2010 Fifth International Conference on Internet Computing for Science and Engineering

Research on Mining Positive and Negative Association Rules Based on Dual Confidence

Xiufeng Piao
College of Computer Science and Technology
Harbin Engineering University
Harbin , China
xfpiao@126.com

Zhanlong Wang
College of Computer Science and Technology
Harbin Engineering University
Harbin , China
wangzhanlong@hrbeu.edu.cn

Gang Liu
College of Computer Science and Technology
Harbin Engineering University
Harbin , China
liugang@hrbeu.edu.cn

***Abstract*—Mining of association rules has become an important area in the research on data mining. However the traditional approaches based on support-confidence framework maybe generate a great number of redundant and wrong association rules. In order to solve the problems , a correlation measure is defined and added to the mining algorithm for association rules. According to the value of correlation measure, association rules are classified into positive and negative association rules. Therefore,the new algorithm can mine the negative-item-contained rules. In the paper, the algorithm which based on the correlation and dual confidence, can mine the positive and negative association rules.The experimental result shows that positive and negative association rules mining algorithm can reduce the scale of meaningless association rules, and mine a lot of interesting negative association rules.**

Keywords- data mining; positive and negative association rules; dual confidence; minimum correlation

I. INTRODUCTION

Association rules was proposed by Rakesh Agrawal et al [1] in 1993 for initially, which is an important technique in data mining. Traditional association rules, mainly for mining customer transaction database relationship between sets of items, that is of the form $A \Rightarrow B$, high frequency, high correlation rules, which we call positive association rules, is a strong correlation of the dominant model, and there are many mining algorithms[2-7]. In fact, many databases which use these mining technology can not find the hidden patterns, one of these important hidden pattern is the negative association rules, which has a low frequency and strong correlations, showing the property of the strong correlations in the data itemsets which is hard to find. The rules tell us that which data items less to occur together, and they share a very strong correlation, and contain very valuable information, such as rules $A \Rightarrow \neg B$ 、 $\neg A \Rightarrow B$ 、 $\neg A \Rightarrow \neg B$,so mining negative association rules is very important.

Currently used for the negative association rules mining algorithm is not too much, such as [3] proposed a positive and negative association rules mining algorithm based on interestingness,a negative association rules mining algorithm based on support, confidence, correlation coefficient in [4], and a negative association rules mining algorithm based on matrix in [5].

Negative association rules of the search space for all non-frequent itemsets, and the non-frequent itemsets in the database D is exponential, which is also the main reason of study negative association rules more difficult than positive association rules.While obtained the support of all non-frequent itemsets is unrealistic, but also meaningless. For example, for the negative association rules $A \Rightarrow \neg B$, $A \cup B$ is the non-frequent itemsets, but both A and B are frequent itemsets, Therefore, a negative association rules mining can be divided into two sub-problems: ① the database D, find all the frequent itemsets L; ② Determine negative association rules based on frequent itemsets L. There are many algorithms can be used directly to solve the first sub-problems, such as the Apriori algorithm, FP-growth algorithm. This paper gives a negative association rules mining algorithm to find hidden rules in massive data. Negative association rules are used to discover the meaningful association beteween the itemsets and others in the large data, giving the interesting relationship between these items.

II. PROBLEM FORMULATION

Set $I=\{i_1,i_2,\ldots,i_m\}$that contains a collection of m different items. Given a transaction database D, where each transaction T is a group itemsets of I. If A is a subset of I,with $A \subseteq T$, we can say that a transaction T contains A. A negative association rules are an implication of the form $\neg A \Rightarrow B$ (or $A \Rightarrow \neg B$,or $\neg A \Rightarrow \neg B$) ,Where $A, B \subseteq T$, and $A \cap B = \varnothing$.

