УДК 681.3:004.8

Н. И. Фирсов

Институт систем информатики им. А. П. Ершова СО РАН
пр. Акад. Лаврентьева, 6, Новосибирск, 630090, Россия
E-mail: nickfirsov@gmail.com

СРАВНЕНИЕ СИСТЕМЫ «DISCOVERY» С MICROSOFT ASSOCIATION RULES [*]

Сравнивается система «Discovery» с алгоритмом Microsoft Association Rules. Показано, что система «Discovery» больше подходит для обнаружения закономерностей и прогнозирования, чем Association Rules, а также что она позволяет обнаруживать знания в сильно зашумленных данных, например финансовых.

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

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 [1; 2]: 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 неспособны 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 [1–3], снимающий практически все ограничения, свойственные KDD&DM-методам. Данный подход использует логику первого порядка, что позволяет работать практически с любыми типами данных и обнаруживать любые виды гипотез, а также извлекать из данных максимально полный объем знаний.

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 [3], обнаруживает все максимально специфические правила, позволяющие предсказывать без противоречий.

Наиболее близким к данной версии системы «Discovery» подходом можно считать поиск ассоциативных правил (Microsoft Association Rules) [4], поскольку закономерности представляются также в форме логических правил. Тем не менее между ними существует ряд принципиальных отличий.

В данной работе ставится задача сравнения системы «Discovery» с алгоритмом Microsoft Association Rules. Мы покажем, что система «Discovery» больше подходит для обнаружения закономерностей и прогнозирования, чем Association Rules, а также то, что, в отличие от ал-

[*] Работа выполнена при финансовой поддержке РФФИ (проект № 11-07-00560а), интеграционных проектов СО РАН № 47, 115, 119, а также Совета по грантам Президента РФ и государственной поддержке ведущих научных школ (проект НШ-3606.2010.1).

горитма Association Rules, метод «Discovery» позволяет обнаруживать знания в сильно зашумленных данных, какими, например, являются финансовые временные ряды [2].

Для экспериментального сравнения Association Rules с системой «Discovery» последняя 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).

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 наборы, 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 Поддержка (f) / Поддержка (x) > Minimum Probability, тогда $x \Rightarrow y$ является ассоциативным правилом с условной вероятностью Prob = Поддержка (f) / 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, т. е. 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.

Система «Discovery»

Общая схема работы алгоритма поиска закономерностей. Алгоритм поиска закономерностей системы «Discovery» реализует метод семантического вероятностного вывода, позволяющего находить все максимально специфические и максимально вероятные закономерности в данных [1]. Определим на высказываниях языка первого порядка вероятность, как описано в [5].

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

1) $C_i = (A_1^i \& \ldots \& A_{k_i}^i \Rightarrow P)$, $i = 1, \ldots, n$;

2) C_i является подправилом правила C_{i+1}, т. е. $\{A_1^i, \ldots, A_{k_i}^i\} \subset \{A_1^{i+1}, \ldots, A_{k_{i+1}}^{i+1}\}$;

3) $\mathrm{Prob}(C_i) < \mathrm{Prob}(C_{i+1})$, $i = 1, 2, \ldots, n-1$, где $\mathrm{Prob}(C_i)$ – условная вероятность (УВ) правила, $\mathrm{Prob}(C_i) = \mathrm{Prob}(P / A_1^i \& \ldots \& A_{k_i}^i) = \mathrm{Prob}(P \& A_1^i \& \ldots \& A_{k_i}^i) / \mathrm{Prob}(A_1^i \& \ldots \& A_{k_i}^i)$;

4) C_i – вероятностные законы (ВЗ), т. е. для любого подправила $C' = (A_1 \& \ldots \& A_j \Rightarrow P)$ правила C_i, $\{A_1, \ldots, A_j\} \subset \{A_1^i, \ldots, A_{k_i}^i\}$ выполнено неравенство $\mathrm{Prob}(C') < \mathrm{Prob}(C_i)$;

5) C_n – сильнейший вероятностный закон (СВЗ), т. е. правило C_n не является подправилом никакого другого вероятностного закона.

Вероятностные неравенства в пунктах 3–4 проверяются на данных с помощью точного критерия независимости Фишера и критерия Юла [6; 7].

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, так как пункты 2 и 4 определения СВВ подразумевают большое пространство перебора. Для уменьшения перебора применяются 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 P, 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 (рис. 1).

Перед началом обучения модели 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 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.

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.

