ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

An Effective Algorithm to Generate Positive and Negative Association Rules

Dr.B.Ramasubbareddy[1], K.Srinivas[2], B.Kavitharani[3]

Professor, SV College of Engineering, Tirupati, India[1]

Associate Professor, Jyothishmathi Institute of Technology and Sciences, Karimnagar, India [2,3]

ABSTRACT: Recently, mining negative association rules has received some attention and been proved to be useful in real world. This paper presents an efficient algorithm (PNAR) for mining both positive and negative association rules in databases. The algorithm extends traditional association rules to include negative association rules. When mining negative association rules, we use same minimum support threshold to mine frequent negative itemsets. With a Yule's Correlation Coefficient measure and pruning strategies, the algorithm can find all valid association rules quickly and overcome some limitations of the previous mining methods. The experimental results demonstrate its effectiveness and efficiency.

KEYWORDS: Association Rule Mining, Data Mining, Frequent Itemsets, Minimum Support, Yule's Correlation Coefficient.

I. INTRODUCTION

The research and application of data mining technology are a hot spot in database and artificial intelligence at recent years. Association Rules Mining introduced by R. Agrawal [1] is an important research topic among the various data mining problems. Association rules have been extensively studied in the literature for their usefulness in many application domains such as market basket analysis, recommender systems, diagnosis decisions support, telecommunication, intrusion detection, and etc. All the traditional association rule mining algorithms were developed to find positive associations between itemsets. Several algorithms has been developed to cope with the popular and computationally expensive task of association rule mining, such as Apriori [1], AIS [2], DHP [3], Partition [4], and etc.. With the increasing use and development of data mining techniques and tools, much work has recently focused on finding negative patterns, which can provide valuable information. However, mining negative association rules is a difficult task, due to the fact that there are essential differences between positive and negative association rule mining. We will attack two key problems in negative association rule mining:
(1) How to effectively search for negative frequent itemsets.
(2) How to effectively identify negative association rules.

Although some researchers pointed out the importance of negative associations, only some groups of researchers ([5], [6], [7] and etc.) proposed an algorithm to mine these types of associations. This not only illustrates the novelty of negative association rules, but also the challenge in discovering them.

II. BASIC KNOWLEDGE

A. CONCEPTS AND DEFINITIONS

Let $I=\{i_1,i_2,...,i_n\}$ be a set of n distinct literals called items. Let DB be a set of transactions, where each transaction T is a set of items, and each transaction is associated with a unique identifier called TID. Let A, called an itemset, be a set of items in I. The number of items in an itemset is the length (or the size) of an itemset. Itemsets of length k are referred to as k-itemsets. A transaction T is said to contain A if $A \subset T$. An association rule is an implication

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

of the form A=>B, where A ⊂ I, B ⊂ I, and A∩B=φ. We call A the antecedent of the rule, and B the consequent of the rule.

The rule A=>B has a support (denoted as supp) **s** in DB if **s%** of the transactions in DB contains A=>B. In other words, the support of the rule is the probability that A and B hold together among all the possible presented cases. i.e.
supp(A=>B)=supp(AUB)=P(AUB)…….. (1)

The rule A=>B has a measure of its strength called confidence (denoted as conf) c if c% of transactions in DB that contain A also contain B. In other words, the confidence of the rule is the conditional probability that the consequent B is true under the condition of the antecedent A. i.e.
conf(A=>B)=P(B|A)=supp(AUB)/supp(A)… (2)

B. CLASSICAL METHOD

The classical method is well known as the support confidence framework for association rule mining [1]. It can be decomposed into the following two issues:

(1) Generate all frequent itemsets: All itemsets that have a support greater than or equal to user-specified minimum support (ms) are generated.

(2) Generate all the rules that have a user-specified minimum confidence (mc) in the following naive way: For every frequent itemset X and any B⊂X, let A=X-B. If the rule A=>B has the mc, then it is a valid rule.

