Теория модулей и модульные сети находятся в начальной стадии развития. Статья отражает их состояние на сегодня и намечает пути дальнейшего развития.

Материалы статьи позволяют высказать предположение, что практическое использование модульных сетей приведет к широкому распространению парадигмы модульного мышления, как нового метода рассуждения людей о модулях и модульных системах.

СПИСОК ЛИТЕРАТУРЫ

1. Grenander U. Lectures in Pattern Theory.— Springer-Verlag, N. Y., Heidenberg Berlin. Vol I. Pattern Synthesis, 1976; Vol II. Pattern Analysis, 1978; Vol. III. Regular Structures, 1981.

2. Grenander U. General Pattern Theory.— Oxford University Press, 1993.— 904 p.

3. Шуткин Л. В. Паттерновое моделирование гипертекстов // НТИ. Сер. 2.— 1995.— № 9.— С. 20–26.

4. Шуткин Л. В. Результаты и перспективы применения теории паттернов к компьютерам // НТИ. Сер. 2.— 1996.— № 12.— С. 13–21.

5. Шуткин Л. В. Новое мышление компьютерного мира // НТИ. Сер. 2.— 2000.— № 12.

6. Шуткин Л. В. Практика способа консультативно-гипертекстового обучения // НТИ. Сер. 1.— 2004.— № 6.

Материал поступил в редакцию 16.06.04.

УДК 004.6:0018

М. В. Максин

Об одном подходе к проблеме комбинированного использования логических и численных методов в интеллектуальном анализе данных

Предлагается подход к проблеме интеллектуального анализа гибридных (структурно-числовых) данных на примере ДСМ-системы прогнозирования контрпродуктивных свойств химических соединений. Даётся краткое описание ДСМ-метода автоматического порождения гипотез и его настройки на анализ числовых данных. Приводятся результаты экспериментов с комбинированной системой.

Интеллектуальный анализ данных (ИАД) как одно из направлений искусственного интеллекта призван решить задачу автоматизации процесса извлечения формализованных знаний из массивов структурированных данных и сократить разрыв в темпах накопления и осмысления данных, вызванный бурным развитием технологий сбора, передачи и хранения информации. Методы ИАД [1] прошли развитие от регрессионного анализа, работающего с простейшим представлением информации — числовыми векторами, — до методов, способных воспринимать такие выразительные конструкции, как графы и пропозициональные формулы, и формализующих средствами математической логики философские представления о природе познания.

Сейчас всё большее внимание уделяется методам, комбинирующим познавательные (логико-комбинаторные) процедуры со статистическими (вычислительными) процедурами. Такие методы позволяют учитывать в анализе как структурные, так и числовые характеристики изучаемых объектов, а также подкрепить сравнительно молодой, но чрезвычайно перспективный аппарат формального логического анализа многолетним опытом разработок в области статистического анализа. При этом числовые характеристики и числовые модели, отражающие "физику" изучаемых явлений и процессов (как, например, энергия активации в задаче "структура-активность"), могут являться важным элементом настройки интеллектуальной системы анализа на конкретную предметную область.

Прототип такой гибридной интегрированной системы, реализованный в Отделе интеллектуальных систем ВИНИТИ, с успехом применялся к задаче прогнозирования канцерогенности химических соединений. В этой системе ДСМ-решатель [2] дополнен квантово-механическим модулем [3], производящим преобразование исходных соединений в соответствии с некоторой моделью метаболизма, а также вносящим в описание изучаемых соединений числовую характеристику — значение энергии образования метаболита из исходного соединения. В ходе экспериментов эта система породила непротиворечивый набор гипотез о причинах наличия (или отсутствия) канцерогенности, способных объяснить каждое из обучающих соединений, чего не удавалось получить раньше.

Рассмотрим подробнее принцип работы данной системы. И начнём с логико-комбинаторного ядра системы — ДСМ-метода автоматического порождения гипотез.

ДСМ-МЕТОД АВТОМАТИЧЕСКОГО ПОРОЖДЕНИЯ ГИПОТЕЗ

ДСМ-метод автоматического порождения гипотез (АПГ) был предложен В. К. Финном в начале 80-х гг. и основывается на применении формального аналога метода сходства Джона Стюарта Милля [4], откуда и получил своё название. Он представляет собой ориентированную на компьютер-

ные приложения формализацию некоторого класса правдоподобных рассуждений, позволяющих на основе анализа имеющихся данных формировать гипотезы о причинах наличия (или отсутствия) у объектов изучаемых свойств, а затем применять эти гипотезы для прогноза свойств у изучаемых объектов.

