \beta_d\} = D(f) \backslash \{\alpha\}$.*

Утверждение 2. *Пусть заданы: g — отображение на T; q, q_1, q_2 — текстовые запросы; $u, v, w \in \mathbb{Z}$. Справедливы следующие утверждения о равносильности предикатов:*

1) $(f'_{(g,u,v)}(S) \neq \alpha) \equiv (f''_{(q^{(g,\alpha)},u,v)}(S) = 0)$;
2) $(f'_{(g,u,v)}(S) = \alpha) \equiv (f''_{(q^{(g,\alpha)},u,v)}(S) = 1)$;
3) $(f''_{(q,u,v)}(S) = 0) \equiv (f''_{(\neg q,u,v)}(S) = 1)$;
4) $(f''_{(q_1,u,v)}(S) = 1) \vee (f''_{(q_2,u,v)}(S) = 1) \equiv$
   $\equiv (f''_{(q_1 \vee q_2,u,v)}(S) = 1)$;
5) $(f''_{(q_1,u,v)}(S) = 0) \vee (f''_{(q_2,u,v)}(S) = 0) \equiv$
   $\equiv (f''_{(q_1 \wedge q_2,u,v)}(S) = 0)$;
6) $(f''_{(q_1,u,v)}(S) = 1) \wedge (f''_{(q_2,u,v)}(S) = 1) \equiv$
   $\equiv (f''_{(q_1 \wedge q_2,u,v)}(S) = 1)$;
7) $(f''_{(q_1,u,v)}(S) = 0) \wedge (f''_{(q_2,u,v)}(S) = 0) \equiv$
   $\equiv (f''_{(q_1 \vee q_2,u,v)}(S) = 0)$;
8) $(f''_{(q_1,u,v)}(S) = 1) \wedge (f''_{(q_2,v+1,w)}(S) = 1) \equiv$
   $\equiv (f''_{(q_1 \to q_2,u,w)}(S) = 1)$;
9) $(f''_{(q_1,u,v)}(S) = 0) \vee (f''_{(q_2,v+1,w)}(S) = 0) \equiv$
   $\equiv (f''_{(q_1 \to q_2,u,w)}(S) = 0)$;
10) $(f''_{(q,u,v)}(S) = 1) \equiv (f''_{(q \to q^{(1)},u,v+1)}(S) = 1) \equiv$
    $\equiv (f''_{(q^{(1)} \to q,u-1,v)}(S) = 1)$.

Замечание 2. Утверждение 1 и пункт 1) утверждения 2 обосновывают замену предикатов $(f(S) \neq \alpha)$ на дизъюнкции простых импликант. Для реализации этой замены, возможно, понадобится добавить новые запросные признаки.

Сокращенная ДНФ для $U^{(K,\mathfrak{S})}(S)$ получается после исключения всех предикатов вида $(f(S) \neq \alpha)$, раскрытия скобок и выполнения правил поглощения по описанному в [7] алгоритму. Затем из сокращенной ДНФ удаляются все элементарные конъюнкции, не удовлетворяющие ни одному $S \in \mathfrak{S} \cap K$.

Замечание 3. Пункты 2)–9) утверждения 2 обосновывают преобразования импликант ДНФ с уменьшением их длины. Равносильность 10) позволяет выполнять «выравнивание границ» запросных признаков и применять равносильности 4)–9). При этом, возможно, также потребуется добавление новых признаков.

Следствие 1. *Для любого класса $K \subset I$ существует конечное число текстовых запросов $q_1, \ldots, q_s$ и целых чисел $u_1, \ldots, v_s$; $v_1, \ldots, v_s$, таких, что равносильны:*

$$U^{(K,\mathfrak{S})}(S) \overset{\mathfrak{S}}{\equiv}$$
$$\overset{\mathfrak{S}}{\equiv} (f''_{(q_1,u_1,v_1)}(S) = 1) \vee \ldots \vee (f''_{(q_s,u_s,v_s)}(S) = 1).$$

Замечание 4. Следствие показывает, что каждый элементарный классификатор в процедуре распознавания может быть представлен в виде текстового запроса.

Замечание 5. Очевидно, запросы, о которых идет речь в следствии, строятся с применением равносильностей из утверждения 2. Однако запросы, построенные таким образом, содержат очень строгие условия.

Утверждение 3. *Пусть заданы: q, q_1, q_2 — текстовые запросы; целые числа u, v, u', v', u'', v'' : $u'' \leqslant u$, $v \leqslant v''$, $u'' \leqslant u'$, $v' \leqslant v''$. Справедливы следующие импликации:*

1) $(f''_{(q,u,v)}(S) = 1) \Rightarrow (f''_{(\uparrow q,u'',v'')} = 1)$;
2) $(f''_{(q_1,u,v)}(S) = 1) \wedge (f''_{(q_2,u',v')}(S) = 1) \Rightarrow$
   $\Rightarrow (f''_{(q_1 \wedge \uparrow q_2,u'',v'')}(S) = 1)$;
3) $(f''_{(q_1,u,v)}(S) = 1) \vee (f''_{(q_2,u',v')}(S) = 1) \Rightarrow$
   $\Rightarrow (f''_{(q_1 \vee \uparrow q_2,u'',v'')}(S) = 1)$.

Замечание 6. Преобразования $U^{(K,\mathfrak{S})}(S)$, осуществляемые заменой левой части верной импликации правой, не всегда приводят к равносильной на $\mathfrak{S}$ формуле. Однако, если включить признаки, построенные по правилам утверждения 3, в состав признаков и повторно вычислить сокращенную ДНФ для $U^{(K,\mathfrak{S})}(S)$, то можно добиться более короткой ДНФ.

Текстовые запросы, построенные с использованием утверждения 3, менее строгие, чем запросы в следствии к утверждению 2.

Утверждение 4. *Пусть функция* $U^{(K,\mathfrak{S})}$ *представляется на* $\mathfrak{S}$ *формулой вида*

$$U^{(K,\mathfrak{S})}(S) \overset{\mathfrak{S}}{\equiv} (h(S) \wedge h'(S)) \vee h''(S). \qquad (1)$$

Обозначим множество $H = \{S \in \mathfrak{S} \mid (h(S) = 1) \wedge (h''(S) = 0)\}$. *Пусть существуют запрос* q *и целые числа* u *и* v *такие, что выполняются свойства:*

1) $f''_{(q,u,v)}(S) = 1, \forall S \in H \cap K$;
2) $f''_{(q,u,v)}(S) = 0, \forall S \in H \backslash K$.

Тогда функция $U^{(K,\mathfrak{S})}$ *представляется на* $\mathfrak{S}$ *формулой:*

$$U^{(K,\mathfrak{S})}(S) \overset{\mathfrak{S}}{\equiv} (h(S) \wedge (f''_{(q,u,v)}(S) = 1)) \vee h''(S).$$

Замечание 7. Представление (1) может быть получено вынесением общей части нескольких импликант ДНФ за скобку:

$$U^{(K,\mathfrak{S})} \equiv \sigma_1 \vee \ldots \vee \sigma_l \vee (\sigma'_0 \wedge (\sigma'_1 \vee \ldots \vee \sigma'_k)),$$

где σ_i, σ'_j — элементарные конъюнкции, $i = 1, \ldots, l$; $j = 0, \ldots, k$.

Утверждение 4 позволяет упрощать ДНФ, уменьшая число импликантов. Интерес представляет конструирование признаков, удовлетворяющих условиям утверждения 4.

Заметим, что если запрос q удовлетворяет условиям утверждения 4, то $f''_{(q,u,v)}(S) = U^{(K,H)}(S)$ на H. Это обстоятельство позволяет рекуррентно строить такой запрос q.

Эксперименты

Предложенный подход тестируется на задаче, поставленной в работе [1]. Результы экспериментов будут изложены в докладе.

Выводы

В работе дается формальное описание конструкции текстовых запросов. Показывается принципиальная возможность представления элементарных классификаторов дискретных процедур распознавания в виде текстовых запросов.

Дальнейшие исследования будут направлены на следующие моменты:

1) расширение и строгое формальное описание языка текстовых запросов;
2) разработка правил формирования коротких запросов при построении элементарных классификаторов;
3) формальное описание алгоритма настройки предложенной процедуры распознавания и исследование его сложности;
4) использование описанных в [5, 8] методов поиска информативных фрагментов описаний объектов при построении процедуры распознавания для сокращения вычислительных затрат и улучшения качества распознавания;
5) разработка алгоритмов вычисления текстовых запросов и исследование алгоритмической сложности.

Литература

[1] *Junker M., Abecker A.* Learning Complex Patterns for Document Categorization // AAAI-98/ICML Workshop on Learning for Text Categorization. Madison, Wisconsin, USA, 1998.

[2] *Reiss F.,Raghavan S., Krishnamurthy R., Zhu H., Vaithyanathan S.* An Algebraic Approach to Rule-Based Information Extraction // ICDE. Cancun, Mexico, 2008.

[3] *Прокофьев П. А.* Использование методов извлечения информации при географической привязке текстов на русском языке // Электронные библиотеки: Перспективные методы и технологии, Электронные коллекции (RCDL). — Петрозаводск, 2009.

[4] *Агеев М. С.* Методы автоматической рубрикации текстов, основанные на машинном обучении и знаниях экспертов. — Диссертация на соискание ученой степени к.ф.-м.н. — М.: МГУ, 2004.

[5] *Дюкова Е. В., Песков Н. В.* Построение распознающих процедур на базе элементарных классификаторов // Математические вопросы кибернетики / Под ред. О. Б. Лупанова. — М.: Физматлит, 2005. — Т. 14.

[6] *Дюкова Е. В.* Дискретные (логические) процедуры распознавания: принципы конструирования, сложность реализации и основные модели. Учебное пособие для студентов математических факультетов педвузов. — М.: Изд-во «Прометей», 2003. — 29 с.

[7] Дискретная математика и математические вопросы кибернетики / Под ред. С. Б. Яблонского, О. Б. Лупанова. — М.: Наука, 1974. — 312 с.

[8] *Песков Н. В.* Поиск информативных фрагментов описаний объектов в задачах распознавания. — Диссертация на соискание ученой степени к.ф.-м.н., М.: ВЦ РАН. — 2004.

Формализация и автоматический анализ понятий при обработке неструктурированной информации*

Майсурадзе А. И.
maysuradze@cs.msu.su
Москва, Факультет ВМК МГУ имени М. В. Ломоносова

Приведены результаты исследования задачи выделения в тексте на русском языке основных и дополнительных понятий. Понятия выделяются в виде групп речевых маркеров, присутствующих в тексте. Предложена методика подготовки обучающего материала. Использованы классические методы индукции правил. Приведенные результаты позволяют сравнить стандартные методы с успехами современных закрытых промышленных решений.

