Теоретические основы информатики

УДК 004.891

Об одном методе анализа больших массивов структур с частично детерминированными свойствами объектов

А. А. Липкин

Отделение интеллектуальных систем в гуманитарной сфере
Российский государственный гуманитарный университет
Миусская пл., д. 6, Москва, Россия, 125267

ДСМ-метод автоматического порождения гипотез — это один из наиболее перспективных методов Интеллектуального анализа данных (ИАД, Data Mining). Цель данного метода состоит в том, чтобы на основании имеющейся базы фактов сделать предположения о причинах наличия или отсутствия определённых свойств (целевые свойства) у объекта. Этот метод предложил В. К. Финн [1]. Примером задач, решаемых ДСМ-методом, может служить выявление причинно-следственных закономерностей вида структура-активность в фармакологии, анализ сложных химических соединений и белковых структур в химии.

В данной статье автор предлагает отличный от канонического подход к описанию ДСМ-метода — теоретико-множественный подход.

Ключевые слова: интеллектуальный анализ данных, ИАД, ДСМ-метод, экспертные системы.

1. Введение

В течение последних двух десятилетий произошёл стремительный рост технических возможностей для сбора и хранения больших массивов данных. И этот рост продолжается. Уже накоплены миллионы баз данных, которые охватывают практически все области человеческого знания. Подобный рост накапливаемых данных и их объёма остро поднимает проблему поиска данных, а также, что ещё более важно, порождает необходимость в средствах, позволяющих автоматически извлекать полезные знания из больших массивов данных. Именно к таким средствам и относится Интеллектуальный анализ данных (ИАД).

2. Простой ДСМ-метод

ДСМ-метод[1] автоматического порождения гипотез — это один из методов ИАД. Цель данного метода состоит в том, чтобы на основании имеющейся базы фактов сделать предположения о причинах наличия или отсутствия определённых свойств (целевые свойства) у объекта. Этот метод предложил В. К. Финн [1, 3, 4]. Примером задач, решаемых ДСМ-методом, может служить выявление причинно-следственных закономерностей вида структура-активность в фармакологии, анализ сложных химических соединений и белковых структур в химии [5].

Статья поступила в редакцию 25 декабря 2007 г.

[1] Формализация структурной индукции ДСМ-метода берет своё начало от известного английского философа, логика, историка и социолога Д.С.Милля, чьи инициалы и составляют название метода [2].

Ниже будет рассмотрена теоретико-множественная формулировка классического ДСМ-метода, предложенная автором[2] . От канонической формулировки теоретико-множественный подход выгодно отличается тем, что существенно упрощает построение математической модели ДСМ-системы, а также облегчает её понимание неспециалистами в данной (достаточно узкой) области.

Характерной чертой ДСМ-метода является сочетание трёх разновидностей правдоподобных рассуждений:

— индуктивные рассуждения определяют некоторый способ обучения на примерах и позволяют сформировать гипотезы о возможных причинах рассматриваемых свойств объектов предметной области;
— с помощью рассуждений по аналогии формируются гипотезы о наличии или отсутствии интересующего нас набора свойств у тех объектов предметной области, для которых информация (о наличии у них этих свойств) неполна или противоречива. Процедуры ДСМ-метода обрабатывают некоторое множество исходных данных (фактов или примеров). В результате исполнения процедур, соответствующих индуктивным рассуждениям и рассуждениям по аналогии, формируется некоторое множество гипотез;
— рассуждения по абдукции, основанные на применении следующего правила: если каждый факт (пример) может быть объяснён с помощью имеющихся гипотез, то гипотезы принимаются, т. е. предполагается, что они сформированы на достаточном основании. В том случае, когда посылка правила абдукции ложна, делается вывод о необходимости расширения исходного набора фактов с помощью внешних источников.

Определение 1. *ДСМ-структурой* будем называть кортеж

$$\mathcal{J} = \langle \mathbf{A}, \mathbf{O}, \mathbf{P}, \mathbf{V}, \mathbf{F}, \mathbf{H} \rangle ,$$

где

(1) $\mathbf{A}$ — непустое конечное множество независимых *атомов*,
(2) $\mathbf{O} \subseteq \mathcal{P}(\mathbf{A})$ — непустое конечное множество *объектов* (каждый объект представлен в виде множества атомов)[3] ,
(3) $\mathbf{P}$ — непустое конечное множество (возможных целевых) *свойств* объектов,[4]
(4) $\mathbf{V} = \{+1, -1, 0, \tau\}$ — множество (типов) *внутренних истинностных значений* ($+1$ — *истинно*, -1 — *ложно*, 0 — *противоречиво*, τ — *неопределённо*),
(5) $\mathbf{F} : \mathbf{O} \times \mathbf{P} \to \mathbf{V}$, отображение $\mathbf{F}$ будем называть *функцией обладания*,
(6) $\mathbf{H} : \mathcal{P}(\mathbf{A}) \times \mathbf{P} \to \mathbf{V}$, отображение $\mathbf{H}$ будем называть *функцией причинности*.

