

*В работе проводится сравнение системы «Discovery» с алгоритмами Microsoft Association Rules и Decision Trees, встроенными в Microsoft SQL Server Analysis Services. Показывается, что система «Discovery», во-первых, обладает теоретическими преимуществами перед этими алгоритмами, и, во-вторых, практически работает лучше на данных, где эти преимущества проявляются явно. Эти данные, в том числе, хорошо демонстрируют и иллюстрируют преимущества системы Discovery.*

**Ключевые слова:** Интеллектуальный анализ данных, извлечение знаний, предсказание, обнаружение закономерностей.

**1. Введение.** В последнее время получили широкое развитие и активно применяются на практике различные KDD&DM-методы (Knowledge Discovery in Data Bases and Data Mining). Однако используемые сейчас KDD&DM-методы имеют серьезные ограничения [Витяев, 2006; Kovalerchuk, Vityaev, 2000]: каждый метод может работать только с определенными типами данных, имеет свой язык оперирования и интерпретации данных, и обнаруживает только определенный класс гипотез. Таким образом, они не способны извлекать из данных все знания в полном объеме, а также могут получать результаты, не интерпретируемые в терминах предметной области.

Система «Discovery» реализует реляционный подход к методам извлечения знаний [Витяев, 2006; Vityaev, Kovalerchuk, 2006; Kovalerchuk, Vityaev, 2000], снимающий практически все ограничения, свойственные KDD&DM-методам.

Система «Discovery» обладает следующими важными теоретическими свойствами: может обнаруживать теорию предметной области, может обнаруживать все правила, имеющие максимальные условные вероятности, может обнаруживать непротиворечивую вероятностную аппроксимацию теории предметной области [Vityaev, Kovalerchuk, 2006], обнаруживает все максимально специфические правила, позволяющие предсказывать без противоречий [Vityaev, 2005].

Наиболее близкими к системе «Discovery» методами можно считать поиск ассоциативных правил (Microsoft Association Rules) [Tang, MacLennan, 2005] и Decision Trees, в виду того, что закономерности в этих методах также представляются в форме логических правил. В данной работе ставится задача сравнения системы «Discovery» с Microsoft Association Rules и Decision Trees, встроенными в Microsoft SQL Server Analysis Services. Мы покажем, что система «Discovery» обладает теоретическими преимуществами перед этими методами и практически работает лучше на данных, где эти преимущества явно проявляются.

Система «Discovery» была реализована в виде плагина, подключаемого к службам Microsoft SQL Server 2005 Analysis Services (SSAS). Это позволяет использовать для сравнения алгоритмов единую среду разработки Business Intelligence Development Studio, единые средства визуализации Data Mining моделей, а также стандартные средства сравнения качества Data Mining моделей: диаграмму роста (Lift Chart) и классификационную матрицу (Classification Matrix).

**2. Association Rules.** Алгоритм Microsoft Association Rules состоит из двух шагов. Первый шаг – это ресурсоемкая фаза нахождения часто встречающихся наборов. Второй шаг – это генерация ассоциативных правил с использованием множества часто встречающихся наборов.

**Нахождение часто встречающихся наборов.** Под набором (itemset) мы понимаем набор предикатов. Например,  $\{A = 1; B = 0; C = 1\}$  – это набор длины 3. Запись таблицы содержит некоторый набор, если на этой записи выполнены все предикаты данного набора. Поддержка набора – это количество записей таблицы, которые содержат данный набор.

Основным параметром, участвующим в нахождении часто встречающихся наборов, является параметр Minimum Support, который определяет, в каком минимальном количестве записей анализируемой таблицы должен содержаться некоторый набор, чтобы он являлся часто встречающимся.

На первой итерации находятся все часто встречающиеся наборы длиной 1. Алгоритм просто сканирует таблицу и подсчитывает поддержку каждого возможного предиката. Предикаты с поддержкой большей, чем Minimum Support, добавляются во множество часто встречающихся наборов длины 1. На второй итерации из часто встречающихся наборов, найденных на первой итерации, строятся всевозможные наборы длины 2, подсчитываются поддержки этих наборов, те набо-

<sup>2</sup> Работа поддержана Российским Гуманитарным научным фондом, проект № 12-01-12026, грантом РФФИ 08-07-00272-а и интеграционными проектами СО РАН № 3,86,136.

ры, которые проходят критерий Minimum Support, добавляются во множество часто встречающихся наборов длины 2. Далее из предикатов, входящих в часто встречающиеся наборы длины 2, строятся всевозможные наборы длины 3 и т.д. Алгоритм повторяется для наборов длины 3, 4, 5 и т.д., пока находятся наборы удовлетворяющие критерию Minimum Support.

Далее проверяется условие, что каждый поднабор часто встречающегося набора, также должен являться часто встречающимся набором.

**Генерация ассоциативных правил.** Следующая процедура генерирует ассоциативные правила:

Для любого часто встречающегося набора  $f$ , генерируем все поднаборы  $x$  и их дополнения  $y = f - x$ .

Если  $\text{Поддержка}(f)/\text{Поддержка}(x) > \text{Minimum Probability}$ , тогда  $x \Rightarrow y$  является ассоциативным правилом с условной вероятностью  $\text{Prob} = \text{Поддержка}(f)/\text{Поддержка}(x)$ .

Параметр Minimum Probability задается перед началом обучения модели.

**Прогнозирование.** Следующий алгоритм по набору предикатов, поданных на вход, предсказывает значение целевого признака, либо выдает множество ( $n$  штук) наиболее вероятных значений целевого признака:

1. На вход подается некоторый набор предикатов. Ищутся все правила, условная часть которых совпадает либо с данным набором, либо с некоторым поднабором данного набора, а целевая часть содержит целевой признак. Найденные правила ( $k$  штук) применяются: целевые части правил и соответствующие условные вероятности добавляются в список рекомендаций.
2. Если подходящих правил не найдено, или их слишком мало ( $k < n$ ), находятся  $n-k$  наиболее популярных значений целевого признака. То есть, среди всех правил вида  $\Rightarrow P = a_i$  (с пустой условной частью и целевым признаком в правой части) находятся  $n-k$  правил с наибольшей условной вероятностью.
3. Предикаты, полученные на первых двух шагах, сортируются по вероятности.

**3. Decision Trees.** Основная идея алгоритма решающих деревьев состоит в рекурсивном разделении данных на подмножества, содержащие более или менее однородные состояния целевого (прогнозируемого) атрибута. При каждом разделении, все входные атрибуты оцениваются по их влиянию на целевой атрибут. Когда этот рекурсивный процесс заканчивается, решающее дерево сформировано.

**Построение таблицы подсчета корреляций.** Пусть для анализа с помощью алгоритма решающих деревьев дана некоторая таблица, с входными колонками  $F0, F1, F2, F3$ , и целевой колонкой  $P$ , размером 3000 записей, содержащую в своих ячейках только 0 или 1.

Тогда таблица подсчета корреляций на первом шаге алгоритма будет выглядеть следующим образом:

		F0		F1		F2		F3	
		0	1	0	1	0	1	0	1
P	0	300	700	700	300	400	600	500	500
	1	200	1800	400	1600	400	1600	1100	900

Каждая колонка таблицы подсчета корреляций соответствует паре атрибут-значение одного из входных атрибутов. Каждая строка соответствует значению целевого атрибута.

