УДК 681.3.00:007

ВЛИЯНИЕ СТАНДАРТНЫХ ПАРАМЕТРОВ АССОЦИАТИВНОГО ПРАВИЛА НА ЕГО ИНФОРМАТИВНОСТЬ

к.т.н. Д.Э. Ситников, Е.В. Титова
(представил д.т.н., проф. О.И. Сухаревский)

Проводится анализ зависимости информативности ассоциативных правил от стандартных параметров – уровней поддержки, доверия и улучшения. Показывается, что правило, обладающее лучшими характеристиками, не всегда является более интересным с точки зрения сообщаемой информации. Предлагается определение R-"интересности" ассоциативного правила на основе его информативности.

Введение. К настоящему времени в различных областях человеческой деятельности накоплено много данных, которые могут стать источником получения новых знаний, коммерчески полезной информации. Эти данные нуждаются в серьезной обработке, так как сами по себе они, как правило, не имеют практической ценности. Таким образом, проблема извлечения знаний из данных (Data mining или Knowledge discovery in databases) привлекает в последние годы большое внимание исследователей в области компьютерных технологий и искусственного интеллекта.

Генерация ассоциативных правил – одна из подзадач, возникающих в процессе извлечения знаний из данных. Ассоциативное правило определяется как выражение (шаблон, паттерн) вида $X \to Y$, где X и Y – некоторые множества атрибутов (признаков объектов в базе данных). Ассоциативные правила выявляют закономерности в данных, позволяющие, например, отличать нормально развивающиеся ситуации от подозрительных и аварийных. Они могут использоваться для анализа данных и принятия решений, либо служить исходным материалом для построения баз знаний.

Анализ литературы. Оценка полученных зависимостей (ассоциативных правил) ставит своей целью выделить во множестве найденных паттернов такие, которые интересны, пригодны для практического использования, несут новые знания. Для оценки паттернов имеется много метрик. Стандартными (принятыми в литературе по Data mining) параметрами ассоциативных правил являются поддержка (Support) и доверие (Confidence) [1 – 4]. Менее распространенной, однако, очень важной, как показано в [5],

характеристикой является улучшение (Improvement) [6]. Support определяется как соотношение количества объектов, обладающих признаками X и Y, к общему количеству объектов в базе данных (БД) – с точки зрения теории вероятности: P(XY). Confidence – как отношение количества объектов, обладающих признаками X и Y к количеству объектов, обладающих признаком X (Confidence еще называют вероятностью правила). С точки зрения теории вероятности Confidence соответствует P(X/Y). Improvement определяют как отношение Confidence правила к поддержке его правой части $\frac{P(X/Y)}{P(Y)} = \frac{P(XY)}{P(X)P(Y)}$. В [5] предложена интегральная характеристика ассоциативного правила – информативность, которая рассчитывается на основании трех вышеприведенных величин.

Цель статьи. Проанализировать зависимость информативности ассоциативного правила от стандартных параметров: Support, Confidence и Improvement. Рассмотреть "интересность" правила с точки зрения теории информации.

Оценка влияния стандартных параметров ассоциативного правила на его информативность. Выражение для интегральной характеристики ассоциативного правила – полной взаимной информации было получено в [5] (для краткости вместо Support (X → Y), Confidence (X → Y), Improvement (X → Y) будем писать Sup, Conf и Imp соответственно):

$$I_{X\leftrightarrow Y} = \mathrm{Sup}\cdot\log_2(\mathrm{Imp}) + \left(\frac{\mathrm{Sup}}{\mathrm{Conf}} - \mathrm{Sup}\right)\cdot\log_2\left(\frac{\mathrm{Imp}\cdot(1-\mathrm{Conf})}{\mathrm{Imp}-\mathrm{Conf}}\right) + \left(\frac{\mathrm{Conf}}{\mathrm{Imp}} - \mathrm{Sup}\right)\cdot\log_2\left(\frac{\mathrm{Conf}-\mathrm{Sup}\cdot\mathrm{Imp}}{\mathrm{Conf}-\mathrm{Sup}}\right) + \left(1+\mathrm{Sup}-\frac{\mathrm{Sup}}{\mathrm{Conf}}-\frac{\mathrm{Conf}}{\mathrm{Imp}}\right)\cdot\log_2\left(1+\mathrm{Sup}-\frac{\mathrm{Sup}}{\mathrm{Conf}}-\frac{\mathrm{Conf}}{\mathrm{Imp}}\right)\Bigg/\left(\left(1-\frac{\mathrm{Sup}}{\mathrm{Conf}}\right)\left(1-\frac{\mathrm{Conf}}{\mathrm{Imp}}\right)\right). \quad (1)$$

Эта характеристика позволяет сравнивать между собой ассоциативные правила, обладающие разными показателями уровней доверия, поддержки и улучшения.

