ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 37:007]:002.001.36

С. О. Кузнецов

СЛОЖНОСТЬ ОБУЧЕНИЯ И КЛАССИФИКАЦИИ, ОСНОВАННЫХ НА ПОИСКЕ ПЕРЕСЕЧЕНИЯ МНОЖЕСТВ

Изучаются проблемы алгоритмической сложности обучения на основе поиска сходства положительных и отрицательных примеров и построения классификации на основе гипотез, полученных в результате обучения. Рассматривается случай использования дескрипторных языков, т. е. представление примеров множествами. При этом гипотезами будут такие пересечения положительных примеров, которые не являются подмножествами отрицательных примеров. Доказывается $\neq P$-полнота проблемы определения числа всех минимальных гипотез, NP-полнота и полиномиальная разрешимость некоторых проблем порождения гипотез с ограничениями на размер и число подтверждающих примеров. Показана трудноразрешимость проблемы классификации примеров на основе гипотез в общем случае, а также полиномиальная разрешимость важных частных случаев.

1. ВВЕДЕНИЕ

Большинство систем Автоматического Обучения использует то или иное понятие сходства как средство, позволяющее выделять закономерности в исследуемых объектах. Сходство также применяется для классификации новых объектов с помощью найденных закономерностей. Сходство обычно определяется либо как отношение, либо метрически, либо как операция, сопоставляющая нескольким исходным объектам подобъект, выражающий их сходство. Такое определение сходства используется, например, в ДСМ-методе автоматического порождения гипотез (ДСМ-АПГ) [1, 2], где сходство понимается как идемпотентная, коммутативная и ассоциативная операция на парах объектов (т. е. задающая на их множестве полурешетку). Эти достаточно естественные свойства операции сходства позволяют однозначно выражать сходство множества объектов через попарные сходства независимо от порядка расположения объектов в базе данных (см., например, [3, 4]). Примерами полурешеточных операций сходства могут быть, в том числе

— Полурешетка на N-множествах гиперграфов с упорядоченными метками вершин и гиперребер, в которой результат действия операции сходства на паре множеств гиперграфов $\mathcal{G}$ и $\mathcal{H}$ есть множество всех максимальных по вложению общих подгиперграфов гиперграфов из $\mathcal{G}$ и $\mathcal{H}$ [4, 5].

— Интерполяционная полурешетка интервалов, в которой минимальным элементом будет интервал, заключенный между минимальным и максимальным допустимыми значениями, а результатом действия операции сходства на пару интервалов будет третий интервал, нижняя граница которого есть минимум нижних границ первых двух, а верхняя — максимум верхних [4].

В данной работе, являющейся продолжением работ [6, 7], будут рассматриваться проблемы алгоритмической сложности поиска сходства и его использования для классификации в случае представления данных булевыми нижними полурешетками вида $\langle 2^U, \cap, \varnothing\rangle$. Использование таких полурешеток, соответствует представлению данных в виде множеств дескрипторов, причем операцией сходства при этом будет операция пересечения множеств. Результаты о трудноразрешимости (об NP- и $\#P$-трудности), полученные для таких полурешеток, будут свидетельствовать о трудноразрешимости аналогичных задач для других представлений, указанных выше, так как булев случай является частным случаем этих представлений. В случае графового представления даже проблема определения вложимости одного объекта в другой трудноразрешима (в силу NP-полноты задачи ИЗОМОРФИЗМ ПОДГРАФУ [8]).

Дальнейшее изложение статьи построено следующим образом: во втором разделе, в соответствии с [1, 2], дается определение гипотез, рассматриваются функционалы качества гипотез, приводится комбинаторная интерпретация и исследуется сложность задач распознавания и перечисления гипотез, оптимальных в смысле предложенных функционалов. В третьей главе приводится комбинаторная интерпретация и исследуется сложность задач прогноза (классификации), производимого на основе гипотез.

2. ГИПОТЕЗЫ И СЛОЖНОСТЬ ИХ ПОРОЖДЕНИЯ

Пусть нами исследуется некоторое свойство $\mathbb{W}$ объектов из $S \subset 2^U$, для некоторого U — множества структурных элементов. Множество всех объектов из S, про которые известно, что они обладают свойством $\mathbb{W}$, будем обозначать S^+, множество объектов, про которые известно, что они не обладают свойством $\mathbb{W}$, будет обозначать S^-. Множество объектов из S, про которые не известно, обладают ли они свойством $\mathbb{W}$ или нет, будет обозначать S^τ. Таким образом, $S^\tau = S/(S^+ \cup S^-)$.

Тройку $\langle U, S^+, S^-\rangle$ будет называть исходными данными, элементы множества S^+ — положительными примерами, а S^- — отрицательными примерами.

Определение 2.1. $\langle h, \{X_1, \ldots, X_n\}\rangle$ есть глобальное сходство относительно множества $Z \subseteq S$, если $\{X_1, \ldots, X_n\} \subseteq Z$, $n > 1$, $X_1 \cap \ldots \cap X_n = h$ и для произвольного Y: $Y \in Z \setminus \{X_1, \ldots, X_n\}$ имеет место $Y \cap h \neq h$ (таким образом, $\{X_1, \ldots, X_n\}$ есть множество всех объектов из Z, содержащих h, и h есть их пересечение).

Заметим, что это определение эквивалентно определению «понятия» (concept) из [9]. В этой работе и в ее продолжениях, однако, не используется представление об отрицательных примерах, как в следующем определении.

Определение 2.2. $\langle h, \{X_1, \ldots, X_n\}\rangle$ — есть положительная (или (+)-) гипотеза (относительно причины свойства W), если $\langle h, \{X_1, \ldots, X_n\}\rangle$ есть глобальное сходство относительно множества S^+ и h не есть подобъект (в смысле $\subseteq$) какого-либо объекта из S^-*). h будем называть головой гипотезы. Отрицательные (или (—)-гипотезы (о причинах отсутствия свойства W) определяются двойственным образом.

Как было показано в [7], задача подсчета всех гипотез $\#P$-полна, поэтому порождение всех гипотез может наталкиваться на трудности, связанные с экспоненциальными затратами памяти и времени работы ЭВМ. В связи с этим оправдана постановка вопроса о порождении одной, нескольких или всех «самых интересных гипотез». Таковыми, например, могут быть признаны гипотезы, с минимальными по вложению головами. Эти гипотезы подтверждаются большим количеством примеров, чем гипотезы с большими по вложению головами и в то же время, являются более «смелыми»: с их помощью можно произвести больше прогнозов (см. раздел 3). В практике ДСМ-метода отбор минимальных по вложению гипотез позволял значительно (иногда в сотни раз) сокращать количество гипотез. Однако пессимистический результат теоремы 2.5 не позволяет надеяться на эффективное применение этого метода отбора в общем случае.

Другим возможным способом отбора был бы отбор гипотез, на которых достигается минимумы или максимумы следующих функционалов, зависящих от размера головы гипотезы, h и числа подтверждающих примеров, n.

1. $|h|$ — «смелость» гипотезы [6]. Чем меньше гипотеза, тем более сильное предположение о предметной области она выражает, так как способна породить большее количество прогнозов. С другой стороны, чем больше гипотеза, тем она меньше отличается от исходных фактов и тем меньше отличаются от них объекты, которые классифицируются с ее помощью. Таким образом, есть основание в некоторых ситуациях считать эту гипотезу более «надежной» («надежность-1»).

