Алгоритмы обучения многоклассовому распознаванию образов по большим массивам данных

Маленичев А.А., Моттль В.В., Середин О.С., Красоткина О.В.

ФГБОУ ВПО Тульский государственный университет

Херсониссос, 2014

CAUTION! BIG DATA!

Метод k-ближайших соседей (k-Nearest Neighbor algorithm)

+ Достоинства:

  Простота, легкая масштабируемость

- Недостатки:

  Использует всю выборку для обучения

Ссылки:

1 P. Jain, A. Kapoor. Active Learning for Large Multi-class Problems. Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

Метод опорных векторов (SVM for Active Learning)

\+ Достоинства:

В процессе обучения часть объектов может быть вообще выброшена из рассмотрения (так называемые неопорные объекты);

\- Недостатки:

Высокая вычислительная сложность: двухклассовый SVM в общем виде имеет алгоритмическую сложность в худшем случае $O(N^3)$, в среднем — $O(N^2)$, в лучшем - $O(N \ln(N))$.

Ссылки:

1 E. Y. Chang, S. Tong, K. Goh, and C. Chang. Support vector machine conceptdependent active learning for image retrieval. IEEE Transactions on Multimedia, 2005.;

2 S. Tong and D. Koller. Support vector machine active learning with applications to text classification. In ICML, 2000.

Постановка задачи многоклассовой классификации для больших массивов данных

- Предположим, что существует множество объектов реального мира Ω.
- Каждый объект $\omega \in \Omega$ может быть охарактеризован числом $y \in \{0, 1, \ldots, m-1\}$, где $m > 2$, и вектором $\mathbf{x} \in R^n$, $n \geq 1$, так называемые класс и вектор признаков соответственно.
- Задача многоклассовой классификации заключается в том, чтобы построить некоторую решающую функцию $f(\mathbf{x})$, которая по вектору $\mathbf{x}_i$ некоторого объекта ω_i будет возвращать его y_i и, говоря нестрогим языком, будет «как можно реже ошибаться».
- Мы будем рассматривать задачу обучения с учителем: предположим, что существует некоторая совокупность объектов $\bar{\Omega} \subset \Omega$, для которых известны как векторы $\mathbf{x}$, так и классы y. Размер такой совокупности будет составлять N
- Для ясности будем считать, что задача анализа больших массивов данных подразумевает $N \geq 10000$.

Алгоритм решения поставленной задачи

- Решение задачи двухклассовой классификации для больших массивов данных
- Вычисление функции степени достоверности принадлежности объекта к классу для каждой пары классов
- Согласование полученных функций степени достоверности и вычисление вектора принадлежности объекта к одному из m классов

Алгоритм двухклассовой классификации для больших массивов данных: модель логистической регрессии

Для разработки алгоритма многоклассовой классификации разработаем сначала алгоритм двухклассовой классификации: в случае двухклассовой классификации $y_i \in \{-1, 1\}$. В качестве основы возьмём алгоритм логистической регрессии.

Примем следующую модель данных: предположим, что объект ω_i относятся к классу y_i с вероятностью

$$Lf_y(\mathbf{x}_i, y_i, \mathbf{a}, b, \gamma) = \frac{1}{1 + \exp[-\gamma y_i(\mathbf{a}^T \mathbf{x}_i + b)]},$$

где $\mathbf{a}, b$ — коэффициенты разделяющей гиперплоскости, γ — параметр, характеризующий меру пересечения объектов первого и минус первого класса.

Модель логистической регрессии: переход в расширенное пространство признаков

Переведём объекты в расширенное пространство признаков при помощи следующей замены

$$\mathbf{z}_j = \begin{pmatrix} \mathbf{x}_j \\ 1 \end{pmatrix}, \mathbf{c} = \begin{pmatrix} \mathbf{a} \\ b \end{pmatrix}.$$

Тогда вероятность отнесения объекта ω_i к классу y_i запишется в следующем виде:

$$Lf_y(\mathbf{z}_i, y_i, \mathbf{c}, \gamma) = \frac{1}{1 + \exp[-\gamma y_i \mathbf{c}^T \mathbf{z}_i]}.$$

Критерий обучения

Задача: требуется найти такой вектор $\hat{\mathbf{c}}$, чтобы вероятность того, что все объекты обучающей совокупности классифицируются правильно, была максимальна; вероятность корректной классификации всех объектов составляет произведение вероятностей корректной классификации каждого объекта.

