УДК 518.74

Инструментальное средство Visual Discovery решения задач интеллектуального анализа данных

А.А. Подберезный, Е.Е. Витяев, А.А. Москвитин[1]

В работе кратко изложен оригинальный реляционный подход к интеллектуальному анализу данных и представлена программная система Visual Discovery, реализующая данный подход с максимальным удобством для пользователей. Главным достоинством разработанной системы является визуальный конструктор отношений и операций над данными, интерпретируемых в онтологии предметной области, а также конструктор гипотез, проверяемых на данных.

Ключевые слова: интеллектуальный анализ данных, искусственный интеллект, data mining, knowledge discovery in data bases.

1. Введение

1. 1. Познание предметной области

Определим, что такое предметная область (ПО). *Предметная область* – это совокупность *объектов предметной области*, рассматриваемых с точки зрения некоторого *предмета исследования* – совокупности *существенных свойств (атрибутов)* и *отношений* объектов исследования, описываемых в некоторой *системе понятий* предметной области. Предмет исследования может быть задан *онтологией предметной области* – специфицирующей в некотором формальном языке множество рассматриваемых объектов, связи между ними, систему понятий, и свойства объектов. Предмет исследования и онтология определяют «взгляд», «точку зрения», с которой рассматриваются (описываются в системе понятий) объекты предметной области, отношения и их свойства.

Предметная область может быть задана эмпирической системой $\Im = \langle A, \Omega \rangle$, где A – объекты ПО, а Ω – множество отношений и операций, интерпретируемых в системе понятий ПО и определяющих взаимосвязь объектов ПО. Система понятий онтологии может быть задана одноместными предикатами, которые также могут входить в Ω. Таким образом, множество Ω представляет собой *онтологию предметной области*, так как является формальной спецификацией связей между объектами, системой понятий и свойствами объектов.

Для осуществления *процесса познания* необходимо *понимание* и *интерпретация* человеком предметной области и её онтологии, т.е. извлечение информации из предметной области. «*Информация* – это понимание (смысл, представление, интерпретация), возникающее в аппарате мышления человека в результате получения им данных, взаимоувязанное с предшествующими знаниями и понятиями» [1]. *Информация о предметной области* и *онтологии* состоит из восприятия и интерпретации человеком объектов предметной области, связей

[1] Работа поддержана грантом РФФИ 08-07-00272-а, Российским гуманитарным научным фондом, проект № 12-01-12026, интеграционными проектами СО РАН № 3, 86, 136 НШ-276.2012.1 и программой Президента Российской Федерации поддержки научных школ НШ-276.2012.1.

между ними, системы понятий и свойств объектов. В результате такой интерпретации получаем *знание* о предметной области. «Знания – это воспринятая, осознанная и ставшая личностно значимой информация» [2].

1.2. Информация, содержащаяся в атрибутах и свойствах объектов

Проанализируем, как следует задавать свойства и атрибуты объектов ПО в терминах онтологии Ω. Чтобы правильно извлекать информацию и знания из свойств и атрибутов, необходимо их интерпретировать в системе понятий ПО. Сами по себе числовые значения величин смысла и информацию не содержат, смысл величин указывается в их интерпретации, например, 5 метров, 5 литров, 5 килограмм и т.д. Интерпретация чисел, в частности, определяет, какие математические действия можно с ними осмысленно проводить, чтобы не получать бессмысленных результатов типа 1.5 дровосека, и т.п.

Как говорилось в [3 – 6] интерпретация числовых значений – метры, литры, килограммы и т.д. – привязана к соответствующей ПО и её онтологии. Физические величины, измеренные в отличной от физики предметной области, теряют свою физическую интерпретацию. Рассмотрим, например, такую физическую величину, как температура. Шкалы температур 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 могут одним и тем же прибором – термометром. Далеко не всегда понимается тот факт, что шкала – это набор отношений и операций, которые имеет смысл производить с числовыми значениями величин в данной предметной области. Точнее, это те отношения и операции, которые интерпретируемы в онтологии ПО. Можно возразить, что термометр не может измерять ничего, кроме температуры. Он действительно во всех случаях измеряет *физическую* температуру. Но зачем мы измеряем температуру? Ведь не затем, чтобы согласно законам физики узнать, сколько в больном содержится тепла, и не затем, чтобы определить среднюю кинетическую энергию молекул почвы или курицы в духовке. Термометр, как и любой другой прибор, нужен для *получения выводов (знаний) в системе понятий* (онтологии) той предметной области, к которой он 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» [7]. Для почв температура интерпретируется в системе понятий физиологии растений и деятельности микроорганизмов. Физическая величина температуры в других предметных областях *является косвенным измерением* некоторой другой величины, интерпретируемой в системе понятий предметной области, которую мы и хотим измерить. Физическая температура больного 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 свёртывания белка и т.д. Какие отношения и операции над числовыми значениями температуры имеют смысл для всех этих величин, определяется уже этими интерпретациями и онтологиями соответствующих ПО. Например, для температуры больного интерпретируемы выделенные значения 36.7, 42.0 и отношение линейного порядка $<$.

Таким образом, для извлечения информации из атрибутов, свойств, признаков и величин ПО нужно определить множество интерпретируемых в онтологии Ω математических отношений и операций и включить их в онтологию Ω. Именно эта процедура извлечения информации из данных является ключевой в системе Visual Discovery.

Зачем нужно такое извлечение информации из атрибутов, свойств и величин ПО?

1.3. Познание предметной области методами интеллектуального анализа данных. Онтология методов интеллектуального анализа данных

Рассмотрим методы интеллектуального анализа данных (Knowledge Discovery in Data Bases and Data Mining (KDD&DM)) и машинного обучения (Machine Learning (ML)) с точки зрения извлечения знаний из данных. В силу определения, знание, получаемое методами интеллектуального анализа, должно быть интерпретируемо в системе понятий ПО. Для этого методы KDD&DM и ML должны правильно использовать содержащуюся в данных информацию. Рассмотрим этот вопрос более подробно.

Анализ методов KDD&DM и ML [3, 5] показывает, что *методы* имеют свою *онтологию*, которая включает:

1. типы данных, с которыми работает метод;
2. язык оперирования и интерпретации данных;
3. класс гипотез, проверяемый методом и сформулированный в языке интерпретации данных.

Для того чтобы применение KDD&DM&ML-методов давало знания – интерпретируемые в онтологии ПО высказывания – необходимо, чтобы онтология метода и онтология ПО были согласованы между собой, т.е.:

1. типы данных, с которыми работает метод, должны интерпретироваться в онтологии Ω предметной области. Поэтому атрибуты, свойства и признаки, используемые в данных метода, должны быть интерпретируемы в онтологии Ω. Тем самым определяется *информация, извлекаемая из данных этим методом*, которая представляется множеством интерпретируемых в онтологии Ω математических отношений и операций;
2. язык оперирования данными, используемый методом в своей работе, также должен интерпретироваться в онтологии Ω. Это значит, что метод должен использовать в своей работе только интерпретируемые в онтологии Ω математические отношения и операции. Если это не так, то метод получает не вполне интерпретируемые и не являющиеся знаниями результаты. Человек не может осознать результаты математических действий, применённых методом, которые для него не имеют интерпретацию и, следовательно, бессмысленны с точки зрения системы понятий ПО;
3. класс проверяемых методом гипотез также должен интерпретироваться в онтологии ПО. Это означает, что класс проверяемых гипотез также должен выражаться через интерпретируемые в онтологии Ω математические отношения и операции. Например, решающие функции в распознавании образов, функции регрессии, формы кластеров в признаковом пространстве и т.д. должны содержать только интерпретируемые математические отношения и операции.