Given support s and confidence c. If D has $(100 \times s)\%$ of the transaction contains B but does not contain A, the support of negative association rules $\neg A \Rightarrow B$ is s,denoted as $sup(\neg A \Rightarrow B)=s$. If the transaction does not contain A, there are $(100 \times c)\%$ of the transaction contains B, the confidence of negative association rules $\neg A \Rightarrow B$ is c, denoted as $conf(\neg A \Rightarrow B)=c$. So the support and confidence of negative association rules $A \Rightarrow \neg B$ 、 $\neg A \Rightarrow \neg B$ can also be defined.

Negative association rule discovery seeks rules of the form $\neg A \Rightarrow B$ (or $A \Rightarrow \neg B$,or $\neg A \Rightarrow \neg B$) with support and

978-0-7695-4339-0/10 $26.00 © 2010 IEEE
DOI 10.1109/ICICSE.2010.28

confidence greater than, or equal to, user-specified minimum support (*minsup*) and minimum confidence (*minconf*) thresholds respectively, where A and B are frequent itemsets. In this paper, introducting the minimum correlation theory, so it can reduce the production of the meaningless association rules.

III. ALGORITHM DESIGN

A. Confidence

With goods A、B, by *sup(A)* and *sup(B)* small (e.g. less than 5%), then sup($A \cup B$) is smaller , while the confidence $conf(A \Rightarrow B)$ may be large, may also be small, but $conf(\neg A \Rightarrow \neg B)$ positive large. If the uniform rules for all the confidence bound, there will be such an embarrassing situation: If the less confidence, would be a lot of rules, the customer can not choose the genuine needs of the rules, even presence of a large $\neg A \Rightarrow \neg B$ type rules are not necessarily meaningful; if confidence is greater, may miss many valuable positive association rules. Therefore, the best solution is to use different confidence thresholds, and we set two confidence thresholds, *P_minconf* and *N_minconf*. *P_minconf* show that the minimum confidence threshold of the rules $A \Rightarrow B$ and $\neg A \Rightarrow \neg B$; *N_minconf* show that the minimum confidence threshold of the rules $A \Rightarrow \neg B$ and $\neg A \Rightarrow B$.

Property 1

Let $A \subset I$, $B \subset I$, $A \cap B = \varnothing$ then

①$sup(\neg A) = 1 - sup(A)$

②$sup(\neg A \cup B) = sup(B) - sup(A \cup B)$

③$sup(A \cup \neg B) = sup(A) - sup(A \cup B)$

④$sup(\neg A \cup \neg B) = 1 - sup(A) - sup(B) + sup(A \cup B)$

Property 2 Let $A \subset I$, $B \subset I$, $A \cap B = \varnothing$ then

$$conf(A \Rightarrow \neg B) = \frac{\sup(A) - \sup(A \cup B)}{\sup(A)} = 1 - conf(A \Rightarrow B)$$

$$conf(\neg A \Rightarrow B) = \frac{\sup(B) - \sup(A \cup B)}{1 - \sup(A)}$$

$$conf(\neg A \Rightarrow \neg B) = \frac{1 - \sup(A) - \sup(B) + \sup(A \cup B)}{1 - \sup(A)} = 1 - conf(\neg A \Rightarrow B)$$

Where $conf(X)$ as the confidence function, such as $conf(\neg A \Rightarrow B)$ is the confidence of negative association rules $\neg A \Rightarrow B$.

The relationship between them as follows:

(1) $conf(A \Rightarrow B) + conf(A \Rightarrow \neg B) = 1$

(2) $conf(\neg A \Rightarrow B) + conf(\neg A \Rightarrow \neg B) = 1$

Proof omitted.