$$\hat{\mathbf{c}} = \arg\max_{\mathbf{c}} \left(\prod_{j=1}^{N} \frac{1}{1 + \exp(-\gamma y_j \mathbf{c}^T \mathbf{z}_j)} \right)$$

После преобразования отсюда можно получить критерий

$$J(\mathbf{Z}, \mathbf{y}, \mathbf{c}, \gamma) = \sum_{j=1}^{N} \ln \left[1 + \exp(-\gamma y_j \mathbf{c}^T \mathbf{z}_j) \right] \to \min$$

Процедура оптимизации критерия обучения

Оптимизация ведётся по методу Ньютона. Каждое следующее приближение $\mathbf{c}^{k+1}$ получается из предыдущего путём решения системы линейных алгебраических уравнений

$$\mathbf{c}^{k+1} = -\left[\nabla^2_{\mathbf{cc}} J(\mathbf{Z}, \mathbf{y}, \mathbf{c}^k, \gamma) + \beta \mathbf{I}\right]^{-1} \nabla_{\mathbf{c}} J(\mathbf{Z}, \mathbf{y}, \mathbf{c}^k, \gamma) + \mathbf{c}^k,$$

где

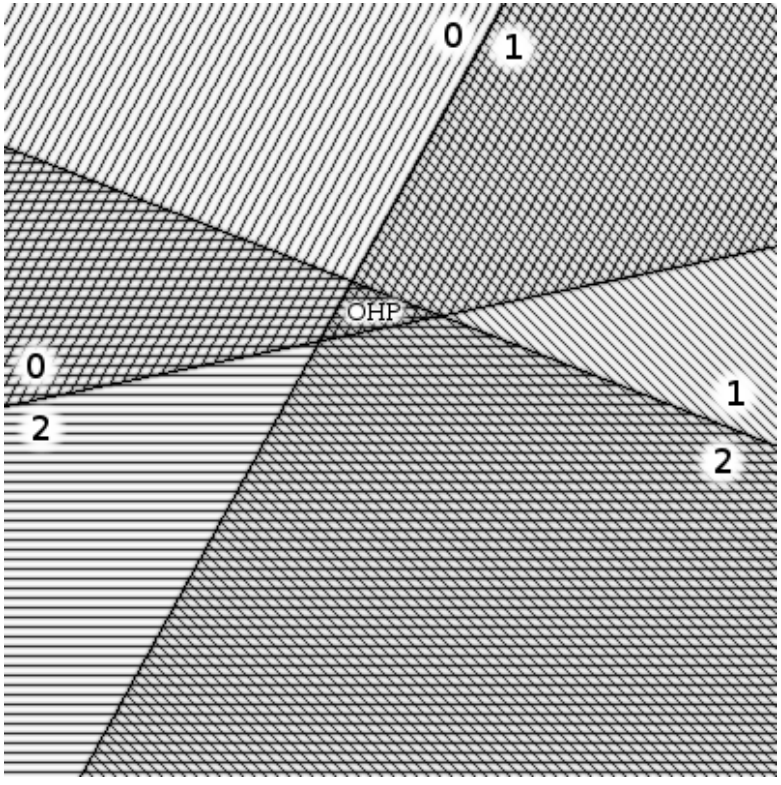
$$\nabla_{\mathbf{c}} J(\mathbf{Z}, \mathbf{y}, \mathbf{c}^k, \gamma) = -\gamma \sum_{j=1}^{N} \frac{y_j}{1 + \exp\left(\gamma y_j (\mathbf{c}^k)^T \mathbf{z}_j\right)} \mathbf{z}_j$$

$$\nabla^2_{\mathbf{cc}} J(\mathbf{Z}, \mathbf{y}, \mathbf{c}^k, \gamma) = \gamma^2 \sum_{j=1}^{N} \frac{\exp\left(\gamma y_j (\mathbf{c}^k)^T \mathbf{z}_j\right)}{\left[1 + \exp\left(\gamma y_j (\mathbf{c}^k)^T \mathbf{z}_j\right)\right]^2} \mathbf{z}_j \mathbf{z}_j^T$$

