Study of an improved Apriori algorithm for data mining of association rules

Xueting Zhang

School of Economic and Management, North China Electric Power University, Hebei, 071000, China.

1007822413@qq.com

Keywords: Improved Apriori algorithm, Association Rules,Data Mining.

Abstract. Data mining of association rules provides the technology for discovering the interesting association or correlation from mass of data. Apriori algorithm can find all the frequent items from transactional databases, and eliminate non-frequent items. But, the Apriori algorithm for data mining of association rules always produces a large number of candidate items, and scans the database repeatedly. Z-Apriori algorithm, the improved Apriori algorithmfor data mining of association rules, is introduced. A numerical example about a supermarket is given to show that Z-Apriori algorithm can dig the weighted frequent items easily and quickly. The association rules and items which are more interested by customers and more profitable can be found by Z-Apriori algorithm, and they are also traditionally supported highly.

Introduction

As the mature of database technology and universal applications of data, the amount of data accumulated in the human being growing rapidly at an exponential rate. Thus, there is no longer confined to a particular area of the database, but the vast expanse of information ocean. Faced with such a huge data from which to keep crude refined, discriminating technology: Data Mining was born. Data mining can be divided into classification, regression, association rules, and clustering of time series models. The association rules are the method that applied the most common, and also the focus in recent years. In the background of data mining technology development and prosperity, association rules has been booming, and is developing toward the more extensive and in-depth orientation.

Association rule mining is to find an interesting link between the associated items in a large database. Its applications extend from a narrow market basket analysis to the website design and optimization, business intelligence and pharmaceutical ingredients correlation analysis. Its theoretical research develop from initial frequent pattern mining to a closed pattern mining, the maximum pattern mining, association rule extended association rules and other types of data mining. In fact, association rules mining has to address 2 questions: one is to identify all frequent item sets from the transaction database; the other is to generate strong association rules from the frequent item sets. Apriori algorithm[1], the most classic algorithms, uses an iterative search method step by step, solves the first problem. Above all, find the 1-item frequent set L_1, then use it to generate 2-item candidate set C_2, and determinate C_2 to mine 2-item frequent set L_2, constantly cycling continues until you can not find more frequent k-items to set up; each excavation layer L_k need to scan the entire database to calculate the support value of a candidate set, eliminate the non-frequent item sets.

The need to produce a large number of candidate sets and repeatedly scan the database are the biggest performance bottlenecks of Apriori algorithm, which limit the more effective application of traditional mining model in a certain extent. In recent years, scholars continue to study a variety of improved algorithms. Equally to solve these problems, this paper proposes an improved algorithm based on Apriori algorithm mining association rules, meantime introduce the development of Apriori algorithm and predict its future applications.

Organization of the Text

Apriori Algorithm introduced. Apriori algorithm was first proposed by Agrawal, its initial motivation is for market basket analysis issues; its purpose is to discover association rules between different commodities and its essence is to use step by step search iteration, repeatedly scan the database, use k-item frequent set to generate (k+1)-item candidate set, identify frequent sets, and build association rules.

Algorithm idea. Get 1-item frequent set L_1 by traversing the database, if L_1 is non-empty, its resulting is 2-item candidate set C_2, and then scan the database, computing support value of C_2 in all candidate subset to select the candidate sets who meet the terms set minsup, denoted 2-frequent item sets; using these processes repeatly to give a new set of frequently until no frequent sets can get.

Apriori algorithm description[2] .set k-candidate item sets as C_k, k-frequent item sets as L_k. Input: transaction database, minsup;Output: result = all the frequent sets;(see Fig. 1)

```
1. L1= find_ frequent_1_itemsets(D);
2. For(k=2;Lk-1 !=null;k++){
3.   Ck=apriori_gen(Lk-1);}
4. For all t∈D{
5.   Ct=subset(Ck,t);
6. For all C∈Ct{
7.    C.sup=C.sup+1;}
8. }
9. Lk={C∈Ck|C.sup≥minsup};
10. }
11. L=UkLk;
```

Fig. 1 The pseudo code of Apriori algorithm

Apriori algorithm performance analysis. Apriori algorithm uses Apriori property to generate candidate sets, greatly reduced the size of the frequent sets, showing good performance, suitable to transaction database and sparse data sets association rules mining. but the problems are still obviously: Frequently scan the database and inefficiently identify the frequent itemsets.