We can see from the above analysis, the two confidence $P\_minconf + N\_minconf = 1$

B. Correlation

Most association rules mining algorithms use the framework of support-confidence. Although it may exclude some meaningless rules, it still produce some rules which the users are not interested in. Suppose there are 10000 sales data items in the set , purchase of goods X denoted as the item A, purchase of goods Y denoted as the item B, purchase of X and Y denoted as item $A \cup B$. Assume the purchase of goods X is 6000, purchase of goods is 5000, purchase of goods X and Y are 2500.Given *minsup*=0.2,*minconf*=0.3,then $\sup(A \cup B) = 0.25 > minsup$, $conf(A \Rightarrow B) = 0.42 > minconf$, so $A \Rightarrow B$ is a strong association rules. Then considering negative association rules $A \Rightarrow \neg B$,since $sup(A \cup B) = 0.35 > minsup$, $conf(A \Rightarrow \neg B) = 0.58 > minconf$,therefore, $A \Rightarrow \neg B$ also is strong association rules. This is in contradiction with the $A \Rightarrow B$, meanwhile, because of $conf(A \Rightarrow \neg B) > conf(A \Rightarrow B)$, so $A \Rightarrow \neg B$ should be more reliable. Because of $conf(A \Rightarrow B) = P(B \mid A) < P(B)$, the occurrence of A lead to probability B is decreased, so A and B should be a negative relation. For this, it introduced another measure of association rules - Correlation. The correlation of association rules show the relationship between the itemset.,through the correlation threshold, we can remove some meaningless negative association rules.

Definition 3.1 If $P(A \cup B) = P(A)P(B)$,then itemsets A and itemsets B are independent; Otherwise, itemsets A and itemsets B are correlated.

The correlation of itemsets A and itemsets B:

$$corr_{A,B} = \sup(A \cup B) - \sup(A)\sup(B)$$

Which $corr_{A,B}$ is the correlation of itemsets A and B; $sup(A \cup B)$、$sup(A)$、$sup(B)$, respectively, the support of itemsets $A \cup B$、A and B.

There are three kinds of $corr_{A,B}$:

If $corr_{A,B} > 0$,then itemsets A and B is positive correlation;

If $corr_{A,B} < 0$,then itemsets A and B is negative correlation;

If $corr_{A,B} = 0$,then itemsets A and B is independent of each other.

The correlation reflects the relationship between the itemsets A and B , when the correlation is equal to 0, indicating that two itemsets appear together does not have special meaning, namely, A and B are independent of each other, called the rules are not relevant rules; when the correlation less than 0, indicating that the possibility one of the itemsets would be decreased by another itemsets, called the rules are negative correlation rules;when the correlation greater than 0, indicating that the possibility one of the itemsets would be increased by another itemsets, called the rules are positive correlation rules.

Because of the meaning expressed by the three cases vary widely, according to the traditional rules generated by Apriori algorithm is no distinction for them, so some rules are contradictory or uninteresting

The paper [6] proposed that if a rules to meet $sup(A \cup B) - sup(A)sup(B) \approx 0$, then the rules are no interesting, in other words, only the correlation of rules is greater than a certain threshold, the user will be interested in, Therefore, we set the minimum correlation threshold (*mincorr*), that only the association rules to meet $sup(A \cup B) - sup(A)sup(B) \geq mincorr$ is our interest. *mincorr* specified by the user or expert. For the negative association rules, the value of $sup(A \cup B) - sup(A)sup(B)$ may be less than 0, therefore, it need to join the absolute value on the left of the formula, that is, if $A \Rightarrow B$ is interested in the rules, if and only if $| sup(A \cup B) - sup(A)\ sup(B) | \geq mincorr$. The following will discuss the relationship between the minimum correlation of four forms of association rules .

Theorem 3.1

If $| sup(A \Rightarrow B) - sup(A)sup(B) | \geq mincorr$, then

(1) $| sup(\neg A \Rightarrow B) - sup(\neg A)sup(B) | \geq mincorr$

(2) $| sup(A \Rightarrow \neg B) - sup(A)sup(\neg B) | \geq mincorr$

(3) $| sup(\neg A \Rightarrow \neg B) - sup(\neg A)sup(\neg B) | \geq mincorr$

Proof. (1) $| sup(\neg A \Rightarrow B) - sup(\neg A)sup(B) |$

$= | sup(B) - sup(A \cup B) - (sup(B) - sup(A)sup(B)) |$

$= | -sup(A \cup B) + sup(A)sup(B) |$

$= | sup(A \Rightarrow B) - sup(A)sup(B) | \geq mincorr$

(2), (3) for the same reason.

