

**Витяев Е.Е.,**      Инструментальное средство Visual Discovery извлечения ин-  
**Москвитин А.А.**      формации из данных и решения задач интеллектуального  
**Подберезный А.А.**      анализа данных<sup>1</sup>

В предыдущей работе, посвященной извлечению информации из данных, аргументировалось, что для получения осмысленного и интерпретируемого результата методами интеллектуального анализа данных необходимо, чтобы онтология предметной области была согласована с онтологией применяемого метода. Для такого согласования необходимо, во-первых, извлечь информацию из данных в терминах онтологии предметной области, а, во-вторых, применить такой метод, который может обнаружить на данных закономерности, интерпретируемые в терминах этой информации. Таким методом является разработанная нами в рамках реляционного подхода система Discovery. Для того, чтобы пользователь мог в этой системе удобно и визуально извлекать информацию из данных и формировать гипотезы в терминах этой информации, нами была разработана система Visual Discovery, которая позволяет это делать в режиме визуального конструктора. В работе дано описание данной системы и пример её использования для решения задачи из области медицины.

**Ключевые слова:** Data Mining, KDD&DM, интеллектуальный анализ данных.

**1. Онтология предметной области.** Как отмечалось в работе [Витяев Е.Е., 2006, 2010], для корректного применения методов интеллектуального анализа данных необходимо определить онтологию предметной области. Кратко напомним основные понятия предметной области и онтологии [Витяев Е.Е., 2010]. *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  $\mathfrak{Z} = \langle A, \Omega \rangle$ , где  $A$  – объекты ПО, а  $\Omega$  – онтология ПО (система понятий онтологии, заданная одноместными предикатами, и множество отношений и операций, интерпретируемых в системе понятий ПО).

**2. Онтология методов интеллектуального анализа данных.** Как отмечалось в работе [Витяев Е.Е., 2010] для того, чтобы методы интеллектуального анализа данных позволяли получать осмысленные и интерпретируемые результаты, необходимо, чтобы они правильно использовали содержащуюся в данных информацию. 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 получения осмысленных и интерпретируемых результатов, необходимо, чтобы онтология метода и онтология ПО были согласованы между собой следующим образом:

1. типы данных, с которыми работает метод, должны интерпретироваться в онтологии  $\Omega$  предметной области. Поэтому атрибуты, свойства и признаки, используемые в данных метода, должны быть интерпретируемы в онтологии  $\Omega$ . Тем самым определяется *информация, извлекаемая из данных этим методом*, которая представляется множеством интерпретируемых в онтологии  $\Omega$  математических отношений и операций;
2. класс проверяемых методом гипотез должен интерпретироваться в онтологии ПО. Это означает, что он должен выражаться через информацию, извлекаемую из данных.

Для того, чтобы знать какая информация содержится в данных, нам необходимо *извлечь информацию из данных*. Для этого надо представить информацию, содержащуюся в данных, множеством отношений и операций, интерпретируемых в онтологии предметной области.

Для этого нами была разработана программная система Visual Discovery, которая позволяет визуально извлекать информацию из данных в режиме конструктора. Описание системы приведено ниже.

**3. Извлечение информации из данных.** Проанализируем, как следует задавать свойства и ат-

<sup>1</sup> Работа поддержана грантом РФФИ 08-07-00272-а и интеграционными проектами СО РАН № 3,86,136.

рибутов объектов ПО в терминах онтологии  $\Omega$ . Чтобы правильно извлекать информацию и знания из свойств и атрибутов, необходимо их интерпретировать в системе понятий ПО. Как говорилось в [Витяев Е.Е., 2010], 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 числовых значений – метры, литры, килограммы и т.д. привязана к онтологии. Для извлечения информации из атрибутов, свойств, признаков и величин ПО нужно определить множество интерпретируемых в онтологии  $\Omega$  математических отношений и операций и включить их в онтологию  $\Omega$ .

