Алгоритм выбора признаков в задаче обучения классификации методом опорных векторов

Гончаров Ю. В., Мучник И. Б., Шварцер Л. В.
Goncharov.Yuri@gmail.com, muchnik@dimacs.rutgers.edu,
Leonid.Shvartser@ness.com
Москва, Артезио; США, Rutgers University; Израиль, Ness Technologies

В предлагаемом алгоритме одновременно осуществляется выбор признаков и построение решающего правила классификации. Решение задачи выбора признаков сводится к поиску минимакса выпукло-вогнутой функции, которая является модифицированной функцией из метода опорных векторов [1].

В методе опорных векторов решается следующая задача оптимизации:

$$\min_{w,b,\delta_i}\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\delta_i\right); \tag{1}$$
$$y_i\bigl(\langle w, x_i\rangle + b\bigr) \geqslant 1-\delta_i,\ i=1,\ldots,l;$$
$$\delta_i \geqslant 0,\ i=1,\ldots,l;$$

где $x_i \in \mathbb{R}^n$, $y_i \in \{-1;+1\}$, $i=1,\ldots,l$, $C>0$.

Двойственной для (1) является следующая задача:

$$\max_{\lambda}\left(\sum_{i=1}^{l}\lambda_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{k=1}^{l}\bigl(y_i y_k\langle x_i, x_k\rangle\lambda_i\lambda_k\bigr)\right); \tag{2}$$
$$\sum_{i=1}^{l}\lambda_i y_i = 0; \quad 0 \leqslant \lambda_i \leqslant C,\ i=1,\ldots,l.$$

Пусть $I=\{1,\ldots,n\}$ — множество координат вектора v, $Q \subseteq I$ — подмножество координат. Обозначим через v^Q вектор с множеством координат Q, $v_i^Q = v_i$, $i \in Q$. Мы предлагаем минимизировать модифицированный функционал вида (1), в котором штрафуется величина размерности пространства признаков:

$$\min_{Q\subseteq I, w^Q, b, \delta}\left(\frac{1}{2}\|w^Q\|^2 + C\sum_{i=1}^{l}\delta_i + A\,|Q|\right); \tag{3}$$
$$y_i\bigl(\langle w^Q, x_i^Q\rangle + b\bigr) \geqslant 1-\delta_i,\ i=1,\ldots,l;$$
$$\delta_i \geqslant 0,\ i=1,\ldots,l.$$

Первые два члена функционала в (3) отвечают за поиск оптимальной гиперплоскости в подпространстве признаков, задаваемом Q. Третий член — шраф на мощность подмножества признаков, $A > 0$. Наша постановка задачи аналогична постановке, предложенной в работе [2].

Рассмотрим задачу:

$$\min_{z,w,b,\delta}\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\delta_i + A\sum_{j=1}^{n} z_j\right); \tag{4}$$

$$y_i\left(\sum_{j=1}^{n} w_j x_i^j \sqrt{z_j} + b\right) \geqslant 1 - \delta_i,\ i = 1, \dots, l;$$

$$\delta_i \geqslant 0,\ i = 1, \dots, l; \quad z_j \in \{0, 1\},\ j = 1, \dots, n;$$

где x_i^j — j-я координата вектора x_i, w_j — j-я координата вектора w. Эта задача содержит булевые переменные z_j. Значение $z_j = 1$ означает, что признак с индексом j выбран, значение $z_j = 0$ означает, что признак удален.

Следующая теорема утверждает, что задачи (3) и (4).

Теорема 1. *Пусть Q, w^Q, b, δ — решение задачи* (3), *значения z, w в задаче* (4) *преобразуются по следующему правилу:*

$$\begin{cases} z_j = 1,\ w_j = w_j^Q, & \text{если } j \in Q; \\ z_j = 0,\ w_j = 0, & \text{если } j \notin Q. \end{cases} \tag{5}$$

тогда z, w, b, δ — решение задачи (4).

Пусть z, w, b, δ — решение задачи (4), *значения Q, w^Q преобразуются по следующему правилу:*

$$Q = \{j\colon z_j = 1;\ j = 1, \dots, n\}, \quad w_j^Q = w_j,\ j \in Q. \tag{6}$$

Тогда Q, w^Q, b, δ — решение задачи (3).

В задаче (4) переменные z_j могут принимать значения 0 или 1. Предлагается ослабить это условие и разрешить переменным z_j принимать значения из отрезка $[0, 1]$. Таким образом, приходим к непрерывному аналогу задачи (4):

$$\min_{z,w,b,\delta}\left(\frac{1}{2}\,\|w\|^2 + C\sum_{i=1}^{l}\delta_i + A\sum_{j=1}^{n} z_j\right); \tag{7}$$

$$y_i\left(\sum_{j=1}^{n} w_j x_i^j \sqrt{z_j} + b\right) \geqslant 1 - \delta_i,\ i = 1, \dots, l;$$

$$\delta_i \geqslant 0,\ i = 1, \dots, l; \quad z_j \in [0, 1],\ j = 1, \dots, n.$$