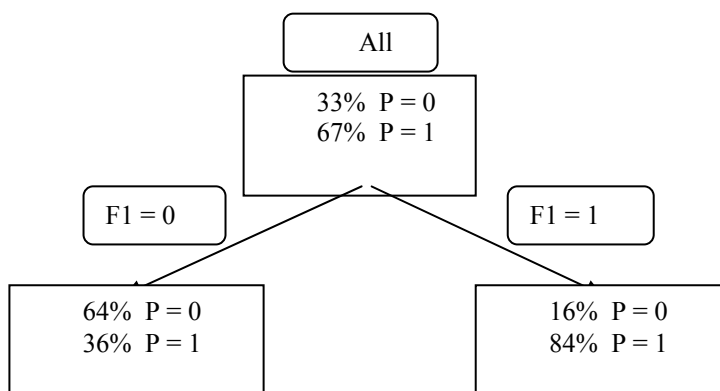
Ячейки таблицы содержат количество корреляций соответствующих пар: входной атрибут-значение, целевой атрибут-значение. Например, пересечение первой строки и первого столбца таблицы подсчета корреляций содержит число 300, т.е. в анализируемой таблице есть 300 записей, у которых одновременно  $F0 = 0$  и  $P = 0$ .

**Нахождение наиболее подходящего для разбиения атрибута.** Одним из широко известных критериев, с помощью которого можно найти необходимый атрибут, является *Энтропия*. Энтропия ( $H$ ) определяется следующим образом:

$$H(p_1, p_2, \dots, p_n) = -p_1 \cdot \log(p_1) - p_2 \cdot \log(p_2) - \dots - p_n \cdot \log(p_n), \quad p_1 + p_2 + \dots + p_n = 1$$

Например, энтропия атрибута  $F1$  равна:  $H(F1) = H(700, 400) + H(300, 1600) = 0.946 + 0.629 = 1.571$

Атрибут с наименьшей энтропией является наиболее подходящим для разбиения. В данном примере первое разбиение будет по атрибуту  $F1$ . Решающее дерево после первого разбиения выглядит следующим образом:



Таким образом, разделив данные входной таблицы на два подмножества, далее рекурсивно применяем первый шаг алгоритма к новообразованным вершинам решающего дерева. Например, новая таблица подсчета корреляций для вершины решающего дерева  $F1 = 0$  выглядит следующим образом:

		F0		F1		F2		F3	
		0	1	0	1	0	1	0	1
P	0	300	700	700	0	400	600	500	500
	1	200	1800	400	0	400	1600	1100	900

**Прогнозирование.** Алгоритм предсказания решающих деревьев достаточно простой и эффективен. Поданный на вход предсказания набор предикатов, например  $(F1 = 1, F2 = 0, F3 = 0, F4 = 1)$ , спускается по дереву от корня к листьям по соответствующему этому набору пути, вершина дерева, находящаяся в конце этого пути и определяет предсказанное значение целевого признака. Таким образом, количество шагов в процессе прогнозирования не превышает максимальной длины пути от корня к листьям решающего дерева.

**4. Система «Discovery».** Алгоритм поиска закономерностей системы «Discovery» реализует метод семантического вероятностного вывода [Витяев, 2006], позволяющего находить все максимально специфические и максимально вероятные закономерности в данных [Витяев, 2006]. Определим на высказываниях языка первого порядка вероятность, как описано в [Halpern, 1990].

Семантическим вероятностным выводом (СВВ) некоторого атома/литерала  $P$  является такая последовательность правил  $C_1, C_2, \dots, C_n$ , что:

- 1)  $C_i = (A_1^i \& \dots \& A_{k_i}^i \Rightarrow P)$ ,  $i = 1, \dots, n$ ;
- 2)  $C_i$  является подправилем правила  $C_{i+1}$ , т.е.  $\{A_1^i, \dots, A_{k_i}^i\} \subset \{A_1^{i+1}, \dots, A_{k_{i+1}}^{i+1}\}$ ;
- 3)  $\text{Prob}(C_i) < \text{Prob}(C_{i+1})$ ,  $i = 1, 2, \dots, n-1$ , где  $\text{Prob}(C_i)$  – условная Вероятность (УВ) правила,  $\text{Prob}(C_i) = \text{Prob}(P/A_1^i \& \dots \& A_{k_i}^i) = \text{Prob}(P \& A_1^i \& \dots \& A_{k_i}^i) / \text{Prob}(A_1^i \& \dots \& A_{k_i}^i)$ ;
- 4)  $C_i$  – Вероятностные Законы (ВЗ), т.е. для любого подправила  $C' = (A_1 \& \dots \& A_j \Rightarrow P)$  правила  $C_i$ ,  $\{A_1, \dots, A_j\} \subset \{A_1^i, \dots, A_{k_i}^i\}$  выполнено неравенство  $\text{Prob}(C') < \text{Prob}(C_i)$ ;
- 5)  $C_n$  – Сильнейший Вероятностный Закон (СВЗ), т.е. правило  $C_n$  не является подправилем никакого другого вероятностного закона.

Вероятностные неравенства в пунктах 3-4 проверяются на данных с помощью точного критерия независимости Фишера и критерия Юла [Кендалл, Стьюарт, 1973; Закс, 1975].

Множество всех цепочек СВВ предиката  $P$  образуют дерево СВВ предиката  $P$ .

Реализовать семантический вероятностный вывод в чистом виде не представляется возможным ввиду требований к производительности алгоритма, т.к. пункты 2 и 4 определения СВВ подразумевает большое пространство перебора. Для уменьшения перебора применяется следующие упрощения.

Во-первых, положим, что при построении цепочки СВВ правило  $C_{i+1}$  получается из правила  $C_i$  добавлением к его условной части только одного предиката. Эксперименты показывают, что крайне редка ситуация, когда добавление в условную часть правила сразу двух предикатов дает ВЗ, а добавление любого из этих двух признаков по отдельности не дает ВЗ. Следовательно, мы можем значительно уменьшить пространство перебора, почти не снижая количество и качество

извлеченных из данных закономерностей.

Во-вторых, для того чтобы уменьшить перебор при проверке условия в пункте 4, используется поуровневая схема генерация правил: сначала генерируются все ВЗ с одним предикатом в условной части и заключением Р, затем с двумя предикатами, тремя и т.д. Таким образом, для проверки, является ли некоторое правило ВЗ, достаточно просмотреть все его подправила, находящиеся на предыдущем уровне дерева СВВ.

Перед началом обучения колонки входной таблицы помечаются атрибутами Input, PredictOnly и Predict, которые указывают, каким образом та или иная колонка участвует в обучении: в качестве входного признака, целевого признака, или в качестве обоих одновременно. Также в качестве параметров модели могут задаваться пороговые величины: условная частота правила, уровни значимости критериев Фишера и Юла, максимальное число интервалов значений для признака и др.

