The Research of Data Mining Algorithm Based on Association Rules

Lei Chen
Basic Courses Teaching Department, The Chinese People's Armed Police Force Academy
Langfang 065000, China
E-mail:chen_nuli@126.com

***Abstract*-Based on in-depth study of the existing data mining and association rule mining algorithms, a new mining algorithm of weighted association rules is proposed. By introducing a support factor of the weighted frequent item sets, a reasonable minimum support is set. The algorithm does not need to repeatedly scan the database in the discovery of frequent item sets, so it greatly reduces the time of input and output, and improves the efficiency of data mining.**

Keywords-Data mining, Association rules, Apriori algorithm, Frequent item set

I. INTRODUCTION

Data mining coming from the rapid growth of information, is the process to extract, identify and find the potentially useful and ultimately understandable knowledge from the data. Data mining technology is to identify patterns of data in the data already existed, to help users understand the existing information and predict for the future conditions on the basis of existing information.

Association rule mining is an important research branch of data mining, mainly used to discover the relevant links between items in the data set. Apriori algorithm and FP-Tree algorithm are classical algorithms in association rule mining, both mining based on that the amount of data in the transaction database will not change and each data item has equal importance[1]. However, in actual application, data in the database is constantly changing, and people's concern for different data items is not the same. If we still use the traditional mining algorithms for association rule mining, the mining efficiency will be very low, and the mining results are not accurate enough. Therefore it is necessary to improve the existing data mining algorithms to meet the needs of association rule mining.

II. CHARACTERISTICS OF ASSOCIATION RULE MINING

A. Mathematical model of assocIation rule mining

The purpose of association rule mining is to find out the hidden relationship between different data item sets in the database. In general, given a transaction database, the problem of association rule mining is the process to find the association rules through a user-specified minimum support and minimum confidence. Let $I = \{i_1, i_2, \cdots, i_m\}$ be the set of all data items. D is the set of all transactions in the transaction database, in which each transaction T is a collection of items, and each transaction T has a unique TID identifier. For the item set $X \subseteq I$, which is called that T contains X only if $X \subseteq T$. If $|X| = k$, then the collection is called k order set. If the items arranged according to the dictionary order, k order item set X can be expressed as $X[1]X[2]\cdots X[k]$. For the item set $C = X \bigcup Y$, if Y is m order item set, then Y is the m order expansion of X .If X, Y are project sets, and $X \bigcap Y = \Phi$, then a containing formula $X \Rightarrow Y$ is known as association rules, and X, Y is separately called the premise and conclusion of the association rule $X \Rightarrow Y$.

Support is a measure of statistical significance of association rules in the entire transaction database, which demonstrates the degree of representation of the rule in all transactions. The greater the support, the more important the rule. If $Support(X)$ is not less than the user-specified minimum support $MinSup$, then X is called frequent item set, referred to as the frequent or large item set, otherwise known as non-frequent item set, referred to as non-frequent or small item set. The confidence is a measure of association rules's accuracy . If $Support(X \Rightarrow \mathrm{Y}) \geq MinSup$ and $conf(X \Rightarrow Y) \geq MinConf$, then the association rule is effective or strong association rule, otherwise, said the association rule to be invalid or weak association rule.

The proportion of the rule $X \Rightarrow Y$ in the transaction database describes the appearance probability of the item set Y in all transactions in the absence of any condition effect. Degree of interest is the ratio between the confidence and expectations[2], which describes the impact of the emergence of item set X on the emergence of item set Y. Under normal circumstances, the degree of interest of a useful association rule is greater than 1. If the interest is no more than 1, this association rule does not make much sense.

B. Process of association rule mining

Association rule mining is to find the association rules met the user-specified minimum support and minimum confidence requirement from the transaction database D. The entire mining process can be decomposed into the following two steps: first, find all frequent item sets, that is, find all the item sets had support greater than the given support threshold; second, based on the obtained frequent item sets, generate a corresponding strong association rule, that is, generate the association rules had support and confidence respectively greater than or equal to the given support threshold and confidence threshold[3]. In addition, interest can be used to help extract valuable association rules, remove the redundant or worthless association rules.

Published by Atlantis Press, Paris, France.

The basic model of association rule mining is shown in figure 1.

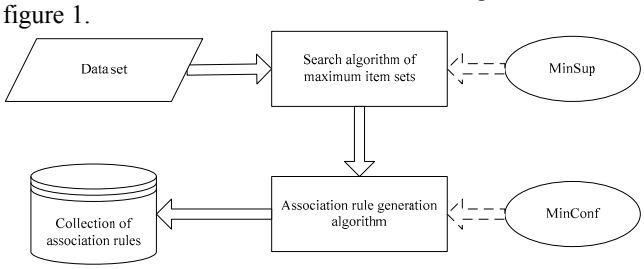

Figure 1. The basic model of association rule mining

In the above two steps, the second step is relatively easy, because it only needs to list all possible association rules based on the frequent item sets have been found, and then use the support threshold and confidence threshold to measure them, and the association rules both met the support threshold and confidence threshold requirements are considered to be interesting association rules.