Seen from the basic idea of the algorithm, Apriori algorithm's obvious flaw is the need to scan transaction database every time to generate and select frequent sets. Thus, reducing the number of database scanning is the same to reducing Apriori algorithm operations, as well as improving Apriori algorithm efficiency.

In practical application, the user will be more inclined to the rules that produced higher value or they own interested. However, Apriori algorithm considered the importance of all data the same. In order to study such problems, this paper introduces the concept of weights, adoptes the weighted association rule mining to improve the traditional Apriori algorithm and to solve the inconsistent data importance problem.

As we said before, apriori algorithm has to scan the database several times to calculate the support value of candidate sets. When the number of candidate sets were relatively long time algorithm will be quite huge overhead, but using matrix-vector operations can improve this problem.

Apriori algorithm present international situation[3]. After Apriori algorithm proposed, many researchers have done a lot of research on the problem of mining association rules, especially those based on Apriori algorithm and optimization. Such as Savasere, who designed the algorithm that based on partition; Park, who made hash-based algorithms; Mannila proposed the sampling method in mining etc. Researchers have recently extended a lot in the depth of study, including: spatial association rule mining, negative association rules mining, fuzzy association rules mining, sequential pattern mining and positive association rule mining. Over the course of the past 30 years, Apriori algorithm has emerged form intuitive and has been well applied in the actual data mining system.

Domestic situation. Currently, China is more mature in the application of Apriori algorithm, there have been many improvements and optimization algorithms. But compared with other countries, our research on the data mining are relatively late, only 10 years. Meantime, the main part of the strength concentrated in relatively strong institutions and research institutions, such as the Chinese Academy of Sciences, Tsinghua University, Xi'an Jiaotong University, Shanghai Jiao Tong University and the

National University of Defense Technology. Although the study of association rules has just started, scholars have made plain looking results. Since then, domestic mining association rules has been extensively explored and involved in many fields, mainly in seeking association rule frequent itemsets algorithm, theoretical research and practical application of association rules mining excavation. Among them, the research projects of great significance are: multi-strategy data mining systems of Chinese Academy of Science, and the developed AR Miner system in Fudan University, these current system have both been made on the practical application of certain achievement. It can be said that our country has being in the rise of Apriori algorithm association rule mining research and application field boom.

Apriori algorithm development and applications forecast. Now, with information and data constantly updated and increased, its potential association rules are also changing. Therefore, the study of related algorithms should equaly go deeper. Through research, author conclude the development and future application prediction of Apriori algorithm as the following points:

In the beginning, improve and optimizate the efficiency furthermore.

After that, try to use the visualization technology to research association rules mining process of interaction with the user function, and build the association rules according to the user's interest.

Then, with the extensive application of database technology development and database management systems, it has been used in electronic commerce. For example, since e-commerce sites accumulate more and more data, using Apriori algorithm in mining association rules can get valuable information from the data ocean.

Finally, it could analyse the data in science field, applying geoscience data analysis association rule mining can reveal linkages among some oceans, land, atmosphere to help scientists study the Earth system better.

Z-Apriori Algorithm Improvement Ideology. Section 1 gives the background of the Apriori algorithm problem which includes its inefficient shortcoming, multi-scanning database and large candidate sets. The main idea of the Z-Apriori algorithm is described in Section 2 while Section 3 describes a concrete realization of the process Z-Apriori algorithm. At present, some scholars have proposed average weighted association rules and maximum weighted association rules[4]. In this paper,we shall present an improved Association rule mining algorithm that based on maximum weighted and New-Apriori algorithm[5].

