Computer and Information Science
www.ccsenet.org/journal.html

Vol. 2, No. 2
May 2009

Study on the Data Mining Algorithm Based on Positive and Negative Association Rules

Jingrong Yang & Chunyu Zhao

The Engineering & Technical College, Chengdu University of Technology, Chengdu 614007, China

Tel: 86-833-383-2599 E-mail: yyoko@126.com

Abstract

In this article, we systematically, deeply and comprehensively analyzed and studied the association rule data mining technology, and induced, analyzed and researched the typical mining algorithms of association rule and their basic principles, and objectively compare the differences among various algorithms. We used to correlation to measure the relations among item sets, and gave the computations of support level and confidence level of negative association rule based on traditional association rules, and analyzed and researched the operation principle and implementation approaches of this algorithm. Through the demonstration test of the algorithm, the results indicated that the algorithm was effective.

Keywords: Data mining, Association rule, Correlation, Positive and negative association rules

1. Introduction

Traditional database application emphasizes the online transaction processing (OLTP), and its main task is to exert the online transactions and inquiry processing. The quick increasing speed of the commercial database which intention is OLTP requires that the data mining technology which could offer the support information for the decision-making develops quickly, i.e. the online analysis processing (OLAP) could acquire and utilize information from the database.

At present, the domestic researches about data mining mainly concentrate in the optimization and improvement of the algorithm. Based on former research results, we study the association rule, i.e. the negative association rule, from another view, and make it combine with traditional association rule to form the positive and negative association rule, which will make the theory of association rule more perfect.

2. Basic concept of association rule

The association rule is one of very important rules which the data mining technology can apply, and it is used in the interest association among item sets of large amount of data, for example, the association rule mining can be used to find the association among different commodities (items) in the transaction database.

Suppose I={i1 i2,…, im} is the set of item, and suppose the data D relative to the task is the set of the transaction in the database, and each transaction T is the set of item, and $T\subseteq I$. Each transaction has one identifier which is called as *TID*. Suppose *X* is the set of item in *I*, and it is called as the item set, and the transaction *T* contains *X* when and only when $\mathrm{X}\subseteq\mathrm{T}$.

The association rule is the formula like $X\Rightarrow Y$, and $X\subseteq I$, $Y\subseteq I$ and $X\bigcap Y=\Phi$.

The rule $X\Rightarrow Y$ comes into existence in the transaction D, and it has the support level, s (support), and when and only when the proportion that D contains the transactions of XUY is s, i.e.

$$\mathrm{s=support(X\Rightarrow Y) = P(XUY) = |\{T|X \bigcup Y \subseteq \in T \wedge T \in D \}|/|D|}$$

The rule $X\Rightarrow Y$ comes into existence in the transaction D, and it has the confidence level, c (confidence), and when and only when the proportion that D contains the transactions of X is c, i.e.

$$\mathrm{C=confidence(X\Rightarrow Y) = p(Y/X) = \{T|\ X \bigcup Y \subseteq T \wedge T \in D \}|/|\{T|X \subseteq T \wedge T \in D\}|}$$

The set of item is called as the item set. The item set contains k items is called as k-item set, for example, {printer, computer} is a 2-item set. The occurrence frequency of the item set is the amount of transaction containing the item set, which is called as the frequency of the item set, the support count or the count. The item set fulfills the minimum support, *min_sup*, when and only when the occurrence frequency of the item set exceeds or equals the product of *min_sup* and the amount of transactions in D. the item set fulfilling *min_sup* is called as the frequent item set. The frequent item set containing k items is called as the frequent k-item set which is generally noted as L_k. The association rules which fulfill the *min_sup* and *min_conf* (minimum confidence level) synchronously are called as the strong rules.

3. Mining algorithm of positive and negative association rules

Traditional association rule (AR) has the form of $A \Rightarrow B$, and it is used to mine the association relation among the item sets in the database of consumer transaction, and it was first proposed by R. Agrawal et al in 1993, and he proposed a sort of quick algorithm in 1994. As one important complement of the association rule of $A \Rightarrow B$, the association rules with three forms such as $A \Rightarrow \neg B$, $\neg A \Rightarrow B$, $\neg A \Rightarrow \neg B$ are studied in this article, and they are called as the negative AR (NAR). We will give a sort of simple and effective method which was used to compute the support level and the confidence level of NAR only by relative information of positive association rule, and an algorithm which could synchronously mine positive and negative association rules. The difference with existing algorithms is that the algorithm in this article can not only synchronously mine the positive and negative association rules in the frequent item set, but test and delete the inconsistent rules.

