УДК 007:510.64:159.955

С. М. Гусакова, С. О. Кузнецов

СХОДСТВО В ОБОБЩЕННОМ ДСМ-МЕТОДЕ И АЛГОРИТМЫ ЕГО ПОРОЖДЕНИЯ

Рассматривается порождение гипотез в «обобщенном» ДСМ-методе, заключающемся в порождении «условных» гипотез о причинно-следственных связях по положительным и отрицательным примерам. Условность гипотез состоит в том, что причины, о которых выдвигаются гипотезы, «срабатывают» (вызывают проявление свойства) в отсутствии некоторых подструктур, называемых «тормозами» этих причин. В отличие от «обычных» отрицательных причин, выражающих сходства отрицательных примеров, тормоза обладают некоторой «причинно-следственной структурой», неся в себе свою противоположность — условные положительные причины. В работе предложено алгебраическое описание сходства в обобщенном ДСМ-методе, а также алгоритмы порождения обобщенных гипотез (вместе с анализом вычислительной сложности).

1. ОБОБЩЕННЫЙ МЕТОД И СХОДСТВО

Основная идея ДСМ-метода [1, 2] — установление причинно-следственных отношений между подструктурами структурированных объектов и подмножеством свойств этих объектов на базе определения существенного сходства объектов — получает естественное расширение в обобщенном ДСМ-методе, принцип которого состоит в учете контекста (в структуре объекта) при установлении причинно-следственного отношения.

Такой учет контекста вызван реалиями решаемых с помощью ДСМ-метода задач из области химии, фармакологии, социологии.

В работе [3] было предложено формальное определение обобщенного ДСМ-предиката. Неформальный смысл его сводится к следующему: при наличии описания некоторых положительных и отрицательных примеров для некоторого класса объектов (характеризуемого определенным свойством) порождаются «условные» гипотезы. Условность их состоит в том, что причины «срабатывают» (вызывают проявление свойства) в отсутствие некоторых подструктур, называемых «тормозами» этих причин. В отличие от «обычных» (см. [1, 2]) отрицательных причин, выражающих сходства отрицательных примеров, тормоза обладают некоторой «каузальной (причинно-следственной) структурой», неся в себе свою противоположность — условные положительные причины. Таким образом бинарное отношение $V \Rightarrow_2 W$, читаемое как «подобъект V есть причина наличия множества свойств W» простого ДСМ-метода заменяется на тернарное отношение $T(V, x, W)$, читаемое как «V есть причина наличия множества свойств W при отсутствии «тормозов» x» в обобщенном ДСМ-методе.

Ограничения объема статьи и нежелание повторять неоднократно изложенный материал вынуждает авторов отослать читателя, незнакомого с ДСМ-методом и имеющего желание познакомиться с ним, к работам [1—4], содержащим все необходимые сведения.

Однако необходимый для прочтения настоящей работы минимум сведений, понятий и обозначений приводится ниже.

ДСМ-метод автоматического порождения гипотез используется для анализа и обработки данных, представленных в базе данных с неполной информацией. Класс задач, решаемый ДСМ-методом характеризуется следующими условиями:

данные предметной области хорошо структурированы;

имеется множество объектов и множество свойств, на каждом из которых заданы определенные операции;

на множестве объектов определено отношение «быть подобъектом объекта»;

на произведении множеств объектов и свойств заданы частично определенные отношения «объект обладает множеством свойств» (обозначается через $\overset{*}{\Rightarrow}_1$) и «подобъект является причиной множеств свойств» (обозначается через $\overset{*}{\Rightarrow}_2$);

имеется множество положительных и отрицательных примеров первого отношения, т. е. множество пар вида (x, y), где x — объект, y — свойство, удовлетворяющих и неудовлетворяющих отношению.

Примеры получены эмпирически.

Наличие свойств у объектов или их отсутствие определяется «положительными» (+) и «отрицательными» (—) причинами, т. е. наличие некоторого явления обусловлено некоторым множеством (+)-причин, а отсутствие явления обусловлено или наличием (—)-причин или отсутствием соответствующих (+)-причин.

Как правило, задачи указанного класса возникают в плохо формализуемых областях знаний.

Исходными данными для ДСМ-системы являются матрицы частично определенных отношений — $\overset{*}{\Rightarrow}_1$ и $\overset{*}{\Rightarrow}_2$. Строки матриц соответствуют объектам в первом случае и подобъектам — во втором, столбцы — элементарным свойствам. На пересечении i-й строки и j-го столбца стоит $+1$, если $x_i \overset{*}{\underset{e}{\Rightarrow}} y_j$ выполнено, -1, если не выполнено и т, если не известно, имеет ли место $x_i \overset{*}{\underset{e}{\Rightarrow}} y_j$ $(l = 1, 2)$. Подстрока матрицы, соответствующая всем $+1$ (-1), называется положительным (отрицательным) примером.

С помощью логико-комбинаторных алгоритмов простой ДСМ-метод порождает гипотезы вида:

$$J_{\langle v,n\rangle}(C \Rightarrow_1 A),\ J_{\langle v,n\rangle}(\overline{C} \Rightarrow_2 A),$$

$$J_{(\tau,n)}(C \Rightarrow_1 A) \text{ и } J_{(\tau,n)}(\overline{C} \Rightarrow_2 A),$$

где J — одноместный оператор, $\nu = \{1, -1, 0\}$ и τ — типы истинностных значений, обозначающих фактическую истину (+1), фактическую ложь (−1), фактическое противоречие (0), недоопределенность (τ).