В настоящее время такого рода проверка на соответствие онтологии ПО и онтологии метода, как правило, не проводится. Для того чтобы знать, какая информация содержится в данных и, следовательно, какой метод KDD&DM и ML мы можем применить для обработки данных, нам необходимо *извлечь информацию из данных*. Разработанная нами система Visual Discovery позволяет это делать в режиме визуального конструктора.

1.4. Извлечение информации из данных

Рассмотрим подробнее, что такое информация, содержащаяся в данных, и как она может быть представлена эмпирическими системами $\Im = \langle A, \Omega \rangle$, являющимися частью эмпирических систем предметной области. Для этого покажем, как информация, содержащаяся в таких известных типах данных как: матричное представление бинарных отношений, матрицы упорядочений, матрицы близости и матрицы объект-признак – могут быть представлены эмпирическими системами $\Im = \langle A, \Omega \rangle$. Эти типы данных встречаются в таких областях, как

экспертное оценивание, социология, психология, психофизика, геология, медицина, сельское хозяйство и т. д. Все эти области характеризуются тем, что в них встречаются признаки и величины самой разнообразной природы.

Для полученных эмпирических систем приведём относящиеся к ним результаты теории измерений [8 – 9], показывающие, как корректно эта информация представляется числами. Эти результаты включают в себя системы аксиом и теоремы существования и единственности соответствующих числовых представлений. Теоремы единственности дают нам группы допустимых преобразований шкал, что позволяет определять методы анализа данных, инвариантные относительно этих групп и, следовательно, применимые к этим данным.

Многоместные отношения возникают естественным образом, если источником информации являются суждения человека [10]. Как показали многие эксперименты, человек более правильно и с меньшими затруднениями отвечает на вопросы качественного, в частности, сравнительного характера, чем количественного. В различных дисциплинах человека называют по-разному: экспертом в экспертных оценках, испытуемым в психологии и психофизике, респондентом в социологии и пациентом в медицине и т.д.

Матричное представление бинарных отношений. Бинарное отношение $P(a,b)$, определённое на множестве объектов $A=\{a_1,\ldots,a_m\}$, задаётся матрицей $(e_{ij}), i, j = 1,\ldots,m$, где $e_{ij} = 1(0)$ означает, что $P(a_i,a_j)$ истинно (ложно). Такой матрицей можно задать произвольное бинарное отношение на множестве *A*. Информация, содержащаяся в матрице бинарного отношения, может быть задана эмпирической системой $\Im=\langle A,P\rangle$, где $P(a_i,a_j)\Leftrightarrow e_{ij}=1$.

Матричное представление бинарных отношений широко используется в работах [11 – 12] ввиду его привычности и простоты. Наиболее часто используются отношения эквивалентности, квазипорядка, частичного порядка и лексикографического порядка.

Приведём результаты теории измерений, относящиеся к бинарным отношениям [8 – 9, 13].

Отношение толерантности. Для любых $a,b\in A$

1. $P(a,a)$;
2. $P(a,b)\Leftrightarrow P(b,a)$.

Числового представления не существует.

Отношение эквивалентности. Для любых $a,b,c\in A$:

1. $P(a,a)$;
2. $P(a,b)\Leftrightarrow P(b,a)$;
3. $P(a,b)\,\&\,P(b,c)\Rightarrow P(a,c)$.

Числовое представление существует в виде нумерации классов.

Отношение частичного порядка. Для любых $a,b,c\in A$:

1. $P(a,a)$;
2. $P(a,b)\,\&\,P(b,c)\Rightarrow P(a,c)$.

Числового представления не существует.

Отношение интервального упорядочения. Для любых $a,b,c,d\in A$:

1. $\neg P(a,a)$;
2. $P(a,b)\,\&\,P(c,d)\Rightarrow(P(a,d)\vee P(c,d))$.

Числовое представление существует [14]. Существуют две вещественнозначные функции $U,V:A\to Re^+$, такие, что для любых $a,b\in A$,

$$P(a,b) \Leftrightarrow (U(a) + V(a)) < U(b).$$

Отношение полупорядка. Отношение P называется отношением полупорядка, если оно является отношением интервального порядка и для любых $a,b,c,d \in A$ удовлетворяет аксиоме:

3. $P(a,b) \& P(b,c) \Rightarrow P(a,d) \vee P(d,c)$.

Числовое представление существует [14]. Существует вещественнозначная функция $U : A \to Re$ такая, что для любых $a,b \in A$,

$$P(a,b) \Leftrightarrow (U(a) + 1) < U(b).$$

Отношение древесного порядка. Отношение P называется отношением древесного порядка, если для любых $a,b,c \in A$ удовлетворяет аксиоме:

1. $(a < b) \& (a < c) \Rightarrow (b < c) \vee (c < b)$;

2. Существует наибольший элемент.

Числового представления не существует.

Отношение слабого порядка (квазисерии, предпорядки). Для любых $a,b,c \in A$ удовлетворяет аксиомам:

1. $P(a,b) \vee P(b,a)$;

2. $P(a,b) \& P(b,c) \Rightarrow P(a,c)$.

Если упорядоченная система $\langle A; P \rangle$ имеет счётную базу, то числовое представление существует.

Не все из приведённых отношений имеют числовые представления. Поэтому не всегда данные, содержащие бинарные отношения, можно представить в некотором числовом пространстве.

Матрицы упорядочений: $(r_{ij}), i = 1,...,m; j = 1,...,n$; r_{ij} – оценка *i*-го объекта по *j*-му признаку. Такие матрицы могут выражать либо упорядочения *k* объектов *n* экспертами, либо упорядочения *k* объектов по *n* ранговым признакам [12]. Такие матрицы обрабатываются методами многомерного шкалирования [16] и методами ранжирования [15], а также некоторыми из методов обработки матричного представления бинарных отношений (см. п. 3).

Информация, содержащаяся в матрице упорядочения, может быть задана эмпирической системой $\Im = \langle A, P_1, ..., P_n \rangle$, где каждому признаку j соответствует отношение P_j, $j = 1,...,n$, определённое следующим образом: $P_j(a_{i_1}, a_{i_2}) \Leftrightarrow r_{i_1 j} < r_{i_2 j}$.

Матрицы близости. Пусть дано некоторое множество объектов $A = \{a_1,...,a_m\}$. Матрицей близости для этих объектов называется матрица $(r_{ij}), i, j = 1,...,m$; r_{ij} – числовые оценки меры близости (сходства или различия) в порядковой шкале (имеет смысл только сравнение величин $r_{i_1 j_1} < r_{i_2 j_2}$). Такие матрицы возникают в различных областях при сравнении или оценке экспертом двух объектов в некотором отношении.

Матрицы близости обрабатываются методами многомерного неметрического шкалирования [15 – 16]. Целью этих методов является представление объектов точками в некотором метрическом пространстве (евклидовом или римановом) минимальной размерности так, чтобы расстояния t_{ij} между ними с точностью до порядка соответствовали величинам r_{ij}. После применения методов многомерного шкалирования мы получаем представление данных в метрическом пространстве.

Определим на множестве пар $A^* \subseteq A \times A$, бинарное отношение упорядочения:

$(a_{i_1}, a_{j_1}) \le (a_{i_2}, a_{j_2}) \Leftrightarrow r_{i_1 j_1} < r_{i_2 j_2}$

Информация, содержащаяся в матрице близости, может быть задана эмпирической системой $\Im = \langle A^*, \leq \rangle$, где $A^* \subseteq A \times A$, $\leq$ – бинарное отношение упорядочения на A^*.

Приведём некоторые результаты теории измерений, относящиеся к таким эмпирическим системам.

