СПЕЦКУРС
Логический анализ данных в распознавании
(Logical data analysis in recognition)

лектор д.ф.-м.н. Елена Всеволодовна Дюкова

Спецкурс посвящён вопросам применения аппарата дискретной математики в задачах интеллектуального анализа данных. Излагаются общие принципы, лежащие в основе логического подхода к задачам машинного обучения. Описываются методы конструирования процедур классификации по прецедентам с использованием понятий теории булевых функций и теории покрытий булевых матриц. Рассматриваются основные модели логических процедур классификации, вопросы сложности их реализации и качества решения прикладных задач.

Спецкурс для бакалавров 2-4 курсов ВМК МГУ им. М.В. Ломоносова.

По спецкурсу издано учебное пособие:
http://www.ccas.ru/frc/papers/djukova03mp.pdf

Лекция 9
Об алгебро-логической коррекции элементарных классификаторов

• В случае целочисленных признаков небольшой значности задача корректного распознавания по прецедентам успешно решается методами логического подхода, среди которых основными являются алгоритм вычисления оценок, голосование по тестам и голосование по представительным наборам (последний часто называют алгоритмом типа КОРА). Позднее в работах Е.В. Дюковой и Н.В. Песковым было предложено понятие элементарного классификатора и построены новые модели. Имеются в виду алгоритм голосования по антипредставительным наборам и алгоритм голосования по покрытиям класса. Была дана общая схема синтеза логических процедур классификации, основанная на голосовании по элементарным классификаторам.

- Напомним, что эл.кл. $(\boldsymbol{\sigma}, \boldsymbol{H})$ называется корректным для класса $\boldsymbol{K}$, если не существует двух прецедентов, содержащих $(\boldsymbol{\sigma}, \boldsymbol{H})$ и таких, что один из них принадлежит $\boldsymbol{K}$, а другой $\boldsymbol{K}$ не принадлежит. Корректный эл.кл. класса $\boldsymbol{K}$, содержащийся хотя бы в одном прецеденте из класса $\boldsymbol{K}$, называется представительным набором этого класса. Будем говорить, что эл.кл. $(\boldsymbol{\sigma}, \boldsymbol{H})$ имеет ранг $\boldsymbol{r}$, если $\boldsymbol{H}$ – набор из $\boldsymbol{r}$ признаков.

- Рассмотрим процедуру голосования по представительным наборам. На этапе обучении для каждого класса строится семейство представительных наборов. Распознавание объекта осуществляется голосованием по представительным наборам построенных семейств. Корректность алгоритма обеспечивается за счёт корректности каждого участвующего в голосовании представительного набора. Основной задачей этапа обучения является поиск информативных представительных наборов. Практика показывает, что, как правило, информативными являются представительные наборы небольшого ранга. Например, алгоритм КОРА, предложенный Вайнцвагом, использует представительные наборы с рангом меньшим или равным 3.

- Однако в задачах с большой значностью признаков, как правило, почти все корректные эл.кл. имеют большой ранг и, как следствие, каждый такой эл.кл. содержится в небольшом числе прецедентов (под значностью признака понимается число его различных значений). Задачи, в которых значность признаков высока, сложны для классических логических алгоритмов распознавания. Существуют различные способы решения этой проблемы.

- Один из этих способов заключается в выполнении «корректной» перекодировки исходных признаков с целью понизить их значность. Требуется разбить множество значений каждого признака на интервалы порогами. Значения признака, попавшие в один интервал, считаются близкими и кодируются одним числом. Корректное перекодирование основано на рассмотрении только таких порогов, которые содержат «различающую» информацию. В результате после перекодирования объекты из разных классов остаются различимыми.

• Задача корректного перекодирования сводится к поиску покрытия специального вида булевой матрицы, построенной по обучающей выборке. Основная проблема – выбор функционала, характеризующего качество перекодировки. Построение наилучшей в смысле качества распознавания корректной перекодировки – труднорешаемая оптимизационная задача.

• Другой способ основан на идеях алгебро-логического подхода (предложен Е.В. Дюковой, К.В. Рудаковым и Ю.И. Журавлёвым в 1996 г.). Этот подход базируется на использовании произвольных эл.кл. (не обязательно корректных) и объединяет идеи логического и алгебраического подходов.