Определение 2. Утверждение о наличии или отсутствии у объекта целевого свойства будем называть *фактом*. Множество фактов образует *базу фактов* (БФ). Математическим представлением базы фактов является отображение $\mathbf{F}$ из определения 1.

Определение 3. Будем говорить, что объект $o \in \mathbf{O}$ *обладает* свойством $p \in \mathbf{P}$, если $\mathbf{F}(o, p) = +1$, т. е., если утверждение о том, что объект o обладает свойством p, *истинно*.

Определение 4. Будем говорить, что объект $o \in \mathbf{O}$ *не обладает* свойством $p \in \mathbf{P}$, если $\mathbf{F}(o, p) = -1$, т. е., если утверждение о том, что объект o обладает свойством p, *ложно*.

Определение 5. Объект o будем называть *плюс-примером* для свойства p, если он обладает этим свойством.

Множество всех плюс-примеров для свойства p будем обозначать через $\mathbf{O}(+p)$.

---

[2] В основу данных разработок легли идеи, заложенные О. М. Аншаковым в [6].
[3] Через $\mathcal{P}(\mathbf{A})$ обозначено множество всех подмножеств множества $\mathbf{A}$.
[4] Атомы, объекты и свойства — сущности произвольной природы. Например, в зависимости от конкретной области в роли атомов могут выступать пары вида атрибут-значение, функциональные группы химических соединений, ключевые слова — любые единицы, выступающие в рамках решаемой задачи как неделимая сущность.

Определение 6. Объект o будем называть *минус-примером* для свойства p, если этот объект не обладает свойством p.

Множество всех минус-примеров для свойства p будем обозначать через $\mathbf{O}(-p)$.

Определение 7. Объект o будем называть *нуль-примером* для свойства p, если

$$\mathbf{F}(o,p)=0,$$

т. е. если утверждение о том, что объект o обладает свойством p, *противоречиво*.

Множество всех нуль-примеров для свойства p будем обозначать через $\mathbf{O}(0p)$.

Определение 8. Объект o будем называть *тау-примером* для свойства p, если

$$\mathbf{F}(o,p)=\tau,$$

т. е. если утверждение о том, что объект o обладает свойством p, *неопределённо*.

Множество всех тау-примеров для свойства p будем обозначать через $\mathbf{o}(\tau p)$.

Определение 9. Любое подмножество множества атомов $\mathbf{A}$ будем называть *фрагментом*.

Заметим, что согласно последнему определению объект является *частным случаем* фрагмента.

Определение 10. Будем говорить, что фрагмент $c\subseteq\mathbf{A}$ является *причиной наличия* свойства $p\in\mathbf{P}$ (*плюс-причиной* для свойства p), если

$$\mathbf{H}(o,p)=+1.$$

Утверждение о том, что c является плюс-причиной для p, назовём *положительной гипотезой* (*плюс-гипотезой*) для p.

Множество всех плюс-причин для свойства p будем обозначать через $\mathbf{S}(+p)$.

Определение 11. Будем говорить, что фрагмент $c\subseteq\mathbf{A}$ является *причиной отсутствия* свойства $p\in\mathbf{P}$ (*минус-причиной* для свойства p), если

$$\mathbf{H}(o,p)=-1.$$

Утверждение о том, что c является минус-причиной для p, назовём *отрицательной гипотезой* (*минус-гипотезой*) для p.

Множество всех минус-причин для свойства p будем обозначать через $\mathbf{S}(-p)$.

Определение 12. Будем говорить, что фрагмент $c\subseteq\mathbf{A}$ является *нуль-причиной* для свойства $p\in\mathbf{P}$, если

$$\mathbf{H}(o,p)=0.$$

Утверждение о том, что c является нуль-причиной для p, назовём *противоречивой гипотезой* (*нуль-гипотезой*) для p.

Множество всех нуль-причин для свойства p будем обозначать через $\mathbf{S}(0p)$.

Определение 13. Будем говорить, что фрагмент $c\subseteq\mathbf{A}$ является *тау-причиной* для свойства $p\in\mathbf{P}$, если

$$\mathbf{H}(o,p)=\tau.$$

Утверждение о том, что c является тау-причиной для p, назовём *неопределённой гипотезой* (*тау-гипотезой*) для p.

Тау-причины будем называть также *кандидатами в причины*.

Множество всех тау-причин для свойства p будем обозначать через $\mathbf{S}(\tau p)$.

Определение 14. Будем говорить, что между двумя объектами $o_1,o_2\in\mathbf{O}$ *имеется сходство*, если $o_1\cap o_2\neq\emptyset$.