В настоящее время на рынке программных средств интеллектуального анализа данных предлагается всё больше систем для анализа неструктурированной информации. В первую очередь речь идет об обработке текстов на естественном языке. Для английского и, в меньшей степени, других западноевропейских языков присутствуют решения с открытым исходным кодом или хотя бы с четко описанными алгоритмами обработки информации. Для русского языка практически все решения являются закрытыми, используемые алгоритмы или точные формальные описания обрабатываемых объектов не разглашаются. Разумеется, во многих случаях это препятствует полноценному рассмотрению методов обработки неструктурированной информации в рамках научной деятельности. Практически, остается лишь сравнивать результаты работы таких систем.

Публикаций, в которых представлены конкретные результаты работы систем, в России выходит не очень много. В качестве одного из регулярных источников информации укажем труды международной конференции по компьютерной лингвистике Диалог [1], в которых разрешается «предложить не только методы автоматического извлечения, но и продемонстрировать полученные в результате их применения конкретные лингвистические данные». Анализ таких публикаций показал, что читателю нередко трудно дать оценку качества достигнутых результатов. Чаще всего это происходит, когда авторы публикаций рассматривают только использование своего метода в своей задаче и не делают сравнения с результатами применения к задаче стандартных методов.

Данное исследование призвано «перераспределить» упомянутые достоинства и недостатки публикаций в рассматриваемой области. Будут даны явные формализации объектов обработки, которые не зависят от предполагаемых для использования алгоритмов. Свойства объектов будут назначены только экспертами. Будут использованы общедоступные методы машинного обучения. Содержательная задача хорошо понятна экспертам. Таким образом, одной из целей данного исследования было ввести в научную деятельность конкретные результаты работы стандартных методов, что даст опору для анализа качества современных специализированных методов.

Работа выполнена при финансовой поддержке РФФИ, проект № 10-01-00131.

Постановка содержательной задачи

В работе рассматривается задача выделения в тексте основных и дополнительных понятий, представленных своими речевыми маркерами[1].

Указанная содержательная задача может быть сформулирована как набор содержательных подзадач. Дан текст на русском языке.

— Требуется перечислить понятия, упомянутые в тексте. Под понятием можно понимать объект (в том числе, собирательный) предметной области, о котором в тексте делаются те или иные утверждения.
— Для каждого понятия требуется указать в тексте речевые маркеры. То есть надо указать в тексте те слова и словосочетания, которые это понятие обозначают. Вообще говоря, речевые маркеры для разных понятий могут входить друг в друга.
— Требуется выделенные понятия классифицировать на основные и дополнительные. Вообще говоря, такая классификация зависит от цели написания и чтения текста. В данном исследовании определение оставлялось на усмотрение экспертов.

План исследования

При планировании было предложено разбить исследование на следующие этапы.

1. Предварительная частичная формализация основных понятий необходима для того, чтобы четко поставить на следующем этапе вопросы экспертам. В тексте могут быть выделены

[1] Отметим, что будет использоваться терминология, принятая в социологии, дискурс-анализе, контент-анализе. Разработчики программных средств нередко пользуются различными другими терминами.

различные единицы: абзацы, предложения, словосочетания, отдельные слова и т. п. После предварительного общения с экспертами было решено, что из составляющих единиц текста будут использованы отдельные слова и словосочетания.

Определение 1. *Группа вхождений для заданного текста обозначает набор попарно непересекающихся слов и словосочетаний из текста с учетом их конкретного положения в тексте.*

Если по сути одно и то же словосочетание входит в текст несколько раз (в разных местах), то в данном исследовании будем считать, что это разные словосочетания. Соответственно, понятия будут выделяться в виде группы вхождений, которая и будет основным объектом обработки в предлагаемых процедурах.

2. Опрос экспертов требовал подготовки набора текстов для анализа, составления анкет, подборки экспертной группы, распределения текстов по экспертам, сбор индивидуальных результатов анкетирования.

3. Формирование обучающей информации заключалось в разработке схемы консолидации индивидуальных мнений экспертов в единую обучающую выборку. Использовалось две схемы консолидации.

4. Выбор и реализация метода представляют интерес из-за особенностей предметной области и выбора составного объекта обработки (группа вхождений).

5. Получение и анализ результатов основаны на различных функционалах качества, применимых к обрабатываемым данным.

Опрос, консолидация ответов

Подготовка корпуса текстов для анализа основывалась на следующих требованиях. Тексты должны быть самодостаточными, содержать целиком одну или несколько идей. Для экспертов удобнее тексты объемом около 1000 символов (примерно 140 слов, 9 предложений, 3 небольших абзаца). Таким совокупным требованиям лучше всего удовлетворяют тексты в жанре «журналистских заметок». Для корпуса была выбрана узкая тематика «Пожары в России летом 2010 года». Такой стиль исследования называется работой в одном дискурсе.

Подборка экспертной группы не относится к сильным сторонам данного исследования. Экспертами были студенты Факультета иностранных языков и регионоведения МГУ им. М. В. Ломоносова. Для них не удалось задать однозначно интерпретируемую всей группой цель чтения текста.

Составление анкет потребовало тщательно отшлифовать формулировки вопросов и используемые термины. Стало понятно, что использование одних и тех же анкет для разных групп экспертов будет затруднительным. Вопросы формулировались в соответствии с содержательными подзадачами: перечислить понятия, указать для каждого понятия группу вхождений, указать среди понятий основные.

Распределение текстов по экспертам потребовалось из-за того, что даже с небольшими текстами эксперты работают медленно. Вместо ожидаемого ускорения обработки каждого последующего текста наблюдалось замедление. В среднем эксперт за час обрабатывал 10 текстов. 100 текстов были распределены по 20 экспертам так, что каждый текст был обработан 4 экспертами.

Сбор индивидуальных результатов анкетирования для каждого текста дал 4 набора групп вхождений, некоторые из которых были помечены как основные. Кроме того, в свободной форме эксперты дали комментарии, почему они выбрали именно такие группы.

Визуализация множества ответов всех экспертов для заданного текста проводится в виде специально разработанной диаграммы «картограмма текста». В настоящее время ещё не разработана аналогичная черно-белая диаграмма.

Консолидация ответов разных экспертов для каждого текста проводилась тремя способами. При первом способе все экспертные ответы просто конкатенировались. При этом предполагалось, что некоторые понятия могут совпадать. Второй способ заключался в отборе таких пересечений экспертных групп, в которых сохраняется большая часть каждой группы. При этом проблемы назначения атрибута «основная / дополнительная» не возникло. Третий способ позволял сохранить все речевые маркеры, утерянные во втором способе, путем назначения им различных весов.

Особенности метода

Было решено использовать стандартные методы машинного обучения, в основе которых лежит принцип индуктивного вывода логических закономерностей. Результат настройки таких информационных моделей пригоден для экспертной интерпретации. Базовый алгоритм, по которому получены нижеприведенные результаты, следует классической работе [2] и в данной работе не исследуется.

Объекты распознавания. Основная особенность принятой формализации состоит в том, что объектами распознавания являются составные сущности из сложной предметной области. Логические закономерности обязаны строиться из предикатов над группами вхождений. В литературе по-

ка не сложилось какого-либо классического набора предикатов над группами словосочетаний. Используемый набор предикатов определялся, по большей части, доступными технологиями.

Сначала для отдельных слов вычислялись морфологические (часть речи, нормальная форма и т. п.) и частотные (количество повторов, положение в тексте) показатели. Отметим, что набранные цифрами числа тоже анализировались. Потом для слов и словосочетаний устанавливалась синтаксическая функция (подлежащее, определение и т. п.). Далее анализировались пары слов из разных объектов одной группы на наличие синонимов и однокоренных слов. Все указанные характеристики вычислялись при помощи готовых общедоступных программ или библиотек.

На следующем этапе на основе полученной информации вычислялись признаки и предикаты на отдельных вхождениях в группу. В итоге по самостоятельным частям речи для каждого вхождения вычислялся вес. Далее вычислялись предикаты для пары вхождений в группу. В итоге для каждой группы вычислялось 11 числовых признаков и 4 предиката. Числовые признаки индуцировали предикат по порогу, который можно было зафиксировать или настроить по обучению.

Генерация групп. Нетрудно заметить, что в исходных данных содержательной задачи предложенных объектов распознавания нет. Следовательно, возникает дополнительная задача генерации допустимых групп. Очевидно, что генерировать для заданного текста всевозможные множества подстрок неэффективно.

В работе предложен эвристический алгоритм 1, который для заданного текста генерирует сравнительно небольшое множество групп. Указанный алгоритм генерирует все группы, присутствующие в собранном в ходе исследования обучении. Эксперты выделяли в тексте от 2 до 5 понятий, эвристический алгоритм давал несколько десятков групп.

Определение 2. *Смежными вхождениями будем называть разные вхождения в разные группы, касающиеся в тексте по границе.*

Этапы решения общей задачи предлагается формализовать следующим образом.

1. По заданному тексту сгенерировать множество групп-кандидатов. В дальнейшем никаких новых групп не появится. В настоящее время задача решается эвристическим алгоритмом.
2. Классифицировать заданные группы-кандидаты на заданном тексте по атрибуту «понятие / не понятие». Получаем стандартную двухклассовую задачу классификации. Эксперты дают положительные прецеденты, остальные сгенерированные группы-кандидаты становятся отрицательными прецедентами.
3. Классифицировать понятия, заданные в виде групп вхождений на тексте, по атрибуту «основное / дополнительное». Получаем стандартную двухклассовую задачу классификации. Эксперты дают положительные и отрицательные прецеденты.

Алгоритм 1. Эвристический алгоритм генерации групп вхождений.

Вход: текст на русском языке;
Выход: множество групп вхождений;
1: имена собственные, если возможно, привязать к определяемому ими существительному;
2: числительные и единицы измерения привязать к определяемым ими существительным, включая возможные промежуточные прилагательные;
3: **для всех** одиночные существительные, не состоящие в именном типе связи с прилагательным
4: поместить существительное в новую «именную» группу;
5: **для всех** сочетание существительного, являющегося подлежащим, с прилагательными
6: поместить сочетание в новую «именную» группу;
7: объединить в одну новую группу безличные сказуемые;
8: **для всех** причастные обороты
9: поместить оборот в новую группу;
10: **повторять**
11: объединять в новые группы именные группы для каждого корня;
12: объединять в новые группы именные группы для каждого синонима;
13: **пока** группы порождаются;
14: **повторять**
15: добавить конкатенации смежных вхождений в группах, порожденных на шаге 11 алгоритма;
16: **пока** группы меняются;

Видно, что предложенная последовательность получения результатов не совсем соответствует тому, в каком порядке людям удобнее давать ответ. Кроме того, люди обычно способны явно назвать выделенное ими из текста понятие. При этом используются слова, не содержащиеся в тексте. В настоящее время такая функциональность выделена в отдельную задачу, выходящую за рамки данного исследования.