Для работы непосредственно с информацией, извлеченной из данных и представленной множеством  $\Omega$  отношений и операций, нами разработан специальный реляционный подход (Relational Data Mining) к методам извлечения знаний и система «Discovery», реализующая его [Витяев, 2006; Витяев, Москвитин, 1993; Kovalerchuk, Vityaev, 2000; Vityaev, Kovalerchuk, 2008]. Однако в этой системе предполагается, что информация уже извлечена и представлена множеством  $\Omega$  отношений и операций. Для удобства пользователей необходимо, чтобы и сам процесс извлечения информации тоже сопровождался системой. С этой целью и разработано инструментальное средство Visual Discovery, описанное далее.

#### 4. Инструментальное средство Visual Discovery извлечения информации из данных и решения задач интеллектуального анализа данных.

Нами разработано инструментальное средство Visual Discovery, позволяющее специалисту ПО работать с онтологией ПО, извлекать информацию из данных, вручную формировать шаблоны предикатов (онтологию) и классы гипотез, и решать следующие три основные задачи:

1. Предсказывать значения признаков объектов на основе проверенных гипотез;
2. Выбирать информативную подсистему признаков на основе гипотез, которые подтвердились на данных;
3. Классифицировать объекты в соответствии с обнаруженными группами подтвержденных гипотез (закономерностей).

Для простоты, работу системы проиллюстрируем на следующем конкретном примере. В области распространения медицинских препаратов было проведено исследование компаний и численно измерены признаки, представленные на рис. 1:

Номер компании	Организационная культура	Товарооборот на 1 специалиста	Средний доход	Прибыль на 1 специалиста	Издержки на 1 специалиста	Доля издержек на весь персонал	Среднее число профессионального обучения, часов	Доля издержек на обучение	Коэффициент текущей	Доля часов на обучение в общем балансе времени
Компания № 1	ПОК (1+4)	3301,5	734,4	163,5	282,3	8,2	25,6	0,05	123,3	1,5
Компания № 2	БОК (2+3)	1851,6	548,3	144,3	191,5	10,3	10,3	0,03	10,6	0,6
Компания № 3	БОК (2+1)	1934,5	648,3	154,6	153,4	9,3	5,4	0,01	8,0	0,3
Компания № 4	ООК (3+1)	2635,4	528,4	126,4	104,3	6,2	15,4	0,05	5,6	0,8
Компания № 5	ПОК (1+2)	2031,2	526,4	113,4	107,4	6,3	12,3	0,01	17,6	0,6
Компания № 6	ПОК (1+2)	1935,4	583,4	139,4	144,3	9,2	30,2	0,07	19,3	1,7
Компания № 7	ПОК (1+3)	3142,4	675,3	159,3	218,3	11,3	12,4	0,04	15,3	0,6
Компания № 8	БОК (2+1)	2023,4	426,4	114,3	123,4	7,9	5,8	0,01	8,6	0,3
Компания № 9	ПОК (1+3)	2673,3	429,3	123,4	98,6	6,4	6,2	0,01	12,6	0,3
Компания № 10	ООК (3+1)	2515,4	335,6	153,4	97,4	8,6	10,6	0,03	10,4	0,5
Компания № 11	БОК (2+1)	2134,7	235,4	104,3	83,4	8,3	35,6	0,05	7,4	2,0
Компания № 12	ООК (3+1)	2435,7	506,7	118,3	84,6	7,1	16,7	0,06	5,3	0,9
Компания № 13	ПОК (1+2)	3004,5	683,2	159,6	243,4	10,6	5,2	0,01	16,3	0,3
Компания № 14	ПОК (1+3)	2947,3	675,4	153,2	204,3	9,4	6,8	0,01	10,2	0,4
Компания № 15	БОК (2+1)	1543,5	254,6	54,6	72,3	4,6	10,3	0,01	2,0	0,6
Компания № 16	ООК (1+2)	3124,6	653,4	160,4	234,3	10,4	20,6	0,05	1,0	1,1

Рис. 1. Таблица Объекты/признаки, характеризующие медицинские компании.

Необходимо было найти закономерности между признаками (выбирать информативную подсистему признаков), характеризующие максимальный товарооборот на одного специалиста.

**4.1. Процесс решения задач.** В общем случае, процесс решения задач 1-3, можно представить в виде диаграммы концептуальной модели деятельности (рис. 2) на языке UML. Эта модель может быть спецификацией прецедента использования, а также, многократно используемых функциональных свойств.