Обобщенный метод, отражающий более глубокие представления о характере причинно-следственных зависимостей, порождает с помощью существенно более сложных алгоритмов гипотезы вида:

$$J_{\langle \nu, n \rangle} T(\overline{C}, \overline{\varkappa}, A) \text{ и } J_{(\tau, n)} T(\overline{C}, \overline{\varkappa}, A),$$

где C и A есть соответственно объект и множество свойств, $\overline{C}$ — подобъект, $\overline{\varkappa}$ — множество тормозов.

Логико-комбинаторные алгоритмы ДСМ-метода, с помощью которых порождаются гипотезы, реализуют в различных усложненных вариантах простую идею Д. С. Миля (отсюда и название ДСМ-метод) — причиной сходства свойств объектов является сходство их структур.

Поэтому понятие сходства является одним из центральных понятий ДСМ-теории, а определение существенного сходства в конкретных прикладных задачах и для конкретных структур данных — одной из центральных проблем создания прикладной ДСМ-системы.

В ДСМ-системах используется операция сходства для выделения причин подобъектов и локальное (отношение) и глобальное (семейство множеств) сходства для конструирования алгоритмов.

Операция сходства понимается в ДСМ-системах как идемпотентная, коммутативная и ассоциативная операция на парах объектов. Эти свойства позволяют однозначно выражать сходство множества объектов через попарные сходства независимо от порядка расположения объектов в базе данных.

Операция сходства обозначается через Π. Строгое ее определение можно найти, например, в [5].

Определение 1. Объекты $X_{i_1}, \ldots, X_{i_n}$ локально сходны, если $\prod_{j=1}^{n} X_{i_j} \neq \varnothing$.

Локальное сходство есть n-арное отношение толерантности (см. [6]).

Определение 2. Объекты $X_{i_1}, \ldots, X_{i_k}$ глобально сходны если

$$\left(\prod_{j=1}^{k} X_{i_j} = h\right) \& \forall X_{i_m} \left(\left(\prod_{j=1}^{k} X_{i_j}\right) \Pi X_{i_m} = h\right) \& \& \left(\&_{r=1}^{k} \neg (m = r)\right).$$

Глобальное сходство объединяет все объекты, содержащие некоторый подобъект (результат операции сходства) и потому не является отношением.

Структура глобального сходства описывается семейством множеств (M, G), где M — рассматриваемое множество объектов, $G = \{g_i\}$, g_i — подмножество глобально сходных объектов. Для ознакомления с теорией глобального и локального сходства рекомендуем читателю [7, 8].

Для дальнейшего изложения нам будет удобно пользоваться операциональной трактовкой локального и глобального сходства (см. [5]), в соответствии с которой:

Определение 1′. h есть локальное сходство $X_1, \ldots, X_n$ из S (S — множество объектов), если $\prod_{i=1}^{n} X_i = h$.

Определение 2′. $\langle h, \{X_1, \ldots, X_k\} \rangle$ есть глобальное сходство, если $X_1, \ldots, X_k \in S$, $\prod_{i=1}^{k} X_i = h$ и $\forall Y \in S \setminus \{X_1, \ldots, X_k\}$, имеет место $Y \Pi h \neq h$.

Логико-комбинаторные алгоритмы, реализующие простой ДСМ-метод, основаны на нахождении глобального сходства объектов, обладающих некоторым свойством.

Алгоритм для реализации обобщенного ДСМ-метода ищет тройки, находящиеся в отношении $T(V, \varkappa, W)$. Множество троек этого отношения есть область истинности обобщенного ДСМ-предиката сходства, обозначаемого через $M^{+}_{ag,n}$ (мы рассматриваем случай положительного предиката, помня, что все рассуждения могут быть без труда перенесены на случай отрицательного предиката $M^{-}_{ag,n}$).

В силу объемности формулировки указанного предиката приведем его здесь в упрощенном виде, отсылая читателя за точной формулировкой к статье [9], которая посвящена изучению этого предиката.

Представим предикат $M^{+}_{ag,n}(V, \varkappa, W)$ в виде конъюнкции нескольких частей, которым для удобства дадим мнемонические обозначения

$$M^{+}_{ag,n}(V, \varkappa, W) = EX \& ED \& CE \& Bl.$$

EX описывает множество примеров вида $I_{(1,\, n)}$ $(X_i \underset{e}{\Rightarrow} Y_i)$, которые являются основанием правдоподобного вывода;

ED — эмпирическая зависимость между V, W и $\varkappa$, выражающая тот факт, что «V является причиной W при отсутствии тормозов $\varkappa = \{V_1, \ldots, V_r\}$.

Эта зависимость имеет вид:

$$\forall X \forall Y ((I_{(1,\, n)}(X \underset{e}{\Rightarrow} Y) \& \forall U (I_{(1,\, n)}(X \underset{e}{\Rightarrow} Y) \to U \subseteq Y) \& V \subset X) \& \neg (V_1 \subset X \vee \ldots \vee V_r \subset X)) \to W \subseteq Y \& W \neq \varnothing).$$

CE — условие исчерпываемости, обеспечивающее рассмотрение всех подходящих примеров из базы данных: $\bigvee_{i=1}^{k} (X = Z_i)$.