Результаты

Поскольку формализация свела содержательную задачу к двум задачам классификации, то ре-

зультатом настройки будет пара наборов правил. Было исследовано две пары наборов правил. Одну пару «субъективных» наборов вручную построил эксперт. Вторая пара наборов была настроена автоматически.

В подобных задачах принято использовать два показателя качества решения: точность и полноту. Полнота показывает долю положительных объектов, найденных алгоритмом, среди всех положительных объектов. Точность показывает долю положительных объектов, найденных алгоритмом, среди всех найденных. Кроме того, с использованием составной природы объекта распознавания, был предложен показатель «минимаксное совпадение», который показывает, насколько полно была выявлена та из групп обучающей выборки, которая представлена в результатах работы алгоритма хуже всего. Показатели качества оцениваются отдельно для каждого текста.

$$\text{совпадение двух групп} := \\ := \frac{\text{число общих слов в группах}}{\text{общее число слов в группах}};$$

$$\text{минимаксное совпадение двух множеств групп} := \\ := \min_{\substack{\text{группа} \\ \text{обучения}}} \max_{\substack{\text{группа} \\ \text{результата}}} \text{совпадение двух групп}.$$

Эксперименты показали, что эксперт получил более сложные наборы правил (5 и 4), чем дала автоматическая настройка (4 и 3). Сложнее были и сами отдельные правила. Точность всегда была меньше полноты, что означает, что правила склонны выявлять лишние группы. Точность варьировалась от 50% до 75%. Полнота варьировалась от 60% до 80%. Минимакс от 43% до 65%. Субъективные наборы оставляли больше понятий, но меньше основных понятий, чем автоматические. Соответственно, точность субъективных наборов для понятий была ниже, а для основных понятий была выше.

В современных системах нормальными считаются показатели полноты и точности от 75% и выше.

Выводы

Была проведена формализация и операционализация актуальной задачи выделения в тексте понятий и соответствующих им речевых маркеров. Использовались стандартные методы машинного обучения и компьютерной лингвистики. Таким образом, в определенном смысле была намечена нижняя планка качества для подобных систем. Как оказалось, эта тривиальная нижняя граница качества лежит сравнительно высоко по сравнению с современными системами.

Возможно совершенствование методов генерации групп-кандидатов. Представляется, что методы генерации должны зависеть от используемых методов классификации и показателей качества.

Представляет самостоятельный интерес задача подбора наименования понятия, заданного набором словосочетаний. На английском языке подобные задачи решаются путём анализа графа связей в Wikipedia ([3]). Утверждается, что русскоязычная Википедия пока не достигла необходимого наполнения.

Литература

[1] `www.dialog-21.ru` — Диалог. Международная конференция по компьютерной лингвистике — 2011.

[2] *Вайнцвайг М. Н.* Алгоритм обучения распознаванию образов «Кора» // Алгоритмы обучения распознаванию образов, М.: Советское радио, 1973. — С. 110–116.

[3] *Korshunov A., Turdakov D., Jeong J., Lee M., Moon C.* A Category-Driven Approach to Deriving Domain Specific Subset of Wikipedia // Proceedings of SYRCoDIS'11: The Seventh Spring Researchers Colloquium on Databases and Information Systems — 2011. — pp. 43-53.

Инкрементное обучение деревьев решений в задаче распознавания структуры статистических таблиц*

Кудинов П. Ю., Полежаев В. А.
pkudinov@gmail.com, walter2kf@gmail.com
Москва, Вычислительный Центр им. А. А. Дородницына РАН, МГУ им. М. В. Ломоносова

Рассматриваются задачи распознавания логической структуры статистических таблиц, возникающие при создании системы поиска статистической информации по разнородным источникам. Предлагается новый корректный алгоритм динамического обучения, основанный на композиции случайных инкрементных деревьев. Описываются результаты экспериментального сравнения этого алгоритма с известным алгоритмом ITI (incremental tree induction) на поставленных задачах распознавания.

Одним из основных способов представления статистических данных является запись их в табличном виде. При этом разные источники, такие как Росстат, ВЦИОМ, OECD, банки и финансовые организации, могут использовать разные названия показателей и структуры таблиц для описания одних и тех же явлений. Для сокращения рутинной работы при поиске и анализе статистических данных создаётся система информационного поиска, основной функцией которой будет вывод агрегированных статистических таблиц по запросам пользователей. Первым шагом на пути реализации такой системы является разработка методов извлечения статистических показателей из таблиц [2, 3].

Составители статистических таблиц используют различные типы логической структуры таблиц и различные средства для разделения таблицы на ячейки и блоки. Постоянная доработка алгоритмов анализа структуры таблиц привела бы к неуправляемому росту их сложности. Поэтому предлагается использовать методы динамического машинного обучения (online learning). Предполагается, что эксперт просматривает каждую обработанную таблицу и исправляет ошибки распознавания. Эти исправления добавляются в обучающую выборку и происходит дообучение алгоритмов. Если ошибки распознавания редки, то динамический режим существенно снижает трудоёмкость обучения в сравнении с обычной практикой, когда обучающая выборка формируется заранее (offline learning).

Статистические таблицы

Рассмотрим квадратную сетку $G^{M\times N}$ из M строк и N столбцов и зададим её полное покрытие непересекающимися прямоугольными областями. Элемент покрытия будем называть *ячейкой*. Множество всех ячеек обозначим через C. Каждой ячейке $c \in C$ поставим в соответствие текстовую строку $text(c)$, возможно пустую, которую назовём *содержимым ячейки* или её *текстом*.

Будем полагать, что каждая ячейка c имеет тип $type(c) \in \{data, key, empty\}$. *Ячейка данных* ($data$) содержит ровно одно числовое значение. *Ячейка описания* (key) содержит текстовую строку, состоящую из словесных описаний одного или нескольких ключей. *Пустая ячейка* ($empty$) не содержит значимой информации, её содержимое игнорируется.

Статистической таблицей T будем называть четвёрку $\langle G^{M\times N}, C_V, C_K, R\rangle$, где C_V — множество ячеек данных, C_K — множество ячеек описаний, отображение $R\colon C_V \to 2^{C_K}$ ставит в соответствие каждой ячейке данных множество ячеек описания, т. е. задаёт структуру таблицы.

Логическая структура таблиц

Отображение R определяется взаимным расположением ячеек данных и описаний и стилевым оформлением ячеек. Для таблицы *простой структуры* (рис. 1) оно определяется следующим образом: для каждой ячейки данных выбираются все ячейки описания, пересекающиеся со строкой или столбцом рассматриваемой ячейки.

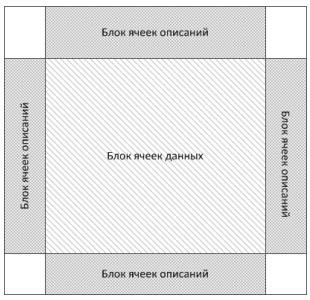


Рис. 1. Статистическая таблица простой структуры.

Первым шагом к построению отображения R является распознавание типов ячеек, т. е. построение множеств C_V и C_K. Сформулируем эту задачу как задачу классификации. В качестве объектов рассмотрим множество C ячеек некоторого множества таблиц. Множество классов в задаче совпадает с множеством $\{data, key, empty\}$.

Положение ячейки $c \in C$ в таблице описывается координатами левого верхнего $(r_1(c), c_1(c))$ и правого нижнего $(r_2(c), c_2(c))$ углов прямоугольника по сетке $G^{M\times N}$. Для каждой ячейки $c \in C$ генерируются следующие признаки:

Работа выполнена при финансовой поддержке РФФИ, проекты № 11-07-00480-а, № 10-07-00609-а, № 10-07-00422-а.

1) $f_1(c)$ — количество чисел;
2) $f_2(c)$ — количество слов;
3) $f_3(c)$ — количество символов;
4) $f_4(c) = (r_1(c)+r_2(c))/2M$ — вертикальное положение;
5) $f_5(c) = (c_1(c) + c_2(c))/2N$ — горизонтальное положение;
6) $f_6(c) = r_2(c) - r_1(c)$ — число вертикально объединённых элементов сетки;
7) $f_7(c) = c_2(c) - c_1(c)$ — число горизонтально объединённых элементов сетки.

Таблица с суперстроками. Таблица может быть разделена на несколько частей в соответствии со значениями некоторого показателя. Например, данные по мужской и женской занятости, данные за разные годы, абсолютные и относительные данные одних и тех же показателей. Для совмещения таких данных в одной таблице составители часто используют суперстроки (рис. 2).

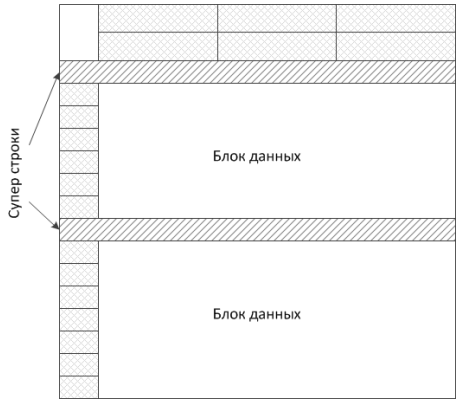


Рис. 2. Статистическая таблица с суперстроками.

Рассмотрим задачу классификации строк таблицы на два класса: «обычная строка» и «суперстрока». Для каждой строки будем строить следующий набор признаков:

1) $f_1(c)$ — количество ячеек в строке;
2) $f_2(c) = N$ — ширина таблицы;
3) $f_3(c)$ — высота строки;
4) $f_4(c)$ — количество пустых ячеек;
5) $f_5(c) = (c_1(c) + c_2(c))/2N$ — количество не пустых ячеек.

Вложенные ячейки. Ещё одним часто встречающимся приёмом оформления таблиц является использование вложенных ячеек, когда несколько ячеек сдвигаются на один уровень вправо (рис. 3). Этот приём часто используется в таблицах Росстата. Его игнорирование приводит к неправильному пониманию сути отражённой в таблице информации. Таблицы с несколькими уровнями вложенности встречаются редко и в настоящей работе не рассматриваются.

Для определения вложенности решается задача классификации, в которой объектами являются пары последовательно идущих ячеек $p = (x_1, x_2)$ в левом блоке ячеек описания, разделённых на три класса: «ячейки находятся на одном уровне», «x_2 сдвинута вправо относительно x_1» и «x_2 сдвинута влево относительно x_1». Для этих объектов вычисляется следующий набор признаков:

1) $f_1(p)$ — текст x_1 заканчивается на «:»;
2) $f_2(p)$ — количество начальных пробельных символов в тексте x_2;
3) $f_3(p)$ — тип первого непробельного символа в x_2: «цифра», «буква» или «знак»;
4) $f_4(p)$ — первая буква в x_1 является прописной;
5) $f_5(p)$ — первая буква в x_2 является прописной;
6) $f_6(p) = r_2(x_1) - r_1(x_1)$ — высота x_1;
7) $f_7(p) = c_2(x_2) - c_1(x_2)$ — высота x_2.