The generative rules are called interesting positive rules. A frequent itemset (denoted as PL) [8] is an itemset that meets the user-specified ms. Accordingly we define an infrequent itemset (denoted as NL) as an itemset that does not meet the user-specified ms. The second sub problem is straight forward and can be done efficiently in a reasonable time. However, the first sub problem is very tedious and computationally expensive for very large database and this is the case for many real life applications.

In order to generate the frequent itemsets, an iterative approach is used to first generate the set of frequent 1-itemsets L_1, then the set of frequent itemsets L_2, and so on until for some value of r the set L_r is empty. At this stage the algorithm can be terminated. During the k-th iteration of this procedure a set of candidates C_k is generated by performing a (k-2)-join on the frequent itemsets L_{k-1}. The itemsets in this set C_k are candidates for frequent itemsets, and the final set of frequent itemsets L_k must be a subset of C_k. Each element of C_k needs to be validated against the transaction database to see if it indeed belongs to L_k.

The validation of the candidate itemset C_k against the transaction database seems to be bottleneck operation for the algorithm. In order to improve the algorithm efficiency, the Apriori property is introduced that all subsets of a frequent itemset A in DB are also frequent in DB, and all supersets of an infrequent itemset A in DB are also infrequent in DB.

C. NEGATIVE ASSOCIATION RULES

The negation of an itemset A is indicated by ¬A, which means the absence of the itemset A. We call a rule of the form A=>B a positive association rule, and rules of the other forms (A=>¬B, ¬A=>B and ¬A=>¬B) negative association rules.

The support and confidence of the negative association rules can make use of those of the positive association rules [9]. The support is given by the following formulas:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

supp(¬A)=1-supp(A)……...…… (3)
supp(A=>¬B)=supp(A)-supp(AUB) ..…… (4)
supp(¬A=>B)=supp(B)-supp(AUB)……(5)
supp(¬A=>¬B)=1-supp(A)-supp(B)+supp(AUB)...(6)

The confidence is given by the following formulas:

$$conf(A \Rightarrow \neg B) = \frac{\sup p(A) - \sup p(A \cup B)}{\sup p(A)}. \quad (7)$$

$$conf(\neg A \Rightarrow B) = \frac{\sup p(A) - \sup p(A \cup B)}{1 - \sup p(A)}. \quad (8)$$

$$conf(\neg A \Rightarrow \neg B) = \frac{1 - \sup p(A) - \sup p(B) + \sup p(A \cup B)}{1 - \sup p(A)}. \quad (9)$$

The negative association rules discovery seeks rules of the three forms with their support and confidence greater than, or equal to, user-specified ms and mc thresholds respectively. These rules are referred to as an interesting negative association rule.

III. PNAR ALGORITHM

A. YULE'S CORRELATION COEFFICIENT

When mining positive and negative association rule at the same time, we will find that the mining rules are contradictory frequently. For example, the rules of the forms A=>¬B and ¬A=>B may be mined together, but the two rules are contradictory. In order to resolve these contradictions, we can judge the types of mining association rules by the correlation coefficient [10]. Let A and B for the two itemsets. The correlation coefficient (denoted as $Q_{A,B}$) can show the relevance of the two itemsets. As follows:

$$Q_{A,B} = \frac{Supp(AB)\,Supp(\neg A\neg B) - Supp(A\neg B)\,Supp(\neg AB)}{Supp(AB)\,Supp(\neg A\neg B) + Supp(A\neg B)\,Supp(\neg AB)}$$

The value of Yule's correlation coefficient exist the following three situations:

(1) If $Q_{A,B}>0$, then A and B are positive correlation. The more A occurs in a transaction the more B will likely also occur in the same transaction and vice versa.

(2) If $Q_{A,B}<0$, then A and B are negative correlation. The more A occurs in a transaction the less B will likely also occur in the same transaction and vice versa.

