ЛОГИЧЕСКИЕ МЕТОДЫ В ИНФОРМАТИКЕ
Вычислительные системы
1993 год Выпуск 148

УДК 53.02+519.7+519.812.2

ВВЕДЕНИЕ В ТЕОРИЮ ОТКРЫТИЙ. ПРОГРАММНАЯ СИСТЕМА DISCOVERY*)

Е.Е.Витяев, А.А.Москвитин

В в е д е н и е

1. В настоящее время в Анализе Данных и Искусственном Интеллекте разработано довольно много различных методов обработки данных. Однако методология этих методов практически не разработана. Давно назрела потребность проанализировать эти методы с точки зрения их связи с процессом познания. Такому анализу и посвящено введение. Более подробно отдельные вопросы изложены в работах [1-11]. В результате анализа мы естественным образом придем к некоторому подходу к Теории Открытий, определенному ниже, разработанному в этих же работах, и программной системе DISCOVERY, реализующей эту вполне определенную часть процесса познания. В процессе анализа основное внимание будет уделяться методам анализа зависимостей.

Из методов Анализа Данных к методам анализа зависимостей относятся методы регрессионного и дискриминантного анализов, методы распознавания образов и обнаружения закономерностей,методы классификации и аппроксимации. Во всех этих методах вид зависимости задается внешним, априорным по отношению к самой зависимости образом. В регрессионном анализе это линейная или нелинейная регрессия, в дискриминантном анализе - дискриминант-

*) Работа выполнена при финансовой поддержке Российского Фонда Фундаментальных исследований (93-011-1506).

ная функция, в распознавании образов - решающее правило, методах классификации - форма кластеров. Какова "истинная" зависимость - такой вопрос не ставится, да и не может быть поставлен в рамках Анализа Данных. В Анализе Данных неизвестная зависимость аппроксимируется некоторыми заданными априори классами функций, моделями, решающими правилами и т.д. Аппроксимируя неизвестную зависимость с требуемой степенью точности и надежности, методы Анализа Данных решают по существу задачу предсказания. Найденная аппроксимация ничего (или почти ничего) [1, § 2] не говорит о том, какова "истинная" зависимость.

Процесс аппроксимации неизвестных зависимостей начинается с переноса способов измерения из точных наук и прежде всего из физики в другие области. Рассмотрим, например, такую физичес - кую величину, как температура [1, § 2]. Шкалы температуры в нефизических областях, например, при измерении температуры тела больного в медицине, температуры почвы в сельском хозяйстве, температуры воздуха в духовке в кулинарии и т.д., должны быть разные, хотя измеряться они могут одним и тем же прибором - термометром. Далеко не всеми понимается тот факт, что шкала - это не только риски делений шкалы на приборе, а это прежде всего тот набор операций и отношений, которые имеет смысл производить с числовыми значениями величин с точки зрения рассматриваемой предметной области, точнее, это те операции и отношения, которые интерпретируемы в системе понятий соответствующей предметной области. Некоторые возражают, что термометр не может измерять ничего, кроме температуры. Он действительно во всех случаях измеряет физическую температуру. Но резонно в таком случае спросить, а зачем собственно мы измеряем температуру? Ведь не затем, чтобы согласно законам физики узнать, сколько в больном содержится тепла и сколько он в состоянии растопить льда, если его положить на лед, и не затем, чтобы определить среднюю

кинетическую энергию молекул почвы или курицы в духовке. Температура, как и любой другой прибор, нужна для получения выводов в системе понятий той предметной области, к которой она относится. Для больного "температурный фактор служит наиболее общим и универсальным регулятором скорости химических реакций и активности ферментов, с повышением температуры в известной мере ускоряются и обменные процессы" [12]. Для почв температура должна интерпретироваться в системе понятий физиологии растений и деятельности микроорганизмов и т.д. Следует понимать, что физическая величина температуры является косвенным измерением некоторой другой величины, интерпретируемой в системе понятий предметной области, которую мы именно и хотим измерить. Физическая температура больного, например, есть косвенное измерение медицинской величины - уровня обмена веществ, температура почвы измеряет состояние биохимических процессов в растениях и микроорганизмах, температура воздуха в духовке измеряет течение процесса свертывания белка и т.д. Какие отношения и операции над числовыми значениями температуры имеют смысл для всех этих величин, определяется уже этими интерпретациями. Поэтому числовые значения величин нельзя слепо переносить из одной области знаний в другую. После такого переноса необходимо заново определить шкалу. Например, для температуры больного интерпретируемы выделенные значения 36.7 и 42.0 и отношение линейного порядка $\leq$, поэтому это будет шкала порядка с выделенными значениями.

На следующем шаге применения методов Анализа Данных также проявляется их аппроксимационный характер. Перед обработкой данные, как правило, преобразуются к одному из известных видов - количественным или качественным. Если они преобразуются к количественным данным (т.е. с числами разрешается производить любые математические операции, вне зависимости от их интерпретации), то в них вносится бессмысленная информация (связанная с

произволом в выборе числового представления), и как следствие проявляющаяся в том, что невозможно обоснованно проинтерпретировать полученные результаты. Если данные преобразуются в количественные за счет использования различного рода (числовых) моделей или дополнительных предположений, которые в этом случае не полностью интерпретируемы, то это также приводит к невозможности обоснованно проинтерпретировать полученные результаты. Если данные преобразуются в дискретные, то это ведет к потере информации. Поэтому неизвестные зависимости не просто аппроксимируются задаваемыми априори видами зависимостей, но и сами данные часто искажаются, чтобы их обработка этими методами была возможной.

Для того чтобы детальнее разобраться с такими понятиями, как числовые значения величин, их интерпретируемость, осмысленность математических операций с величинами, "истинная" зависимость и т.д., необходимо обратиться к Теории Измерений [13-16]. Теория Измерений основана на известном принципе: свойства определяются отношениями. Из Теории Измерений следует, что числовые значения величин и функциональные выражения для законов являются лишь удобным и математически хорошо разработанным способом числового кодирования элементов эмпирических систем.Число, например 5,само по себе смысла не имеет, оно приобретает смысл лишь при его интерпретации в некоторой эмпирической системе,например,если мы говорим: 5 метров, 5 баллов, 5 деталей и т.д.Интерпретация чисел,в частности,определяет,какие математические действия с ними можно осмысленно проводить,чтобы не получить бессмысленных результатов типа 1.5 дровосека,1 м + 1 кг и т.д. Эмпирическая система - это множество (идеализированных) объектов с заданным на нем множеством интерпретируемых в системе понятий отношений и операций, удовлетворяющих некоторой системе аксиом. Такой семантический уровень рассмотрения с необходимостью возникает из того факта, что интерпретирует че-

ловек всегда качественно. Поэтому, интерпретируя количествен - ные значения величин, модели, функции и т.д.,он интерпретирует их качественно - в системе понятий предметной области - и в качестве промежуточной стадии такой интерпретации - на семан - тическом уровне - в (многосортной) эмпирической системе. Семантический уровень возникает не только из-за требования интерпретируемости, но он и исторически является первичным и представляет собой целостное (модельное) представление той исходной деятельности над объектами, которая привела в свое время к возникновению теорий и чисел [13-16].

Если в Анализе Данных зависимости аппроксимируются, то в Теории Измерений определяются в определенном смысле "истинные" величины и зависимости.

Числовые представления величин, получаемые из систем ак - сиом Теории Измерений, дают "истинные" шкалы величин, интер - претируемые в системе понятий соответствующей предметной области. "Истинные" шкалы это те, которые интерпретируются в системе понятий предметной области и являются числовыми кодами значений величины соответствующей эмпирической системы.

Числовые представления законов, получаемые на основании какой-либо системы аксиом, являются "истинными" законами данной предметной области в том смысле, что они, во-первых,интерпретируемы в системе понятий данной предметной области и являются числовыми кодами взаимосвязи величин из эмпирической системы,и, во-вторых, их числовые представления получаются либо одновременно (единой процедурой шкалирования) с числовыми представлениями величин, либо согласованы с числовыми представле - ниями величин. В [13-16] показано, что уравнения для физичес - ких законов просты только потому, что они получаются процеду - рой одновременного шкалирования всех, входящих в закон величин так,чтобы взаимосвязь этих величин выражалась заданной, опре - деляемой системой аксиом, простой функциональной зависимостью.

Следующий вывод, который следует из Теории Измерений, состоит в том, что числовой способ представления данных, используемый в Анализе Данных, не является адекватным для целей обнаружения законов и зависимостей, хотя он вполне может быть адекватен для целей предсказания, где вид зависимости для нас не важен. Можно хорошо предсказывать, используя только двоичную кодировку данных. Цель анализа зависимостей совсем другая - познать предметную область. Для достижения этой цели интерпретируемость данных и результатов анализа данных в системе понятий предметной области является необходимым условием получения хоть какого-нибудь полезного результата, вносящего вклад в теорию предметной области. Так как числа сами по себе смысла не имеют, то интерпретируемость данных и результатов счета означает, согласно Теории Измерений, их интерпретируемость на семантическом уровне в системе понятий предметной области без использования чисел. Поэтому для целей анализа зависимостей необходим тот способ представления данных, который принят в Теории Измерений - в виде (многосортных) эмпирических систем. Системы аксиом, которым удовлетворяют эти эмпирические системы, представляют собой логическую эмпирическую теорию предметной области. Системы аксиом как логические высказывания очевидно интерпретируемы в системе понятий предметной области. Поэтому анализ зависимостей должен состоять в обнаружении зависимостей на данных, представленных (многосортными) эмпирическими системами. Каков при этом возможен вид закономерностей? Богатство языка логики первой ступени, а также тот факт, что его достаточно для Теории Измерений показывают, что этот язык вполне приемлем для выражения зависимостей. Таким образом, задача анализа зависимостей сводится к задаче усиления (в логическом смысле) логической эмпирической теории за счет обнаружения зависимостей в логике первого порядка. Числовые представления величин и функциональные зависимости для законов должны получаться из

обнаруженных систем аксиом в результате применения Теории Измерений. Полученные шкалы величин и законы, связывающие величины, дают Количественную Теорию Предметной Области. Для физики этот переход продемонстрирован в [14](см.также [4,8]). В [14] показано,как можно строить Количественную Теорию Предметной Области - систему величин, связанных между собой (фундаментальными) законами.

Таким образом, задача анализа зависимостей разбивается на два этапа: сначала надо построить Логическую Эмпирическую Теорию, а затем, применяя Теорию Измерений, построить Количественную Теорию Предметной Области. Такое разбиение отражает естественный процесс перехода теории из качественного состояния, представленного логической эмпирической теорией, в количественное. Теория Измерений и является теорией такого перехода. Для физики, например, этот процесс протекал достаточно долго.