В [5] также было показано, что на три стандартные характеристики накладываются следующие ограничения:

$$1)\ 0 \le \mathrm{Sup} \le 1; \quad (2)$$

$$2)\ \mathrm{Sup} \le \mathrm{Conf} \le 1; \quad (3)$$

$$3)\ \frac{\mathrm{Conf}^2}{\mathrm{Sup}\cdot\mathrm{Conf}+\mathrm{Conf}-\mathrm{Sup}} \le \mathrm{Imp} \le \frac{\mathrm{Conf}}{\mathrm{Sup}}. \quad (4)$$

Рассмотрим, какое влияние оказывают Sup, Conf и Imp на информа-

тивность правила. Традиционно считается, что чем больше у ассоциации Sup, Conf и Imp, тем правило лучше. В [3] вводится понятие "интересности" правила, основанное на том, что если Sup или Conf выше в R раз, чем ожидаемое значение, то такое ассоциативное правило "интересно" (R – показатель "интересности"). Ожидаемые значения Sup и Conf рассчитываются с учетом таксономии признаков, т.е. для признаков-потомков, исходя из значений Sup и Conf для признаков-предков. В качестве примера приводятся два ассоциативных правила, отражающие зависимость продаж продуктов для супермаркета:

1) Молоко → Каша (8% Sup, 70% Conf).

Если признак "Молоко" является предком признака "Сливки" и около 25% продаж молочных продуктов составляют сливки, то ожидаемое правило:

2) Сливки → Каша (2% Sup, 70% Conf).

Если реальные показатели второго правила близки к ожидаемым, то такое правило считается "неинтересным" (избыточным), т.е. не сообщающим никакой дополнительной информации и менее общим.

Покажем, что для случая, когда речь идет о вероятности (уровне доверия), т.е., когда Conf одного правила превышает Conf другого, мы не можем однозначно утверждать, что вторая ассоциация менее "интересна" и сообщает меньше информации.

Построим график функции полной взаимной информации при фиксированных значениях Sup и Imp, изменяя Conf в пределах, определенных неравенством (3) – рис. 1.

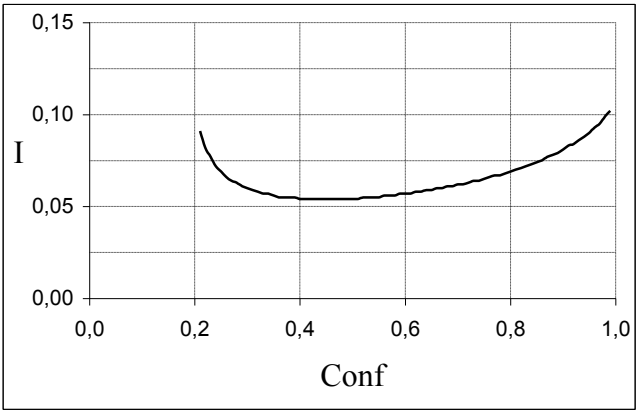

Рис. 1. Зависимость информативности правила от уровня доверия (при Sup = 0,1; Imp = 2)

Видно, что график имеет минимум. Чтобы найти значение Conf, соответствующее минимуму информации, продифференцируем функцию

информации, полагая Sup и Imp параметрами

$$\frac{\partial I_{X\leftrightarrow Y}}{\partial Conf} = \frac{Sup}{Conf^2} \times \log_2\left(\frac{Conf\cdot Imp + Sup\cdot Conf\cdot Imp - Sup\cdot Imp - Conf^2}{Conf\cdot Imp + Sup\cdot Conf\cdot Imp - Sup\cdot Imp - Conf^2\cdot Imp}\right) +$$

$$+\frac{1}{Imp}\cdot\log_2\left(\frac{Conf\cdot Imp + Sup\cdot Conf\cdot Imp - Sup\cdot Imp^2 - Conf^2}{Conf\cdot Imp + Sup\cdot Conf\cdot Imp - Sup\cdot Imp - Conf^2}\right).$$

Производная обращается в нуль в точке $Conf = \sqrt{Sup\cdot Imp}$. Для нашего примера (Sup = 0,1; Imp = 2) минимальная информация будет при Conf ≈ 0,447.