2. n — «надежность-2» [6]. Чем больше подтверждающих примеров, тем надежнее гипотеза.

Поскольку «надежность-1», с одной стороны, и «надежность-2», с другой, находятся в соотношении trade-off, представляется осмысленными также следующие характеристики качества гипотез:

3. $|h|+n$;

4. $q \cdot |h|+n$, $0<q<1$;

5. $|h|+qn$, $0<q<1$;

6. $|h| \cdot n$.

Имеющиеся до настоящего времени результаты о сложности проблем существования гипотез определенного вида относились к случаю с множеством исходных примеров лишь одного знака [6, 7, 10]. Эти результаты можно представить следующей таблицей:

f \ R	$\leqslant$	$=$	$\geqslant$
$\lvert h\rvert$	P	NP	P
n	P	NP	P
$\lvert h\rvert+n$	?	NP	P

где P означает наличие полиномиального алгоритма, решающего проблему, NP—NP-полноту проблемы, ? — открытость проблемы. Так, самый левый верхний элемент таблицы означает, что задача «существует ли гипотеза, у которой $|h| \leqslant K$» имеет полиномиальный разрешающий алгоритм, а элемент, находящийся в нижней строке и среднем столбце означает, что задача «существует ли гипотеза, такая что для нее $|h|+n=K$» NP-полна.

Будем обозначать массовую проблему о существовании гипотезы знака z, с фиксированным ограничением на значения функционала, с множеством входных данных, состоящих, либо из положительных, либо из отрицательных, либо из положительных и отрицательных примеров, четверкой вида $\langle z, f, R, s\rangle$, где

$z \in \{+, -\}$ — знак гипотезы,

$f \in \{|h|, n, |h|+n, q|h|+n, |h|+qn, |h|n\}$ — вид функционала,

$R \in \{\leqslant, =, \geqslant\}$ — тип отношения между значением функционала и параметром,

$s \in \{\{+\}, \{-\}, \{+, -\}\}$ — характеристика множества примеров (включающего либо только положительные примеры, либо только отрицательные примеры, либо и те и другие, соответственно).

Так, кортеж вида $\langle +, q|h|+n, \leqslant, \{+\}\rangle$ соответствует следующей проблеме.

Дано: Множество (+)-примеров S^+, $S^+ \subseteq 2^U$, натуральное число $K \leqslant |U|$.

Определить: Существует ли (+)-гипотеза с головой h и числом примеров n, такая, что $q|h|+n \leqslant K$.

Теорема 2.1. Проблема $\langle +, q|h|+n, \leqslant, \{+\}\rangle$ NP-полна для любого $q: 0<q<1$.

Доказательство. Принадлежность проблемы классу NP очевидна: для предложенного решения в качестве проверки пересекаются все (+)-примеры, содержащие h, полученное пересечение сравнивается с h, а значение $q|h|+n$ сравнивается с K. На эти действия потребуется $0(|U| \cdot |S^+|)$ операций.

Сведем к нашей задаче задачу «3-Сочетание (3—С)» [8]:

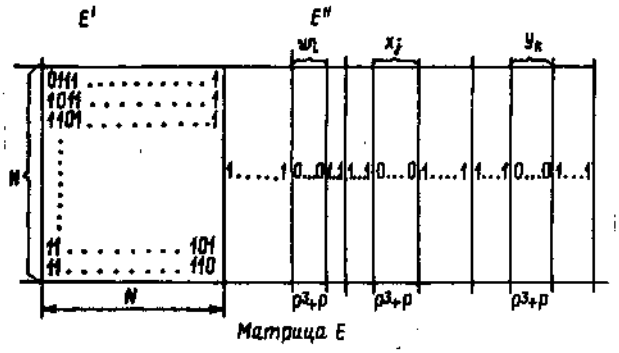
Дано: Множество $M \subseteq W \times X \times Y$, где W, X, Y — непересекающиеся множества, $|W| = |X| = |Y| = P$, $|M| = N$.

Определить: Существует ли множество $M' \subseteq M$ такое, что $|M'| = P$ и никакие два разных элемента M' не имеют равных компонент (M' называется трехмерным сочетанием).

По входным данным задачи 3-С построим следующую бинарную матрицу E размера $N \times (N+3p(p^3+p))$. Правая часть матрицы E — подматрица E' состоит из $3p$ групп столбцов, в каждой группе по p^3+p столбцу. Каждая группа столбцов взаимно-однозначно соответствует некоторому элементу множеств W, X, Y. t-му элементу множества M, т. е. $m_t = (w_i, x_j, y_k)$ из задачи 3-С соответствует t-я строка матрицы E, причем в подматрице E'' в ячейках, соответствующих элементам w_i, x_j, y_k стоят нули, а в остальных ячейках — единицы. Таким образом, подматрица E'' получена из матрицы задачи 3-С [8, с. 83]. (p^3+p)-кратным дублированием

* В терминологии ДСМ-метода «(+)-гипотеза, полученная по правилу с запретом на контрпример». В данной статье нами будут исследоваться только гипотезы такого типа.

столбцов. В левой $N\times N$-подматрице E' матрицы E в t-й ячейке t-й строки стоят нули ($1\leqslant t\leqslant N$), а в остальных ячейках — единицы.



Покажем, что задача 3-С с указанными параметрами сводится к проблеме $\langle +,\ q|h|+n,\ \leqslant,\ \{+\}\rangle$ $(0<q<1)$, где S^+ состоит из N-примеров, каждый из которых соответствует строке матрицы E, единица в строке означает наличие соответствующего элемента из U, ноль — отсутствие, причем $|U|=3p(p^3+p)$, $|S^+|=N$, $K=q(N-p)+p$.

Пусть исходная задача 3-C имеет решение, тогда для некоторых строк матрицы E правые (соответствующие E'') подстроки в произведении образуют нулевой вектор, так как нулевая строка получилась бы в исходной (не «распухшей» в (p^3+p) раз) матрице задачи 3-C. При этом произведение левых (соответствующих матрице E') подстрок этих p строк даст строку с $(N-p)$ единицами и функционал $q|h|+n$ примет значение $q(N-p)+p$, т. е. проблема имеет решение, и наоборот, пусть проблема $\langle +,\ q|h|+n,\ \leqslant,\ \{+\}\rangle$ с параметрами $N+3p(p^3+p)$ (размер U), N (число примеров), $K=q(N-p)+p$ (ограничение значения функционала) имеет решение, т. е. в матрице E найдется r таких строк, что их произведение даст строку, сумма числа единиц в которой с числом r не превышает $q(N-p)+p\leqslant qN+p\leqslant qp^3+p$. Так как $qp^3<p$ при $0<q<1$, то число единиц e в правой части произведения, относящегося к E'', заведомо равно нулю (так как, по построению матрицы E'' число единиц e не может быть таким, что $0<e<p^3+p$. При этом r не может быть меньше p, так как тогда соответствующая задача 3-C имела бы решение: 3-сочетание размера $r<p$, что невозможно. r не может быть больше p, тогда значение функционала $q|h|+n$ было бы $q(N-r)+r=qN+(1-q)r>qN+(1-q)p=q(N-p)+p$, что противоречит нашему предположению о том, что значение функционала не более $q(N-p)+p$. Значит, $r=p$ и для задачи 3-C найдено 3-сочетание размера p. Сводимость доказана. Полиномиальность ее следует непосредственно из полиномиальности размеров матрицы E: матрица E' имеет размер, не превосходящий $p^3\times p^3$ ячеек, а матрица E'' — не превосходящий $p^3\times 3p(p^3+p)$.