By the definition of correlation coefficient, we can conclude the below lemmas:

Lemma 1: If the itemset A and B are positive correlation, then the forms of A=>B or ¬A=>¬B will be mined.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

Lemma 2: If the itemset A and B are negative correlation, then the forms of A=>¬B or ¬A=>B will be mined.

B. PRUNING STRATEGIES

As we have seen, there can be an exponential number of infrequent itemsets in a database, and only some of them are useful for mining interesting association rules. Therefore, pruning strategy is critical to efficient search for interesting frequent negative itemsets. When mining negative association rules, we can adopt same minimum support (ms) and minimum confidence (mc) threshold to improve the usability of the rules. Through the experimental analysis, we found that the association rules of the forms A=>B and ¬A=>¬B have considerable proportion when mining both positive and negative association rules. In particular, the number of the form ¬A=>¬B is very large, and these rules including pure negative itemsets are usually of little use in real application. For example, we assume that the database DB in a supermarket contain n transactions. Now we concern the sale of tea (t) and coffee (c). Suppose we mine the rule of the form ¬t=>¬c, which means customers not to buy tea and coffee in a transaction, the result is not useful to our market basket analysis. So we adopt a pruning strategy that we will not to consider the part negative association rules of the form ¬A=>¬B to improve mining efficiency. The search space can be significantly reduced by the pruning strategy.

In addition, we are only interested in those absence itemsets whose positive counterparts are frequent for market basket analysis when mining negative association rules. For example, the absence itemset ¬A is not show if the itemset A is not frequent. The pruning strategy is more benefit to generate frequent 1-itemset. Because of reducing the number of frequent 1-itemset, the number of frequent and infrequent k-itemset is reduced accordingly.

C. PNAR ALGORITHM

As mentioned before, the process of mining both positive and negative association rules can be decomposed into the following three sub problems, in a similar way to mining positive rules only.

(1)Generate the set PL of frequent itemsets and the set NL of infrequent itemsets.
(2)Extract positive rules of the form A=>B in PL.
(3)Extract negative rules of the forms A=>¬B and ¬A=>B in NL.

Let DB be a database, and ms, mc, dms and dmc given by the user. Our algorithm for extracting both positive and negative association rules with a correlation coefficient measure and pruning strategies is designed as follows:

Algorithm: **Positive and Negative Association Rules**

Input: TDB-Transactional Database

MS-Minimum Support

MC-Minimum Confidence

Output: Positive and Negative Association Rules

Method:

1. P← Φ, N←Φ
2. Find F_1← Set of frequent 1- itemsets
3. for (k=2;F_{k-1}!=Φ; k++)
4. {

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

5. $C_k = F_{k-1}$ join F_{k-1}
6. // Prune using Apriori Property
7. for each i ε C_k, any subset of i is not in F_{k-1} then $C_k = C_K$ - { i }
8. for each i ε C_k find Support(i)
9. for each A,B (A U B= i)
10. {
11. $Q_{A,B}$= Association(A,B);
12. if Q>0
13. if(**supp**(A→B)≥MS &&**conf**(A→B)≥MC) then
14. P← P U { A→ B}
15. if Q<0
16. {
17. if(**supp**(A→┐ B)≥MS&&**conf**(A→┐ B)≥MC) then
18. N← N U { A→ ┐ B}
19. if(**supp**(┐ A→B)≥MS&&**conf**(┐ A→B)≥MC) then
20. N← N U { ┐ A→ B}
21. }
22. }
23. }
24. AR← P U N

PNAR generates not only all positive association rules in PL, but also negative association rules in NL. When mining negative association rules, we use same threshold to improve the usability of the frequent negative itemsets. With a Yule's correlation coefficient measure and pruning strategies, the algorithm can find all valid association rules quickly. An example of mining positive and negative itemsets is given below for illustrative purposes.

IV. RESULTS

For the convenience of comparison, we conducted our experiments on the synthetic dataset to study the behaviors of the algorithm.