3.1 Computations of support level and confidence level in negative association rule

The confidence level (c) of the rule $A \Rightarrow B$ in the transaction database D means the ratio of the amount of transaction containing A and B and the amount of transaction containing A, i.e. $c(A \Rightarrow B)$. The negative association rule contains non-existing-items such as $\neg A$ and $\neg B$, and because it is difficult to directly compute their support level and confidence level, so we give following theorems and computation methods.

Theorem 1 Suppose $A, B \subset I$, $A \cap B = \Phi$, so we have

a. $s(A)=1-s(\neg A)$;

b. $s(A \cup \neg B)=s(A)-s(A \cup B)$;

c. $s(\neg A \cup B)=s(B)-s(A \cup B)$;

d. $(\neg A \cup \neg B)=1-s(A)-s(B)+s(A \cup B)$.

To prove this theorem, we need to re-denote the support level and the confidence level from the view of set, i.e. changing the set operation of item set into the set operation of transaction set, which can better apply some theorems and characters and be easy to be understood.

Suppose *As* denotes the set of transactions containing the item set A, and its base |*As*| is the amount of transaction in *As*. In the same way, suppose *Bs* denotes the set of transactions containing the item set B, and its base |*Bs*| is the amount of transaction in *Bs*. The database D is the set of all transactions in the database, i.e. the total set which is denoted by D, its base |*D*| is the amount of all transactions, so the corresponding conversions are

a. $s.count(A \cup B)=|As \cap Bs|$;

b. $s(A)=s.count(A)/|D|=|As|/|D|$;

c. $s(A \cup B)=s.count(A \cup B)/|D|=|As \cap Bs|/|D|$;

d. $c(A \Rightarrow B)=s(A \cup B)/S(A)=|As \cap Bs|/|As|$.

Deduction 1 Suppose $A,B,I,A \cap B=\Phi$, so we have

a. $c(A \Rightarrow \neg B)=(s(A)s(A \cup B))/s(A)=1-c(A \Rightarrow B)$;

b. $c(\neg A \Rightarrow B)=(s(B)-s(A \cup B))/(1-s(A))$;

c. $c(\neg A \Rightarrow \neg B)=(1-s(A)-s(B)+s(A \cup B))/(1-s(A))=1-c(\neg A \Rightarrow B)$.

According to Theorem 1 and the definition of the confidence level, we can easily prove Deduction 1 which can be used to compute the confidence level of the negative association rule.

3.2 Algorithm of mining positive and negative association rules

In the algorithm, suppose that the frequent item set has been solved and stored in the set L.

Algorithm 1 Mining_PAR&NAR

Input L: frequent item set; *min_conf*: minimum support level;

Output: The set of positive and negative association rules A_R;

```
a. $A_R=\Phi$ ;
b. // generate the positive and negative association rules in L.
For any itemset $X$ in $L$ do{
    for any itemset $A \cup B=X$ and $A \cap B= \Phi$ do
    {
        corr=$s(A \cup B)/(s(A)s(B))$
        if corr>1then{
        // generate the rule of $A \Rightarrow B$ and the rule of $\neg A \Rightarrow \neg B$
      if $c(A \Rightarrow B) \geq$ min_conf then
      $AR=AR \cup \{A \Rightarrow B\}$;
      if $c(\neg A \Rightarrow \neg B) \geq$ min_conf then
      $AR=AR \cup \{\neg A \Rightarrow \neg B\}$;
         }
    if corr<1 then{
   // generate the rule of $A\neg \Rightarrow B$ and the rule of $\neg A \Rightarrow B$
   if $c(A \Rightarrow \neg B) \geq$ min_conf then
   $AR=AR \cup \{A \Rightarrow \neg B\}$;
   if $c(\neg A \Rightarrow B) \geq$ min_conf then
   $AR=AR \cup \{\neg A \Rightarrow B\}$;
   }
   }
   }
```

c. Return AR.