Шкала положительных разностей, определяемая системой аксиом *S* [8; с. 147]. Если система аксиом *S* выполнена на эмпирической системе $\Im = \langle A^*, \leq \rangle$, то существует гомоморфизм $\phi : A^* \to Re$, $A \neq \varnothing$, такой, что для любых $(a, b), (b, c), (c, d) \in A^*$:

1) $(a,b) \leq (c,d) \Leftrightarrow \phi(a,b) \leq \phi(c,d)$;

2) $\phi(a, c) = \phi(a, b) + \phi(b, c)$.

Отображение ϕ единственно с точностью до положительного множителя (шкала отношений).

Шкала алгебраических разностей [8; с. 151]: Определяется системой аксиом *S* эмпирической системы $\Im = \langle A^*, \leq \rangle$, $A^* = A \times A$ такой, что, если она выполнена, то существует гомоморфизм $\phi : A \to Re$, удовлетворяющий для любых $a,b,c,d \in A$ условию:

$(a,b) < (c,d) \Leftrightarrow \phi(a) - \phi(b) < \phi(c) - \phi(d)$.

Отображение ϕ единственно с точностью до лог-линейных преобразований (шкала интервалов).

Шкала разностей равных конечных промежутков [8; с. 168]. Определяется системой аксиом S эмпирической системы $\Im = \langle A^*, \leq \rangle$, $A^* = A \times A$, A – конечно, $A^* \neq \varnothing$, для которой существует гомоморфизм $\phi : A \to N$ в натуральные числа, такой, что для любых $a,b,c,d \in$ A:

$(a,b) \leq (c,d) \Leftrightarrow \phi(a) - \phi(b) \leq \phi(c) - \phi(d)$.

Отображение ϕ единственно с точностью до линейных преобразований (шкала интервалов).

Шкала абсолютных разностей: [8; с. 172]. Определяется системой аксиом *S* эмпирической системы $\Im = \langle A^*, \leq \rangle$, $A^* = A \times A$ для которой существует гомоморфизм $\phi : A \to Re$ такой, что:

$(a,b) < (c,d) \Leftrightarrow |\phi(a) - \phi(b)| < |\phi(c) - \phi(d)|$.

Отображение ϕ единственно с точностью до линейных преобразований (шкала интервалов).

Матрица объект-признак представляет собой матрицу (x_{ij}), $i = 1,...,m, j = 1,...,n$, где $x_{ij} = x_j(a_i)$ – числовые значения *j*-го признака x_j на *i*-ом объекте. Признаки могут быть любыми: количественными, качественными, ранговыми и т.д. Тот факт, что такая матрица получена в результате некоторых измерений (опросов, экспериментов, обследований и т. д.), говорит о том, что существует *n* измерительных процедур x_j. Такие измерения называют приборными или косвенными измерениями. Рассмотрим, как можно определить эмпирическую систему приборных измерений.

Для каждого прибора x_j и некоторого числового отношения $R(y_1, \ldots, y_k)$, определённого на *Re*, можно определить следующее эмпирическое отношение на множестве объектов *A*:

$P_j^R(a_1,...,a_k) \Leftrightarrow R(x_j(a_1),...,x_j(a_k))$.

Это отношение (или операция) может не иметь эмпирической интерпретации, например, нельзя складывать метры и килограммы. Прибор x_j имеет эмпирическую интерпретацию,

но отношение *R*, определённое на нём, может уже не иметь эмпирической интерпретации. Поэтому нужно найти такие числовые отношения $R(y_1,...,y_k)$ на *Re*, для которых отношение P_j^R интерпретируемо. Предположим, что мы перебрали некоторые, наиболее распространённые числовые отношения (и операции) и нашли множество $\{P_j^{R_1},...,P_j^{R_k}\}$ интерпретируемых отношений для приборного измерения x_j. Оно не пусто, так как, по крайней мере, отношение $P_j^{=}(a_1,a_2) \Leftrightarrow x_j(a_1) = x_j(a_2)$ имеет эмпирическую интерпретацию, состоящую в том, что на объектах a_1 и a_2 величина x_j принимает одно и то же значение. Отношение $P_j^{=}$, как правило, является отношением эквивалентности. В теории измерений известно много систем аксиом, использующих только отношение эквивалентности и приводящих, тем не менее, к сильным шкалам.

Таким образом, эмпирической системой для матрицы объект-признак будет система $\Im = \langle A, P_1^{R_1^1},...,P_1^{R_{k_1}^1},...,P_1^{R_1^n},...,P_n^{R_{k_n}^n}\rangle$, включающая объединение всех отношений для всех приборных измерений x_j.

1.5. Реляционный подход к извлечению знаний из информации, содержащейся в данных

В существующих методах KDD&DM&ML информация, извлекаемая из данных, явно не выделяется. Кроме того, эта информация может быть зафиксирована в виде эмпирических систем, которые требуют язык логики первого порядка. В настоящее время только методы, разрабатываемые в рамках направлений Probabilistic Inductive Logic Programming и Rule extraction, могут работать с такой информацией.

Поэтому нами разработан оригинальный подход (Relational Data Mining, [3 – 6]), ориентированный на обнаружение знаний путём прямого использования информации, извлечённой из данных и представленной эмпирическими системами в логике первого порядка. Он состоит в том, чтобы:

1) представить предметную область (решаемую задачу) эмпирической системой $\Im = \langle A, \Omega\rangle$, где *A* – объекты ПО (выборка), а Ω – множество отношений и операций, интерпретируемых в системе понятий ПО;
2) извлечь информацию из имеющихся данных *D* и представить её эмпирической системой $\Im_D = \langle D, \Omega_D\rangle$, являющейся подсистемой эмпирической системы $\Im = \langle A, \Omega\rangle$;
3) сформулировать проверяемые гипотезы в языке Ω_D интерпретируемых в терминах ПО отношений и операций на данных;
4) проверить эти гипотезы на эмпирической системе $\Im_D = \langle D, \Omega_D\rangle$ и получить закономерности на этих данных как множество подтвердившихся на этих данных гипотез. Эти закономерности и будут тем знанием, которое проверено на данных и интерпретируемо с онтологии ПО.

Данный подход обладает следующими преимуществами перед известными подходами в KDD&DM&ML:

1) может обрабатывать данные, содержащие величины различных типов:
   a. различного рода отношения: предпочтения, частичного порядка, решётки, древовидные структуры, сети, графы и т.д.;
   b. величины, измеренные в различных шкалах: наименований, порядка, отношений и т.д.;
   c. использовать не только сами данные, как они заданы, а также различного рода преобразованные величины (вторичные признаки), в соответствии с тем, что для пользователя имеет смысл и интерес;
2) использовать не сами данные, а только то, что интересует пользователя в данных. Например, никто не использует стоимости ценных бумаг напрямую, существуют сотни различного рода индексов, которые вычисляются по этим стоимостям и которые используются для принятия решений о сделках;
3) может обнаруживать и проверять любые классы гипотез, представленных в логике первого порядка, в терминах извлечённой из данных информации;
4) можно показать [3], что в результате проверки гипотез и обнаружения закономерностей в рамках данного подхода можно обнаружить:
   a. теорию предметной области;
   b. правила, имеющие максимальные условные вероятности;
   c. непротиворечивую вероятностную аппроксимацию теории предметной области;
   d. максимально специфические правила, позволяющие предсказывать без противоречий.

Разработана достаточно «универсальная» версия программной системы «Discovery», реализующей данный подход и позволяющая пользователю самому задавать класс обнаруживаемых закономерностей, извлекать из данных множество закономерностей заданного класса и использовать найденные закономерности для прогноза и принятия решений.

