

# О задаче монотонизации выборки

Р.С. Таханов

## Аннотация

Рассматривается задача выделения максимальной подвыборки некоторой обучающей выборки, состоящей из пар вида «объект - ответ», не противоречащей ограничениям монотонности. Показывается, что данная задача является NP-трудной и равносильна задаче о максимальном независимом множестве в специальных орграфах. Подробно рассматриваются очень важные практически случаи, когда частичный порядок, заданный на множестве ответов, является полным порядком либо имеет размерность 2. Показывается, что второй случай сводится к максимизации квадратично-выпуклой функции на выпуклом множестве. Для этого случая строится приближенный полиномиальный алгоритм, основанный на выпуклой оптимизации.

## 1 Постановка задачи монотонизации выборки

Требования к классифицирующим правилам в задачах обучения по прецедентам состоят из двух частей — требования согласования 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 2 типа ограничений могут быть взаимно противоречивыми и возникает задача минимальной коррекции прецедентных данных. Рассмотрим эту задачу.

Пусть заданы множества  $X, Y$  и на них частичные порядки  $\geq^X, \geq^Y$  соответственно. Предположим также, что частичный порядок  $\geq^Y$  является решеткой. При заданном отображении  $o : X' \rightarrow Y$ , где  $X' \subseteq X, |X'| < \infty$ , возникает задача нахождения функции  $f : X \rightarrow Y$ , монотонной относительно частичных порядков  $\geq^X, \geq^Y$  и минимизирующей функционал согласованности:  $Er_o(f) = |\{x|f(x) \neq o(x)\} \cap X'|$ .

Обозначим  $M(\geq^X, \geq^Y)$  множество монотонных функций из  $X$  в  $Y$ . Тогда, при заданном отображении  $o : X' \rightarrow Y$ , задача заключается в следующем:

$$Er_o(f) \rightarrow \min_{f \in M(\geq^X, \geq^Y)}$$

Всякое монотонное на подмножестве  $X' \subseteq X$  отображение  $f' : X' \rightarrow Y$  может быть продолжено до монотонного на всем  $X$ , так как множество  $(Y, \geq^Y)$  есть решетка. Действительно, на решетке  $(Y, \geq^Y)$  у всякого конечного подмножества существует  $\sup$  и функция  $f(x) = \sup \{f'(x') \mid x' \in X', x' \leq^X x\}$  обладает свойством монотонности и  $f(x) = f'(x)$ ,  $x \in X'$ . Отсюда следует, что в поставленной задаче можно всегда полагать  $X' = X$ . Из сказанного также следует, что эта задача равносильна нахождению максимального подмножества  $X'' \subseteq X'$ , такого, что функция  $o$ , ограниченная на множество  $X''$ , является монотонной.

Итак, рассмотрим следующее обобщение этой задачи, которую обозначим как  $\text{MaxCMS}$  (Maximal Consistent with Monotonicity Set).

**MaxCMS.** Заданы конечные множества  $B_n, B_m$ , где  $B_r = \{1, \dots, r\}$  и на них частичные порядки  $\geq^1, \geq^2$  соответственно и функция  $\varphi : B_n \rightarrow B_m$ . 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  $\varphi$ , ограниченная на  $B$ , является монотонной, то есть  $\forall i, j \in B [i \geq^1 j \rightarrow \varphi(i) \geq^2 \varphi(j)]$ .

**Определение 1.** Множества  $B \subseteq B_n$ , такие, что функция  $\varphi$ , ограниченная на  $B$ , является монотонной, называются допустимыми.

**Определение 2.** Максимальное допустимое  $B \subseteq B_n$  обозначим  $\text{MaxCMS}(\geq^1, \geq^2, \varphi, w)$  (в некоторых случаях так обозначается его вес).

В остальной части работы мы будем рассматривать именно эту задачу.

## 2 Задача монотонизации выборки и максимальные независимые подмножества

Покажем, что  $\text{MaxCMS}$  представляет собой задачу нахождения максимального независимого подмножества (либо минимального вершинного покрытия) в орграфах специального вида.

**Определение 3.** Пусть задан орграф  $G = (V, E)$  и каждая вершина  $v$  орграфа имеет положительный целочисленный вес  $w_v$ . Под независимым множеством  $G$  будем понимать подмножество вершин орграфа, каждая пара элементов которого не соединена дугой. Обозначим  $IS(G, w)$  наибольшее по весу

независимое множество орграфа (в некоторых случаях так обозначается и сам вес).

Введем на множестве  $B_n$  частичный предпорядок (напомним, что так называется транзитивный и рефлексивный бинарный предикат):

$$i \succ j \Leftrightarrow \varphi(i) \geq^2 \varphi(j).$$

Рассмотрим орграф  $G = (V, E)$ , где  $V = B_n$ , а  $E = \{(i, j) \mid i \geq^1 j, \varphi(i) \not\geq^2 \varphi(j)\}$ . Орграф  $G$  также может быть задан равенствами:  $V = B_n$  и  $E = \geq^1 \cap \overline{\succ}$ , где  $\overline{\succ}$  - дополнение бинарного предиката.

**Определение 4.** Всякий орграф, множество дуг которого может быть представлено как пересечение некоторых частичного порядка и дополнения частичного предпорядка на вершинах орграфа называется специальным.

**Предложение 1.** Максимальное допустимое множество равно максимальному независимому множеству специального орграфа  $G$ , то есть  $MaxCMS(\geq^1, \geq^2, \varphi, w) = IS(G, w)$ .

**Доказательство.** Всякое независимое множество  $B$  орграфа  $G$  обладает тем свойством, что если  $i, j \in B$  и  $i \geq^1 j$ , то  $\varphi(i) \geq^2 \varphi(j)$ , то есть функция  $\varphi$ , ограниченная на  $B$ , является монотонной. Верно и обратное, если ограничение  $\varphi$  на  $B$  монотонно, то  $B$  — независимое множество в  $G$ . Отсюда следует утверждение предложения.