• Алгебраический подход, развиваемый школой Ю. И. Журавлева, применяется тогда, когда требуется скорректировать работу нескольких алгоритмов, каждый из которых безошибочно классифицирует лишь часть обучающих объектов. Цель коррекции сделать так, чтобы ошибки одних алгоритмов были скомпенсированы другими алгоритмами и качество результирующего алгоритма оказалось лучше, чем качество каждого из базовых алгоритмов в отдельности. Об алгебро-логическом подходе говорят, когда каждый базовый алгоритм однозначно определяется некоторым эл.кл. и корректирующие функции являются булевыми.

• Основным понятием алгебро-логического подхода является понятие корректного набора эл.кл.

• Рассмотрим набор эл.кл. $\boldsymbol{U = \{(\sigma_1, H_1), \dots, (\sigma_r, H_r)\}}$. Обозначим через $\boldsymbol{P_{(\sigma_i, H_i)}(S)}$, $\boldsymbol{S \in M}$, $\boldsymbol{i \in \{1, 2, \dots, r\}}$, предикат, равный 1 в том и только в том случае, если объект $\boldsymbol{S}$ содержит эл.кл. $\boldsymbol{(\sigma_i, H_i)}$.

• Бинарный вектор $\boldsymbol{W(U,S) = \{P_{(\sigma_1,H_1)}(S), \dots, P_{(\sigma_r,H_r)}(S)\}}$ называется откликом набора $\boldsymbol{U}$ на объекте $\boldsymbol{S}$.

• Набор эл.кл. $\boldsymbol{U}$ называется ***корректным для класса*** $\boldsymbol{K}$, если не существует двух прецедентов с одинаковыми откликами и таких, что один из них принадлежит $\boldsymbol{K}$, а другой не принадлежит $\boldsymbol{K}$.

• Очевидно, что для корректного набора эл.кл. $\boldsymbol{U}$ всегда существует частичная булева функция, выполняющая роль корректирующей функции. Эта функция определена на откликах прецендентов и принимает значение 1, если прецедент из $\boldsymbol{K}$, и принимает значение 0 в остальных случаях. Особо следует отметить корректные наборы эл.кл., имеющие монотонную корректирующую функцию. Такие наборы называются монотонными.

- Очевидным является следующее

- **Утверждение.** Набор эл.кл. $\boldsymbol{U}$ является монотонным корректным набором для класса $\boldsymbol{K}$ тогда и только тогда, когда для любых двух обучающих объектов $\boldsymbol{S'}$ и $\boldsymbol{S''}$ таких, что $\boldsymbol{S' \in K}$, $\boldsymbol{S'' \notin K}$, можно указать $\boldsymbol{i \in \{1, 2, \dots, q\}}$ такое, что $\boldsymbol{P_{(\sigma_i, H_i)}(S') = 1}$ и $\boldsymbol{P_{(\sigma_i, H_i)}(S'') = 0}$.

- Требование монотонности можно убрать, если условие утверждения 1 заменить на условие $\boldsymbol{P_{(\sigma_i, H_i)}(S') \neq P_{(\sigma_i, H_i)}(S'')}$.

- Модели алгоритмов распознавания, основанные на построении корректных наборов эл.кл. получили название логических корректоров.

- В самых общих чертах работа логического корректора заключается в следующем.

- На этапе обучения исходная выборка обучающих объектов делится на базовую и настроечную. По базовой подвыборке для каждого класса строятся корректные для этого класса наборы эл.кл., по настроечной подвыборке оценивается распознающая способность этих наборов. При этом перечисление корректных наборов эл.кл сводится к перечислению покрытий специальной булевой матрицы (в частности, к перечислению неприводимых покрытий). Распознающая способность построенных корректных наборов эл.кл. проверяется путём проведения процедуры голосования для объектов из настроечной подвыборки. Затем с помощью генетического алгоритма отбираются корректные наборы эл.кл. с распознающей способностью близкой к максимальной.

- Таким образом, для каждого класса K строится семейство, состоящее из наиболее информативных для этого класса корректных наборов эл.кл., которое в дальнейшем используется для распознавания новых объектов.