Метод сходства Д. С. Милля звучит следующим образом [5]: *если два или более случаев подлежащего исследованию явления имеют общим лишь одно обстоятельство, то это обстоятельство, в котором только и согласуются все эти случаи, есть причина (или следствие) данного явления.* Понятно, что рассуждение, заключённое в методе сходства, является *правдоподобным*, т. е. следующим "логике здравого смысла", но не достоверным. Однако именно такие рассуждения способны сделать некоторое индуктивное обобщение, и, таким образом, возможно, получить *новое* знание, в то время как при достоверном рассуждении, реализуемом дедуктивным выводом, происходит лишь применение имеющихся знаний. В ДСМ-методе мера правдоподобия некоторого заключения выражается посредством многозначных логик и порождается конструктивно посредством аппарата рассуждения, комбинирующего абдукцию, индукцию и аналогию [6].

Из природы ДСМ-метода АПГ вытекают условия его применимости:

(1) *Исследуемое явление* есть частично определённое отношение "обладать множеством свойств", при этом в качестве *случаев явления* выступают конкретные объекты и их свойства, а под *обстоятельствами случаев* понимаются фрагменты описания объектов.

(2) Для формализации понятия "общее обстоятельство двух или более случаев явления" необходимо, чтобы для фрагментов описания объектов было определено отношение вложимости, а для конструктивного нахождения "общих обстоятельств" — операция сходства, сопоставляющая двум объектам третий, выражающий их общий фрагмент.

(3) В массиве данных имеются положительные и отрицательные примеры отношения "обладать множеством свойств" и примеры неопределённости этого частично определённого отношения.

(4) В этих примерах также содержатся в неявном виде зависимости причинно-следственного типа, представимые отношениями "быть причиной наличия множества свойств" и "быть причиной отсутствия множества свойств".

(5) Для получения содержательных индуктивных обобщений (гипотез) выдвигается дополнительное требование интерпретируемости результата операции сходства.

Изложим общий принцип работы систем, реализующих ДСМ-метод АПГ (такие системы известны как системы типа ДСМ или ДСМ-системы) на более формальном уровне.

Входом системы типа ДСМ является база данных с неполной информацией (БДНИ) в виде частично определённых отношений $C \Rightarrow_1^* A$ "объект C обладает множеством свойств A" (примеров, как положительных, так и отрицательных) и $C \Rightarrow_2^* A$ "подобъект C является причиной наличия множества свойств A" (уже известных причин как наличия свойств, так и их отсутствия). Эти отношения представлены посредством предикатов бесконечнозначной логики $\tilde{V}_\tau^{(\infty)}$ [7]. В этой логике истинностное значение формулы $\langle v, n\rangle$ означает приписывание типа истинностного значения $v \in \{1$ (фактическая истина), -1 (фактическая ложь), 0 (фактическая противоречивость)$\}$ на n-м шаге применения правил правдоподобного вывода (п. п. в.). Эти отношения отражают тем самым степень правдоподобия, с которой данная формула "истинна", "ложна" или "противоречива": при $n = 0$ — это факт, при $n > 0$ — это гипотеза, тем менее правдоподобная, чем больше n. Результатом работы является БДНИ, пополненная гипотезами о наличии свойств у конкретных объектов, доопределяющими отношение $\Rightarrow_1^*$, и база знаний (БЗ), образованная гипотезами о причинно-следственных зависимостях, доопределяющими отношение $\Rightarrow_2^*$.

Для формального представления БДНИ и БЗ используется аппарат квазиаксиоматических теорий (КАТ) [8]. КАТ $\mathfrak{J}$ имеет вид

$$\mathfrak{J} = \langle \Sigma, \Sigma', \mathfrak{R} \rangle,$$

где Σ — множество аксиом, заведомо неполно характеризующих предметную область; Σ' — открытое множество полуфактов (фактов и гипотез), пополняемое в процессе обучения; $\mathfrak{R} = \mathfrak{R}^0 \cup \mathfrak{R}'$ — множество правил вывода, где $\mathfrak{R}^0$ — множество правил достоверного (дедуктивного) вывода, а $\mathfrak{R}'$ — множество правил правдоподобного вывода.

