УДК 681.3.00:007

Д.Э. Ситников, С.В. Титов, Е.В. Титова

Харьковская государственная академия культуры, Харьков

СЕМАНТИЧЕСКИЕ СВОЙСТВА ИНФОРМАТИВНОСТИ АССОЦИАТИВНЫХ ЗАВИСИМОСТЕЙ МЕЖДУ ПРИЗНАКАМИ ОБЪЕКТОВ В БАЗАХ ДАННЫХ

Извлечение логических зависимостей (IF-THEN правил) из данных является одной из типичных задач интеллектуального анализа данных (Data Mining). Одной из проблем, возникающих при решении данной задачи, является оценка полученных правил с помощью ряда характеристик. Число предлагаемых характеристик достаточно велико, что обуславливает необходимость рассмотрения их семантических свойств для выделения наиболее понятных и информативных.

Ключевые слова: *логические правила, ассоциация, уровень поддержки, уровень доверия, информативность логической зависимости.*

Введение

Постановка проблемы. Под термином Data Mining понимают центральный этап процесса выявления знаний в данных, т.е. собственно обнаружение зависимостей, регулярностей и отношений, который опирается на определенную форму повторяемости информации в базах данных. Операции этого этапа включают поиск и конструирование зависимостей, их оценку и отсеивание.

Одной из типичных задач Data Mining является задача нахождения (генерации) правил ассоциации. Ассоциативное правило определяется как утверждение вида "Если A, то B", где A и B – некоторые множества признаков объектов (записей в базе данных), причем данное утверждение должно иметь меру определенности не ниже заданного уровня. Для оценки полученных зависимостей используется ряд характеристик, самые известные из которых это уровень поддержки Support и уровень доверия Confidence [1 – 4]. Эти две характеристики позволяют оценить качество полученного правила и существенно ограничить количество генерируемых зависимостей. Однако использование только двух характеристик для оценки ассоциативного правила явно недостаточно и в ряде случаев ведет к потере интересных зависимостей. На данный момент известно большое количество дополнительных характеристик ассоциации. Этот факт приводит к тому, что, во-первых, требуется рассмотрение существующих характеристик с точки зрения семантики, во-вторых, необходима выработка определенных рекомендаций в отношении попыток определения интегральных (синтетических) характеристик логических зависимостей.

Анализ последних исследований и публикаций. Наиболее известными алгоритмами для нахождения ассоциативных правил являются Apriori, AprioriTid, AprioriHybrid [1-3], предложенные в 1993-1994 г. американским ученым Агравалом. Для оценки получаемых ассоциативных правил используются две характеристики: уровень поддержки, равный Support=P_{AB} (отношение количества объектов, обладающих признаками A и B к общему количеству объектов в базе данных) и уровень доверия Confidence=$P_A(B)$ (отношение количества объектов, обладающих признаками A и B к количеству объектов, обладающих признаком A).

Однако, эти две характеристики, будучи очень важными, не позволяют в достаточной степени оценить качество ассоциативного правила, что привело к появлению большого количества дополнительных характеристик, таких как корреляция [5], обобщенная мера ассоциации [6], уровень улучшения и т.д. В [7, 8] было доказано, что для полной оценки ассоциативной зависимости необходимо и достаточно три характеристики: поддержка Support (Sup), доверие Confidence (Conf) и уровень улучшения Improvement (Imp), который определяется как отношение вероятности правила к вероятности результата:

$$\mathrm{Imp}(A \to B) = \frac{P_A(B)}{P(B)} = \frac{\mathrm{Conf}(A \to B)}{\mathrm{Sup}(B)}.$$

Там же [7,8] было предложено использовать интегральный показатель качества ассоциации – полную взаимную информацию.

Наиболее полное исследование характеристик (мер) ассоциативных правил было проведено в [9]. Авторами были рассмотрены более 60 параметров, предлагаемых для оценки ассоциативных зависимостей. Параметры оценивались по таким свойствам, как чувствительность к асимметричности правила, к независимости A и B, к размерам A и B, к ситуациям равновесия $P(AB) = P(\overline{AB})$ и импликации $P(AB) = P(A)$, к реакции на несовместимость $P(AB) = 0$, чувствительность к шуму (ошибочным данным), чувствительность к размеру данных.

Однако в данном исследовании не рассматриваются интегральные показатели, такие как информативность ассоциации.

Цель статьи. Оценить семантические свойства информативности ассоциации и выработать рекомендации по составу характеристик для оценки ассоциативного правила.