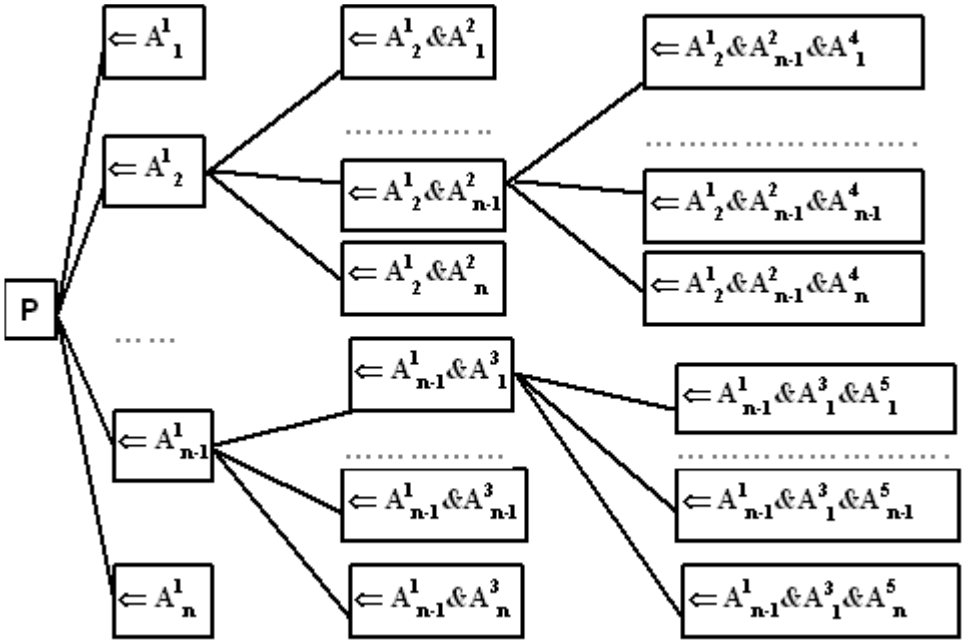

Рис. 1. Дерево СВВ, включающее все СВВ, содержащие в заключении атом *P*

Множество всех МСЗ обладает таким важным свойством, как потенциальная непротиворечивость [1].

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 *T*; 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 *T*, добавляем в очередь *Q*;
- процесс повторяется, пока очередь непуста, т. е. еще можно получить новые ВЗ путем добавления 1-го предиката в условную часть правила. Правила, соответствующие элементы 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 [1. 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.

Теоретическое сравнение

В качестве модельных данных используются следующие тестовые таблицы. Тестовая таблица 1 имеет три значимых колонки F_1, F_2, F_3 определяемые следующим выражением:

$$F_1(i) = \begin{cases} 1, \text{при i} \in (1, 512k), \\ 0, \text{при i} \in (512k+1, 1024k), \end{cases}$$

$1024k$ – количество записей в таблице;

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

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

и колонку *P*, используемую в качестве целевого признака:

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

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:

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

$2187k$ – количество записей в таблице;

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

$$F_3(i) = \begin{cases} 1, \text{ при } i \in \bigcup\limits_{j=0}^{2} \begin{pmatrix} (1+729kj, 81k+729kj) \cup (1+243k+729kj, 324k+729kj) \\ \cup (1+486k+729kj, 567k+729kj) \end{pmatrix}, \\ 2, \text{ при } i \in \bigcup\limits_{j=0}^{2} \begin{pmatrix} (1+81k+729kj,\ 162k+729kj) \cup (1+324k+729kj, 405k+729kj) \\ \cup (1+567k+729kj, 648k+729kj) \end{pmatrix}, \\ 3, \text{ при } i \in \bigcup\limits_{j=0}^{2} \begin{pmatrix} (1+162k+729kj,\ 243k+729kj) \cup (1+405k+729kj, 486k+729kj) \\ \cup (1+648k+729kj, 729k+729kj) \end{pmatrix}; \end{cases}$$

и колонку P, используемую в качестве целевого признака:

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

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, так как в обоих подходах закономерности представляются в форме логических правил, тем не менее между ними существуют принципиальные отличия.

1. 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 & … & $D \Rightarrow C$, которые получаются из правила A & $B \Rightarrow C$ добавлением дополнительных условий $D, F, \ldots$: 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 $(F_1 = 1 \wedge F_2 = 1 \wedge F_3 = 1) \Rightarrow P = 1$ с УВ = 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, следовательно, ложно предсказывать 1, когда колонка *P* 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 %.

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

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

так как 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.