Bl — описывает множество примеров, порождающих тормоза и включает в себя две части Bl_1 и Bl_2, описывающие множества соответственно положительных и отрицательных примеров:

$$Bl_1: \&_{j=1}^{s} W_j ((\neg J_{(1,n)}(Z_j \Rightarrow_l W_j) \& W \subseteq W_j) \& V \subset Z_j,$$

$$Bl_2: (J_{(1,n)}(Z_j \Rightarrow_l W_j) \& \neg (W \subseteq W_j)) \& V \subset Z_j,$$

$$Bl: (Bl_1 \vee Bl_2) \& \left(V_q = \bigcap_{p=1}^{l} Z_{J_p}\right) \& V \subset V_q.$$

Остальные фрагменты предиката $M^{+}_{ag,n}$ характеризуют условия исчерпываемости и минимальности для тормозов.

Суть предиката $M^{+}_{ag,n}$ в том, что подобъект, найденный как операциональное сходство объектов $X_1, \ldots, X_k$, является причиной свойства W, если ни один из указанных объектов не содержит тормозов из $\varkappa = \{V_1, \ldots, V_r\}$, $V \subset V_i$, $i = 1, \ldots, r$. Тормоза, в свою очередь, находятся из объектов $Z_1, \ldots, Z_s$ как для отрицательных примеров, так и для положительных примеров, содержащих V, но не обладающих свойством W.

Причем, в случае, если тормозов для V из имеющихся в базе данных примеров извлечь нельзя, то V не является причиной свойства W в обобщенном смысле.

В алгоритмическом плане основные трудности порождает тот факт, что при последовательном исполнении действий по проверке условий предиката, найденный как кандидат в причины свойства W подобъект V, может быть забракован. Следующий шаг в этом случае будет проведен над множеством объектов $X_1, \ldots, X_n$ без тех объектов, которые содержат тормоза, и может привести к появлению нового кандидата в причины, а это, в свою очередь, повлечет поиск новых тормозов. Наличие в предикате $M^+_{ag,n}$ условия, по которому тормоза ищутся и среди положительных примеров, не обладающих свойством W, еще более усложняют алгоритм.

Исследование сходств, определяемых на множестве объектов предикатом обобщенного сходства, дает возможность выделить различные ситуации, возникающие в зависимости от структуры базы данных, и среди этих ситуаций найти те, для которых алгоритм, реализующий обобщенный ДСМ-метод, проще, чем в общем случае.

Поэтому мы и приступим к этому исследованию.

Анализируя предикат $M^+_{ag,n}$ обобщенного сходства, с точки зрения характера сходств, порождаемых им на множестве объектов (так называемые процедурные сходства), мы должны отметить следующие существенные моменты:

1) процедурные сходства на множестве объектов — это глобальные сходства в силу условия исчерпываемости;

2) процедурное сходство, определяющее причину, есть некоторая композиция простых сходств. Это следует из тернарности предиката и соотношений между υ и $\varkappa$.

Введем следующие обозначения:

$W=\{w_1, \ldots, w_a\}$ — множество элементарных свойств.

$\tilde{W}=2^W=\{\tilde{w}_1, \ldots, \tilde{w}_{\tilde{a}}\}$ — множество всех подмножеств множества W.

$\overset{\sigma}{G}{}^l=\{\overset{\sigma}{g}_k\}$, $k=1, \ldots, \bar{k}$; $l=1, \ldots, \tilde{a}$; $\sigma=\{+, -, \pm\}$;

$\overset{\sigma}{g}_k=\{\overset{\sigma}{h}_k{}^l, \overset{\sigma}{S}_k{}^l\}$ — глобальное сходство, причем $\overset{+}{S}_k{}^l$ состоит из объектов, взятых из подмножества положительных примеров вида $J_{\langle 1,0\rangle}(X \Rightarrow_1 \tilde{W}_l)$, $\bar{S}_k{}^l$ — из объектов подмножества отрицательных примеров также для свойства $\tilde{w}_l$, а $\overset{\pm}{S}_k{}^l$ — из объектов подмножества как положительных, так и отрицательных примеров для этого же свойства.

Таким образом $\{\overset{+}{h}_l\}$ — множество всех максимальных положительных «пересечений» (точнее результатов операции сходства), $\{\bar{h}^i\}$ — всех отрицательных, т. е.

$$\overset{+}{h}_k{}^l = \prod_{x_n \in \overset{+}{S}_k{}^l} X_n, \quad \overset{+}{S}_k{}^l = \{X_n \mid (J_{\langle 1,0\rangle}(X_n \Rightarrow_1 Y) \& (\tilde{W}_l \subseteq Y)\};$$

$$D^{l,j}=\{d_m^{l,j}\}, \quad m=1, \ldots, \bar{m}; \quad l, j=1, \ldots, \tilde{a};$$

$$l \neq j; \quad \tilde{W}_l \subseteq \tilde{W}_j;$$

$$d_m^{l,j}=\{\overset{+}{h}{}_m^{l,j}, \overset{+}{S}{}_m^{l,j}\}; \quad \overset{+}{h}{}_m^{l,j} = \prod_{x_n \in \overset{+}{S}{}_m^{l,j}} X_n;$$

$$\overset{+}{S}{}_m^{l,j}=\{X_n \mid ((J_{\langle 1,0\rangle}(X_n \Rightarrow_1 Y_1) \& (\tilde{W}_l \subseteq Y_1)) \vee$$

$$\vee ((J_{\langle 1,0\rangle}(X_n \Rightarrow_1 Y_2) \& (\tilde{W}_j \subseteq Y_2)\},$$

т. е. $\overset{+}{S}{}_m^{l,j}$ состоит из объектов, взятых из положительных примеров для свойств $\tilde{w}_l$ и $\tilde{w}_j$, причем $\tilde{w}_l \not\subseteq \tilde{w}_j$.