Получившаяся задача имеет невыпуклые ограничения по совокупности всех переменных z, w, b, δ. Однако, возможно заменить задачу минимизации по совокупности всех переменных на эквивалентную задачу последовательной минимизации:

$$\min_{\substack{0\leqslant z_j\leqslant 1\\ j=1,\dots,n}} \psi(z), \tag{8}$$

где значение $\psi(z)$ получается в результате решения следующей задачи:

$$\psi(z) = \min_{w,b,\delta}\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\delta_i + A\sum_{j=1}^{n} z_j\right); \tag{9}$$

$$y_i\left(\sum_{j=1}^{n} w_j x_i^j \sqrt{z_j} + b\right) \geqslant 1-\delta_i,\ i=1,\dots,l;$$

$$\delta_i \geqslant 0,\ i=1,\dots,l.$$

Пусть z^* — решение задачи (8) и минимум в (9) при вычислении значения $\psi(z^*)$ достигается на w^*, b^*, δ^*. Тогда z^*, w^*, b^*, δ^* будет решением (7). Используя переход от прямой задачи (1) к двойственной задаче (2), можно записать (9) в следующем двойственном виде:

$$\psi(z) = \max_{\lambda}\left(\sum_{i=1}^{l}\lambda_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{k=1}^{l}\left(y_i y_k \sum_{j=1}^{n} z_j x_i^j x_k^j\right)\lambda_i\lambda_k + A\sum_{j=1}^{n} z_j\right); \tag{10}$$

$$\sum_{i=1}^{l}\lambda_i y_i = 0; \quad 0 \leqslant \lambda_i \leqslant C,\ i=1,\dots,l.$$

Теорема 2. *Функция $\psi(z)$ в задаче (10) является выпуклой.*

Запишем функцию из задачи (8)–(9) в виде функции двух переменных:

$$L(z,\lambda) = \sum_{i=1}^{l}\lambda_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{k=1}^{l}\left(y_i y_k \sum_{j=1}^{n} z_j x_i^j x_k^j\right)\lambda_i\lambda_k + A\sum_{j=1}^{n} z_j, \tag{11}$$

которая определена на произведении двух компактных множеств:

$$Z = \left\{z\colon 0 \leqslant z_j \leqslant 1,\ j=1,\dots,n\right\}; \tag{12}$$

$$\Lambda = \left\{\lambda\colon \textstyle\sum_{i=1}^{l}\lambda_i y_i = 0,\ 0\leqslant\lambda_i\leqslant C,\ i=1,\dots,l\right\}. \tag{13}$$

Заметим, что $L(z,\lambda)$ — выпукло-вогнутая функция, т. е. выпуклая по z при фиксированном значении λ и вогнутая по λ при фиксированном z.

Рассмотрим задачу поиска седловой точки $(z^*, \lambda^*) \in Z \times \Lambda$:

$$L(z^*, \lambda) \leqslant L(z^*, \lambda^*) \leqslant L(z, \lambda^*), \quad \forall z \in Z, \quad \forall \lambda \in \Lambda. \tag{14}$$

Для седловой точки справедливо следующее равенство:

$$\min_{z \in Z} \max_{\lambda \in \Lambda} L(z, \lambda) = \max_{\lambda \in \Lambda} \min_{z \in Z} L(z, \lambda) = L(z^*, \lambda^*). \tag{15}$$

Опираясь на это равенство, мы можем заменить задачу поиска минимакса на задачу поиска седловой точки. Следующая теорема утверждает существование седловой точки у функции (11) и показывает условия, при которых координаты оптимального z отличны от 0 и 1.

Теорема 3. *1. Существует седловая точка (z^*, λ^*) в задаче (14) и справедливы следующие равенства:*

$$\{z^* \colon (z^*, \lambda) \text{ — седловая точка}\} = \arg\min_{z \in Z} \max_{\lambda \in \Lambda} L(z, \lambda); \tag{16}$$

$$\{\lambda^* \colon (z, \lambda^*) \text{ — седловая точка}\} = \arg\max_{\lambda \in \Lambda} \min_{z \in Z} L(z, \lambda). \tag{17}$$

2. Если в седловой точке выполнено $\sum_{i=1}^{l} \sum_{k=1}^{l} y_i y_k x_i^j x_k^j \lambda_i^* \lambda_k^* > 2A$, то $z_j^* = 1$. Если в седловой точке выполнено $\sum_{i=1}^{l} \sum_{k=1}^{l} y_i y_k x_i^j x_k^j \lambda_i^* \lambda_k^* < 2A$, то $z_j^* = 0$. Если $0 < z_j^* < 1$, то

$$\sum_{i=1}^{l} \sum_{k=1}^{l} y_i y_k x_i^j x_k^j \lambda_i^* \lambda_k^* = 2A.$$

Для поиска седловой точки функции предлагается использовать дискретный вариант алгоритма из [3].

Литература

[1] *Vapnik V.* The Nature of Statistical Learning Theory. Second Edition. — New York: Springer, 2000. — 314 p.

[2] *Weston J., Mukherjee S., Chapelle O., Pontil M., Poggio T. and Vapnik V.* Feature Selection for SVMs // Advances in Neural Information Processing Systems, 13. — 2000.

[3] *Antipin A.* From optima to equilibria // Proc. of Institute for Systems Analysis. «Dynamics of non-homogeneous systems». Vol. 3. — Moscow, 2000. — Pp. 35–64.