Из всего сказанного следует, что Теория Измерений является той теорией, которую надо использовать для разработки методов обнаружения законов и шкал. Однако точность аксиоматического анализа законов, проведенная в Теории Измерений, оборачивается сложностью применяемого ею аппарата, недостаточным разнообразием систем аксиом, чтобы можно было браться искать любые зависимости, отсутствием общего метода обнаружения систем аксиом в данных и другими проблемами. Прежде чем получить Теорию Открытий, основанную на Теории Измерений, необходимо заполнить эти пробелы или, по крайней мере, указать пути, по которым эти пробелы могут быть заполнены. Для получения такой Теории Открытий необходимо было:

1) разработать достаточно общий метод обнаружения систем аксиом на различных данных [2] ;

2) найти обобщение Теории Измерений,позволяющее получать числовые представления практически для любых систем аксиом,которые может обнаружить этот метод [9-10] ;

3) найти путь к построению систематик возможных законов природы [8].

Этим вопросам и посвящена серия работ [1-11], которая приводит в результате к логико-статистическому (и в этом смысле объективному, в отличие от подходов к Теории Открытий в Искусственном Интеллекте, см. п.2) подходу к Теории Открытий, основанному на Теории Измерений, и программной системе обнаружения систем аксиом DISCOVERY.

Что можно сказать о возможностях построения Количественных Теорий такой Объективной Теорией Открытий в других областях знания? Возможность построения Количественных Теорий в психологии, психофизике, социологии, психолингвистике, принятии решений, этике и т.д. подтверждают результаты, полученные в функциональной теории измерений [17,18]. Хотя методы функциональной теории измерений менее обоснованы и требуют более сильных и менее очевидных предположений, чем методы Теории Измерений (сравнение этих двух подходов обсуждено в [14]), тем не менее, благодаря своей простоте, они позволили получить довольно много интересных результатов в этих областях. Только результаты, полученные в психологии, позволили их авторам заметить, что "мышление человека, кажется, подчиняется довольно общей когнитивной алгебре".

2. Сравним предлагаемый подход с существующими подходами к Теории Открытий в Искусственном Интеллекте (см.обзор [19]). Цель этих подходов формулируется следующим образом: "... можно рассматривать программные системы открытий Искусственного Интеллекта как вычислительные модели исторического процесса открытий". Моделировать исторический процесс открытия законов, не зная что такое Закон, вполне соответствует духу Искусственного Интеллекта (Ньюелл и Саймон обосновали полезность достаточных моделей поведения (см.[19]). Но нами предлагается другой подход: не моделировать историю открытий, а выяснить, как следует

с современной точки зрения обнаруживать законы. Понятие Закона еще исследуется и в самой физике, и в Теории Измерений, и в Теории Физических Структур [20-22] и в других работах (см., например, [4,8,23]), и эти исследования еще не закончены. Наш подход опирается на Теорию Измерений и Теорию Физических Структур. С точки зрения этих теорий, существующие подходы к Теории Открытий являются аппроксимационными - вид закона должен быть определен априори. Из Теории Измерений также следует, что, научившись моделировать процесс открытия законов на исторических физических примерах, нельзя применить их в нефизических областях. Для самой же физики эти методы не нужны. Как утверждается в [19], они, возможно, будут полезны для истории открытий: "В той степени, в какой шаги системы будут теми же, что и были сделаны исторически, можно будет сделать вывод, что эта система представляет собой подходящую модель исторических открытий".

3. Метод обнаружения "полного" множества интерпретируемых зависимостей (систем аксиом) в данных - одна из первых задач, которую необходимо было решить. Первый вариант этого метода практически апробирован и опубликован в [2]. Приведем основную нить рассуждений, аргументы и ссылки, приводящие к построению метода и программной системы DISCOVERY.

Из Теории Измерений следует, что эмпирическая теория предметной области может быть представлена многосортной эмпирической системой $\mathcal{M} = \langle U, \{A_i\}_{i\in I}; V, \Omega \rangle$ и системой аксиом $S^{V \cup \Omega}$, истинной на $\mathcal{M}$, где U - генеральная совокупность объектов, $\{A_i\}_{i\in I}$ - множества объектов различных типов I (например, множество значений некоторой величины; наборы величин, связанных некоторой зависимостью; множество действительных чисел и т.д.), $V = \langle P_1, \ldots, P_k, \rho_1, \ldots, \rho_l \rangle$ - множество эмпирических отношений и операций, Ω - множество идеализированных

отношений и операций. Отношения и операции словаря V эмпирической системы $\mathcal{M}$ эмпирически интерпретируемы, т.е. представляют собой некоторые измерительные процедуры (в том числе анкетирование, тестирование, экспертные оценки и т.д.) интерпретируемые (вместе с результатами) в системе понятий предметной области. Эмпирическая система $\mathcal{M}$ дает нам интерпретируемое семантическое представление предметной области, а система аксиом S^V - аксиоматическое описание предметной области.

В Теории Измерений существуют два уровня рассмотрения - эмпирический и теоретический. На эмпирическом уровне рассматриваются реальные множества объектов $A \subset U$ (выборки из U), данные, представляемые эмпирическими системами $\mathcal{D} = \langle A, V \rangle$, и системы аксиом S^V реальных приборов из V, а на теоретическом уровне - идеализированные эмпирические системы Теории Измерений $\mathcal{M} = \langle A, \Omega \rangle$ и системы аксиом S^Ω для идеализированных приборов из Ω и теоретические результаты о существовании, единственности и адекватности числовых представлений. Оба этих уровня вместе с взаимосвязью между ними с помощью правил соответствия [24], а также одновременное представление эмпирической системы, числовой системы и сильного гомоморфизма эмпирической системы в числовую систему могут быть представлены в многосортной эмпирической системе $\mathcal{M}$ и системе аксиом $S^{V \cup \Omega}$ за счет совместного представления реальных, идеализированных и числовых сортов. Многосортную эмпирическую систему $\mathcal{M}$ вместе с системой аксиом $S^{V \cup \Omega}$ будем называть эмпирической аксиоматической теорией. Заметим, что в эмпирической аксиоматической теории числа играют вспомогательную роль и над ними можно производить только те математические действия, которые при обратном отображении относительно сильного гомоморфизма, преобразуются в определимые в словаре V отношения и операции.

В [1, §4] приводится методика извлечения из наиболее распространенных типов данных - парных сравнений, множественных

сравнений, матричного представления бинарных отношений, матриц упорядочений и близости, матриц объект-признак - всей интерпретируемой в них информации и представление ее в эмпирических аксиоматических теориях. Извлеченная информация представляется в виде эмпирических систем и систем аксиом, взятых как на эмпирическом,так и на теоретическом уровнях.

Как было сказано, без ограничения общности можно считать, что анализ зависимостей на данных, представленных эмпирическими аксиоматическими теориями, сводится к обнаружению зависимостей на эмпирической системе $\mathcal{M} = \langle U, V \rangle$ в языке первого порядка в словаре V. Такой анализ по существу представляет собой аксиоматический анализ предметной области.

Зависимости в виде утверждений в языке первого порядка могут быть детерминированными (истинными на $\mathcal{M}$) и недетерминированными. Детерминированные зависимости являются системами аксиом. Таким образом, "полный" анализ детерминированных зависимостей на $\mathcal{M}$ сводится к определению полного множества $S^V_{\mathcal{M}}$ истинных на $\mathcal{M}$ формул в языке первого порядка.

Рассмотрим множество $S^V_{\mathcal{M}}$. В [1, § 5] аргументируется, что практически достаточно ограничиться рассмотрением зависи - мостей в виде универсальных формул (формул, содержащих только кванторы всеобщности). Там же находится эмпирически интерпре - тируемое свойство измерительных процедур, из которого следует универсальная аксиоматизируемость экспериментальной зависимости $S^V_{\mathcal{M}}$. Таким образом, если это свойство выполнено для измерительных процедур из V (а оно является достаточно слабым и, как показано в [1,§5], практически не ограничивает "пол - ноту" анализа зависимостей), то задача "полного" анализа детерминированных зависимостей сводится к определению множества из $US^V(\mathcal{M})$ всех универсальных формул языка первого порядка, истинных на $\mathcal{M}$.

Известно, что любая конечная совокупность универсальных формул логически эквивалентна совокупности формул вида:

$$\forall x_1,\ldots,x_m(A_1 \,\&\, \ldots \,\&\, A_n \Rightarrow A_o), \qquad (1)$$

где $A_o, A_1, \ldots, A_n$ - атомарные формулы вида $P^{\varepsilon}(t_1,\ldots,t_k)$, $(t = g)^{\varepsilon}$; $t, g, t_1, \ldots, t_k$ - термы; $x_1, \ldots, x_m$ - переменные; $\varepsilon = 1(0)$ - отсутствие (наличие) отрицания. Этот результат сводит задачу "полного" анализа детерминированных зависимостей на $\mathcal{M}$ к задаче нахождения множества $R^{V}[u1](\mathcal{M})$ всех формул вида (1), истинных на $\mathcal{M}$.

Заметим, что формулы (1) "выделяют" из универсальных формул всю их способность к предсказанию - возможность предсказывать по условию формулы (1) заключение A_o. Такое выделение необходимо, так как зависимости нас интересуют ровно в той мере, в какой они интерпретируемы, поддаются экспериментальной проверке и способны предсказывать. Для осуществления предсказаний более удобна форма записи формул (1) через индивидные константы. Вместо кванторов всеобщности и связанных ими переменных введем константы $z_1, z_2, \ldots$. Тогда формулы вида (1) преобразуются в формулы вида:

$$A_1 \,\&\, \ldots \,\&\, A_n \Rightarrow A_o, \qquad (2)$$

где $A_o, A_1, \ldots, A_n$ - атомарные формулы вида $P^{\varepsilon}(t_1,\ldots,t_k), (t=g)^{\varepsilon}$; $t(z_1,\ldots,z_1), g(z_1,\ldots,z_1), t_1(z^1_1,\ldots,z^1_{11}),\ldots,t_k(z^k_1,\ldots,z^k_{1k})$ - термы; $z_1,\ldots,z_1$, $z^1_1,\ldots,z^1_{11}$, $z^k_1,\ldots,z^k_{1k}$ - индивидные константы; $\varepsilon = 1(0)$ определяет отсутствие (наличие) отрицания. Смысл формулы (2) состоит в том, что при любой замене индивидных констант на объекты некоторой модели ($\mathcal{M}$ или $\mathcal{D}$) , из истинности посылки следует истинность заключения. Множество формул $R^{V}[u1](\mathcal{M})$ после такого преобразования переходит в неко-

торое множество формул вида (2), истинных на $\mathcal{M}$, которое обозначим через $RC^V(\mathcal{M})$.

Для обнаружения недетерминированных зависимостей в языке первого порядка необходимо ввести вероятностную меру на формулах языка первого порядка. Такая мера μ может быть введена различными способами [25].

