Поиск оптимальной метрики в задачах классификации с порядковыми признаками

Иофина Г. В., Кропотов Д. А.
giofina@gmail.com, dkropotov@yandex.ru
Москва, МФТИ, ВЦ РАН

Рассматривается класс задач распознавания, в которых объекты описываются n порядковыми признаками, представленными набором целых чисел от 0 до $N-1$. Для решения таких задач предполагается применять метрические алгоритмы классификации, такие как метод ближайших соседей, метод потенциальных функций или алгоритмы вычисления оценок [1]. Для этого необходимо вводить метрики как на отдельных признаках, так и на векторах признаковых описаний. В работе находится наилучшая метрика, при которой взвешенная разность между межклассовым и средним внутриклассовым расстояниями максимальна. Таким образом, найденная метрика отображает структуру задачи, т. е. дает маленькие расстояния для объектов из одного класса и большие — для объектов из разных классов.

Пусть на множестве $\tilde{N} = \{0, 1, \ldots, N-1\}$ допустимых значений признаков задано естественное отношение порядка $0 \leqslant 1 \leqslant \ldots \leqslant N-1$. Произвольная метрика $\rho(i, j)$ на множестве $\tilde{N}$ задается матрицей $\{c_{ij}\}$ размера $N \times N$, симметричной, с нулевой диагональю, элементы которой, находящиеся выше главной диагонали, не убывают по строкам и не возрастают по столбцам (удовлетворяют отношению порядка).

Пусть даны два конечных множества объектов:
$a_u = (a_u^1, \ldots, a_u^n) \in \tilde{N}^n$, $u = 1, \ldots, m$ — объекты класса K_1;
$b_v = (b_v^1, \ldots, b_v^n) \in \tilde{N}^n$, $v = 1, \ldots, l$ — объекты класса K_2.

Под расстоянием между объектами $a_u \in K_1$ и $b_v \in K_2$ будем понимать величину $\rho(a_u, b_v) = \sum\limits_{i=1}^{n} \rho_i(a_u^i, b_v^i)$, где ρ_i — метрика, заданная на i-ом признаке.

Под внутриклассовыми и межклассовым расстояниями будем понимать величины

$$\alpha_1 = 1/N_1 \sum_{u=1}^{m} \sum_{v=1}^{m} \rho(x_u, x_v), \quad x_u, x_v \in K_1;$$

$$\alpha_2 = 1/N_2 \sum_{u=1}^{l} \sum_{v=1}^{l} \rho(x_u, x_v), \quad x_u, x_v \in K_2;$$

$$\beta = 1/M \sum_{u=1}^{m} \sum_{v=1}^{l} \rho(x_u, x_v), \quad x_u \in K_1, x_v \in K_2;$$

где N_1, N_2 и M — нормировочные множители.

Задачу максимизации взвешенной разности между межклассовым и средним внутриклассовым расстояниями можно представить в виде следующей оптимизационной задачи:

$$\beta - 0.5\lambda(\alpha_1 + \alpha_2) \to \max_{\rho_i, i=1,\ldots,n}, \tag{1}$$

где λ можно рассматривать как отношения весов межклассового и среднего внутриклассового расстояний.

Считается, что на каждом признаке задана своя функция расстояния, и признаки не зависят друг от друга, поэтому функции расстояний для разных признаков можно искать независимо. Далее в работе рассматривается один признак и, следовательно, ищется одна функция расстояния.

Обозначим количество нулей, единиц, двоек, троек, и т. д. среди значений признака у объектов первого класса через $\xi_0, \xi_1, \ldots, \xi_{N-1}$, а у объектов второго класса — через $\eta_0, \eta_1, \ldots, \eta_{N-1}$. При попарном сравнении объектов из класса K_1 количество сравнимаемых пар (i, j) можно представить в виде матрицы внутриклассовых расстояний первого класса $A_1 = \{a_1^{ij}\}$, где $a_1^{ij} = \xi_i\xi_j,\ i, j = 0, \ldots, N-1,\ i \neq j;\ a_1^{ii} = \xi_i(\xi_i - 1)/2,$ $i = 0, \ldots, N-1$. Для объектов из второго класса матрица внутриклассовых расстояний $A_2 = \{a_2^{ij}\}$, где $a_2^{ij} = \eta_i\eta_j,\ i, j = 0, \ldots, N-1,\ i \neq j;$ $a_2^{ii} = \eta_i(\eta_i - 1)/2,\ i = 0, \ldots, N-1$. Аналогично, матрица межклассовых расстояний $B = \{b^{ij}\}$, где $b^{ij} = \eta_i\xi_j + \xi_i\eta_j,\ i, j = 0, \ldots, N-1,\ i \neq j;$ $b^{ii} = \eta_i\xi_i,\ i = 0, \ldots, N-1$.