**Предложение 2.** Пусть произвольный специальный орграф  $G'$  задается множеством вершин  $V' = B_n$  с весами  $w'_i$  и дуг  $E' = \geq' \cap \overline{\succ'}$ , причем заданы по отдельности  $\geq'$  и  $\overline{\succ'}$  (то есть множество дуг  $E'$  не нужно декомпозировать). Тогда задача нахождения максимального независимого множества в таком орграфе полиномиально сводится к MaxCMS.

**Доказательство.** Разобьем множество  $V'$  на классы эквивалентности по отношению  $x \sim y \Leftrightarrow x \succ y \& y \succ x$ . Тогда этому разбиению естественно соответствует отображение  $\varphi' : V \rightarrow V / \sim$ . На фактор-множестве  $V / \sim$  естественно индуцируется частичный порядок  $\bar{x} \geq'' \bar{y} \Leftrightarrow x \succ y$ . Как легко видеть,  $IS(G', w') = MaxCMS(\geq', \geq'', \varphi', w')$ . Сводимость осуществляется за  $O(n^2)$  шагов.

### 3 NP-полнота MaxCMS.

В предыдущей главе было показано, что MaxCMS равносильно поиску максимального независимого множества (или минимального вершинного покрытия) в орграфах специального вида. Задачу о существовании допустимого

множества мощности большей чем  $C$  обозначим CMS. Очевидно, она принадлежит классу NP.

**Теорема 1.** CMS — NP-полная задача.

**Доказательство.** Сведем задачу 3-ВЫП к CMS. Используем прием примененный в [1] для сведения 3-ВЫП к Вершинному покрытию. Напомним, что вершинным покрытием называется подмножество вершин, которое для любой дуги орграфа содержит одну из ее вершин. Легко видеть, что дополнение вершинного покрытия является независимым множеством.

Пусть задана 3-КНФ и  $U = \{u_1, \dots, u_n\}$  — множество переменных, использованных в ней, а  $C = \{c_1, \dots, c_m\}$  — множество ее дизъюнктов, каждый из которых содержит ровно 3 различных по переменным литерала (литерал это  $u_i$  либо  $\bar{u}_i$ ). Для каждого дизъюнкта упорядочим вхождения литералов в него. Тогда факт вхождения литерала  $l$  в дизъюнкт  $c_r$  на  $s$ -м месте обозначим как  $lc_r^s$ . Рассмотрим граф, вершинами которого являются литералы и тройные копии дизъюнктов  $V = \{u_1, \bar{u}_1, \dots, u_n, \bar{u}_n\} \cup \{c_1^1, c_1^2, c_1^3, \dots, c_m^1, c_m^2, c_m^3\}$ . Определим множество дуг равным  $E = E_1 \cup E_2$ , где  $E_1 = \{(u_i, \bar{u}_i)\}_{i=1}^n \cup \{(u_k, c_m^l) \mid u_k c_m^l\} \cup \{(c_m^l, \bar{u}_k) \mid \bar{u}_k c_m^l\}$  и  $E_2 = \{(c_j^1, c_j^2), (c_j^2, c_j^3), (c_j^1, c_j^3)\}_{j=1}^m$  (это разбиение множества ребер на 2 подмножества понадобится нам в дальнейшем).

Вершинное покрытие орграфа  $G = (V, E)$  длины  $n + 2m$  существует тогда и только тогда, когда исходная 3-КНФ выполнима. Действительно, из каждой пары вершин  $u_i, \bar{u}_i$  одна вершина и из каждой тройки  $c_j^1, c_j^2, c_j^3$  не меньше 2 вершин должны войти в вершинное покрытие, так как они попарно соединены. Итак, мощность вершинного покрытия не меньше  $n + 2m$ .

Пусть такое вершинное покрытие существует. Если в него вошел литерал  $u_i$ , полагаем  $u_i = true$ , иначе  $u_i = false$ . Все переменные получают так свои значения, так как из вышесказанного ясно, что, либо  $u_i$ , либо  $\bar{u}_i$  присутствуют в покрытии, причем не одновременно. Тогда этот набор, как легко видеть, и будет выполняющим для исходной 3-КНФ. Это рассуждение может быть обратимо и из существования выполняющего набора следует наличие вершинного покрытия мощности  $n + 2m$ .

Рассмотрим теперь граф  $G' = (V, E^* \setminus E)$ , где  $E^*$  - транзитивное замыкание  $E$ . Предположим, что граф  $G'$  транзитивен. Тогда, положив  $\geq = E^*$  и  $\succ = E^* \setminus E$ , получим, что  $\geq \cap \succ = E$ . Это означает, что наша задача свелась к нахождению минимального вершинного покрытия, а следовательно, максимального независимого множества для специального орграфа  $G = (V, E)$ , которая по предложению 2 равносильна MaxCMS, или CMS, при  $C = 2n + 3m - (n + 2m) = n + m$ .