Информативность ассоциации и ее семантические свойства

Система двух величин (в данном случае двух признаков, двух частей ассоциативного правила) с точки зрения теории вероятностей характеризуется следующими значениями вероятностей: P_{AB}, $P_{\overline{A}B}$, $P_{A\overline{B}}$, $P_{\overline{A}\overline{B}}$. Как было показано в [7] использование для оценки ассоциативной зависимости A→B трех характеристик: Sup, Conf и Imp позволяет однозначно определить вероятности P_{AB}, $P_{\overline{A}B}$, $P_{A\overline{B}}$, $P_{\overline{A}\overline{B}}$. Так же был предложен интегральный показатель качества ассоциации – полная взаимная информация, который рассчитывается при любых показателях системы двух величин и позволяет количественно сравнивать полученные правила:

$$I_{A\leftrightarrow B} = \sum_{i=1}^{n}\sum_{j=1}^{m} P_{ij} \log \frac{P_{ij}}{P_i P_j}.$$

Для информативности ассоциации была получена формула, позволяющая рассчитать ее значение на основе уровня доверия, поддержки и уровня улучшения [7]:

$$\begin{aligned} I_{A\leftrightarrow B} = {} & \mathrm{Sup}(A\to B)\cdot\log(\mathrm{Imp}(A\to B))+ \\ & +\mathrm{Sup}(A\to \overline{B})\cdot\log(\mathrm{Imp}(A\to \overline{B}))+ \\ & +\mathrm{Sup}(\overline{A}\to B)\cdot\log(\mathrm{Imp}(\overline{A}\to B))+ \\ & +\mathrm{Sup}(\overline{A}\to \overline{B})\cdot\log(\mathrm{Imp}(\overline{A}\to \overline{B})). \end{aligned}$$

Использование такой характеристики расширяет и дополняет систему показателей ассоциации. Кроме этого, информативность правила (полная взаимная информация) может задаваться в качестве ограничения при генерации ассоциативных зависимостей во время выделения знаний из данных.

Оценим свойства информативности ассоциации согласно перечню, предложенному в [9].

1. **Асимметричность правила**. Предполагается, что характеристика ассоциативной зависимости обладает свойством асимметричности, если ее значения не равны для правил A→B и B→A. Полная взаимная информация по своему определению является характеристикой симметричной.

2. **Чувствительность к размеру левой части правила**. Характеристика ассоциативной зависимости A→B обладает чувствительностью к B, если ее значение уменьшается при уменьшении B (или P_B). При уменьшении P_B вероятность $P_{\overline{A}B}$ также уменьшается, а вероятность $P_{\overline{A}\overline{B}}$ возрастает (предполагается, что P_{AB} не меняется). Таким образом, значение информативности ассоциации будет меняться, но прямой зависимости между размером левой части правила и значением информативности не существует.

3. **Чувствительность к размеру правой части правила**. Прямой зависимости между размером правой части правила и значением информативности также не существует.

4. **Чувствительность к независимости правой и левой части ассоциативного правила.** Как было показано в [7, 8] при независимости A и B информативность ассоциации обращается в 0. Таким образом, данная характеристика чувствительна к ситуации независимости правой и левой частей правила.

5. **Чувствительность к ситуации импликации**. Данное свойство предполагает, что характеристика ассоциативного правила принимает значение const, если имеет место $P(AB) = P(A)$. Информативность ассоциации этим свойством не обладает, т.к. ее значение зависит от таких вероятностей как $P_{\overline{A}B}$, $P_{A\overline{B}}$, $P_{\overline{A}\overline{B}}$, которые могут разные значения при $P(AB) = P(A)$.

6. **Чувствительность к ситуации равновесия**. Данное свойство предполагает, что характеристика ассоциативного правила принимает значение const, если имеет место $P(AB) = P(\overline{AB})$. Информативность ассоциации этим свойством не обладает, т.к. ее значение зависит от таких вероятностей как $P_{\overline{A}B}$, $P_{A\overline{B}}$, которые могут разные значения при $P(AB) = P(\overline{AB})$.

7. **Реакция на несовместимость.** Данное свойство предполагает, что характеристика ассоциативного правила принимает значение const, если имеет место $P(AB) = 0$. Информативность ассоциации этим свойством не обладает, т.к. ее значение зависит от таких вероятностей как $P_{\overline{A}B}$, $P_{A\overline{B}}$, $P_{\overline{A}\overline{B}}$, которые могут разные значения при $P(AB) = 0$.

8. **Чувствительность к размеру данных**. Данное свойство предполагает увеличение значения характеристики при увеличении размера анализируемых данных. Информативность ассоциации этим свойством не обладает.

9. **Чувствительность к шуму (ошибочным данным)**. Данное свойство предполагает нечувствительность к появлению обратных $A\to\overline{B}$ примеров. Характеристика ассоциативного правила считается выпуклой, если она уменьшается незначительно с появлением малого количества обратных примеров и уменьшается резко с возрастанием числа обратных примеров. Полная взаимная информация по своей природе не является характеристикой ассоциативного правила A→B, а характеризует систему двух величин A и B. Увеличение количества обратных примеров $A\to\overline{B}$ означает улучшение показателей

"обратного" ассоциативного правила. Такие обратные ассоциации могут быть достаточно интересны в некоторых случаях. Например, для торговой компании, занимающейся изучением покупательского спроса путем анализа БД супермаркета, правила вида: "если покупатель приобрел товар А, то с большой вероятностью он не приобретет товар В" вряд ли будут представлять интерес. Однако при выделении ассоциативных правил из медицинских БД, правила вида "если у пациента наблюдается симптом А, то с большой вероятностью можно утверждать, что он не болен В" достаточно ценны.