Matrix-vector operations: First, store data in the database as a matrix structure, namely the line matrix represents projects and column indicates the transaction. For example, if “ i ” projects occurs in the “ j ” transaction, then mark "1" in the corresponding position of matrix, otherwise marked "0". “ Di ” represents the vector by a matrix of rows to facilitate representation within the vector product.(see Fig. 2)

$$D = \begin{pmatrix} D1 \\ D2 \\ \cdots \\ Dm \end{pmatrix}$$

Fig. 2 Row vector matrix representation

Just use the internal vector product operation of D1 and D2, the result is 2-candidate sets' support value. When a candidate set of item is greater than 2, we only need to focus on the candidate vector element for each column, logical "and" calculation, then statistics "1" in the number as 2-candidate's traditional support value, then together with the weighted formula to get the maximum weighted association rules.

Z-Apriori Algorithm. Z-Apriori algorithm is similar to Apriori algorithm. The procedure is as follows three steps: firstly, set the project elements weight value; secondly use matrix-vector to find all frequent item sets whose weighted support no less than the user-specified minimum support minsup constraint and thirdly generate association rules which meet the set minimum confidence constraint.

(1) The weighted association rule mining[6]

a). Definitions 1: Given item set X={i1, i2, i3, ..., ik}, ij weights set to wj, which 0≤wj≤1, j={1,2,..., k}. item set X's weighted support value as follow:

wsup(X)=max{w1,w2, ...,wk}×support(X);

b). Definitions 2: weighted confidence is same with the traditional sense of confidence.

wconf(X→Y)=(support(X∪Y)) / (support(X))

c). Definitions 3: For projects set X, if wsup (X) ≥minsup, then X is weighted frequent sets.

d). Theorem 1: Suppose X is a weighted frequent set, any superset of X Y, if wsup (Y) ≥wsup (X), then Y is a weighted frequent sets.

e). Theorem 2: Assuming that X, Y are not weighted frequent sets, X, Y definitely not weighted frequent sets.

(2) matrix-vector computing support level[7]

The transaction data in the database stored as the matrix structure(see Fig. 3): where the transaction represented by T, item collection by I:

$$D = \begin{array}{c} \\ I1 \\ I2 \\ \dots \\ Im \end{array} \begin{array}{c} \begin{array}{cccc} T1 & T2 & & Tn \end{array} \\ \begin{pmatrix} d11 & d12 & \dots & d1n \\ d21 & d22 & \dots & d2n \\ \dots & \dots & \dots & \dots \\ dm1 & dm2 & \dots & dmn \end{pmatrix} \end{array} \begin{array}{c} I1 \\ I2 \\ \dots \\ Im \end{array}$$

$$dij = \begin{cases} 0, & Ii \neq Tj \\ 1, & Ii = Tj \end{cases}$$

Fig. 3 Matrix data structure of the transaction database

After convert the database into a matrix structure, the total number of "1" in each row is the traditional support value of 1-itemset, and so will be able to find the 1-item frequent sets quickly.

Matrix row vector “Di” inner vector product operation, “Di” and “Dj” inner product:

[Di, Dj] = di1 * dj1 + di2 * dj2 + ... + din * djn;

These can be the traditional support level of 2-item candidate sets.And when items are greater than 2,use logical "and" calculation.

(3) Z-Apriori algorithm description

Algorithms: Z-Apriori algorithm;

Input: A transaction database D, the minimum support value minsup;

Output: all frequent sets of L (see Fig. 4).

```
 1. L1=find_ frequent_1_itemsets(D);
 2. For(k=2;Lk-1!=null;k++){
 3. Ck=apriori_gen(Lk-1,min_sup);
 4. For each c∈Ck{
 5.    C.count=Σ^n_{j=1}d_{1j}∧d_{2j}∧d_{5j}∧d_{7j;}}
 6. Lk={c∈Ck|c.countmin_sup}
 7. L=UkLk;
 8. For each t∈Lk{
 9. For each item∈t
10. Item.count+=Ik.Wk（weight）×sup(X);}
11. For each item.count<k
12. Delete t(item∈t)from Lk;
13. Return L;
14. Procedure apriori_gen(Lk-1:frequent(k-1)-itemsets)
15. For each  L1∈Lk-1
16. For each  L2∈Lk-1
17.  If((I1[1]=I2[1])&&(I1[2]=I2[2])&& and
&& (I1[k-2]=I2[k-2])&&(I1[k-1]<I2[k-1]))  then{
18.      c ={I1[1],I1[2] and I1[k-1],I2[k-1]};
19. Add c to Ck'} ;
20. For each Ii∈Lk-1
21. For each c∈Ck'
22. If ( Ii is the subset of Ck')  then
23. C.number ++;
24. Ck={c∈Ck'|c.number=k};
25. Return Ck;
```