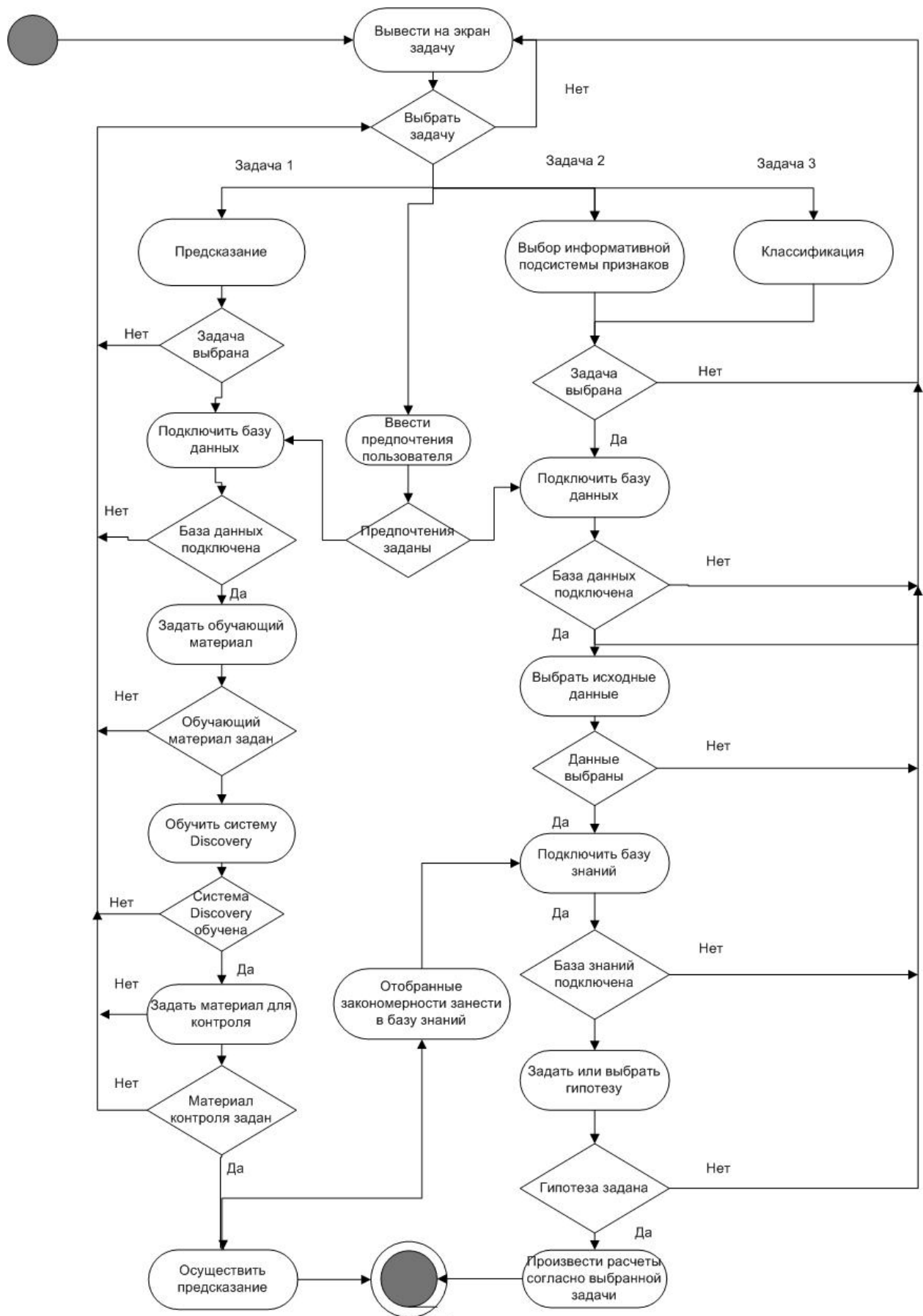


Рис. 2. Диаграмма прецедентов в системе Visual Discovery

Модель деятельности описывает поведение, в которое вовлечено несколько элементов системы. В системе Visual Discovery такими элементами являются: основные данные, шкала, справочник по

шкалам, рабочие гипотезы, описание предпочтений пользователя и знания экспертов. Диаграмма деятельности показывает шаги вычисления в системе Visual Discovery. За каждым классом деятельности скрывается определенная процедура вычислений.

