Available online at www.sciencedirect.com
ScienceDirect
Knowledge-Based Systems 19 (2006) 739–743
Knowledge-Based SYSTEMS
www.elsevier.com/locate/knosys

Novel measurement for mining effective association rules

Jin-Mao Wei *, Wei-Guo Yi, Ming-Yang Wang

Institute of Computational Intelligence, Northeast Normal University, Changchun, Jilin 130024, China

Received 25 November 2005; accepted 4 May 2006
Available online 18 July 2006

Abstract

Mining association rules are widely studied in data mining society. In this paper, we analyze the measure method of *support–confidence* framework for mining association rules, from which we find it tends to mine many redundant or unrelated rules besides the interesting ones. In order to ameliorate the criterion, we propose a new method of *match* as the substitution of *confidence*. We analyze in detail the property of the proposed measurement. Experimental results show that the generated rules by the improved method reveal high correlation between the antecedent and the consequent when the rules were compared with that produced by the *support–confidence* framework. Furthermore, the improved method decreases the generation of redundant rules.

Keywords: Data mining; Association rules; Correlation; Match

1. Introduction

It is an important issue to mine association rules from transaction databases in data mining. Mining association rules aims at finding the correlation between the different items in a database. It can be used to find the purchase patterns of customers such as how the transaction of buying some goods will impact on the transaction of buying others. The rules can be utilized to design the merchandise shelves, to manage the stock and to classify the customers according to the purchase patterns.

Assume $I = \{i_1, i_2, \ldots, i_m\}$ is a binary set, in which parameters $i_1, i_2, \ldots, i_m$ are called items. Define transaction T as the item set with the restriction of $T \subseteq I$. Define D as the transaction set. Supposing that X is the set containing some items in I, transaction T includes X if $X \subseteq T$. The length of the item set is defined as the number of the items in it. Association rules are those implications that can be depicted as $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \emptyset$. The *support* of the rule $X \Rightarrow Y$ in the transaction database D is the ratio of the number of transactions containing X and Y in the transaction sets to the number of all transactions. It is written as $support(X \Rightarrow Y)$, that is to say

$$support(X \Rightarrow Y) = |\{T : X \cup Y \subseteq T,\ T \in D\}| / |D|.$$

The *confidence* of the rule $X \Rightarrow Y$ in the transaction sets is the ratio of the number of transactions including X and Y to the number of those including X. It is written as $confidence(X \Rightarrow Y)$, that is to say

$$confidence(X \Rightarrow Y) = |\{T : X \cup Y \subseteq T,\ T \in D\}| / |\{T : X \subseteq T,\ T \in D\}|.$$

For a certain transaction set D, mining association rules means to figure out the association rules whose *support* and *confidence* are higher than the minimum *support*(*minsup*) and *confidence*(*mincon*), respectively. Consequently, mining association rules is divided into two subissues as follows:

(1) Find out all the item sets, the *supports* and *confidences* of which are larger than or equal to the *minsup* required by the customers. The item sets with *minsup* are called frequent sets.
(2) Form association rules from frequent sets. First, we find out all M's non-empty subsets ms to each frequent set M. Association rule $m \Rightarrow (M - m)$ is

* Corresponding author. Tel.: +86 431 5099665; fax: +86 431 5099789.
E-mail address: weijm374@nenu.edu.cn (J.-M. Wei).

0950-7051/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2006.05.011

generated when $support(M)/support(m) \geqslant mincon$. $support(M)/support(m)$ is defined as the *confidence* of the rule $m \Rightarrow (M - m)$ where m is the antecedent of the rule and $M - m$ is the consequent of the rule.

In the process of mining association rules, the efficiency of the algorithm is very significant because it needs to scan the datasets many times to create the frequent sets. Therefore, most current research works focus on the improvement of the efficiency of the algorithms for producing frequent sets. Agrawal et al. proposed the apriori algorithm [1]; Han et al., developed DBMiner for mining knowledge from large databases [2]; Park and his colleagues proposed PHD algorithm, etc. There are a lot of other popular research issues, such as the research on the apriori method; the incremental renovation of the association rules; mining effective association rules; mining association rules based on Neural Networks and so on [3–8]. Our work is focused on the improvement on the measurement of mining effective association rules.