Теорема 2.2. Существует алгоритм, решающий проблему $\langle +,\ |h|,\ \geqslant,\ \{+,\ -\}\rangle$ за время $O(|U|\cdot|S^+|^2\times|S^-|)$.

Доказательство. Дадим описание алгоритма, находящего решение за указанное время.

Шаг 1. Строим множество J^+ всех попарных пересечений множеств из S^+.

Шаг 2. Для каждого $X\in J^+$ определяем, существуют ли в S^--множество s, такое, что $s\supset X$. Составляем множество $Y=\{X|X\in J^+,\ \neg\exists s\colon s\in S^-,\ s\supset X\}$.

Шаг 3. Ищем в Y множества по мощности не меньше, чем параметр K. Если такие множества есть, то алгоритм отвечает «да», если нет то алгоритм отвечает «нет». Алгоритм заканчивает работу.

Комментарии к алгоритму. Все максимальные по вложению пересечения содержатся среди попарных пересечений (совпадают с каким-либо из них), так как пересечения большего количества объектов (подмножеств U) могут лишь уменьшить результат пересечения. Если Y содержит множества мощности не меньше K, то они определяют гипотезы со значением $|h|\geqslant K$.

Подсчитаем временную сложность алгоритма.

Шаг 1. $O(|U|\cdot|S^+|^2)$ — поиск всех пар положительных примеров и нахождение их пересечений.

Шаг 2. $O(|U|\cdot|S^+|^2\cdot|S^-|)$ — для каждой пары положительных примеров просматривается все множество отрицательных примеров.

Шаг 3. $O(|U|\cdot|S^+|^2)$ — проверка всего Y, ограниченного по величине множеством пар.

Итоговая временная сложность алгоритма есть $O(|U|\cdot|S^+|^2\cdot|S^-|)$.

Следствие. Проблема $\langle +,\ n,\ \leqslant,\ \{+,\ -\}\rangle$ имеет решающий алгоритм с временной сложностью $O(|U|\cdot|S^+|^2\cdot|S^-|)$.

Доказательство следует из теоремы 2.2 и того факта, что наибольшие по мощности пересечения соответствуют наименьшему числу пересекаемых множеств.

Заметим также, что предложенный алгоритм, дает ответ и на более общий вопрос «Существует ли хотя бы одна гипотеза для заданных множеств положительных и отрицательных примеров?», так как, если все попарные пересечения положительных примеров включаются в отрицательные, то и пересечения большего количества примеров заведомо включаются в отрицательные примеры.

Для дальнейших рассуждений нам потребуется введение некоторых вспомогательных конструкций. Исходной для них будет задача о вершинном покрытии графа $G=\langle V, E\rangle$.

Определение 2.3 Трехдольный граф, ассоциированный с произвольным графом $G=\langle V, E\rangle$ есть граф T следующего вида:

$$T=\langle W^1\cup W^2\cup W^3,\ E'\rangle,\quad |W^1|=|W^2|=|V|,\quad |W^3|=|E|,$$
$$E'\subseteq W^1\times W^2\cup W^2\times W^3.$$

Пара вершин (w_i^1, w_i^2), $w_i^1\in W^1$, $w_i^2\in W^2$ взаимно-однозначно соответствует вершине $v_i\in V$. $(w_i^1, w_j^2)\in E'$, если $i\neq j$. Вершина $w_k^3\in w^3$ взаимно-однозначно соответствует ребру $e_k\in E$. $(w_j^2, w_k^2)\in E'$, если вершина $v_j\in V$ инцидентна ребру $e_k\in E$.

Будем говорить, что в двудольном графе $B=\langle X\cup Y, Z\rangle$ множество вершин $X'\subseteq X$ доминирует над вершинами из $Y'\subseteq Y$, если каждая вершина из Y' смежна с какой-либо вершиной из X'. Общей тенью множества вершин $X'\subseteq X$ назовем множество $Y''\subseteq Y$ всех вершин, с которыми связана каждая вершина из множества X'.

Лемма 2.3. Каждое вершинное покрытие размера K в графе $G=\langle V, E\rangle$ соответствует в трехдольном графе T тройке $\langle C, Z, W^3\rangle$, где $C\subseteq W^1$, $Z\subseteq W^2$, Z есть общая тень вершин из C, которая доминирует над всеми вершинами из W^3, причем $|C|=|W^1|-K=|V|-K$, $|Z|=K$.

Доказательство следует непосредственно из построения графа T. В самом деле, множество вершин Z размера K доминирует над всеми вершинами из W^3 тогда и только тогда, когда оно соответствует подмножеству вершин в графе G, являющимся вершинным покрытием размера K. При этом множество вершин Z будет общей тенью множества вершин C, соответствующего множеству вершин графа G, дополнительному к множеству вершин, соответствующего множеству Z. Следовательно, $C=|W^1|-K=|V|-K$. □

Определение 2.4. Исходными данными, соответствующими трехдольному графу T: $\langle W^1\cup W^2\cup W^3, E'\rangle$, назо-

вем тройку $\langle U, S^+, S^-\rangle$, где $U=W^1\cup W^2\cup W^3$, элементы S^+ соответствуют вершинам из W^1, а элементы S^- — вершинам из W^3, причем положительный пример s_i, взаимно-однозначно соответствующий вершине $w_i{}^1\in W^1$, состоит из объединения множества вершин из W^2, с которыми смежна вершина $w_i{}^1$ и $\{w_i{}^1\}$, т. е. $s_i=\{w_i{}^1\}\cup\{w^2\mid w^2\in W^2,\ (w_i{}^1),\ w^2)\in E'\}$, а отрицательный пример s_k, взаимно-однозначно соответствующий вершине $w_k{}^3\in W^3$, состоит из объединения множества вершин из W^2, с которыми не смежна вершина $w_k{}^3$ и $\{w_k{}^3\}$, т. е. $s_k=\{w_k{}^3\}\cup\{w^2\mid w^2\in W^2,\ (w_k{}^3,\ w^2)\notin E'\}$.

Лемма 2.4. Пусть тройка $\langle C, Z, W^3\rangle$ множеств вершин графа T из определения 2.3 такова, что $C\subseteq W^1$, $|C|>1$, $Z\subseteq W^2$ есть общая тень вершин из C, которая доминирует над всеми вершинами из W^3, а C максимальное по вложению множество вершин, общая тень которых есть Z. Тогда пара $\langle Z, \{\{w^2\mid(w^2,\ w_i{}^1)\ E'\}: w_i{}^1\in C\}\rangle$ есть (+)-гипотеза, полученная при исходных данных, соответствующих трехдольному графу T по определению 2.4.