Тем не менее, в разработанной версии системы «Discovery» предполагается, что данные уже заданы в виде эмпирической системы, либо есть некоторое множество отношений и операций, которые мы можем определять на данных. До настоящего времени не было самостоятельного интерфейса извлечения информации из данных. Система Visual Discovery разработана с целью устранения этого пробела.

1.6. Применения реляционного подхода

Реляционный подход наиболее широко применялся для решения задач финансового прогнозирования [3, 6, 17], медицины [18 – 19] и биоинформатики [21 – 23]. В каждой из этих задач, в соответствии с реляционным подходом, нужна была настройка на предметную область. Это приводило к тому, что множество отношений и операций в каждой из этих задач были разные. Когда эти задачи решались, то не было возможности программно, тем более визуально, как это реализовано в системе VisualDiscovery, конструировать соответствующие отношения, операции и предикаты, поэтому программа каждый раз переписывалась заново.

Приведём кратко отношения и операции, которые использовались в этих задачах, а также наиболее важные классы гипотез, которые формулировались в их терминах и проверялись на данных.

Финансовое прогнозирование [3, 6, 17]. Мы использовали следующие отношения и операции, определяемые для финансового временно́го ряда – индекса SP500C:

а) первая разность $\Delta_{ij}(a_t) = (\mathrm{SP500C}(a_t^j) - \mathrm{SP500C}(a_t^i)) / \mathrm{SP500C}(a_t^i),\ i < j,\ i, j = 1,...,5$

Эта функция представляет собой разность между SP500C для *i*-го и *j*-го дней рассматриваемого пятидневного промежутка, нормализованных относительно SP500C для *i*-го дня;

b) вторая разность $\Delta_{ijk}(a_t) = \Delta_{jk}(a_t) - \Delta_{ij}(a_t)$, являющаяся разностью между двумя первыми разностями;

c) функция $\mathrm{wd}(a) = \langle 1,2,3,4,5 \rangle$, отображающая пять календарных дней в числа. Эта запись означает, что a представляет собой пять последовательных дней недели с понедельника по пятницу. Запись $\mathrm{wd}(a) = \langle d_1,...,d_5 \rangle, d_1,...,d_5 \in \{1,2,3,4,5\}$ будет означать циклическую перестановку чисел $1,...,5$;

d) $\mathrm{goal}(a)$ – целевое значение, которое надо научиться предсказывать, является изменением значения индекса SP500C за последующие пять дней, по сравнению с последним днём пятидневки a;

e) будем рассматривать всевозможные сравнения $(\Delta(a) \le \Delta(b))$ между первыми и вторыми разностями, где дельта означает любую из первых или вторых разностей. Разность может быть с отрицанием $(\Delta(a) \le \Delta(b))^{\varepsilon_1}, \varepsilon_1 = 0$ или без него $(\Delta(a) \le \Delta(b))^{\varepsilon_1}, \varepsilon_1 = 1$;

f) для целевого значения также анализируется сравнение $(\mathrm{goal}(a) \le \mathrm{goal}(b))^{\varepsilon_0}$ значений относительно двух разных пятидневок a,b.

Для анализа индекса SP500C использовался следующий класс гипотез:

$$\mathrm{wd}(a) = \mathrm{wd}(b) = \langle d_1,...,d_5 \rangle \,\&\, (\Delta(a) \le \Delta(b))^{\varepsilon_1} \,\&\, ... \,\&\, (\Delta(a) \le \Delta(b))^{\varepsilon_k} \Rightarrow (\mathrm{goal}(a) \le \mathrm{goal}(b))^{\varepsilon_0}$$

Примером обнаруженной закономерности является следующее высказывание:

«**Если** текущая пятидневка a индекса SP500C заканчивается в понедельник и существует пятидневка b в истории (1984 – 1996 гг.), которая также заканчивается в понедельник,

И относительная разность между вторником и четвергом для пятидневки b не больше, чем такая же разность (между вторником и четвергом) для текущей пятидневки,

И относительная разность между вторником и понедельником для пятидневки b строго больше, чем такая же разность для текущей пятидневки,

И вторая разность между вторником, средой и четвергом для пятидневки b не больше, чем такая же разность для текущей пятидневки,

И вторая разность между вторником, четвергом и пятницей для пятидневки b строго больше, чем такая же разность для текущей пятидневки,

ТО индекс SP500C в следующий понедельник (через пять дней по сравнению с текущей пятидневкой) вырастет не больше, чем в следующий понедельник по отношению к пятидневке b».

Разработка диагностической системы рака груди [18 – 20]. В приложениях по разработке диагностической системы рака груди использовались различные признаки, определённые экспертом. Они включали в себя количественные, ранговые, номинальные и булевы признаки. Обнаруженные закономерности включали в себя сочетания этих признаков, а также простейшие интерпретируемые отношения на них.

Примеры обнаруженных закономерностей:

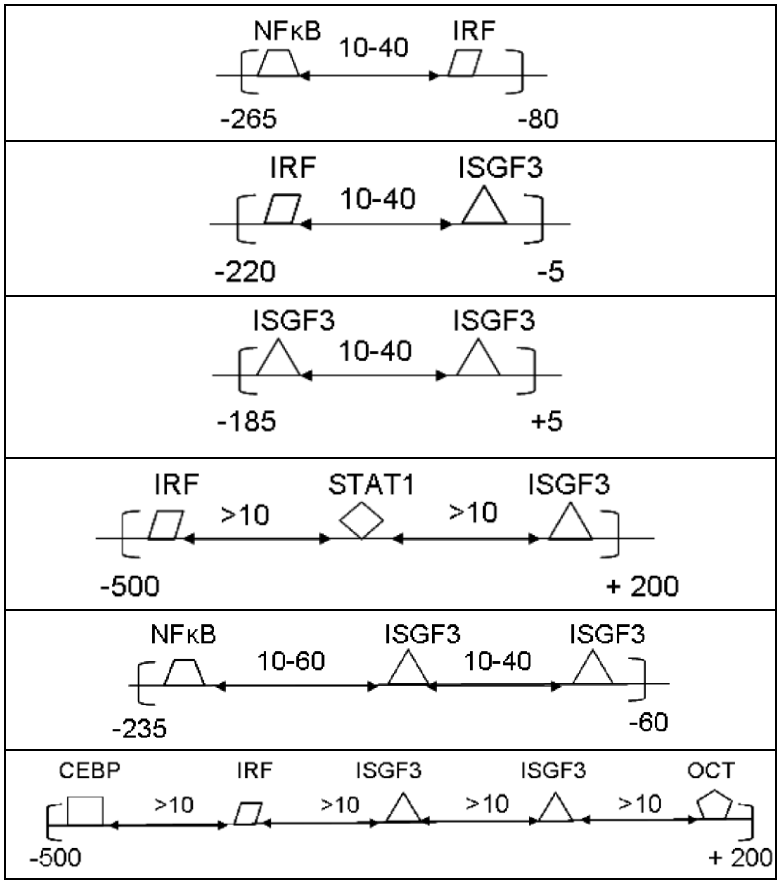
ЕСЛИ количество кальцинозов в см^2 между 10 и 20
И объём опухоли больше 5 см^3,
ТО подозрение на злокачественное развитие с оценкой вероятности 93%.

ЕСЛИ общее количество кальцинозов больше 30
И объём опухоли больше 5 см^3
И плотность кальциноза средняя,
ТО подозрение на злокачественное развитие с оценкой вероятности около 100%.

ЕСЛИ вариации в форме кальциноза значительны
И количество кальцинозов в см^2 между 10 и 20
И нерегулярность в форме кальциноза средняя,
ТО подозрение на злокачественное развитие с оценкой вероятности около 100%.