Fig. 4 The pseudo code of Z-Apriori algorithm

Z-Apriori algorithm analysis. The following example illustrates the use of a concrete realization of the process Z-Apriori algorithm. Example: A shopping mall transaction database shown in Table 1, the data of corresponding commodities are shown in Table 2.

Table 1. A mall transaction database

TID	Items
1	I1, I2, I5
2	I2, I4
3	I2, I3
4	I1, I2, I4
5	I1, I3
6	I2, I3
7	I1, I3, I6
8	I1, I2, I3,I5
9	I1, I2, I3

Table 2. Product weight for each item in the database

Items	Weights
I1	0.6
I2	0.7
I3	0.5
I4	1
I5	1
I6	0.7

Assume minsup = 0.2, then calculate weighted support value of each " I " project. The calculation of weighted frequent set is as follows:

a). matrix vector calculate 1-term support value of a candidate set(see Fig. 5 and Fig. 6)

$$D = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Fig. 5 The example of Matrix

$$D = \begin{pmatrix} D1 \\ D2 \\ D3 \\ D4 \\ D5 \\ D6 \end{pmatrix} = \begin{pmatrix} D1=6 \\ D2=7 \\ D3=6 \\ D4=2 \\ D5=2 \\ D6=1 \end{pmatrix} = \begin{pmatrix} D1=6*0.6=3.6 \\ D2=7*0.7=4.9 \\ D3=6*0.5=3 \\ D4=2*1=2 \\ D5=2*1=2 \\ D6=1*0.9=0.9 \end{pmatrix}$$

Fig. 6 The calculation of 1-item candidate set support value

1-item frequent set: L1 = {I1, I2, I3, I4, I5};

2-item candidate sets: C2 = {I1, I2}, {I1, I3}, {I1, I5}, {I2, I3}, {I2, I4};

b). matrix-vector calculate 2-item support value of candidate set(see Fig. 7)

D1*D2 = 4	4*0.7 = 2.8
D1*D3 = 4	4*0.6 = 2.4
D1*D5 = 2	2*1 = 2
D2*D3 = 4	4*0.7 = 2.8
D2*D4 = 2	2*1 = 2

Fig. 7 The calculation of 2-item candidate set support value

2-item frequent sets: {I1, I2}, {I1, I3}, {I1, I5}, {I2,I3},{I2, I4};

3-item candidate sets: {I1, I2, I3}, {I1, I2, I5}, {I1,I3,I4}, {I1, I3, I5}, {I1, I4, I5};

c). For each candidate set of row vectors corresponding“ I1, I2, I3, I4, I5 ”, logical "and" computing, and the statistics the number of "1", with the weighted association rules are weighted support.(see Fig. 8)

$$D = \begin{array}{c} \quad T1\ T2\ T3\ T4\ T5\ T6\ T7\ T8\ T9 \\ \begin{pmatrix} I1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\ I2 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 \\ I3 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ I4 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ I5 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ I6 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

{I1, I2, I3} : 0 0 0 0 0 0 0 1 1 :2*0.7=1.4
{I1, I2, I5} : 1 0 0 0 0 0 0 1 0 :2*1=2
{I1, I3, I4} : 0 0 0 0 0 0 0 0 0 :0*1=0
{I1, I3, I5} : 0 0 0 0 0 0 0 1 0 :1*1=1
{I1, I4, I5} : 0 0 0 0 0 0 0 0 0 :0*1=0

Fig. 8 The calculation of 3-item candidate set support level

Z-Apriori algorithm performance analysis. Analysis the results of the algorithm shows that: the algorithm can easily and quickly dig out a weighted frequent set containing at least one project of larger weight, as well as the users more interested or higher profits project association rules, with its traditional support value also high.

