УДК 681.3.00:007

СРАВНИТЕЛЬНАЯ ХАРАКТЕРИСТИКА ПРОСТЫХ И РАСШИРЕННЫХ АССОЦИАТИВНЫХ ПРАВИЛ ДЛЯ ПРИЗНАКОВ ОБЪЕКТОВ В БАЗАХ ДАННЫХ

Е.В. Титова
(представил д.т.н., проф. В.М. Бильчук)

Проводится анализ уровней поддержки и доверия простых и расширенных ассоциативных правил. Оценивается влияние расширения ассоциативного правила на его точность и поддержку. Формулируются критерии генерации дискриминирующих ассоциаций с максимальной точностью и характеристических ассоциаций с максимальной поддержкой.

Введение. В сфере компьютерных технологий недавно оформилось новое направление – выявление знаний в данных (Knowledge discovery in databases (KDD) and Data mining). Обнаружение знаний в базах данных определяют как процесс получения знаний путем концентрации и обобщения информации, установление новых, потенциально полезных зависимостей между объектами (записями в БД) [4]. Потребность в новой методологии обусловлена нерешенностью ряда важных задач, таких как прогнозирования процессов и объяснения явлений. На основе старых технологий далеко не везде удавалось повысить эффективность управления и принятия решений. Таким образом, встала **проблема выделения неявной, предварительно неизвестной и полезной информации из данных**.

Одной из типичных задач KDD является задача открытия (вывода) ассоциативных правил (association rules). Ассоциативное правило определяется как утверждение вида $A \to B$, где A и B – некоторые множества признаков объектов в базе данных (БД), причем данное утверждение должно обладать определенным уровнем поддержки (Support) и доверия (Confidence).

Для ассоциативного правила (АП) поддержка определяется как количество записей в данной БД, удовлетворяющих этому правилу, а доверие – как отношение между количеством записей, удовлетворяющих правилу и количеством записей, удовлетворяющих его левой части. Если $S_1 \to S_2$ является ассоциативным правилом, то должны выполняться следующие условия:

$$\mathrm{Sup}(S_1 \cap S_2) \geq \mathrm{minSup};$$

$$\frac{\mathrm{Sup}(S_1 \cap S_2)}{\mathrm{Sup}(S_1)} \geq \mathrm{minConf},$$

где Sup(S) = M(S) – количество объектов (записей в БД), обладающих признаком S (т.е. мощность множества объектов, обладающих признаком S); minSup, minConf – установленные минимальные уровни поддержки и доверия.

Для нахождения АП разработана группа алгоритмов, одним из которых является алгоритм построения дерева покрытий [1]. **Анализ литературы** показывает, что данные алгоритмы позволяют генерировать ассоциативные правила только для бинарных признаков. В реляционных же БД признаки объектов нередко являются категорированными величинами, т.е. могут принимать значения из множества возможных (разбиты на категории). Например, признак "возраст" может иметь следующие категории: до 20 лет, 20 – 30 лет, 30 – 50 лет, старше 50. В [3] предлагается усовершенствование алгоритма построения дерева покрытий для случая категорированных признаков.

Безусловно, генерация абсолютно всех ассоциативных зависимостей для текстовой или реляционной БД является чрезвычайно громоздкой задачей. Поэтому перед началом нахождения АП задается минимум управляющей информации, например, целевые переменные, необходимые уровни поддержки и доверия, группы признаков, для которых будет устанавливаться ассоциация. Однако, даже при небольшом количестве признаков, количество сгенерированных АП может быть достаточно велико [3]. Поэтому на этапе построения ассоциаций встает **задача ограничения количества правил**, исходя из потребностей максимизации уровня доверия и поддержки.

Целью статьи является сравнительный анализ простых и расширенных ассоциативных правил, который позволяет отбирать АП, представляющие больший интерес с точки зрения точности и поддержки.

1. Описание признаков объектов в базах данных с помощью конечных предикатов. В случае, когда признаки объектов (записей) в БД являются категорированными величинами, удобным способом их описания являются конечные предикаты.

Пусть признак X_j объекта может принимать значения из множества $A=\{a_1, a_2, \ldots, a_n\}$, тогда такой признак может быть описан с помощью конечного предиката следующим образом:

$$X_j^{a_i} = \begin{cases} 1, & \text{если } X_j = a_i; \\ 0 & \text{если } X_j \neq a_i. \end{cases}$$

Предикат $X_j^{a_i}$ как бы "узнает" значение a_i признака X_j среди всех возможных значений $a \in A$ [2].

