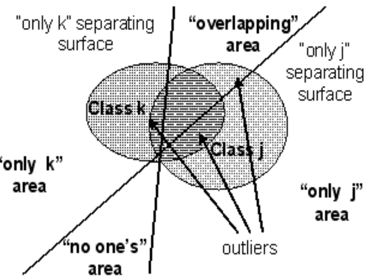
Метод многотемной (multi-label) классификации на основе попарных сравнений с отсечением наименее релевантных классов

Петровский М. И., Глазкова В. В.
Москва, МГУ им. М. В. Ломоносова

Расширением традиционной постановки задачи классификации с несколькими классами (multi-class) является постановка задачи *много-темной* (multi-label) классификации, в которой классифицируемый объект может принадлежать нескольким классам одновременно, и сами классы являются не взаимоисключающими (возможно, даже вложенными). Задачи такого типа возникают при анализе, рубрикации и классификации текстов, изображений, цепочек ДНК, и т. д. В случае multi-label классификации в исходной обучающей совокупности $S = (x_i, y_i)_{i=1}^n$ для каждого примера x_i задан не единственный класс, а множество релевантных классов $y_i \subset \{1, \ldots, q\}$, и целью алгоритма машинного обучения является построение классификатора $f_S \colon X \to 2^q$, предсказывающего все релевантные классы, где X — исходное пространство признаков, а q — число классов. Традиционным подходом к решению таких задач является декомпозиция типа «*каждый против остальных*» исходной multi-label задачи в q задач бинарной классификации с целью независимо определить, является ли каждый из q классов релевантным x. Для каждого класса l формируется тренировочный набор, где x_i помечен как «положительный», если в S выполнялось $l \in y_i$, иначе x_i помечен как «отрицательный». Далее строится q бинарных классификаторов $f_S^l \colon X \to \{0, 1\}$. Этот подход критикуется за низкую точность, поскольку не учитываются корреляции между классами, и за ресурсоемкость, поскольку при обучении необходимо решать q задач размера n вместо одной.

В настоящей работе исследуется возможность использования для решения задачи multi-label классификации подхода на основе декомпозиции типа «*каждый против каждого*». Предлагается новый алгоритм, основанный на модифицированном *методе попарных сравнений* с помощью набора *бинарных классификаторов*. Следует отметить, что ранее этот подход не давал позитивных результатов для задач multi-label классификации, поскольку не удавалось построить точный классификатор, разделяющий два существенно перекрывающихся класса. Поэтому в нашем методе каждая пара возможно классов j и k разделяется с помощью двух бинарных классификаторов. Используя их, можно выделить четыре области: область «только класса j»; область «только класса k»; «перекрывающаяся область» (j и k); область, не принадлежащую «ни классу j, ни классу k»:



Для каждой пары классов j и k формулируется две задачи обучения, которые включают только примеры, помеченные в S либо j, либо k, либо обоими классами одновременно (число таких примеров, как правило, существенно меньше n). В первой подзадаче примеры, помеченные только классом k, рассматриваются как «положительные», все остальные — как «отрицательные». В результате, построенный классификатор предсказывает вероятность: $r^+_{kj}(x) = P(k \in f_S(x) \wedge j \notin f_S(x) | x \in k \cup j)$ и дополнительную вероятность: $r^-_{kj}(x) = 1 - r^+_{kj}(x) = P(j \in f_S(x) | x \in k \cup j)$. Вторая подзадача формулируется и решается аналогично, но для класса j: $r^+_{jk}(x) = P(j \in f_S(x) \wedge k \notin f_S(x) | x \in k \cup j)$ и $r^-_{jk}(x) = P(k \in f_S(x) | x \in k \cup j)$. Вероятности принадлежности x каждой из областей вычисляются так:

$$\begin{aligned} P(x \in \text{ overlapping}\, k, j) &= r^-_{kj}(x)\, r^-_{jk}(x); \\ P(x \in \text{ only}\, k) &= r^+_{kj}(x)\, r^-_{jk}(x); \\ P(x \in \text{ no one's}\, k, j) &= r^+_{kj}(x)\, r^+_{jk}(x); \\ P(x \in \text{ only } j) &= r^-_{kj}(x)\, r^+_{jk}(x). \end{aligned}$$