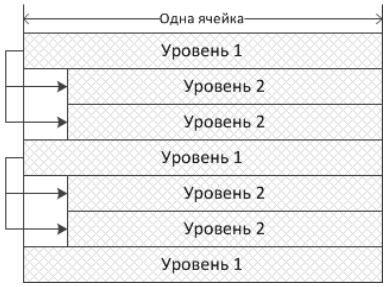


Рис. 3. Фрагмент таблицы с вложенными ячейками.

Алгоритмы динамического обучения

При динамическом обучении объекты появляются по одному, причём каждый объект обрабатывается только один раз. Динамическое обучение состоит из последовательности чередующихся шагов классификации и дообучения. На стадии классификации очередного объекта алгоритму не известно значение целевого признака, однако непосредственно после классификации ему передаётся правильный ответ, который используется для модификации параметров алгоритма (дообучения).

Обычно допускается возможность ошибок классификации на обучающей выборке. Однако в рассматриваемой задаче эксперт, обучающий систему, ожидает, что все сделанные им исправления будут сохранены. Поэтому к алгоритму классификации предъявляется *требование корректности* [1]: при последующих модификациях он не должен делать ошибок на ранее классифицированных объектах.

В [3] авторами исследовались алгоритмы динамического обучения, удовлетворяющие требованию корректности. По результатам экспериментов в качестве основного был выбран алгоритм инкрементного построения бинарного решающего дерева ITI (Incremental Tree Induction) [6].

Incremental Tree Induction

Алгоритм ITI [6] является модификацией алгоритма ID3 [5] для задач инкрементного обучения решающих деревьев (рис. 4). В начале работы алгоритма происходит обучение дерева по обучающей выборке аналогично ID3, но с сохранением объектов (или ссылок на них) в вершинах дерева. Таким образом, на каждом уровне дерева хранится информация обо всей выборке.

При инкрементном добавлении объектов они пропускаются по дереву от вершины к листьям аналогично классификации. Если метка листа совпадает с меткой объекта, то он сохраняется в листе. В противном случае строится поддерево, разделяющее новый объект и объекты, сохранённые в листе ранее. Полученное в результате дерево зависит от порядка добавления объектов и может сильно отличаться от оптимального. Для решения этой проблемы в [6] использовалась *операция транспозиции* дерева, заключающаяся в периодическом поиске лучшего предиката для каждого узла дерева. Временны́е затраты на транспозицию в среднем меньше, чем на полное перестроение дерева, но её выполнение затрудняет практическое применение алгоритма.

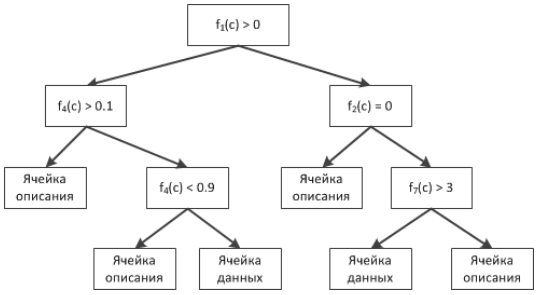


Рис. 4. Пример решающего дерева.

Новый алгоритм Random Incremental Forest

В композиции решающих деревьев Random Forest [4] каждое последующее дерево строится по некоторой подвыборке с повторением из обучающей выборки (бэггинг) в некотором случайном признаковом подпространстве.

Эту идею можно применить к инкрементному обучению. Будем строить композицию решающих деревьев, в которой за каждым деревом закреплено случайное признаковое подпространство. При построении деревьев поиск каждого предиката осуществляется только по одному случайно выбранному признаку из подпространства признаков. Для обеспечения корректности композиции от принципа бэггинга приходится отказаться. Объекты добавляются в каждое дерево композиции, аналогично алгоритму ITI. Операция транспозиции в данном случае также неприменима.

Каждое дерево композиции «слабее» дерева ITI, но этот недостаток компенсируется с помощью композиции. Более того, RIF работает значительно быстрее, чем ITI на больших задачах или с небольшими периодами транспозиции.

Операция транспозиции в некоторых задачах существенно улучшает качество работы алгоритма ITI, поэтому RIF может уступать ему в качестве классификации. Эта проблема решается добавлением некоторой стратегии перестроения деревьев в RIF. Например, можно полностью перестраивать композицию по всей выборке с заданной периодичностью.

Отбор деревьев

Идея заключается в том, чтобы сначала построить как можно больше деревьев в композиции, затем удалять худшие деревья и строить новые на основе признаков, используемых в лучших деревьях. Процедуры удаления и добавления выполняются с определённой периодичностью, при условии, что накопленного числа объектов достаточно для применения статистических критериев.

Каждое дерево в таком процессе строится по всей выборке, доступной на момент построения. Это начальные объекты обучения и объекты, добавленные в процессе динамического обучения. Каждому дереву ставится в соответствие пара чисел (n, m), где n — число объектов, классифицированных деревом после последнего перестроения, m — число ошибок. Поскольку разные деревья перестраиваются в разные моменты времени, числа n у них, как правило, отличаются. После начального построения дерева и после каждого перестроения числа n и m обнуляются, и далее в процессе динамического обучения обновляются после добавления каждого нового объекта.

Если допустить, что ошибка дерева является случайной величиной из распределения Бернулли с параметром p, то верхняя доверительная граница для p с уровнем доверия α, согласно теореме Муавра–Лапласа, равна

$$\bar{p} = \frac{m}{n} + \Phi_\alpha \sqrt{\frac{m(n-m)}{n^3}},$$

где Φ_α — α-квантиль стандартного нормального распределения. Эта оценка позволяет сравнивать деревья с различными значениями числа n.

Заметим, что отбор деревьев может приводить к отбору признаков. По мере увеличения длины выборки n малоинформативные признаки будут всё реже использоваться деревьями композиции, и в конечном итоге могут быть вовсе исключены.

Эксперименты

Алгоритмы RIF и ITI сравнивались на коллекции из 1000 таблиц. Начальная обучающая выборка составляла 50 объектов, после чего запускалось динамическое обучение. Все эксперименты проводились 20 раз со случайно перемешанной выборкой. Композиция RIF состояла из 50 деревьев. Кривые обучения алгоритмов (графики зависимости доли ошибок от числа обучающих объектов) представлены на рис. 5, 6, 7, доверительные интервалы приведены в Таблице 1.

	ITI			RIF		
	CT	SR	CI	CT	SR	CI
Мин.	0,05	0,02	0,6	0,04	0,02	1,3
Сред.	0,08	0,04	1,7	0,07	0,03	2,8
Макс.	0,13	0,08	2,5	0,11	0,05	4,4

Таблица 1. Доверительные интервалы частоты ошибок алгоритмов (в процентах) на последнем объекте.

Задача распознавания типа ячеек (CT) представлена 28 624 объектами, 8 числовыми признаками. На этой задаче ITI использовался без транспозиции из-за больших вычислительных затрат. Оба алгоритма справились с задачей достаточно успешно, средняя ошибка составила менее 0,1%. Следует заметить, что доверительный интервал ошибки у RIF меньше чем у ITI.

Задача распознавания суперстрок (SR) представлена 10 217 объектами, 5 числовыми признаками. На ней алгоритм ITI также использовался без транспозиции, для этой задачи справедливы выводы, аналогичные выводам по предыдущей задаче. Ошибка составила менее 0,1%.

Задача распознавания вложенных ячеек (CI) представлена 370 объектами, 4 номинальными и 7 числовыми признаками. Относительно небольшая длина выборки позволила использовать ITI с транспозицией (с периодом 50), в результате ITI справился с задачей лучше — 1,7% ошибок против 2,8% ошибок у RIF. Наличие транспозиции оказалось существенным для этой задачи.

Выводы

В статье рассмотрены задачи динамического обучения, возникающие при распознавании структуры статистических таблиц. Предложен алгоритм динамического обучения композиций случайных деревьев. В экспериментах он показал лучшее качество распознавания на задачах с большой длиной выборки. На задачах, где заметен выигрыш в качестве от использования транспозиции, улучшить качество работы композиции можно, добавив стратегию перестроения деревьев. Например, включить отбор деревьев, причем строить и удалять одинаковое число деревьев.

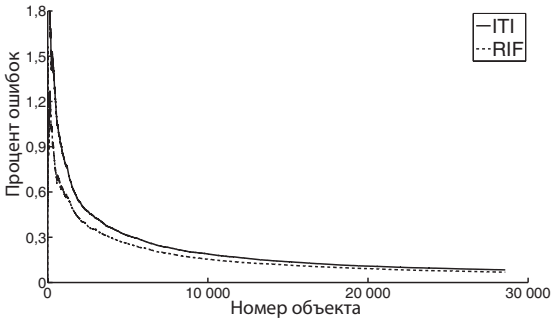


Рис. 5. Задача распознавания типа ячеек (CT).

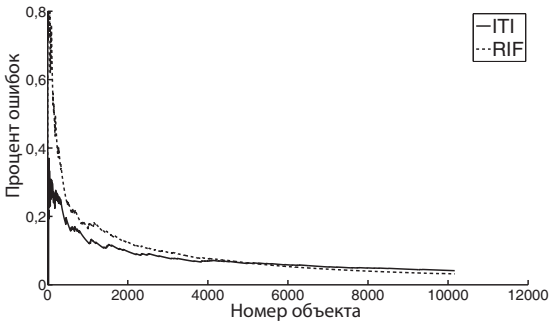


Рис. 6. Задача классификации суперстрок (SR).

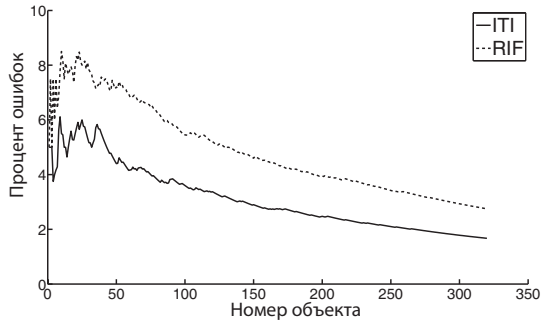


Рис. 7. Задача классификации вложенных ячеек (CI).

Литература

[1] *Журавлев Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики, 1978. — Т. 33. — С. 5–68.

[2] *Кудинов П. Ю.* Задача распознавания статистических таблиц // 14-я всеросс. конф. «Матем. методы распознавания образов», 2009. — С. 552–555.

[3] *Кудинов П. Ю., Полежаев В. А.* Динамическое обучение распознаванию статистических таблиц // 10-я межд. конф. «Интеллектуализация обработки информации», Кипр, 2010. — С. 512–515.