Реально мы никогда не имеем эмпирической системы $\mathcal{M}$ всей генеральной совокупности, а только лишь некоторую выборку (данные) $\mathcal{D} = \langle A,V \rangle$, $A \subset U$ из U, представляющую собой подмодель многосортной эмпирической системы $\mathcal{M}$. Искать зависимости можно только по данным $\mathcal{D}$. Задача "полного" анализа зависимостей с учетом данного ограничения преобразуется в задачу "полного" анализа зависимостей на данных $\mathcal{D}$. Взаимосвязь между этими двумя задачами устанавливается следующими двумя требованиями:

1. Нас интересуют такие зависимости на $\mathcal{D}$ (как детерминированные, так и недерминированные), которые обладают предсказательной способностью по отношению к другим случайно выбранным из U объектам.

2. "Полное" множество зависимостей на $\mathcal{D}$ (детерминированных и недетерминированных) должно в пределе (при увеличении выборки) стремиться к "полному" множеству зависимостей на $\mathcal{M}$.

Нетрудно видеть, что $RC^V(\mathcal{M}) \subset RC^V(\mathcal{D})$, так как $\mathcal{D} \subset \mathcal{M}$. Поэтому детерминированные зависимости могут удовлетворять условиям 1 и 2, если из множества $RC^V(\mathcal{D})$ взять только те зависимости, которые в соответствии с определенным статистическим критерием, при некотором доверительном уровне α, давали бы требуемую надежность предсказания (оценку условной вероятности формулы (2)). Обозначим это множество через $RCP^V_\alpha(\mathcal{D})$.

Рассмотрим недетерминированные зависимости и множество $\mathcal{A}^V = \{A_j\}_{j \in J}$ всех атомарных высказываний в словаре $V \cup C$,

где $C = \{c_k\}_{k \in K}$ - множество всех индивидных констант. Задачу "полного" анализа недетерминированных зависимостей на $\mathcal{M}$ определим как задачу нахождения всех наиболее сильных (дающих максимальную оценку условной вероятности) условных зависимостей между атомарными высказываниями из $\mathcal{A}$. Нетрудно видеть, что такие зависимости должны удовлетворять следующим условиям:

а) $\mu(A_1,\ldots,A_n) > 0$ (в противном случае условная вероятность не определена);

б) если из посылки $A_1,\ldots,A_n$ удалить одно или несколько атомарных высказываний, то условная вероятность $\mu(A_o/A_1,\ldots,A_n)$ уменьшится (т.е. все атомарные высказывания существенны (повышают условную вероятность) для предсказания A_o).

Зависимости, удовлетворяющие условиям "а","б", называются в [26-29] вероятностными закономерностями. Множество вероятностных закономерностей обозначим через REG[ularity] ($\mathcal{M}$). Определение вероятностной закономерности относится ко всей генеральной совокупности $\mathcal{M}$ (в предположении, что нам известна вероятность μ). В [26-29] устанавливается связь между множеством детерминированных закономерностей $RC^V(\mathcal{M})$ и множеством вероятностных закономерностей REG($\mathcal{M}$). Доказывается, что детерминированные закономерности, удовлетворяющие некоторым естественным дополнительным условиям, которые логически не сужают множества $RC^V(\mathcal{M})$ (посылка не всегда ложна;при удалении какого-либо атомарного высказывания из посылки закономерность становится ложной; при замене заключения на ложь закономерность также становится ложной), являются частным случаем вероятностных закономерностей. Таким образом, множество REG($\mathcal{M}$) является в этом смысле расширением множества $RC^V(\mathcal{M})$.

Для обнаружения вероятностных закономерностей по выборке $\mathcal{D}$ условия "а","б" проверяются для формул (2) с помощью определенных статистических критериев. Формулы вида (2), для ко-

торых эти условия с некоторым уровнем доверия α выполнены, назовем закономерностями. Множество закономерностей, имеющихся на $\mathfrak{D}$, обозначим через $RG^{V}_{\alpha}(\mathfrak{D})$. В силу того, что детерминированные закономерности являются в определенном смысле частным случаем вероятностных закономерностей, обнаружение множества детерминированных закономерностей $RCP^{V}_{\alpha}(\mathfrak{D})$ происходит автоматически при обнаружении $RG^{V}_{\alpha}(\mathfrak{D})$. Во множестве закономерностей $RG^{V}_{\alpha}(\mathfrak{D})$ детерминированные закономерности выделяются тем, что они не имеют исключений (при истинности посылки заключение всегда истинно).

Таким образом, задача "полного" анализа зависимостей сведена к задаче обнаружения множества закономерностей $RG^{V}_{\alpha}(\mathfrak{D})$. Эта задача решается методом обнаружения закономерностей, который приводится в [2].

Метод обнаружения закономерностей реализован программной системой DISCOVERY.

Пользователь в диалоге с программной системой задает некоторое параметрическое семейство формул:

$$A_1 \ \& \ldots \ \& \ A_n \Rightarrow A_o, \qquad (3)$$

где $A_o, A_1, \ldots, A_n$ - логические выражения (логические выражения включают логические связки AND, OR, NOT, скобки и произвольные арифметические выражения с параметрами). Параметры могут быть произвольными и изменяться в цикле. Параметрами могут быть номера признаков, интервалы изменений признаков, выделенные значения признаков, параметры, модифицирующие признак (подвергающие его различным преобразованиям), и т.д. Для каждого набора значений параметров мы получаем конкретную формулу вида (3), которая проверяется на закономерность на многосортной эмпирической системе $\mathfrak{D}$.

§ 1. Стратегии поиска "полного" множества закономерностей

Цель каждой стратегии - получить по возможности "полное" множество зависимостей в данных. Полный перебор всех правил вида (3), как правило, невозможен (только для отдельных типов данных, например булевых или наименований, он, в принципе, возможен, и то, если ограничить его требуемой надежностью предсказания). Поэтому необходим направленный перебор. Направленный перебор должен преследовать следующие цели:

1) не пропустить статистически значимые закономерности;

2) среди статистически значимых закономерностей найти наиболее точные в смысле максимальности полноты условий в посылке и обеспечивающие максимальную оценку предсказания (см. ниже определение отношения $\supset$ - "быть более точным правилом");

3) среди закономерностей, удовлетворяющих первым двум условиям, т.е. статистически значимых и наиболее точных, найти закономерности, использующие наиболее тонкие свойства шкалы и соответствующие им отношения и операции, которые с максимальной точностью уже содержательной, учитывающей интерпретацию и содержательную взаимосвязь отношений и операций, выражают зависимость в данных. Как правило, такие закономерности более точны и в смысле п.2 (имеют большую оценку условной вероятности), хотя могут быть и несравнимы по отношению $\supset$.

Программная система DISCOVERY позволяет реализовывать направленный перебор с помощью стратегий направленного и все более точного анализа эмпирического содержания данных. При этом точность в соответствии с пп. 2 и 3 понимается в двух смыслах: в смысле отношения $\supset$ и в смысле богатства информации. Шкалы величин упорядочены в соответствии с богатством информации, содержащейся в значениях величин - от шкал наименований и шкал порядка к шкалам интервалов, отношений и абсолютным шкалам. В программной системе уточнения правил осуществляются следующим образом:

4) путем добавления новых условий в посылку либо применением подстановок. Это могут делать сами исследователь или эксперт, опираясь на свою интуицию и определение отношения ⊐ - "быть более точным правилом", введенного и исследованного в работах [28,29] и приведенного в HELP программной системы;

5) такое уточнение, кроме того, как показано в работах [28, 29], будет представлять собой некоторый вероятностный вывод, определяющий поиск наиболее точного (имеющего максимальную оценку условной вероятности) правила;

6) путем использования более "тонких" свойств величин, представленных соответствующими операциями и отношениями более сильных шкал.

Уточнения 4-6 позволяют удовлетворить соответствующим целям 2,3. Цель 1 - не пропустить статистически значимой закономерности - в практически достаточной мере достигается тем, что для реальных задач, как правило, всегда выполняется условие: если есть закономерность со сложной посылкой, то такую закономерность всегда можно получить в результате последовательного уточнения посылки, начиная с простейшей, так, что все последующие уточнения также будут закономерностями (эта процедура на генеральной совокупности представляет собой вероятностный вывод, этот вероятностный вывод введен и исследован в [28-29]). Кроме того, для достижения этой же цели следует начинать с обнаружения "полного" множества закономерностей сначала на бедных шкалах - булевых и наименований, а затем уже переходить к более сильным шкалам, т.е. стратегия должна опираться на упорядоченность шкал по богатству информации. Это следует из того, что для бедных шкал в большей мере возможен полный перебор, с которого и начинается процесс последовательного уточнения закономерностей.

Таким образом, для того чтобы получить стратегию, нужно следовать пп.4-6. Но кроме того следует учитывать упорядочен-

ность шкал, результаты Теории Измерений и те типы данных, которые у Вас имеются. В соответствии с этим ниже описаны различные возможные стратегии.

Приведем определение отношения $\sqsupset$ - "быть более точным" [28]. Обозначим множество всех подстановок, не являющихся перестановками, через θt (тождественная подстановка принадлежит θt).

ОПРЕДЕЛЕНИЕ. *Отношение* $C \sqsupset C'$, $C = A_1 \& \ldots \& A_n \Rightarrow A_o$; $C' = A'_1 \& \ldots \& A'_n \Rightarrow A'_o$, $n, n' \geq 0$, имеет место тогда и только тогда, когда существует подстановка $\theta \in \theta t$ такая, что $A_o\theta = A'_o$ и $\{A_1\theta, \ldots, A_n\theta\} \subset \{A'_1, \ldots, A'_{n'}\}$ и либо θ не тождественная подстановка, либо $n < n'$.

Прежде чем привести стратегии, необходимо описать типы данных (не шкалы), встречающиеся в различных предметных областях. Эти типы данных определяют множество отношений и операций, которые для этих данных определены, но не аксиомы, которым эти отношения и операции удовлетворяют. Типы данных могут быть следующими: логическими, наименований, порядка и отношений. Такими типами данных могут быть соответственно: булевы матрицы $L(x_1, \ldots, x_n)$ со значениями во множестве $\{0,1\}$; матрицы в шкале наименований $N(x_1, \ldots, x_n)$ со значениями во множестве натуральных чисел; матрицы в шкале порядка $P(x_1, \ldots, x_n)$ с порядковыми значениями (числовые значения сравниваются только отношением порядка) и матрицы в шкале отношений (интервалов, абсолютной) $A(x_1, \ldots, x_n)$ со значениями во множестве вещественных чисел. Здесь $x_1, \ldots, x_n$ - переменные не только по объектам, но и по признакам, номерам экспериментов, числу градаций признаков и т.д. Мы всегда будем считать, что у нас есть переменные двух типов - переменные по объектам из множества $\{a_1, a_2, \ldots\}$ и переменные-параметры из множества $\{par_1, par_2, \ldots\}$. Каждый параметр есть переменная, пробегающая некоторое конечное множество натуральных чисел. Эти множества могут быть раз-

личны для различных переменных. Всегда далее будем иметь в виду, что на определенных местах в матрице могут стоять переменные только строго определенного типа. Кортеж переменных будем обозначать через $\bar{x}$. Аксиомы, которым должны удовлетворять эти отношения и операции, должны быть получены уже в результате анализа этих данных, например, системой DISCOVERY.