2. 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 приведет 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, так как 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

$$(F_1=1 \wedge F_2=1 \wedge F_3=1) \Rightarrow P=1 \text{ с УВ} = p,$$

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.

Экспериментальное сравнение

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 табл. 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 табл. 1 (228 на тестовой таблице 2). На рис. 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. Как видим, количество правил, найденных Association Rules, экспоненциально растет при добавлении новых колонок.

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

Отметим, что на тестовых табл. 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.

На рис. 3, 4 приводится сравнение качества предсказания алгоритма Association Rules и системы «Discovery». 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.

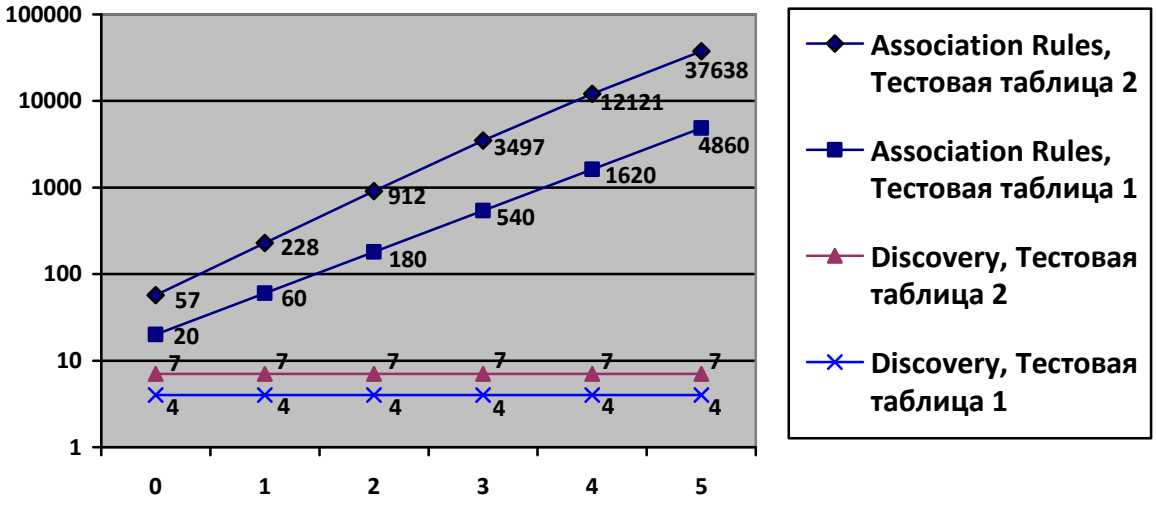

Рис. 2. Количество правил с УВ = 1, обнаруженных алгоритмом Association Rules. На горизонтальной оси отмечено количество колонок $R_1 - R_5$ со случайными данными, используемых (в дополнение к F_1, F_2, F_3) в качестве входных данных

Таблица 1

Процент правильно предсказанных алгоритмом Association Rules значений на тестовой табл. 1, %

Association Rules	0	1	2	3	4	5
n = 256	100	87,5	69,53	45,70	29,29	19,53
n = 1 024	100	87,59	70,41	45,31	30,56	23,92
n = 4 096	100	88,15	69,14	47,07	32,86	24,43

Таблица 2

Процент правильно предсказанных алгоритмом Association Rules значений на тестовой табл. 2, %

Association Rules	0	1	2	3	4	5
n = 243	100	91,81	74,16	58,46	25,68	16,26
n = 2 187	100	91,95	75,76	55,05	28,30	17,92
n = 1 9683	100	91,06	76,85	55,19	29,71	18,46