The theorem show that select an appropriate minimum correlation can be prune to four association rules.

C. Algorithm Design

In the PAR&NAR_Mining algorithm, assuming that the frequent itemsets have been obtained and saved in the set *L*.

Input: frequent itemsets *L*, minimum confidence *minconf*, the minimum correlation *mincorr*;

Output: positive association rules set of PAR, negative association rules set of NAR;

PAR=∅ ,NAR=∅ ;

For any frequent itemsets *X* in *L* do begin

For any itemsets $A \cup B=X$ and $A \cap B=\varnothing$ do begin

If $| sup(A \cup B) - sup(A)sup(B) | \geq$ mincorr do begin

If $corr_{A,B} > 0$ then begin // Produce rules forms such as $A \Rightarrow B$ and $\neg A \Rightarrow \neg B$

If $conf(A \Rightarrow B) \geq P\_minconf$ then

PAR=PAR $\cup \{A \Rightarrow B\}$;

If $conf(\neg A \Rightarrow \neg B) \geq P\_minconf$ then

NAR=NAR $\cup \{\neg A \Rightarrow \neg B\}$;

End;

If $corr_{A,B} < 0$ then begin // Produce rules forms such as $A \Rightarrow \neg B$ and $\neg A \Rightarrow B$

If $conf(A \Rightarrow \neg B) \geq N\_minconf$ then

NAR=NAR $\cup \{A \Rightarrow \neg B\}$;

If $conf(\neg A \Rightarrow B) \geq N\_minconf$ then

NAR=NAR $\cup \{\neg A \Rightarrow B\}$;

End;

End;

End;

End;

Return PAR and NAR;

The algorithm to generate positive association rules set PAR and negative association rules set NAR. PAR and NAR is initialized to empty set. First, check $| sup(A \cup B) - sup(A)sup(B) |$, whether to meet the minimum correlation, and then, determine its relevance based on $corr_{A,B}$,and generate rules. Which, if *A* and *B* are positive correlation, production rules of the form $A \Rightarrow B$ and $\neg A \Rightarrow \neg B$; if *A* and *B* are negative correlation, production rules of the form $\neg A \Rightarrow B$ and $A \Rightarrow \neg B$. Finally back to PAR and NAR, and end of the algorithm.

IV. Algorithm Analysis

Let *n* frequent items in *L*, frequent itemsets X contains two frequent item, where $X \subseteq L$、$A \cup B=X$ and $A \cap B=\Phi$, there are c_n^2 frequent itemsets *X* in *L*, each *X* requires two comparative analysis, so the operation frequency of the algorithm is $2 * c_n^2 = n(n\text{-}1)$, that is the time complexity is $O(n^2)$.The operation frequency of Apriori algorithm also reached the square level ,or even higher. Therefore, in theory, the algorithm is feasible and effective.

To illustrate the algorithm efficiency and accuracy, in the same experimental conditions, to achieve PAR&NAR_ Mining algorithm and the positive and negative association rules mining algorithm in [7], and using the same experimental data set to analyze and compare the performance of the algorithm.

This experimental platform is the Intel (R) P4 dual-core processor, 1.5G RAM, WindowsXP operating system, programming language is JAVA, programming environment is JBuilder2006. Data set used in the experiments is the social security audit data set, which is a real data set.

There are two comparison for the experiments:

a) Positive and negative association rules mining algorithm in [7];

b) PAR&NAR_Mining algorithm based on correlation and dual confidence thresholds.

Given *minsupp*=0.2,*P_minconf*=0.2,*N_minconf*=0.8. the comparison results shown in Figure 1、2.