1.1. В соответствии с упорядоченностью шкал, сначала следует провести обработку данных в шкале наименований. Имеющиеся числовые значения следует разбить на интервалы, которые в формуле (3) можно задать параметрами. Приведем общий вид законо - мерностей такого вида. Чтобы формула была легко обозримой, мы будем всячески избегать индексов:

$$\forall a(L^{\omega}(\bar{x}) \,\&\, \ldots \,\&\, L^{\omega}(\bar{x}) \,\&$$

$$\&\ (N(\bar{x}) = FN)^{\omega} \,\&\, \ldots \,\&\ (N(\bar{x}) = FN)^{\omega} \,\&$$

$$\&\ (F \le P(\bar{x}) \le F)^{\omega} \,\&\, \ldots \,\&\, (F \le P(\bar{x}) \le F)^{\omega} \,\&$$

$$\&\ (F \le A(\bar{x}) \le F)^{\omega} \,\&\, \ldots \,\&\, (F \le A(\bar{x}) \le F)^{\omega} \Rightarrow$$

$$\Rightarrow \{L^{\omega}(\bar{x}) \mid (N(\bar{x}) = FN)^{\omega} \mid (F \le P(\bar{x}) \le F)^{\omega} \mid$$

$$(F \le A(\bar{x}) \le F)^{\omega})\}\ , \qquad (4)$$

где a - одна переменная по объектам, входящая в кортежи $\bar{x}$; $\omega \in \{0,1\}$ определяет отсутствие или наличие отрицания; FN - целочисленные арифметические выражения, не содержащие перемен - ной по объектам и включающие произвольное число переменных-параметров; F - арифметические выражения, не содержащие переменных по объектам и включающие переменные-параметры; $\mid$ - означает, что в заключении правила (4) стоит только одно из указанных в фигурной скобке выражений.

Закономерности вида (4), обнаруженные для количественных переменных, дают качественный ответ на вопрос: существует ли вообще количественная зависимость. Если для некоторой зависимости (4) в дальнейшем будет найдена более точная зависимость и в

смысле прогноза (оценки условной вероятности), и в смысле вида (логически более сильная зависимость), то первую зависимость можно будет удалить.

1.2. Для отношения порядка и для сочетания типов наименований и порядка существует несколько различных классов закономерностей.

1. Прежде всего это аксиоматический анализ самого отношения порядка - определить, какой системе аксиом отношение порядка удовлетворяет. Само отношение порядка часто задается матрицами бинарного отношения типа L(a,b). В настоящее время известно несколько десятков различных систем аксиом для отношений порядка. Программная система DISCOVERY содержит HELP систем аксиом, в котором есть достаточно большое число систем аксиом для отношений порядка. Они довольно разнообразны. Их следует просто проверить.

2. Закономерности между отношениями порядка (участки монотонной зависимости). Монотонные зависимости, как правило, не распространяются на все множество объектов, поэтому необходимо находить участки монотонной зависимости, ограничивая области монотонности формульно определимыми одноместными предикатами. Приведем общий вид таких закономерностей. Определим сначала формульно определимые одноместные предикаты. Обозначим посылку правила (4) через $\Phi(a)$:

$$\Phi(a) \equiv (L^{\omega}(\bar{x}) \,\&\, \ldots \,\&\, L^{\omega}(\bar{x}) \,\&\, \&\, (N(\bar{x})=FN)^{\omega} \,\&\, \ldots \,\&\, (N(\bar{x})=FN)^{\omega} \,\&\, \&\, (F \le P(\bar{x}) \le F)^{\omega} \,\&\, \ldots \,\&\, (F \le P(\bar{x}) \le F)^{\omega} \,\&\, \&\, (F \le A(\bar{x}) \le F)^{\omega} \,\&\, \ldots \,\&\&\, (F \le A(\bar{x}) \le F)^{\omega}). \qquad (5)$$

Формула $\Phi(a)$ определяет некоторый участок (область) в обобщенном признаковом пространстве, включающем признаки разных шкал, относительно переменной по объектам a. Сформулируем сна-

чала свойство монотонности непосредственно для отношений порядка $\leq$, $>$, которые могут быть заданы, например, типом данных $L(a,b,n)$ или $P(a,b,n)$, где n - номер признака, эксперимента или эксперта:

$$\forall a,b(\Phi(a) \& \Phi(b) \& (a \leq_{i1} b)^{\omega} \& \ldots \& (a \leq_{im} b)^{\omega} \Rightarrow \Rightarrow (a \leq_{io} b)^{\omega}), \qquad (6)$$

где $i1,\ldots,im,i0$ - номера признаков; $\omega \in \{0,1\}$, $(a \leq b)^{0} = (a > b)$, $(a \leq b)^{1} = (a \leq b)$.

Напомним, что мы всячески избегаем индексов и поэтому ω может принимать различные значения для различных отношений и вместо отношения $\leq$ в формуле (6) в произвольных местах может стоять отношение $>$. Отношение порядка также часто задается баллами посредством типа данных $N(a,b,n)$.

3. Системы аксиом для законов в виде простых полиномов [14]. В [14] не только получены системы аксиом для основных физических законов, но и разработан целый класс систем аксиом для законов в виде простых полиномов. Особенность этого класса в том, что системы аксиом сформулированы относительно простейших отношений - отношений равенства или эквивалентности для всех величин, и только для одной из величин - отношение линейного порядка. Если эти отношения удовлетворяют одной из разработанных в [14] систем аксиом, то существуют сильные (логинтервальные) шкалы для всех величин, связанные между собой заданным полиномом. Шкалы и полином получаются процедурой одновременного шкалирования. Эти результаты показывают, что для получения сильных шкал и законов не нужны какие-то особые отношения и операции для величин, а нужна прежде всего определенная их взаимосвязь. Эти результаты подтверждают возможность построения Количественных Теорий в различных предметных областях. Как было отмечено во введении, при переносе измерительных процедур из одной области знания в другую множество интерпре -

тируемых отношений и операций следует пересматривать. Как правило, их становится меньше, шкалы величин обедняются и бытует мнение, что тем самым какая-то важная информация теряется. Упомянутые выше результаты показывают, что это заблуждение, так как отношения равенства, эквивалентности и порядка при этом, как правило, всегда остаются. Используя их и определяя систему аксиом, связывающую эти отношения, можно, в принципе, построить новые более адекватные для данной предметной области сильные шкалы величин и определить связывающий их закон. Поэтому дело не в богатстве отношений, а в их закономерных связях. Мы не будем приводить здесь эти системы аксиом, просто сформулируем в наиболее общем виде полученные из них шкалы и законы.

Для любого простого полинома $y = f(x_1, \ldots, x_n)$ существует (либо может быть построена) система аксиом S^W в словаре $W = \langle \leq_y, =_{x_1}, \ldots, =_{x_n} \rangle$, из истинности которой на некоторой эмпирической системе вытекает существование числовых представлений (сильных гомоморфизмов относительно не только отношений и операций словаря W, но и некоторых определимых через них отношений и операций) $\varphi_y : Y \to Re$; $\varphi_{x_1} : X_i \to Re; \ldots; \varphi_{x_n} : X_n \to Re$ величин $y, x_1, \ldots, x_n$, связанных данным простым полиномом $\forall r(\varphi_y(y) = f(\varphi_{x_1}(x_1), \ldots, \varphi_{x_n}(x_n))), r = \langle y, x_1, \ldots, x_n \rangle$.

4. Суммируем стратегии для шкал порядка. Если есть отношения порядка, системы аксиом которых неизвестны, то их следует установить. Для этого надо проверить содержащиеся в HELP системы аксиом. Если какая-то система аксиом окажется выполненной, то там же будет указано, какой это порядок и какое числовое или конструктивное числовое [9] представление для него существует, либо таковое неизвестно. Если числовое представление неизвестно, то можно попытаться самому построить конструктивное числовое представление, как это определено в [9].

Если среди величин есть отношение линейного порядка, то следует проверить системы аксиом для простых полиномов, содержащиеся в HELP. Если какая-то из систем аксиом для простых полиномов окажется выполненной (как при этом обходиться с аксиомами, содержащими кванторы существования, показано в [2]), то это укажет способ совместного шкалирования всех входящих в зависимость величин так, чтобы они были связаны законом в виде соответствующего простого полинома. Это даст нам новые адекватные шкалы величин и связывающий их закон.

Если все порядки линейные, но ни одна из систем аксиом для простых полиномов не выполнена, то следует провести аксиоматический анализ свойств функциональной зависимости. Для шкал порядка - это свойство монотонности, приведенное в п.2. Общий вид таких зависимостей содержится в HELP.

Если ни одна из упомянутых выше систем аксиом, содержащихся в HELP, не выполнена, то можно продолжить аксиоматический анализ неизвестной зависимости, как это указано в п.4 (с. 133). Для этого надо самому исследователю или эксперту, опираясь на свою интуицию, строить все более точные параметрические семейства правил, уточняющие закономерности в соответствии с определением отношения $\supset$.

1.3. Рассмотрим типы данных L(a,b,c,d) и P(a,b). Тип данных P(a,b) называется матрицей близости (см. [2] или обзоры [30,31]). В [2] показано, что этот тип данных может быть представлен типом данных L(a,b,c,d). Пусть дано некоторое множество объектов $A = \{a_1,\ldots,a_m\}$. Матрицей близости для этих объектов называется матрица (r_{ij}), $i,j = 1,\ldots,m$; r_{ij} - числовые оценки меры близости (сходства или различия) в порядковой шкале (имеют смысл только сравнения величин $r_{ij} \le r_{kl}$). Такие матрицы возникают, например, при сравнении или оценке экспертом всех пар объектов из A в некотором отношении.

Матрицы близости обрабатываются методами многомерного неметрического шкалирования [30,31]. Целью этих методов является представление объектов точками в некотором метрическом пространстве (евклидовом или римановом) минимальной размерности так, чтобы расстояния d_{ij} между ними с точностью до порядка соответствовали бы величинам r_{ij}. После применения методов многомерного шкалирования получается представление данных в метрическом пространстве, после чего, как правило, применяются методы Анализа Данных. Недостатки такого подхода обсуждены в [1].

Чтобы представить эти данные в терминах Теории Измерений, определим на множестве A отношение P типа L(a,b,c,d):

$$P(a,b,c,d) \Leftrightarrow r_{ab} \leq r_{cd}.$$

В Теории Измерений эмпирические системы, включающие подобные четырехместные отношения, обозначаются как $\mathcal{M} = \langle A^*; \leq \rangle$, где $A^* \subset A \times A$, $\leq$ - бинарное отношение упорядочения, определенное на A^*, и называются шкалами разностей [1,14] (положительных разностей, алгебраических разностей, равных конечных промежутков, абсолютных разностей и т.д.).