- Для распознаваемого объекта $\boldsymbol{S}$ процедура голосования по корректному набору эл.кл. $\boldsymbol{U}$ заключается в следующем.

- Для каждой пары объектов $\boldsymbol{(S,\ S')}$, где $\boldsymbol{S'}$ - объект из обучающей подвыборки, принадлежащий классу $\boldsymbol{K}$, выписываются двоичные наборы $\boldsymbol{W(U,S)}$ и $\boldsymbol{W(U,S')}$ (отклики объектов $\boldsymbol{S}$ и $\boldsymbol{S'}$ на наборе $\boldsymbol{U}$). Набор $\boldsymbol{W(U,S')}$ сравнивается с набором $\boldsymbol{W(U,S)}$. Возможны случаи:

- 1) $\boldsymbol{W(U,S) = W(U,S')}$.

- 2) Набор $\boldsymbol{W(U,S')}$ предшествует набору $\boldsymbol{W(U,S)}$.

- 3) Набор $\boldsymbol{W(U,S')}$ следует за набором $\boldsymbol{W(U,S)}$.

- 4) $\boldsymbol{W(U,S) \neq W(U,S')}$ и наборы несравнимы.

- Если $\boldsymbol{U}$ – монотонный, то объект $\boldsymbol{S}$ получает голос за принадлежность к классу $\boldsymbol{K}$ в случаях 1) и 3). Если же $\boldsymbol{U}$ не является монотонным, то $\boldsymbol{S}$ получает голос за принадлежность к классу $\boldsymbol{K}$ только в случае 1).

- Корректность распознающего алгоритма обеспечивается только на базовой подвыборке и эта корректность обеспечивается корректностью каждого голосующего набора эл.кл.

- На практике проверено, что целесообразно использовать корректные наборы эл.кл. небольшой длины, в частности, тупиковые. Корректный набор эл.кл. называется тупиковым для класса K, если любое его собственное подмножество не является корректным для K. Решение прикладных задач показало, что монотонный логический корректор работает лучше логического корректора с произвольными корректными наборами эл.кл.

- На этапе построения семейств корректных наборов эл.кл. также, как и при построении логических процедур распознавания, основанных на голосовании по корректным эл.кл., приходится решать сложные дискретные задачи. Это задачи перечисления покрытий булевой матрицы. Каждый (монотонный) корректный набор эл.кл. для класса K однозначно соответствует покрытию булевой матрицы (расширенной матрицы сравнения), построенной специальным образом по прецедентной информации.

- **Случай 1.** (Построение монотонных корректных наборов эл.кл.).

- Выпишем всевозможные эл.кл., каждый из которых порождается хотя бы одним обучающим объектом класса $\boldsymbol{K}$. Пусть это будет множество эл.кл. $\boldsymbol{B_K} = \{\boldsymbol{P_1}, \dots, \boldsymbol{P_u}\}$. Фактически $\boldsymbol{B_K}$ - это множество всех подописаний обучающих объектов из $\boldsymbol{K}$.

- Паре обучающих объектов $\boldsymbol{S'}$ и $\boldsymbol{S''}$ таких, что $\boldsymbol{S'} \in \boldsymbol{K}$, $\boldsymbol{S''} \notin \boldsymbol{K}$ поставим в соответствие строку $\boldsymbol{C(S', S'')} = (\mathbf{c_1}, \dots, \mathbf{c_u})$, в которой $\mathbf{c_j} = \mathbf{1}$, если $\boldsymbol{S'}$ содержит $\boldsymbol{P_j}$, и $\mathbf{c_j} = \mathbf{0}$ в противном случае, $\boldsymbol{j} = \mathbf{1}, \mathbf{2}, \dots, \boldsymbol{u}$. Составим булеву матрицу $\boldsymbol{L_K}$ из всех строк $\boldsymbol{C(S', S'')}$ таких, что $\boldsymbol{S'} \in \boldsymbol{K}$, $\boldsymbol{S''} \notin \boldsymbol{K}$. По построению каждому столбцу в $\boldsymbol{L_K}$ соответствует некоторый эл.кл. из $\boldsymbol{B_K}$.

- Нетрудно видеть, что набор эл.кл. является монотонным (тупиковым) корректным набором для $\boldsymbol{K}$ тогда и только тогда, когда соответствующий набор столбцов матрицы $\boldsymbol{L_K}$ является (тупиковым) покрытием.