[4] *Breiman L., Schapire E.* Random forests // Machine Learning, 2001.

[5] *Quinlan J. R.* Induction of Decision Trees // Machine Learning, 1986.

[6] *Utgoff P. E., Berkman N. C., Clouse J. A.* Decision tree induction based on efficient tree restructuring // Machine Learning, 1997.

Синтез правил коррекции документов в формате LaTeX с помощью сопоставления синтаксических деревьев

Чувилин К. В.
kirill.chuvilin@gmail.com
Московский физико-технический институт (государственный университет)

Рассматривается задача автоматизации коррекции документов в формате LaTeX. Каждый документ представляется в виде синтаксического дерева. С помощью модифицированного алгоритма Zhang-Shasha строится отображение вершин дерева изначального документа в вершины дерева отредактированного документа, соответствующее минимальному редактирующему расстоянию. Отображения вершины в вершину составляют обучающую выборку, по которой генерируются правила замены для автоматической коррекции.

Работа относится к области автоматической обработки текстов. Поводом для нее послужила подготовка сборников трудов конференций ММРО и ИОИ в 2007–2011 годах. Многие конференции и издательства принимают материалы от авторов в формате LaTeX. В каждом издательстве есть определенные традиции и требования к оформлению публикуемого материала. К ним относится оформление заголовков, списков, таблиц, библиографии, формул, чисел, и многое другое. Ошибки, связанные с несоблюдением этих правил, называются *типографическими*. Обычно авторские тексты содержат значительное количество (десятки на страницу) таких ошибок, исправление которых производится корректорами вручную. Обработка одной страницы занимает до двух часов времени.

Предлагаемый подход направлен на значительное сокращение объёма рутинной работы путём создания автоматизированной системы, которая определяет в исходном тексте возможные места исправлений и предлагает корректору вариант замены. Если он согласен с заменой, он подтверждает её нажатием одной кнопки. Иначе ему приходится править текст по-прежнему вручную.

Одним из стандартных способов описания правил изменения текстовых данных являются регулярные выражения [2]. Однако они позволяют задать лишь вид окружающего текста, тогда как для описания правки часто требуется знать контекст рассматриваемого фрагмента в логической структуре документа, в том числе текущее состояние синтаксического анализатора LaTeX. Кроме того, регулярные выражения обладают некоторыми ограничениями. Например, невозможно описать структуру из парных скобок произвольной степени вложенности.

Файлы формата LaTeX обладают естественной древовидной структурой (*синтаксическим деревом*), исследуя которую, можно получить всю необходимую информацию для описания корректорской правки. Узлы этой структуры называются *токенами*. Корнем является окружение `document`. Примерами типов нетерминальных токенов являются: команда с параметрами (потомками являются все параметры), окружение (потомками являются команда начала, тело, команда конца), элемент списка, строка таблицы, ячейка таблицы, формула (различаются выключенная и включенная), группа (текст в фигурных скобках). Примерами типов терминальных токенов являются: команда без параметров, пробел, пустая строка (начало нового абзаца), слово, число, символ, имя файла, метка, параметр размера, текст внутри окружения `verbatim`.

Синтаксическое дерево взаимнооднозначно определяет документ LaTeX. Поэтому правила замены удобно формулировать именно для деревьев.

Правила замены можно задавать вручную, непосредственно на основе практического опыта корректоров. Однако ввиду значительно числа и разнообразия правил, это приведёт скорее к увеличению трудозатрат, особенно на начальном этапе. Поэтому предлагается строить правила по корпусу уже обработанных пар документов. Документы, не прошедшие корректуру, будем называть *черновиками*, прошедшие — *чистовиками*. Соответствующие синтаксические деревья — *чистовыми* и *черновыми*.

В данной статье рассматривается задача автоматической генерации правил преобразования документов по обучающей выборке, составленной из пар «черновик–чистовик».

Редактирующее расстояние между деревьями

Рассматриваются деревья, обладающие следующими свойствами: каждая вершина содержит *ключ* (элемент из заранее определенного набора), выбрана вершина, которая является корнем дерева, вершины, имеющие общего родителя, упорядочены.

К дереву разрешается последовательно применять следующие операции: удаление вершины (все ее потомки переходят родителю), вставка новой вершины в произвольное место, изменение ключа вершины.

Определение 1. *Редактирующим расстоянием между двумя деревьями называется минимальное количество операций удаления вершины, вставки вершины и изменения ключа, позволяющих получить из первого дерева второе.*

Алгоритм Zhang-Shasha

Этот алгоритм позволяет вычислять редактирующее расстояние между двумя деревьями и, кроме того, определять, какую операцию нужно применить к каждой вершине для реализации такого расстояния [1].

Определение 2. *В произвольном дереве каждой вершине можно сопоставить наиболее левую для нее вершину: каждой терминальной вершине — ее саму, любой другой — ту же, что и для самого левого ее потомка.*

Определение 3. *В произвольном дереве вершина, для которой наиболее левая вершина отличается от наиболее левой вершины для ее родителя называется ключевым корнем.*

Обратная нумерация вершин. Пусть у дерева n вершин. Каждой взаимно однозначно сопоставляется номер от 1 до n так, чтобы для любого поддерева выполнялись следующие условия:

— корень имеет номер больший, чем все остальные вершины,
— для любых двух потомков корня все вершины поддерева, образованного более левым, имеют меньшие номера, чем вершины поддерева, образованного более правым.

Такой порядок нумерации называется *обратным*. Оказывается, что поддерево, образованное произвольным ключевым корнем, состоит из вершин, номера которых не превосходят номера корня, и только из них.

Далее каждая вершина дерева будет обозначаться числом, равным ее номеру.

Отображения деревьев.

Определение 4. *Пусть заданы два дерева. Отображением первого дерева во второе называется правило, которое некоторым вершинам первого дерева взаимно однозначно сопоставляет некоторые вершины второго дерева так, чтобы порядок следования вершин сохранялся.*

Такие отображения принято записывать с помощью набора пар номеров вершин (прообраз и образ). Пусть отображение содержит пары (a, b) и (c, d). Тогда требуемые условия запишутся следующим образом:

$$a = c \Leftrightarrow b = d, \quad a < c \Leftrightarrow b < d.$$

Каждое такое отображение соответствует набору операций, используемых для построения редактирующего расстояния:

— если вершина первого дерева не имеет образа, то ее нужно удалить;
— если вершина второго дерева не имеет прообраза, то ее нужно добавить;
— если вершине первого дерева соответствует вершина второго с другим ключом, то нужно изменить ключ.

Таким образом, отображение, соответствующее минимальному количеству операций, реализует редактирующее расстояние.

Пример 1. На рисунке 1 показан пример обратной нумерации и отображения двух деревьев, соответствующего редактирующему расстоянию. Формально оно записывается следующим образом:

$$(1,1),\ (2,2),\ (3,3),\ (4,5),\ (6,6).$$

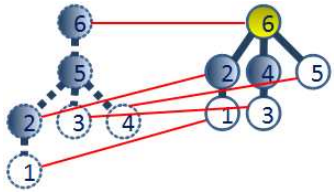


Рис. 1. Отображение деревьев.

Вычисление расстояний. В следующих формулах символы $\hat{\triangle}$ и $\hat{\blacktriangle}$ обозначают деревья с корнями $\circ$ и $\bullet$ соответственно, $\triangle$, $\blacktriangle$ — леса, образованные удалением корней этих деревьев, $\triangle\triangle$ и $\blacktriangle\blacktriangle$ — произвольные леса.

Расстояние между деревьями определяется рекуррентной формулой с помощью расстоянии между лесами:

$$\delta(\hat{\triangle}, \hat{\blacktriangle}) = \min \begin{cases} \delta(\triangle, \hat{\blacktriangle}) + 1; \\ \delta(\hat{\triangle}, \blacktriangle) + 1; \\ \delta(\triangle, \blacktriangle) + \delta(\circ, \bullet); \end{cases} \tag{1}$$

где $\delta(\circ, \bullet)$ равно 1, если корни деревьев имеют разный ключ, и равно 0, если одинаковый.

Расстояние между лесами или деревом и лесом, в свою очередь, определяется рекуррентной формулой:

$$\delta(\triangle\hat{\triangle}, \blacktriangle\hat{\blacktriangle}) = \min \begin{cases} \delta(\triangle\triangle, \blacktriangle\hat{\blacktriangle}) + 1; \\ \delta(\triangle\hat{\triangle}, \blacktriangle\blacktriangle) + 1; \\ \delta(\triangle, \blacktriangle) + \delta(\hat{\triangle}, \hat{\blacktriangle}). \end{cases} \tag{2}$$

Замечание 1. Вообще говоря, стоимость операций (аддитивный штраф в формулах) вставки, удаления и изменения ключа вершины могут быть не равны единице.

Через $K^1 = \{k_1^1, \ldots, k_{m_1}^1\}$ и $K^2 = \{k_1^2, \ldots, k_{m_2}^2\}$ обозначим упорядоченные по возрастанию наборы ключевых корней первого и второго деревьев соответственно, то основной цикл алгоритма запишется следующим образом.

1: **для** $i = k_1^1, \ldots, k_{m_1}^1$
2: **для** $j = k_1^2, \ldots, k_{m_2}^2$
3: `treeDist`(i, j);

Функция `treeDist`(i, j) вычисляет расстояние между поддеревьями первого и второго дерева с корнями i и j соответственно.

Построение отображений. Во время вычисления расстояний для ключевых корней заполняются две матрицы:

— матрица расстояний между деревьями, где в ячейке (i, j), образованной пересечением i-й строки и j-го столбца, стоит расстояние между поддеревом первого дерева с корнем i и второго с корнем j;
— матрица расстояний между лесами, где в ячейке (i, j) стоит расстояние между лесами, образованными удалением корней из соответствующих поддеревьев.

В правом нижнем углу таблицы расстояний между деревьями будет число, равное редактирующему расстоянию.

Для каждой ячейки двух таблиц можно вычислить, из каких других можно перейти в нее, согласно формулам (1) и (2). Другими словами, определить, какая операция производилась с соответствующей вершиной (не всегда однозначно, в таких случаях можно выбрать любую). Таким образом строится маршрут из правого нижнего угла таблицы расстояний между деревьями в левый верхний. Ячейки этой таблицы, которые попали в маршрут, зададут пары чисел, соответствующих отображению.

Итак, результат работы алгоритма: пары (прообраз и образ) не измененных вершин, пары (прообраз и образ) измененных вершин, множество удаленных вершин, множество добавленных вершин.

Применение для документов LaTeX

Если для терминальных токенов синтаксического дерева в качестве ключа выбрать текст, которому они соответствуют, а для нетерминальных — тип, то оно будет полностью удовлетворять условиям применимости алгоритма Zhang-Shasha. Но, как оказалось, правки, совершаемые корректорами, не могут быть заданы тремя вышеописанными действиями.