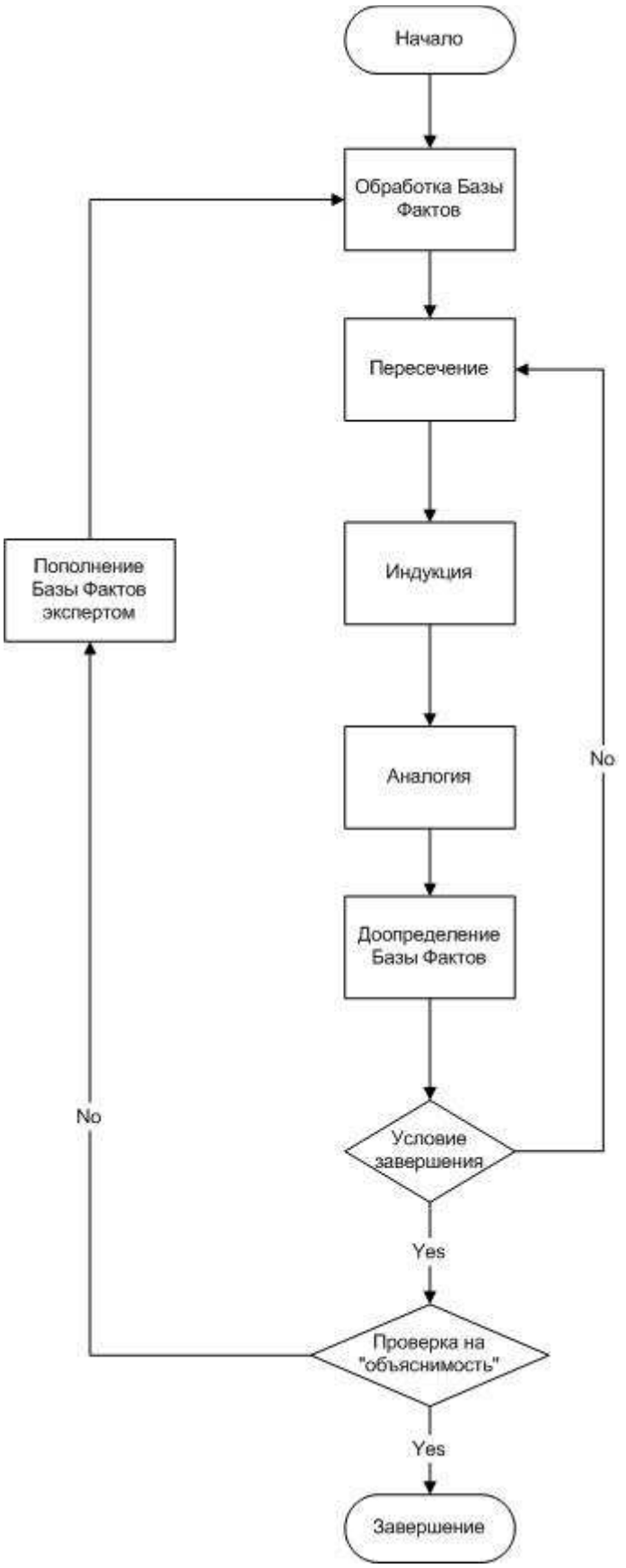

Рис. 1. Схема работы ДСМ-системы

Напомним, что любой объект мы представляем в виде множества атомов (см. определение 1).

Кратко обозначенные на схеме (рис. 1) этапы можно описать следующим образом:

Обработка базы фактов. Получение данных из БФ, формирование множеств плюс-примеров и минус-примеров.

Пересечение. Построение множества всех возможных непустых пересечений плюс-примеров и построение множества всех возможных непустых пересечений минус-примеров.

Индукция. Формирование плюс-гипотез, минус-гипотез и нуль-гипотез.

Аналогия. Формирование гипотез о наличии/отсутствии целевого свойства у объекта на основании их сходства, доопределение базы фактов.

Условие завершения. Процесс работы ДСМ-алгоритма может быть как одношаговым, так проходящим в несколько шагов. В таком случае требуется некоторое условие, означающее прекращение итераций. Подробнее это будет рассмотрено ниже.

Проверка на «объяснимость». Проверка каузальной полноты, все факты в изначальной базе должны объясняться при помощи порождённых системой гипотез. Если этого не происходит, значит требуется изменение исходной базы фактов экспертом.

Данная схема отражает основной принцип работы ДСМ-системы. На вход такой системы поступает база фактов, заданная специалистом, а на выходе выдаётся доопределённая база фактов, т. е. БФ с устранённой неполнотой данных.

Ниже мы остановимся на наиболее интересных этапах подробнее. Но предварительно договоримся об обозначениях.

Соглашение 1. Введём нижний индекс для обозначения номера шага алгоритма ДСМ-метода, на котором получен результат. Например, $\mathbf{O}_n(+p)$ будет означать множество плюс-примеров свойства p, имеющихся после завершения n-го шага. Аналогично понимаются обозначения $\mathbf{O}_n(-p)$, $\mathbf{O}_n(0p)$ и $\mathbf{O}_n(\tau p)$.

Следуя данному соглашению, мы будем обозначать через $\mathbf{S}_n(+p)$ множество плюс-причин свойства p, известных после завершения n-го шага. Аналогично понимаются обозначения $\mathbf{S}_n(-p)$, $\mathbf{S}_n(0p)$ и $\mathbf{S}_n(\tau p)$.

Обозначение 1. Пусть $Y \subseteq \mathcal{P}(\mathbf{A})$. Положим по определению

$$\cap Y = \bigcap_{c \in Y} c.$$

Здесь, как и в определении 1, $\mathcal{P}(\mathbf{A})$ есть множество всех подмножеств множества $\mathbf{A}$.