C. Weighted association rule mining algorithms

Association rules reflecting the interesting connections between data items in transaction database, is an important subject in knowledge discovery in databases. The attributes in the database are equal and consistent. That is, among the items in the database there is no importance and subordination distinction, and their importance is to be measured by calculating the frequency of their occurrence. Allocation of the items in the database is uniform, that is, the same or similar frequency. However, in the transaction database with actual transactions the items are often unevenly distributed, and some even the frequency may vary greatly, for example, some product project is new listed on or improved goods, some goods may be going promotions[4]. These all will lead to uneven distribution of the data items in the database. However, when this happens, it will result in problem when setting the minimum support high or low. If set too high, the association rules found may not involve the items with lower frequency; too low, there will be too many meaningless or even false association rules, but also may lead to combinatorial explosion, thereby reducing the efficiency of the algorithm. In order to reflect the different importance of various items, we have introduced the concept of the weighted item, to assign different weights for different items, to reflect the different importance of the items in the database, consequently to extend the issue model, the so-called weighted association rules.

Make D as the transaction database, a collection $I=\{i_1,i_2,\cdots,i_m\}$ of items is all the item sets in it, in which each transaction actually occurred is a subset of I, and given a transaction ID TID, $W=\{w_1,w_2,\cdots,w_m\}$ for the weight set corresponding to I, $X=\{x_1,x_2,\cdots,x_p\}$ and $Y=\{y_1,y_2,\cdots,y_q\}$ for subsets of I, $X,Y\subset I$, and $X\bigcap Y=\Phi$, $Spport(X)$ known as the support of X in D, $Confidence(X)$ as the confidence, as well as $WMinSup$ for the minimum weighted support, then the issue of weighted association rules to be discussed is formalized as $X\Rightarrow Y$. With the introduction of the concept of weights, the weighted frequent item set does not have the anti-monotony of the frequent item sets in general association rule mining, in other words, in the mining of weighted frequent item sets, because of different weights, the subset of them may no longer be weighted frequent item sets, so can not be the same as an ordinary association rule mining, but must seek new solutions for the mining of weighted association rules[5].

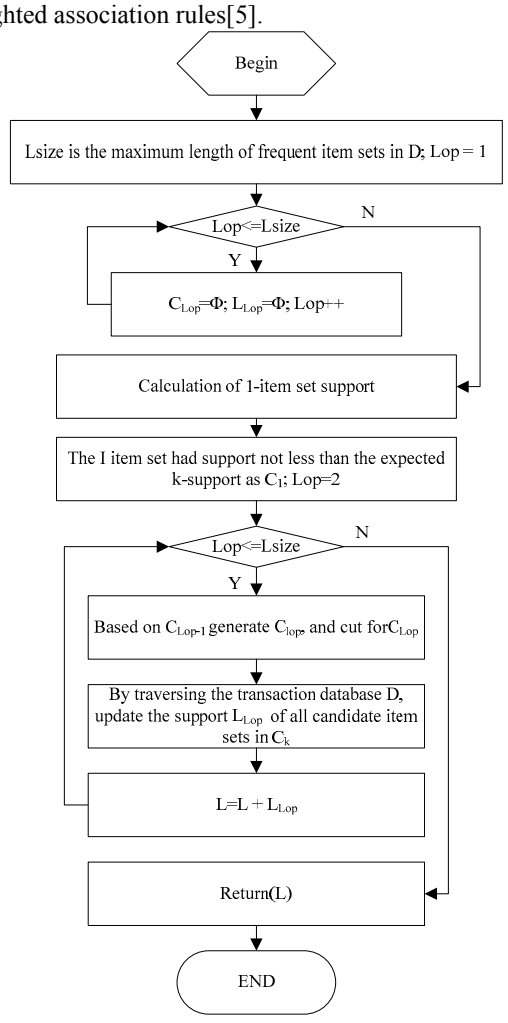

Figure 2. Implementation process of weighted data mining algorithm

Weighted association rule mining algorithm has a framework similar to Apriori algorithm, both through frequent item sets to generate candidate items, but in some specific details is clearly different. The subset of weighted association rule frequent item set may not be frequent, so we can not simply generate candidate k-item set only by the

first (k-1) frequent item set like the Apriori algorithm. For the production of 1-item set, to extract such k-item set from the database, and use J-support expectations, according to all candidate k-item sets to obtain by calculation. J is between the maximum possible length of the k and frequent item set, if the support of a certain k-item set is lower than all the J-support expectations, we can conclude that in the subsequent traversal it is impossible the subset of any frequent item set, thus it can be removed. All the item sets which may be frequent item subset constitute the candidate set C_k. The mining of weighted association rules makes MINWAL (0) as the research object. Set transaction database as D, weight set as W, weighted minimum support as $WMinSup$ and weighted minimum confidence as $WMinConf$, then MINWAL (0) algorithm is shown as in figure 2.