Доказательство. Рассмотрим пару $\langle Z, \{\{w^2\mid(w^2,\ w_i{}^1)\ E'\}: w_i{}^1\in C\}\rangle$. Элементы вида $w_i{}^1$, $w_k{}^3$ были введены в U для того, чтобы сделать различными примерами (это требуется определением 2.2) те, которые соответствуют вершинам из w^1 (или w^3), смежные с одними и теми же вершинами из w^2. Так как элементы $w_i{}^1$ и $w_k{}^3$ различны у всех примеров, то они не входят ни в одно пересечение. Пересечением всех множеств вида $\{w^2\mid(w^2,\ w_i{}^1)\in E'\}$ для $w_i\in C$ будет в точности множество $\{w^2\mid w^2\in Z\}$, т. е. множество Z, поскольку Z есть общая тень вершин из C. С другой стороны, среди вершин из W^1 нет других вершин, связанными со всеми вершинами из Z, так как C максимально по вложению в силу условия леммы 2.4. Таким образом, пара $\langle Z, \{\{w^2\mid(w^2,\ w_i{}^1)\in E'\}: w_i{}'\in C\}\rangle$ есть глобальное сходство (+)-примеров из S^+. Так как каждый элемент Z по условию леммы 2.4 не содержится хотя бы в одном (—)-примере, то Z не содержится ни в одном (—)-примере, и указанная пара есть (+)-гипотеза в соответствии с определением 2.2. □

Очевидно, что выполнимо также и обратное: по исходным данным $\langle U, S^+, S^-\rangle$ можно построить трехдольный граф T, в котором гипотезам будут соответствовать максимальные по вложению полные двудольные подграфы (при доминировании вершин правой доли).

Пример. Рассмотрим граф T, изображенный на рис. 1, где вершины средней доли помечены как A, B, C, D, E, F, G.

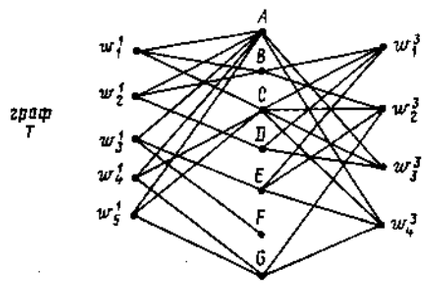


Рис. 1

Тогда в соответствующей задаче о гипотезах множество (+) — примеров будет $S^+=\{X_1, X_2, X_3, X_4, X_5\}$, $S^-=\{Y_1, Y_2, Y_3, Y_4\}$, где

$X_1=\{A, B, C, w_1{}^1\}$, $X_2=\{A, B, D, w_2{}^1\}$, $X_3=\{A, E, F, w_3{}^1\}$, $X_4=\{A, C, G, w_4{}^1\}$, $X_5=\{A, C, G, w_5{}^1\}$;

$Y_1=\{A, F, G, w_1{}^3\}$, $Y_2=\{A, D, F, w_2{}^3\}$, $Y_3=\{B, E, F, G, w_3{}^3\}$, $Y_4=\{B, D, F, w_4{}^3\}$. Глобальными сходствами (+)-примеров будут пары $\langle A, \{X_1, X_2, X_3, X_4, X_5\}\rangle$, $\langle AB, \{X_1, X_2\}\rangle$, $\langle AC, \{X_1, X_4, X_5\}\rangle$, $\langle ACG, \{X_4, X_5\}\rangle$. Из них (+)-гипотезами будут вторая, третья и четвертая пары. В случае первой пары, вершина средней доли с пометкой A не доминирует первую и вторую вершины правой доли.

Теорема 2.5. Следующая задача «Число гипотез, минимальных по вложению» $\#P$-полна (определение $\#P$-полноты см. в [11]).

Дано: S^+, S^- — множества (+)- и (—)-примеров.

Найти: $\#\{H=\langle h, \{X_1, \ldots, X_n\}\rangle$: (+)-гипотеза и не существует (+)-гипотезы $H'=\langle h', \{X_1', \ldots, X_s'\}\rangle$ такой, что $h'\subseteq h\}$.

Доказательство. Сведем к нашей задаче задачу о числе минимальных по вложению вершинных покрытий [12]:

Дано: Граф $G=\langle V, E\rangle$.

Найти: $\#\{V'\subseteq V\mid$ «$(u, v)\in E\to u\in A$ или $v\in A$» имеет место для $A=V'$, но ни для одного $A\subset V'\}$.

По построению, проводимому в лемме 2.3, минимальное по вложению вершинное покрытие в графе G будет соответствовать такой тройке $\langle C, Z, W^3\rangle$ подмножеств вершин трехдольного графа T, что Z, будучи общей тенью вершин из C, в то же время является минимальным по вложению множеством вершин из W^2, доминирующих над W^3. И наоборот, каждая тройка указанного вида соответствует минимальному по вложению вершинному покрытию в графе G. По лемме 2.4 каждая тройка указанного вида взаимно-однозначно соответствует гипотезе, удовлетворяющей запрету на контрпример при входных данных $\langle U, S^+, S^-\rangle$, где U, S^+, S^- такие, как указано в лемме 2.4. Минимальности (по вложению) Z при этом будет соответствовать минимальность (по вложению) h. □

Теорема 2.6 Проблема $\langle +,\ |h|,\ \leqslant,\ \{+, -\}\rangle$ NP-полна.

Доказательство. Принадлежность проблемы классу NP очевидна: для каждого потенциального решения, т. е. предложенной гипотезы, достаточно пересечеть все (+)-примеры, содержащие h, сравнить полученное пересечение с h и, в случае совпадения, проверить невхождение h во все (—)-гипотезы и сравнить $|h|$ с K. Все эти действия можно осуществить за время $O(|U|\cdot(|S^+|+|S^-|))$. Сведем к указанной задаче задачу «минимальное вершинное покрытие» из [8]:

Дано: Граф $G=\langle V, E\rangle$, натуральное число $K\leqslant|V|$.

Определить: Существует ли множество $V'\subseteq V$, такое, что $|V'|\leqslant K$, и для произвольного $e=(v_i, v_j)\in E$ имеет место «$v_i\in V'$ или $v_j\in V'$».

Построим по графу G трехдольный граф T способом, указанным в определении 2.3. При этом по лемме 2.3 вершинному покрытию размера K графа G в графе T будет соответствовать тройка $\langle C, Z, W^3\rangle$ такая, что $|C|=|V|-K$, $|Z|=K$, $|W^3|=|E|$, причем множество Z есть общая тень множества вершин C, которая доминирует над множеством W^3. По лемме 2.4 эта тройка соответствует гипотезе с запретом на контрпример размера K, образованной $|V|-K$ положительными примерами на входных данных, соответствующих графу T по определению 2.4. Сводимость осуществляется за время $O(|V|+|E|)$.

Следствие. Проблема $\langle +,\ n,\ \geqslant,\ \{+, -\}\rangle$ NP-полна.

Доказательство следует из теоремы 2.6 и того факта, что наибольшие по мощности пересечения соответствуют наименьшему числу пересекаемых множеств.

Теорема 2.7. Проблема $\langle +,\ |h|+n,\ \geqslant,\ \{+, -\}\rangle$ NP-полна.

Напомним, что частный случай этой задачи — при отсутствии отрицательных примеров сводится к полиномиально разрешимой задаче поиска размера максимального паросочетания [7, 10]. Полиномиальный алгоритм нахождения гипотез с максимальным значением $|h|+n$ для случая $S^-=\varnothing$ приводится в [10].

Доказательство. В соответствии с леммой 2.4 данная проблема эквивалентна следующей.

Дано: Трехдольный граф $T'=(V_1 \cup V_2 \cup V_3, E'')$, $E'' \subseteq V_1 \times V_2 \cup V_2 \times V_3$, натуральное число $k \leqslant |V_1|+|V_2|$.