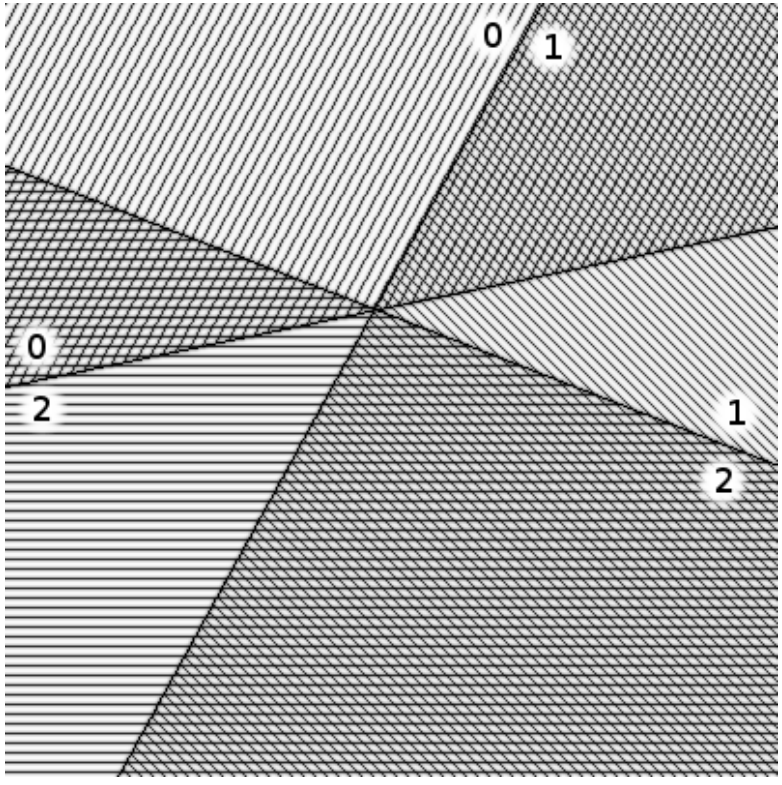
Идея алгоритма многоклассовой классификации

- Идея: если существует алгоритм, позволяющий произвести классификацию в каждой паре классов и построить своё решающее правило между ними, то тогда для любого нового поступившего объекта немедленно находится доминирующий класс в каждой из пар, по нему находится общий доминирующий класс.
- Однако иногда выбрать доминирующий класс не представляется возможным: к примеру, в случае трёх классов правила могут войти в так называемый парадокс Кондорсе: по предпочтительности классы будут иметь вид $0 \prec 1$, $1 \prec 2$, $2 \prec 0$ и найти наиболее предпочтительный класс не представляется возможным. Область пространства признаков, где наблюдается такая несовместность, называется областью непринятия решений (ОНР).
- Но в случае расширенного пространства признаков парадокса Кондорсе возникнуть не может, поскольку все линейные гиперплоскости проходят через начало координат.

Область непринятия решений



Наличие ОНР в стандартном пространстве признаков



Отсутствие ОНР в расширенном пространстве признаков

Вычисление вероятности отнесения объекта к одному из m классов

Представленный выше алгоритм двухклассовой классификации выгоден тем, что он может оценивать не только класс нового объекта, но и вероятность его отнесения к данному классу. Вероятность отнесения объекта с расширенным вектором признаков $\mathbf{z}_i$ к классу y_i составляет

$$Lf_y(\mathbf{z}_i, y_i, \tilde{\mathbf{c}}, \tilde{\gamma}) = \frac{1}{1 + \exp[-\tilde{\gamma} y_i \tilde{\mathbf{c}}^T \mathbf{z}_i]}.$$

Это позволит в качестве результата обучения не только указать наиболее предпочтительный класс, но и вернуть вектор вероятностей отнесения объекта к каждому из классов $\boldsymbol{\pi}(\mathbf{z})$. Очевидно, что

$$\sum_{k=1}^{m} \pi_k(\mathbf{z}) = 1.$$

Процедура обучения многоклассовому распознаванию образов

Из всей обучающей совокупности выберем все объекты классов k и l.
После обучения каждому новому объекту каждое попарное правило сравнения поставит в соответствие вероятность P_1^{kl} — вероятность отнесения объекта к классу k в пределах пары классов (k, l).
Тогда вероятности принадлежности объекта $\mathbf{z}$ к одному из m классов могут быть вычислены следующим образом:

$$\pi^k(\mathbf{z}) = \left(\sum_{l=1}^{k-1} \frac{1 - P_1^{lk}(\mathbf{z})}{P_1^{lk}(\mathbf{z})} + 1 + \sum_{l=k+1}^{m} \frac{1 - P_1^{kl}(\mathbf{z})}{P_1^{kl}(\mathbf{z})} \right)^{-1}$$