Пример 2. На рисунке 2 схематично показано отображение деревьев, которое дожно возникнуть при правке фрагмента

```
$(k-1)/(8L^2),$ где $L$ ’’--- параметр алгоритма.
```

в результате которой должно получиться

```
$(k-1)/(8L^2)$, где $L$ ’’--- параметр алгоритма.
```

В этом случае должно произойти перемещение токена, соответствующего запятой, что вызовет нарушение порядка.

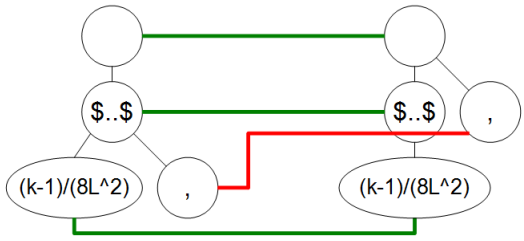


Рис. 2. Отображение, не задаваемое тремя видами операций.

Модификация алгоритма Zhang-Shasha. Для выделения подобных перемещений набор команд был расширен операциями *поднятия* и *опускания*.

В предположении, что найдено отображение произвольного дерева на другое, введены обозначения: $\mathbb{D}$ — множество удаленных вершин, $\mathbb{I}$ — множество добавленных вершин, $p(x)$ — родитель вершины x, $f(x)$ — образ вершины x ($\forall x \in \mathbb{D}\ f(x) = \varnothing$), $k(x)$ — ключ вершины x.

Определение 5. *Поднятые вершины — это удаленные вершины $\{x_1, \ldots, x_k\} \subset \mathbb{D}$ чернового дерева такие, что для $i = 1, \ldots, k$ выполняется: $x_i = x_1 + i - 1$ (последовательные), $p(x_i) = p(x_1)$ (имеют общего родителя), $p(x_i) = x_k + 1$ (являются последними потомками). При этом существуют добавленные вершины $\{y_1, \ldots y_k\} \in \mathbb{I}$ чистового дерева такие, что для $i = 1, \ldots, k$ выполняется: $k(x_i) = k(y_i)$ (ключи соответствуют удаленным вершинам), $y_i = y_1 + i - 1$ (последовательные), $p(y_i) = p(f(x_k + 1))$ (имеют общего родителя), $f(x_k + 1) = y_1 - 1$ (следуют за образом родителя $x_1, \ldots, x_k$).*

Определение 6. *Опущенные вершины — это удаленные вершины $\{x_1, \ldots, x_k\} \subset \mathbb{D}$ чернового дерева такие, что для $i = 1, \ldots, k$ выполняется: $x_i = x_1 + i - 1$, $p(x_i) = p(x_1)$ (имеют общего родителя). При этом существуют добавленные вершины $\{y_1, \ldots y_k\} \in \mathbb{I}$ чистового дерева такие, что для $i = 1, \ldots, k$ выполняется: $k(x_i) = k(y_i)$ (ключи соответствуют удаленным вершинам), $y_i = y_1 + i - 1$ (последовательные), $p(y_i) = f(x_1 - 1)$ (являются потомками образа вершины, предшествующей $x_1, \ldots, x_k$).*

Для всех поднятых и опущенных вершин отображение деревьев дополняется парами (x_i, y_i), $i = 1, \ldots, k$.

Для конструктивного поиска отображения с учетом введения новых операций можно воспользоваться следующим утверждением.

Теорема 1. *Если сначала применять алгоритм поиска отображения, реализующего редактирующее расстояние, для операций удаления, вставки и изменения вершин, а после найти все поднятые*

и опущенные вершины, то полученное отображение будет соответствовать наименьшему расстоянию среди всех отображений между двумя заданными деревьями, которые могут быть получены этими пятью операциями.

Выделение закономерностей

Каждое построенное правило для изменения дерева характеризуется *шаблоном* (последовательностью соседних токенов с общим родителем) и типом *локализатора* (токена, к потомкам которого применяется шаблон).

Определение 7. *Токен чернового дерева, к которому применяется одна из пяти вышеописанных операций в процессе перехода к чистовому дереву называется измененным.*

Определение 8. *Левая (правая) шаблонная цепочка радиуса r — это последовательность соседних токенов с общим родителем, длиной не больше r, такая, что если она содержит измененный токен, то только один, причем он самый левый (правый).*

Синтез шаблонов. Пусть токен x чернового дерева удален или у него изменен ключ. Тогда локализатор — это $p(x)$, шаблон составляется из левой и правой шаблонных цепочек, наиболее близких к x и самого токена x.

Пусть в чистовое дерево добавлен токен y. Тогда локализатор — это прообраз $p(y)$, если он существует; шаблон составляется из левой шаблонной цепочки, начинающейся в прообразе левого соседа y, если он существует, и аналогичной правой.

Пусть токены $x_1, \ldots, x_k$ чернового дерева были подняты. Тогда локализатор — это $p(x_1)$, шаблон состоит из токенов $x_1, \ldots, x_k$.

Пусть токены $x_1, \ldots, x_k$ чернового дерева были опущены. Тогда локализатор — это $p(x_1)$, шаблон состоит из токенов $x_1, \ldots, x_k$ и токена $x_0 = x_1 - 1$, если $p(x_0) = p(x_1)$.

Сразу после того, как для какой-то операции найден шаблон, эта операция применяется к дереву. Дальнейший поиск происходит в уже измененном дереве.

Использование шаблонов. Для каждого токена синтаксического дерева документа выбираются шаблоны, тип локализаторов которых совпадает с типом рассматриваемого токена. Среди потомков токена ищется цепочка соседних, совпадающая с шаблоном по следующим правилам:

— для всех измененных токенов из шаблона соответствующие потомки должны иметь такие же ключи,

— для всех остальных токенов из шаблона соответствующие потомки должны иметь такие же типы.

Сразу после того, как совпадение найдено, выполняется соответствующая операция. Дальнейший поиск происходит в уже измененном дереве.

Эксперимент

В качестве данных использовались 135 пар черновых и чистовых статей, вошедших в сборник трудов конференции ИОИ-8. Радиус шаблонных цепочек был выбран $r = 5$.

Для каждого количества n пар статей, используемых для обучения, 45 раз случайным образом генерировалась выборка. Определялось количество синтезированных шаблонов и количество мест в чистовиках, удовлетворяющих шаблонам (количество срабатываний). На рисунке 3 показаны графики с усредненными значениями для каждого n.

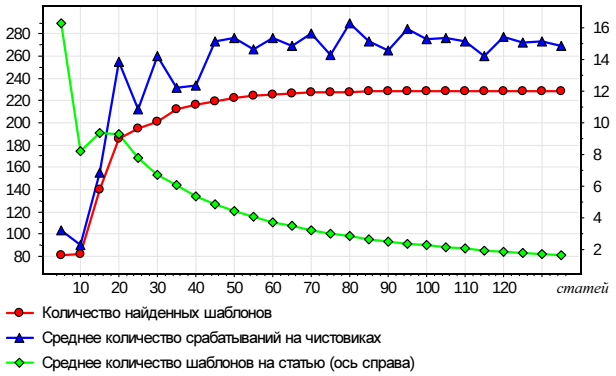


Рис. 3. Результаты эксперимента.

Поскольку чистовики, с точки зрения корректора, не должны быть подвержены правкам, для корректных шаблонов (полностью соответствующих описаниям правок) количество срабатываний на чистовиках должно стремиться к нулю при увеличении числа статей, используемых для обучения. С построенными шаблонами такого не произошло. Это дает направление дальнейшим исследованиям, связанным с модификацией построенных шаблонов.

Литература

[1] *Zhang K., Shasha D.* Simple fast algorithms for the editing distance between trees and related problems // SIAM Journal of Computing, 1989. — No. 18. — Pp. 1245–1262.

[2] *Фридл Д.* Регулярные выражения, 3-е издание. Пер. с англ. — СПб.: Символ-Плюс, 2008.

Классификационный метод идентификации имитационных моделей транспортных потоков*

Ивкин Н. П., Чехович Ю. В.
ivkinnikita@gmail.com, chehovich@forecsys.ru
Москва, Вычислительный Центр им. А. А. Дородницына РАН

В настоящей работе рассматривается приложение классификационного подхода к имитационному моделированию транспортных потоков. Дается краткий обзор классического подхода и его сравнение с предложенным. Предлагается программный стенд, как способ идентификации имитационных моделей.

Введение

Настоящая работа посвящена вопросам идентификации микроскопических моделей транспортных потоков. Авторы предлагают использовать для этого классификационный подход к имитационному моделированию сложных социально-технических систем, предложенный в [1, 2]. Термин «микроскопическое моделирование», используемый в работе, соответствует классификации на макро-, микро- и мезоскопическое моделирование транспортных потоков, предложенной в [3]. Очевидно, что моделируемый объект, а именно транспортный поток, является сложной социально-технической системой в терминах, введенных в [1], а именно, является системой, состоящей из большого числа субъектов, действующих друг на друга и на систему в целом. Водители транспортных средств, составляющих поток, принимают решения, которые полностью определяют характер движения управляемого ими транспортного средства. В дальнейшем, пару ⟨транспортное средство, водитель⟩ будем называть субъектом и, говоря, что субъект принимает решение, будем иметь в виду, что решение принимает водитель.

Классификационный подход

В классификационном подходе [1, 2] предполагается, что каждый субъект имеет информацию о себе и субъектах из некоторой своей «окрестности», то есть «описание ситуации», и на основе анализа описания ситуации принимает решение о своем дальнейшем поведении. Чаще всего решение выбирается субъектом из конечного, скорее всего «относительно небольшого» множества решений. Стоит отметить, что для каждой конкретной задачи необходимо формально ввести параметры «описания ситуаций» и описать (чаще всего перечислить все элементы) множество решений.

Другим важным предположением является возможность сравнения ситуаций друг с другом, то есть наличие некоторой меры сходства (или даже метрики) в пространстве описаний «ситуаций». Это позволяет проводить классификацию «ситуаций» Предполагается, что субъект, в ситуациях одного класса принимает одни и те же решения. Продолжением этой идеи будет введение меры близости и в пространстве решений и предположение, что субъект, в ситуациях одного класса принимает не одни и те же, но схожие решения.

Наконец, третьим важным предположением является наличие типов субъектов. Субъекты, принадлежащие одному типу, в схожих ситуациях принимают одинаковые (или близкие) решения. Целесообразно рассматривать типы субъектов только в том случае, если субъектов относительно много, а типов сравнительно мало.