Необходимость рассмотрения сходства $D^{l,j}$ определяется частью Bl_2 предиката для поиска тормозов.

Определим два тернарных предиката:

$$\zeta_1(\overset{+}{g}_k{}^l, \bar{g}_n{}^l, \overset{\pm}{g}_m{}^l) =$$

$$= 1 \Leftrightarrow (\overset{+}{h}_k{}^l \subset \bar{h}_n{}^l) \& (\bar{h}_n{}^l = \overset{\pm}{h}_m{}^l) \& \forall l' \neg (\bar{h}_l{}^l \subset \bar{h}_n{}^l);$$

$$\zeta_2(\overset{+}{g}_k{}^l, \overset{+}{g}_{n'}{}^j, d_m^{l,j}) =$$

$$= 1 \Leftrightarrow (\overset{+}{h}_k{}^l \subset \overset{+}{h}_{n'}{}^j) \& (\overset{+}{h}_{n'}{}^j = h_m^{l,j}) \& \forall l' \neg (\overset{+}{h}_l{}^j, \subset \overset{+}{h}_{n'}{}^j).$$

Ясно, что ζ_1 и ζ_2 равны 0, если хотя бы один из членов конъюнкции равен 0 (1, 0 — истинностные значения «истина» и «ложь»).

Глобальное сходство $\overset{+}{P}{}^l=\{\overset{+}{p}_k{}^l\}$, где $\overset{+}{p}_k{}^l=(\overset{+}{\upsilon}_k{}^l, \overset{+}{\Sigma}_k{}^l)$ определим следующим образом:

$$\overset{+}{\upsilon}_k{}^l = \overset{+}{h}_k{}^l,$$

$$\overset{+}{\Sigma}_k{}^l = \overset{+}{S}_k{}^l \setminus \left[\left(\bigcup_{m,n} (\overset{\pm}{S}_m{}^l \setminus \bar{S}_n{}^l) \mid \zeta_1(\overset{+}{g}_k{}^l, \bar{g}_n{}^l, \overset{\pm}{g}_m{}^l)=1\right) \cup \right.$$

$$\left. \left(\bigcup_{m',n',j} (S_m^{l,j} \setminus \overset{+}{S}_{n'}{}^j) \mid \zeta_2(\overset{+}{g}_k{}^l, \overset{+}{g}_n{}^j, d_m^{l,j})=1\right)\right];$$

$$\overset{+}{p}_k{}^l = \varnothing, \text{ если } (\forall n \forall m \zeta_1(\overset{+}{g}_k{}^l, \bar{g}_n{}^l, \overset{\pm}{g}_m{}^l)=0) \&$$

$$\& (\forall n' \forall m' \zeta_2(\overset{+}{g}_k{}^l, \bar{g}_n{}^j, d_m^{l,j})=0) \vee \exists r (\overset{+}{\Sigma}_k{}^l = \overset{+}{S}_r{}^l).$$

Утверждение 1. Для любого $\overset{+}{p}_k{}^l=(\overset{+}{\upsilon}_k{}^l, \overset{+}{\Sigma}_k{}^l)$ имеет место $T(\overset{+}{\upsilon}_k{}^l, \varkappa, \tilde{w}_l)=1$, где

$$\varkappa = \{(\bar{h}_n{}^l \mid \exists m \zeta_1(\overset{+}{g}_k{}^l, \bar{g}_n{}^l, \overset{\pm}{g}_m{}^l)=1\} \cup$$

$$\cup \{\bar{h}_{n'}{}^j \mid \exists m' \zeta_2(\overset{+}{g}_k{}^l, \overset{+}{g}_{n'}{}^j, d_m^{l,j})=1)\}\}.$$

Доказательство.

1. Из определения сходства $\overset{+}{G}{}^l$ следует, что для каждых l, k множество $\overset{+}{S}_k{}^l$ представляет из себя как раз множество примеров, описываемых в части Ex предиката $M^+_{ag,n}$. Условие исчерпываемости (CE) обеспечивается глобальностью сходства $\overset{+}{g}_k{}^l$. Следовательно, $\overset{+}{\upsilon}_k{}^l=\overset{+}{h}_k{}^l$ удовлетворяет условиям $M^+_{ag,n}$, так как является максимальным операциональным сходством положительных примеров.

2. Из определения $\bar{g}_n{}^l$ следует, что $\bar{h}_n{}^l$ является максимальным операциональным сходством отрицательных примеров.

Условие минимальности $\bar{h}_n{}^l$ по вложению, накладываемое на тормоза, обеспечивается выполнением предиката ζ_1.

3. Из определения $d_m^{l,j}$ следует, что $\overset{+}{h}_{n'}{}^j$ является максимальным операциональным сходством положительных примеров, не обладающих свойством $\tilde{w}_l$.