Имеют место следующие тождества:

$$X^{a_1} \vee X^{a_2} \vee \ldots \vee X^{a_n} \equiv 1; \quad (1)$$

$$X^{a_i} X^{a_j} \equiv 0. \quad (2)$$

Действительно, признак X принимает какое-либо значение из возможного набора, а, следовательно, всегда один из дизъюнктивных членов, а вместе с ним и вся дизъюнкция (1) обращается в 1, и признак X не может принимать одновременно два различных значения (2).

Подобное описание позволяет не использовать отрицание признаков: если признак X_j не принимает значение a_i, то он принимает любое значение из оставшегося набора признаков, т.е.

$$\overline{X^{a_i}} = X^{a_1} \vee \ldots X^{a_{i-1}} \vee \vee X^{a_{i+1}} \ldots \vee X^{a_n}.$$

В случае бинарных признаков для их описания используется унарный предикат, различающий два значения: X^0 и X^1.

2. Анализ уровней доверия и поддержки простых и расширенных ассоциативных правил. Простым ассоциативным правилом назовем утверждение вида

$$X_i^{a_1} \to Y_k^{b_3}.$$

Расширенным ассоциативным правилом назовем правило, получаемое из простого путем добавления новых переменных и (или) путем расширения области значения некоторых переменных, т.е.

$$(X_i^{a_1} \vee X_i^{a_3}) Z^{c_2} \to Y_k^{b_3} (T^{d_3} \vee T^{d_7}).$$

Рассмотрим, как связаны уровни поддержки и доверия простых и расширенных АП, и в каких случаях те или иные АП представляют больший интерес.

Теорема 1. Введение новых переменных в левую или правую часть ассоциативного правила не ведет к увеличению поддержки.

Доказательство. Если $S_1 \to S_2$ – является ассоциативным правилом, то $Sup(S_1 \to S_2) = M(S_1 \cap S_2)$. При введении переменной S_3:

$$Sup(S_3 S_1 \to S_2) = Sup(S_1 \to S_2 S_3) = M(S_1 \cap S_2 \cap S_3).$$

Но $M(S_1 \cap S_2 \cap S_3) \le M(S_1 \cap S_2)$, что и требовалось доказать.

Теорема 2. Введение новых переменных в правую часть АП не ведет к увеличению уровня доверия.

Доказательство. Пусть $X^{a_i} \to Y^{b_j}$ является ассоциативным правилом. Введем новую переменную Z следующим образом: $X^{a_i} \to Y^{b_j} Z^{c_k}$. Сравним уровни доверия этих правил:

$$\text{Confidence } (X^{a_i} \to Y^{b_j}) = \frac{\text{Sup}(X^{a_i} \to Y^{b_j})}{\text{Sup}(X^{a_i})} = \frac{M(X^{a_i} \cap Y^{b_j})}{M(X^{a_i})};$$

$$\text{Confidence } (X^{a_i} \to Y^{b_j} Z^{c_k}) = \frac{\text{Sup}(X^{a_i} \to Y^{b_j} Z^{c_k})}{\text{Sup}(X^{a_i})} =$$

$$= \frac{M(X^{a_i} \cap Y^{b_j} \cap Z^{c_k})}{M(X^{a_i})},$$

но $M(X^{a_i} \cap Y^{b_j} \cap Z^{c_k}) \le M(X^{a_i} \cap Y^{b_j})$.

Следовательно, $\text{Confidence } (X^{a_i} \to Y^{b_j}) \ge \text{Confidence } (X^{a_i} \to Y^{b_j} Z^{c_k})$, что и требовалось доказать.

Рассмотрим пример, когда признак X объектов может принимать значения a_1, a_2, a_3, признак Y – значения b_1, b_2, b_3 и признак Z – значения c_1, c_2, c_3 (табл. 1).

Таблица 1

Фрагмент реляционной БД

	1	2	3	4	5	6	7	8
X	a_1	a_2	a_1	a_3	a_1	a_2	a_3	a_1
Y	b_1	b_3	b_2	b_2	b_1	b_3	b_2	b_2
Z	c_1	c_1	c_2	c_3	c_1	c_1	c_2	c_1

Утверждение $X^{a_1} \to Z^{c_1}$ является ассоциативным правилом с уровнем поддержки Support = 3 и уровнем доверия Confidence = 3/4. Введение новой переменной Y следующим образом: $X^{a_1} \to Z^{c_1} Y^{b_1}$ приводит нас к ассоциативному правилу с Support = 2 и уровнем доверия Confidence = 1/2.