Покажем, что граф  $G'$  обладает свойством транзитивности. Так как

$G^* = (V, E^*)$  транзитивен, транзитивности  $G'$  может помешать лишь наличие  $(u, v), (v, t) \in E^* \setminus E$ , что  $(u, t) \in E$ . Пусть  $(u, t) \in \{(u_i, \bar{u}_i)\}_{i=1}^n$ . Но легко видеть, что любая цепочка в графе  $G$  начинающаяся с литерала может  $u_i$  не может закончиться в литерале  $\bar{u}_i$ , так как иначе должен существовать дизъюнкт содержащий оба литерала  $u_i$  и  $\bar{u}_i$ . Рассмотрим теперь случай, когда  $(u, t) \in \{(c_j^1, c_j^2), (c_j^2, c_j^3), (c_j^1, c_j^3)\}$ . В этом случае цепочка начинающаяся с  $c_j^\alpha$  и заканчивающаяся в  $c_j^\beta$ , не может содержать элемент не принадлежащий  $\{c_j^1, c_j^2, c_j^3\}$ , откуда следует, что  $(u, v), (v, t) \in E$ , что противоречит  $(u, v), (v, t) \in E^* \setminus E$ . Наконец, последний случай,  $(u, t) \in \{(u_k, c_m^l) | u_k c_m^l\}$ . Но, легко видеть, что любой путь в орграфе  $G$ , который начинается в  $u$ , а заканчивается в  $t$ , тождественен дуге  $(u, t)$ , что показывает, что  $(u, v) \notin E^* \setminus E$ . Также рассматривается и случай  $(u, t) \in \{(c_m^l, \bar{u}_k) | \bar{u}_k c_m^l\}$ . Итак, граф  $G'$  транзитивен и сведение 3-ВЫП к CMS осуществлено.

## 4 1-MaxCMS

Всякий частичный порядок на конечном множестве можно представить как пересечение полных порядков на нем.

**Определение 5.** Пусть на конечном множестве  $M$  задан частичный порядок  $\geq$ . Размерностью частичного порядка  $\geq$  назовем минимальное число  $d$  полных порядков  $\geq_1, \dots, \geq_d$ , что  $\geq = \geq_1 \cap \dots \cap \geq_d$ .

Рассмотрим задачу MaxCMS с входом  $(\geq^1, \geq^2, \varphi, w)$  для случая, когда размерность частичного порядка  $\geq^2$  равна  $d$ . В этом случае  $\geq^2 = \geq_1 \cap \dots \cap \geq_d$ . Для соответствующего этой задаче специального орграфа  $G = (V, E)$  будет справедливо:  $V = B_n$  и  $E = \geq^1 \cap \overline{\succ}$ , где  $i \succ j \Leftrightarrow i \succ_1 j \& \dots \& i \succ_d j$  и  $i \succ_s j \Leftrightarrow \varphi(i) \geq_s \varphi(j)$ . И тогда,

$$E = \geq^1 \cap \overline{\succ_1} \cap \dots \cap \overline{\succ_d} = \geq^1 \cap (\overline{\succ_1} \cup \dots \cup \overline{\succ_d}) = (\geq^1 \cap \overline{\succ_1}) \cup \dots \cup (\geq^1 \cap \overline{\succ_d}).$$

Так как каждый из предикатов  $\overline{\succ_s}$  является транзитивным, то  $E$  можно представить как объединение  $d$  транзитивных предикатов.

**Определение 6.** Задача MaxCMS с входом  $(\geq^1, \geq^2, \varphi, w)$  для случая, когда размерность частичного порядка  $\geq^2$  равна  $d$ , называется  $d$ -MaxCMS.

Фактически доказано

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

Из предложения следует, что 1-MaxCMS сводится к нахождению максимального независимого множества в орграфе  $G = (V, E)$  без циклов, удовлетворяющему условию транзитивности дуг: если  $(u, v), (v, t) \in E$ , то  $(u, t) \in E$ . Эта задача имеет полиномиальный алгоритм решения, так как граф, получаемый из орграфа  $G$  преобразованием дуг в неориентированные ребра, является графом сравнимости некоторого частичного порядка, то есть совершенным. Приведем один из алгоритмов ее решения, следуя [7].

**Теорема 2.** 1-MaxCMS полиномиально разрешима.

**Доказательство.** Полагая  $x \triangleright y \Leftrightarrow (x, y) \in E$ , оргграф можно рассматривать как частично упорядоченное множество  $(V, \triangleright)$ . Алгоритм заключается в сведении к задаче нахождения минимального потока в некотором графе. Покажем как происходит сведение.

Обозначим  $\min G$  и  $\max G$  соответственно множество минимальных и максимальных элементов  $(V, \triangleright)$ . Для каждой вершины  $v \in V$  графа  $G$  создадим 2 копии  $v^+, v^-$  (для выходящих ребер и входящих ребер). Положим  $V' = \{v^+, v^-\}_{v \in V} \cup \{s, t\}$  и  $E' = \{(v^+, v^-)\}_{v \in V} \cup \{(x^-, y^+) \mid (x, y) \in E\} \cup \{(s, a^+) \mid a \in \min G\} \cup \{(b^-, t) \mid b \in \max G\}$ . Получим оргграф  $G' = (V', E')$ . Положим минимальные пропускные способности ребер  $(v^+, v^-)$  равными весам соответствующих элементов  $w_v$ , а для остальных ребер равными 0. Максимальные пропускные способности всех дуг положим равными  $\infty$ .

Легко видеть, что для любой дуги  $e \in E'$  в орграфе  $G'$  найдется путь из  $s$  в  $t$  проходящий через  $e$ . Данное условие, очевидно, гарантирует существование допустимого конечного потока через сеть (допустимый поток определяется как поток с весами ребер, большими минимальных пропускных способностей, и равенством 0 дивергенции для каждой внутренней вершины). Следовательно для данной сети можно использовать модифицированный алгоритм Форда-Фалкерсона, с той лишь разницей от обычного, что через ненасыщенные пути из  $s$  в  $t$  поток нужно уменьшать. Полученный в результате поток будет минимальным, и ему будет соответствовать разрез с максимальным  $P$ -весом, где под  $P$ -весом разреза  $S, \bar{S}$  понимается (граф ориентированный и вес не сумма всех его ребер!)

$$\sum_{(u,v) \in E, u \in S, v \in \bar{S}} c_{\min}(e) - \sum_{(u,v) \in E, v \in S, u \in \bar{S}} c_{\max}(e).$$

Минимальному потоку данной сети, найденному модифицированным алгоритмом Форда-Фалкерсона (обычный находит максимальный поток), будет соответствовать  $P$ -максимальный разрез.