Следует особо подчеркнуть, что несмотря на то, что разница между шкалой порядка и шкалой разностей небольшая - в обоих случаях мы имеем только одно отношение порядка, определенное на объектах либо на парах объектов, - тем не менее в первом случае мы имеем только шкалу порядка, а во втором случае уже сильную шкалу - шкалу отношений (как и для простых полиномов). Например, для шкалы положительных разностей [1,14] существует шкала Φ такая, что для любых пар $(a,b),(b,c),(c,d) \in A^*$:

$$(a,b) \leq (c,d) \Leftrightarrow \Phi(a,b) \leq \Phi(c,d);$$
$$\Phi(a,c) = \Phi(a,b) + \Phi(b,c).$$

Отображение Φ единственно с точностью до положительного множителя (шкала отношений). Разница в том, что, имея возможность

сравнивать отрезки, можно проводить различные процедуры шкалирования, откладывая равные отрезки. В Теории Принятия Решений известны десятки процедур одномерного шкалирования для субъективных оценок Лица, Принимающего Решение. Эти же процедуры можно применять и для перешкалирования физических приборов, применяемых за пределами физики.

Для проверки систем аксиом разностей необходимо проверить имеющиеся в HELP системы аксиом.

1.4. Как уже было сказано, отношения порядка, эквивалентности и равенства, как правило, всегда остаются при применении физических приборов за пределами физики и получают новую интерпретацию в системе понятий соответствующей предметной области. *Поэтому системы аксиом для простых полиномов, шкал разностей и шкал порядка являются основными при пересмотре шкал и обнаружении законов в новых предметных областях.*

1.5. Системы аксиом многих физических величин основаны на том факте (хотя есть довольно много исключений, см. [14]), что для этих физических величин существует эмпирически интерпретируемая операция $\bullet$, обладающая свойствами операции сложения. Для времени, например, эта операция определяет время $t_1 \bullet t_2$, равное двум следующим друг за другом промежуткам времени t_1 и t_2; для массы эта операция $a_1 \bullet a_2$ интерпретируется как совместное взвешивание двух объектов a_1 и a_2; для длин она интерпретируется как длина $l_1 \bullet l_2$, равная общей длине двух положенных вдоль одной прямой отрезков l_1 и l_2. Такие величины называются экстенсивными. Если для какой-либо из величин в данных есть интерпретируемая двуместная операция $\bullet$, то, используя системы аксиом экстенсивных структур, заложенных в HELP, можно проверить, являются ли эти величины экстенсивными. Для экстенсивных величин можно построить шкалу отношений такую же, как и для упомянутых физических величин.

1.6. Все физические величины (как показано в [14]) разбиваются на две группы: те, для которых существует операция ●, и те, для которых такой операции нет. Последние получают сильную шкалу за счет законов, связывающих эти величины с величинами, имеющими сильные шкалы. Системы аксиом таких законов приведены в [14] (см. также [4]). Они описывают законы вида $y^n = x^m \cdot z^k$ и, кроме того, дают сильные шкалы для величин, у которых нет операции ●. Почти все простейшие физические законы, связывающие между собой всю совокупность физических величин в единую систему физических величин, имеют именно такой вид [14]. За счет этих законов все физические величины имеют сильные шкалы. Проверить эти системы аксиом можно системой DISCOVERY, взяв эти системы аксиом из HELP.

1.7. Если ни одна из систем аксиом п.1.6 не выполнена, а операции • и + для некоторых из величин интерпретируемы, то можно исследовать свойства функций путем последовательного уточнения правил с помощью отношения ⊐, как указано в п.4 (см. с.133).

Поскольку функции исследованы достаточно хорошо, то можно предложить сразу несколько вариантов дальнейшего уточнения вида функций. Уточнения, в частности, могут делаться для выделенных в п.2 (с.136) участков монотонности.

1. Приведем сначала целый класс свойств, которые обычно не исследуются, а предполагаются заданными априори - свойства метрики для образов, кластеров или функций:

$$\forall a,b \ \{\Phi(a)\&\Phi(b)\&(NO(a)=FN)\&(|x(b)-x(a)|\le A)\&\ldots \\ \ldots\&(|x(b)-x(a)|\le A)\Rightarrow(NO(b)=FN)\}, \qquad (7)$$

где $\Phi(a)$ - формула (5); $x(a), x(b)$ - (разные) признаки объектов a,b; $|-|$ - некоторая метрика или норма для одного признака; NO - признак номера образа; FN - целочисленное арифметическое выражение.

2. Приведем простейшее свойство для классов функций - линейность функции (интерпретировать его лучше как свойство пропорциональных изменений для некоторых областей):

$$\forall a,b \{\Phi(a) \& \Phi(b) \& |(x(b)-x(a)| \le A_{i_1}) \& \ldots$$

$$\ldots \& |(x(b)-x(a)| \le A_{i_k}) \Rightarrow$$

$$\Rightarrow (A_3 \le A_2*(y(b)-y(a))/(x(b)-x(a)) \le A_4)\}, \quad (8)$$

где $\Phi(a)$, $\Phi(b)$ - формулы (5); $A_{i_1},\ldots,A_{i_k}$, A_2,A_3,A_4 - арифметические выражения с параметрами.

3. Участки монотонности с допустимыми отклонениями описываются следующей формулой:

$$\forall a,b(\Phi(a) \& \Phi(b) \& (a\pm A_1 \le_{i1} b)^{\omega} \& \ldots$$

$$\ldots \& (a\pm A_k \le_{ik} b)^{\omega} \Rightarrow (a\pm A_o \le_{io} b)^{\omega}, \quad (9)$$

где $\Phi(a)$ и $\Phi(b)$ - формулы вида (5); i1,...,ik,i0 - номера признаков; $A_1,\ldots,A_k,A_o$ - арифметические выражения с параметрами; $\omega \in \{0,1\}$, $(a\pm A_1 \le_i b)^1 = (a- A_1 \le_i b), (a\pm A_1 \le_i b)^0 = (a+ A_1 >_i b)$.

Нетрудно видеть, что продолжать уточнять вид функциональной зависимости можно самыми разнообразными способами и привести их все здесь просто невозможно. Список видов зависимостей, содержащихся в HELP, может постоянно пополняться самими пользователями, поэтому в процессе эксплуатации он будет непрерывно обогащаться.

Укажем теории, на которые можно опираться при анализе свойств функций. Известно, что любую непрерывную функцию можно приблизить рядом Тейлора. В качестве свойств функций можно исследовать поведение приближений к производным. Есть функции, задающиеся дифференциальными уравнениями. Для них можно исследовать соответствующие этим уравнениям приближения в виде раз-

ностных уравнений. Только на разностные уравнения нужно смотреть не как на способ приближенного вычисления, а как на свойства этих функций.

Может показаться, что исследование свойств функций ничем не отличается от тех аппроксимаций, которые находятся методами Анализа Данных. Напомним, что есть два принципиальных отличия: во-первых, свойства функций всегда интерпретируемы в системе понятий предметной области (только такие и рассматриваются) и, во-вторых, свойства функций могут быть установлены со сколь угодно большой степенью достоверности. Это позволяет говорить о них как об "истинных", а не как об аппроксимациях.

§ 2. DISCOVERY - инструмент процесса познания (первая версия)

Как отмечалось выше, перед специалистами в различных областях человеческих знаний часто возникает задача обнаружения полного множества зависимостей в данных по определенным стратегиям. Чем более мощным инструментом пользуется специалист для этих целей, тем большее количество информации он может извлечь из имеющихся данных. Мы будем рассматривать только класс задач и типы данных, описанные выше.

Решать указанные задачи мы будем с помощью системы DISCOVERY (первая версия). При этом в системе DISCOVERY имеются следующие возможности:

- проверять выполнимость на данных некоторых утверждений (далее гипотез), формируемых экспертом или извлекаемых из базы данных с помощью специального HELP;
- определять выполнимость систем аксиом Теории Измерений на данных по определенным стратегиям исполнения;
- проводить аксиоматический анализ данных.

Полученные в результате этих действий знания могут быть использованы: для принятия решений, предсказания и прогноза; для

создания баз знаний экспертных систем; построения предметной области, а также использованы в Теории Открытий.

При работе с данным инструментом, в соответствии с выше сказанным, мы будем исходить из понятия задачи. С понятием задачи мы будем связывать следующее: определение массивов исходных данных; задание параметров вычислений; формирование гипотезы специалистом и выбор им стратегии анализа данных и результатов счета.

Будем исходить из следующей предпосылки: специалист (далее пользователь), берущийся за решение своей задачи системой DISCOVERY, знает свою задачу и,следовательно, может выделить некоторый критерий, по которому он будет отделять решение задачи от ее нерешения.

Сформированную таким образом задачу можно: сохранить на определенный период работы; базифицировать (т.е. занести ее в базу задач и сделать образцом использования другими пользователями); удалить из дальнейшего рассмотрения без возможности дальнейшей работы с ней. При этом каждая задача должна иметь уникальное имя.

Специфика рассматриваемого инструмента такова, что процесс решения задачи по извлечению знаний из данных требует многократного анализа последних с применением некоторых вариаций вводимых предпосылок (далее называемых гипотезами) о характере имеющихся в данных знаний.

Другой особенностью указанного инструмента является то, что конструирование гипотез пользователем напрямую связано с программным обеспечением, которое тоже должно конструироваться одновременно с гипотезой. Причем возможны несколько ситуаций: 1) полное конструирование программного обеспечения по гипотезе пользователя; 2) изменение параметров настройки программной системы без генерации программной системы; 3) подстройка программной системы под измененные входные данные; 4) накопление

и анализ результатов счета в различном объеме в соответствии с заданными настроечными параметрами программной системы. Стратегии работы с системой DISCOVERY приведены выше. Наконец, разрабатываемый инструмент специалиста должен быть максимально к нему дружествен.

Перечисленные специфические особенности указанного инструмента налагают серьезные ограничения на программное обеспечение системы DISCOVERY. Во-первых, это связано с разработкой удобного интерфейса с пользователем по формированию исходной гипотезы и определению априорных сведений и некоторых управляющих и настроечных параметров. Во-вторых, с решением проблем синтеза программных средств по входной гипотезе, задаваемой в виде формулы (3). В-третьих, с созданием, введением и изменением специфической базы данных и знаний, содержащей исходные данные, априорные сведения экспертов, параметры настройки и управления системой, конструируемые гипотезы и их образцы, результаты счета, различные виды подсказок, HELP и многое другое.

В результате проектирования указанной программной системы выделилось пять самостоятельных, но взаимосвязанных блока: 1) блок формирования исходных и вспомогательных данных; 2) блок формирования настроечных, управляющих параметров и конструирования гипотезы; 3) блок генерации программных средств анализа входных данных; 4) блок анализа результатов счета; 5) блок базификации всех фрагментов и самой задачи, если это требуется пользователю.