Важно отметить, что при такой формулировке оба бинарных классификатора разделяют взаимно исключающие суперклассы, и для решения и оценки попарных вероятностей могут быть использованы стандартные алгоритмы бинарной классификации, например, на основе Support Vector Machines или Kernel Fisher Discriminant. Формулируя и решая таким образом $q(q-1)$ задач бинарной классификации, каждая размера меньшего, чем n, мы получаем попарные вероятности сравнения $r^*_{jk}(x)$.

Далее необходимо, используя результаты попарных сравнений, оценить вероятности принадлежности $p_l(x)$ каждому из q невзаимоисключающих классов. Для этого мы предлагаем использовать обобщённую *модель ранжирования Бредли-Терри с «ничьёй»*, которую сформулировали Рао и Купер (1967). Учитывая, что в нашем случае вероятность «ничьи»

$$P(j \text{ ties } k) = P(x \in \text{ overlapping}\, k, j) + P(x \in \text{ no one's } k, j),$$

$P(k\,\text{beats}\,j) = P(x \in \text{only } k)$, $P(j\,\text{beats}\,k) = P(x \in \text{only } j)$, получаем оптимизационную задачу:

$$\min_{\bar{p},\theta} l(\bar{p},\theta), \quad p_k \geqslant 0;$$

$$l(\bar{p},\theta) = -\frac{1}{2}\sum_k\sum_j\biggl[2r^+_{kj}r^-_{jk}\ln\frac{p_k}{p_k+\theta p_j} + \\ + (r^-_{kj}r^-_{jk} + r^+_{kj}r^+_{jk})\ln\frac{(\theta^2-1)p_jp_k}{(\theta p_k+p_j)(p_k+\theta p_j)}\biggr].$$

Для решения данной задачи можно использовать итеративный minorization-maximization алгоритм [1], детально описанный для нашего случая в [2].

Решив сформулированную задачу оптимизации, мы получаем оценки релевантности классов $p_l(x)$, и теперь должны выбрать наиболее релевантные классы в качестве решения исходной multi-label задачи: $\{l | p_l(x) > t, l \in \{1,\ldots,q\}\}$. Порог отсечения t обычно зависит от классифицируемых объектов, поэтому возникает необходимость построения пороговой функции $t(x)$. Для этих целей используется подход, предложенный нами в [3] для сведения задачи ранжирования к задаче multi-label классификации. В существующих методах сложные пороговые функции строятся в пространстве признаков X, в нашем подходе предлагается строить простые линейные пороговые функции в пространстве релевантностей классов, используя результат работы алгоритмов ранжирования как новое множество признаков анализируемого объекта:

$$t(\bar{p}(x)) = \sum_{j=1}^{q} a_j p_j(x_i) + a_0.$$

Коэффициенты a_j определяются методом наименьших квадратов на примерах из обучающей совокупности. Детально подход описан в [3].

Все разработанные алгоритмы были экспериментально проверены на эталонных тестовых наборах данных Yeast 2K и Reuters-2000, где показали высокие результаты как по точности классификации так и скорости обучения [2, 3].

Работа выполнена при поддержке РФФИ, проект № 06-01-00691, гранта Президента РФ МК-4264.2007.9, а также в рамках госконтракта с Федеральным агенством по науке и инновациям № 02.514.11.4026.

Литература

[1] *Hunter D. R.* MM algorithms for generalized Bradley-Terry models. — Annals of Statistics, 32(1), 2004. — pp. 384–406.

[2] *Petrovskiy M.* Paired Comparisons Method for Solving Multi-label Learning Problem. — Hybrid Intelligent Systems, IEEE Press, 2006. — pp. 42–48.

[3] *Petrovskiy M., Glazkova V.* Linear Methods for Reduction from Ranking to Multilabel Classification. — Springer-Verlag, 2006. — LNAI, vol. 4304 — pp. 1152–1156.