Множество полуфактов имеет следующую структуру: $\Sigma' = \bar{\Gamma}' \cup \bar{\Gamma}_0 \cup \bar{\Delta}$, где $\bar{\Gamma}_0$ есть множество фактов о причинно-следственных зависимостях (начальное определение отношения $\Rightarrow_2^*$), $\bar{\Delta} = \bigcup\limits_{n=0} \bar{\Delta}_n$ — множество полуфактов о наличии свойств у конкретных объектов (текущее определение отношения $\Rightarrow_1^*$), $\bar{\Gamma}' = \bigcup\limits_{n=1} \bar{\Gamma}'_n$ — множество гипотез о причинно-следственных зависимостях (текущее доопределение отношения $\Rightarrow_2^*$). Множество аксиом имеет следующую структуру:

$$\Sigma = \bigcup_{m=0} \Sigma^{(m)},\ \Sigma^{(m)} = \Sigma_{core} \cup \Sigma_{tune}^{(m)},\ \Sigma_{core} =$$

$$= \Sigma_{pr} \cup \Sigma_C \cup \Sigma_{mix},\ \Sigma_{tune}^{(m)} = \Sigma_{DS} \cup \Sigma_{UD}^{(m)},$$

где Σ_{core} — аксиомы ядра КАТ, независимые от предметной области (ПО), включающие:

Σ_{pr} — процедурные аксиомы, представляющие п. п. в. в декларативной форме;

Σ_C — декларативные аксиомы связи исходных предикатов $\Rightarrow_1$ и $\Rightarrow_2$, характеризующие класс решаемых задач (например, "аксиомы казуальной непротиворечивости", утверждающие, в частности, что если объект X *фактически обладает* набором свойств Y, то никакой его подобъект $Z \subset X$ не может быть объявлен *причиной отсутствия* подмножества свойств $W \subseteq Y$);

Σ_{mix} — смешанные процедурно-декларативные аксиомы, контролирующие "качество" правдоподобного вывода, например, критерий достаточного основания принятия гипотез (к. д. о. п. г.), требующий, чтобы каждый из исходных фактов был объяснён финальным множеством гипотез;

Σ_{tune} — ПО-зависимые аксиомы, позволяющие настраивать (fine tune) систему на конкретную предметную область (ПО):

Σ_{DS} — аксиомы используемой структуры данных;

$\Sigma_{UD}^{(m)}$ — аксиомы предметной области, пополняемые индуктивными обобщениями, порождаемыми в процессе обучения.

Таким образом, состояние знания о ПО на каждом этапе обучения описывается КАТ $\Im_{m,n} = \langle \Sigma^{(m)}, \Sigma'_n, \Re \rangle$, причём состоянию БДНИ отвечает множество формул $\bar{\Gamma}_0 \cup \bar{\Delta}_n$, а структура БЗ характеризуется $\Im_{m,n} = \langle \Sigma^{(m)}, \bar{\Gamma}'_n, \Re \rangle$.

Ядром системы типа ДСМ является метод ДСМ-АПГ, образованный правилами правдоподобного вывода двух родов — $(I_x^{(\sigma)})$ и $(II^{(\sigma)})$, где $\sigma \in \{+, 0, \tau, -\}$. П. п. в. основаны на решающих предикатах, которые существенно используют операцию "сходства" на множестве объектов — операцию, которая для двух данных объектов возвращает максимальный подобъект, принадлежащий им обоим. Это по сути требует, чтобы на множестве объектов была задана нижняя полурешётка [9]. Простейшим способом задания операции сходства служит булева операция пересечения $\cap$, задающая на произвольном множестве U нижнюю полурешётку $\langle 2^{\mathrm{U}}, \cap, \varnothing \rangle$. Приведем алгоритм работы системы типа ДСМ:

• Начальное состояние: $n = 0$, $m = 0$, $\bar{\Gamma}' = \bar{\Gamma}_0$, $\bar{\Delta} = \bar{\Delta}_0$, $\Sigma_{UD} = \Sigma_{UD}^{(0)}$.

• Пока не установлена "сходимость" ("все факты объяснены" или "некоторое множество фактов объяснено и остаётся устойчивым") или не обнаружена "расходимость" (появляются новые "необъяснённые" факты):

∘ $m = m + 1$.