Такая организация программной системы позволяет гибко использовать имеющиеся и наработанные средства, а также имеющиеся ресурсы компьютера. Рассматриваемая версия системы DISCOVERY реализована средствами технологии решения задач [32] на персональном компьютере типа IBM PC/AT-286 в операционной системе MS DOS 3.3.

Система DISCOVERY оформлена в единой технологической среде и по своему внешнему виду напоминает работу в таких распространенных системах, как, например, Turbo Pascal, LOTUS 1-2-3 и т.п. В системе DISCOVERY вызов и запуск на выполнение каждого из перечисленных блоков осуществляется через главное меню,

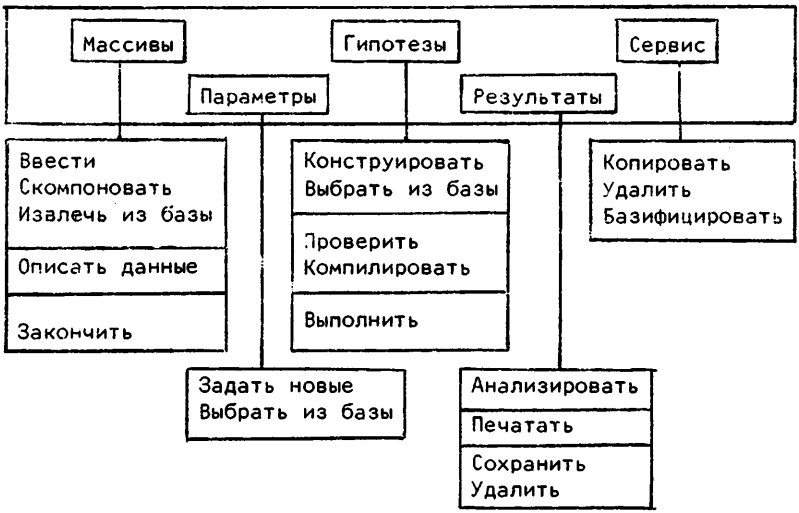


Рис. 1

а вся дальнейшая работа осуществляется с помощью подменю нижних уровней. Рассмотрим более подробно технологию решения задач в системе DISCOVERY (см.рис.1).

1. Технология решения задач в системе DISCOVERY. Цель анализа данных системой DISCOVERY в общем случае состоит в исследовании некоторого объекта, явления или предметной области. Исходные данные пользователь воспринимает как индикаторы состояния, функционирования, изменения и поведения интересующего объекта, явления или предметной области.

Согласно всему выше сказанному, решение задач в системе DISCOVERY можно разделить на несколько этапов.

Этап 1. Подготовка исходных данных для анализа. Данный этап имеет несколько возможностей: первичный ввод исходных данных и определение их структуры; извлечение из базы данных некоторых массивов, подлежащих обработке; компоновка из имеющихся в базе данных новых массивов данных.

В первом случае пользователь вначале должен описать структуру данных (на рис.1 пункт меню: "массивы" → "описать данные"), а затем осуществить ввод данных (пункт меню: "массивы" → "ввести").

Во втором случае пользователь из предоставляемого каталога выбирает файлы с данными, подлежащими обработке, и объявляет их активными (на рис.1 пункт меню: "извлечь из базы").

В третьем случае пользователь в редакторе таблиц отмечает необходимые строки и столбцы в выбранных файлах данных и компонует из них новые файлы требуемой структуры (на рис.1 пункт меню: "скомпоновать").

Этап 2. Формирование настроечных и управляющих параметров. На данном этапе пользователь задает параметры для первоначальной настройки системы на анализ данных. Эта информация не влияет на процесс синтеза программных средств системы и с ее помощью пользователь приводит в соответствие свои априорные знания о характере имеющихся в исходных данных знаний с результатами счета. Таковыми параметрами являются: количество объектов обучения; уровень критерия отбора закономерных связей; количество объектов, участвующих в гипотезе, и количество случайных выборок. Кроме этого, задаются параметры, определяющие полноту выдачи результатов счета, точность вычислений, тип и формат исходных данных, размерность обрабатываемых массивов и их вид (объект-признак, объект-объект, вспомогательный) и некоторые другие.

Указанные выше параметры можно извлечь из базы данных в виде готовых значений (на рис.1 пункт меню: "выбрать из базы")

либо задать новые значения (рис.1, пункт меню: "задать новые").

Этап 3. Конструирование гипотез и генерация по ним программных средств. Данный этап работ в системе DISCOVERY является наиболее ответственным для пользователя, поскольку именно здесь он формирует свое видение исследуемого объекта или явления и представляет его в виде специальным образом сформулированных в виде формул (1) гипотез. Для этих целей он пользуется специальным экранным редактором формул вида (2), полученных из (1) введением индивидных констант, и специальным формульным интерпретатором. Для задания индивидных констант, а также учета специфики исходных данных (тип шкалы, допустимые преобразования, порядок выборки исходных данных, количество задействованных в гипотезе объектов и т.п.) разработан специальный редактор логических формул вида (2) и удобный интерфейс с пользователем (рис.2), в котором действия пользователя целенаправлены и контролируемы. Контроль осуществляет синтаксический анализатор формирования гипотез (рис.2).

Этап 4. Анализ результатов работы системы DISCOVERY. Выше мы сделали замечание о том, что пользователь, решая свою задачу в системе DISCOVERY, знает критерий, по которому он может отличить решение от нерешения. Для удобства анализа результатов счета и проверки критерия правильности вычислений вся результирующая информация в системе DISCOVERY может быть выдана на экран дисплея или получена в виде специального вида твердой копии в виде табл.1. Способов выдачи результатов счета несколько:

- полная выдача всех вариантов гипотезы;

- выдача вариантов-закономерностей (знаний) с обстановочными признаками и возможными сокращениями гипотезы (при условии, что сокращение является гипотезой);

- выдача вариантов-закономерностей с возможными сокращениями гипотезы;

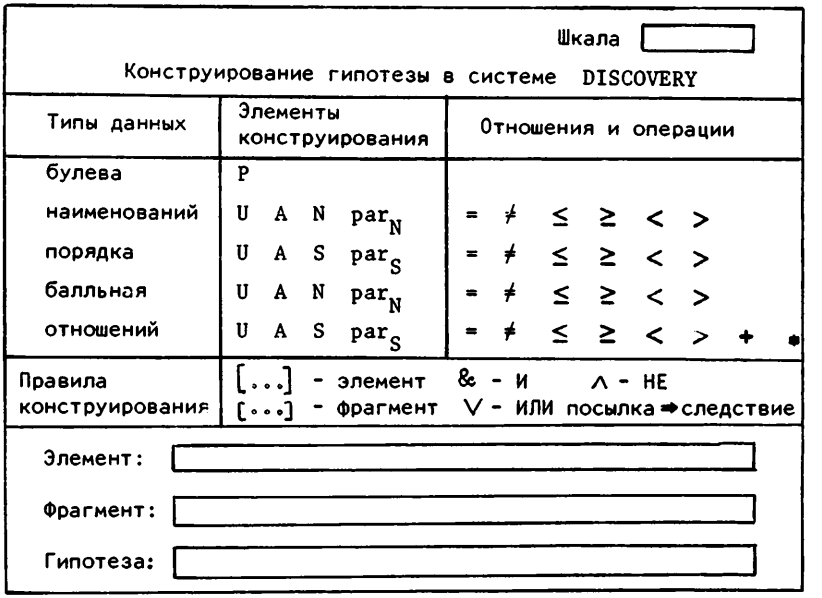


Рис.2

Таблица 1

Форма выдачи результатов счета

Гипотеза в текстовом виде с параметрами										
par_1	par_2	par_3	par_4	...	$лог_1$	$лог_2$	$лог_3$	$лог_4$	...	оцвер
Значения параметров					Индикация статистической оценки логических выражений					Оценка условной вероятности
					обст					
					сущ					
					нез					
					неуд					

"обст" - логическое выражение является обстановочным признаком; "сущ" - логическое выражение существенно для предсказания заключения; "нез" - логическое выражение независимо от заключения; "неуд" - логическое выражение в соответствии со статистическим критерием уменьшает условную вероятность заключения и, следовательно, должно быть удалено из гипотезы.

- выдача вариантов-закономерностей с возможными сокращениями гипотезы в текстовом виде (с подставленными значениями параметров).

Каждая строка таблицы результатов счета представляет собой вариант гипотезы с указанием всех значений параметров, статистических характеристик всех логических выражений (фрагментов гипотезы), представленных в виде одного из четырех утверждений: "обст", "сущ", "нез", "неуд" и оценкой условной вероятности.

Работа данного этапа решения задач в системе DISCOVERY осуществляется из главного меню в пункте "результаты" → "анализировать" с выходом на один из перечисленных выше способов выдачи через меню следующих уровней. Кроме этого, в данном пункте меню возможно выполнить работы по сохранению и удалению результатов счета или же их выдачи в виде твердой копии (см.рис.1).

2. Особенности организации программного обеспечения системы DISCOVERY. В системе задействованы все три типа программного обеспечения: прикладное, системное и инструментальное. Прикладное программное обеспечение в системе DISCOVERY базируется на программах статистических расчетов критериев. В состав системного программного обеспечения входят средства операционной системы MS DOS 3.3 и некоторые средства СУБД. К инструментальному программному обеспечению можно отнести специализированный редактор данных табличного вида и редактор логических формул вида (1), а также специализированный интерпретатор формул вида (2).

Прикладное и системное программное обеспечение в системе DISCOVERY скрыто от глаз пользователя, и поэтому мы не будем задерживать на этом внимание пользователя, а вот инструментальное программное обеспечение рассмотрим более подробно.

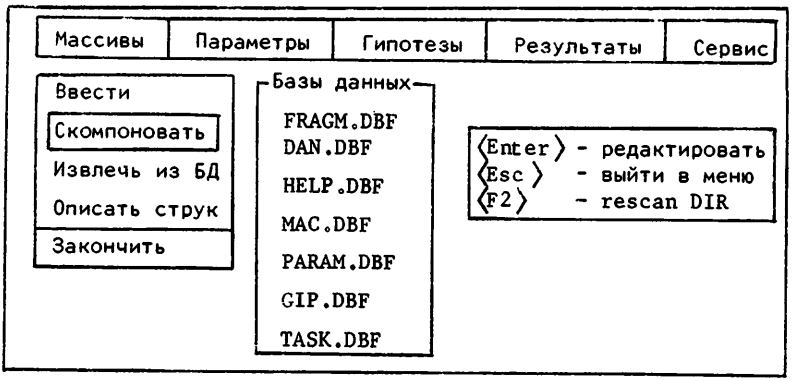
Для удобства работы пользователя его инструментальные средства реализованы в интерактивном режиме с применением следующих типов диалога: меню, команды, подсказки, ответы на вопросы вида ДА/НЕТ и заполнение бланков.

