Применение обобщенного метода наименьших квадратов к задаче построения разделяющей гиперплоскости

Матросов В. Л., Горелик В. А., Жданов С. А., Муравьева О. В.
gorelik@ccas.ru, saj@mpgu.edu.ru, muraveva@mpgu.edu.ru
Москва, МПГУ

Пусть задан набор объектов в n-мерном пространстве признаков. Требуется определить разрешающее правило, разделяющее множество точек на два класса. На практике признаки объектов меряются приближенно, поэтому границы между классами (выборками) имеют весьма причудливую форму и даже размыты. Вместе с тем, желательно, чтобы решающие правила были попроще. Особенно удобны линейные правила, когда границы между классами представляют собой гиперплоскости. Покажем, как можно использовать методы коррекции данных для построения разделяющих гиперплоскостей в пространстве признаков.

Класс K_1 представлен объектами с векторами признаков $x^1, \ldots, x^k$, класс K_2 — выборкой $y^1, \ldots, y^l$, $m = k + l$. Требуется найти коэффициенты линейной решающей функции $F(x) = (a, x) - b$, т. е. найти коэффициенты $a \in \mathbb{R}^n$ и $b \in \mathbb{R}$ такие, что выполняется система

$$\begin{cases} (a, x^i) \leqslant b, \ i = 1, \ldots, k; \\ (a, y^j) \geqslant b, \ j = 1, \ldots, l. \end{cases}$$

Полагая, что эта система неравенств относительно a, b несовместна, рассмотрим задачу минимальной коррекции всех объектов выборки по критерию суммы квадратов расстояний от заданных точек до их образов при коррекции.

Обозначим матрицу входных данных $X = \left[x^1, \ldots, x^k, -y^1, \ldots, -y^l\right]^{\mathrm{T}}$; $X' = \left[x'^1, \ldots, x'^k, -y'^1, \ldots, -y'^l\right]^{\mathrm{T}}$ — матрица скорректированных значений признаков, $H = X' - X$ — матрица коррекции, вектор $(b, \ldots, b, -b, \ldots, -b)^{\mathrm{T}} = bp$, где $p = (1, \ldots, 1, -1, \ldots, -1)^{\mathrm{T}} \in \mathbb{R}^m$. Получим задачу коррекции несовместной системы линейных неравенств

$$v = \inf_{H,a,b} \left\{ \|H\|^2 \colon (X + H)a \leqslant bp \right\}, \quad \text{где } \|H\|^2 = \sum_{i,j=1}^{m,n} h_{ij}^2.$$

Обозначим $I' \subset I = \{1, \ldots, m\}$, $\bar{X}$ — подматрица матрицы X, состоящая из части строк X с номерами из I', $\hat{X}$ — подматрица X, состоящая из остальных строк, $\bar{p}$, $\hat{p}$ — вектора с координатами p, и с номерами, соответственно, I' и $I \setminus I'$, $P_{\bar{p}} = \frac{\bar{p}\bar{p}^{\mathrm{T}}}{(\bar{p}, \bar{p})}$ — матрица проектирования на векторное пространство с базисом $\bar{p}$. Можно показать, что значение задачи

определяется формулой [1]:

$$v = \min_{I' \subset I\colon \hat{X}e - b\hat{p} \leqslant 0} \lambda_{\min}(\bar{X}^{\mathrm{T}}(E - P_{\bar{p}})\bar{X}),$$

где E — единичная матрица, $\lambda_{\min}(\bar{X}^{\mathrm{T}}(E - P_{\bar{p}})\bar{X})$ — минимальное собственное число матрицы $\bar{X}^{\mathrm{T}}(E - P_{\bar{p}})\bar{X}$, e — соответствующий собственный вектор. Решение задачи коррекции, в том числе коэффициенты a, можно выразить через e.

В результате получим следующий метод построения решающей функции. Рассматриваются все подсистемы неравенств системы $Xa - bp \leqslant 0$, в порядке возрастания мощности. Для рассматриваемой подсистемы $\bar{X}a - b\bar{p} \leqslant 0$ определяется минимальная матрица коррекции H такая, что совместна система уравнений $(\bar{X} + H)a - b\bar{p} = 0$. Если, кроме того, выполняются остальные неравенства, $\hat{X}a - b\hat{p} \leqslant 0$, то 1) полученное значение $\|H\|$ сравнивается с текущим наилучшим значением задачи коррекции; 2) все подмножества, содержащие данное, можно исключить из рассмотрения.

Геометрически алгоритм можно интерпретировать следующим образом: для выбранного подмножества I' точек строится аппроксимирующая гиперплоскость по полному методу наименьших квадратов, т. е. координаты этих точек корректируются так, чтобы через них можно было провести гиперплоскость.

Работа выполнена при поддержке Программы Федерального агенства по образованию «Развитие потенциала высшей школы».

Литература

[1] *Матросов В. Л., Горелик В. А., Жданов С. А., Муравьева О. В.* Применение методов коррекции несобственных задач линейного программирования к задаче классификации // Научные труды Мос. пед. гос. ун-та. Серия: Естественные науки, М: Прометей, 2005. — С. 55–60.