Нормировка вероятностей

Вообще, строго говоря, это верно не всегда, поскольку попарные вероятности $P_1^{kl}(\mathbf{z})$ получены независимо друг от друга, плюс ко всему являются лишь оценками истинных вероятностей: их совокупность может оказаться несовместной. Однако опыт показывает, что если несовместность и наблюдается, она присутствует лишь в очень малых областях пространства признаков. Однако в случае, если несовместность всё-таки наблюдается, ищется наилучшая аппроксимация решения. Можно показать, что такой наилучшей аппроксимацией будет нормировка вероятностей к единице:

$$\hat{\pi}^k(\mathbf{z}) = \frac{\pi^k(\mathbf{z})}{\sum_{k=1}^{m} \pi^k(\mathbf{z})}$$

Оценка вычислительной сложности предложенного алгоритма

Произведена оценка сложности алгоритма многоклассовой классификации в целом: при размере обучающей выборки в N объектов, при количестве признаков n и количестве классов m вычислительная сложность по временному ресурсу занимает

$$O(N, n, m) = N \cdot \ln(N) \cdot n^2 \cdot m,$$

по памяти —

$$O(N, n, m) = N \cdot n \cdot m^{-1}.$$

Структура модельных данных

Для первоначальной оценки корректности работы программы исходная выборка объектов была сгенерирована следующим образом: N объектов с $n = 2$ признаками (размерность расширенного признакового пространства равна 3) и $m = 5$ различными классами. Каждый объект соответствующего класса — реализация случайного события, распределённого по нормальному закону со среднеквадратическим отклонением σ и математическим ожиданием $(0, 0)$ для нулевого класса, $(0, 2)$ для первого, $(2, 0)$ для второго, $(0, -2)$ для третьего и $(-2, 0)$ для четвёртого.

Реальные данные

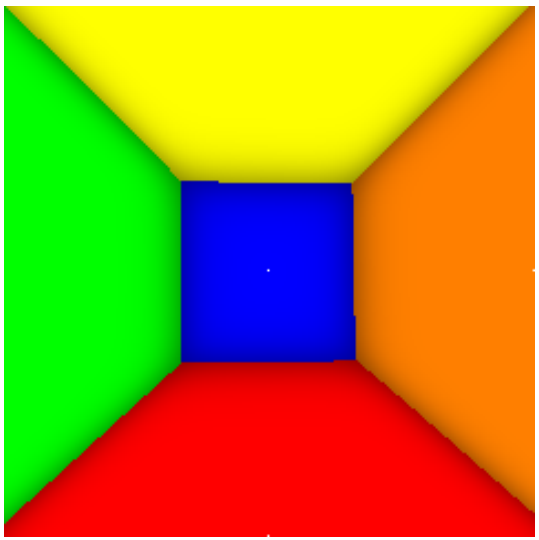
В качестве экспериментальной выборки для проверки корректности работы двухклассовой классификации была выбрана база данных `http://archive.ics.uci.edu/ml/datasets/SUSY`. Это база данных физического эксперимента, нацеленного на решение проблемы классификации процессов, порождающих суперсимметричные частицы, от процессов, не порождающих их.

Сводная таблица результатов экспериментов

Выборка	Клас-сов	Кол-во объектов	Приз-наков	Время обучения	Процент класс-ии	Ожидае-мый %
Модельная	5	50000	2	мало	72.18	72.35
Модельная	5	50000	2	мало	72.19	72.35
Модельная	5	50000	2	мало	71.77	72.35
Модельная	5	35000000	2	7м 2с	72.21	72.35
SUSY	2	4500000	18	3м 46с	79	—
PUC-rio	5	165500	17	43с	90	—

Результаты экспериментального исследования для модельных данных

Тестовая выборка — совокупность точек, равномерно распределённых в интервале $(-3, 3)$ по обеим осям. При $N = 50000$ и $\sigma = 1$ тестовая выборка была классифицирована следующим образом (разными цветами обозначены разные классы):