∘ Пока образуются новые гипотезы ($\bar{\Gamma}'_n \neq \varnothing \vee \bar{\Delta}_n \neq \varnothing$), применять ДСМ-АПГ:

♦ $n = n + 1$,

♦ получить $\bar{\Gamma}'_n$ применением п. п. в. $(II^{(\sigma)})$ к $\bar{\Delta}$,

♦ получить $\bar{\Delta}_n$ применением п. п. в. $(I_x^{(\sigma)})$ к $\bar{\Gamma}' \cup \bar{\Gamma}'_n$,

♦ исключить из $\bar{\Gamma}'_n \cup \bar{\Delta}_n$ гипотезы, противоречащие аксиомам

$$\Sigma_C \cup \Sigma_{DS} \cup \Sigma_{UD},$$

♦ $\bar{\Gamma}' = \bar{\Gamma}' \cup \bar{\Gamma}'_n$, $\bar{\Delta} = \bar{\Delta} \cup \bar{\Delta}_n$.

∘ Проверить $\bar{\Gamma}' \cup \bar{\Delta}_0$ на непротиворечивость с Σ_{mix} (к. д. о. п. г.).

∘ В случае противоречия предложить пользователю пополнить $\bar{\Gamma}_0 \cup \bar{\Delta}_0$ (БДНИ) фактами нужного вида; иначе, построить индуктивное обобщение $\Sigma_{UD}^{(m)}$ гипотез из $\bar{\Gamma}'$ и пополнить множество аксиом предметной области $\Sigma_{UD} = \Sigma_{UD} \cup \Sigma_{UD}^{(m)}$.

Хочется отметить, что на разных шагах могут применяться различные решающие предикаты I рода (получаемые добавлением дополнительных условий к "основному" предикату, реализующему прямой метод сходства Д. С. Милля), образуя тем самым некоторую *стратегию* правдоподобного рассуждения. Формальное определение стратегии как некоторого алгоритмического управления п. п. в. и упорядочивание множества стратегий позволяют использовать более тонкие оценки формул, претендующие на формализацию *рефлексии* как оценки способа получения гипотез [7].

ДСМ-СИСТЕМА ПРОГНОЗИРОВАНИЯ КОНТРПРОДУКТИВНЫХ СВОЙСТВ ХИМИЧЕСКИХ СОЕДИНЕНИЙ

Одной из задач, отвечающих условиям применимости ДСМ-метода, является задача "структура-активность", в которой требуется предсказать химические свойства (активности) соединения исходя из его структурной формулы. Природа объектов изучения (химических соединений) и существующие формальные описания их структуры в виде пространственных химических графов позволяют естественным образом определить понятие "фрагментов описания объектов" как химических подграфов, отношение вложимости как вхождение подграфа в граф и операцию сходства как нахождение общих подграфов двух графов. Выполняется также требование интерпретируемости сходств подграфов, более того, такая интерпретация соответствует теперешним представлениям химической науки о том, что химическая активность вещества определяется наличием в его структуре определённых подфрагментов — так называемых "фармакофоров" и "анти-фармакофоров" — вызывающих или, наоборот, блокирующих проявление тех или иных свойств.

Такая ДСМ-система была реализована в Отделе интеллектуальных систем ВИНИТИ и стала лауреатом международного конкурса Predictive Toxicology Challenge'2000 в трёх из четырёх категорий [1]. В этой системе для представления химических соединений применялся язык ФКСП (фрагментарный код суперпозиции подструктур) [10], специально разработанный для задачи "структура-активность". В этом языке химическое соединение представляется в виде набора символьных кодов его подструктур, являющихся центрами локализации π-электронов. Выбор такого представления основан на концепции, в соответствии с которой биологическая активность веществ определяется характером слабых связей, возникающих между активными центрами биологического рецептора и соответствующими активными центрами вещества, являющихся центрами локализации π-электронов. Это представление данных, выражающее некоторое видение проблемы специалистами в данной предметной области, является важным элементом настройки ДСМ-метода АПГ на конкретную задачу — задачу "структура—активность".