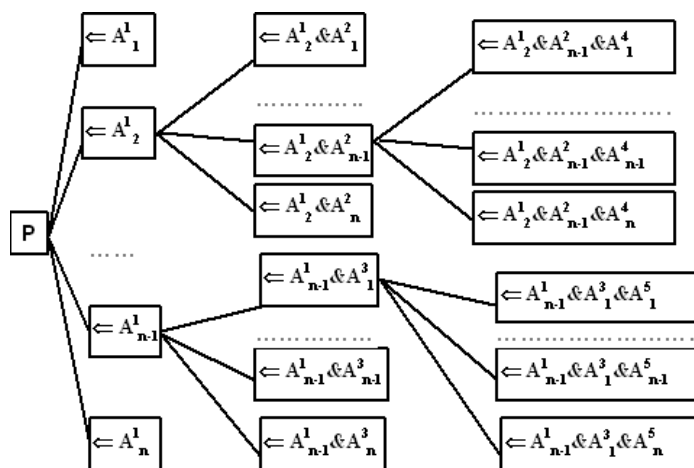


Рис. 1. дерево СВВ, включающее все СВВ, содержащие в заключении атом Р.

Результатом работы алгоритма является:

- 1) дерево СВВ для каждого целевого предиката;
- 2) множество ВЗ и СВЗ этих деревьев;
- 3) *Максимально Специфический Закон* (МСЗ) для каждого целевого предиката, определяемый как СВЗ, обладающий наибольшей условной вероятностью среди других СВЗ дерева вывода этого предиката.

Множество всех МСЗ обладает таким важным свойством как потенциальная непротиворечивость [Витяев, 2006].

#### Алгоритм поиска закономерностей:

1. Generate\_First\_Tree\_Level (Queue Q, Tree T)

– Создаются все возможные правила, состоящие только из целевой части (они все по определению ВЗ). Для них сразу рассчитывается вся необходимая статистика на основе входных данных. Все элементы добавляются в корень дерева Т; элементы, содержащие Predict предикаты, добавляются в очередь Q.

2. Generate\_Subsequent\_Tree\_Level (Queue Q, Tree T)

Обход дерева в ширину:

– Берем элемент из начала очереди Q, для него генерируем потомков, т.е. уточняем правило путем добавления 1-го нового предиката в условную часть.

– Проверяем, является ли новое правило ВЗ (см. Check\_If\_Probability\_Low). Если является ВЗ, добавляем в дерево Т, добавляем в очередь Q.

– Процесс повторяется, пока очередь не пуста, т.е. еще можно получить новые ВЗ, путем добавления 1-го предиката в условную часть правила. Правила, соответствующие элементам которых не имеют потомков, являются СВЗ.

3. Check\_If\_Probability\_Low (Rule R)

– Проверяем, является ли статистически значимым правило R с помощью критериев Фишера и Юла [Витяев, 2006, с. 117-120]. Если нет, то R не является ВЗ;

– просматриваем все подправила Sr длины Length(R) - 1

– Если Prob(Sr) > Prob(R), то R не является ВЗ;

– иначе R является ВЗ.

#### 4. Extract\_MSL (Tree T)

- Для каждого целевого предиката просматриваем его дерево СВВ. Сортируем множество СВЗ этого дерева по вероятности. Правила с наибольшей условной вероятностью являются Максимально Специфическими Законами (МСЗ) рассматриваемого целевого предиката.

**Прогнозирование.** Следующий алгоритм по набору предикатов, поданных на вход и множествам обнаруженных закономерностей ВЗ, СВВ и МСЗ, предсказывает значение целевого признака:

1. На вход подается некоторый набор предикатов. Ищутся все максимально специфичные закономерности, условная часть которых совпадает с данным набором либо с некоторым поднабором данного набора, а целевая часть содержит целевой признак. Максимально специфичная закономерность с наибольшей УВ определяет предсказанное значение целевого признака.
2. Если подходящих МСЗ не найдено, то рассматриваются все ВЗ с дерева СВВ целевого признака. Среди рассмотренных ВЗ ищутся те правила, условная часть которых совпадает с входным набором либо с некоторым поднабором входного набора. Правило с наибольшей УВ определяет предсказанное значение целевого признака.

**5. Теоретическое сравнение с Association Rules и Decision Trees.** В силу определения семантического вероятностного вывода в посылку правила всегда добавляются только такие предикаты, которые статистически значимо строго увеличивают условную вероятность правила. Такая проверка в других методах не проводится. В этом случае в правила могут включаться признаки, которые не имеют отношения к предсказанию и в том числе случайные. Опора на случайные признаки в предсказании может значительно ухудшить его точность.

Важность отсева случайных признаков критична, если нам надо не просто получить предсказание, а еще и проинтерпретировать полученные результаты. Специалист предметной области, интерпретируя правило, должен быть уверен, что признаки, входящие в правило действительно имеют отношение к предсказываемому признаку. Иначе интерпретация будет невозможна, либо специалист скажет, что правила бессмысленны и полученные предсказания – гадание на кофейной гуще и отчасти будет прав.

В Decision Trees для выбора разделяющих признаков используется энтропия, но на малых данных или в концах веток дерева могут включаться признаки, которые случайны и минимум энтропии признака получается чисто случайно. В Decision Trees нет статистического критерия проверки случайности добавляемых признаков. Поэтому правила, получаемые в Decision Trees, могут содержать случайные признаки, не имеющие отношение к предсказываемому признаку.

Для сравнения с алгоритмами Association Rules в качестве модельных данных возьмем следующие тестовые таблицы. Тестовая таблица 1 имеет три значимых колонки  $F_1, F_2, F_3$  определяемые следующим выражением:

$$F_1(i) = \begin{cases} 1, i \in (1, 512k) \\ 0, i \in (512k+1, 1024k) \end{cases}, \quad 1024 - \text{количество записей в таблице.}$$

$$F_2(i) = \begin{cases} 1, i \in (1, 256k) \cup (512k+1, 768k) \\ 0, i \in (256k+1, 512k) \cup (768k+1, 1024k) \end{cases}$$

$$F_3(i) = \begin{cases} 1, i \in (1, 128k) \cup (256k+1, 384k) \cup (512k+1, 640k) \cup (768k+1, 896k) \\ 0, i \in (128k+1, 256k) \cup (384k+1, 512k) \cup (640k+1, 768k) \cup (896k, 1024k) \end{cases}$$

и колонку P, используемую в качестве целевого признака:

$$P(i) = \begin{cases} 1, i \in (1, 128k) \\ 0, i \in (128k+1, 1024k) \end{cases},$$

а также 5 колонок  $R_1 - R_5$  с независимыми Бернуллиевскими случайными значениями.

Тестовая таблица 2 также имеет три значимых колонки  $F_1, F_2, F_3$  определяемые следующим выражением:

$$F_1(i) = \begin{cases} 1, i \in (1, 729k) \\ 2, i \in (729k+1, 1458k) \\ 3, i \in (1458k+1, 2187k) \end{cases}$$

$$F_2(i) = \begin{cases} 1, i \in (1, 243k) \cup (729k+1, 972k) \cup (1458k+1, 1701k) \\ 2, i \in (243k+1, 486k) \cup (972k+1, 1215k) \cup (1701k+1, 1944k) \\ 3, i \in (486k+1, 729k) \cup (1215k+1, 1458k) \cup (1944k+1, 2187k) \end{cases}$$

$$F_3(i) = \begin{cases} 1, i \in \bigcup_{j=0}^2 \left( (1+729kj, 81k+729kj) \cup (1+243k+729kj, 324k+729kj) \right) \\ \quad \cup (1+486k+729kj, 567k+729kj) \\ 2, i \in \bigcup_{j=0}^2 \left( (1+81k+729kj, 162k+729kj) \cup (1+324k+729kj, 405k+729kj) \right) \\ \quad \cup (1+567k+729kj, 648k+729kj) \\ 3, i \in \bigcup_{j=0}^2 \left( (1+162k+729kj, 243k+729kj) \cup (1+405k+729kj, 486k+729kj) \right) \\ \quad \cup (1+648k+729kj, 729k+729kj) \end{cases}$$

(2187 – количество записей в таблице) и колонку Р, используемую в качестве целевого признака:

$$P(i) = \begin{cases} 1, i \in (1, 81k) \\ 0, i \in (1+81k, 2187k) \end{cases},$$

а также 5 колонок  $R_1 - R_5$  с независимыми Бернуллиевскими случайными значениями.