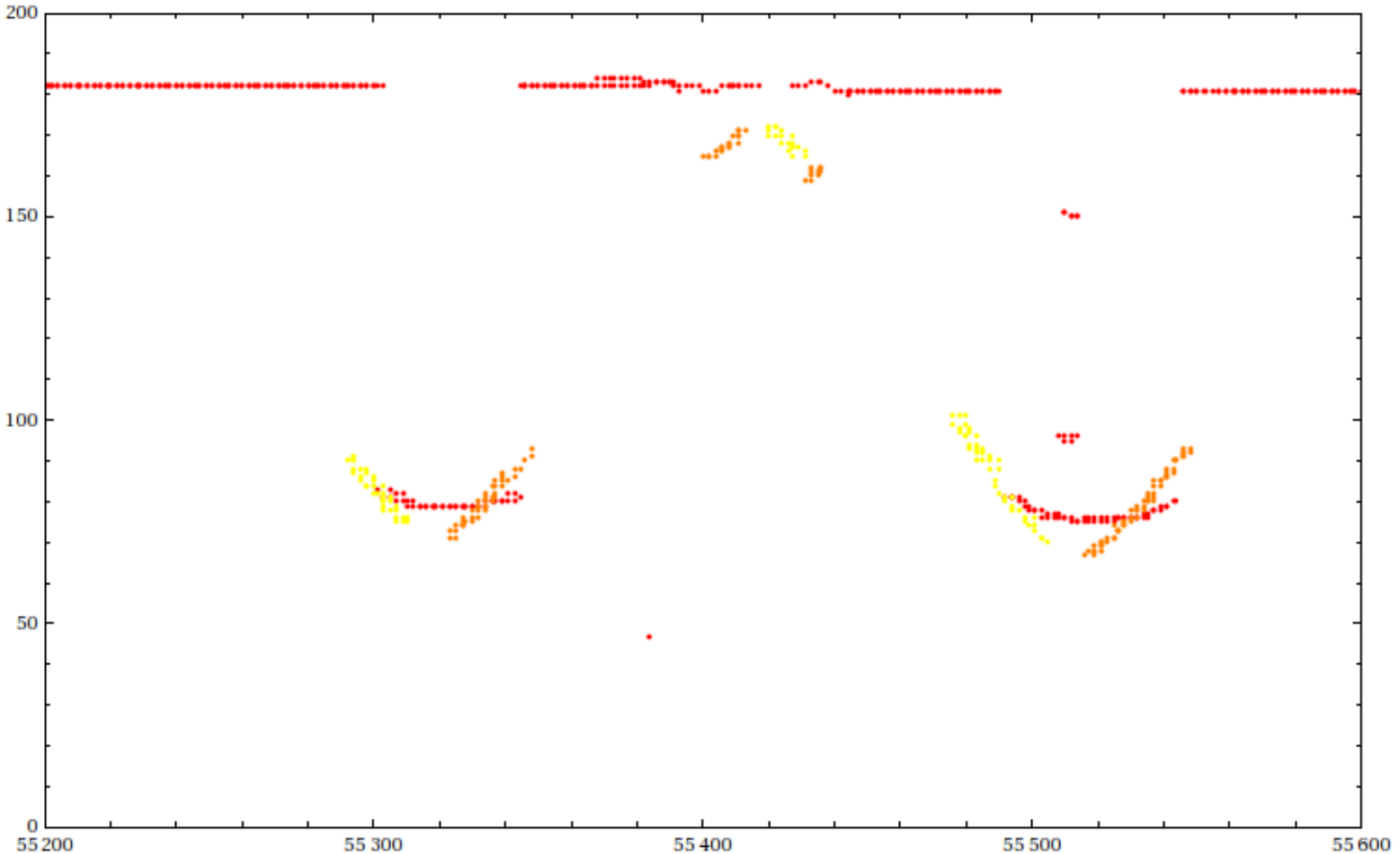
Классификация тестовой выборки

Практическое применение — поиск стыков рельсового пути по дефектограмме

В качестве альтернативы путевым обходчикам при задаче анализа рельсового пути на дефекты предложен способ ультразвуковой дефектоскопии. В качестве идеи данного метода положено следующее рассуждение: при создании ультразвуковой волны на поверхности рельса, направленной строго (или нестрого) вниз, она отразится от встреченной ей особенности (в том числе и дефекта) и снова вернётся к поверхности рельса, где может быть уловлена датчиком.

В данной части работы будет рассматриваться только классификация выборки на три класса: в данной точке дефектограммы присутствует стык, присутствует болтовое отверстие или этих особенностей не имеется.

Пример участка ультразвуковой дефектограммы в области стыка рельсов



Результаты эксперимента

В качестве обучающей была использована выборка небольшого размера $N = 935$ объектов с $n = 769$ признаками. Результат классификации — 80% тестовой выборки было классифицировано корректно.

Для полного решения задачи распознавания стыков рельсового пути требуется после классификации добавить механизм марковской цепи, а также механизм совмещения точек рельсового пути, полученных после прогонки дефектоскопа по одному и тому же участку пути несколько раз.

Направление дальнейших исследований

- Применение метода стохастической аппроксимации для задач классификации больших массивов данных
- Применение марковской цепи для улучшения качества определения стыков на ультразвуковой дефектограмме