Рассмотрим произвольный разрез  $V' = S \cup \bar{S}$ , где  $s \in S, t \in \bar{S}$ , с весом отличным от  $-\infty$ . Так как максимальные пропускные способности ребер равны  $\infty$ , то ребер  $(u, v) \in E'$  вида  $v \in S, u \in \bar{S}$  быть не может. Ребра же  $(u, v) \in E'$  вида  $u \in S, v \in \bar{S}$  только в том случае дадут вклад в вес разреза, если  $u = r^+, v = r^-$ . Обозначим  $R = \{r | r^+ \in S, r^- \in \bar{S}\}$ . Очевидно, что элементы  $R$  представляют собой независимое множество в  $G$  и вес разреза будет в точности равен весу этого множества. Обратное также справедливо, любому независимому в  $G$  множеству  $R$  соответствует разрез  $S = \{u^+, u^- | u \notin R \& \exists r \in R [r \triangleright u]\} \cup \{r^+ | r \in R\} \cup \{s\}$  вес которого равен весу  $R$ . Из этого следует, что результату этого алгоритма, то есть  $S$ -максимальному разрезу будет соответствовать максимальное независимое множество в  $G$ . Теорема доказана.

Запишем задачу нахождения минимального потока как задачу линейного программирования:

$$\begin{aligned} x(\Gamma) &\geq 0, \Gamma \in \mathbb{G}(s, t) \\ \sum_{\Gamma \in \mathbb{G}(v)} x(\Gamma) &\geq w_v \\ \sum_{\Gamma \in \mathbb{G}(s, t)} x(\Gamma) &\rightarrow \min \end{aligned}$$

где  $\mathbb{G}(s, t)$  - множество всех путей в графе  $G' = (V', E')$  из  $s$  в  $t$ , а  $\mathbb{G}(v) \subset \mathbb{G}(s, t)$  — множество путей проходящих через ребро  $(v^+, v^-)$ . Двойственная ей:

$$\begin{aligned} y(v) &\geq 0, v \in V \\ \sum_{(v^+, v^-) \in \Gamma} y(v) &\leq 1, \Gamma \in \mathbb{G}(s, t) \\ \sum_{v \in V} w_v y(v) &\rightarrow \max \end{aligned}$$

Из доказанного выше следует, что у двойственной задачи всегда существует целочисленное решение. Политоп двойственной задачи обозначим как  $\Pi(G)$ .

## 5 2-MaxCMS

Рассмотрим теперь задачу 2-MaxCMS. Как было показано, она сводится к задаче нахождения максимального независимого множества в орграфе  $G = (V, E)$ , где  $E = \succ^1 \cup \succ^2$  и предикаты  $\succ^s$  транзитивны, причем в  $G$  нет циклов. В дальнейшем будет рассматриваться именно эта задача.

Заметим, что дуги орграфа в теореме 1 также разбиты на 2 множества  $E_1$  и  $E_2$ , каждое из которых транзитивно. Отсюда следует, что рассматриваемая задача NP-трудна.

Рассмотрим 2 орграфа:  $G_1 = (V, \succ^1)$  и  $G_2 = (V, \succ^2)$ . Заметим, что максимальное независимое множество орграфа  $G = (V, E)$  является независимым множеством в обоих  $G_1$  и  $G_2$ , откуда очевидна следующая теорема.

**Предложение 4.** Множество решений задачи

$$\begin{aligned} \bar{x} &\in \Pi(G_1) \\ \bar{y} &\in \Pi(G_2) \\ \psi(\bar{x}, \bar{y}) &= \sum_{v \in V} w_v x_v y_v \rightarrow \max \end{aligned}$$

содержит и такие вектора  $\bar{x}^*, \bar{y}^*$ , что  $\{v | x_v^* y_v^* = 1\}$  является максимальным независимым множеством  $G$ .

**Доказательство.** Так как при фиксированном  $\bar{x}$  максимум  $\sum_{v \in V} w_v x_v y_v$  по полиэдру  $\bar{y} \in \Pi(G_2)$  достигается и на некотором целочисленном векторе  $\bar{y}$  (и наоборот), можно считать, что максимум функционала достигается на целочисленных векторах.

**Предложение 5.** Справедливо

$$\max_{\bar{x} \in \Pi(G_1), \bar{y} \in \Pi(G_2)} \psi(\bar{x}, \bar{y}) = \max_{\bar{x} \in \Pi(G_1), \bar{y} \in \Pi(G_2)} \gamma(\bar{x}, \bar{y}),$$

где

$$\gamma(\bar{x}, \bar{y}) = \frac{1}{2} \sum_{v \in V} w_v (x_v + y_v)^2 - w_v (x_v + y_v)$$

**Доказательство.**

$$\begin{aligned} \max_{\bar{x} \in \Pi(G_1), \bar{y} \in \Pi(G_2)} \sum_{v \in V} w_v x_v y_v &= \max_{\bar{x} \in \Pi(G_1), \bar{y} \in \Pi(G_2)} \frac{1}{2} \sum_{v \in V} w_v (x_v + y_v)^2 - w_v (x_v^2 + y_v^2) \geq \\ &= \max_{\bar{x} \in \Pi(G_1), \bar{y} \in \Pi(G_2)} \frac{1}{2} \sum_{v \in V} w_v (x_v + y_v)^2 - w_v (x_v + y_v) \end{aligned}$$

Однако, так как максимум левой части неравенства достигается на целочисленных векторах, ясно, что здесь необходимо равенство. Учитывая, что функционал  $\gamma(\bar{x}, \bar{y})$  является выпуклым, наша задача свелась к максимизации выпуклой функции на выпуклом множестве.