Однако большинство моделей в данной задаче являются числовыми, например, модель канцерогенности полиароматических углеводородов (ПАУ) [3]. В этой модели контрпродуктивная — канцерогенная или мутагенная — активность ПАУ будет тем выше, чем ниже энергия образования соответствующего ему метаболита, т. е. вещества, получающегося в результате превращений, которые претерпевает исходное вещество в организме. Однако в процессе эксплуатации этой модели были обнаружены соединения, выпадающие из общего ряда. Оказалось, что такие соединения имеют специфическую структуру, и качество результатов значительно повышается, если провести предварительный отбор соединений по структуре, а уж затем применять регрессионный анализ. Таким образом, возникает задача анализа гибридных — структурно-числовых — данных.

МОДУЛЬ ДСМ-АНАЛИЗА ЧИСЛОВЫХ ХАРАКТЕРИСТИК (МЕТОД ИНТЕРВАЛОВ)

Описываемый здесь подход является первым приближением (в рамках ДСМ-метода) к проблеме анализа гибридных данных и состоит в попытке

использования в анализе так называемых мульти-ДСМ-объектов, т. е. ДСМ-объектов, представляющих упорядоченный набор компонентов, каждый из которых, в свою очередь, — это ДСМ-объект.

Под ДСМ-объектом или ДСМ-фрагментом, понимается структура данных, для которой выполняется условие (2) применимости ДСМ-метода. Изложим это условие чуть более формально. Операцией сходства на множестве объектов S, представляющем некоторую предметную область, будем называть операцию $\cap$, удовлетворяющую следующим свойствам (где s_i, s_j, s_k — произвольные объекты из S):

(1) $s_i \cap s_j = s_i$ — рефлексивность,

(2) $s_i \cap s_j = s_j \cap s_i$ — симметричность,

(3) $(s_i \cap s_j) \cap s_k = s_i \cap (s_j \cap s_k)$ — ассоциативность,

(4) $s_i \cap s_0 = s_0$ — для некоторого s_0 из S, который называется пустым объектом и представляет собой "неинформативное" сходство.

Отношение вложимости '$\subset$' на множестве объектов S определим как

$$s_j \subset s_j := (s_i \cap s_j = s_i),$$

т. е. один объект вкладывается в другой, если он является их сходством. Нужно заметить, что на практике операция нахождения сходства двух объектов может быть довольно тяжеловесной, так что использовать её для проверки вложимости следует только тогда, когда нет более простого способа.

Тогда, для мульти-ДСМ-объектов вида $\{C_1, C_2, \ldots, C_k\}$, где C_i — ДСМ-объект, представляющий компонент описания объекта некоторой предметной области:

$$\{C_1, C_2, \ldots, C_k\} \subset \{D_1, D_2, \ldots, D_k\} :=$$

$$= \bigwedge_{i=1}^{k} (C_i \subset D_i),$$

$$\{C_1, C_2, \ldots, C_k\} \cap \{D_1, D_2, \ldots, D_k\} :=$$

$$= \{C_1 \cap D_1, C_2 \cap D_2, \ldots, C_k \cap D_k\},$$

при этом сходство считается пустым, если пуст хотя бы один его компонент.

Следующим шагом является формализация числовой величины как ДСМ-объекта.

Существует несколько вариантов задания операции сходства для числовых величин [9,11]. Можно разбить всю область значений числовой величины на непересекающиеся интервалы и считать, что внутри одного интервала I значения сходны. Для таких значений сходство выражается как "интервал I", а для значений из разных интервалов сходство пусто. Недостатком такого подхода является то, что предопределённое разбиение может оказаться неадекватным и в результате близкие "по смыслу" значения могут оказаться "несходными".

Избежать "произвола" при разбиении числовых значений на интервалы в рамках ДСМ-АПГ можно, например, введя следующую алгебру интервалов. Пусть $[a_i, b_i]$ — интервалы, и $A \leq a_i, b_i \leq B$, где A и B — минимальное и максимальное возможные значения числового параметра. Операция сходства определяется как $[a_i, b_i] \cap [a_j, b_j] := [\min(a_i, a_j), \max(b_j, b_j)]$, а $[A, B]$ играет роль пустого объекта. Нетрудно проверить, что эта операция обладает свойствами (1)–(4). Исходные "точечные" значения параметра x_i представляются в виде интервалов $[x_i, x_i]$. Отношение вложимости в этом случае по смыслу противоположно традиционному — при таком подходе "размер" ДСМ-фрагмента обратно пропорционален длине соответствующего интервала (так, длина "целого" объекта равна 0, а пустого — $|A - B|$) : $[a_i, b_i] \subset [a_j, b_j] := a_i \leq a_j \wedge b_j \leq b_i$.