Теперь опишем этапы ДСМ-метода более подробно.

Пересечение. Пересечение осуществляется любым из известных методов. Традиционно используют метод Норриса или модифицированный метод Норриса, однако есть и другие варианты построения множества пересечений, подробнее об этих методах и их сравнительных характеристиках можно прочитать в [7]. В результате пересечения нужно построить множество пересечений отдельно для плюс-примеров и для минус-примеров. Для полученных множеств пересечений введём специальные обозначения.

Обозначение 2.

$$\mathbf{X}_n^+(p) = \{\cap Y \mid Y \subseteq \mathbf{O}_n(+p), |Y| \geqslant 2\},$$
$$\mathbf{X}_n^-(p) = \{\cap Y \mid Y \subseteq \mathbf{O}_n(-p), |Y| \geqslant 2\}.$$

Здесь, как и в соглашении 1, нижний индекс n понимается как номер шага ДСМ-метода, на котором пересечения были получены.

Индукция. В ходе работы этого модуля идёт поиск сходства на объектах, обладающих (плюс-примеры) и не обладающих (минус-примеры) искомым свойством. На основании найденного сходства находятся в плюс- и минус-причины свойства.

Сначала вводятся вспомогательные предикаты (они же — *решающие предикаты*) $\mathbf{M}_n^+$ и $\mathbf{M}_n^-$.

Обозначение 3.

$$\mathbf{M}_n^+(s,p) \;\Leftrightarrow\; s \in \mathbf{X}_n^+(p),$$
$$\mathbf{M}_n^-(s,p) \;\Leftrightarrow\; s \in \mathbf{X}_n^-(p).$$

Используя введённые обозначения, мы можем определить правила индукции.

Определение 15. Следующие правила называются *правилами индукции*, или *правилами первого рода*:

$$(\mathrm{I}^+) \quad s \in \mathbf{S}_n(\tau p) \;\&\; \mathbf{M}_n^+(s,p) \;\&\; \neg\,\mathbf{M}_n^-(s,p) \quad\to\quad s \in \mathbf{S}_{n+1}(+p),$$

$$(\mathrm{I}^-) \quad s \in \mathbf{S}_n(\tau p) \;\&\; \mathbf{M}_n^-(s,p) \;\&\; \neg\,\mathbf{M}_n^+(s,p) \quad\to\quad s \in \mathbf{S}_{n+1}(-p),$$

$$(\mathrm{I}^0) \quad s \in \mathbf{S}_n(\tau p) \;\&\; \mathbf{M}_n^+(s,p) \;\&\; \mathbf{M}_n^-(s,p) \quad\to\quad s \in \mathbf{S}_{n+1}(0p),$$

$$(\mathrm{I}^\tau) \quad s \in \mathbf{S}_n(\tau p) \;\&\; \neg\,\mathbf{M}_n^+(s,p) \;\&\; \neg\,\mathbf{M}_n^-(s,p) \quad\to\quad s \in \mathbf{S}_{n+1}(\tau p).$$

Словами эти правила можно выразить так: если до текущего момента про некоторый фрагмент неизвестно (не было выявлено во время предыдущих шагов), включается ли он в какую-либо гипотезу, и этот фрагмент принадлежит множеству пересечений плюс-примеров и не принадлежит множеству пересечений минус-примеров, то данный фрагмент принадлежит множеству кандидатов в плюс-причины.

Двойственным образом задаётся правило для получения минус-причин. Иногда их также называют *антипричинами*.

Если фрагмент принадлежит одновременно и множеству пересечений плюс-примеров, и минус-примеров, то выдвигается гипотеза, что наличие этого фрагмента ведёт к противоречию (противоречивая причина).

Если фрагмент не принадлежит ни множеству пересечений плюс-примеров, ни множеству пересечений минус-примеров, то он остаётся в числе неопределённых.

Аналогия. В данном блоке на основании найденных ранее причин происходит выдвижение гипотез о наличии / отсутствии искомого свойства у объекта. Или же о имеющем место фактическом противоречии.

Обозначение 4. Вводятся вспомогательные выражения, *решающие предикаты второго рода*, имеющие значением наличие плюс/минус/противоречивой причины для данного объекта и данного свойства:

$$\mathbf{G}_{n+1}^+(o,p) \;\Leftrightarrow\; \exists s \in \mathbf{S}_{n+1}(+p)\,(s \subseteq o),$$
$$\mathbf{G}_{n+1}^-(O,p) \;\Leftrightarrow\; \exists s \in \mathbf{S}_{n+1}(-p)\,(s \subseteq o),$$
$$\mathbf{G}_{n+1}^0(O,p) \;\Leftrightarrow\; \exists s \in \mathbf{S}_{n+1}(0p)\,(s \subseteq o).$$