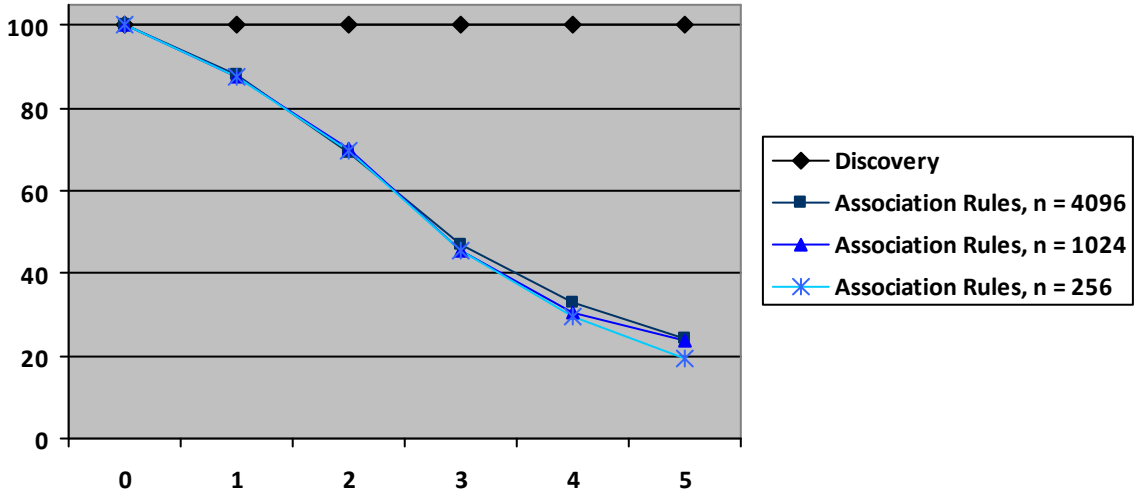

Рис. 3. Процент правильно предсказанных значений при анализе тестовой таблицы 1

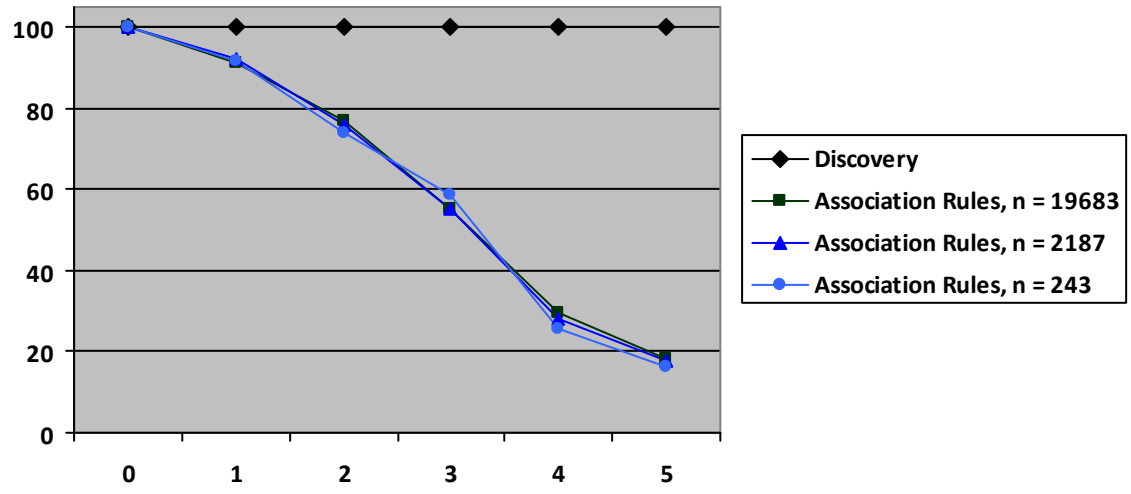

Рис. 4. Процент правильно предсказанных значений при анализе тестовой таблицы 2

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.

Рассмотрим тестовые табл. 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 табл. 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 табл. 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 табл. 2 Association Rules обнаруживает 60 правил с УВ > 0,85, в том числе и все правила, найденные системой «Discovery». 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. Анализируются тестовые таблицы с наложением 3 % шума.

Сравним качество предсказания алгоритма Association Rules и системы «Discovery» на тестовой табл. 2 с шумом 0, 2 и 3 % (рис. 6).