## 6 Приближенный алгоритм для 2-MaxCMS

Рассмотрим функционал

$$\varphi(\bar{x}, \bar{y}) = -\frac{1}{2} \sum_{v \in V} w_v (x_v - y_v)^2 - w_v (x_v + y_v)$$

**Предложение 6.** Справедливо

$$\max_{\bar{x} \in \Pi(G_1), \bar{y} \in \Pi(G_2)} \varphi(\bar{x}, \bar{y}) \geq \max_{\bar{x} \in \Pi(G_1), \bar{y} \in \Pi(G_2)} \psi(\bar{x}, \bar{y}),$$

причем на целочисленных точках политопа  $\bar{x} \in \Pi(G_1), \bar{y} \in \Pi(G_2)$  значения  $\varphi(\bar{x}, \bar{y})$  и  $\psi(\bar{x}, \bar{y})$  совпадают.

**Доказательство.** Проверка второго утверждения очевидна. Из него следует первое, в связи с тем, что максимум правой части по предложению 4 достигается и на целочисленных векторах.

Рассмотрим теперь следующую оптимизационную задачу:

$$\begin{aligned} \bar{x} &\in \Pi(G_1) \\ \bar{y} &\in \Pi(G_2) \\ \varphi(\bar{x}, \bar{y}) &\rightarrow \max \end{aligned}$$

Будем называть ее выпуклой.

**Определение 7.** Для выпуклой задачи  $\varepsilon$ -решением называется пара  $\bar{x}^* \in \Pi(G_1), \bar{y}^* \in \Pi(G_2)$  такая, что  $\max_{\bar{x} \in \Pi(G_1), \bar{y} \in \Pi(G_2)} \varphi(\bar{x}, \bar{y}) - \varphi(\bar{x}^*, \bar{y}^*) \leq \varepsilon$

**Предложение 7.** Выпуклая задача для любого  $\varepsilon$  может быть  $\varepsilon$ -разрешена за полиномиальное от длины входа время. Под длиной входа подразумевается длина описания графов  $G_1 = (V, \succ^1)$  и  $G_2 = (V, \succ^2)$  и целочисленных весов  $w_v$ . Причем полученное  $\varepsilon$ -решение  $(\bar{x}^*, \bar{y}^*)$  удовлетворяет условиям  $|x_i^* - y_i^*| \leq \frac{1}{2}$ .

**Лемма.** Точка  $\bar{\xi}^{opt} = (\bar{x}^{opt}, \bar{y}^{opt}) = \arg \max_{(\bar{x}, \bar{y}) \in \Pi(G_1) \times \Pi(G_2)} \varphi(\bar{x}, \bar{y})$  удовлетворяет условию  $|x_i^{opt} - y_i^{opt}| \leq \frac{1}{2}$ .

**Доказательство леммы.** Квадратичный функционал  $\varphi(\bar{x}, \bar{y})$  не является ограниченным в  $R^{2n}$  и потому, максимум на множестве  $\Pi(G_1) \times \Pi(G_2)$  достигается на границе полиэдра. Пусть  $\bar{a}_1^T \bar{\xi} \leq b_1, \dots, \bar{a}_s^T \bar{\xi} \leq b_s$ —те самые неравенства в определении полиэдра, которые обращаются в равенства. Из оптимальности  $(\bar{x}^{opt}, \bar{y}^{opt})$  ясно, что конус  $\{\bar{\xi} | \bar{a}_1^T \bar{\xi} \leq 0\} \cap \dots \cap \{\bar{\xi} | \bar{a}_s^T \bar{\xi} \leq 0\} \cap \{\bar{\xi} | \nabla_{\bar{\xi}} \varphi(\bar{\xi}^{opt})^T \bar{\xi} > 0\} = \emptyset$ . Отсюда, из теоремы Фаркаша-Минковского следует, что  $\varphi(\bar{\xi}^{opt})$  разлагается в положительную комбинацию векторов  $\bar{a}_1, \dots, \bar{a}_s$ .

Но учитывая, что компоненты этих векторов положительны, получаем, что и  $\varphi(\bar{\xi}^{opt}) = \left\| x_1^{opt} - y_1^{opt} + \frac{1}{2}, y_1^{opt} - x_1^{opt} + \frac{1}{2}, \dots, x_n^{opt} - y_n^{opt} + \frac{1}{2}, y_n^{opt} - x_n^{opt} + \frac{1}{2} \right\|^T \geq \bar{0}$ . Лемма доказана.

**Доказательство предложения.** Так как функция  $\varphi(\bar{x}, \bar{y})$  является вогнутой, множество пар

$$\begin{aligned} \bar{x} &\in \Pi(G_1) \\ \bar{y} &\in \Pi(G_2) \\ \varphi(\bar{x}, \bar{y}) &\geq c \\ -\frac{1}{2} &\leq x_i - y_i \leq \frac{1}{2}, i = \overline{1, n} \end{aligned}$$

является выпуклым.

Заметим, что по заданной паре векторов  $\bar{x}', \bar{y}'$ , задача определения их принадлежности множеству  $\Pi(G_1) \times \Pi(G_2)$  решается за полиномиальное время. Действительно, алгоритм Флойда-Уоршола позволяет нам найти самый длинный путь из вершины  $s$  в вершину  $t$  орграфов  $G_1$  и  $G_2$ , где под длиной пути подразумевается сумма весов вершин пути. И сравнение этих цифр с 1, дает нам ответ о принадлежности  $\bar{x}', \bar{y}' \in \Pi(G_1) \times \Pi(G_2)$ . Кроме того, если  $\bar{x}', \bar{y}' \notin \Pi(G_1) \times \Pi(G_2)$ , то найденный путь длины большей чем 1 и даст нам нарушенное линейное неравенство в определении политопа  $\Pi(G_1) \times \Pi(G_2)$ .