2. Measure standards of the association rules

Researchers have been applying the framework of *support–confidence* to set up association rules in the process of producing association rules. However, a lot of redundant and unrelated rules are also generated when the framework of *support–confidence* is applied to find rules. We use an example to show the shortcomings of the framework. A set of transaction data is shown in Table 1.

We only discuss the item sets with length of 2, and assume the *minsup* and *mincon* are 0.2 and 0.5, respectively. From the table we learn that C and K are always present or absent at the same time, which can form an effective association rule. The *support* and *confidence* of $C \Rightarrow K$ are 0.3 and 1 individually by calculation. Meanwhile, the *support* and *confidence* of $C \Rightarrow R$ are also 0.3 and 1 individually. Hence, the rules $C \Rightarrow K$ and $C \Rightarrow R$ have the same *support* and *confidence*, and we can draw the conclusion that both of them are effective association rules. However, by observation we notice that R will always present whatever C is present or not. Therefore, we would draw the different conclusion that $C \Rightarrow R$ is not an effective association rule.

Table 1
A set of transaction data

TID	Items
01	*R*, *I*, *J*, *C*, *K*, *H*, *M*, *N*
02	*R*, *I*, *C*, *K*, *H*, *M*, *N*
03	*R*, *I*, *J*, *C*, *E*, *K*
04	*R*, *I*, *J*, *E*, *F*, *H*, *N*
05	*R*, *I*, *J*, *E*, *F*, *H*
06	*R*, *I*, *J*, *E*, *F*
07	*R*, *I*, *J*, *E*
08	*R*, *I*, *J*, *F*
09	*R*, *J*, *E*
10	*R*, *J*, *F*

Sequentially, we analyze the rule $F \Rightarrow E$. The *support* and *confidence* of it are 0.3 and 0.6, respectively, which are larger than the *minsup* and the *mincon*. From this we can also draw the conclusion that the rule is an effective association rule. With further calculation we find that $P(EF)=P(E) \times P(F)$. E and F are unrelated to each other from the point of mathematics. Mathematically, E and F are positive correlated to each other if and only if $P(EF) > P(E) \times P(F)$. Otherwise, they are negative correlated. In this paper, we are only interested in mining the rules from the positive correlated item sets. Many scholars and experts have investigated the correlations among item sets, and defined the correlation threshold in order to reduce the emergence of irrelated rules [9,10]. From the above discussion, we can learn that not all of the rules conforming to the *minsup* and *mincon* are all effective rules.

Now we further analyze the framework of *support–confidence*. From the definitions, *support* denotes the frequency of the occurrence of item sets. The regularities exist only when item sets occur frequently. Otherwise, it is hard to find out the regularities included in the item sets. Meanwhile, *confidence* denotes the probability that the emergence of some item sets will lead to the occurrence of the others. However, we notice that the *confidence* of association rule $F \Rightarrow E$ only takes into consideration the possibility of the case when E and F occur simultaneously, and fails to take into account the possibility of the case when only E occur and the case whether E and F are correlated. Consequently, many association rules obtained in accordance with the framework of *support–confidence* tend to be ineffective.

3. Improvement of measure standards

According to those problems discussed above, we suggest that the depiction of *confidence* is not consummate. It is inadequate to describe the correlation among item sets. In [8], the authors applied *validity* to substitute *confidence* to generate association rules. *Validity* is defined as follows:

validity = (probability that X and Y occur simultaneously in database D) − (probability that $\overline{X}$ and Y occur simultaneously in database D), that is,

$$validity = P(XY) - P(\overline{X}Y).$$

The introduction of *validity* will reduce the occurrence of some redundant rules but it does not work on eliminating irrelated rules. Take the rule $E \Rightarrow F$ in Table 1 for example. The *support* and *validity* are $support = 0.3$ and $validity = 0.3 - 0.2 = 0.1$. The rule turns to be an effective association rule according to the method reported in [7]. But, from the above analysis we have $P(EF)=P(E) \times P(F)$, which shows that E and F are irrelated to each other. Analysis of the rule $I \Rightarrow J$ with the $support = 0.7$ and the $validity = 0.5$ shows that $I \Rightarrow J$ is also an effective association rule according to literature [8]. But, the calculation of $P(IJ) - P(I) \times P(J) = 0.7 - 0.8 \times 0.9 = -0.02$ indicates that there is a negative correlation between I and J. So there are also some drawbacks if *validity* is used to substitute *confidence*.