Приложения в биоинформатике. Для решения задачи анализа последовательностей регуляторных районов генов понадобилось определение гипотез в виде структур – комплексных сигналов, обнаруживаемых на генетических последовательностях. Мы не будем приводить формальное определение комплексных сигналов, ограничимся примерами найденных сигналов.

Таблица 1. Комплексные сигналы, обнаруженные в cis-regulatory modules интерферон индуцированных генов. Разными фигурами обозначены и названы сайты связывания транскрипционных факторов, которые в указанном порядке и расположении должны находиться в промоторных районах интерферон индуцированных генов



2. Инструментальное средство Visual Discovery

Инструментальное средство Visual Discovery, позволяет специалисту ПО работать с онтологией ПО, извлекать информацию из данных, вручную формировать шаблоны предикатов и операций над данными (онтологию) и классы гипотез. Тем самым Visual Discovery позволяет реализовать возможности описанного выше реляционного подхода с учётом удобства пользователей.

В соответствии с порядком работы в рамках реляционного подхода, описанного в 1.5, решение задачи системой Visual Discovery состоит из следующих шагов:

1. Выбрать исходные данные – объекты/признаки;
2. Задать онтологию ПО в виде шаблонов предикатов;
3. На основе онтологии и исходных данных сформировать класс проверяемых гипотез;
4. Задать основные параметры работы системы;
5. Получить найденные закономерности;

6. Проинтерпретировать найденные закономерности и, тем самым, получить результат.

Рассмотрим последовательно порядок работы на каждом шаге.

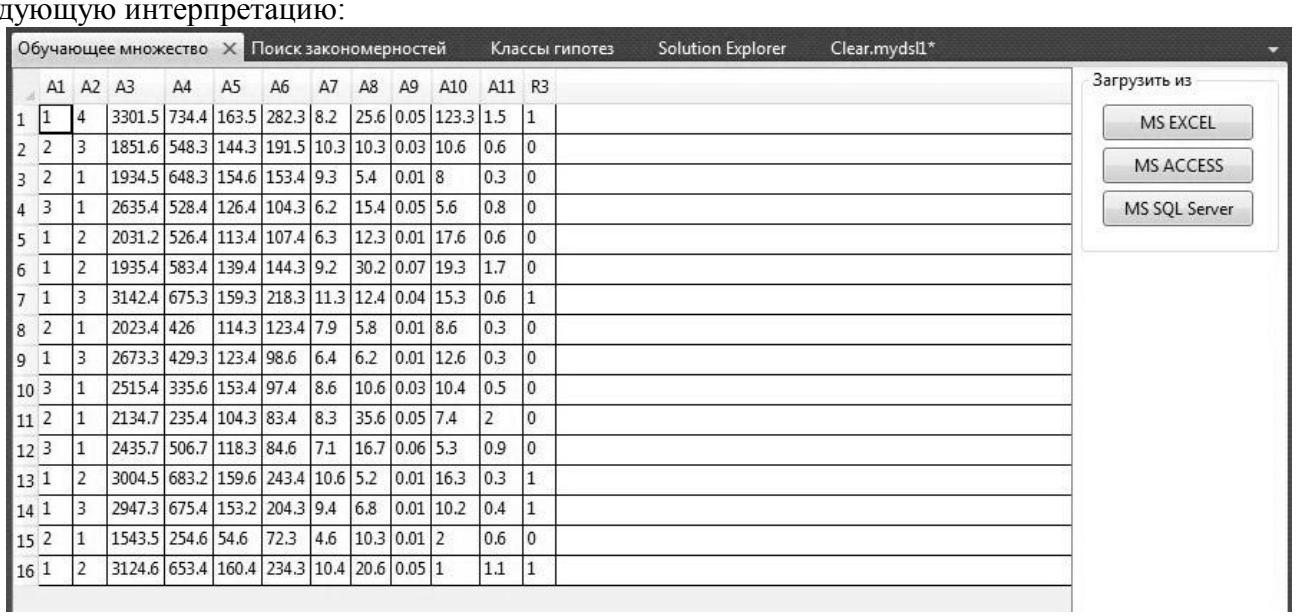
2.2. Выбор исходных данных

Данные в системе Visual Discovery представляются виде таблицы, объекты в которой представлены строками, а признаки – столбцами. Для простоты работу системы проиллюстрируем на следующем конкретном примере. Данные были получены путём анализа распространения медицинских препаратов. Необходимо найти закономерности между признаками, характеризующие максимальный товарооборот на одного специалиста.

Система позволяет загружать данные из следующих источников данных:

1. MS Excel. Данные берутся из выбранной таблицы пользователем при открытии файла;
2. MS Access. Данные извлекаются с помощью соответствующего SQL запроса из файла;
3. MS SQL Server. Данные извлекаются из БД сервера с помощью соответствующего SQL запроса.

Для поставленной задачи данные были внесены в Excel файл. Признаки $A_1,...,A_{11}$ имеют следующую интерпретацию:


Обучающее множество | Поиск закономерностей | Классы гипотез | Solution Explorer | Clear.mydsl1*

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	R3
1	1	4	3301.5	734.4	163.5	282.3	8.2	25.6	0.05	123.3	1.5	1
2	2	3	1851.6	548.3	144.3	191.5	10.3	10.3	0.03	10.6	0.6	0
3	2	1	1934.5	648.3	154.6	153.4	9.3	5.4	0.01	8	0.3	0
4	3	1	2635.4	528.4	126.4	104.3	6.2	15.4	0.05	5.6	0.8	0
5	1	2	2031.2	526.4	113.4	107.4	6.3	12.3	0.01	17.6	0.6	0
6	1	2	1935.4	583.4	139.4	144.3	9.2	30.2	0.07	19.3	1.7	0
7	1	3	3142.4	675.3	159.3	218.3	11.3	12.4	0.04	15.3	0.6	1
8	2	1	2023.4	426	114.3	123.4	7.9	5.8	0.01	8.6	0.3	0
9	1	3	2673.3	429.3	123.4	98.6	6.4	6.2	0.01	12.6	0.3	0
10	3	1	2515.4	335.6	153.4	97.4	8.6	10.6	0.03	10.4	0.5	0
11	2	1	2134.7	235.4	104.3	83.4	8.3	35.6	0.05	7.4	2	0
12	3	1	2435.7	506.7	118.3	84.6	7.1	16.7	0.06	5.3	0.9	0
13	1	2	3004.5	683.2	159.6	243.4	10.6	5.2	0.01	16.3	0.3	1
14	1	3	2947.3	675.4	153.2	204.3	9.4	6.8	0.01	10.2	0.4	1
15	2	1	1543.5	254.6	54.6	72.3	4.6	10.3	0.01	2	0.6	0
16	1	2	3124.6	653.4	160.4	234.3	10.4	20.6	0.05	1	1.1	1

Рис. 1. Исходная таблица объектов/признаков в Visual Discovery

1. A_1, A_2 – организационная культура;
2. A_3 – товарооборот на 1 специалиста;
3. A_4 – средний доход;
4. A_5 – прибыль на 1 специалиста;
5. A_6 – издержки на 1 специалиста;
6. A_7 – доля издержек на весь персонал;
7. A_8 – среднее число профессионального обучения, часов;
8. A_9 – доля издержек на обучение;
9. A_{10} – коэффициент текучести;
10. A_{11} – доля часов на обучение в общем балансе времени.

2.3. Задание онтологии

Отличительной особенностью Visual Discovery от других систем интеллектуального анализа данных является графическая модель задания онтологии (задания отношений и операций на исходных данных).

Самая сложная часть работы специалиста предметной области сводится к заданию онтологии для информации, извлекаемой из данных путём создания диаграммы шаблонов предикатов в интуитивно понятном графическом интерфейсе. Эта задача решается визуальным интерфейсом системы Visual Discovery рис. 4.