- **Случай 2.** (Построение немонотонных корректных наборов эл.кл.).

- В этом случае множество $\boldsymbol{D_K}$ всех рассматриваемых эл.кл. для класса K образуется как подописаниями объектов из K, так и подописаниями объектов из других классов. Пусть $\boldsymbol{D_K = \{Q_1, \dots, Q_t\}}$.

- Паре обучающих объектов $\boldsymbol{S'}$ и $\boldsymbol{S''}$ поставим в соответствие строку $\boldsymbol{D(S', S'') = (d_1, \dots, d_t)}$, в которой $\boldsymbol{d_j = 1}$, если один из рассматриваемых объектов содержит $\boldsymbol{Q_j}$, а другой не содержит, и $\boldsymbol{d_j = 0}$ в противном случае, $\boldsymbol{j = 1, 2, \dots, t}$. Составим булеву матрицу $\boldsymbol{L_K}$ из всех строк $\boldsymbol{D(S', S'')}$ таких, что $\boldsymbol{S' \in K,\ S'' \notin K}$. Дальнейшие рассуждения те же, что и в случае 1.

- Как правило, даже для задач небольшой размерности число эл.кл. велико и процедура построения корректных наборов эл.кл. с использованием матрицы $\boldsymbol{L_K}$ требует значительных вычислительных ресурсов. В простейших моделях логических корректорах для снижения вычислительной сложности используются эл.кл. ранга 1. При этом среди корректных наборов с эл.кл., имеющими ранг отличный от 1, могут оказаться наиболее информативные.

- В связи с этим для рассматриваемой задачи представляет интерес разработка не только точных, но и приближённых алгоритмов, что может оказаться необходимым для увеличения скорости её решения. С этой целью разрабатываются более сложные модели, в которых используются эл.кл. произвольного ранга и для сокращения временных затрат строятся так называемые локальные базисы классов.

• Локальному базису соответствует покрытие матрицы $\boldsymbol{L_K}$ достаточно большой мощности, которое используется для построения семейства корректных наборов эл.кл. Локальные базисы строятся либо итеративным образом с использованием бустинга, либо стохастическим образом (обрабатываются случайные подматрицы матрицы $\boldsymbol{L_K}$).

• Последние исследования в данной области посвящены вопросам разработки общих подходов к синтезу корректных распознающих алгоритмов с использованием логических и алгебраических методов анализа данных и построения на этой основе новых более совершенных моделей логических корректоров.

• Исследованы вопросы повышения качества решения прикладных задач (на основе расширения класса корректирующих функций) и вопросы сокращения временных затрат (на основе применения новейших технологий к организации вычислений). Полученные результаты свидетельствуют об эффективности новых подходов к обработке больших данных методами алгебро-логического анализа.

- Кроме того решена важная в методологическом плане задача создания общей схемы синтеза корректных логических процедур классификации. В рамках предложенной схемы описаны классические логические алгоритмы распознавания и ранее построенные логические корректоры. При построении логического корректора общего вида в качестве корректирующих функций предложены функции более широкого класса, чем ранее используемые, а именно поляризуемые булевы функции. Булева функция называется поляризуемой, если она монотонно возрастает или убывает по каждой отдельной переменной. .

- Наиболее полно основные результаты по разработке практических моделей логических корректоров и создания общей схемы синтеза корректных логических процедур классификации представлены в статье:

- *Дюкова Е.В., Журавлёв Ю.И., Прокофьев П.А.* Логические корректоры в задаче классификации по прецедентам // Ж. вычисл. матем. и матем. физ. 2017. Т. 57. № 11. С. 1906–1927 .

УПРАЖНЕНИЯ

1. Пусть обучающая выборка из шести объектов разбита на два класса, равных по мощности. Классы представлены множествами объектов:

{(0, 1, 0, 0), (0, 0, 1, 0), (1, 0, 0, 0)} и {(1, 1, 0, 0), (0, 1, 1, 0), (0, 1, 1, 1)}.

Для каждого класса построить все корректные наборы из эл.кл. ранга 1 и все монотонные корректные наборы из эл.кл. ранга 1.