На тестовых таблицах можно увидеть две простые закономерности:

$$(F_1 = 1 \wedge F_2 = 1 \wedge F_3 = 1) \Rightarrow P = 1, \quad (F_1 \neq 1 \wedge F_2 \neq 1 \wedge F_3 \neq 1) \Rightarrow P = 0.$$

Под тестовой таблицей с n-процентным шумом мы подразумеваем тестовую таблицу, в которой в n процентах ячеек, выбранных случайным образом, значение заменено на противоположное.

Хотя система «Discovery» и алгоритм Microsoft Association Rules достаточно похожи, в виду того, что в обоих подходах закономерности представляются в форме логических правил, тем не менее, между ними существуют принципиальные отличия.

В детерминированном случае, когда нет шума в данных, система «Discovery» обнаружит одно правило  $A \& B \Rightarrow C$ , истинное на данных. В то же время алгоритм, обнаруживающий ассоциативные правила, обнаружит все правила вида  $A \& B \& \dots \& D \Rightarrow C$ , которые получаются из правила  $A \& B \Rightarrow C$  добавлением дополнительных условий  $D, F, \dots$ :  $A \& B \& D \Rightarrow C$ ,  $A \& B \& F \Rightarrow C$ ;

Например, при анализе тестовой таблицы 1, где в качестве входных колонок использовались  $F_1, F_2, F_3$ , а также колонка  $R_1$  со случайными данными, алгоритмом Association Rules было обнаружено правило  $(F_1 = 1 \wedge F_2 = 1 \wedge F_3 = 1) \Rightarrow P = 1$ , а также следующие 2 правила с УВ = 1:

$$(F_1 = 1 \wedge F_2 = 1 \wedge F_3 = 1 \wedge R_1 = 1) \Rightarrow P = 1,$$

$$(F_1 = 1 \wedge F_2 = 1 \wedge F_3 = 1 \wedge R_1 = 0) \Rightarrow P = 1.$$

Таким образом, в случае, когда цель анализа – найти закономерности в данных, эксперт, использующий алгоритм Association Rules, получит три противоречивых правила с УВ = 1. Доверия к таким результатам не будет.

Кроме того, алгоритмом Association Rules было обнаружено множество правил следующего вида:

$$(F_1 = 1 \wedge R_1 = 1) \Rightarrow P = 1, \quad (F_1 = 1 \wedge R_1 = 0) \Rightarrow P = 1,$$

$$(F_2 = 1 \wedge R_1 = 1) \Rightarrow P = 1, \quad (F_2 = 1 \wedge R_1 = 0) \Rightarrow P = 1,$$

$$(F_1 = 1 \wedge F_2 = 1 \wedge R_1 = 1) \Rightarrow P = 1, \quad (F_1 = 1 \wedge F_2 = 1 \wedge R_1 = 0) \Rightarrow P = 1.$$

Последние правила могут иметь приоритет над правилами с целевым предикатом  $P = 0$ , и, соответственно, ложно предсказывать 1, когда колонка Р содержит 0. Например, в случае, когда на вход подаются колонки  $F_1, F_2, F_3$  и только одна колонка со случайными данными  $R_1$ , процент правильно предсказанных алгоритмом Association Rules значений составит около 87%, когда на

вход подаются  $F_1, F_2, F_3$  плюс две колонки  $R_1$  и  $R_2$  точность падает до 70% (подробнее см. параграф 5, таб. 1).

Система «Discovery» в данном случае обнаружила только следующие правила:

$$(F_1 = 1 \wedge F_2 = 1 \wedge F_3 = 1) \Rightarrow P = 1, (F_1 = 0) \Rightarrow P = 0,$$

$$(F_2 = 0) \Rightarrow P = 0, (F_3 = 0) \Rightarrow P = 0,$$

т.к. они уже имеют  $УВ = 1$ , и добавление каких-либо предикатов в условную часть правила не может увеличить условную вероятность правила. Предсказание, основанное на этих правилах, будет 100% точным.

Когда есть шум в данных, система «Discovery» может обнаружить одно правило  $A \& B \Rightarrow C$ , представляющее собой вероятностный закон с определенным уровнем статистической значимости. В то же время алгоритм, обнаруживающий ассоциативные правила, должен обнаружить все детерминированные правила вида  $A \& B \& D \Rightarrow C$ ,  $A \& B \& F \Rightarrow C$ , включающие случайные признаки, что приведёт к ухудшению предсказания.

Например, при анализе тестовой таблицы 1 с 3% шумом, и колонками  $F_1, F_2, F_3, R_1$  в качестве входных колонок, системой «Discovery» были обнаружены только 4 правила, являющиеся СВЗ:

$$(F_1 = 1 \wedge F_2 = 1 \wedge F_3 = 1) \Rightarrow P = 1, (F_1 = 0) \Rightarrow P = 0,$$

$$(F_2 = 0) \Rightarrow P = 0, (F_3 = 0) \Rightarrow P = 0.$$

Эти правила не содержат колонку со случайными данными, т.к. любое правило, имеющее в условной части колонку  $R_1$  не пройдет проверку критерием Юла-Фишера и будет удалено.

Алгоритм Association Rules в данном примере обнаружил правило  $(F_1 = 1 \wedge F_2 = 1 \wedge F_3 = 1) \Rightarrow P = 1$  с  $УВ = p$ , а также правила:

$$(F_1 = 1 \wedge F_2 = 1 \wedge F_3 = 1 \wedge R_1 = 1) \Rightarrow P = 1, (F_1 = 1 \wedge F_2 = 1 \wedge F_3 = 1 \wedge R_1 = 0) \Rightarrow P = 1,$$

причем одно из них имеет  $УВ > p$ . Таким образом, в случае, когда цель анализа – найти закономерности в данных, эксперт, использующий алгоритм Association Rules, получит три противоречивых правила. При этом правило с наибольшей  $УВ$  может содержать колонку со случайными данными.

Кроме того, из-за шума в данных алгоритм Association Rules обнаруживает множество правил следующего вида:

$$(F_1 = 1 \wedge F_2 = 1 \wedge F_3 = 0 \wedge R_1 = 1) \Rightarrow P = 1, (F_1 = 1 \wedge F_2 = 1 \wedge F_3 = 0 \wedge R_1 = 0) \Rightarrow P = 1,$$

$$(F_1 = 1 \wedge F_2 = 0 \wedge F_3 = 1 \wedge R_1 = 1) \Rightarrow P = 1, (F_1 = 1 \wedge F_2 = 0 \wedge F_3 = 1 \wedge R_1 = 0) \Rightarrow P = 1,$$

$$(F_1 = 0 \wedge F_2 = 1 \wedge F_3 = 1 \wedge R_1 = 1) \Rightarrow P = 1, (F_1 = 0 \wedge F_2 = 1 \wedge F_3 = 1 \wedge R_1 = 0) \Rightarrow P = 1,$$

которые могут иметь приоритет над правилами с целевым предикатом  $P = 0$  и ложно предсказывать 1, когда колонка  $P$  содержит 0. Это приводит к ухудшению предсказания (что показано на рис. 6 в приложении 4).