Таким образом, ассоциативное правило, имеющее больший уровень доверия, не всегда будет более "интересным" с точки зрения информативности. Например, при Sup = 0,3 и Imp = 2, правило с Conf = 0,6 будет сообщать больше информации ($I_{X\leftrightarrow Y} \approx 0{,}4$), чем правило с Conf = 0,78 ($I_{X\leftrightarrow Y} \approx 0{,}3$), хотя уровень доверия второго в 1,3 раза выше.

Рассмотрим влияние такой характеристики как Sup на информативность ассоциативного правила. Продифференцируем функцию полной взаимной информации, полагая Conf и Imp параметрами

$$\frac{\partial I_{X\leftrightarrow Y}}{\partial Sup} = \frac{1}{Conf} \times \log_2\left(\frac{Imp\,(1-Conf)(Conf-Sup)}{Conf\cdot Imp + Sup\cdot Conf\cdot Imp - Sup\cdot Imp - Conf^2}\right) -$$

$$-\log_2\left(\frac{(1-Conf)(Conf-Sup\cdot Imp)}{Conf\cdot Imp + Sup\cdot Conf\cdot Imp - Sup\cdot Imp - Conf^2}\right).$$

При Imp > 1: $\partial I_{X\leftrightarrow Y} \big/ \partial Sup > 0$.

Действительно, выражение под первым логарифмом при Imp > 1 будет больше выражения под вторым логарифмом. В дополнение к этому, первый логарифм умножается на величину, не меньшую единицы: 1/Conf.

График функции взаимной информации в зависимости от Sup (при Conf и Imp > 1 – параметрах) не имеет экстремумов (рис. 2).

Можно сделать вывод, что ассоциативное правило с более высокой поддержкой (при равных Conf и Imp) несет больше информации. Таким образом, для примера, приведенного в [3], второе правило: Сливки → Каша (2% Sup, 70% Conf) действительно является менее информативным, чем правило Молоко → Каша (8% Sup, 70% Conf) при условии, что Imp > 1 (т.к. Conf (первого правила) = Conf (второго правила) и правые части равны, то Imp (первого правила) = Imp (второго правила)). Влияние такой характеристики как Imp на информативность ассоциативного правила оценим аналогично: продифференцировав функцию полной взаимной информации, полагая Sup и Conf параметрами

$$\frac{\partial I_{X\leftrightarrow Y}}{\partial Imp} = \frac{Conf}{Imp^2} \times \log_2\left(\frac{Conf \cdot Imp + Sup \cdot Conf \cdot Imp - Sup \cdot Imp - Conf^2}{Imp - Conf}\right) -$$

$$- \frac{Conf}{Imp} \cdot \log_2(Conf - Sup \cdot Imp).$$

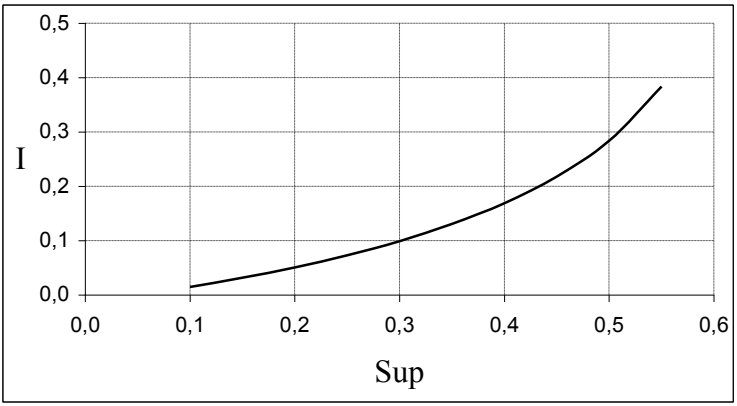

Рис. 2. Зависимость информативности правила от уровня поддержки (при Conf = 0,8; Imp = 1,5)

Производная обращается в нуль при Imp = 1, что соответствует независимости признаков (полная взаимная информация равна нулю). График зависимости информативности правила от Imp показан на рис. 3. Imp изменяется в пределах, определенных неравенством (4).

Если Imp > 1, то, как видно из графика (рис. 3), информативность правила растет с его увеличением. Однако, правило с Imp < 1 также может оказаться весьма "интересным" с точки зрения сообщаемой информации.

На рис. 4 показан график полной взаимной информации при Sup = 0,1; Conf = 0,2.

При малых значениях Imp правило будет обладать высокой информативностью.