Наконец, при заданных  $\bar{x}', \bar{y}'$ , удовлетворение условия  $\varphi(\bar{x}', \bar{y}') \geq c$ , а также в случае его нарушения, гиперплоскость разделяющая пару  $\bar{x}', \bar{y}'$  и множество  $\{(\bar{x}, \bar{y}) \mid \varphi(\bar{x}, \bar{y}) \geq c + \varepsilon\}$  может быть найдена за полиномиальное время.

Действительно,

$$\begin{aligned} \{(\bar{x}, \bar{y}) \in \Pi(G_1) \times \Pi(G_2) \mid (\nabla_{\bar{x}'} \varphi(\bar{x}', \bar{y}'), \bar{x} - \bar{x}') + (\nabla_{\bar{y}'} \varphi(\bar{x}', \bar{y}'), \bar{y} - \bar{y}') \geq \varepsilon\} &\supseteq \\ &\supseteq \{(\bar{x}, \bar{y}) \in \Pi(G_1) \times \Pi(G_2) \mid \varphi(\bar{x}, \bar{y}) \geq c + \varepsilon\} \end{aligned}$$

Это следует из следующих неравенств для квадратично-вогнутой функции  $\varphi$  и всяких таких точек  $(\bar{x}, \bar{y}), (\bar{x}', \bar{y}')$ , что  $\varphi(\bar{x}, \bar{y}) \geq c + \varepsilon$  and  $\varphi(\bar{x}', \bar{y}') \leq c$ :  $\varepsilon \leq \varphi(\bar{x}, \bar{y}) - \varphi(\bar{x}', \bar{y}') \leq (\nabla_{\bar{x}'} \varphi(\bar{x}', \bar{y}'), \bar{x} - \bar{x}') + (\nabla_{\bar{y}'} \varphi(\bar{x}', \bar{y}'), \bar{y} - \bar{y}')$ .

Тогда округляя компоненты  $\nabla_{\bar{x}'} \varphi(\bar{x}', \bar{y}')$  и  $\nabla_{\bar{y}'} \varphi(\bar{x}', \bar{y}')$  до  $2(\log n + |\log \varepsilon| + 1)$  знаков в двоичном представлении и обозначая их как  $c_x$  и  $c_y$ , получим следующую разделяющую гиперплоскость

$$\left\{ (\bar{x}, \bar{y}) \mid (c_x, \bar{x} - \bar{x}') + (c_y, \bar{y} - \bar{y}') \geq \frac{\varepsilon}{2} \right\}.$$

Согласно [5, 2], в этом случае пара векторов  $\bar{x}', \bar{y}'$  удовлетворяющих усло-

виям:

$$\begin{aligned}\bar{x}' &\in \Pi(G_1) \\ \bar{y}' &\in \Pi(G_2) \\ \varphi(\bar{x}', \bar{y}') &\geq c \\ -\frac{1}{2} &\leq x'_i - y'_i \leq \frac{1}{2}, i = \overline{1, n}\end{aligned}$$

может найдена за полиномиальное время методом эллипсоидов[3, 4], либо будет показано, что

$$\left\{ (\bar{x}, \bar{y}) \mid \bar{x} \in \Pi(G_1), \bar{y} \in \Pi(G_2), \varphi(\bar{x}, \bar{y}) \geq c + \varepsilon, -\frac{1}{2} \leq x_i - y_i \leq \frac{1}{2}, i = \overline{1, n} \right\} = \emptyset.$$

С учетом того, что  $|\varphi(\bar{x}, \bar{y})| \leq 2 \sum_{v \in V} w_v$ , методом половинного деления мы быстро приходим к такой константе  $c$ , что множество  $\Omega = \{(\bar{x}, \bar{y}) \mid \bar{x} \in \Pi(G_1), \bar{y} \in \Pi(G_2), \varphi(\bar{x}, \bar{y}) \geq c, -\frac{1}{2} \leq x_i - y_i \leq \frac{1}{2}, i = \overline{1, n}\} \neq \emptyset$  и  $\{(\bar{x}, \bar{y}) \mid \bar{x} \in \Pi(G_1), \bar{y} \in \Pi(G_2), \varphi(\bar{x}, \bar{y}) \geq c + \varepsilon, -\frac{1}{2} \leq x_i - y_i \leq \frac{1}{2}, i = \overline{1, n}\} = \emptyset$ . Из леммы следует, что  $\bar{\xi}^{opt} \in \Omega$  и  $\varphi(\bar{\xi}^{opt}) < c + \varepsilon$ . И любая пара из  $\Omega$  дает  $\varepsilon$ -решение оптимизационной задачи. Предложение доказано.

Рассмотрим следующий приближенный алгоритм для 2-МахСМS.

1. Найти пару  $(\bar{x}', \bar{y}')$  такую, что  $\max_{(\bar{x}, \bar{y}) \in \Pi(G_1) \times \Pi(G_2)} \varphi(\bar{x}, \bar{y}) \leq \varphi(\bar{x}', \bar{y}') + \varepsilon$  и  $|x'_i - y'_i| \leq \frac{1}{2}$ , где  $\varepsilon = \frac{1}{16}$ .

2. Найти  $\bar{x}^* = \arg \max_{\bar{x} \in \Pi(G_1)} \psi(\bar{x}, \bar{y}')$  и  $\bar{y}^* = \arg \max_{\bar{y} \in \Pi(G_2)} \psi(\bar{x}^*, \bar{y})$ . Здесь  $\bar{x}^*, \bar{y}^*$  целочисленны.