4. Из определения $\overset{+}{\Sigma}{}^l$ следует, что

$$\forall n \forall X \in \overset{+}{\Sigma}_k{}^l \neg (\bar{h}_n{}^l \subset X) \& \forall n' \forall X' \in \overset{+}{\Sigma}_k{}^l \neg (\overset{+}{h}_{n'}{}^j \subset X').$$

Выполнение условий (1—4) обеспечивает $T(\overset{+}{v}_k^l, \varkappa, \tilde{w}_l)=1$ при $\varkappa$, описанном в формулировке утверждения.

Покажем теперь, что, если $T(v', \varkappa', \tilde{w}_l)=1$, то $\exists k$ $v'=\overset{+}{h}_k^l$,

$$\exists n \forall m \exists n' \exists m' \varkappa' = \{\{\overline{h}_n^l \mid \zeta_1(\overset{+}{g}_k^l, \overline{g}_n^l, \overset{\pm}{g}_m^l)=1\} \cup$$

$$\cup\{\overline{h}_{n'}^l \mid \zeta_2(\overset{+}{g}_k^l, \overset{+}{g}_{n'}^j, d_{m'}^{l,j})=1\}\}).$$

По определению отношения T, v есть максимальное операциональное сходство некоторого подмножества положительных примеров, но все такие сходства входят в глобальное сходство $\overset{+}{G}^l$, следовательно, $\exists k$ $v=\overset{+}{h}_k^l$.

Рассуждая аналогично, получаем, что

$$\forall v_q \in \varkappa' (\exists n (v_q=\overline{h}_n^l)) \vee (\exists j \exists n' (v_q=\overset{+}{h}_{n'}^j)).$$

Выполнение предикатов ζ_1 и ζ_2 следует из определения сходств $\overset{\delta}{G}$ и $\overset{+}{P}$.

Собственно говоря, утверждение 1 непосредственно следует из определения отношения T, предиката $M_{ag,n}^+$ и сходств $\overset{\delta}{G}$ и $\overset{+}{P}$, что и видно непосредственно из доказательства.

Таким образом, глобальное сходство $\overset{+}{P}^l$ порождается всеми теми и только теми объектами, операциональное сходство которых выражается причиной $\overset{+}{v}^l$ свойства $\tilde{w}^l$ при отсутствии тормозов $\varkappa$.

Глобальное сходство $\overset{+}{P}^l$, как видно из его определения, существенно зависит от свойства $\tilde{w}^l$, поэтому структура множества свойств в базе данных определяет во многом сложность алгоритмической процедуры нахождения $\overset{+}{P}^l$. Так очевидно, что для случая: $W=\{w\}$, $\tilde{W}=W$, процедура эта значительно упрощается. В самом деле, прежде всего упрощается предикат сходства $M_{ag,n}^+$, так как из него убирается добавка Bl_2.

Сходство $\overset{+}{P}=\{\overset{+}{p}_j\}$ принимает для этого случая вид:

$$\overset{+}{P}_j=\{\overset{+}{v}=\overset{+}{h}_j;\ \overset{+}{\Sigma}_j=\overset{+}{S}_j \setminus$$

$$\setminus\left(\bigcup_{m,n}(\overset{\pm}{S}_m^i \setminus \overline{S}_n^i) \mid \zeta_1(\overset{+}{g}_j^l, \overline{g}_n^l, \overset{\pm}{g}_m^l)=1\right\}.$$

$\overset{+}{P}$ образуется с помощью сходств $\overset{+}{G}=\{\overset{+}{g}_l\}$, $\overline{G}=\{\overline{g}_n\}$ и $\overset{\pm}{G}=\{\overset{\pm}{g}_m\}$.

Если сходство P найдено, то найдены все тройки $\{v, \varkappa, w\}$, удовлетворяющие предикату T.

2. АЛГОРИТМЫ ПОСТРОЕНИЯ ОБОБЩЕННЫХ ГИПОТЕЗ

В конструктивности определения сходства $\overset{+}{P}$ заложен, по существу, алгоритм 1 его нахождения, заключающийся в последовательности следующих действий:

1. Нахождение сходств $\overset{+}{G}$ и $\overline{G}$, осуществляемое с помощью, так называемого алгоритма максимальных пересечений, реализованного в версии ДСМ-системы для простого метода.

2. Проверка предиката ζ_1.

3. Конструирование сходства $\overset{+}{P}$.

4. Проверка условия $\forall r(\overset{+}{\Sigma}_k^l \neq \overset{\pm}{S}_r^l)$.

Продемонстрируем работу алгоритма для случая одного свойства на примере:

Множество объектов: $\{x_1, \ldots, x_{11}\}$: $\Omega^+=\{x_1, x_2, x_3, x_4, x_5\}$, $\Omega^-=\{x_6, x_7, \ldots, x_{11}\}$; $x_1=vaa'$, $x_2=vbdd'$, $x_3=vbd'd''m$, $x_4=vd'cm$, $x_5=vd''c'm$, $x_6=vae$, $x_7=vae'$, $x_8=vbdf$, $x_9=vbdf'$, $x_{10}=vd'gg'$, $x_{11}=vd'gg''$.

Находим сходство $\overset{+}{G}$ из положительных примеров:

$\overset{+}{g}_1=\{v;\ x_1, x_2, x_3, x_4, x_5\}$;

$\overset{+}{g}_2=\{vbd';\ x_2, x_3\}$;

$\overset{+}{g}_3=\{vd';\ x_2, x_3, x_4\}$;

$\overset{+}{g}_4=\{vm;\ x_3, x_4, x_5\}$;

$\overset{+}{g}_5=\{vd''m;\ x_3, x_5\}$.