Из этого можно сделать вывод, что увеличение количества переменных в правой части АП в любом случае ведет к "ухудшению" правила (в общем случае к "не улучшению"). При введении дополнительных переменных в левую часть правила уровень доверия может меняться как в большую, так и в меньшую сторону. Требование максимизации доверия, таким образом, ведет к минимизации количества переменных в правой части АП.

Обратимся к рассмотрению уровней доверия и поддержки АП, получаемых путем расширения области значения переменных в левой и правой частях правила.

Как было показано в [3], уровень доверия (Confidence) правила вида $S \to Z^{c_1} \vee Z^{c_2} \vee \ldots \vee Z^{c_k}$ равен сумме уровней доверия правил, полученных из него путем последовательного отбрасывания из правой части различных k–1 дизъюнктивных членов. Таким образом, расширение правой части АП за счет увеличения области значения переменных ведет к повышению (не уменьшению) уровня доверия. Поддержка в этом случае также не уменьшается:

$$\mathrm{Sup}(S \to Z^{c_1} \vee Z^{c_2} \vee \ldots \vee Z^{c_k}) = M(S \cap (Z^{c_1} \cup Z^{c_2} \cup \ldots \cup Z^{c_k}) =$$

$$= M(S \cap Z^{c_1}) + M(S \cap Z^{c_2}) + \ldots + M(S \cap Z^{c_k}).$$

Для нашего примера $X^{a_1} Z^{c_1} \to Y^{b_1}$ имеет Support=2 и Confidence = 2/3, $X^{a_1} Z^{c_1} \to (Y^{b_1} \vee Y^{b_2})$ имеет Support = 3 и Confidence =1.

Отсюда можно сделать вывод, что расширение области значения переменных в правой части АП ведет к "улучшению" правила (в общем случае к "не ухудшению").

Теорема 3. Уровень доверия расширенного ассоциативного правила вида $Z^{c_1} \vee Z^{c_2} \vee \ldots \vee Z^{c_k} \to S$ может быть получен путем сложения уровней доверия простых ассоциативных правил $Z^{c_1} \to S$; $Z^{c_2} \to S$; …; $Z^{c_k} \to S$ следующим образом: числитель равен сумме числителей простых АП, а знаменатель – сумме знаменателей.

Доказательство. Рассмотрим, чему равен уровень доверия правила вида $Z^{c_1} \vee Z^{c_2} \vee \ldots \vee Z^{c_k} \to S$.

$$\mathrm{Confidence}\,(Z^{c_1} \vee Z^{c_2} \vee \ldots \vee Z^{c_k} \to S) =$$

$$= \frac{M((Z^{c_1} \cup Z^{c_2} \cup \ldots \cup Z^{c_k}) \cap S)}{M(Z^{c_1} \cup Z^{c_2} \cup \ldots \cup Z^{c_k})}.$$

Так как, $Z^{c_1}, Z^{c_2}, .. Z^{c_k}$ описывают взаимно непересекающиеся множества, то

$$M((Z^{c_1} \cup Z^{c_2} \cup \ldots \cup Z^{c_k}) \cap S) =$$

$$= M(Z^{c_1} \cap S) + M(Z^{c_2} \cap S) + \ldots + M(Z^{c_k} \cap S),$$

$$M(Z^{c_1} \cup Z^{c_2} \cup \ldots \cup Z^{c_k}) = M(Z^{c_1}) + M(Z^{c_2}) + \ldots + M(Z^{c_k}).$$

Следовательно,

$$\text{Confidence}\ (Z^{c_1} \vee Z^{c_2} \vee \ldots \vee Z^{c_k} \to S) =$$

$$= \frac{M(Z^{c_1} \cap S) + M(Z^{c_2} \cap S) + \ldots + M(Z^{c_k} \cap S)}{M(Z^{c_1}) + M(Z^{c_2}) + \ldots + M(Z^{c_k})}.$$

Но $M(Z^{c_1} \cap S) + M(Z^{c_2} \cap S) + \ldots + M(Z^{c_k} \cap S)$ является суммой числителей уровней доверия простых АП $Z^{c_1} \to S$, $Z^{c_2} \to S$, … , $Z^{c_k} \to S$, а $M(Z^{c_1}) + M(Z^{c_2}) + \ldots + M(Z^{c_k})$ – суммой знаменателей.