**6. Экспериментальное сравнение Discovery с Association Rules.** В качестве анализируемых данных используются тестовые таблицы, описанные выше. Для анализа данных таблиц применим систему «Discovery». В качестве входных колонок используем колонки  $F_1, F_2, F_3, R_1 - R_5$ , в качестве целевого признака – колонку  $P$ . В качестве критериев статистической значимости используемая реализация применяет точный критерий Фишера с пороговым значением 0,05 и критерий Юла с пороговым значением 0,1.

Система «Discovery» обнаруживает следующие правила с условной вероятностью ( $УВ$ ), равной 1.

На тестовой таблице 1:

УВ 1: IF (F1 = 1) AND (F2 = 1) AND (F3 = 1) THEN (P = 1);

УВ 1: IF (F3 = 0) THEN (P = 0);

УВ 1: IF (F2 = 0) THEN (P = 0);

УВ 1: IF (F1 = 0) THEN (P = 0);

На тестовой таблице 2:

УВ 1: IF (F1 = 1) AND (F2 = 1) AND (F3 = 1) THEN (P = 1);

УВ 1: IF (F3 = 2) THEN (P = 0);

УВ 1: IF (F3 = 3) THEN (P = 0);

УВ 1: IF (F2 = 2) THEN (P = 0);  
 УВ 1: IF (F2 = 3) THEN (P = 0);  
 УВ 1: IF (F1 = 2) THEN (P = 0);  
 УВ 1: IF (F1 = 3) THEN (P = 0);

Теперь проанализируем тестовые таблицы с помощью Association Rules. В качестве входных колонок используем колонки  $F_1, F_2, F_3$ , в качестве целевого признака – колонку P.

Алгоритм Association Rules обнаруживает 20 правил с УВ, равной 1, на тестовой таблице 1 (57 на тестовой таблице 2), в том числе и все правила найденные системой «Discovery». При добавлении колонки  $R_1$  ко входным колонкам Association Rules обнаруживает уже 60 правил с УВ, равной 1, на тестовой таблице 1 (228 на тестовой таблице 2). В Приложении 1 на рисунке 2 показано как растет количество правил с УВ, равной 1, обнаруженных алгоритмом Association Rules, при добавлении к  $F_1, F_2, F_3$  колонок  $R_1 - R_5$  в качестве входных колонок.

Далее посмотрим, как добавление в модель колонок со случайными данными ухудшает качество предсказания «Ассоциативных правил», а также покажем, что количество записей в таблице незначительно влияет на качество предсказания.

**Таблица 1.** Процент правильно предсказанных алгоритмом Association Rules значений на тестовой таблице 1

	0	1	2	3	4	5
Association Rules, n = 256	100%	87.5%	69.53%	45.70%	29.29%	19.53%
Association Rules, n = 1024	100%	87.59%	70.41%	45.31%	30.56%	23.92%
Association Rules, n = 4096	100%	88.15%	69.14%	47.07%	32.86%	24.43%

**Таблица 2.** Процент правильно предсказанных алгоритмом Association Rules значений на тестовой таблице 2

	0	1	2	3	4	5
Association Rules, n = 243	100%	91.81%	74.16%	58.46%	25.68%	16.26%
Association Rules, n = 2187	100%	91.95%	75.76%	55.05%	28.30%	17.92%
Association Rules, n = 19683	100%	91.06%	76.85%	55.19%	29.71%	18.46%

Отметим, что на тестовых таблицах 1 и 2 без шума система «Discovery» имеет 100% правильно предсказанных значений целевой колонки P при любом количестве случайных колонок  $R_1 - R_5$ , участвующих в обучении модели.

В Приложении 2 на рис. 3,4 проводится сравнение качества предсказания алгоритма Association Rules и системы «Discovery».

Рассмотрим тестовые таблицы 1 и 2 с наложением 3% шума. В качестве входных колонок используем колонки  $F_1, F_2, F_3, R_1 - R_5$ , в качестве целевого признака – колонку P.

В результате система «Discovery» обнаруживает следующие правила, являющиеся СВЗ.

На тестовой таблице 1:

УВ 0,854: IF (F1 = 1) AND (F2 = 1) AND (F3 = 1) THEN (P = 1)

УВ 0,959: IF (F2 = 0) THEN (P = 0)

УВ 0,944: IF (F1 = 0) THEN (P = 0)

УВ 0,945: IF (F3 = 0) THEN (P = 0)

Первые два из них являются МСЗ.

На тестовой таблице 2:

УВ 0,910: IF (F1 = 1) AND (F2 = 1) AND (F3 = 1) THEN (P = 1)

УВ 0,988: IF (F1 = 2) AND (F2 = 3) AND (F3 = 2) THEN (P = 0)

УВ 0,962: IF (F2 = 2) AND (F3 = 3) THEN (P = 0)



УВ 0,967: IF (F1 = 3) AND (F2 = 2) THEN (P = 0)

УВ 0,971: IF (F1 = 3) AND (F3 = 3) THEN (P = 0)

УВ 0,977: IF (F1 = 3) AND (F2 = 3) THEN (P = 0)

Первые два из них также являются МСЗ.

Проанализируем тестовые таблицы 1 и 2 с наложением 3% шума с помощью Association Rules. В качестве входных колонок используем колонки  $F_1, F_2, F_3$ , в качестве целевого признака – колонку P. В результате, на тестовой таблице 1 Association Rules обнаруживает 29 правил, из них 20 правил с УВ > 0.85, в том числе и все правила, найденные системой «Discovery». На тестовой таблице 2 Association Rules обнаруживает 60 правил с УВ > 0.85, в том числе и все правила, найденные системой «Discovery». В Приложении 3 на рис. 5 показано, как растет количество правил с УВ > 0.85, обнаруженных алгоритмом Association Rules, при добавлении колонок  $R_1 - R_5$  в качестве входных колонок.

В Приложении 4 проводится сравнение качества предсказания алгоритма Association Rules и системы Discovery на тестовой таблице 2 с шумом 0%, 2% и 3%.

**7. Экспериментальное сравнение Discovery с Decision Trees.** В качестве анализируемых данных используются таблицы следующего вида:

F0	F1	F2	F3	F4	P
1	1	0	0	0	1
1	1	1	0	0	1
1	1	1	1	0	1
0	1	1	1	1	1
0	0	1	1	1	1
0	0	0	1	1	1
1	0	1	0	0	0
1	0	1	0	0	0
1	0	1	0	0	0
1	0	1	0	0	0
1	0	0	1	0	0
1	0	0	1	0	0
0	1	0	1	0	0
0	1	0	1	0	0
0	1	0	1	0	0
0	1	0	1	0	0
0	1	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0
0	0	1	0	1	0
0	0	1	0	1	0
0	0	1	0	1	0
0	0	0	1	0	0
0	0	0	1	0	0
0	0	0	0	1	0

Для обучения алгоритмов Discovery и Decision Trees использовались таблица1, полученная из таблицы, приведенной выше, масштабированием в 8 раз (и удалением нескольких записей в конце таблицы), так, что итоговая длина тестовой таблицы1 равна 197.