Сходство $\overline{G}$ находим из отрицательных примеров:

$\overline{g}_1=\{v;\ x_6, x_7, x_8, x_9, x_{10}, x_{11}\}$;

$\overline{g}_2=\{va;\ x_6, x_7\}$;

$\overline{g}_3=\{vbd;\ x_8, x_9\}$;

$\overline{g}_4=\{vd'g;\ x_{10}, x_{11}\}$.

Сходство $\overset{\pm}{G}$ представлено следующим образом:

$\overset{\pm}{g}_1=\{v;\ x_1, \ldots, x_{11}\}$;

$\overset{\pm}{g}_2=\{va;\ x_1, x_6, x_7\}$;

$\overset{\pm}{g}_3=\{vb;\ x_2, x_3, x_8, x_9\}$;

$\overset{\pm}{g}_4=\{vbd';\ x_2, x_3\}$;

$\overset{\pm}{g}_5=\{vbd;\ x_2, x_8, x_9\}$;

$\overset{\pm}{g}_6=\{vd';\ x_2, x_3, x_4, x_{10}, x_{11}\}$;

$\overset{\pm}{g}_7=\{vd''m;\ x_3, x_5\}$;

$\overset{\pm}{g}_8=\{vd'g;\ x_{10}, x_{11}\}$.

Множество троек $\{\overset{+}{g}_i, \overline{g}_j, \overset{\pm}{g}_m\}$, для которых предикат $\overset{*}{\zeta}$ истинен, обозначим через Th_{ζ}.

Оно в данном примере следующее:

$Th_{\zeta}=\{(\overset{+}{g}_1, \overline{g}_2, \overset{\pm}{g}_3), (\overset{+}{g}_1, \overline{g}_3, \overset{\pm}{g}_5), (\overset{+}{g}_1, \overline{g}_4, \overset{\pm}{g}_8),$

$(\overset{+}{g}_3, \overline{g}_4, \overset{\pm}{g}_8)\}$.

Теперь можно построить сходства $\overset{+}{P}$:

$\overset{+}{P}_1=\{v;\ \overset{+}{\Sigma}_1(\{x_1, x_2, x_3, x_4, x_5\} \setminus [\{x_1, x_6, x_7 \setminus x_6, x_7\} \cup$

$\cup\{x_2, x_3, x_3 \setminus x_8, x_9\} \cup \{x_{10}, x_{11} \setminus x_{10}, x_{11}\}]\}=$

$=\{x_3, x_4, x_5\}\}$.

Но так как $\overset{+}{\Sigma}_1=\overset{+}{S}_4$, то $\overset{+}{P}_1=\varnothing$.

$\overset{+}{P}_2=\varnothing$, так как $\forall j \neg(\overset{+}{h}_2 \subset \overline{h}_j)$ $j=1, \ldots, 4$.

$\overset{+}{P}_3=\{vd'; \overset{+}{\Sigma}_3\{\{x_2, x_3, x_4\} \setminus \{x_{10}, x_{11} \setminus x_{10}, x_{11}\}\} = \{x_2, x_3, x_4\}$.

$\forall i \overset{+}{\Sigma}_3 \neq \overset{+}{S}_i$ $(i=1, \ldots, 5)$, следовательно, $\overset{+}{P}_3 \neq 0$.

$\overset{+}{P}_4=\varnothing$ и $\overset{+}{P}_5=\varnothing$ по той же причине, что и $\overset{+}{P}_2=\varnothing$.

В результате работы алгоритма мы найдем одну тройку: $\{vd', \varkappa=vd'g, w\}$, удовлетворяющую предикату $M^{+}_{ag,n}$ и, следовательно, принадлежащую отношению T.

Увеличение множества элементарных свойств хотя бы до двух элементов влечет за собой существенное увеличение разнообразия ситуаций, и следовательно, усложнение алгоритма.

Подсчитаем оценку сверху количества машинных операций, необходимых для порождения обобщенных гипотез с помощью указанной процедуры, для произвольных множества положительных примеров Ω^+, множества отрицательных примеров Ω^- и множества элементарных структур U (подмножествами которого являются положительные и отрицательные примеры). В работе [5] предложен алгоритм ЗО поиска пересечений (сходств), временная трудоемкость которого линейна по отношению к числу порождаемых пересечений. Предположив, что сходства (пересечения) ищутся с помощью алгоритма ЗО, получим, что для нахождения множества $\overset{+}{G}$ всех положительных сходств необходимо затратить

$$O(|\overset{+}{G}|\cdot|U|\cdot|\Omega^+|)$$

машинных операций; для нахождения множества $\overline{G}$ всех отрицательных сходств необходимо затратить

$$O(|\overline{G}|\cdot|U|\cdot|\Omega^-|)$$

машинных операций; для нахождения множества сходств $\overset{\pm}{G}$ всех примеров (положительных и отрицательных) необходимо затратить

$$O(|\overset{\pm}{G}|\cdot|U|\cdot(|\Omega^+|+|\Omega^-|))=$$

$$=O(|\overset{+}{G}|\cdot|\overline{G}|\cdot|U|\cdot(|\Omega^+|+|\Omega^-|))$$

машинных операций. Для порождения каждого множества $\overset{+}{\Sigma}_j$ в худшем случае необходимо

$$O(|\overset{\pm}{G}|\cdot|\overline{G}|\cdot \mathrm{pol}_1(|U|, |\Omega^+|, |\Omega^-|))$$