Определение 16. Сами правила второго рода выглядят следующим образом:

$$(\mathrm{II}^+) \quad o \in \mathbf{O}_n(\tau p) \;\&\; \mathbf{G}_{n+1}^+(s,p) \;\&\; \neg\,\mathbf{G}_{n+1}^-(s,p) \quad\to\quad s \in \mathbf{O}_{n+1}(+p),$$

$$(\mathrm{II}^-) \quad o \in \mathbf{O}_n(\tau p) \;\&\; \mathbf{G}_{n+1}^-(s,p) \;\&\; \neg\,\mathbf{G}_{n+1}^+(s,p) \quad\to\quad s \in \mathbf{O}_{n+1}(-p),$$

$$(\mathrm{II}_1^0) \quad o \in \mathbf{O}_n(\tau p) \;\&\; \mathbf{G}_{n+1}^+(s,p) \;\&\; \mathbf{G}_{n+1}^-(s,p) \quad\to\quad s \in \mathbf{O}_{n+1}(0p),$$

$$(\mathrm{II}_2^0) \quad o \in \mathbf{O}_n(\tau p) \;\&\; \mathbf{G}_{n+1}^0(s,p) \quad\to\quad s \in \mathbf{O}_{n+1}(0p),$$

$$(\mathrm{II}^\tau) \quad o \in \mathbf{O}_n(\tau p) \;\&\; \neg\,\mathbf{G}_{n+1}^+(s,p) \;\&\; \neg\,\mathbf{G}_{n+1}^-(s,p) \quad\to\quad s \in \mathbf{O}_{n+1}(\tau p).$$

Словами это можно описать следующим образом: если обладание искомым свойством до текущего момента для объекта не установлено (не было выявлено на предыдущих шагах), имеются плюс-причины (причины, указующие на наличие свойства) и не имеется минус-причин, то утверждается, что данный объект искомым свойством обладает.

Двойственно, если установлено, что для такого объекта есть минус-причины и нет плюс-причин, утверждается, что объект искомым свойством не обладает.

Если для данного объекта имеются как плюс-причины, так и минус-причины, то для данного объекта устанавливается фактическое противоречие. Также противоречие устанавливается[5] , если для данного объекта существуют противоречивые причины (см. сноску 3).

Ну и наконец, если для объекта не удаётся обнаружить никаких причин, он остаётся среди тех, для кого обладанием искомым свойством не установлено.

Условие завершения. Процесс работы ДСМ-алгоритма может быть как одношаговым, так и итерационным, т. е. имеющим несколько шагов, в ходе каждого из которых процедура «пересечение» → «индукция» → «аналогия» работает заново, беря за основу данные, имеющиеся на текущий момент, т. е. как исходные, так и полученные на предыдущих шагах работы. В таком случае условием завершения будет то, что на очередном шаге исходная и доопределённая база фактов совпадают, т. е. не удалось установить факт наличия/отсутствия целевого свойства ни для одного объекта.

Проверка на «объяснимость» (*проверка каузальной полноты*) состоит в том, что все уже известные факты должны быть *объяснимы*. В противном случае утверждается, что база фактов нуждается в пополнении извне.

Ранее был рассмотрен теоретико-множественный подход к классическому (простому) ДСМ-методу. Однако это далеко не единственный возможный вариант построения ДСМ-системы. Покажем, что теоретико-множественная формулировка возможна и для иных вариаций ДСМ-метода и будет выглядеть следующим образом:

Например, существует ещё ДСМ-метод с запретом на контр пример. Существенное его отличие от простого заметно на этапе индукции: Если *решающие предикаты* M_n^+ и M_n^- определяются так же:

$$\mathbf{M}_n^+(s,p) \;\Leftrightarrow\; s \in \mathbf{X}_n^+(p)\,,$$
$$\mathbf{M}_n^-(s,p) \;\Leftrightarrow\; s \in \mathbf{X}_n^-(p)\,,$$

то само рассуждение (правила первого рода) имеет существенное отличие: фрагмент определяется как плюс-причина (минус-причина) в том случае, если он входит в множество пересечений плюс-примеров (минус-примеров) и не существует такого минус-примера (плюс-примера), в который бы данный фрагмент входил:

$$(\mathrm{I}^+) \qquad s \in \mathbf{S}_n(\tau p)\ \&\ \mathbf{M}_n^+(s,p)\ \&\ \neg\exists o \in O(-p)\,(s \subseteq o) \quad \rightarrow \quad s \in \mathbf{S}_{n+1}(+p),$$

$$(\mathrm{I}^-) \qquad s \in \mathbf{S}_n(\tau p)\ \&\ \mathbf{M}_n^-(s,p)\ \&\ \neg\exists o \in O(+p)\,(s \subseteq o) \quad \rightarrow \quad s \in \mathbf{S}_{n+1}(-p).$$

Правило для противоречивости не вводится, а правило, сохраняющее неопределённость, выглядит также:

$$(\mathrm{I}^\tau) \qquad s \in \mathbf{S}_n(\tau p)\ \&\ \neg\,\mathbf{M}_n^+(s,p)\ \&\ \neg\,\mathbf{M}_n^-(s,p) \quad \rightarrow \quad s \in \mathbf{S}_{n+1}(\tau p)$$

Рассуждение по аналогии (правила второго рода) для ДСМ-метода с запретом на контрпример выглядят так же, как и в простом методе.