Ответ алгоритма—множество вершин  $\{v \mid x_v^* y_v^* = 1\}$ . Далее везде  $(\bar{x}', \bar{y}')$  и  $(\bar{x}^*, \bar{y}^*)$  будут обозначать пары найденные на первом и втором шагах алгоритма соответственно.

Легко видеть, что все этапы алгоритма полиномиальны. Исследуем его решение.

Введем обозначения  $W = \sum_{v \in V} w_v$  и  $\varphi(\bar{x}', \bar{y}') = \alpha W$ . Ясно, что  $0 \leq \alpha \leq 1$ .

**Предложение 8.** Справедливо

$$\max_{(\bar{x}, \bar{y}) \in \Pi(G_1) \times \Pi(G_2)} \psi(\bar{x}, \bar{y}) - \psi(\bar{x}^*, \bar{y}^*) \leq \left( \frac{1}{4} - \left( \alpha - \frac{1}{2} \right)^2 \right) W + \varepsilon,$$

если  $\alpha \geq \frac{1}{2}$ . А также

$$\max_{(\bar{x}, \bar{y}) \in \Pi(G_1) \times \Pi(G_2)} \psi(\bar{x}, \bar{y}) - \psi(\bar{x}^*, \bar{y}^*) \leq \frac{1}{4} W + \varepsilon,$$

при  $\frac{3}{8} \leq \alpha \leq \frac{1}{2}$ . И

$$\max_{(\bar{x}, \bar{y}) \in \Pi(G_1) \times \Pi(G_2)} \psi(\bar{x}, \bar{y}) - \psi(\bar{x}^*, \bar{y}^*) \leq \left( \frac{1}{4} - \left( \alpha - \frac{3}{8} \right)^2 \right) W + \varepsilon,$$

если  $\alpha \leq \frac{3}{8}$ .

**Доказательство.** Оценим сверху величину  $\varphi(\bar{x}', \bar{y}') - \psi(\bar{x}', \bar{y}')$ , используя факт вогнутости функции  $f(x) = x - x^2$ :

$$\begin{aligned} \varphi(\bar{x}', \bar{y}') - \psi(\bar{x}', \bar{y}') &= \sum_{v \in V} \frac{1}{2} w_v (x'_v - x_v'^2) + \frac{1}{2} w_v (y'_v - y_v'^2) \leq \sum_{v \in V} w_v \frac{(x'_v + y'_v)}{2} \left( 1 - \frac{(x'_v + y'_v)}{2} \right) = \\ &= \sum_{v \in V} w_v \left( \frac{1}{4} - \left( \frac{x'_v + y'_v - 1}{2} \right)^2 \right) \end{aligned}$$

При  $\alpha \geq \frac{1}{2}$ :

$$\alpha W = \varphi(\bar{x}', \bar{y}') = \sum_{v \in V} -\frac{1}{2} w_v (x'_v - y'_v)^2 + \frac{1}{2} w_v y'_v + \frac{1}{2} w_v x'_v \leq \sum_{v \in V} \frac{1}{2} w_v y'_v + \frac{1}{2} w_v x'_v$$

и отсюда:

$$\sum_{v \in V} w_v \frac{(x'_v + y'_v - 1)}{2} \geq \left( \alpha - \frac{1}{2} \right) W.$$

Тогда

$$\varphi(\bar{x}', \bar{y}') - \psi(\bar{x}', \bar{y}') \leq \sum_{v \in V} w_v \left( \frac{1}{4} - \left( \frac{x'_v + y'_v - 1}{2} \right)^2 \right) \leq \frac{1}{4} W - t,$$

где  $t = \min_{\sum_{v \in V} w_v t_v \geq (\alpha - \frac{1}{2}) W} \sum_{v \in V} w_v t_v^2$ . Легко видеть, что  $t = (\alpha - \frac{1}{2})^2 W$ . Итак, получаем, что

$$\varphi(\bar{x}', \bar{y}') - \psi(\bar{x}', \bar{y}') \leq \left( \frac{1}{4} - \left( \alpha - \frac{1}{2} \right)^2 \right) W.$$

Наконец, используя  $\varphi(\bar{x}', \bar{y}') \geq \max_{(\bar{x}, \bar{y}) \in \Pi(G_1) \times \Pi(G_2)} \psi(\bar{x}, \bar{y}) - \varepsilon$  и  $\psi(\bar{x}^*, \bar{y}^*) \geq \psi(\bar{x}', \bar{y}')$ , окончательно получаем:

$$\max_{(\bar{x}, \bar{y}) \in \Pi(G_1) \times \Pi(G_2)} \psi(\bar{x}, \bar{y}) - \psi(\bar{x}^*, \bar{y}^*) \leq \left( \frac{1}{4} - \left( \alpha - \frac{1}{2} \right)^2 \right) W + \varepsilon.$$

Почти аналогично, при  $\alpha \leq \frac{3}{8}$ ,

$$\alpha W = \varphi(\bar{x}', \bar{y}') = \sum_{v \in V} -\frac{1}{2} w_v (x'_v - y'_v)^2 + \frac{1}{2} w_v y'_v + \frac{1}{2} w_v x'_v \geq \sum_{v \in V} \frac{1}{2} w_v y'_v + \frac{1}{2} w_v x'_v - \frac{1}{8} W$$

и отсюда:

$$\sum_{v \in V} w_v \frac{(x'_v + y'_v - 1)}{2} \leq \left( \alpha - \frac{3}{8} \right) W.$$

Тогда,

$$\varphi(\bar{x}', \bar{y}') - \psi(\bar{x}', \bar{y}') \leq \frac{1}{4}W - s,$$

где  $s = \min_{\sum_{v \in V} w_v t_v \leq (\alpha - \frac{3}{8})W} \sum_{v \in V} w_v t_v^2 = (\alpha - \frac{3}{8})^2 W$ . И, наконец,

$$\max_{(\bar{x}, \bar{y}) \in \Pi(G_1) \times \Pi(G_2)} \psi(\bar{x}, \bar{y}) - \psi(\bar{x}^*, \bar{y}^*) \leq \varphi(\bar{x}', \bar{y}') - \psi(\bar{x}', \bar{y}') + \varepsilon \leq \left( \frac{1}{4} - \left( \alpha - \frac{3}{8} \right)^2 \right) W + \varepsilon.$$

Утверждение теоремы для случая  $\frac{3}{8} \leq \alpha \leq \frac{1}{2}$  очевидно. Предложение доказано.