машинных операций. Всего по всем j (т. е. для нахождения всех положительных обобщенных гипотез) необходимо

$$O(|\overset{+}{G}_j|\cdot|\overset{\pm}{G}|\cdot \mathrm{pol}_1(|U|, |\Omega^+|, |\Omega^-|))=$$

$$=(|\overset{+}{G}|^2\cdot|\overline{G}|^2 \mathrm{pol}_1(|U|, |\Omega^+|, |\Omega^-|))$$

машинных операций, где $\mathrm{pol}_1(\cdot)$ — некоторый полином небольшой степени, зависящий от $|U|$, $|\Omega^+|$, $|\Omega^-|$ и определяемый вычислительной моделью.

Можно предложить более эффективные алгоритмы построения обобщенных гипотез. Быстродействие этих алгоритмов существенным образом зависит от типа операции сходства (даже если считать время выполнения таких операций одинаковым). Вначале дадим описание такого алгоритма для построения обобщенных положительных гипотез в случае представления данных множествами и наличия лишь одного свойства.

Алгоритм 2

1. Строятся все (+)-пересечения с помощью алгоритма ЗО [5].

2. Для каждого (+)-пересечения V:

2.1. Находятся отрицательные примеры его содержащие,

2.2. Ищутся все минимальные по вложению пересечения отрицательных примеров, содержащие V (т. е. тормоза): Пусть N есть объединение всех отрицательных примеров, содержащих V. Тогда произвольное минимальное пересечение отрицательных примеров, содержащее V, имеет вид $X^i=\bigcap_j X_{ij}$, где $X_{ij} \in X_i=\{X \mid X \in \Omega^-, (V \cup \{q_i\}) \subseteq X\}$, где $q_i \in N$. Естественно, что $\bigcup_j X_{ij}$ определено, если $j \geqslant 2$.

Чтобы исключить порождение одинаковых пересечений и, тем самым, ускорить процесс их порождения, при построении очередного множества $\bigcup_j X_{ij}$, соответствующую «добавку» к v можно выбирать не из N, а из N за вычетом всех предыдущих «добавок» и порожденных пересечений, или формально, $N_1=N$, $N_{k+1}=N_k \setminus (\{q_{k+1}\} \cup \bigcap_j X_{k_j})$.

2.3. Из множества (+)-примеров, содержащих V, удаляются все такие, которые содержат какие-либо тормоза, оставшиеся (+)-примеры вновь пересекаются. Если результат совпадает с V, то V вместе с найденными минимальными пересечениями составляет обобщенную гипотезу, если нет, то (+)-обобщенной гипотезы относительно пересечения V не существует, переходим к следующему (+)-пересечению.

Временная сложность построения одной обобщенной положительной гипотезы для фиксированного (+)-пересечения V на шагах 2.1—2.3 такого алгоритма есть $O(N\cdot|\Omega^-|)$.

Таким образом, сложность построения всех положительных обобщенных гипотез есть

$$O(|\overset{+}{G}|\cdot N\cdot|\Omega^-|).$$

Рассмотрим действие алгоритма для множеств положительных и отрицательных примеров, приведенных выше. Множество N в этом случае есть $N=\{a, b, d, d', e, f, f', g, g', g'', v\}$. Пусть на втором этапе действия алгоритма нами выбрано (+)-пересечение v. Тогда минимальные пересечения (—)-примеров, содержащие v, находятся как

$X^a=\bigcap_j X_j{}^a$, где $X_j{}^a \in X_a=\{X \in \Omega^- \mid \{v, a\} \subset X\}=\{v, a\}$,

$N_2=N \setminus \{a\}=\{b, d, d', e, e'f, f', g, g', g''\}$;

$X^b=\bigcap_j X_j{}^b$, где $X_j{}^b \in X_b=\{X \in \Omega^- \mid \{v, b\} \subset X\}=\{v, b, d\}$,

$N_3=N_2 \setminus \{b, d\}=\{d', e, e'f, f', g, g', g''\}$.

Пересечение, соответствующее $\{v, d\}$, не порождается, так как на этом шаге $d \notin N$.

$X^{d'}=\bigcap_j X_j^{d'}$, где $X_j^{d'} \in X_{d'}=\{X \in \Omega^- \mid \{v, d'\} \subset X\}=$

$=\{v, d', g\}$, $N_4=N_3 \setminus \{d', g\}=\{e, e'f, f', g', g''\}$;

$X^e=\bigcap_j X_j{}^e$, где $X_j{}^e \in X_e=\{X \in \Omega^- \mid \{v, e\} \subset X\}$,

X^e не определено, $N_5=N_4 \setminus \{e\}=\{e'f, f', g', g''\}$;

$X^{e'}=\bigcap_j X_j^{e'}$, где $X_j^{e'} \in X_{e'}=\{X \in \Omega^- \mid \{v, e'\} \subset X\}$,

$X^{e'}$ не определено, $N_6=N_5 \setminus \{e'\}=\{f, f', g', g''\}$

$X^f = \bigcap_j X_j^f$, где $X_j^f \in X_f = \{X \in \Omega^- \mid \{v, f\} \subset X\}$,

X^f не определено, $N_7 = N_6 \setminus \{f\} = \{f', g', g''\}$;

$X^{f'} = \bigcap_j X_j^{f'}$, где $X_j^{f'} \in X_{f'} = \{X \in \Omega^- \mid \{v, f'\} \subset X\}$,

$X^{f'}$ не определено, $N_8 = N_7 \setminus \{f'\} = \{g', g''\}$.