Передача управления от одного состояния вида деятельности к другому задает поток управления, который определяет, как выполнение действия в одном узле влияет на выполнение действия в других узлах и, одновременно, испытывает их влияние.

Рассмотрим поток управления для нашей задачи – поток №2 (задача 2 в диаграмме) – выбор информативной подсистемы признаков. Для этого необходимо последовательно выполнить следующие действия (см. диаграмму):

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.

**4.2. Выбор исходных данных** (блок «выбор исходных данных» в потоке №2 диаграммы). Данные в системе Visual Discovery представляются виде таблицы, объекты в которой представлены строками, а признаки столбцами (см. screen shot на рис. 3). 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 A1-A11 имеют следующую интерпретацию:

1. A1 и A2 – организационная культура;
2. A3 – товарооборот на 1 специалиста;
3. A4 – средний доход;
4. A5 – прибыль на 1 специалиста;
5. A6 – издержки на 1 специалиста;
6. A7 – доля издержек на весь персонал
7. A8 – среднее число профессионального обучения, часов;
8. A9 – доля издержек на обучение;
9. A10 – коэффициент текучести;
10. A11 – доля часов на обучение в общем балансе времени.

Обучающее множество X 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. 3. Исходная таблица объектов/признаков в Visual Discovery.

**4.3. Задание онтологии** (см. блок «подключить базу знаний» в потоке №2 диаграммы). 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 путем 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 путем создания диаграммы шаблонов предикатов;
3. Получены системой путем 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 ( $T1 \text{ in } [T2, T3]$ ), ( $T1 \text{ in } (T2, T3)$ ), ( $T1 \text{ in } (T2; T3]$ ), ( $T1 \text{ in } [T2; T3]$ ).

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.

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 A3 (товарооборот на 1 специалиста) был разбит на три интервала:

1.  $T3 \text{ in } [1851,6; 2023,4)$ ;
2.  $T3 \text{ in } [2435,7; 2947,3)$ ;
3.  $T3 \geq 2947,3$

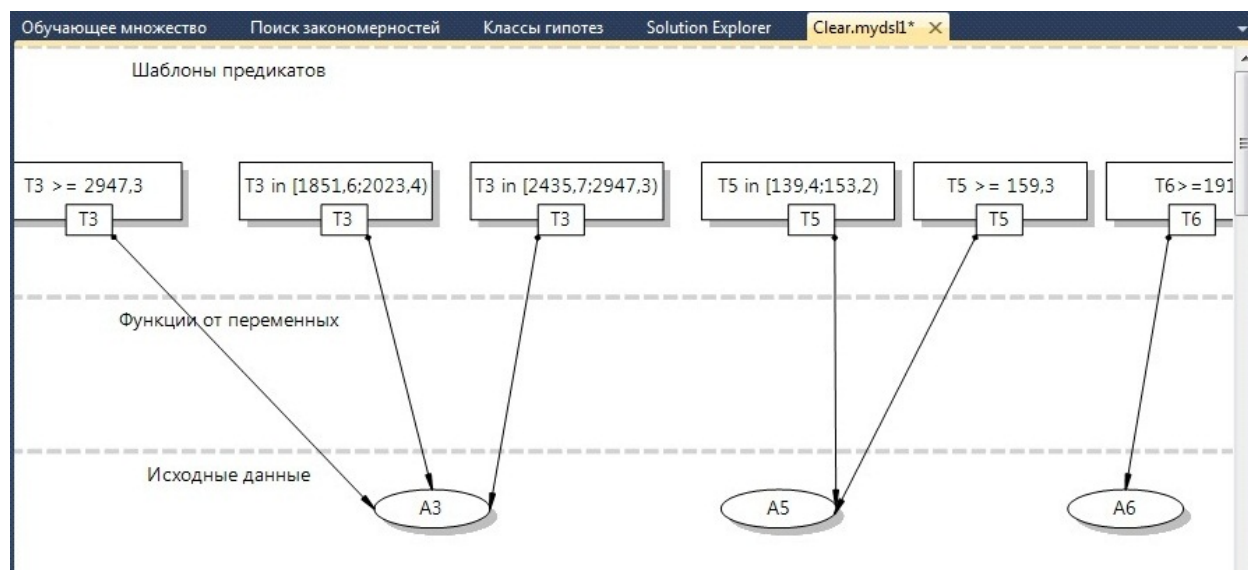


Рис. 4. Формирование шаблонов предикатов.