Процедура построения модели

Описание пространств. Рассмотрим одну из моделей классического подхода к имитационному моделированию транспортных потоков, на ее примере покажем ключевые различия классического и классификационного подходов. В классическом, как и в классификационном подходе, моделью является отображение из пространства «описаний ситуаций» в пространство решений. Рассмотрим модель ускорения (1), предложенную Газизом в [7]:

$$a_n(t) = \alpha \frac{V_n(t)^\beta \Delta V_n^{front}(t-\tau_n)}{\Delta X_n^{front}(t-\tau_n)^\gamma} \qquad (1)$$

где a_n — ускорение моделируемого автомобиля, ΔV_n^{front} — относительная скорость впереди идущего автомобиля, ΔX_n^{front} — расстояние до впереди идущего автомобиля, V_n — скорость моделируемого автомобиля, τ_n — время реакции водителя, α, β, γ — внутренние параметры.

Параметрами «описания ситуации» субъекта для данной модели являются характеристики ближайших транспортных средств и самого автомобиля, что может быть представлено в виде вектора микропараметров $(\Delta V_n^{front}, V_n, \Delta X_n^{front})$. При построении модели согласно классификационному подходу предлагается пользоваться теми же соображениями.

Множество решений описывается всевозможными значениями ускорения, которые может предпринять моделируемое ТС. Для рассматриваемой нами модели (1) ускорение может принимать значения

Работа выполнена при частичной финансовой поддержке Российского фонда фундаментальных исследований (проект №11-07-00424) и гранта Президента РФ для поддержки молодых российских ученых — кандидатов наук МК-7954.2010.9

из множества действительных чисел, т. е. для нее множество принимаемых решений не конечно. Для классификационного подхода предлагается взять конечное и достаточно небольшое множество решений, например: сильное и слабое ускорения, поддержание скорости, сильное, слабое и экстренное торможение. Ключевым моментом является конечность множества принимаемых решений.

Построение отображения. Классический подход к построению отображения состоит из построения с помощью экспертного вмешательства интерпретируемой формулы (как, например, в (1)) и настройки внутренних параметров для конкретной прикладной задачи.

Согласно классификационному подходу, когда получено формальное описание пространства ситуаций и задано множество принимаемых решений, процедура построения модели разбивается на два этапа (см. рис. 440):

1. **Описание поведения каждого субъекта зонной диаграммой.** Для каждого субъекта на прецедентной базе вида ⟨описание ситуации, решение⟩ решается задача классификации [4]. Построенный классификатор разбивает все пространство ситуаций на зоны различного поведения (зонная диаграмма). Таким образом, каждому вектору локального описания ситуации сопоставляется определенное решение из множества возможных решений.
2. **Описание типов субъектов.** При наличии большой выборки из субъектов, описанных своими зонными диаграммами, решается задача кластеризации [4] на множестве зонных диаграмм, тем самым субъекты со схожим выбором решений соотносятся к одному типу.

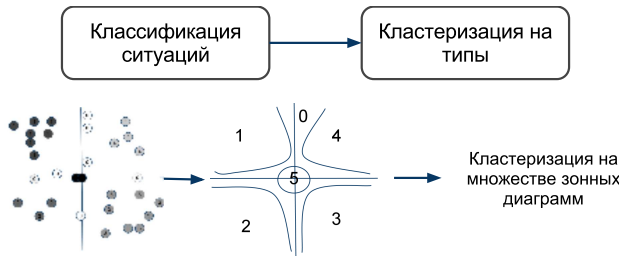


Рис. 1. Этапы построения модели

Для создания прецедентной базы вида ⟨описание ситуации, решение⟩ авторы предлагают использовать программный стенд, имитирующий движение автомобиля, управляемого оператором, в транспортном потоке. Все локальные описания ситуаций и принятые в них решения фиксируются программой, затем по полученной прецедентной базе идентифицируется модель.

Более правильным инструментом для создания прецедентной базы являлось бы реальное транспортное средство, оборудованное датчиками, фиксирующими описания ситуаций и решения, принятые водителем в этих ситуациях. Программный стенд позволяет продемонстрировать основные качества предложенной модели.

Приведем пример полученной с помощью программного стенда зонной диаграммы(см. рис. 2).

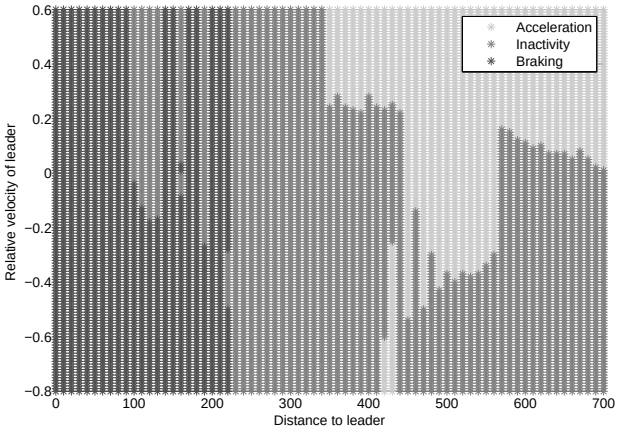


Рис. 2. Зонная диаграмма, построенная с помощью стенда.

Предложенный в работе подход применялся к двум базовым классам микроскопических моделей транспортных потоков: модели следования за лидером и модель ускорения. Следует отметить,что модели этих базовых классов, построенные по классическому подходу, обладают рядом недостатков:

1. Большинство моделей настраивается на некоторого «среднего» водителя («средний» по внутренним параметрам), что чаще всего неинформативно.
2. Большинство программных комплексов, основанных на микро-моделировании, не предполагают переключения моделей для описания одного и того же субъекта. Хотя в различных ситуациях этот объект может вести себя по-разному.
3. Исторически сложилось, что большинство предлагаемых моделей являлись усложнениями и обобщениями уже существующих, например цепочка моделей Дженерал Моторс (в использованных ранее обозначениях):

$$a_n = \alpha \Delta V_n^{front}(t - \tau_n) \text{ — Чендлер, 1958 [5];}$$

$$a_n = \alpha \frac{\Delta V_n^{front}(t - \tau_n)}{\Delta X_n^{front}(t - \tau_n)} \text{ — Газис, 1959 [6];}$$

$$a_n = \alpha \frac{V_n(t)^{\beta} \Delta V_n^{front}(t - \tau_n)}{\Delta X_n^{front}(t - \tau_n)} \text{ — Газис, 1961 [7].}$$

Каждая последующая модель имеет новые внутренние и внешние параметры. Тогда для

каждой конкретной ситуации встает вопрос о том, какую модель выбрать. Необходимо взять с одной стороны не слишком простую, но с другой стороны не слишком сложную, чтобы не допустить переобучения.

Классификационный подход устраняет эти недостатки:

1. Модель, построенная в соответствии с классификационным подходом, предполагает разделение субъектов на типы. То есть если на дороге присутствуют несколько групп субъектов, поведение которых сильно отличается, то во время обучения на этапе решения задачи кластеризации число кластеров-типов будет равно числу групп.
2. В отличии от классического подхода, где построение каждого отображения(модели) связано только с варьированием внутренних параметров, в классификационном подходе отображение(модель) строится только на основе обучающей выборке. Поэтому если в различных ситуациях на обучающей выборке субъект ведет себя не в рамках одной классической модели, то классификационный подход построит более «гибкую» модель.
3. В модели, построенной в соответствии с классификационным подходом, проблема выбора сложности переходит в задачу отбора признаков при решении задачи классификации. А для задач классификации уже существуют методы редукции признакового описания (описания локальных ситуаций в нашем случае)[4].

Таким образом, предложенный подход позволяет элиминировать указанные недостатки. Так же в данной работе доказывается, что любая адекватная классическая модель может быть сколь угодно точно приближена моделью, построенной согласно классификационному подходу.

Выводы

В настоящей работе рассмотрено приложение классификационного подхода предложенного в [1, 2] к транспортным потокам, предложен программный стенд, используемый в демонстрационных целях для идентификации модели, рассмотрены основные преимущества предложенной модели над классическим подходом. Доказана сводимость классических моделей к модели, построенной согласно классификационному подходу. С помощью программного стенда получена простая модель.

Данная работа является применением классификационного подхода к конкретной отрасли, но только на базовых задачах. Поэтому задача создания универсального подхода, пока до конца не написан формальный язык, остается нерешенной.

Литература

[1] *Чехович Ю. В.* Об идентификации имитационных моделей сложных социально-технических систем по агрегированным данным. // Интеллектуализация обработки информации: 8-я международная конференция. Республика Кипр, г. Пафос, 17–24 октября 2010 г.: Сборник докладов. — М.: МАКС пресс, 2010. — С 539–540.

[2] *Чехович Ю. В.* Применение алгебраического подхода к имитационному моделированию сложных социально-технических систем. // Сборник докладов третьей всероссийской научно-практической конференции «Имитационное моделирование. Теория и практика», Санкт-Петербург, 2007. — Том I. — С. 249–252.

[3] *Гасников А В., Кленов С. Л., Нурминский Е. А., Холодов Я. А., Шамрай Н. Б.* Введение в математическое моделирование транспортных потоков: учеб. пособие. — М.: МФТИ, 2010. — 362 с.

[4] *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning. — Springer, 2001. — 533 p.

[5] *Chandler R., Herman R. and Montroll C.* Traffic dynamics: Studies in car following. // Operations Research, 1958. — No. 6. — Pp. 165–184.

[6] *Gazis D., Herman R. and Potts B.* Car following theory of steady-state traffic flow. // Operations Research, 1959. — No. 7. — Pp. 499–505.

[7] *Gazis D., Herman R. and Rothery R.* Nonlinear follow-the-leader models of traffic flow. // Operations Research, 1961. — No. 9. — Pp. 545–567.

Визуализация многомерных данных методом проецирования на пространства малой размерности

Саакян Р. Р., Шпехт И. А.
rsahakyan@yahoo.com, shpekht@mail.ru
Анапа, филиал Российского государственного социального университета в г. Анапе

В статье предложен новый метод визуализации многомерных данных, где для восприятия многомерного пространства реализовано проецирование многомерных данных на пространства малой размерности (двумерных, трёхмерных) путём их разбиения выходящими из начала координат лучами, равных по количеству размерности исходного пространства данных. Для каждой точки исходного пространства определяются точки-отображения как центры тяжести плоских многоугольников, полученных в результате отложения величины соответствующих координат на лучах. Полученные результаты могут использоваться при построении решающих правил в задачах принятия решений в трудноформализуемых технических и информационных системах.

Более полное изучение и развитие сложной системы управления возможно с помощью многомерного анализа данных, представляющего её многопараметрическое поведение. Решение задач анализа системы в целом получается весьма эффективным, если удаётся на основе экспериментальных данных проводить первичный разведочный анализ многомерных данных, представляющих систему.