Определить: Существует ли максимальный по вложению полный двудольный подграф $B'=(V_1' \cup V_2', E_1)$ графа T' такой, что $V_1' \subseteq V_1$, $V_2' \subseteq V_2$, $E_1=V_1' \times V_2'$, $|V_1'|+|V_2'| \geqslant k$, а V_2' доминирует над V_3.

Сведем к этой проблеме проблему о «минимальном вершинном покрытии» (см. теорему 2.6). По графу G строим ассоциированный с ним трехдольный граф $T=(W_1 \cup W^2 \cup W^3, E')$ в соответствии с определением 2.3. По графу T строим следующий трехдольный граф $T'=(V_1 \cup V_2 \cup V_3, E'')$, $E'' \subseteq V_1 \times V_2 \cup V_2 \times V_3$, $|V_1|=n \cdot |W^1|$, $|V_2|=|W^2|$, $|V_3|=|W^3|$, $V_1=V_1^1 \cup \ldots \cup V_1^n$, где для любого i: $1 \leqslant i \leqslant n$, $|V_1^i|=|W^1|$ и подграф, индуцированный множествами вершин V_1^i, V_2, V_3 изоморфен графу T. Таким образом, максимальному по вложению полному двудольному подграфу (МВПДП) графа T на вершинах $A \subseteq W^1$, $B \subseteq W^2$ соответствует МВПДП графа T' на вершинах $A' \subseteq V_1$, $B' \subseteq V_2$, где $|A'|=n \cdot |A|$.

Покажем, что в произвольном графе G существует вершинное покрытие размера не более $K \leqslant |V|=n$ тогда и только тогда, когда в построенном по G трехдольном графе T' имеется полный двудольный подграф $B=(V_1' \cup V_2', E_1)$ такой, что $V_1' \subseteq V_1$, $V_2' \subseteq V_2$, $E_1=V_1' \times V_2'$, $|V_1'|+|V_2'| \geqslant k=n \cdot (n-K)+1$, а V_2' доминирует над V_3.

В самом деле, пусть в графе G нашлось вершинное покрытие размера не более K. При этом в графе T'' найдется МВПДП $(V_1' \cup V_2', V_1' \times V_2')$, такой, что $V_1' \subseteq V_1$, $V_2' \subseteq V_2$, $1 \leqslant |V_2'| \leqslant K$, и V_2' доминирует над V_3. При этом V_1' будет не менее $n \cdot (n-K)$, а $|V_1'|+|V_2'| \geqslant n \cdot (n-K)+1$.

Наоборот, пусть в графе T' нашелся МВПДП $(V_1' \cup V_2', V_1' \times V_2')$ такой, что $V_1' \subseteq V_1$, $V_2' \subseteq V_2$, $1 \leqslant V_2' \leqslant K$, V_2' доминирует над V_3, а $|V_1'|+|V_2'| \geqslant n \cdot (n-K)+1$. Поскольку $|V_2'| \leqslant n$, то $|V_1'| \geqslant n \cdot (n-K)-n+1$. Данному МВПДП графа T' в графе T будет соответствовать МВПДП $B=(W_1' \cup W_2', W_1' \times W_2')$, такой, что $W_1' \subseteq W^1$, $W_2' \subseteq W^2$ и $|W_1'|=|W_1'|/n$. Следовательно, $|W_1'| \geqslant [(n \cdot (n-K)-n+1)/n]+1 \geqslant n-K$. Значит, в силу определения графа T (определение 2.3) $|W_2'| \leqslant K$ и, в силу леммы 2.3, в графе G есть вершинное покрытие размером не более K.

Теорема 2.8. Проблема $\langle +, |h|+n, \leqslant, \{+\}\rangle$ NP-полна.

Доказательство. Принадлежность проблемы классу NP очевидна. Графовой интерпретацией [7, лемма 1] данной проблемы будет следующая проблема: в произвольном двудольном графе $B=\langle V_1 \cup V_2, E\rangle$ найти максимальный по вложению полный двудольный подграф (МВПДП) с числом вершин не более K, т. е. максимальный по вложению граф вида $\langle V_1' \cup V_2', E'\rangle$, где $V_1' \subseteq V_1$, $V_2' \subseteq V_2$, $E'=V_1' \times V_2' \subseteq E$. Сведем к указанной задаче NP-полную задачу «минимальное по мощности максимальное паросочетание» [2, с. 239]:

Дано: Двудольный граф $B=\langle W^1 \cup W^2, E\rangle$, натуральное число $K \leqslant |E|$.

Определить: Существует ли максимальное по вложению паросочетание M размера $|M| \leqslant K$.

По графу $B=\langle W^1 \cup W^2, E\rangle$ строим двудольный граф $B'=\langle V_1 \cup V_2, E'\rangle$, $|V_1|=|V_2|=E$. Ребру e_i графа B в графе B' будет взаимно-однозначно соответствовать пара вершин (v_1^i, v_2^i): $v_1^i \in V_1$, $v_2^i \in V_2$, $(v_1^i, v_2^i) \in E'$ тогда и только тогда, когда либо ребра e_i и e_j из E не инцидентны, либо $i=j$. При этом произвольному паросочетанию в B будет соответствовать полный двудольный подграф в B' и наоборот. При сведении максимальность по вложению сохраняется и, таким образом, максимальные по вложению паросочетания графа B, по мощности не большие K, взаимно-однозначно соответствуют максимальным по вложению полным двудольным подграфам графа B' с числом вершин не большим $2K$. Сводимость осуществляется за $O(|V|+|E|)$ операций.

3. СЛОЖНОСТЬ АЛГОРИТМОВ ПРОГНОЗА

В этой главе будут рассмотрены вопросы алгоритмической сложности, относящиеся к проблеме прогноза или классификации объектов из S^τ на основе порожденных (+)- и (—)-гипотез. Определения даются в соответствии с [1].

Определение 3.1. Объект $P \in S^\tau$ назовем (+)-прогнозом*, если существует (+)-гипотеза $\langle h, \{X_1, \ldots, X_n\}\rangle$, такая, что $h \subseteq P$, и для любой (—)-гипотезы $\langle h', \{Y_1, \ldots, Y_k\}\rangle$ имеет место $h' \not\subseteq P$.

Отрицательный прогноз ((—)-прогноз) определяется двойственным образом. Для удобства введем следующие вспомогательные определения.

Определение 3.2. (+)-гипотеза $\langle h_1, \{X_1, \ldots, X_n\}\rangle$ есть гипотеза в пользу положительного прогноза для объекта $P \in S^\tau$, если $h_1 \subseteq P$.

Определение 3.3. (—)-гипотеза $\langle h_2, \{Y_1, \ldots, Y_m\}\rangle$ есть гипотеза против положительного прогноза для объекта $P \in S^\tau$, если $h_2 \subseteq P$. Таким образом, объект $P \in S^\tau$ будет (+)-прогнозом, если для него есть гипотеза в пользу положительного прогноза и нет гипотез против положительного прогноза.

Определение 3.1 имеет очевидную алгоритмическую реализацию: сперва порождаются множества (+) и (—) гипотез, затем анализируются вхождения полученных гипотез в объекты из S^τ. Такая реализация, однако, страдает не менее очевидным недостатком: если число гипотез экспоненциально (напомним, что задача «число всех гипотез» $\#P$-полна [6]), то количество времени и памяти, необходимые для классификации даже одного объекта из S^τ, заведомо экспоненциальны.