1. Редактор табличных данных. Рассмотрим работу специализированного редактора табличных данных. Выбор в главном меню системы DISCOVERY пункта "массивы" → "скомпоновать" приводит в действие механизм "Pop Up" меню следующего уровня и результатом этого является появление на экране сцен, показанных на рис.3.

Пользователь курсором выбирает имена массивов для редактирования (по очереди) (рис.3а) и клавишами F5 и F6 отмечает выбранные строки и столбцы в таблице (рис.3б). Посредством клавиши F2 он сохраняет отмеченные данные в файле с новым именем, запрашиваемом системой. При выходе из данного режима работ все вновь скомпонованные файлы данных сохраняются в Базе Данных под своими именами и каталогизируются системой для дальнейшей работы с ними.

2. Специализированный редактор формул и формульный интерпретатор. Как уже отмечалось выше, основная нагрузка на пользователя ложится в момент формирования им гипотез относительно скрытых закономерностей в данных. Естественно, что основное внимание он должен уделять именно процессу формирования гипотез и не очень заботиться о том, как это сделать средствами системы. Система же должна взять на себя функции качественного и ответственного исполнителя, а именно, не заставлять пользователя совершать лишние и ошибочные действия; вести его по технологической цепочке с предоставлением необходимой помощи в критических местах; правильно реагировать на неверные шаги пользователя, помогать ему исправлять их и многое другое.

Перечисленные требования являются очень сильными и их полная реализация требует затраты значительных ресурсов и раз-



а) Первое меню редактора табличных данных

p1	p2	p3	p4	p5	p6	p7	p8
08	02	08	10	05	09	00	53
06	04	00	10	01	01	08	79
04	01	03	05	00	03	08	53
08	00	09	08	09	01	05	30
07	08	04	15	05	00	02	83
09	01	08	10	09	07	00	53
09	05	05	14	03	03	09	75
05	08	02	13	07	01	06	83
01	09	04	10	00	01	04	64
01	02	03	03	04	00	03	31
07	05	05	12	06	08	00	66

F5 - строки F6 - столбцы ESC - выйти
F2 - записать отмеченное

б) Основной экран редактора табличных данных

Рис. 3

решения некоторых противоречий (например, между простотой ввода и эффективностью синтаксического и семантического анализов). С целью уменьшения сложности реализации и снятия некоторых неоднозначностей в понимании вводимой пользователем информации в системе DISCOVERY вводится иерархия объектов конструирова-

ния гипотезы и устанавливается определенный порядок конструирования формул. При этом вся необходимая для конструирования формул информация выдается на экран дисплея в виде, удобном для применения.

В качестве объектов конструирования формул (2) выделены:

- элемент - часть формулы (2), содержащая константы либо обращение к исходным массивам данных и связанная знаками арифметических операций. Элемент гипотезы всегда выделяется квадратными скобками;

- фрагмент - состоит из элементов гипотезы, соединенных знаком логической операции (& - И, V - ИЛИ, ! - НЕ). Фрагмент гипотезы заключается в фигурные скобки;

- гипотеза - считается сформированной, если она состоит из посылки и заключения, сформированных из фрагментов и соединенных знаками логических операций.

В системе предусмотрены две возможности конструирования формул (2):

- по жесткому сценарию (для неподготовленных пользователей), когда нарушения в синтаксисе конструируемой формулы не позволяют продвинуться дальше до тех пор, пока ошибки не будут исправлены;

- по произвольному сценарию (для подготовленного пользователя), когда проверка правильности синтаксиса полностью лежит на самом пользователе.

Кроме того, в базе данных системы DISCOVERY имеется набор образцов гипотез (формул вида (2)), которые доступны пользователю для редактирования или первоначального применения. Образцы гипотез можно вызвать и использовать для работы.

В качестве дополнительного контроля правильности ввода формул (2), перед выходом из экранного редактора, пользователю еще раз предлагается проверить правильность введенной гипотезы и подредактировать ее перед запуском интерпретатора.

Для работы формульного интерпретатора от пользователя активного участия не требуется. Вся необходимая информация генерируется в экранном редакторе формул и подается на вход интерпретатора по команде, выдаваемой из главного меню (рис.1,пункт меню: "гипотезы" → "проверить").

На вход формульного интерпретатора подаются:

- сконструированная формула (2) с включенными элементами и фрагментами;

- имена исходных и вспомогательных массивов данных;

- управляющие и настроечные параметры,определенные ранее пользователем,и некоторые другие вспомогательные аргументы.

Отработав, формульный интерпретатор и блок статистических проверок выдают несколько массивов результатов счета для их дальнейшего анализа на этапе 4.

3. Анализатор результатов счета. Анализатор результатов счета в системе DISCOVERY представляет из себя специализированный редактор, подключенный к базе данных системы. Он может работать в оперативном режиме,и тогда все результаты его работы видны на экране дисплея, и также в фоновом режиме,и тогда все результаты его работы в окончательном виде записываются в соответствующие файлы базы данных, которые доступны СУБД, входящей в состав системы DISCOVERY.

Внешний вид экрана оперативного режима анализа результатов представлен табл.1.

Блочная организация системы DISCOVERY оказалась удобной с точки зрения использования ее как инструмента исследователя, поскольку зачастую у пользователя возникает потребность прервать процесс работы и возобновить его при более подходящей ситуации. Именно это и позволяет сделать система DISCOVERY. Вы можете независимо сформировать исходные данные, определить базовые параметры вычислений, сконструировать гипотезу, провести вычисления и проанализировать результат. И вместе с тем, у Вас

есть возможность фиксировать каждый свой шаг в базе данных и многократно возвращаться к тем или иным данным по мере возникновения такой необходимости. Кроме этого, Вы можете базифицировать каждый из выше приведенных этапов вычислений и использовать его в дальнейшем как образец (стандарт) решения аналогичных задач или использовать в процессе обучения пользователя работе с системой.

3. Модельный пример работы системы DISCOVERY. В заключение рассмотрим модельный пример работы системы DISCOVERY по выявлению закономерных связей в данных и их гипотетическую интерпретацию.

Модельный пример, с одной стороны, должен содержать заранее известные закономерности, которые должен обнаружить метод, с другой стороны, эти закономерности должны иметь некоторую гипотетическую интерпретацию в решаемых задачах и, с третьей стороны, эти закономерности не должны обнаруживаться другими методами. Рассматриваемый далее пример удовлетворяет указанным требованиям.

Приведенная ниже табл.2 получена следующим образом: признаки 1-3 и 5-9 сгенерированы датчиком случайных чисел. Признаки 1-3 моделируют некоторые базисные независимые между собой показатели деятельности производства. Признак 4 является суммой первых двух, а признак 10 (целевой, интересующий нас) равен некоторому случайному монотонному преобразованию F от четвертого минус третий (т.е. условно можно записать 10 = F(4-3)= = F(1+2-3)). Признаки 5-9 моделируют случайные, не имеющие отношения к решаемой задаче признаки. Гипотетической интерпретацией зависимости F(1+2-3) может быть, например, следующая: эффективность работы предприятия (например, рост прибыли) выше, если у предприятия больше начальный капитал (признак 1 выражается в баллах), выше деловые качества директора (признак 2 - в баллах) и ниже налоги (признак 3 - в баллах). Монотонное пре-

Таблица 2

Признаки									
1	2	3	4	5	6	7	8	9	10
08	02	08	10	05	09	00	04	01	53
06	04	00	10	01	01	08	01	05	79
04	01	03	05	00	03	08	07	01	53
08	00	09	08	09	01	05	00	00	30
07	08	04	15	05	00	02	08	06	83
09	01	08	10	09	07	00	03	04	53
09	05	05	14	03	03	09	07	07	75
05	08	02	13	07	01	06	06	08	83
01	09	04	10	00	01	04	07	01	64
01	02	03	03	04	00	03	02	07	31
07	05	05	12	06	08	00	02	04	66
02	01	06	03	03	08	01	07	02	17
04	06	05	10	06	06	07	05	04	61
01	05	08	06	04	09	00	09	05	24
07	03	09	10	01	04	05	08	02	51
02	06	09	08	09	05	03	05	02	30
01	08	09	09	06	06	03	08	02	31
09	03	07	12	03	09	01	02	01	61
00	04	04	04	04	02	00	03	04	31
09	01	05	10	05	04	02	06	01	61
08	07	06	15	01	02	07	01	01	75
07	09	06	16	09	06	07	03	01	79
08	09	02	17	01	05	06	03	09	96
08	02	07	10	09	04	01	01	06	56
00	01	07	01	06	09	00	06	06	09
07	07	01	14	05	00	05	02	05	91
02	09	07	11	02	01	01	05	03	58
05	05	01	10	02	04	05	08	08	75
01	07	00	08	06	03	06	09	07	72
04	06	00	10	00	09	08	02	07	79
00	06	01	06	03	07	08	06	01	61
08	04	04	12	02	06	06	04	06	72
03	04	06	07	07	02	01	07	06	51
02	07	08	09	03	03	08	04	04	51
06	05	09	11	07	02	04	05	05	53
09	03	09	12	08	05	08	06	00	56
01	06	02	07	00	07	06	00	08	61
06	00	00	06	00	02	03	07	05	64
02	07	09	09	00	01	06	09	08	31
04	08	02	12	00	09	05	01	05	79
01	04	03	05	07	09	06	06	07	53
03	08	09	11	05	06	02	04	02	53

Продолжение таблицы 2

02	09	04	11	04	07	04	00	01	66
04	01	09	05	09	00	02	08	05	14
09	01	05	10	04	00	07	09	02	61
05	08	04	13	08	06	02	05	08	75
05	08	06	13	06	00	05	09	09	66
09	02	05	11	02	01	09	07	02	64
05	08	08	13	03	06	02	01	07	61
03	07	09	10	02	01	06	04	07	51
08	04	08	12	08	02	04	09	00	58
07	00	08	07	06	00	08	01	05	30
04	01	06	05	09	02	08	09	07	30
08	06	08	16	04	08	08	02	08	72
01	05	05	06	02	03	07	06	08	51
01	03	09	04	02	04	05	06	04	12
07	03	00	10	06	03	07	06	06	79
09	00	04	09	03	08	00	00	05	61
05	07	04	12	01	06	08	06	08	72
05	01	09	06	01	04	08	08	04	17
03	00	05	03	07	06	05	04	01	24
07	07	07	14	07	09	07	05	07	66
00	02	09	02	05	08	05	01	08	05
03	04	04	07	00	09	09	07	08	56
09	01	05	10	01	08	00	06	03	61
08	06	06	14	06	02	06	07	02	72
00	08	09	08	09	07	02	01	01	30
04	04	03	08	09	02	08	02	05	61
09	07	08	16	04	04	05	00	09	72
06	04	07	10	04	07	04	09	06	56
05	02	04	07	07	07	07	07	03	56
02	02	09	04	03	06	02	05	02	12
01	07	08	08	05	00	06	05	06	31
09	01	08	10	03	06	05	05	04	53
08	01	05	09	05	02	07	03	01	58
02	07	07	09	08	00	04	08	09	53
03	05	06	08	05	06	02	04	02	53
05	05	01	10	08	07	09	01	03	75
03	05	07	08	08	03	03	06	07	51
04	05	04	09	08	06	04	08	01	61

образование взято по следующим причинам.