Шаблоны предикатов могут быть заданы двумя способами:

1. Загружены из файла;
2. Созданы непосредственно в системе путём создания диаграммы шаблонов предикатов;
3. Получены системой путём решения одной из 3-х задач (п. 4).

Диаграмма шаблонов предикатов разбита на три поля, которые содержат:

- Шаблоны предикатов;
- Функции от переменных;
- Исходные данные.

На поле «Шаблоны предикатов» помещаются предикаты с соответствующими термами и отношениями между ними. В данный момент поддерживаются следующее множество отношений, соответствующее шкале порядка:

- отношения сравнения $=$, $<$, $>$, $\leq$, $\geq$;
- отношения принадлежности предиката к множеству или интервалу значений $T_1 \in (T_2;T_3)$, $T_1 \in [T_2;T_3)$, $T_1 \in (T_2;T_3]$, $T_1 \in [T_2;T_3]$;
- Внешние предикаты и функций в виде модулей, подключаемых пользователем.

На поле «Функции от переменных» помещаются функции, которые являются интерпретацией термов из поля «Шаблоны предикатов». Функции определяют переменные и операции над ними. Функция может быть задана арифметическим выражением или любой другой математической функцией.

На поле «Исходные данные» помещаются признаки объектов или константы, на которые ссылаются переменные из функций или термов предикатов.

Помимо того, что пользователю даётся возможность самостоятельно конструировать проверяемые гипотезы, пользователь может использовать абсолютно любые предикаты и функции, не ограничиваясь уже встроенными в систему. Предикаты и функции можно взять готовыми или создать в виде подключаемых модулей к системе Visual Discovery. Один модуль соответствует одному предикату или функции. Таким образом, пользователь не ограничен возможностями системы для задания онтологии ПО.

Модуль являются программой, написанной на любом языке программирования, которую может запускать операционная система, в которой работает система Visual Discovery. Модуль должен принимать параметры на входе и заканчивать свою работу с возвратом подсчитанного значения.

Пример исходного кода модуля сравнения двух чисел на языке C#:

```
using System;
namespace VisualDiscoveryModules
{
   class GreateThen
   {
      public static double Main(string[] args)
          {
              if (args.Length<2) return 0;
```

```
            double a = Convert.ToDouble(args[0]);
            double b = Convert.ToDouble(args[1]);
            if (a>b) return a;
            return b;
        }
    }
}
```

Для того чтобы подключить модуль к системе Visual Discovery, необходимо в окне «Внешние операции» (рис. 2) добавить файл модуля, например GreateThen.exe , после чего он появится в списке подключенных модулей.

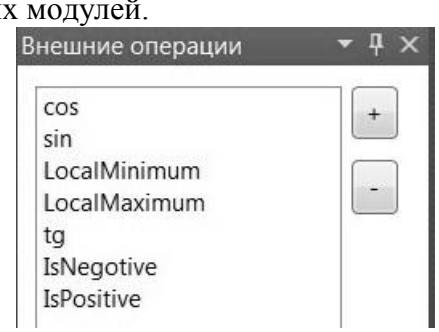


Рис. 2. Окно «Внешние операции» для подключения внешних операций и предикатов

После того как модуль подключен, он может быть использован в диаграмме построения шаблонов предикатов и конструкторе гипотез.

Для использования модуля необходимо в формуле функции или предиката указать название модуля в фигурных скобках вместе с параметрами через точку с запятой (рис.3).

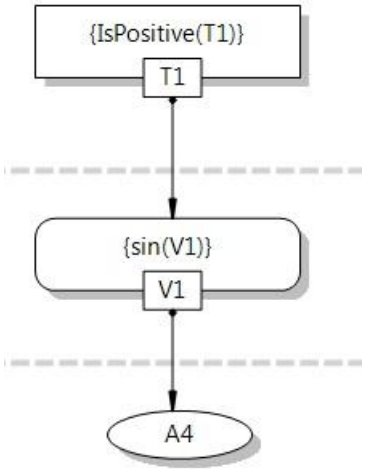


Рис. 3. Использование подключенных модулей при построении шаблонов предикатов

Для решения поставленной задачи была создана диаграмма шаблонов предикатов, представленная на рис. 4. Каждый признак был разбит на интервалы некоторым алгоритмом, выделяющим «сгустки» значений, и по этим интервалам были сформированы предикаты, отвечающие за принадлежность признака некоторому интервалу значений.

Например, признак A_3 (товарооборот на 1 специалиста) был разбит на три интервала:

1. $A_3 \in [1851.6; 2023.4)$;
2. $A_3 \in [2435.7; 2947.3)$;
3. $A_3 \geq 2947.3$.

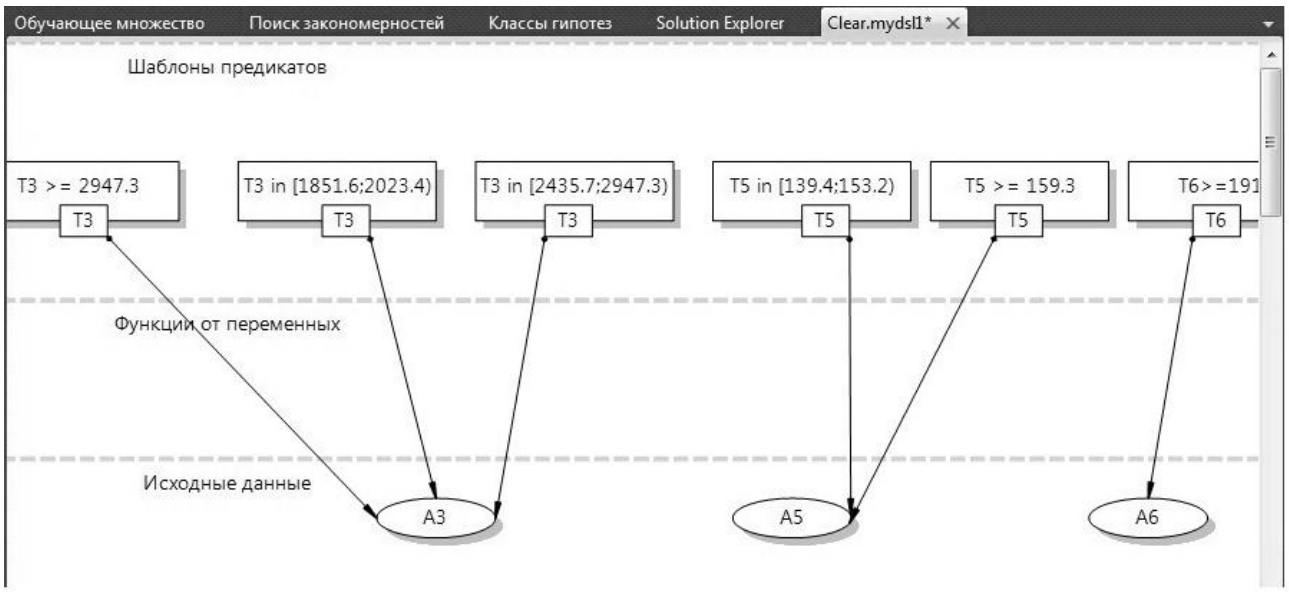


Рис. 4. Формирование шаблонов предикатов

Соответствующие шаблоны предикатов для признака A_3 введены в раздел «Шаблоны предикатов» (рис.4). Каждый предикат интерпретируется в некотором признаке и связывается с ним направленной стрелкой. Аналогично для признаков A_5 и A_6 выбираем другие шаблоны. Интервалы значений могут находиться автоматически программой, вводиться пользователем в соответствии с интерпретацией признаков, редактироваться и удаляться. Шаблоны предикатов фиксируют информацию, извлекаемую из этих признаков. Гипотезы будет формироваться с использованием только этой информации. Поэтому нужно определить столько шаблонов предикатов, сколько нужно для выражения всей интересующей нас информации.