ЭКСПЕРИМЕНТЫ

Описанный модуль анализа гибридных данных был реализован и состыкован с ДСМ-системой прогнозирования контропродуктивных свойств соединений и с полученной гибридной системой проведён ряд экспериментов [12]. В качестве числовой характеристики во всех экспериментах использовалась энергия образования конечного метаболита, рассчитанная с помощью квантово-механического модуля [3] в соответствии с описанной выше моделью канцерогенности ПАУ. С помощью этого же модуля порождалась структура собственно конечного метаболита, которая по желанию экспериментатора могла использоваться вместо структуры исходного соединения. Канцерогенная активность соединений была задана в виде индексов Бэджера (уровней активности от "неактивно" до "сверхактивно"), которая для выполнения условия (1) применимости ДСМ-метода была оттранслирована в четыре свойства:

(а) "быть (по крайней мере) слабоактивным",

(б) "быть (по крайней мере) среднеактивным",

(в) "быть (по крайней мере) сильноактивным",

(г) "быть сверхактивным".

Таким образом, каждое соединение может быть (+)-примером по одному набору свойств и (−)-примером по другому (так, среднеактивное соединение будет (+)-примером для свойств (а)—(б) и (−)-примером для свойств (в)—(г); неактивное — (−)-примером для всех свойств).

Один из экспериментов с системой проводился на массиве из 25-ти ПАУ, для которых уровень канцерогенности был установлен. Три соединения, для которых данные из разных источников расходились (табл. 1, где рассчитанные значения энергии образования метаболита приведены в условных единицах), были закрыты для последующего доопределения, остальные — образовали обучающий массив. Первым отличительным результатом применения гибридной системы явилось то, что полученная система гипотез объясняла все обучающие примеры (т. е. выполнялся критерий достаточности оснований принятия этих гипотез), чего не удавалось достичь, рассматривая только структуру соединений. Далее, если сравнить результаты прогноза трех тестовых соединений "традиционной" и гибридной системы (табл. 2 и 3), то видно, что учёт числовой характеристики соединений сделал прогноз более полным по числу доопределённых свойств и, таким образом, более точным по описанию свойств соединения в целом.

Таблица 1

Результаты экспериментальной оценки канцерогенной активности тестовых соединений и значения рассчитанной числовой характеристики ΔE

Соединение	1	2	3	4	ΔE
Бензо(е)пирен	+/−	−	−	−	0,826
Дибенз(а,с)антрацен	+/−	−	−	−	0,833
Дибензо(а,е)пирен	+	+	+/−	+/−	0,846

Таблица 2

Результат ДСМ-прогноза без учета числовой характеристики

Соединение	1	2	3	4
Бензо(е)пирен	+	?	?	?
Дибенз(а,с)антрацен	+	?	?	?
Дибензо(а,е)пирен	+	+	?	?

Таблица 3

Результат ДСМ-прогноза с учётом числовой характеристики

Соединение	1	2	3	4
Бензо(е)пирен	?	–	–	–
Дибенз(а,с)антрацен	+	?	–	–
Дибензо(а,е)пирен	+	+	–	–

Рассмотрим подробнее процесс получения прогноза для соединения № 18 — дибензо(а,е)пирена. На этапе порождения гипотез о причинах наличия/отсутствия канцерогенной активности (этап обучения), среди прочих, были получены две следующие. Первая из них (рис. 1), полученная совместным рассмотрением соединений №№ 16, 20 и 22 (дибенз(а,h)антрацен, бензо(а)пирен и трибензо(а,е,i)пирен), интерпретируется так: наличие в соединении данного структурного фрагмента, если энергия образования его конечного метаболита лежит внутри данного интервала, является причиной того, что это соединение проявляет среднюю активность (по крайней мере), т. е. обладает свойствами 1 и 2. Вторая гипотеза (рис. 2), полученная на основании соединений №№ 16 и 17 (дибенз(а,h)антрацен и бенз(а)антрацен), предполагает, что наличие указанного структурного фрагмента при условии, что энергия образования его конечного метаболита лежит внутри указанного интервала, блокирует сильную и сверхсильную активности (т. е. такое соединение не обладает свойствами (в) и (г)). Далее, на этапе порождения гипотез о наличии/отсутствии активности у тестовых соединений (или этапе прогнозирования) на основании этих двух гипотез была выдвинута гипотеза о том, что соединение № 18 является в точности среднеактивным, т. е. обладает свойствами (а) и (б) и не обладает свойствами (в) и (г) (рис. 3 и 4).