## 6.1 Обсуждение

Как было замечено выше, MaxCMS может рассматриваться как задача нахождения максимального независимого множества в специальном орграфе. Рассмотрим ее, напротив, как задачу минимального вершинного покрытия. Вспомним, что MaxCMS изначально рассматривалась как обобщение задачи поиска монотонной функции наименее отклоняющейся от данных обучающей выборки. 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.

Сравним теоретическую аппроксимирующую способность нашего алгоритма с основным стандартным приближенным алгоритмом для вершинного покрытия в общем случае. Как хорошо известно[6], 2-аппроксимирующий полиномиальный алгоритм для последней задачи может быть построен сведением к линейному программированию. Сведение осуществляется следующим образом: каждой вершине  $v$  (веса  $w_v$ ) графа (орграфа)  $G = (V, E)$  ставим в соответствие переменную  $x_v$  и рассмотрим задачу линейного программирования

$$\begin{aligned} x_i + x_j &\geq 1, (i, j) \in E \\ \sum_v w_v x_v &\rightarrow \min \end{aligned}$$

Тогда множество  $\{v | x_v \geq \frac{1}{2}\}$  будет результирующим вершинным покрытием. Заметим, что существование приближенного алгоритма для вершинного покры-

тия в константой аппроксимации меньшей чем 2 является известной открытой задачей.

Ясно, что слагаемое  $\varepsilon$  в предложении 8 можно сделать сколь угодно малым и в приведенных оценках оно не играет никакой роли, так как оцениваемая величина целочисленна. Чтоб не загромождать запись будем полагать, что  $\varepsilon = 0$ . Обозначим  $\varphi(\bar{x}', \bar{y}') = \alpha W \geq W - \Delta \stackrel{def}{=} \max_{(\bar{x}, \bar{y}) \in \Pi(G_1) \times \Pi(G_2)} \psi(\bar{x}, \bar{y})$ . Здесь  $\Delta$ —вес минимального вершинного покрытия.

Ясно, что 2-аппроксимирующий алгоритм теоретически никак не обоснован в случае, если максимальное допустимое множество имеет вес меньше половины суммы всех весов вершин. Поэтому рассмотрим случай, когда  $\alpha \geq \alpha' \stackrel{def}{=} \frac{MaxCMS}{W} \geq \frac{1}{2}$ . Из предложения 8 получим:

$$\max_{(\bar{x}, \bar{y}) \in \Pi(G_1) \times \Pi(G_2)} \psi(\bar{x}, \bar{y}) - \psi(\bar{x}^*, \bar{y}^*) \leq \alpha(1 - \alpha)W \leq \alpha'(1 - \alpha')W = \alpha' \Delta$$

что означает, что алгоритм аппроксимирует минимальное вершинное покрытие с константой  $1 + \alpha' \leq 2$ . Если рассматривать задачу MaxCMS как задачу удаления «шума» из почти монотонной выборки, то это означает, что при малой зашумленности, то есть при  $\alpha' \approx 1$ , алгоритм будет в худшем случае удалять в 2 раза больше объектов, чем минимально возможное число, равное  $\Delta$ . Это полностью соответствует стандартной оценке аппроксимации. Однако, как показывает оценка, наш алгоритм, при сколь угодно сильно зашумленных выборках, например, если шум составляет половину всех элементов выборки, не удалит больше чем  $\Delta + \frac{1}{4}W$ . Более того, предложение 8 позволяет получить оценку излишне удаленных объектов и для случая, когда шум составляет большую часть выборки ( $\alpha \leq \frac{3}{8}$ ). Таким образом, ограничение задачи MaxCMS на 2-MaxCMS позволяет получить приближенный алгоритм с лучшими оценками аппроксимации чем у классического. Заметим, что этот алгоритм приближенно находит минимальное вершинное покрытие для любого орграфа, множество вершин которого может быть разбито на 2 частичных порядка.

## Список литературы

- [1] Гэри М., Джонсон Д. Вычислительные машины и труднорешаемые задачи. М. Мир. 1982.
- [2] Схрейвер А. Теория линейного и целочисленного программирования: В 2-х т. М.: Мир, 1991. 360 с.

- [3] Хачиян Л.Г. Полиномиальный алгоритм в линейном программировании // Доклады АН СССР, 1979, том 244, с.1093-1096.
- [4] Юдин Д.Б., Немировский А.С. Оценка информационной сложности задач математического программирования // Экономика и математические методы, 1976, т.12, вып. 2, с. 357
- [5] Grotshel M., Lovasz L., Schrijver A. (1988). Geometric algorithms and combinatorial optimization. Springer-Verlag, Berlin Geidelberg New York.
- [6] Hochbaum D. S. Approximation algorithms for the set covering and vertex cover problems // SIAM Journal on Computing, 11:555–556, 1982.
- [7] Mohring R.H. (1985). Algorithmic aspects of comparability graphs and interval graphs. In Graphs and Order, pp.41-101. Dordrecht: Reidel.