Z-Apriori algorithm is the effective weighted frequent item sets mining Apriori improved algorithm, presented with the actual situation. It uses the ideology that derived from Apriori algorithm, and improves the connecting and the pruning step to quickly and efficiently mine the

weighted frequent item sets, making the mining results more valuable. Compared with the classical Apriori algorithm, its superiority is mainly reflected in the following three areas:

a). The discovery of maximum weighted frequent set association rules, better reflect the algorithm's attitude towards the degree of user interest considerations.

b). To generate a k-item candidate set from classic Apriori algorithm, one need to scan Lk-1 to confirm the conditions that (k-1)-item frequent set, which used to generate k-item candidate set, can be connected. If they meet the connection, just generate a set of k-item candidate sets. Whereas, if connection is not satisfied, then can not be connected. The above shows that, Z-Apriori algorithm prune effectively reduces the scanning number of Lk-1.

c). Classic Apriori algorithm, when pruning the K-item candidate sets to generate frequent itemsets, should calculate the support value for each K-item candidate set. In other words, for every K-item candidate set, one need to scan a database for the entire transaction. Nevertheless, Z-Apriori algorithm only need to operate row I1, I2.

To sum up, comparing with the classical Apriori algorithm: Z-Apriori algorithm is more focused on the user's interest degree in mining, the efficiency and speed of mining frequent sets and rules has also been improved, and the more the number of transactions, it appears more obvious.

Summary

Section 4 provides a summary and a discussion of some extensions of the paper. Association rule mining is an important data mining techniques, in which Apriori algorithm is the most classic, the other association rule mining algorithm proposed mostly based on Apriori algorithm. In this paper, we introduced a brief introduction of Apriori algorithm and performance analysis, and for two issues of its existence, we propose an improved Z-Apriori algorithm.

However, the improved algorithm proposed by the author only add the user-interest association rules by weighting methods and reducing the number of scanning database by matrix-vector calculations. Indeed, in the era of ever-expanding amount of data, the algorithm’s execution speed, scalability, and the intelligibility of the output characteristics are equally significant, so we should further developed a association rule mining algorithm which is very good in all aspects characteristic.

Acknowledgment

This paper is supported by “ the Humanities and Social Sciences project of the Education Ministry No.14YJC630187), and the national Higher-education Institution General Research and Development Project of North China Electric Power University (No.2014QN43)”.

References

[1] Agrawal. R, and Srikant. R, In Proc of the 20th Int’l Conf on Very Large Database, Fast algorithms for mining association [C]. Santiago. (1994) 487-499.

[2] Hongyin Zhao, Yuecai Cai and Xianjie Li, Summary of Apriori Algorithm Mining Association Rules, Sichuan Institute of Technology (Natural Science) . [J]. (2011).

[3] Zhengchan Rao. and Nianbai Fan, A review of associative rule mining Apriori algorithm, Computer Era, [J]. (2012)

[4] Cai.CH, FuAda. WC and Cheng. CH, Proc of the Int’l Database Engineering and Applications Symposium, Mining association rules with weighted items[C]. Cardiff. (1998) 68-77.

[5] Zhang Zhijun, Fang Ying. and Xu Wentao, Based on average weighted association rules Apriori algorithm for mining, Computer Engineering and Applications [J]. (2003)

[6] Wang Yan. and Wang Hongxia, Based on weighted association rule mining algorithm Apriori algorithm, Zhengzhou University of Light Industry (Natural Science) [J]. (2007)

[7] Wang Wei, Research and Improvement Association Rules of Apriori algorithm, Ocean University of China (Wanfang Database) [J]. (2012)

[8] Hongyan Liu, Jian Chen and Guoqing Chen, Review the Data Mining Classification Algorithm, Tsinghua University (Natural Science) [J]. (2002)

[9] Lu Lu and Caichang Yu, Association Rules Apriori Algorithm Improvement Research, Yangtze University (Natural Science) [J]. (2009)

[10] Srikant. R, Vu. Q and Agrawal. R, Mining Association Rules with Item Constraints, San Jose: In Proc of the Third Int'l Conf in knowledge Discovery in Databases and Data Mining. (1997) 67-73.