To validate the validity of Algorithm 1, we do a test in the synthetic data, and the test is performed under the environment of Celeron 2.5, 256RAM, WIN2000, VC++. There are experimental data containing 200 transactions, and the maximum item set number is 5. Suppose that *min_supp* is 0.20, *min_conf* is 0.40, and Table 1 lists the comparison of experiment results of two algorithms.

From Table 1, the positive association rule number obtained by Algorithm 1 is significantly less than the positive association rule number obtained by traditional Apriori algorithm, which indicates some inconsistent rules have been deleted, and many negative association rules have been mined, which indicates Algorithm 1 is effective.

4. Research of P-S interest in positive and negative association rules

Only the rule of $A \Rightarrow B$ accords with the condition of $supp(AUB)-supp(A)supp(B) \geq$ miniuterest>0, it is interesting, but for negative association rule, supp(AUB)-supp(A)supp(B) may be less than 0, so we use its absolute value as the condition, i.e. only if the rule of $A \Rightarrow B$ accords with the condition of $|supp(AUB)-supp(A)supp(B)| \geq$ mininterest, it is interesting. What relations exist in the minimum interests of four sorts of association rule?

Theorem 2 If $|supp(AUB)-supp(A)supp(B)| \geq$ mininterest, so

(1) $|supp(AU\neg B)-supp(A)supp(\neg B)| \geq$ mininterest;

(2) $|supp(\neg AUB)-supp(\neg A)supp(B)| \geq$ mininterest;

(3) $|supp(\neg AU\neg B)-supp(\neg A)supp(\neg B)| \geq$ mininterest.

Theorem 2 indicates that only if the mini-interest is reasonably set up, some rules without interest can be avoided effectively, and four sorts of association rule can be restricted by one minimum interest P.

When studying the positive and negative association rules at the same time, the problem of conf($\neg$A =>B)>conf(A=>B)>min_conf may occur, so the application of correlation is the effective method to solve this problem. The correlation of association rule can be measured by supp(AUB)/(supp(A)supp(B), where $s(A) \neq O, s(B) \neq 0$. In fact, if we improve the P-S interest little, it can be used in the correlation judgment of association rules, i.e. using corr(A,B)=supp(AUB)-supp(A)supp(B) to measure the correlation.

There are three possible instances for corr(A,B).

(1) If corr(A,B)>0, so A and B are positively correlated, i.e. transaction A occurs more, transaction B occurs more too.

(2) If corr(A,B)=0, so A and B are independent each other, the occurrence of transaction B is independent of transaction A.

(3) If corr(A,B)=0, so A and B are negatively correlated, i.e. transaction A occurs more, transaction B occurs less.

Theorem 3 If corr(A,B)>0, so

(1) corr($\neg$ A,B)<0;

(2) corr(A, ¬ B)<0;

(3) corr(¬ A, ¬ B)>0;

the results will be contrary.

Theorem 3 indicates that the rule of A =>B (or ¬ A=>¬ B) and A=>¬ B(or ¬ A=>B) can not be the effective rules simultaneously, so the dissociable rules will be effectively prevented.

References

Dong, Xiangjun, Song, Hantao & Jiang, He et al. (2004). Minimum Interestingness Based on Method for Discovering Positive and Negative Association Rules. *Computer Engineering and Applications.* No.27. P.24-25, 31.

Dong, Xiangjun, Wang, Shujing & Song, Hantao et al. (2004). Study on Negative Association Rules. *Transactions of Beijing Institute of Technology.* No.11(24). P.978-981.

Huang, Jin & Yin, Zhiben. (2003). Improvement of Apriori Algorithm for Mining Association Rules. *Journal of University of Electronic Science and Technology of China.* No.32(1). P.76-79.

Lu, Jingli, Xu, Zhangyan & Liu, Meiling et al. (2004). An Improved Algorithm for Identifying Negative Association Rules. *Journal of Guangxi Normal University (Natural Science Edition).* No.22 (2). P.41-46.

Peng, Mugen. (2002). The Database Technology and Its Implementation. Beijing: Electronic Industry Press.

Table 1. Comparison of association rule number

Algorithm	Association rule number
Apriori	+365
Algorithm 1	+282
	-603