Сначала для анализа таблицы1 применим систему «Discovery». В качестве входных колонок используем колонки  $F_0 - F_4$ , в качестве целевого признака – колонку P. В качестве критериев статистической значимости применим точный критерий Фишера с пороговым значением 0,13 и критерий Юла с пороговым значением минус 0,03.

Система «Discovery» обнаруживает следующие правила с условной вероятностью (УВ) равной 1 предсказывающие  $P = 1$

УВ 1: IF (F0 = 1) AND (F1 = 1) THEN (P = 1);

УВ 1: IF (F1 = 1) AND (F2 = 1) THEN (P = 1);

УВ 1: IF (F2 = 1) AND (F3 = 1) THEN (P = 1);

УВ 1: IF (F3 = 1) AND (F4 = 1) THEN (P = 1).

А также несколько правил, предсказывающие  $P = 0$ . Далее проанализируем таблицу1 с помощью Decision Trees. В качестве входных колонок используем колонки  $F_0 - F_4$ , в качестве целевой – колонку P.

Параметр COMPLEXITY\_PENALTY для Decision Trees установим равным 0.001. Алгоритм Decision Trees обнаружил следующие правила, предсказывающие  $P = 1$ :

УВ 0.965: IF (F0 = 1) AND (F1 = 1) THEN (P = 1);

УВ 0.982: IF (F1 = 0) AND (F3 = 1) AND (F4 = 1) THEN (P = 1);

А также несколько правил, предсказывающие  $P = 0$ . Нетрудно заметить, что найденных Decision Trees правил не достаточно для точного предсказания  $P = 1$ .

Для сравнения качества предсказания Discovery и Decision Trees будем использовать тестовую таблицу2, полученную масштабированием таблицы1 в 4 раза. Введем следующую меру ошибки: если алгоритм ошибочно предсказывает 1 вместо 0, то стоимость ошибки равна 1, если ошибочно предсказывает 0 вместо 1, то стоимость ошибки равна 3. Данную меру ошибки можно представить в виде таблицы:

Actual\Predicted	0	1
0	0	1
1	3	0

Максимальная ошибка, возможная на тестовой таблице2 имеет стоимость 1172. В результате выполнения предсказания для записей тестовой таблицы 2, Discovery не допустила ни одной ошибки, т.е. предсказала абсолютно точно. Decision Trees на тестовой таблице2 допустила ошибку общей стоимостью 96, или 8.2% от стоимости максимальной ошибки.

Проанализируем, как добавление шума в таблицы влияет на обнаруживаемые закономерности

и ухудшает качество предсказания алгоритмов Discovery и Decision Trees.

Под таблицей с  $n$ -процентным шумом мы подразумеваем таблицу, в которой в  $n$  процентах ячеек, выбранных случайным образом, значение признака заменено на противоположное.

На тестовой таблице 1 с шумом 1% система «Discovery» обнаруживает следующие правила, предсказывающие  $P = 1$ :

УВ 0.958: IF ( $F_0 = 1$ ) AND ( $F_1 = 1$ ) THEN ( $P = 1$ );

УВ 0.962: IF ( $F_1 = 1$ ) AND ( $F_2 = 1$ ) THEN ( $P = 1$ );

УВ 0.960: IF ( $F_2 = 1$ ) AND ( $F_3 = 1$ ) THEN ( $P = 1$ );

УВ 1.000: IF ( $F_3 = 1$ ) AND ( $F_4 = 1$ ) THEN ( $P = 1$ );

А также правила, предсказывающие  $P = 0$ . На этой же таблице Decision Trees обнаруживает следующие правила, предсказывающие  $P = 1$ :

УВ 0.982: IF ( $F_1 = 0$ ) AND ( $F_3 = 1$ ) AND ( $F_4 = 1$ ) THEN ( $P = 1$ );

УВ 0.982: IF ( $F_0 = 1$ ) AND ( $F_1 = 1$ ) AND ( $F_2 = 1$ ) THEN ( $P = 1$ );

УВ 0.878: IF ( $F_0 = 0$ ) AND ( $F_1 = 1$ ) AND ( $F_2 = 1$ ) THEN ( $P = 1$ );

А также несколько правил, предсказывающие  $P = 0$ . В следующей таблице показаны результаты предсказаний, выполненных алгоритмами Discovery и Decision Trees. Алгоритмы обучались на таблице 1 с шумами в 1, 2 и 3 процента. Предсказание выполнялось для записей таблицы 2 с теми же шумами 1, 2, 3 процента.

	Discovery		Decision Trees	
	Стоимость	% от max	Стоимость	% от max
шум 0%	0	0	96	8.19
шум 1%	40	3.38	128	10.83
шум 2%	92	7.7	179	14.99
шум 3%	123	10.12	198	16.28

Во всех приведенных случаях система Discovery работала лучше Decision Trees.

## ЛИТЕРАТУРА

**Витяев Е.Е.** Извлечение знаний из данных. Компьютерное познание. Модели когнитивных процессов. – НГУ, Новосибирск, 2006. – 293 с.

**Закс Ш.** Теория статистических выводов – М.: Мир, 1975. – 776 с.

**Кендалл М. Дж., Стьюарт А.** Статистические выводы и связи – М.: Наука, 1973. – 899 с.

**Halpern J. Y.** An analysis of first-order logic of probability. – Artificial Intelligence. 1990. – С. 311–350.

**Kovalerchuk, B.Y., Vityaev E.E.** Data mining in finance: Relational and hybrid methods. Kluwer, 2000. 308 p.

**Tang Z., MacLennan J.** Data Mining with SQL Server 2005. Wiley Publishing, Inc., 2005. 483 p.

**Vityaev E., Kovalerchuk B.** Empirical Theories Discovery based on the Measurement Theory // Mind and Machine, 2006. V.14. N4, P. 551-573.

**Vityaev E.** The logic of prediction. In: Mathematical Logic in Asia // Proceedings of the 9th Asian Logic Conference (August 16-19, 2005, Novosibirsk, Russia). World Scientific, Singapore, 2006. P. 263-276.

## Приложение 1

На следующем графике показано, как растет количество правил с УВ, равной 1, обнаруженных алгоритмом Association Rules, при добавлении к  $F_1, F_2, F_3$  колонок  $R_1 - R_5$  в качестве входных с колонок.

Здесь и далее на горизонтальной оси отмечено количество колонок  $R_1 - R_5$  со случайными данными, используемых (в дополнение к  $F_1, F_2, F_3$ ) в качестве входных данных.

Как видим, количество правил, найденных Association Rules, экспоненциально растет при добавлении новых колонок.