3. Несимметричный ДСМ-метод и его модификации

Теперь же рассмотрим теоретико-множественный подход в наиболее интересной, с нашей точки зрения, модификации ДСМ-метода — несимметричном ДСМ-методе [6] .

[5] В классическом ДСМ-методе данного правила нет, однако для полноты системы кажется естественным определить его и использовать полученный выше $G_{n+1}^0(o,p)$.

[6] Его каноническая формулировка была изложена Д. В. Виноградовым [8].

Основным отличием несимметричного ДСМ-метода от простого является то, что обобщённый ДСМ-метод является методом «контекстным». Это значит, что если в простом методе ищутся абсолютные минус-гипотезы (антипричины), то обобщённый метод вместо этого ищет локальные тормоза для каждой конкретной гипотезы, которые работают только для той причины, для которой они были найдены. Таким образом, несложно заметить, что асимметричный ДСМ-метод лучше доопределяет матрицу свойств в тех случаях, когда для предметной области важна именно контекстность.

Обозначение 5. O(+s, +p) — множество плюс-примеров свойства **p**, содержащих фрагмент **s**.

Обозначение 6. O(+s, -p) — множество минус-примеров свойства **p**, содержащих фрагмент **s**.

Обозначение 7. O(+s, τp) — множество объектов, содержащих фрагмент **s**, для которых факт наличия / отсутствия свойства **p** не установлен.

Определение 17. Фрагмент $\mathbf{s}_1$ называется *тормозом* фрагмента **s** для свойства **p**, если выполняется следующее условие: он является подмножеством хотя бы в двух объектах, содержащих s и не обладающих p.

Формально это можно записать следующим образом:

$$\exists o_1 o_2 (o_1 \neq o_2 \& o_1 \in O_n(+s, -p) \& o_2 \in O_n(+s, -p) \& s_1 \subseteq o_1 \& s_1 \subseteq o_2),$$

где O — множество всех объектов.

Замечание. Возможен также подход, при котором тормоз не включает в себя причину (из полученного выше пересечения вычитается сам фрагмент **s**, т. е. $\mathbf{s} \cap \mathbf{s}_1 = \emptyset$). Выбор формулировки не оказывает существенного влияния на работу обобщённого ДСМ-метода и никак не меняет результата его работы.

Правила правдоподобного вывода ДСМ-метода делятся на две группы. Правила первой группы служат для порождения гипотез о структурных причинах возникновения данных свойств у объектов. Такие правила называются правилами первого рода. Правила второй группы позволяют выдвигать гипотезы о наличии/отсутствии свойств у объектов и называются правилами второго рода.

Рассмотрим модификацию несимметричного ДСМ-метода, которая, как и сам несимметричный ДСМ-метод, порождает только плюс-причины и их тормоза, однако имеет несколько отличий. Понятия минус-гипотезы не вводится как такового, а правила первого рода порождают только кандидатов в плюс-гипотезы. Таких правил два[7] :

Правило I_1^+: $s \in S_n(\tau p) \& \exists o_1 o_2 (o_1 \neq o_2 \& o_1 \in O_n(+s, +p) \& o_2 \in O_n(+s, +p)) \& O_n(+s, -p) = \emptyset \to s \in S_{n+1}(+p)$. Этим правилом порождаются кандидаты в абсолютные плюс-гипотезы, т. е. в гипотезы, не имеющие тормозов.

Обозначение 8. Введём обозначение $T_n^+(s, p)$, которое следует понимать как множество тормозов возможной причины **s** свойства **p**, где n — номер шага ДСМ-метода, не позднее которого были обнаружены эти тормоза.

Через $T_n(s, p)$ обозначим множество кандидатов в тормоза для гипотезы **s** и свойства **p**, где n — номер шага ДСМ-метода, вплоть до которого не было выяснено, являются ли эти фрагменты тормозами: $t \in T_n^\tau(s, p) \& \exists o_1 o_2 (o_1 \neq o_2 \& o_1 \in O_n(+s, -p) \& o_2 \in O_n(+s, -p) \& t \subseteq o_1 \& t \subseteq o_2) \to t \in T_{n+1}^+(s, p)$.

Правило I_2^+: $s \in S_n(\tau p) \& \exists o_1 o_2 (o_1 \neq o_2 \& o_1 \in O_n(+s, +p) \& o_2 \in O_n(+s, +p)) \& O_n(+s, -p) \neq \emptyset \& T_{n+1}^+(s, p) \neq \emptyset \to s \in S_{n+1}(+p)$.

Существенная модификация данного метода заключается в том, что при такой формулировке ППВ первого рода отслеживается случай, когда множество

[7] В несимметричном ДСМ-методе, сформулированном Д. В. Виноградовым [8], такое правило одно, без выделения отдельного правила для абсолютных гипотез:
$s \in S_n(\tau p) \& \exists o_1 o_2 (o_1 \neq o_2 \& o_1 \in O_n(+s, +p) \& o_2 \in O_n(+s, +p)) \to s \in S_{n+1}(+p)$.
Ниже будет показано, почему факт наличия или отсутствия контр-примеров для данной гипотезы важен.