С другой стороны, одной из проблем существующих систем классификации и распознавания является представление промежуточных и итоговых результатов исследования в виде, удобном для визуального анализа пользователя, так как данные в них зачастую имеют многомерный характер, в то время как пространство размерностью более трёх переменных визуально трудно представимо. Поэтому анализ многопараметрических задач, как правило, проводят без надлежащего графического отображения данных.

Постановка задачи

Существуют различные методы обработки многомерных данных, например, факторный анализ, кластерный анализ, таксономия и т. п. Однако решения, найденные разными методами, могут существенно отличаться друг от друга, в силу ограничений, накладываемых разными методами, как на исходную выборку, так и на количество и форму полученных кластеров, что может привести к неадекватным и неустойчивым результатам [1–3].

В связи с этим важную роль играет разработка когнитивных методов визуализации многомерных данных, позволяющих исследователю в наглядном виде визуально обнаружить характерные особенности массива изучаемых данных. Программное обеспечение, разработанное с использованием таких методов, позволяет анализировать многомерные данные тем специалистам в различных предметных областях, которые слабо знакомы с методами многомерного анализа.

Методы и подходы решения

К сожалению, существующие методы визуализации данных не свободны от недостатков. Поскольку человек не способен непосредственно воспринимать образы пространства с числом измерений более 3-х, возникает необходимость отображать многомерные данные на пространство малой размерности, что так или иначе приводит к некоторым искажениям. Наиболее характерны следующие проблемы:

— наложение двух и более кластеров;
— искажение топологии (в некоторых методах близкие точки многомерного пространства могут проецироваться в далёкие точки пространства малой размерности, и одновременно, далёкие точки в близкие).

Для усовершенствования существующих методов визуализации многомерных данных с целью предварительной разведки их структуры, обеспечения надёжного разделения кластеров и уменьшения искажений расстояний между точками многомерного пространства при их проецировании в пространство малой размерности (искажений топологии), авторами был разработан метод визуализации многомерных данных на двумерную плоскость применительно к непрерывности признакового пространства (алгоритм «ЛИЛИЯ»). Метод позволяет частично решить проблему наложения различных кластеров при проецировании многомерных данных в пространство малой размерности.

Исходной информацией для применения метода является числовая таблица (матрица) многомерных данных типа «объект–признак», где строки отражают информацию об изучаемых объектах или явлениях, а столбцами являются свойства (признаки, характеристики), описывающие эти объекты и явления. Здесь объектами могут быть живые организмы, ситуации, социальные процессы и т. д.

Содержание алгоритма «ЛИЛИЯ» опишем пошагово.

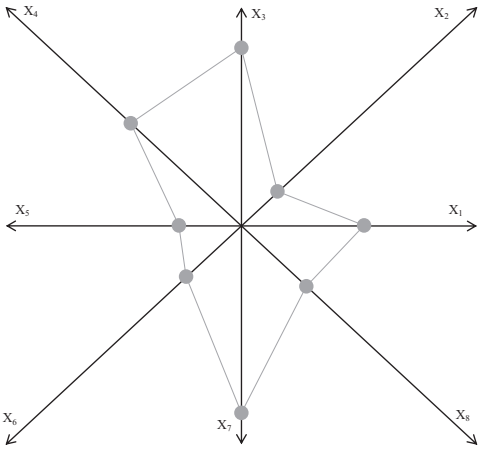


Рис. 1. Представление объекта на многолучевой плоскости для 8-мерного исходного пространства.

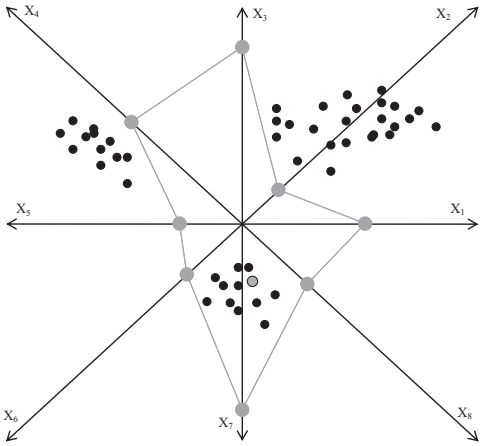


Рис. 2. Первичная разведка многомерных данных алгоритмом «ЛИЛИЯ».

1. Начало координат исходного многомерного пространства смещается таким образом, чтобы любая точка (объект) исходного набора данных имела только положительные координаты и проводится их нормирование.
2. Двумерная плоскость разбивается на равные сектора лучами, исходящими из координатного центра (число лучей равно размерности исходного пространства данных).
3. Каждый луч представляется в виде оси координат, на которых и откладываются представленные в числовом виде нормализованные величины свойства объектов для каждой точки исходного многомерного пространства.
4. Полученные точки на лучах соединяются отрезками прямой, и каждый исследуемый многомерный объект отображается на плоскости в виде плоского многоугольника, состоящего из последовательно соединённых точек — нормализованных атрибутов объекта (рис. 1). На данном шаге получаем отображение объектов без потери информации, но отображение большого числа объектов затрудняет их визуальное восприятие (многоугольниками накладываются друг на друга и мешают их восприятию).
5. Строятся точки-отображения для исходных объектов многомерного пространства в виде центров тяжести полученных ранее фигур (плоских многоугольников). При построении точек-отображений имеет место потеря исходной информации, однако это позволяет существенно упростить анализ большого объёма данных. В большинстве случаев исследователь может визуально выделить отдельные кластеры.

Итогом работы алгоритма «ЛИЛИЯ» является массив точек-отображений на плоскости (приведённые графические представления данных), которые характеризуют исследуемые объекты лучшим образом с точки зрения их разделения на кластеры. Данное представление удобно для первичной разведки многомерных данных (рис. 2).

Однако в процессе работы алгоритма визуализации «ЛИЛИЯ» могут возникать некоторые ошибки или неточности, связанные с приведённым графическим представлением многомерного объекта. Так, например, при визуализации объектов с пропорциональными, но разными по абсолютной величине характеристиками, получаются близкие точки-отображения на плоскости, как это изображено на рис. 3.

При этом исключаются ситуации, когда близкие точки отображаются как далёкие, а обратная ситуация — отображение далёких точек в близкие — возможна. Подобные искажения могут приводить к наложению некоторых кластеров.

Один из способов исключения указанных ошибок заключается в переходе к трёхмерному представлению с соответствующим изменением расположения лучей в пространстве. Они, как и в случае двумерного представления, исходят из координатного центра, но под углом к исходной плоскости, который задаётся пользователем, образуя тем самым перевернутую пирамиду с вершиной в начале координат. Из-за сходства полученной картины представления данных с закрываемым зонтом данный алгоритм получил название «ЗОНТ» (рис. 4).

В этом случае отметки, сделанные на лучах (соответствующие координатам исходных точек многомерного пространства) сместятся относительно нормали исходной плоскости, образуя трёхмерные многоугольники, центры тяжести которых и будут представлять собой точки-отображения для трёхмерного пространства.

Таким образом, расположенные в исходном многомерном пространстве на больших расстояниях точки (спроецированные в близкие или совпадающие точки при двумерном отображении) оказы-

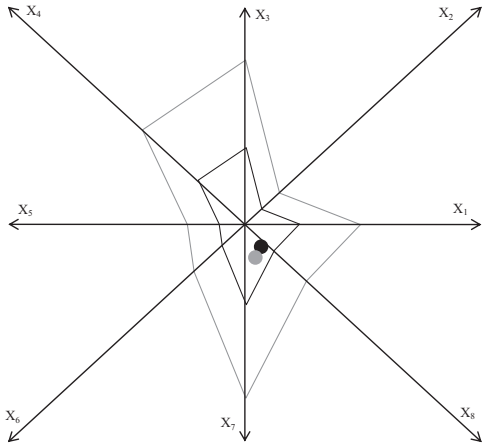


Рис. 3. Представление близких приведённых данных.

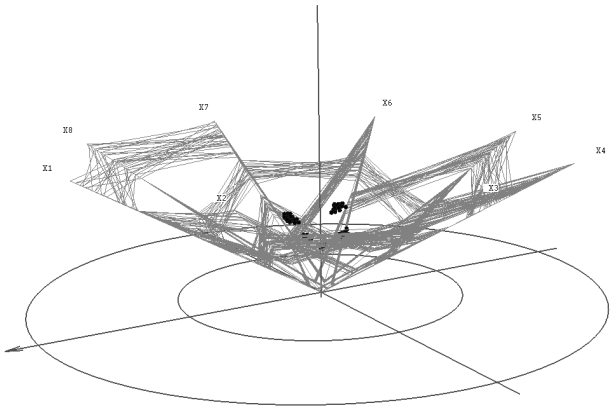


Рис. 4. Графическое представление алгоритма «ЗОНТ».

ваются в новом трёхмерном отображении на большом удалении друг от друга, что позволяет визуально разделить наложенные кластеры. В итоге при трёхмерном представлении приведённых данных происходит расхождение наложенных кластеров, что соответствует реальному представлению данных.

Заключение

Для реализации предложенных алгоритмов визуализации многомерных данных методом проецирования в пространства малой размерности был разработан программный комплекс визуализации (алгоритмы «ЛИЛИЯ», «ЗОНТ») предназначенный для извлечения характерных особенностей и выявления кластерной структуры в многомерном массиве данных. Программный комплекс упрощает проблему разведочного анализа многомерных данных и первичную классификацию объектов в условиях априорной неопределённости в отношении, как самих объектов, так и условий их наблюдения.

Форматом входных данных для программы являются электронные таблицы с нормированными значениями непрерывнозначных признаков объектов представления.

Работоспособность алгоритмов «ЗОНТ» и «ЛИЛИЯ» была подтверждена на примере гипотетического массива многомерных данных, а также на классическом наборе многомерных данных «Ирисы Фишера».

Литература

[1] *Журавлев Ю. И.* Распознавание образов // Избранные научные труды, Москва: Магистр, 1998. — 415 с.

[2] *Загоруйко Н. Г.* Прикладные методы анализа данных. — Новосибирск: Издательство института Математики, 1999. — 270 с.

[3] *Зиновьев А. Ю.* Визуализация многомерных данных. — Новосибирск: Изд.КГТУ, 2000. — 320 с.

Содержание

Авторский указатель

Научное издание

Математические методы
распознавания образов

Сборник докладов
15-й Всероссийской конференции

Напечатано с готового оригинал-макета

Издательство ООО «МАКС Пресс»
Лицензия ИД №00510 от 01.12.1999
Подписано к печати 09.08.2011
Печать офсетная. Бумага офсетная.
Формат 60×88 1/8. Усл. печ. л. 77,5. Тираж 300 экз. Изд. № 331 Заказ

119992, ГСП-2, Москва, Ленинские горы, МГУ им. М. В. Ломоносова,
2-й учебный корпус, 627 к.
Тел. 939-3890, 393-3891, Тел./Факс. 939-3891.