2.4. Формирование классов гипотез

Гипотезы задаются на основе шаблонов предикатов и определяют то знание, которое мы хотели бы получить в результате анализа данных. Гипотезы, и тем самым будущее знание, задаётся правилами, содержащими посылки и следствия. Задание гипотез также осуществляется визуально и представлено на рис.5.

В нашей задаче классы гипотез задаются по шаблонам предикатов и целевому предикату, выбранному из шаблонов предикатов. Например, целевой предикат $A_3 \in [2435.7; 2947.3)$ выбирается из множества предикатов, задающихся шаблоном предиката $T_3 \in [2435.7; 2947.3)$.

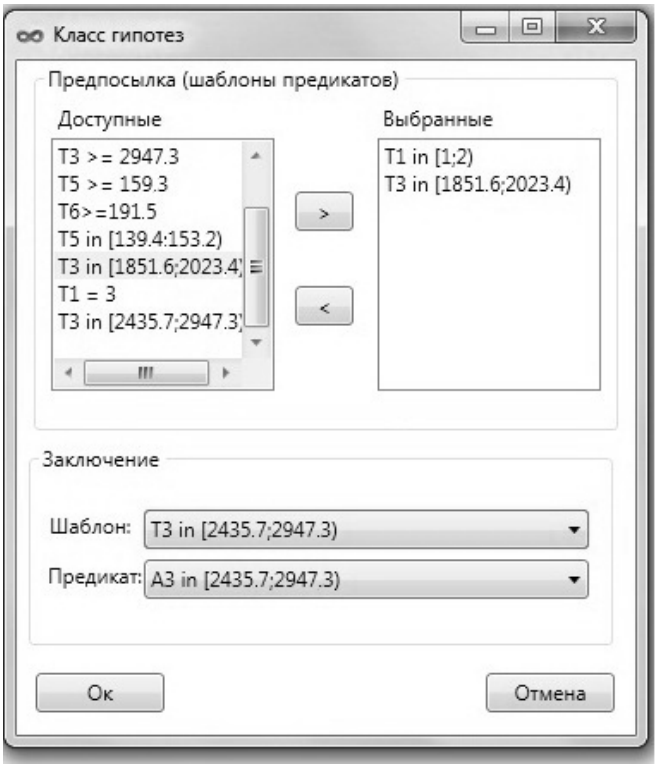


Рис. 5. Формирование класса гипотез

В нашей задаче были сформированы классы гипотез, представленные на рис.6.

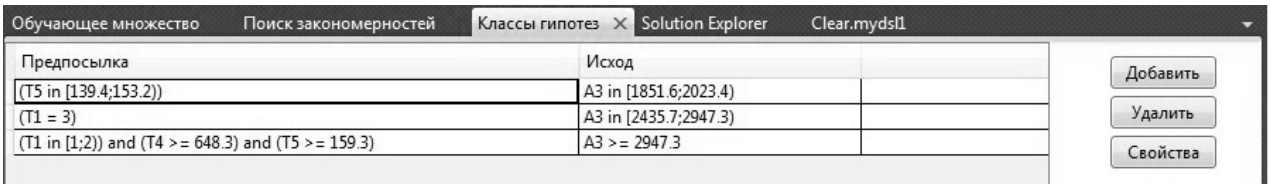


Рис. 6. Заданные классы гипотез

2.5. Проверка классов гипотез

Далее сформированные классы закономерностей проходят проверку на данных. Если какая-то гипотеза подтверждается на данных, пройдя серию статистических тестов, то она фиксируется как закономерность в данных и выдаётся программной системой Visual Discovery. Для обнаружения закономерностей программной системой Visual Discovery нужно задать следующие параметры [3, 6]:

- доверительный уровень для критерия Фишера;
- доверительный уровень для критерия Юла;
- количество объектов обучения;
- глубина базового перебора.

После чего были получены результаты, представленные на рис. 7. На этом рисунке:

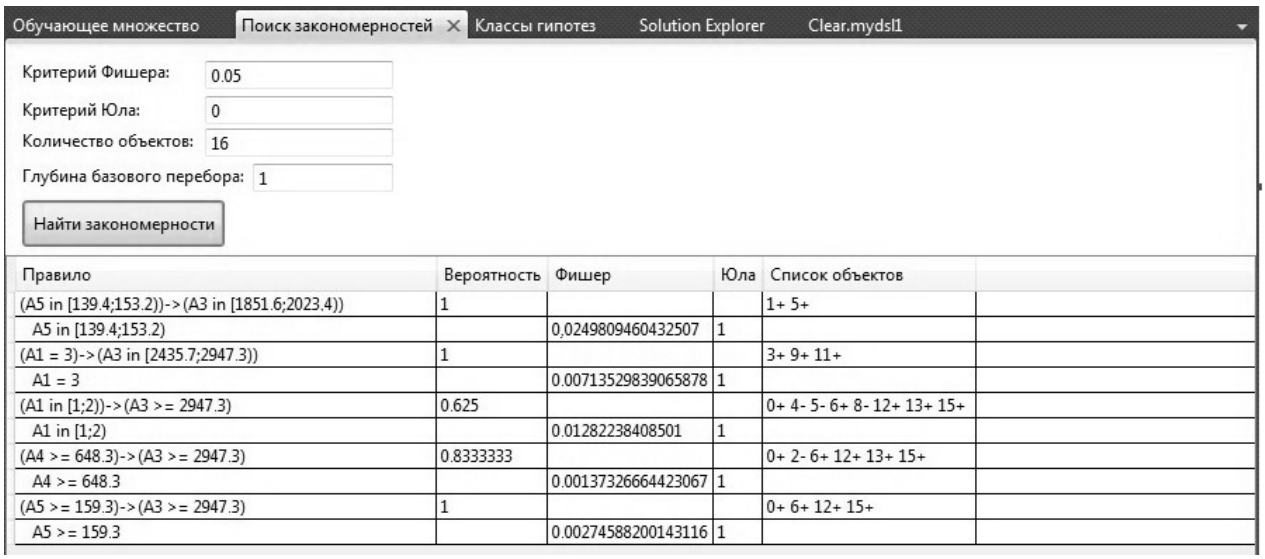
В поле «Правило» записаны закономерности, связывающие признаки объектов.

В поле «Вероятность» приведена условная вероятность правила.

В поле «Фишер» указан критерий Фишера для предиката, содержащегося в правиле.

В поле «Юла» указан критерий Юла для предиката, содержащего в правиле.

В поле «Список объектов» приведены номера объектов, на которых выполняется правило.



Правило	Вероятность	Фишер	Юла	Список объектов
(A5 in [139.4;153.2))->(A3 in [1851.6;2023.4))	1			1+ 5+
A5 in [139.4;153.2)		0,0249809460432507	1	
(A1 = 3)->(A3 in [2435.7;2947.3))	1			3+ 9+ 11+
A1 = 3		0.00713529839065878	1	
(A1 in [1;2))->(A3 >= 2947.3)	0.625			0+ 4- 5- 6+ 8- 12+ 13+ 15+
A1 in [1;2)		0.01282238408501	1	
(A4 >= 648.3)->(A3 >= 2947.3)	0.8333333			0+ 2- 6+ 12+ 13+ 15+
A4 >= 648.3		0.00137326664423067	1	
(A5 >= 159.3)->(A3 >= 2947.3)	1			0+ 6+ 12+ 15+
A5 >= 159.3		0.00274588200143116	1	

Рис. 7. Полученные результаты

Знак «+» после номера объекта означает положительный исход правила, знак «-» означает отрицательный исход соответственно.