Минимальное пересечение, соответствующее $\{v, g\}$, не порождается, так как $g \notin N_8$,

$X^{g'} = \bigcap_j X_j^{g'}$, где $X_j^{g'} \in X_{g'} = \{X \in \Omega^- \mid \{v, g'\} \subset X\}$,

$X^{g'}$ не определено, $N_9 = N_8 \setminus \{g'\} = \{g''\}$;

$X^{g''} = \bigcap_j X_j^{g''}$, где $X_j^{g''} \in X_{g''} = \{X \in \Omega^- \mid \{v, g''\} \subset X_j\}$,

$X^{g''}$ не определено, $N_{10} = N_9 \setminus \{g''\} = \varnothing$.

Осталось произвести действия, относящиеся к шагу 2.3. Из множества (+)-примеров, содержащих пересечение v, т. е. из $\{x_1, x_2, x_3, x_4, x_5\}$, удалим все примеры, содержащие минимальные пересечения (—)-примеров, содержащие v, т. е. va, vbd, $vd'gg'$. Таковыми (+)-примерами будут $x_1 = vaa'$, $x_2 = vbdd'$. Пересечем оставшиеся (+)-примеры, т. е. x_3, x_4, x_5, и получим $vm \neq v$. Значит, обобщенной (+)-гипотезы, соответствующей сходству (+)-примеров v, не будет.

Проделаем вычисления шагов 2.1—2.3 для (+)-пересечения vd'. (—)-примерами, его содержащими, будут $x_{10} = vd'gg'$ и $x_{11} = vd'gg''$. Значит единственное минимальное пересечение (—)-примеров, содержащее vd', будет $vd'g'$. Так как ни один (+)-пример его не содержит, то на шаге 2.3 ни один (+)-пример не будет отброшен, и тройка $\{vd', k = vd'g, w\}$ будет признана обобщенной (+)-гипотезой.

Как уже говорилось выше, алгоритм 2 предназначен для данных, представленных множествами. Последнее обстоятельство позволяет весьма эффективно осуществлять шаг 2.2. В случае, когда данные представлены объектами из произвольной полурешетки с операцией сходства П, дело обстоит хуже: мы не можем образовать минимальное по вложению пересечение (—)-примеров, содержащее данное пересечение (+)-примеров v путем «наращивания» v атомами решетки (—)-пересечений (о решетке пересечений см., например, [5]) — мы можем просто не знать этих атомов. Таким образом, для произвольной операции сходства, вместо шага 2.2 алгоритма 2 мы должны искать минимальные пересечения «сверху — вниз»: беря пересечения нарастающего количества (—)-примеров, содержащих v. Как только на очередном шаге пересечение (—)-примеров становится равным v, мы отходим к предыдущему пересечению (—)-примеров: оно и будет минимальным среди содержащих v. При этом верхняя оценка числа машинных операций, необходимых для порождения всех минимальных пересечений, содержащих v, будет $O(|\bar{G}| \mathrm{pol}_2(|U|, |\Omega^+|, |\Omega^-|))$, следовательно, общая сложность алгоритма будет $O(|\bar{G}| \cdot |\dot{G}| \mathrm{pol}_2(|U\Omega|, |\Omega^+|, |\Omega^-|))$, где $\mathrm{pol}_2(\cdot)$ — некоторый полином небольшой степени, зависящий от $|U|$, $|\Omega^+|$, $|\Omega^-|$ и определяемый вычислительной моделью.

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. Кузнецов С. О. ДСМ-метод как система автоматического обучения // Итоги науки и техники. Сер. Информатика. Т. 15.— М.: ВИНИТИ, 1991.— С. 17—54.
3. Финн В. К. Об обобщенном ДСМ-методе автоматического порождения гипотез // Семиотика и информатика.— 1989.— Вып. 29.— С. 93—123.
4. Гусакова С. М., Финн В. К. Сходство и правдоподобный вывод // Изв. АН СССР. Сер. Техн. кибернетика.— 1987.— Т. 5.— С. 42—63.
5. Кузнецов С. О. Быстрый алгоритм построения всех пересечений объектов из конечной полурешетки // НТИ. Сер. 2.— 1993.— № 1.— С. 17—20.
6. Гусакова С. М. Канонические представления сходств // НТИ. Сер. 2.— 1987.— № 9.— С. 19—21.
7. Гусакова С. М., Финн В. К. О новых средствах формализации локального и глобального сходств // НТИ. Сер. 2.— 1987.— № 10.— С. 14—22.
8. Гусакова С. М. Формализация понятия сходства и его применение в интеллектуальных информационных системах: Автореф. дис... к. ф.-м. н.— М.: ВИНИТИ, 1988.— 21 с.
9. Финн В. К., Михеенкова М. А. Некоторые проблемы обобщенного ДСМ-метода автоматического порождения гипотез // Семиотика и информатика. — 1983.— Вып. 33.— С. 136—163.

Материал поступил в редакцию 05.05.95.