Для реализации определения 3.1 может быть предложен другой алгоритм. Опишем его частный случай при $S^-=\varnothing$.

Пусть $P \in S^\tau$—вопрос, т. е. объект, для которого нужно провести прогноз. Пусть $P=\{p_1, \ldots, p_t\} \subseteq U$.

Шаг 0. $i:=1$.

Шаг 1. Находим все (+)-примеры, содержащие p_i, вычисляем их пересечение h_i. Если нет по крайней мере двух (+)-примеров, содержащих p_i, то переходим к шагу 4.

Шаг 2. Если $h_i \subseteq P$, то P классифицируется по определению 3.1 положительно. Алгоритм прекращает работу.

Шаг 3. Если пересечение всех примеров, содержащих p_i не является подмножеством P, то P не может быть классифицировано на основе примеров, содержащих p_i, так как пересечения меньшего количества примеров тем более не могут быть подмножествами P.

* Или (+)-гипотезой II рода в терминологии ДСМ-метода [1].

Шаг 4. Если $i=t$, то классификация P невозможна и алгоритм прекращает работу, иначе переходим к шагу 5.

Шаг 5. $i:=i+1$, переходим к шагу 1.

Данный алгоритм позволяет проводить (+)-прогноз (или делать заключение о том, что прогноз не возможен) для объекта P за время $O(t\cdot|S^+|\cdot|U|)$. Можно ли ограничиться порождением полиномиального подмножества гипотез в случае, когда $S^-=\varnothing$? Для облегчения ответа на этот вопрос предложим следующую комбинаторную интерпретацию задачи получения прогноза.

Определение 3.4 Задачей «о доминировании долями полных графов (ДДПГ)» назовем следующую задачу

Дано: Четырехдольный граф $G=\langle V_1\cup V_2\cup V_3\cup V_4, E\rangle$, $E\subseteq(V_1\times V_2)\cup(V_2\times V_3)\cup(V_3\times V_4)$. Графы B_1, B_2, B_3 суть подграфы графа G, индуцированные множествами вершин $(V_1\cup V_2)$, $(V_2\cup V_3)$, $(V_3\cup V_4)$, соответственно.

Определить: Существует ли в подграфе B_2 графа G полный подграф $B'=\langle V_2'\cup V_3',\ V_2'\times V_3'\rangle$, максимальный по вложению, такой, что $V_2'\subseteq V_2$, $V_3'\subseteq V_3$, и множество вершин V_2' доминирует над V_1, а множество вершин V_3' доминирует над V_4, $|V_2'|>1$, $V_3\neq\varnothing$.

Определение 3.5. Задачей «о гипотезе в пользу положительного прогноза (ГППП)», соответствующей задаче ДДПГ, назовем следующую задачу.

Дано. Входные данные $\langle U, S^+, S^-\rangle$, вопрос $P\in S^\tau$, где $U=V_1\cup V_3\cup V_2\cup V_4$, $S^+=\{X_i=\{v_i^2\}\cup\{v_1,\ldots,v_r\}\,|\,\{v_1,\ldots,v_r\}$ — объединение множества всех вершин из V_3, смежных с вершиной $v_i^2\in V_2$ и множества всех вершин из V_1, не смежных с вершиной $v_i^2\in V_2$; $S^-=\{Y_k=\{v_k^4\}\cup V_3\setminus\{w_i^3,\ldots,w_q^3\},\ \{w_i^3,\ldots,w_q^3\}$ — множество всех вершин из V_3, смежных с вершиной $v_k^4\in V_4\}$, $P=V_3$.

Определить. Существует ли (+)-гипотеза $\langle h, \{X_1,\ldots,X_n\}\rangle$, такая, что $h\subseteq P=V_3$, ((т. е. h есть гипотеза в пользу положительного прогноза для вопроса P).

Лемма 3.1 Для четырехдольного графа G, имеющего вид, указанный в определении 3.4 задача о ДДПГ имеет решение тогда и только тогда, когда имеет решение соответствующая задача о ГППП.

Доказательство. Заметим сначала, что, как и в лемме 2.4, множества элементов вида v_i^2, v_k^4 введены в U для того, чтобы в S^+ и S^- не было одинаковых примеров. В пересечения эти элементы не входят.

1. Пусть $\langle h, \{X_1,\ldots,X_n\}\rangle$ есть (+)-гипотеза, $h\subseteq P$. Тогда в графе G подграф, индуцированный вершинами $v_{i_1}^2,\ldots,v_{i_n}^2\in V_2$, соответствующими $\{X_1,\ldots,X_n\}$ и вершинами из V_3, соответствующими h, будет максимальным по вложению полным двудольным подграфом [7, лемма 1]. Множество вершин $v_{i_1}^2,\ldots,v_{i_n}^2\in V_2$ будет доминировать над V_1. В самом деле, пусть некоторая вершина $v_1\in V_1$ не смежна ни с одной вершиной из $\{v_{i_1}^2,\ldots,v_{i_n}^2\}$. Тогда, по определению (+)-примеров, $v_1\in X_1,\ldots,v_1\in X_n$ и $X_1\cap\ldots\cap X_n\not\subseteq P$ (так как $v_1\notin P$). Пусть h соответствует вершинам $w_1^3,\ldots,w_{|h|}^3$ в графе G. Тогда множество вершин $\{w_1^3,\ldots,w_{|h|}^3\}$ доминирует над множеством V_4. Пусть это не так и некоторая вершина $v_j^4\in V_4$ не смежна ни с одной вершиной из $\{w_1^3,\ldots,w_{|h|}^3\}$. Тогда, по определению (—)-примеров, для произвольного (—)-примера Y_j имеет место $w_1^3\in Y_j,\ldots,w_{|h|}^3\in Y_j$ и $h\subseteq Y_j$, что противоречит тому, что гипотеза $\langle h, \{X_1,\ldots,X_n\}\rangle$ была получена по правилу «с запретом на контрпример» (определение 2.2).

2. Пусть $V_2'\subseteq V_2$, $V_3'\subseteq V_3$ — такие множества вершин графа, что индуцированный этими вершинами двудольный граф является полным максимальным по вложению, множество вершин V_2' доминирует над V_1, а V_3' — над V_4. Тогда, по определению (+)- и (—)-примеров, задаваемых графом G, т. е. $\langle V_2', V_3'\rangle$ будет соответствовать некоторой (+)-гипотезе $\langle h, \{X_1,\ldots,X_n\}\rangle$, полученный по определению 2.2. В самом деле, так как двудольный граф, индуцируемый вершинами V_2', V_3', есть МВДП (см. Теорему 2.8), то он соответствует глобальному сходству (+)-примеров. Осталось показать, что это сходство лежит в множестве P и не имеет контрпримеров. Первое выполняется в силу определения (+)-примеров по графу G из определения 3.4 и того, что V_2' доминирует над V_1. В самом деле, пусть $X_1\cap\ldots\cap X_n=h\not\subseteq P$. Тогда в множестве V_1 найдется вершина $v\notin h$. Но тогда, по определению X_i, вершина v не связана ни с одной вершиной из V_2', что противоречит тому, что V_2' доминирует над V_1. То, что гипотеза $\langle V_2', V_3'\rangle$ не имеет контрпримеров, следует непосредственно из определения (—)-примеров по графу G и того, что V_3' доминирует над V_4. □