III. IMPROVEMENT OF WEIGHTED ASSOCIATION RULE MINING ALGORITHM

The weighted support of the above weighted association rule algorithms may be greater than 1, which is in contradiction with that the actual support should be less than 1. Algorithms need to frequently scan the database and generate a large number of candidate sets, and this tends to greatly affect the performance of the algorithm. For these reasons it needs to improve the existing data mining algorithms of weighted association rules[6]. The improving idea is as follows: use the classical algorithm of association rules to obtain the frequent item sets without weights, after weight treatment, the sum of weights is less than or equal to 1, so these frequent sets are calculated weighted superset; calculate the weighted support of all the attribute sets in this superset, and remove the attribute sets with weighted support less than the minimum weighted support, resulting in all the weighted large attribute sets; use them to generate the required weighted association rules. As the definition of the weighted confidence is the same to the definition of common association rule confidence, so the Boolean association rules can be used directly to generate algorithm.

In the process of weighted association rule mining, most of the weighted association rule mining algorithms are based on Apriori algorithm, but Apriori algorithm needs to frequently scan the database, resulting in a large number of candidate items, which seriously affected the efficiency of the algorithm. In order to improve the efficiency of the algorithm, in this paper the FP-Growth algorithm is as the basis for mining of weighted item sets. However, it must ensure that the frequent set generated by using the FP-Growth algorithm is a superset of the weighted frequent sets. In other words, it needs to set a reasonable minimum support for FP-Growth algorithm. If the minimum support sets too low, the efficiency of the algorithm will be certainly affected; if the minimum support sets too high, it can not guarantee that the frequent set generated by an ordinary association rule algorithm is a superset of the weighted frequent sets[7]. To this end, the introduction of a support factor of the weighted frequent item sets, represented by δ, δ is the reciprocal of the sum of all trading weights. Just making $\delta WMinSup$ as the minimum support of ordinary association rule mining, then the obtained frequent item set L is certainly the superset of the weighted frequent item set. The implementation process of improved data mining algorithm of weighted association rules is shown as in figure 3.

The traditional association rule mining algorithm makes each record, each item with equal importance. However, in reality it is often not. To solve this problem, we have introduced weights. The improved algorithm needs not to repeat database scans when discovering the frequent item sets, the mining efficiency is greatly improved, such increase will be more apparent with the transaction data increase.

IV. CONCLUSION

Discovery of association rules is the most successful and important task in data mining, and its goal is to discover all the frequent patterns in the data set. Association rules can be used to find the link between different products (items) in the transaction database, through which the buying behavior patterns of customers can be found out. Because different items have different importance, this paper introduces the concept of weighted association rule mining, and a new mining algorithm for weighted association rules is presented too. The algorithm does not need to repeatedly scan the database when discovering frequent item sets, therefore can greatly improve the efficiency of data mining.

REFERENCES

[1] Sankar K. Pal, Sanghamitra Bandyopadhyay, Shubhra Sankar Ray. Evolutionary Computation In Association Rules: A Review. IEEE Transactions on Systems, Man, and Cybernetics-PartC: Applications And Reviews, 2006, 36(5): 601-615.

[2] Pan Hua, Xiang Tongde. Data Warehouse and Data Mining principles, tools and applications,China Electric Power Press, 2007, 12(96): 101-106.

[3] Costa Grahne and Jianfei Zhu. Efficiently Using Prefix-trees in Mining Frequent Item set. In Proceeding of the IEEE ICDM Workshop on Frequent Item set Mining Implementations, Nov. 2003.

[4] Jacques Cohen. Bioinformatics — an Introduction for Computer Scientists, Data mining Computer Surveys, 2004, 36(2): 122-158.

[5] Agrawal R, Aggarwal C, Prasad V. A tree pmjection algorithm for generation of frequent item sets [J]. Journal of Parallel and Distributed Computing, 2001, 61(3): 350-371.

[6] Chen Lei-da etc. Data mining methods, applications, tools, Information Systems Management [M].2000, 17(1): 6570.

[7] Kay C. Wiese, Edward Glen. An association rules based genetic algorithm for RNA secondary structure prediction. Soft Computing Systems: Design, Management and Application, 2002, 4(1): 173-182.

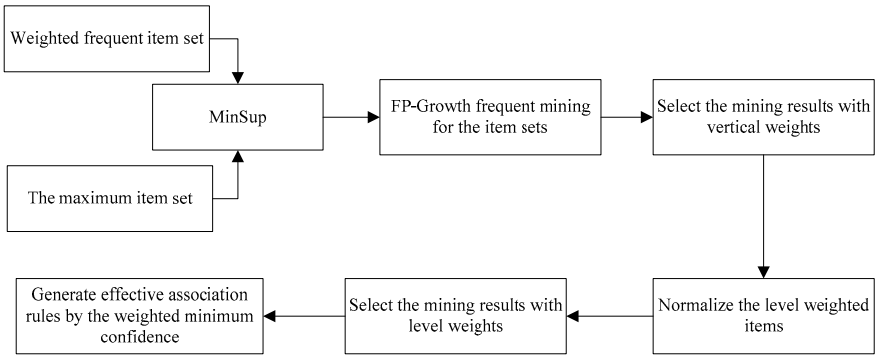

Figure 3. The implementation process of improved algorithm