Example: Let us consider a small transactional table with 10 transactions and 6 items. In Table1 a small transactional database is given.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

Table1: A Transaction Database TD

TID	Items	TID	Items
1	A,C,D	6	E
2	B,C	7	B,F
3	C	8	B,C,F
4	A,B,F	9	A,B,E
5	A,C,D	10	A,D

In the following tables, Lk is denoted as all frequent k-itemset. Given that ms=0.3, all the positive and negative frequent itemsets can then be discovered. And in Table2, we compare three algorithms in the same TD.

From Table2 we discover that the PNAR algorithm can reduce the number of the positive and negative frequent itemsets efficiently. Especially the number of frequent 2-itemset is less than before in a certain extent. From Table2 we can detect the number of frequent 2-itemset is less 19 itemsets than SRM algorithm [6]. So the PNAR algorithm can reduce the search space efficiently and improve the efficiency, and can overcome some limitations of the previous mining methods. The proposed algorithm can generate negative association rules without calculating negative frequent itemsets.

Table2: Comparison of the Three Algorithms

	MS	**L1**	**L2**	**L3**	**L4**
Apriori	0.3	5	2	0	0
SRM	0.3	10	21	10	1
PNAR	0.3	10	2	0	0

V. CONCLUSION

In this paper, we have designed a new algorithm for efficiently mining positive and negative association rules in databases. Our approach is novel and different from existing research. We have designed pruning strategies for reducing the search space and improving the usability of mining rules, and have used the Yule's correlation coefficient to judge which form association rule should be mined. It is shown by empirical studies that the proposed approach is effective, efficient and promising.

REFERENCES

[1] R. Agrawal, T. IMIELINSKI, and A. SWAMI, "Mining association rules between sets of items in massive databases," In Proc. of the 1993 ACM SIGMOD International Conference on Management of Data, ACM, Washington D.C., 1993, pp. 207-216.
[2] R. Agrawal, R. Srikant, "Fast Algorithms for Mining Association Rules," In Proc. of the 20th Int. Conf. on Very Large Databases(VLDB '94), Santiago, Chile, 1994, pp. 487-499.
[3] J. S. Park, M. S. Chen, and P. S. Yu, "An Effective Hash-based Algorithm for Mining Association Rules," In Proc. of the ACM SIGMOD Int. Conf. on Management of data (ACM SIGMOD '95), San Jose, California, 1995, pp. 175-186.
[4] A. Savasere, E. Omiecinski, and S. Navathe, "An efficient algorithm for mining association rules in large databases," In Proc.1995 Int. Conf. Very Large Database (VLDB'95), Zurich, Switzerland, 1995, pp. 1-24.
[5] A. Savasere, E. Omiecinski, and S. Navathe, "Mining for strong negative associations in a large database of customer transactions," In Proc. of ICDE, 1998, pp. 494-502.
[6] W. Teng, M. Hsieh, and M. Chen, "On the mining of substitution rules for statistically dependent items," In Proc. of ICDM, 2002, pp. 442-449.
[7] X. Wu, C. Zhang, and S. Zhang, "Efficient Mining of Both Positive and Negative Association Rules," ACM Transactions on Information Systems, Vol. 22, No. 3, 2004, pp. 381–405.
[8] M. CHEN, J. HAN, and P. YU, "Data mining: An overview from a database perspective," IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, 1996, pp. 866-883.
[9] X. Dong, S. Wang, H. Song, and Y. Lu, "Study on Negative Association Rules," Transactions of Beijing Institute of Technology, Vol. 24, No. 11, 2004, pp. 978-981.
[10] S. Brin, R. Motwani, and C. Silverstenin, "Beyond market baskets: Generalizing association rules to correlations," In Proc. of the 1997 ACM SIGMOD International Conference on Management of Data, ACM, Tucson, Arizona, 1997, pp. 265–276.
[11] Honglei Zhu and Zhigang Xu "An Effective Algorithm for Mining Positive and Negative Association Rules," 2008 International Conference on Computer Science and Software Engineering.