Следствие 1. Уровень доверия правила вида $Z^{c_1} \vee Z^{c_2} \vee \ldots \vee Z^{c_k} \to S$ меньше или равен сумме уровней доверия простых АП $Z^{c_1} \to S$, $Z^{c_2} \to S$,… , $Z^{c_k} \to S$ (при сложении числителей и знаменателей нескольких дробей результирующая дробь не увеличивается).

Следствие 2. Если уровень доверия правила вида $(Z^{c_1} \vee Z^{c_2} \vee \ldots \vee Z^{c_k})(X^{a_1} \vee X^{a_2} \vee .. X^{a_i}) \ldots \to S$ равен 1, то уровни доверия всех простых правил, полученных путем отбрасывания из каждой дизъюнкции вида $Z^{c_1} \vee Z^{c_2} \vee \ldots \vee Z^{c_k}$ любых k–1 дизъюнктивных членов, равны 1 или 0.

Проиллюстрируем это на следующих примерах:

Пример 1.

$Y^{b_1} \to Z^{c_1}$ – Confidence = 2/2 = 1;

$Y^{b_2} \to Z^{c_1}$ – Confidence = 1/4;

$(Y^{b_1} \vee Y^{b_2}) \to Z^{c_1}$ – Confidence = (2+1)/(2+4) = 3/6 = 0,5.

Пример 2.

Рассмотрим АП: $(X^{a_1} \vee X^{a_3}) \to (Y^{b_1} \vee Y^{b_2})$. Его уровень доверия: Confidence = 1. Следовательно, уровни доверия правил $X^{a_1} \to (Y^{b_1} \vee Y^{b_2})$ и $X^{a_3} \to (Y^{b_1} \vee Y^{b_2})$ также равны 1, в чем можно убедиться, обратившись к табл. 1.

Можно сделать следующий вывод: максимизация уровня доверия АП требует минимизации области значения переменных в левой части правила.

Уровень поддержки расширенных АП, получаемых путем увеличения области значения переменных в обеих частях правила, не уменьшается:

$$M(Z^{c_1} \cap S) + M(Z^{c_2} \cap S) + \ldots + M(Z^{c_k} \cap S) \geq M(Z^{c_i} \cap S), \quad i = \overline{1,k}.$$

Выводы. Таким образом, для получения АП с максимальными уровнями доверия необходимо стремиться к уменьшению количества переменных в правой части правила и к расширению области их значения. Что касается левой части правила, то, увеличивая количество переменных, мы будем проигрывать в поддержке, однако, можем выигрывать в уровне доверия. И наоборот, расширяя область значения переменных, мы увеличиваем поддержку, но проигрываем в уровне доверия. Для получения максимально точных правил с минимальной поддержкой (генерация "особых" случаев) необходимо максимизировать количество переменных в правой части и минимизировать их области значения, а также количество переменных в левой части АП (дискриминирующее правило). Для получения характеристических правил, от которых требуется простота и наглядность, необходимо уменьшать количество переменных, расширяя при этом области их значения. Это приведет к получению правил с максимальной поддержкой. Доказанные теоремы позволяют рассчитывать уровни доверия расширенных ассоциативных правил.

В заключение следует отметить, что выявление регулярностей в базах данных и формирование "обобщенных видов" данных, к которым относятся ассоциативные правила, позволяет пользователю обозреть огромный объем информации. Средства вывода импликативных правил обеспечивают логическое уплотнение информации и могут выявить полезные и понятные связи. Технологии KDD могут найти применение не только в крупном бизнесе и исследовательских организациях, но и в органах управления.

ЛИТЕРАТУРА

1. *A. Amir, R. Feldman, R. Kashi. A new and versatile method for association generation // Information Systems. – 1997. – Vol. 22, № 6/7. – P. 333 – 347.*
2. *Шабанов-Кушнаренко Ю.П. Теория интеллекта. Математические средства. – Х.: Вища шк., 1984. – 143 с.*
3. *Ситников Д.Э., Титова Е.В. Метод поиска обобщенных ассоциативных зависимостей между дискретными признаками // Системи обробки інформації. – Х.: НАНУ, ПАНМ, ХВУ. – 2002. – Вип. № 6 (22). – С. 194 – 202.*
4. *Балабанов А.С. Выделение знаний из баз данных – передовые компьютерные технологии интеллектуального анализа данных. // Математичні машини і системи. – 2001. – № 1, 2. – С. 40 – 54.*

Поступила 17.03.2003

***ТИТОВА Елена Витольдиевна**, младший научный сотрудник научного центра при ХВУ. В 1988 году окончила Харьковский институт радиоэлектроники. Область научных интересов – выделение знаний из баз данных.*