Пример. Рассмотрим граф, изображенный на рис. 2, где вершины первой доли помечены как C, F, G, в

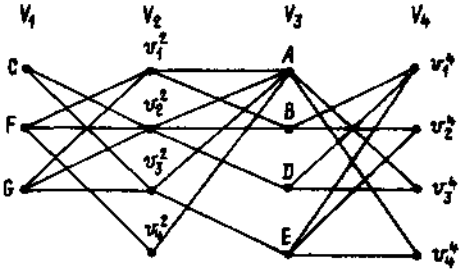


Рис. 2

вершины третьей доли помечены как A, B, D, E. Тогда в соответствующей задаче о прогнозе $P=\{A, B, D, E\}$, множество (+)-примеров и множество (—)-примеров будут, соответственно

$S^+=\{X_1, X_2, X_3, X_4\}$, $S^-=\{Y_1, Y_2, Y_3, Y_4\}$, где

$X_1=\{A, B, C, v_1^2\}$, $X_2=\{A, B, D, v_2^2\}$, $X_3=\{A, E, F, v_3^2\}$, $X_4=\{A, C, G, v_4^2\}$,

$Y_1=\{A, v_1^4\}$, $Y_2=\{A, D, v_2^4\}$, $Y_3=\{B, E, v_3^4\}$, $Y_4=\{B, D, v_4^4\}$.

Глобальными сходствами (+)-примеров будут пары $\langle A, \{X_1, X_2, X_3, X_4\}\rangle$, $\langle AB, \{X_1, X_2\}\rangle$, $\langle AC, \{X_1, X_4\}\rangle$. Из них лишь вторая пара будет свидетельствовать в пользу положительного прогноза для P, т. к. в первом случае не доминируются первая и вторая вершины четвертой доли (голова гипотезы входит в (—)-пример: нарушается определение 2.2), а в третьем случае не доминируется вершина с меткой C (голова гипотезы не входит в P).

Покажем, что при произвольных входных данных и произвольном вопросе $P\in S^\tau$ задача «существует (+)-гипотеза, в пользу положительного прогноза для P» NP-полна. В силу двойственности (+)- и (—)-гипотез это будет также означать, что задача «Объект $P\in S^\tau$, представленный на прогноз, доопределяется положительно с помощью некоторой гипотезы, удовлетворяющей запрету на контрпример» D_P-полна [13]. Для определения (—)-гипотез (не двойственного определе-

нию (+)-гипотез), приведенного в работе [2] все (−)-гипотезы находятся за полиномиальное время и проблема прогноза является *NP*-полной.

Установленная в лемме 3.1 эквивалентность, позволяет переформулировать задачу о прогнозе в виде задачи о четырехдольном графе.

Теорема 3.2. Задача ДДПГ *NP*-полна.

Доказательство. Рассмотрим следующий частный случай задачи ДДПГ. Пусть $|V_2| = |V_3| = n$; $\forall i, j: 1 \leqslant i, j \leqslant n$; $v_j^2 \in V_2$, $v_i^3 \in V_3$; $(v_i^2, v_j^3) \in E$ тогда и только тогда, когда $i \neq j$, а двудольные графы, индуцированные множествами вершин $V_1 \cup V_2$ и $V_3 \cup V_4$, изоморфны. В этом случае все множество максимальных по вложению полных двудольных подграфов будет состоять из графов вида $\langle \{v_{i_1}^2, \ldots, v_{i_k}^2\} \cup \{v_{j_1}^3, \ldots, v_{j_m}^3\}, E' \rangle$, где $E' = \{v_{i_1}^2, \ldots, v_{i_k}^2\} \{v_{j_1}^3, \ldots, v_{j_m}^3\}$, а $\{j_1, \ldots, j_m\} = \{1, \ldots, n\} \setminus \{i_1, \ldots, i_k\}$, т. е. множество индексов вершин из V_2 дополнительно к множеству индексов вершин из V_3. Учитывая, что двудольные графы, индуцированные множествами вершин V_1, V_2, V_3, V_4 изоморфны, указанный частный случай задачи ДДПГ эквивалентен следующей задаче ДВМВ.

Задача «доминирование взаимодополняющими множествами вершин (ДВМВ)»

Дано: Двудольный граф $B = \langle W_1 \cup W_2, E \rangle$, $E \subseteq W_1 \times W_2$.

Определить: Существует ли множество $W_1' \subseteq W_1$ такое, что оба множества W_1', $W_1 \setminus W_1'$ доминируют над W_2.

Лемма 3.3. Задача ДВМВ *NP*-полна.

Доказательство [А. А. Карзанов]. Сведем к задаче ДВМВ задачу о выполнимости КНФ [8]:

Дано: КНФ $C = D_1 \wedge \ldots \wedge D_n$, $D_i = (\neg) x_{i_1} \vee \ldots \vee \vee \ldots \vee (\neg) x_{i_k}$, где для произвольных i, j $x_{i_j} \in X = \{x_1, \ldots, x_m\}$.

Определить: Существует ли булев набор, выполняющий C.

По КНФ F построим двудольный граф $B = \langle V_1 \cup V_2, E \rangle$, $E \subseteq V_1 \times V_2$, $|V_1| = 2m+1$, $|V_2| = n+m$. В множестве вершин V_1 каждой переменной x_i взаимно-однозначно сопоставлена пара вершин (v_{i_1}, v_{i_2})—для x_i и для $\neg x_i$, соответственно. В множестве вершин V_2 каждой дизъюнкции D_j сопоставляется вершина v_j, $1 \leqslant j \leqslant n$, а каждой переменной x_i сопоставляется вершина v_i, $n+1 \leqslant i \leqslant n+m$. Пара вершин (v_i, v_j), где $v_i^1 \in V_1$, $v_j^2 \in V_2$ связана ребром тогда и только тогда, когда имеет место один из следующих случаев:

1) v_i^1 соответствует литералу $(\neg) x_{i_1}$, входящему в дизъюнкцию D_{j_2}, которая соответствует вершине v_{j_2};

2) v_i^1 соответствует литералу $(\neg) x_{i_1}$, а вершина v_{j_2} соответствует переменной x_{i_1} (т. е. $j_2 = n + i_1$);

3) $i = 2m+1$, $1 \leqslant j_2 \leqslant n+m$.

Других ребер нет.

Покажем, что булевой набор, выполняющий КНФ C, существует тогда и только тогда, когда в графе B найдется такое множество вершин $V_1' \subset V_1$, что оба множества V_1' и $V_1 \setminus V_1'$ доминируют над V_2, т. е. соответствующая задача ДВМВ имеет решение.

В самом деле, пусть C выполняется на наборе $(a_1, \ldots, a_n)$, где элементы набора $a_{i_1}, \ldots, a_{i_k}$ единичные, а остальные — нулевые. Тогда все вершины из V_2 будут доминироваться вершинами из $V_1' \subset V_1$, которые соответствуют литералам, выполняющим соответствующие дизъюнкции. Поскольку вершина v_{2m+1} связана со всеми вершинами из $\{v_1^2, \ldots, v_n^2\}$, а вершины $v_{n+1}^2, \ldots, v_{n+m}^2$ доминируются теми вершинами из $V_1 \setminus V_1'$, которые соответствуют остальным литералам, то и $V_1 \setminus V_1'$ будет доминировать над V_2.