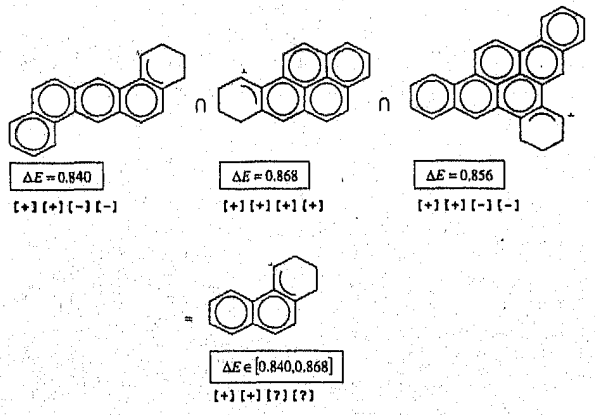

Рис. 1. Порождение первой гипотезы о причинах наличия/отсутствия свойств

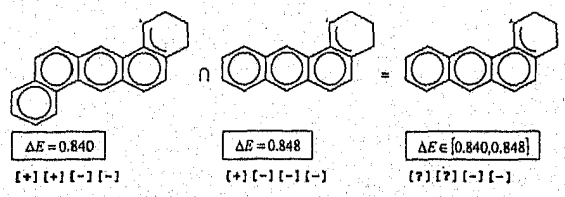

Рис. 2. Порождение второй гипотезы о причинах наличия/отсутствия свойств

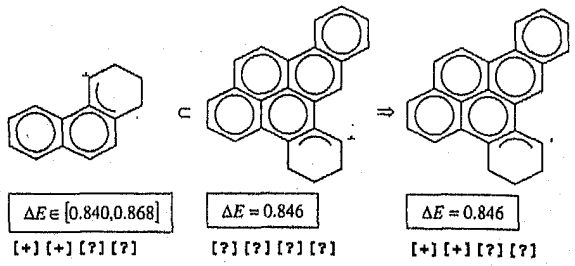

Рис. 3. Порождение первой гипотезы о наличии/отсутствии свойств

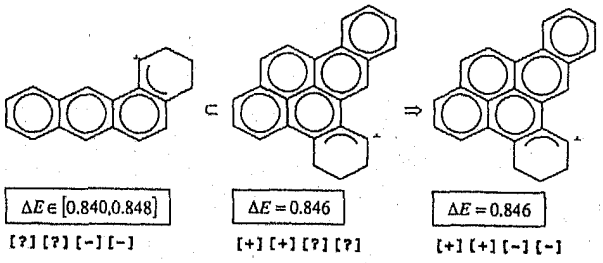

Рис. 4. Порождение второй гипотезы о наличии/отсутствии свойств

Другой эксперимент был проведён на массиве из 63-х ПАУ, приведённом в [13]. Сначала были правильно доопределены два соединения, исключённые из обучающей выборки, а затем была сделана попытка доопределить свойства 14-ти ещё не изученных соединений. Для девяти из них мы получили прогнозы, совпадающие с прогнозами, сделанными с помощью группы правил, предложенной экспертами в [13].

ЗАКЛЮЧЕНИЕ

Предложенное расширение ДСМ-системы прогнозирования контропродуктивных свойств химических соединений позволило значительно улучшить качество получаемых прогнозов, что подтверждает перспективность данного направления. Среди недостатков этого подхода к анализу гибридных данных в рамках ДСМ-метода АПГ следует отметить значительное увеличение числа порождаемых гипотез, связанное с тем, что фактически сходство двух числовых ДСМ-объектов, не бывает пусто (для разрешения этой проблемы, по-видимому, нужно определять "пустой" интервал в терминах его длины). Затем, такой подход, в котором переопределяется только локальное сходство (т. е. сходство двух объектов), не позволяет использовать статистические методы и, таким образом, не вполне отвечает поставленной задаче. Переопределение операции нахождения глобального сходства (т. е. сходства более чем двух объектов) требует внесения существенных изменений в существующую реализацию ДСМ-метода, что и определяет направление развития этой комбинированной системы.