Во-первых,чтобы этот пример не выглядел абстрактно-математическим. В реальных задачах, если есть какая-то скрытая математическая зависимость между признаками, то она в измеряе-

мых показателях сильно искажена. Опыт решения задач показывает, что с точностью до порядка, т.е. с точностью до монотонного преобразования эти зависимости проявляются в реальных показателях.

Во-вторых, существующие методы анализа данных не могут (и для них это принципиально невозможно) обнаружить зависимость, искаженную случайной монотонной зависимостью. Поэтому данный пример иллюстрирует уникальные возможности предлагаемого метода. Единственная возможность обнаружить зависимость после ее случайного монотонного преобразования (и это можно доказать) это обнаружить зависимость в терминах интерпретируемых отношений порядка для всех признаков $\leq_1, \leq_2, \leq_3, \ldots, \leq_{10}$. Такие зависимости, в частности, обнаруживаются предлагаемым методом.

В-третьих, отношение порядка, а также все другие используемые данным методом отношения, интерпретируемы в терминах решаемой задачи и поэтому полученный результат в виде закономерности $\forall a,b(a \leq_1 b \,\&\, a \leq_2 b \Rightarrow a \leq_{10} b)$ также будет интерпретируемым в понятиях задачи в отличие от, например, методов регрессии, которые дают функцию, как правило, не интерпретируемую в системе понятий решаемой задачи.

Таким образом, в данной матрице по построению заложены следующие закономерности:

$$\begin{aligned}
&\forall a,b(a >_3 b \,\&\, a \leq_4 b \Rightarrow a \leq_{10} b),\\
&\forall a,b(a \leq_3 b \,\&\, a >_4 b \Rightarrow a >_{10} b),\\
&\forall a,b(a \leq_1 b \,\&\, a \leq_2 b \,\&\, a >_3 b \Rightarrow a \leq_{10} b),\\
&\forall a,b(a >_1 b \,\&\, a >_2 b \,\&\, a \leq_3 b \Rightarrow a >_{10} b).
\end{aligned} \qquad (10)$$

Системой DISCOVERY можно параметрически задать обнаружение всех закономерностей типа монотонной зависимости так, что будут найдены все требуемые закономерности.

Простейшим видом монотонных зависимостей является следующее параметрическое семейство формул: $\forall a,b(a \leq_i b \Rightarrow a >_{10} b)$, $i = 1,\ldots,9$. Проверив эти закономерности с уровнем 0,0001,можно обнаружить закономерность: $\forall a,b(a \leq_4 b \Rightarrow a \leq_{10} b)$. Этой закономерности нет в списке (1), хотя она является закономерностью.

Следующим видом монотонных закономерностей является параметрическое семейство $\forall a,b(a >_i b \,\&\, a \leq_j b \Rightarrow a \leq_{10} b)$, $i,j = 1,\ldots,9$. При проверке этих закономерностей с уровнем 0,025 будет обнаружена закономерность: $\forall a,b(a >_3 b \,\&\, a \leq_4 b \Rightarrow a \leq_{10} b)$.

Аналогично задается другое параметрическое семейство формул для обнаружения второй закономерности.

Для обнаружения закономерностей, содержащих три предиката в посылке, действуем аналогично: задаем закономерность соответствующего вида, заменяя конкретные номера признаков параметрами. Например, для обнаружения закономерности:

$$\forall a,b(a \leq_1 b \,\&\, a \leq_2 b \,\&\, a >_3 b \Rightarrow a \leq_{10} b)$$

задаем параметрическое семейство: $\forall a,b(a \leq_i b \,\&\, a \leq_j b \,\&\, a >_k b \Rightarrow a \leq_{10} b)$, $i,j,k = 1,\ldots,9$. Проверяем эти закономерности с уровнем 0,1. Можно убедиться,что нужная закономерность будет обнаружена. После этого надо менее сильные закономерности исключить из списка,тогда останутся только закономерности (10).

Литература

1. ВИТЯЕВ Е.Е. Обнаружение закономерностей (методология, метод, программная система SINTEZ).1. Методология// Методологические проблемы науки.- Новосибирск,1991.- Вып.138: Вычислительные системы.- С.26-60.

2. ВИТЯЕВ Е.Е. Метод обнаружения закономерностей и метод предсказания// Эмпирическое предсказание и распознавание образов.- Новосибирск,1976.-Вып.67: Вычислительные системы.- С.54-68.

3. ВИТЯЕВ Е.Е. Закономерности в языке эмпирических систем// Эмпирическое предсказание и распознавание образов.- Новосибирск, 1978.- Вып.76: Вычислительные системы.-С.3-14.

4. ВИТЯЕВ Е.Е. Закономерности в языках эмпирических систем и законы классической физики// Эмпирическое предсказание и распознавание образов.- Новосибирск,1979.- Вып.79: Вычислительные системы.- С.45-56.

5. ВИТЯЕВ Е.Е. Обнаружение функциональных зависимостей с одновременным формированием понятий // Тезисы Второй Всесоюз. конф. по автоматизации поискового конструирования.-Новосибирск, 1980.- С.171-172.

6. ВИТЯЕВ Е.Е. Упрощение функциональных зависимостей за счет перешкалирования величин// Вторая Всесоюз. школа-семинар по "Программно-алгоритмическому обеспечению прикладного многомерного статистического анализа".- М., 1983.- С.260-262.

7. ВИТЯЕВ Е.Е., МОСКВИТИН А.А. ЛАДА - программная система логического анализа данных// Методы анализа данных.- Ново - сибирск,1985.- Вып.111: Вычислительные системы.- С.38-58.

8. ВИТЯЕВ Е.Е. Числовое, алгебраическое и конструктивное представление одной физической структуры// Логико-математические основы МОЗ.- Новосибирск, 1985. - Вып.107: Вычислительные системы.- С.40-51.

9. ВИТЯЕВ Е.Е. Конструктивное числовое представление ве - личин// Методы анализа данных.- Новосибирск,1985.-Вып.111: Вычислительные системы.- С.23-32.

10. ВИТЯЕВ Е.Е. Шкала экстенсивных величин как абстрактный тип данных// Всесоюз. конф. по прикладной логике: Тезисы докл. - Новосибирск, 1985.- С.37-39.

11. ВИТЯЕВ Е.Е. Логико-операционный подход к анализу дан - ных// Комплексный подход к анализу данных в социологии. Труды Института Социологических исследований АН.- М.,1989.- С. 113-122.

12. Малая Медицинская Энциклопедия.- М., 1967.- С.352-359.

13. Психологические измерения/ Под ред.Л.Д.Мешалкина.- М.: Мир, 1967.- 120 с.

14. Foundations of measurement.Vol.1 /Krantz D.H., Luce R.D., Suppes P., Tversky A. - New York, London: Academic Press, 1971. - 577 p.

15. ПФАНЦАГЛЬ И. Теория измерений.-М.: Мир,1976.-248 с.

16. ФИШБЕРН П.С. Теория полезности для принятия реше - ний. - М.: Наука, 1978. - 352 с.

17. ANDERSON N.H. Integration theory, functional measurement and the psychological law //Advances in psychophysics /Ed.Geissler, Yu.Zabrodin. - Berlin, 1976. - P.93-130.

18. ANDERSON N.H. Algebraic Rules in Psychological Measurement //Amer.Soientist. - 1979. - Vol. 67. - P.555-563.

19. LANGLEY P., ZYTKOW J.M. Data-Driven Approaches to Empirical Discovery //Artificial Intelligence. - 1989. - Vol. 40, N.1-3. - P. 283-312.

20. КУЛАКОВ Ю.И.Новая формулировка теории физических структур// Методологические и технологические проблемы информацион - но-логических систем.- Новосибирск,1988.- Вып.125: Вычислительные системы.- С.3-32.

21. САМОХВАЛОВ К.Ф. К обоснованию теории физических структур // Там же.- С.33-41.

22. МИХАЙЛИЧЕНКО Г.Г. Решение функциональных уравнений в теории физических структур //Докл. АН СССР. - 1972. - Т.206,№5. - С. 1056-1058.

23. ВИТЯЕВ Е.Е., КОСТИН В.С. Естественная классификация как закон природы // Интеллектуальные системы и методология: Материалы научно-практического симпозиума "Интеллектуальная поддержка деятельности в сложных предметных областях".- Ново - сибирск,7-9 апреля 1992 г. - Вып.4.- Новосибирск,1992.- С.107-115.

24. PZELECKI M. The logic of empirical theories.- London: Routledge Kogan Paul, 1969. - 109 p.

25. HALPERN J.Y. An Analysis of First-Order Logic of Probability //Artificial Intelligence. - 1990. - Vol. 46. - P. 311-350.

26. ВИТЯЕВ Е.Е. Обнаружение закономерностей, выраженных универсальными формулами // Эмпирическое предсказание и распознавание образов.- Новосибирск, 1979. - Вып.79: Вычислительные системы.- С.57-59.

27. ВИТЯЕВ Е.Е. Классификация как выделение групп объек - тов, удовлетворяющих разным множествам согласованных законо - мерностей // Анализ разнотипных данных.- Новосибирск, 1983. - Вып.99: Вычислительные системы.- С.44-50.

28. Prediction and inductive synthesis of PROLOG-programs by a probabilistic model of data. Предсказание и индуктивный синтез ПРОЛОГ-программ по вероятностной модели данных //Institute of Mathematics SD AS USSR,Тр. Института математики СО АН СССР, 1990.

29. ВИТЯЕВ Е.Е. Семантический подход к созданию баз знаний. Семантический вероятностный вывод наилучших для предсказания ПРОЛОГ-программ по вероятностной модели данных// Логика и семантическое программирование.- Новосибирск, 1992.-Вып.146: Вычислительные системы.- С.19-49.

30. КАМЕНСКИЙ В.С. Модели и методы неметрического многомерного шкалирования. (Обзор) // Автоматика и телемеханика. - 1977.- № 8.- С.118-156.

31. ТЕРЕХИНА А.Ю. Методы многомерного шкалирования и визуализация данных. (Обзор) // Автоматика и телемеханика.-1973. - № 7.- С.80-94.

32. КОСАРЕВ Ю.Г., МОСКВИТИН А.А. Проблемно-инструментальная технология построения программных систем // Методы анализа данных.- Новосибирск, 1985.- Вып.111: Вычислительные системы.- С.59-75.

Поступила в ред.-изд.отд.

24 мая 1993 года