Обратно, пусть некоторое множество $V_1' \subset V_1$ таково, что V_1' и $V_1 \setminus V_1'$ доминируют над V_2. Пусть одно из этих множеств, скажем $V_1 \setminus V_1'$, содержит вершину v_{2m+1}. Вершины, соответствующие противоположным литералам, при этом могут входить либо в V_1' либо в $V_1 \setminus V_1'$ (иначе не будут доминироваться вершины v_j, $n+1 \leqslant j \leqslant n+m$, которые связаны лишь с парой вершин). Поэтому можно построить булев набор, положив единичным каждый литерал, соответствующий вершине, входящей в V_1', а остальные литералы положив нулевыми. Так построенный набор будет выполнять КНФ C. В самом деле, пусть это не так, тогда найдется невыполненная дизъюнкция D. Но вершину, ей соответствующую, доминирует некоторая вершина из V_1, и соответствующий этой вершине литерал должен быть положительным, т. е. выполнять D. Сводимость доказана. Ее полиномиальность, а также принадлежность задачи классу *NP* очевидны. □

Заметим, что в любом из вырожденных случаев, когда

либо $V_1 = \varnothing$ $(U = P)$,

либо $V_2 = V_3$ $(S^+ = P)$,

либо $V = \varnothing$ $(S^- = \varnothing$, см. выше),

когда четырехдольный граф становится трехдольным, для задачи ДДПГ имеется разрешающий алгоритм полиномиальной сложности.

Укажем еще одну ситуацию, когда возможен полиномиальный алгоритм прогноза.

Предположим, что размер множества P фиксирован. Это предположение достаточно оправдано во многих практических ситуациях, например в задаче «структура—активность» (см., например, [1, 7]), в которой прогнозируется принадлежность определенного химического соединения (представленного множеством дескрипторов) классу активных или неактивных соединений. При этом размер соединений оправдано считать ограниченным и уж во всяком случае это так, когда рассматривается последовательность прогнозов для одного соединения при растущем множестве примеров и элементов описания (т. е. элементов U).

Тривиальным алгоритмом, решающим за полиномиальное время задачу о прогнозе, был алгоритм, который последовательно перебирал бы все подмножества P, вычислял пересечения всех положительных примеров, содержащих данное подмножество, проверял бы такие пересечения на вложение в отрицательные примеры. В случае нахождения пересечения, не входящего ни в один отрицательный пример, алгоритм преступал бы к аналогичным действиям с отрицательными примерами. Сложность такого алгоритма не превышает $O(2^{|P|}(|S^+| + |S^-|) \cdot |U|)$. Более эффективный алгоритм, квадратичный от числа гипотез, головы которых содержатся в P, можно построить на основе алгоритма МП [14].

ЗАКЛЮЧЕНИЕ

Результаты относительно сложности проблем определения существования гипотез с ограничениями на размер и число подтверждающих примеров, полученные в данной и более ранних работах [6, 7, 10] можно представить в следующей таблице:

Здесь, как и прежде, P означает наличие полиномиального алгоритма, *NP*—*NP*-полноту, ? — открытость проблемы. Элементы таблицы, написанные через запятую, означают состояние проблемы в случае только примеров одного знака и в случае наличия примеров обоих знаков, соответственно (если проблема *NP*-полна в случае наличия примеров только одного

знака, то она NP-полна и для случая с примерами обоих знаков), при этом в таблице стоит только одна запись «NP».

f \ R	$<$	$=$	$>$
$\lfloor \frac{n}{2} \rfloor$	P, NP	NP	P, P
n	P, P	NP	P, NP
$\lfloor \frac{n}{2} \rfloor + n$	NP	NP	P, NP
$q \cdot \lfloor \frac{n}{2} \rfloor + n$	NP	?	?

В разделе 2 была показана также $\neq P$-полнота задачи «число минимальных по вложению гипотез».

В разделе 3 была показана Dp-полнота задачи о прогнозе в общем случае и полиномиальная разрешимость ее в случае фиксированного размера объекта, представленного для классификации.

* * *

Автор выражает признательность А. А. Карзанову за представление доказательства леммы 3.3 и Д. П. Скворцову за указание ряда неточностей в начальном варианте статьи.

СПИСОК ЛИТЕРАТУРЫ

1. Финн В. К. Правдоподобные выводы и правдоподобные рассуждения // Итоги науки и техники. Сер. Теория вероятностей. Математическая статистика. Теоретическая кибернетика. Т. 28.— М.: ВИНИТИ, 1988.— С. 3—84.
2. Финн В. К. Правдоподобные рассуждения в интеллектуальных системах типа ДСМ // Итоги науки и техники. Сер. Информатика. Т. 15 (Интеллектуальные информационные системы).— М.: ВИНИТИ, 1991.— С. 54—101.
3. Гусакова С. М., Финн В. К. Сходство и правдоподобный вывод // Изв. АН СССР. Сер. Техническая кибернетика.— 1987.— № 5.— С. 42—63.
4. Кузнецов С. О. ДСМ-метод как система автоматического обучения // Итоги науки и техники. Сер. Информатика. Т. 15 (Интеллектуальные информационные системы).— М.: ВИНИТИ, 1991.— С. 17—54.
5. Кузнецов С. О., Финн В. К. Распространение процедур экспертных систем типа ДСМ на графы // Изв. АН СССР Сер. Техническая кибернетика.— 1988.— № 5.— С. 4—11.
6. Забежайло М. И. О некоторых переборных задачах, возникающих при автоматическом порождении гипотез ДСМ-методом // НТИ. Сер. 2.— 1988.— № 1.— С. 28—31.
7. Кузнецов С. О. Интерпретация на графах и сложностные характеристики задач поиска закономерностей определенного вида // НТИ. Сер. 2.— 1989.— № 1.— С. 23—28.
8. Гэри М., Джонсон Д. Вычислительные машины и труднорешаемые задачи.— М.: Мир, 1982.— 416 с.
9. Wille R. Restructuring lattice theory: an approach based on hierarchies of concepts // Ordered sets / Ed. I. Rival, Riedel, Dordrecht—Boston, 1982,— P. 445—470.
10. Левит В. Е. Алгоритм поиска подматрицы максимального периметра, состоящей из единиц на 0—1 матрице // Система передачи и обработки информации: Сб. тр. / ИППИ АН СССР.— М., 1988.— С. 42—45.
11. Valiant L. G. The Complexity of Computing the Permanent // Theoretical Computer Science.— 1979.— № 8.— P. 189—201.
12. Valiant L. G. The Complexity of Enumeration and Reliability Problems // SIAM J. Comput.— 1979.— Vol. 8., № 1.— P. 410—421.
13. Papadimitriou C. H., Yannakakis M. The complexity of facets (and some facets of complexity) // J. Comput. Syst. Sci.— 1984.— Vol. 28.— P. 244—259.
14. Забежайло М. И., Ивашко В. Г., Кузнецов С. О., Михеенкова М. А., Хазановский К. П., Аншаков О. М. Алгоритмические и программные средства ДСМ-метода автоматического порождения гипотез // НТИ. Сер. 2.— 1987.— № 10.— С. 1—14.

Материал поступил в редакцию 16.07.91.