минус-примеров не пусто, но при этом сходства между минус-примерами найти не удаётся. Это означает, что, несмотря на наличие контрпримеров, тормоза выделить не удаётся, множество тормозов пусто. В подобном случае утверждается, что для выявления тормозов недостаточно данных, а значит недостаточно данных и для выдвижения гипотезы с такой фрагмент-причиной. В результате, данный кандидат в гипотезы отвергается.

При этом противоречия не порождаются ни одним из правил вывода.

При доопределении матрицы свойств работают два правила:

Правило A^+:

$$o \in O_n(+s, \tau p) \& s \in S_{n+1}(+p) \& \neg \exists t (t \in T^+_{n+1}(s,p) \& t \subseteq o) \to o \in O_{n+1}(+s, +p).$$

Правило A^- имеет два варианта: для одношагового метода и итеративного.

Для одношагового метода это правило выглядит просто: если не A^+, то A^+. т. е. если к данному объекту не удалось ни разу применить ни одну из гипотез и правило A^+ ни разу не отработало, то данный объект считается не обладающим искомым свойством.

Для итеративного метода данное правило менее строго:

$$o_n \in O_n(+s, \tau p) \& \neg (s_n \in S_{n+1}(+p)) \vee \exists t_n (t \in T^+_{n+1}(s,p) \& t \subseteq o) \to o_n O_{n+1}(+s, \tau p),$$

т. е. если не удаётся применить ни одной гипотезы о наличии свойства, сохраняется неопределённость, которая переходит на следующую итерацию, и только на последнем шаге итерации оно работает аналогично тому, как для одношагового случая.

Опишем применение модифицированного несимметричного ДСМ-метода поэтапно:

1. Найти все непустые пересечения двух и более объектов из множества плюс-примеров. В результате получаем множество $X^+_n(p)$ кандидатов в возможные причины.
2. Для каждой возможной причины $s \in X^+_n(p)$:
   - 2.1. Если множество минус-примеров пусто ($O_n(+s, -p) = \emptyset$), то **s** — абсолютная причина **p**.
   - 2.2. Если множество минус-примеров не пусто ($O_n(+s, -p) \neq \emptyset$), то найдём все непустые пересечения двух и более объектов из множества $O_n(+s, -p)$.
   - 2.3. Если таких пересечений нет и множество $T^+_{n+1}(s,p)$ пусто (в частности, если множество $O_n(+s, -p)$ содержит всего один объект), то мы не можем выдвигать никаких гипотез касательно данного фрагмента, так как не смогли выявить тормозов.
3. Строим $T^+_{n+1}(s,p)$ по уже приведённой выше формуле.
4. Доопределяем матрицу фактов. При этом используются сформулированные выше **Правило $\mathbf{A}^+$** и **Правило $\mathbf{A}^-$**.

При желании можно сделать итерацию всего процесса, то есть после того, как алгоритм отработает один раз, снова запустить поиск гипотез, но уже с учётом доопределённых фактов. Таким образом получатся новые гипотезы с новыми тормозами, которые могут помочь доопределить факт наличия или отсутствия целевого свойства для ряда объектов, которые не были доопределены после первого применения описанной выше процедуры.

4. Некоторые оптимизирующие модификации

При практической реализации этой процедуры целесообразно произвести ряд упрощений, направленных на уменьшение объёма хранимых кандидатов в гипотезы.

Во-первых, имеет смысл объединить этапы 1 и 2, то есть включить этап проверки наличия тормозов (а заодно и их выявления) уже на этапе поиска кандидатов на гипотезы.

Во-вторых, представляется возможным уменьшение объёма хранимых гипотез, если не просто проверять на дубликаты, чтобы избежать хранения одинаковых гипотез, но и выполнять проверку на более общие гипотезы. Предположим, имеется гипотеза $\langle \mathbf{s}, \mathbf{T}^+_{n+1}(\mathbf{s},\mathbf{p})\rangle$, где $\mathbf{s} \in \mathbf{S}_{n+1}(+\mathbf{p})$, и найдена гипотеза $\langle \mathbf{fr}, \mathbf{T}^+_{n+1}(\mathbf{fr},\mathbf{p})\rangle$, где $\mathbf{fr} \in \mathbf{S}_{n+1}(+\mathbf{p})$, такая, что $\mathbf{fr} \subseteq \mathbf{s}$. Тогда все, что нам удастся породить при помощи первой гипотезы, будет порождено и посредством второй. Это означает, что гипотезу $\langle \mathbf{s}, \mathbf{T}^+_{n+1}(\mathbf{s},\mathbf{p})\rangle$ можно исключить без потерь (нетрудно показать, что $\mathbf{T}^+_{n+1}(\mathbf{s},\mathbf{p}) \subseteq \mathbf{T}^+_{n+1}(\mathbf{fr},\mathbf{p})$).

Таким образом может быть введено понятие силы гипотез, где гипотеза $\langle \mathbf{fr}, \mathbf{T}^+_{n+1}(\mathbf{fr},\mathbf{p})\rangle$ называется более сильной чем другая гипотеза $\langle \mathbf{s}, \mathbf{T}^+_{n+1}(\mathbf{s},\mathbf{p})\rangle$ тогда, когда она, среди прочих, порождает и все то, что можно породить при помощи гипотезы $\langle\ \mathbf{s}, \mathbf{T}^+_{n+1}(\mathbf{s},\mathbf{p})\rangle$, которую будем называть более слабой.