Рассмотрим, какой вклад в величину полной взаимной информации вносят в данном случае ее слагаемые, соответствующие частной информации о событиях "система Y находится в состоянии y_j", содержащейся в информации о событиях, "система X находится в состоянии x_i":

$$I_{(X\sim 1)\to(Y\sim 1)} = -0{,}151;\ I_{(X\sim 1)\to(Y\sim 0)} = 0{,}36;$$

$$I_{(X\sim 0)\to(Y\sim 1)} = 0{,}341;\ \ I_{(X\sim 0)\to(Y\sim 0)} = -0{,}083.$$

Видно, что в данном случае значительно большую информацию мы

получаем от "обратных" сообщений: $(X \sim 0) \rightarrow (Y \sim 1)$ и $(X \sim 1) \rightarrow (Y \sim 0)$. Таким образом, при малых значениях Sup, Conf и Imp информативность правила может быть достаточно велика, что говорит о хороших показателях "обратных" ассоциаций, т.е. $\overline{X} \rightarrow Y$ и $X \rightarrow \overline{Y}$.

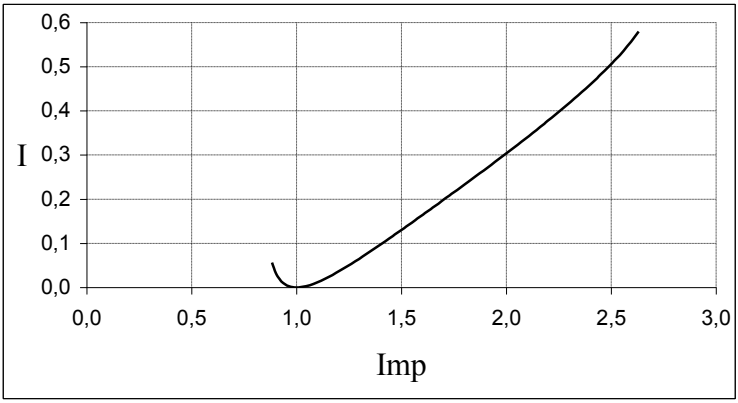

Рис. 3. Зависимость информативности правила от улучшения (при Sup = 0,3; Conf = 0,8)

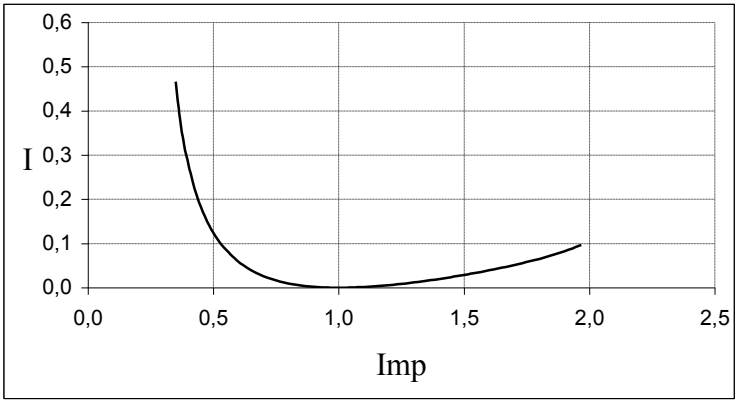

Рис. 4. Зависимость информативности правила от улучшения (при Sup = 0,1; Conf = 0,2)

Можно сделать вывод, что отбраковка ассоциативных правил с малыми значениями Sup и Conf, что происходит при работе распространенных алгоритмов Apriori и AprioriTid [1 – 3], может вести к потере некоторых достаточно интересных зависимостей. В данном случае предпочтительнее оказывается алгоритм, предлагаемый в [4], [7], который позволяет находить ассоциации с малой поддержкой и малой вероятностью.

"Интересность" ассоциативного правила. Следует отметить, что подход к оценке качества ассоциативных правил с точки зрения теории информации представляется весьма продуктивным. Он помогает преодолеть трудности, возникающие при оценке ассоциаций с помощью стандартных характеристик. Например, оценка "интересности" правила, данная в [3], основана, как уже было сказано, на превышении ожидаемых значений Sup или Conf в R раз. Возникает вопрос, как оценить "интересность" правила, Sup которого в R_1 раз больше ожидаемого, а Conf в R_2 раза меньше и наоборот. Расчет интегральной характеристики, включающей в себя все три показателя: Sup, Conf и Imp, позволяет это сделать.

С помощью формулы полной взаимной информации можно дать более широкое определение R-"интересности" ассоциативного правила. Ожидаемые значения уровней поддержки и доверия рассчитываются по формулам, приведенным в [3].

Пусть $X' = \{x'_1, \ldots, x'_j, x_{j+1}, \ldots, x_n\}$ является предком $X = \{x_1, \ldots, x_n\}$, $Y' = \{y'_1, \ldots, y'_j, y_{j+1}, \ldots, y_n\}$ является предком $Y = \{y_1, \ldots, y_n\}$ (с учетом таксономии признаков) и $Z = X \cup Y$.

