Задача монотонизации выборки

Таханов Р. С.
takhanov@mail.ru
Москва, МФТИ, ЗАО «Форексис»

Требования к классифицирующим правилам в задачах обучения по прецедентам состоят из двух частей — требования согласованности с прецедентными данными и удовлетворения некоторым заранее установленным дополнительным ограничениям. Одним из популярных типов подобных дополнительных ограничений являются ограничения монотонности. В некоторых случаях, однако, эти два типа ограничений могут быть взаимно противоречивыми, тогда возникает задача минимальной коррекции прецедентных данных.

Итак, рассмотрим следующее обобщение этой задачи, которую обозначим как MaxCMS (Maximal Consistent with Monotonicity Set).

MaxCMS. Заданы конечные множества B_n, B_m, где $B_r = \{1, \dots, r\}$, на них отношения частичного порядка $\succcurlyeq^1, \succcurlyeq^2$ соответственно, и функция $\varphi\colon B_n \to B_m$. Для каждого элемента $i \in B_n$ задан положительный целочисленный вес w_i. Требуется найти максимальное по весу подмножество $B \subseteq B_n$, такое, что функция φ, ограниченная на B, является монотонной, то есть для любых $i, j \in B$ из $i \succcurlyeq^1 j$ следует $\varphi(i) \succcurlyeq^2 \varphi(j)$. Его вес обозначим MaxCMS.

Введем на множестве B_n частичный предпорядок (транзитивный и рефлексивный бинарный предикат): $i \succ j \Leftrightarrow \varphi(i) \succcurlyeq^2 \varphi(j)$. Рассмотрим орграф $G = (V, E)$, где $V = B_n$, а $E = \left\{(i,j) \;\middle|\; i \succcurlyeq^1 j,\ \varphi(i) \not\succcurlyeq^2 \varphi(j)\right\}$. Орграф G также может быть задан равенствами $V = B_n$ и $E = (\succcurlyeq^1 \cap \overline{\succ})$, где $\overline{\succ}$ — дополнение бинарного предиката.

Определение 1. *Всякий орграф, множество дуг которого может быть представлено как пересечение некоторых частичного порядка и дополнения частичного предпорядка на вершинах орграфа, называется специальным.*

Теорема 1. *Решение* MaxCMS *равно максимальному независимому множеству специального орграфа G.*

Теорема 2. MaxCMS — *NP-трудная задача.*

Определение 2. *Задача* MaxCMS *с входом* $\left(\succcurlyeq^1, \succcurlyeq^2, \varphi, w\right)$ *для случая, когда размерность частичного порядка $\succcurlyeq^2$ равна d, называется* d-MaxCMS.

Теорема 3. *d*-MaxCMS *сводится к нахождению максимального независимого множества в орграфе* $G = (V, E)$, *где* $E = (\succ^1 \cup \cdots \cup \succ^d)$, *и предикаты* $\succ^s$ *транзитивны, причем в* G *нет циклов.*

Теорема 4. 1-MaxCMS *полиномиально разрешима [4].*

1-MaxCMS сводится к следующей задаче линейного программирования, решаемую посредством метода эллипсоидов Хачияна [1]:

$$\begin{aligned}
&\sum\nolimits_{v\in V} w_v y(v) \to \max;\\
&\sum\nolimits_{v\in \Gamma} y(v) \leqslant 1, \quad \Gamma \in \mathbb{G}(s,t);\\
&y(v) \geqslant 0, \quad v \in V;
\end{aligned}$$

где $\mathbb{G}(s,t)$ — множество всех путей в орграфе $G = (V, E)$ из некоторого минимального элемента в некоторый максимальный элемент частичного порядка, определяемого орграфом G. Данный политоп обозначим через $\Pi(G)$.

Рассмотрим теперь задачу 2-MaxCMS. Рассмотрим два орграфа: $G_1 = \left(V, \succ^1\right)$ и $G_2 = \left(V, \succ^2\right)$. Заметим, что максимальное независимое множество орграфа $G = (V, E)$ является независимым множеством в обоих G_1 и G_2.

Рассмотрим следующую оптимизационную задачу:

$$\begin{aligned}
&\varphi(\overline{x}, \overline{y}) \to \max;\\
&\overline{x} \in \Pi(G_1); \quad \overline{y} \in \Pi(G_2);
\end{aligned}$$

где $\varphi(\overline{x}, \overline{y}) = -\frac{1}{2} \sum\limits_{v\in V} w_v (x_v - y_v)^2 - w_v (x_v + y_v)$. Будем называть ее выпуклой.

Рассмотрим следующий приближенный алгоритм для 2-MaxCMS.

1. Найти методом эллипсоидов для выпуклой оптимизации [2, 1] пару $\left(\overline{x'}, \overline{y'}\right)$ такую, что

$$\max_{(\overline{x},\overline{y})\in\Pi(G_1)\times\Pi(G_2)} \varphi(\overline{x}, \overline{y}) \leqslant \varphi(\overline{x}', \overline{y}') + \varepsilon, \quad |x'_i - y'_i| \leqslant \tfrac{1}{2}, \quad \varepsilon = \tfrac{1}{16}.$$

2. Найти $\overline{x}^* = \arg \max\limits_{\overline{x}\in\Pi(G_1)} \psi(\overline{x}, \overline{y}')$ и $\overline{y}^* = \arg \max\limits_{\overline{y}\in\Pi(G_2)} \psi(\overline{x}^*, \overline{y})$, где $\psi(\overline{x}, \overline{y}) = \sum\limits_{v\in V} w_v x_v y_v$. Здесь $\overline{x}^*, \overline{y}^*$ целочисленны.

Ответ алгоритма — множество вершин $\{v | x^*_v y^*_v = 1\}$.

Рассмотрим MaxCMS как задачу минимального вершинного покрытия. Фактически это означает, что задача заключается в удалении «шума» в обучающей выборке для построения корректного классификатора.

Очевидно, что этим самым «шумом» и является соответствующее минимальное вершинное покрытие.

Введем обозначения $W = \sum_{v \in V} w_v$ и $\mathrm{MaxCMS} = \alpha W$.

Ясно, что $0 \leqslant \alpha \leqslant 1$.

Теорема 5. *Алгоритм является полиномиальным.*

Теорема 6. *Алгоритм аппроксимирует минимальное вершинное покрытие G с константой $1 + \alpha \leqslant 2$, при $\alpha \geqslant \frac{1}{2}$.*

С учетом того, что стандартный алгоритм для вершинного покрытия [3] имеет константу аппроксимации 2, предложенный алгоритм является более точным.

Литература

[1] *Хачиян Л. Г.* Полиномиальный алгоритм в линейном программировании // Доклады АН СССР. — 1979. — Т. 244. — С. 1093–1096.

[2] *Grotshel M., Lovasz L., Schrijver A.* Geometric algorithms and combinatorial optimization. — Springer-Verlag, 1988.

[3] *Hochbaum D. S.* Approximation algorithms for the set covering and vertex cover problems // SIAM Journal on Computing. — 1982. — No 11. — Pp. 555–556.

[4] *Mohring R. H.* Algorithmic aspects of comparability graphs and interval graphs // Graphs and Order. — Dordrecht: Reidel, 1985. — Pp. 41–101.