Соответствующие шаблоны предикатов для признака A3 введены в раздел «Шаблоны предикатов» (рис.4). Каждый предикат интерпретируется в некотором признаке и связывается с ним направленной стрелкой. Аналогично для признаков A5 и A6 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 будут формироваться с использованием только этой информации. Поэтому нужно определить, столько шаблонов предикатов, сколько нужно для выражения всей интересующей нас информации.

**4.4. Формирование классов гипотез** (блок «задать или выбрать гипотезу» рис.2.). Гипотезы задаются на основе шаблонов предикатов и определяют то знание, которое мы бы хотели получить в результате анализа данных. Гипотезы, и тем самым будущее знание, задается 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 «A3 in [2435,7; 2947,3]» выбирается из множества предикатов, задающихся шаблоном предиката «T3 in [2435,7; 2947,3]»;

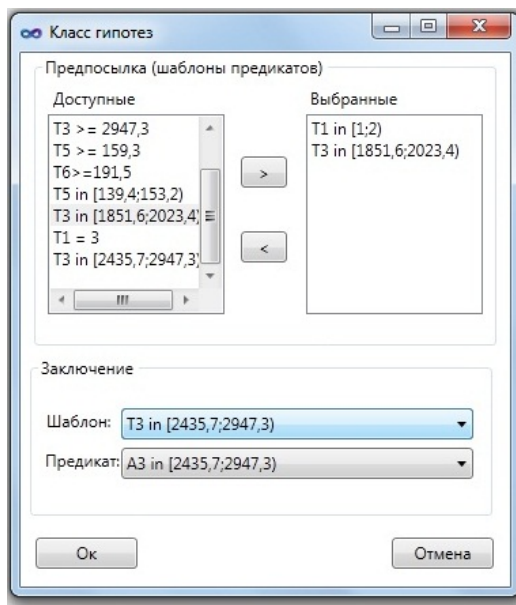


Рис. 5. Формирование класса гипотез.

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

Обучающее множество   Поиск закономерностей <b>Классы гипотез</b> Solution Explorer   Clear.mydsl1			
Предпосылка	Исход		
(T5 in [139,4;153,2))	A3 in [1851,6;2023,4)		Добавить
(T1 = 3)	A3 in [2435,7;2947,3)		Удалить
(T1 in [1;2)) and (T4 >= 648,3) and (T5 >= 159,3)	A3 >= 2947,3		Свойства

Рис.6 Заданные классы гипотез.

**4.5. Проверка классов гипотез** (блок «произвести расчеты согласно выбранной задаче», рис. 2.). 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 выдается программной системой Visual Discovery. Для обнаружения закономерностей программной системой Visual Discovery нужно задать следующие параметры [Kovalerchuk, Vityaev, 2000; Витяев, 2006]:

- 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.



Обучающее множество					
Поиск закономерностей					
Классы гипотез					
Solution Explorer					
Clear.myds1					
Критерий Фишера: 0,05					
Критерий Юла: 0					
Количество объектов: 16					
Глубина базового перебора: 1					
Найти закономерности					
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.

**4.6. 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  $(A1 \in [1;2]) \Rightarrow (A3 \geq 2947,3)$  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  $(A5 \geq 159,3) \Rightarrow (A3 \geq 2947,3)$  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  $(A4 \geq 648,3) \Rightarrow (A3 \geq 2947,3)$  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

- Витяев Е.Е. Извлечение знаний из данных. Компьютерное познание. Модели когнитивных процессов. Новосибирск, 2006. 293с.
- Витяев Е.Е. Извлечение информации из данных // Информационные технологии в гуманитарных исследованиях, Вып. 15, ИАЭТ СО РАН, Новосибирск, 2010, 9-16.
- Витяев Е.Е., Москвитин А.А. Введение в теорию открытий. Программная система DISCOVERY. // Логические методы в информатике (Вычислительные системы, вып. 148), Новосибирск, 1993, с.117-163
- Е. 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
- 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.