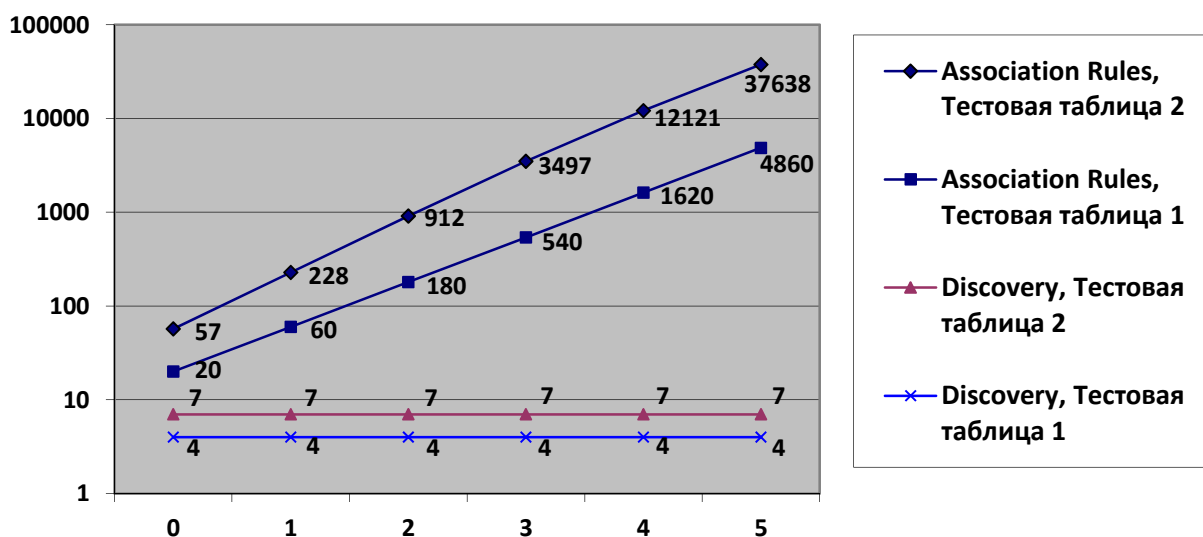


Рис. 2. Количество правил с УВ = 1, обнаруженных алгоритмом Association Rule

## Приложение 2

На следующем графике показан процент правильно предсказанных значений алгоритма Association Rules и системы «Discovery» в зависимости от количества колонок  $R_1 - R_5$  со случайными данными, используемых в качестве входных данных, а также размера тестовой таблицы.

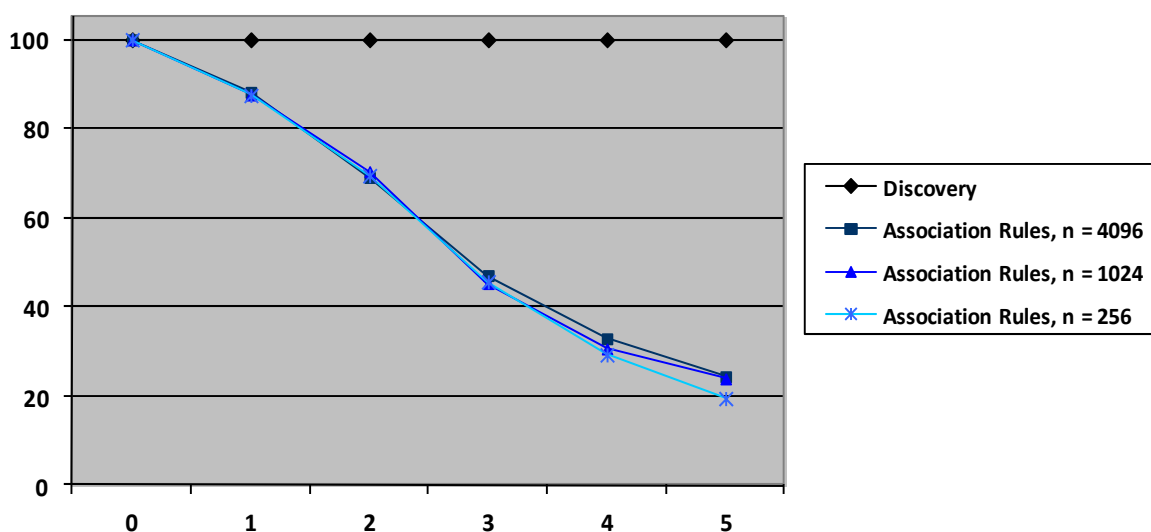


Рис. 3. Процент правильно предсказанных значений при анализе тестовой таблицы 1.

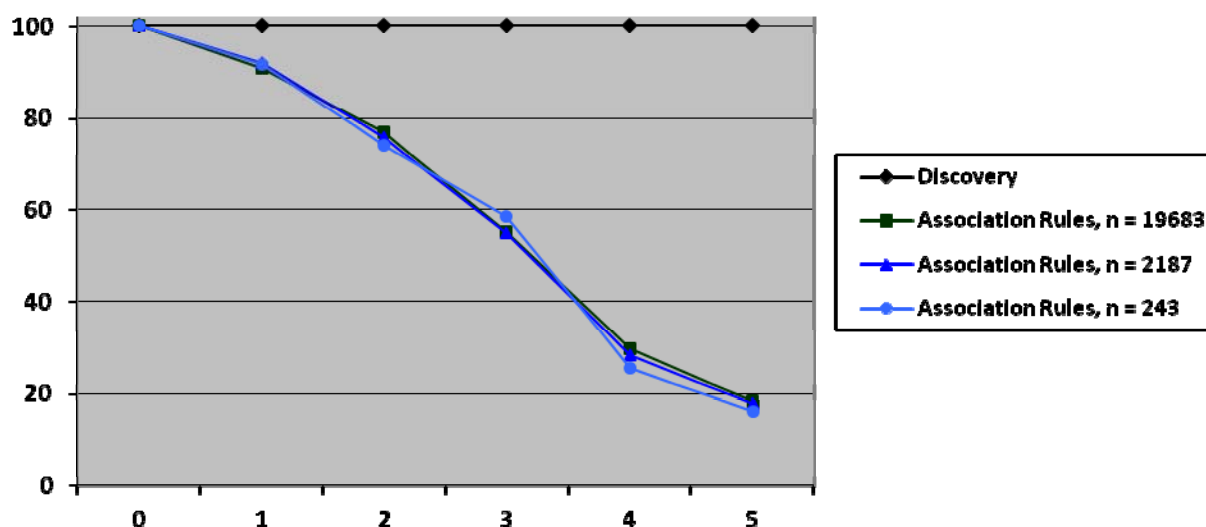


Рис. 4. Процент правильно предсказанных значений при анализе тестовой таблицы 2.

Как видим, при увеличении числа колонок  $R_1 - R_5$  со случайными данными, используемых в качестве входных данных, качество предсказания алгоритма Association Rules значительно падает. Размер тестовой таблицы незначительно влияет на результат.

### Приложение 3

На следующем графике показано, как растет количество правил с  $UB > 0.85$ , обнаруженных алгоритмом Association Rules, при добавлении колонок  $R_1 - R_5$  в качестве входных колонок. Анализируются тестовые таблицы с наложением 3% шума.