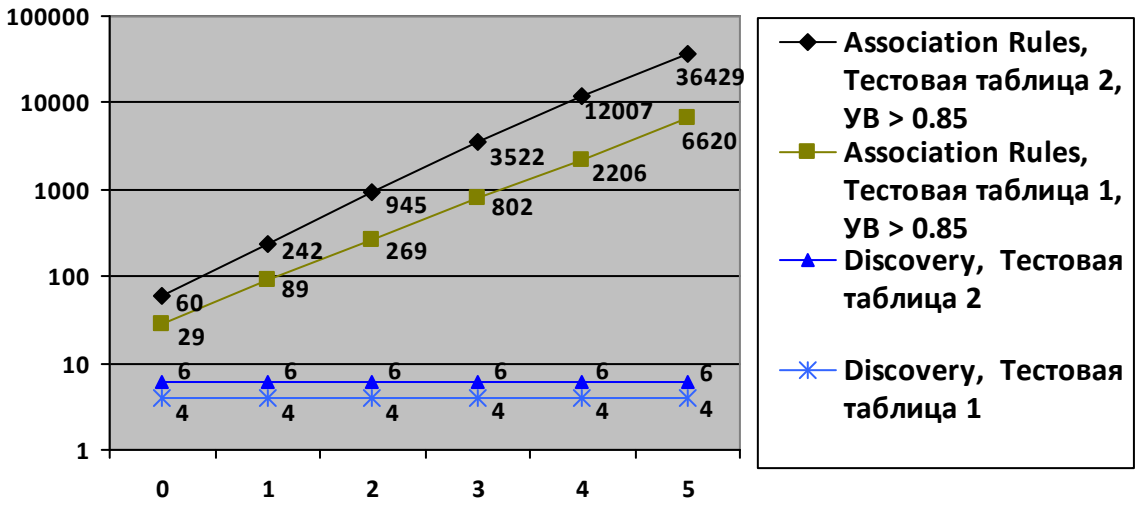

Рис. 5. Количество правил, обнаруженных алгоритмом Association Rules и системой «Discovery»

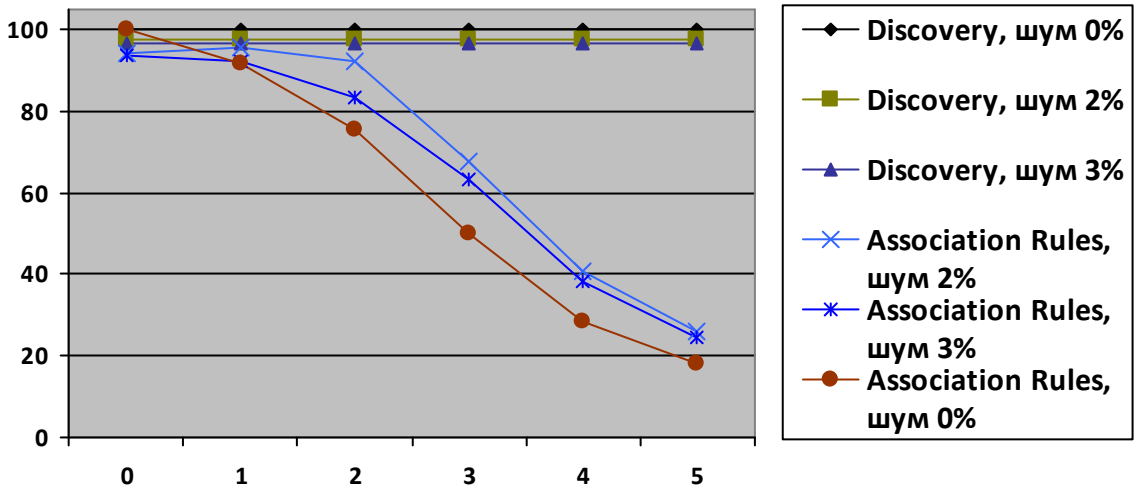

Рис. 6. Процент правильно предсказанных значений при анализе тестовой таблицы 2

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.

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 УВ = 1, выбрать верное из которых не представляется возможным.

Список литературы

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

2. *Halpern J. Y.* An Analysis of First-Order Logic of Probability // Artificial Intelligence. 1990. С. 311–350.

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

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

5. *Vityaev E., Kovalerchuk B.* Empirical Theories Discovery Based on the Measurement Theory // Mind and Machine. 2006. Vol. 14. No. 4. P. 551–573.

6. *Vityaev E.* The Logic of Prediction // Mathematical Logic in Asia: Proc. of the 9th Asian Logic Conference (August 16–19, 2005, Novosibirsk, Russia). Singapore: World Scientific, 2006. P. 263–276.

7. *Kovalerchuk B. Y., Vityaev E. E.* Data Mining in Finance: Relational and Hybrid Methods. Kluwer, 2000. 308 p.

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

Материал поступил в редколлегию 19.06.2011

N. I. Firsov

COMPARISON OF «DISCOVERY» SOFTWARE SYSTEM VERSUS MICROSOFT ASSOCIATION RULES

We have developed a relational approach (Relational Data Mining) to methods of data mining, and the Discovery software system which relaxes almost all restrictions peculiar to the KDD & DM techniques. The goal of this paper is to provide both theoretical and experimental comparisons of Discovery versus Microsoft Association Rules**.** We show that Discovery fits better for pattern detection and prediction tasks than the Association Rules**,** and, unlike the latter, it can detect knowledge in high-noised data such as financial time series.

Keywords: intelligent data analysis, data mining, regularities detection, knowledge «Discovery».