Матрица расстояний $\{c_{ij}\}$ в пространстве $\tilde{N}$ определяется $N(N-1)/2$ числами. Поэтому её можно представить вектором $x = (x_1, \ldots, x_{N(N-1)/2})$, где $x_k = c_{ij}$, $k = (2N - 1 - i)i/2 + j - i,\ i \leqslant j$.

Критерий оптимизации задачи (1) запишется в виде следующей задачи целочисленного линейного программирования [2]:

$$\begin{cases} \sum\limits_{k=1}^{N(N-1)/2} \gamma_k x_k \to \max\limits_{\{x_k\}}; \\ 1 \leqslant x_k \leqslant N-1, \quad k = 1, \ldots, N(N-1)/2; \\ x_k \text{— целые и удовлетворяют отношению порядка;} \end{cases} \tag{2}$$

где $\gamma_k = \frac{1}{M}\left(\xi_i\eta_j + \eta_i\xi_j\right) - \frac{0.5\lambda}{N_1}\xi_i\xi_j - \frac{0.5\lambda}{N_2}\eta_i\eta_j,\ k = (2N - 1 - i)i/2 + j - i,$ $i \leqslant j$ — соответствующие коэффициенты.

Следующая теорема сильно упрощает решение задачи.

Теорема 1. *Решением оптимизационной задачи*

$$\begin{cases} \sum\limits_{k=1}^{N(N-1)/2} \gamma_k x_k \to \max\limits_{\{x_k\}}; \\ X_{\min} \leqslant x_k \leqslant X_{\max}, \quad k = 1, \ldots, N(N-1)/2; \\ x_k \text{ — действительные и удовлетворяют отношению порядка;} \end{cases} \tag{3}$$

могут являться только векторы $b = (b_1, \ldots, b_{\frac{N(N-1)}{2}})$, *в которых* $b_k = X_{\min}$ *или* $b_k = X_{\max}$, $k = 1, \ldots, \frac{N(N-1)}{2}$.

Теорема верна и для целочисленных x_k, $k = 1, \ldots, N(N-1)/2$. Верно и обратное — решение задачи целочисленного линейного программирования будет также решением задачи линейного программирования в непрерывном случае с теми же ограничениями на x_k. Поэтому вначале можно решить задачу линейного программирования в бинарном случае, а затем элементарным линейным преобразованием координат получить решение исходной задачи.

Теорема 2. *Число матриц расстояний размерности* N *можно представить следующей рекуррентной формулой:*

$$f(N) = \sum_{i=0}^{N-1} f(i) f(N-1-i);$$
$$f(0) = 1, \quad f(1) = 1.$$

Из данной рекуррентной формулы можно получить явное выражение для числа допустимых матриц $f(N) = \frac{C_{2N}^N}{N+1}$ [3].

Так как алгоритмы решения задач целочисленного линейного программирования имеют сложность $O(2^{N^2})$ в худшем случае, теорема (2) позволяет уменьшить сложность задачи по крайней мере до $O(2^{(NlogN)/2})$, даже если решать задачу простым перебором допустимых метрик.

Работа выполнена при поддержке РФФИ, проект № 05-01-00332.

Литература

[1] *Журавлев Ю. И.* Избранные научные труды — М.: Магистр, 1998 — 420 с.

[2] *Галеев Э. М., Тихомиров В. М.* Краткий курс теории экстремальных задач. — М.: Изд. Московского университета, 1989. — 204 с.

[3] *Садовничий В., Григорьян А., Конягин С.* Задачи студенческих математических олимпиад МГУ. — М.: Изд. Московского университета, 1987. — 310 с.

[4] *Иофина Г. В.* Выбор наилучшей метрики в алгоритме распознавания по ближайшему соседу // Труды 49-й научной конференции МФТИ, Москва-Долгопрудный, 2006 — С. 266–267.