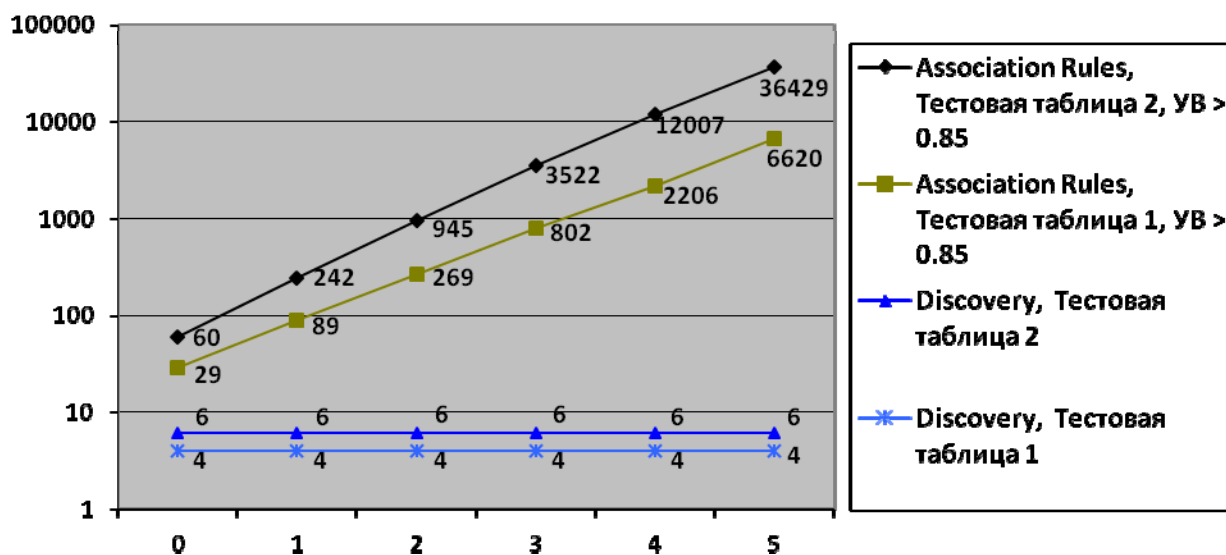


Рис. 5. Количество правил, обнаруженных алгоритмом Association Rules и системой Discovery.

### Приложение 4

Сравним качество предсказания системы “Discovery” и алгоритма Association Rules на тестовой таблице 2 с шумом 0%, 2% и 3%. Как видим, система “Discovery” дает наиболее близкое к идеальному предсказание, причем качество прогнозирования не зависит от количества колонок со случайными данными, используемых в качестве входных данных, а зависит только от величины шума. Заметим, что модель, обученная с помощью алгоритма “Discovery” на данных с шумом, будет давать 100% верные предсказания на данных без шума. Алгоритм Association Rules дает аналогичное Discovery качество предсказания в случае, когда случайные колонки не участвуют в

обучении модели. При добавлении в модель случайных колонок  $R_1 - R_5$  качество прогнозирования Association Rules значительно падает.

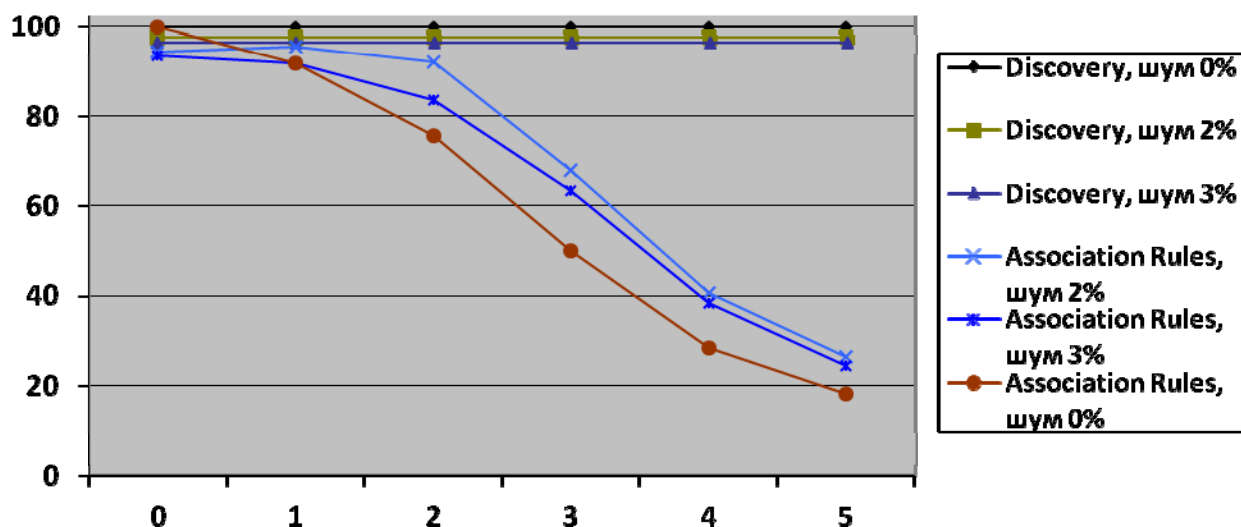


Рис. 6. Процент правильно предсказанных значений при анализе тестовой таблицы 2

Заметим, что качество предсказания Association Rules на данных без шума при добавлении 2 и более колонок  $R_1 - R_5$  заметно хуже, чем на данных с шумом 2% или 3%. Это объясняется тем, что на данных без шума Association Rules обнаруживает огромное количество «равноправных» правил с  $UB = 1$ , выбрать верное из которых не представляется возможным.

Витяев Е.Е.<sup>3</sup>,  
Неупокоев Н.В.

## Математическая модель восприятия и образа

Существующие формализации образов и зрительного восприятия никак не связаны с психологией восприятия и с когнитивными процессами порождения образов. Тем самым нет связи между воспринимаемыми объектами и теми образами, которые они порождают. Для моделирования процессов порождения образов, их изменений и развития в процессе восприятия внешнего мира нужна адекватная формализация образа и восприятия, которая бы основывалась на психологии восприятия. В данной работе предлагается такая формализация, где образ и восприятие рассматриваются, в соответствии с существующими представлениями, как непрерывный процесс предвосхищения (предсказания) образом поступающих стимулов и проверка предсказаний на соответствие реальным стимулам. Показывается, что формализацией этого процесса являются неподвижные точки предсказаний. В работе приведена математическая модель неподвижных точек, алгоритм их обнаружения и эксперименты, демонстрирующие возможности данной модели. Проведено сравнение с сетями Хопфилда, которые являются наиболее похожими на неподвижные точки нейронными сетями.

**Ключевые слова:** Образ, восприятие, предсказание, интеллектуальный анализ данных, Data Mining.

**1. Введение.** Теория Функциональных Систем П.К.Анохина начинается с принципа опережающего отражения действительности. Мозг непрерывно во времени предвосхищает события окружающей среды и одновременно контролирует акцептором результатов действия правильность сделанных предсказаний.

В восприятии предвосхищение (антиципация) непрерывно во времени сравнивает «образ» («образ мира») с наличной стимуляцией и является процессом активного движения от «образа» к внешнему миру – непрерывным во времени процессом проверки предсказаний «образа» на соответствие стимулам внешнего мира. Только если все многочисленные предсказания будут совпадать с реальными стимулами непрерывно во времени, только тогда есть восприятие [1]. «Все это позволяет нарисовать следующую картину хода познавательной деятельности на уровне воспри-

<sup>3</sup> Эта работа поддержана Российским Гуманитарным научным фондом, проект № 12-01-12026, Российским Фондом Фундаментальных Исследований № 11-07-00560-а, интеграционными проектами СО РАН № 3, 87, 136 и программой президента Российской Федерации поддержки научных школ НШ-276.2012.1.