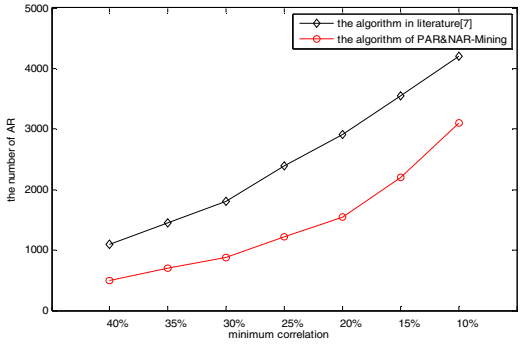

Figure 1. The number of AR change as minimum correlation

As can be seen from Figure 1, the number of association rules generated by PAR&NAR_Mining algorithm is

significantly less than [7], this is because the algorithm introduction of the minimum correlation and double confidence theory, filter out the association rules of meaningless, allowing the rules to become more objective and reasonable.

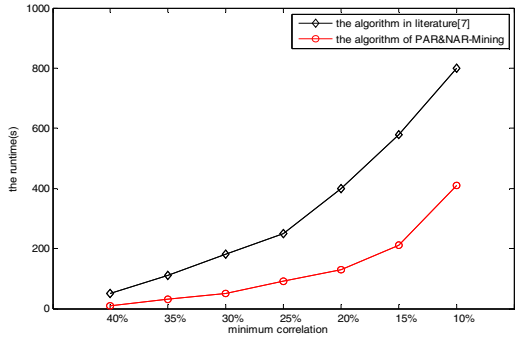

Figure 2. The runtime change of algorithems as minimum correlation

As can be seen from Figure 2, PAR& NAR_Mining algorithm execution time is much better than [7] .These experiments fully demonstrate that the algorithm is effective.

V. CONCLUSION

Since the topic of association rules mining proposed, has proposed a number of association rules mining algorithm, but can be used for negative association rules mining are few. This paper *presents* a both positive and negative association rules mining algorithm based on relevance and confidence ,experiments show that the algorithm is effective and feasible. In addition, this paper only discusses the problem of mining positive and negative association rules, in the future, we will study the following two aspects:

a) To improve the efficiency of the algorithm.

b) Depth study of the issue of its updated .

ACKNOWLEDGMENT

This work is sponsored by the National Natural Science Foundation of China under grant number 60873038, the Fundamental Research Funds for the Central Universities of China under grant number HEUCF100603 and the National Science & Technology Pillar Program under grant number 2009BAH42B02.

REFERENCES

[1] Agrawal R,Imielinski T,Swami A, “Mining Association Rules between Sets of Items in Large Databases, ”In:Proc of the ACM SIGMOD International conference on Management of Data,Washington DC,1993,pp,207-216.

[2] Agrawal R,Srikant R, “Fast algorithms for mining association rule, ”In:Proe.20th Int. Conf.on VLDB.Santiago,Chile,1994,pp,487-499.

[3] Mojdeh Jalali-Heravi, Osmar R. Zaïane, “A Study on Interestingness Measures for Associative Classifiers, ” Proceedings of the 2010 ACM Symposium on Applied Computing,2010,pp,1039-1046.

[4] He Jiang, Yuanyuan Zhao, Chunhua Yang, Xiangjun Dong, “Mining Both Positive and Negative Weighted Association Rules with Multiple Minimum Supports,” 2008 International Conference on Computer Science and Software Engineering,2008,pp,407-410.

[5] Ling Zhou, Stephen Yau, “Association Rule and Quantitative Association Rule Mining among Infrequent Items,” Proceedings of the 8th international workshop on Multimedia data mining: associated with the ACM SIGKDD 2007.

[6] Liqiang Geng,Howard J.Hamilton ,“Interestingness measures for data mining: A survey.” ACM Computing Surveys ,2006,pp,1-32.

[7] SHANG Shi-ju,DONG Xiang-jun,LI Jie, “Algorithms for mining negative association rules in multi-database, ”Computer Engineering and Applications,2009,45(24),pp,150-152.