Таким образом, можно утверждать, что информативность ассоциации не обладает свойством выпуклости, однако анализ отдельных ее слагаемых может привести к обнаружению новых интересных зависимостей.

Отметим также, что такое слагаемое как $I_{A\to B}$ (информация о событии В, получаемую в результате сообщения о событии А), характеризующее информативность не ассоциации, а именно ассоциативного правила, обладает таким свойством как выпуклость. Известные характеристики ассоциативного правила: уровень поддержки Support, доверие Confidence и уровень улучшения Improvement не обладают свойством выпуклости (уменьшаются линейно с увеличением ошибочных данных).

Выводы

В статье рассмотрены семантические свойства информативности ассоциации. Полученные результаты позволяют сделать следующие выводы:

1. Как любая обобщенная характеристика информативность обладает меньшим количеством свойств (если рассматривать первые 8 свойств, то информативность обладает одним – чувствительностью к независимости А и В), чем более простые характеристики, такие как уровень поддержки Support (1 из 8), доверие Confidence (5 из 8) и уровень улучшения Improvement (4 из 8).

2. Рассмотрение информативности ассоциации в качестве дополнительной характеристики к системе Support-Confidence-Improvement позволяет получать новые интересные зависимости, касающиеся "обратных" ассоциативных правил.

3. Анализ такого слагаемого как информативности ассоциативного правила $I_{A\to B}$ позволяет установить наличие ошибочных данных (шума) при оценке ассоциативного правила $A\to B$.

Список литературы

1. Agrawal R. Mining association rules between sets of items in large databases / R. Agrawal, T. Imielinski, A. Swami // Proc. of the ACM SIGMOD Conference Washington DC, USA, May 1993. – P. 207 – 216.

2. Agrawal R. R.Srikant. Fast algorithms for mining association rules / R. Agrawal, R. Srikant // Proc. of the 20th VLDB Conference Santiago, Chile, September 1994.

3. Srikant R. Mining generalized association rules / R. Srikant, R .Agrawal // Proc. of the 21th VLDB Conference Zurich, Swizerland, September 1995. – P. 407 – 419.

4. Amir A., R.Feldman, R.Kashi. A new and versatile method for association generation / A. Amir, R. Feldman, R. Kashi. // Information Systems. – 1997. – Vol. 22, № 6/7. – P. 333 – -347.

5. Балабанов А.С. Выделение знаний из баз данных – передовые компьютерные технологии интеллектуального анализа данных / А.С.Балабанов // Математичні машини і системи. – 2001. – № 1,2. – С. 40 – 54.

6. Edwards A.W.F. The measure of association in 2×2 table / A.W.F. Edwards // Journal of the Royal Statistical Society. ser. A29. – P. 109 – 114.

7. Ситников Д.Э. Взаимная информация как обобщенный показатель качества ассоциативных зависимостей / Д.Э. Ситников, Е.В. Титова // Системи обробки інформації. – Х.: НАНУ, ПАНМ, ХВУ, 2004. – Вип. 4. – С. 20 – 28.

8. Ситников Д. Э., Титова Е. В. Влияние стандартных параметров ассоциативного правила на его информативность // Системи обробки інформації – Х.: НАНУ, ПАНМ, ХВУ, 2004. – № 4. – С. 182 – 189.

9. Maddouri M. On Semantic Properties of Interestingness Measures for Extracting Rules from Data / M. Maddouri, J. Gammoudi. // ICANNGA '07 Proc. of the 8th international conference on Adaptive and Natural Computing Algorithms, Part I. – Berlin, Heidelberg, 2007. – P. 148 – 158.

Поступила в редколлегию 1.11.2011

Рецензент: д-р техн. наук, проф. И.В. Гребенник, Харьковский национальный университет радиоэлектроники, Харьков.

СЕМАНТИЧНІ ВЛАСТИВОСТІ ІНФОРМАТИВНОСТІ АСОЦІАТИВНИХ ЗАЛЕЖНОСТЕЙ МІЖ ОЗНАКАМИ ОБ'ЄКТІВ В БАЗАХ ДАНИХ

Д.Е. Ситніков, С.В. Тітов, О.В. Тітова

Проведено аналіз семантичних властивостей такої характеристики асоціативних правил як повна взаємна інформація.
Ключові слова: *логічні правила, асоціація, рівень підтримки, рівень довіри, інформативність логічної залежності.*

INFORMATION SEMANTIC PROPERTIES OF ASSOCIATION DEPENDENCE BETWEEN OBJECTS ATTRIBUTES IN DATABASE

D.E. Sitnikov, S.V. Titov, E.V. Titova

We analyze some semantic properties such characteristic for association rules as full mutual information.
Keywords: *logical rules, association, level of support, level of trust, informing of logical dependence.*