2.6. Получение результатов

Закономерности на рис.7, полученные в системе Visual Discovery, были проинтерпретированы и проанализированы специалистом, и в результате были сделаны следующие заключения о связи между признаками:

- Из закономерности $(A_1 \in [1;2]) \Rightarrow (A_3 \geq 2947.3)$ следует, что между типом организационной культуры и величиной товарооборота существует закономерность. Так, максимальный товарооборот 2950 – 3300 тыс. руб. на одного человека в год обеспечивает рыночная культура, ориентированная на стабильность. Иерархическая культура обеспечивает товарооборот на одного специалиста на уровне 2435 – 2950 тыс. руб.
- Из закономерности $(A_5 \geq 159.3) \Rightarrow (A_3 \geq 2947.3)$ следует, что между максимальным товарооборотом и прибылью на одного человека существует закономерность. Максимальный товарооборот 2950 – 3300 тыс. руб. на одного человека в год обеспечивает максимальную прибыль – от 159 до 163.5 тыс. руб. на одного специалиста. При товарообороте на уровне 1850 – 2020 тыс. руб. максимальная прибыль составит 139 – 153 тыс. руб. на одного специалиста.
- Из закономерности $(A_4 \geq 648.3) \Rightarrow (A_3 \geq 2947.3)$ следует связь между средним доходом и максимальным товарооборотом. Максимальный средний доход от 648.3 до 734 тыс. руб. обеспечивается при максимальном товарообороте 2950 – 3300 тыс. руб. на одного специалиста.

Литература

1. *Фридланд А.Я.* Информатика: процессы, системы, ресурсы. М.: БИНОМ. Лаборатория знаний, 2003.
2. *Бешенков С.А., Ракитина Е.А.* Моделирование и формализация. Методическое пособие. М.: Лаборатория Базовых Знаний, 2002. 336с.
3. *Витяев Е. Е.* Извлечение знаний из данных. Компьютерное познание. Модели когнитивных процессов. Новосибирск, 2006. 293с.

4. *Витяев Е.Е.* Извлечение информации из данных // Информационные технологии в гуманитарных исследованиях, Вып. 15, ИАЭТ СО РАН, Новосибирск, 2010, 9-16.
5. *E. Vityaev, B.Y. Kovalerchuk*, Relational Methodology for Data Mining and Knowledge Discovery // Intelligent Data Analysis. Special issue on «Philosophies and Methodologies for Knowledge Discovery and Intelligent Data Analysis» eds. Keith Rennolls, Evgenii Vityaev. v.12(2), IOS Press, 2008, pp. 189-210.
6. *Kovalerchuk B., Vityaev E.* Data Mining in Finance: Advances in Relational and Hybrid methods. (Kluwer international series in engineering and computer science; SECS 547), Kluwer Academic Publishers, 2000, p.308.
7. Лихорадка // Малая медицинская энциклопедия, М.
8. *Krantz D.H., Luce R.D., Suppes P., Tversky A.* Foundations of Measurement. Acad. Press, N.Y.; L. 1971; 1989; 1990. Vol. 1–3.
9. *Пфанцагль И.* Теория измерений. М.: Мир, 1976. 248 с.
10. *Сатаров Г. А., Каменский В. С.* Общий подход к анализу экспертных оценок методами не метрического многомерного шкалирования // Статистические методы анализа экспертных оценок. М.: Наука, 1977. С. 251–266.
11. *Миркин Б.Г.* Анализ качественных признаков и структур. М.: Статистика, 1980. 316 с.
12. *Шмерлинг Д. С.* О построении моделей парных и множественных сравнений со связями // Прикладной многомерный статистический анализ. М., 1978. С. 164–189.
13. *Шрейдер С. А.* Систематика, типологии, классификация // Теория и методология биологических классификаций, М.: Наука, 1983.
14. *Фишберн П.С.* Теория полезности для принятия решений. М.: Наука, 1978. 352 с.
15. *Девид Г.* Метод парных сравнений. М.: Статистика, 1978. 150 с.
16. *Терехина А. Ю.* Методы многомерного шкалирования и визуализации данных // Автоматика и телемеханика. 1973. № 7. С. 80–94.
17. *Демин А.В., Витяев Е.Е.* Реализация универсальной системы извлечения знаний «Discovery» и её применение в задачах финансового прогнозирования. Информационные технологии работы со знаниями: обнаружение, поиск, управление (Вычислительные системы, вып. 175), Новосибирск, 2008, стр. 3-47.
18. *Витяев Е.Е., Ковалерчук Б.Я.*, Методика извлечения знаний из эксперта // Информационные технологии в гуманитарных исследованиях, Вып. 13, ИАЭТ СО РАН, Новосибирск, 2009, стр. 75-81
19. *Kovalerchuk, B., Vityaev E., Ruiz J.F.*, Consistent and Complete Data and «Expert» Mining in Medicine. In: Medical Data Mining and Knowledge Discovery, Springer, 2001, pp. 238-280.
20. *B. Kovalerchuk, E. Vityaev, J. Ruiz*. Consistent knowledge discovery in medical diagnosis. Special issue of the journal: IEEE Engineering in Medicine and Biology Magazine: «Medical Data Mining», 2000, pp.26-37.
21. *Хомичева И. В., Витяев Е.Е., Игнатьева Е.В., Ананько А.Е., Шипилов Т.И.* Применение программной системы ExpertDiscovery для поиска закономерностей структурно-функциональной организации регуляторных районов генов. Вестник НГУ, серия: Информационные технологии, Т. 8, вып. 1, Новосибирск, 2010, стр. 12-26, 2010.
22. *Vityaev E.E., Lapardin K.A., Khomicheva I.V., Proskura A.,L.* Transcription factor binding site recognition by regularity matrices based on the natural classification method. Intelligent Data Analysis, Special issue on «Machine learning and bioinformatics» eds. Nikolai Kolchanov, Evgenii Vityaev. v.12(5), IOS Press, 2008, pp. 495-512.
23. *E.E. Vityaev, T.I. Shipilov, M.A. Pozdnyakov, O.V. Vishnevsky, A.L., Proscura, Yu.L. Orlov, P. Arrigo* Software for analysis of gene regulatory sequences by knowledge discovery methods. In: Bioinformatics of Genome Regulation and Structure II. (Eds. N.Kolchanov and R. Hofestaedt) Springer Science+Business Media, Inc. 2006, pp. 491-498.

Статья поступила в редакцию 30.11.2012

Подберезный Андрей Александрович
аспирант кафедры вычислительных систем СибГУТИ
тел. моб. +7(923)243-90-30, e-mail: andreal@inbox.ru

Витяев Евгений Евгеньевич
д.ф.-м.н., профессор кафедры дискретной математики и информатики НГУ,
в.н.с., Института математики им. С.Л.Соболева СО РАН, тел. (383) 363-4562,
e-mail: vityaev@math.nsc.ru.

Москвитин Анатолий Алексеевич
д.ф.-м.н., профессор кафедры вычислительных систем СибГУТИ
тел. служ. (383) 269-82-75, e-mail: moskvit47@mail.ru

«Visual Discovery» instrument for intelligent data analysis problem solving

Podbereznyy A.A., Vityaev E.E., Moskvin A.A.,

In this paper, a short description of the original relational approach to intelligent data analysis is presented and the program system «Visual Discovery» implementing this approach with maximal ease for users is described. The main advantage of the system is visual constructor of data relations and operations being interpreted in the ontology of domain, and also hypotheses constructor being checked up on data.

Keywords: intelligent data analysis, artificial intelligence, data mining, knowledge discovery in data bases.