СПИСОК ЛИТЕРАТУРЫ

1. Максин М. В. Интеллектуальный анализ данных в науках о жизни // НТИ. Сер. 2.— 2000.— № 9.— С. 16–27.

2. 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.

3. Максин М. В., Харчевникова Н. В. Квантово-механический модуль системы, реализующей комбинаторно-численный подход к проблеме прогнозирования свойств химических соединений // НТИ. Сер. 2.— 2002.— № 6.— С. 25–31.

4. Милль Д. С. Система логики силлогистической и индуктивной.— М.: Книжное дело, 1900.— 781 с.

5. Аншаков О. М. Об одной интерпретации ДСМ-метода автоматического порождения гипотез // НТИ. Сер. 2.— 1999.— № 1–2.— С. 45–53.

6. Финн В. К. Синтез познавательных процедур и проблема индукции // НТИ. Сер. 2.— 1999.— № 1–2.— С. 8–45.

7. Аншаков О. М., Скворцов Д. П., Финн В. К. Логические средства экспертных систем типа ДСМ // Семиотика и информатика.— 1986.— Вып. 28.— С. 65–101.

8. Финн В. К. Правдоподобные выводы и проблемы автоматического порождения теорий из базы фактов // Интенсиональные логики и логическая структура теорий: Тез докл. IV советско-финского коллоквиума по логике, Телави, 1985.— М., 1985.— С. 108–114.

9. Кузнецов С. О. ДСМ-метод как система автоматического обучения // Итоги науки и техники. Сер. Информатика.— М.: ВИНИТИ, 1991.— Вып. 15.

10. Blinova V. G., Dobrynin D. A. Languages for Representing Chemical Compounds for Intelligent Systems of Chemical Design // Automated Documentation and Mathematical Linguistics.— 2000.— № 3.

11. Маневич С. И., Харчевникова Н. В., Дьячков П. Н. Прогнозирование контрпродуктивных свойств химических соединений при комбинированном использовании структурных формул и численных энергетических параметров // НТИ. Сер. 2.— 2000.— № 5.

12. Максин М. В., Харчевникова Н. В., Блинова В. Г., Добрынин Д. А., Жолдакова З. И. Прогноз канцерогенности полициклических ароматических углеводородов с использованием квантово-химического модуля генерации метаболитов интеллектуальной ДСМ-системы // НТИ. Сер. 2.— 2003.— № 11.— С. 12–17.

13. Flesher J. W., Horn J., Lehner A. F. Molecular modeling of carcinogenic potential in polycyclic hydrocarbons // J. Molec. Struct. (Theochem).— 1996.— Vol. 362.— P. 29–49.

Материал поступил в редакцию 22.06.04.

УДК 004.738.5:[002.2:001.32]

Л. И. Госина, Н. С. Солошенко

Представление малотиражной издательской научной продукции в Интернете как отражение ее доступности для научного и библиотечного сообществ

В последние годы при практически полном отсутствии традиционных тематических планов и значительной неполноте обязательного экземпляра, признаваемой издателями и библиотеками, в "серую" литературу попадают не только малотиражные издания научных и учебных центров, труды научных мероприятий, но и научные монографии достаточно крупных издательств. При соблюдении законодательства и наличии правильно организованной системы информирования в Интернете такого рода издания могут стать доступными для научного и библиотечного сообществ. В предлагаемом исследовании рассматривается текущее состояние некоторых элементов этой системы: наличия издательской информации о малотиражной научной литературе; отражения информации об этих изданиях в справочно-поисковых аппаратах федеральных библиотечных центров и академических библиотек; доступность первоисточников. Анализ состояния проблемы рассматривается на примере малотиражных математических изданий.

По данным Российской книжной палаты, в 2003 г. около трети всего массива изданий (31,2% названий) имели тираж менее 500 экз., а 7,4% имели тираж до 1000 экз. [1]. Таким образом, в масштабах всей страны около 40% книжной продукции (в названиях) является малотиражной. Анализ продукции "Физматлита", долгое время входившего в состав крупнейшего академического издательства "Наука", показал, что в 2002 г. примерно половину его книжного потока составляли научные монографии. Тираж абсолютного большинства научных монографий не превышал 300–400 экз. Можно утверждать, что существует достаточно выраженная тенденция к превращению всей научной издательской продукции в малотиражную. Сопоставив тираж изданий с количеством научных библиотек