Уровень поддержки $\mathrm{Sup}(Z) = \mathrm{Sup}(X \to Y)$.

$Z' = \{z'_1, \ldots, z'_j, z_{j+1}, \ldots, z_n\}$ – предок $Z = \{z_1, \ldots, z_n\}$, $1 \le j \le n$.

$\mathrm{Sup}_{Z'}(Z) = \frac{P(z_1)}{P(z'_1)} \times \ldots \frac{P(z_j)}{P(z'_j)} \times P(Z')$ – ожидаемый уровень поддержки Sup(Z), рассчитываемый на основе Z'.

$\mathrm{Conf}_{X' \to Y'}(X \to Y) = \frac{P(y_1)}{P(y'_1)} \times \ldots \frac{P(y_j)}{P(y'_j)} \times P(Y'/X')$ – ожидаемый уровень доверия правила $X \to Y$, рассчитываемый на основе правила $X' \to Y'$.

Подобным образом мы можем дать определение ожидаемого уровня улучшения $\mathrm{Imp}_{X' \to Y'}(X \to Y) = \frac{\mathrm{Conf}_{X' \to Y'}(X \to Y)}{P_{Y'}(Y)} = \frac{\mathrm{Conf}(X' \to Y')}{P(Y')}$, так как $P_{Y'}(Y) = \frac{P(y_1)}{P(y'_1)} \times \ldots \frac{P(y_j)}{P(y'_j)} \times P(Y')$.

На основе ожидаемых значений $\mathrm{Sup}_{X' \to Y'}(X \to Y)$, $\mathrm{Conf}_{X' \to Y'}(X \to Y)$ и $\mathrm{Imp}_{X' \to Y'}(X \to Y)$ можно рассчитать ожидаемое значение информативности правила по формуле (1). Сравнивая это значение с реальной информативностью правила $X \to Y$, мы получаем "интересность" правила. Таким образом, можно дать следующее определение "интересности" ассоциации: правило

$X \to Y$ является R-"интересным" относительно своего предка $X' \to Y'$, если его информативность в R раз выше ожидаемой. Если ассоциативное правило $X \to Y$ не имеет предков, то оно также считается "интересным".

Выводы. В статье дан анализ зависимости информативности ассоциативных правил от стандартных параметров – уровней поддержки, доверия и улучшения. Показывается, что эта зависимость не является прямо пропорциональной (чем больше характеристики, тем больше информативность). Правило с "плохими" значениями Sup, Conf и Imp также может представлять интерес с точки зрения сообщаемой информации.

Предлагается определение "интересности" ассоциативного правила с точки зрения теории информации.

ЛИТЕРАТУРА

1. *Agrawal R., Imielinski T., Swami A. Mining association rules between sets of items in large databases // Proc. of the ACM SIGMOD Conference Washington DC. – USA, Washington. – May 1993. – P. 207 – 216.*
2. *Agrawal R., Srikant R. Fast algorithms for mining association rules / Proc. of the 20th VLDB Conference Santiago, Chile, September 1994. – 112 p.*
3. *Srikant R., Agrawal R. Mining generalized association rules / Proc. of the 21th VLDB Conference Zurich, Swizerland, September 1995. – P. 407 – 419.*
4. *Amir A., Feldman R., Kashi R.. A new and versatile method for association generation // Information Systems. – 1997. – Vol. 22. – № 6/7. – P. 333 – 347.*
5. *Ситников Д.Э., Титова Е.В. Взаимная информация как обобщенный показатель качества ассоциативных зависимостей // Системи обробки інформації. – Х.: НАНУ, ПАНМ, ХВУ. – 2004. – Вип. 2. – С. 78 – 87.*
6. *Mei-Ling Shyu, Shu-ching Chen. Generalized affinity-bases association rule mining for multimedia database queries // Knowledge and Information system (KAIS): An International Journal. – 2001. – Vol. 3, № 3. – P. 319 – 337.*
7. *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.*

Поступила 5.04.2004

***СИТНИКОВ Дмитрий Эдуардович**, канд. техн. наук, доцент, зав. кафедрой информационно-документных систем ХГАК. В 1988 году окончил Харьковский институт радиоэлектроники. Область научных интересов – выделение знаний из данных, нечеткие множества.*

***ТИТОВА Елена Витольдиевна**, младший научный сотрудник научного центра при ХВУ, аспирантка кафедры информационно-документных систем ХГАК. В 1988 году окончила ХИРЭ. Область научных интересов – выделение знаний из данных.*