Хотелось бы сказать пару слов относительно того, стоит ли рассматривать отсутствие признака также как признак. Дело в том, что в каноническом подходе рассмотрению подлежат только наличествующие признаки, что, в принципе, является логичным. Однако в случае с контекстными методами подобный подход не всегда полезен, иногда отсутствие признаков, взятое наравне с их присутствием, позволяет существенно расширить наши возможности.[8]

Возьмём в качестве иллюстрации, обосновывающей необходимость такого подхода, небольшой искусственный пример (**a,b,c,d,e,f** — признаки, **P** — целевое свойство):

a	b	c	d	e	f	P
1	1	1	0	0	0	+
1	1	1	1	0	0	+
1	1	0	0	1	0	+
1	1	0	1	0	0	-
1	1	0	0	0	1	-

Давайте рассмотрим кандидата в гипотезы **ab**. Для него имеется два контрпримера. Однако при этом, если мы будем рассматривать только наличие признаков, ни одного тормоза выявить не удастся — оба имеющихся контрпримера пересекаются только по гипотезе, и кандидат в гипотезу будет отбракован. Если же принять к рассмотрению и отсутствие признаков, получается, что в данном примере имеется две гипотезы: абсолютная причина **abc** и гипотеза **ab**, для которой тормозом будет являться $\neg\mathbf{c}\neg\mathbf{e}$.

Таким образом, получается, что если считать значимым только наличие признака, то это может привести к нежелательным потерям среди гипотез: как несложно заметить, при таком подходе здесь мы потеряем гипотезу **ab** (наличие минус-примеров при невозможности вычленить тормоза) и остаёмся только с гипотезой **abc**, являющейся абсолютной причиной. Однако, как несложно заметить, эта гипотеза уже не в состоянии объяснить того факта, что у третьего объекта наличествует свойство **p**, а значит, не будет выполнена проверка на объяснимость.

Кроме того, если у нас появится объект, скажем, **abef**, который надо будет доопределить исходя из имеющейся на данный момент теории, то без гипотезы $\langle \mathbf{ab},\mathbf{p},\{\ \neg\mathbf{c}\neg\mathbf{e}\ \}\rangle$ доопределить его мы будем просто не в силах, в то время, как на самом деле согласно данной гипотезе легко получается, что на самом деле объект **abef** обладает свойством **p**.

Литература

1. *Финн В. К.* Базы данных с неполной информацией и новый метод автоматического порождения гипотез // Диалоговые и фактографические системы информационного обеспечения. — М., 1981.

[8] В практических компьютерных реализациях подобный подход используется достаточно часто, однако он является неформальным, т.к. не прописан в теоретическом аппарате.

2. *Милль Д. С.* Система логики силлогистической и индуктивной. — М.: Книжное дело, 1900.
3. *Финн В. К.* О возможностях формализации правдоподобных рассуждений средствами многозначных логик // Всесоюзный симпозиум по логике и методологии науки. — Киев: Наукова думка, 1976.
4. *Финн В. К.* О машинно-ориентированной формализации правдоподобных рассуждений в стиле Ф. Бэкона–Д. С. Милля // Семиотика и информатика. — Вып. 20. — 1983.
5. Применение ДСМ-метода порождения гипотез для прогноза противоопухолевой активности и токсичности соединений, принадлежащих к различным классам химических соединений / Е. С. Панкратова, В. Г. Ивашко, В. Г. Блинова, Д. В. Попов // Экспертные системы: состояние и перспективы / Под ред. Д. А. Поспелова. — М.: Наука, 1989.
6. *Аншаков О. М.* Об одной интерпретации ДСМ-метода автоматического порождения гипотез // НТИ, сер 2., № 1–2. — М.: ВИНИТИ РАН, 1999.
7. *Обедков С. А.* Алгоритмические аспекты ДСМ-метода и формального анализа понятий. — М.: РГГУ, 1999.
8. *Виноградов Д. В.* Несимметричный ДСМ-метод с учетом контекста // Пятая национальная конференция с международным участием «Искусственный интеллект-96». — Казань: 1996. — КИИ-96: Сб. науч. тр.: В 3 т.— Казань : Ассоц. искусств. интеллекта, 1996.

UDC 004.891

One Method of Analysis for Large Sets of Partly Deterministic Data

A. A. Lipkin

Intelligent System Department
Russian State University of Humanities
Miusskaya sq., 6, Moscow, Russia, 125267

JSM-method of automatic hypothesis generation is one of the most promising methods in Data Mining. The goal of this method is the following: from the given facts' database as a starting point make suggestions on the cause of an object possessing or not possessing some properties. This method was introduced by V.K. Finn [1]. Searching for causal-investigatory regularities in pharmacology and analysis of complicated compounds and protein formations in chemistry might serve an example of its usage.

The author suggests another approach, different from the canonical one, in this article. That is set-theoretic approach.