Advances in Intelligent Systems Research, volume 146
2nd International Conference on Artificial Intelligence: Technologies and Applications (ICAITA 2018)

Analysis and Improvement of Classification Based on Multiple Association Rules

Jing Lin
School of Mathematics and Computer Science, Wuhan Polytechnic University, Wuhan, China

***Abstract*—An excellent algorithm is critical to classify data. The algorithm of Classification based on Multiple Association Rules combines association rules with classification. The limitations of traditional CMAR algorithm are analyzed and an improved algorithm is presented. In the new algorithm, each attribute of sample data has a weight according to its importance. The weight of every attribute revises the basic support threshold. So the unique support threshold of each attribute is calculated. The support of each attribute value is compared with the unique support threshold of the attribute. If the support of an attribute value is greater than the unique support threshold, a set NF-List will contain this attribute value. A new FP tree is built on the basis of the set NF-List. The new FP tree is traversed to create classification association rules by Apriori. Experiments show that the improved CMAR algorithm can generate more association rules about important attributes, but not the ordinary rules in the traditional CMAR algorithm.**

Keywords—classification based on Multiple Association Rules; support threshold; FP tree; classification association rules

I. INTRODUCTION

Data mining is the core of knowledge discovery. It defined as extracting hidden information and knowledge from a large number of noisy ambiguous, incomplete, random data. Association rules and data classification are both import methods in data mining. Association rules mining can discover the relation between items in database and identify potential patterns of behavior. Classification is to construct a classification model based on the input data and use the model to map other data items to some categories. How to use association rules to implement classification is a new research field in recent years. Association rules mining and classification can be integrated. New algorithms appear. Now these methods include CMAR(Classification based on Multiple Association Rules), CBA(Classification Based on Associations), etc. The CMAR algorithm is based on the FP-growth method to produce classification association rules. It enhances the efficiency of mining frequent item sets. This algorithm is well suited for handling large amounts of data. But it has some disadvantages, such as long rules prior and requiring large amount of memory. Particularly, the importance of each attribute is the same in traditional CMAR. But in application, users often pay different attention to each attribute. In this paper, we consider that every attribute has different importance and improve the CMAR algorithm. It will generate more rules related to attributes that users care about. The classification is more accurate.

II. IDEA OF CMAR ALGORITHM

A. Algorithm Step

The CMAR algorithm describes those potential classification rules in transactions. The rules are based on frequent patterns. The general form of a rule is "Condition→Class identifier". The condition is a frequent item set.

The CMAR usually consists of two stages.

First, a classification association rule set is created from sample data. The steps are as follows. Firstly, all sample data are scanned. Supports of some attribute values are larger than the support threshold. These attribute values are chosen to form a set F. In the set F, attribute values are listed in descending order by support. Then FP tree will be built on the set F. In FP tree, each leaf node is appended to a classification value. By traversing the FP tree from leaf nodes to the root, we can find frequent patterns and output the associated classification values. Finally, these classification association rules are stored by the CR tree. Rules often have to be pruned. In the CR tree, only rules that exceed the threshold are preserved. These rules are the final result of mining.

Then we can classify new data according to the above rules. Those rules that match the new data are chosen. If these selected rules have the same class symbol, the new data belongs to this class. Otherwise, the rules of maximum confidence are chosen to match new data.

B. Disadvantages of CMAR

In CMAR algorithm, each attribute must be compared with the support threshold. But the differences between each attribute are not taken into account. In other words, each attribute is considered equally important. It is not true in practical applications. Take the bank business as an example. When a bank makes a judgment on a customer's credit, the judgment criteria may be the customer's characteristics, such as age, sex, job, income, etc. Obviously, the importance of every characteristic is different in evaluating a customer's credit. The job and income have a greater impact on the credit. In sample data, if these more important attributes have many values, each attribute value appears less frequently. If the algorithm is based on the same importance of all attributes, the rules probably are ordinary because the number of important attribute values is too small in the sample data. It is difficult to get rules related to important attributes. So we improve the traditional CMAR method in this paper. According to the importance, each

attribute have a weight. The support threshold is revised with the weight of attribute. In this way, the final classification rules can reflect the difference in the importance of attributes.

III. IMPROVEMENT OF CMAR ALGORITHM

A. Concepts and Definitions

The attribute set of sample data is $(A_1,A_2,\ldots A_n)$. N is the number of attributes. A sample data set is described as T. In T, each record has a class symbol. C is a set of class. $C=\{c_1, c_2,\ldots,c_n\}$.

Definition 1 $P=\{a_{i1},\ldots,a_{ik}\}((1<= j<=k) \&\&(a_{ij}\in A_{ij}))$

P is called pattern. It is composed of attribute values. In pattern P, each attribute value appears only once.

Definition 2 On the basis of T, a classification association rule is derived from pattern to class. It is expressed as: $R: P\rightarrow c$

Definition 3 $R: P\rightarrow c$

Confidence(R) = RsupCount/PsupCount * 100%

Support(R) = RsupCount

Rsupcount: the number of rules that are consistent with the pattern P and the class symbol is c.

Psupcount: the number of rules that are consistent with the pattern P.

B. Algorithm Steps

In this paper, we mainly improve the strategy of generating rules in the algorithm. In the sample data, each attribute has a weight. The weight represents the importance of the attribute. A basic support threshold is given at the beginning of the algorithm. For each attribute, the basic support threshold is modified by the weight. And each attribute has a unique support threshold. When judging whether an attribute value belongs to a frequent item set, we compare the support of this attribute value to the support threshold of this attribute instead of the basic support threshold. In this way, frequent item sets will contain more important attribute values. With this improvement, the steps for the generation of rules are as follows.

T is a data set.

$T=(A_1,A_2,\ldots A_n)$

$R(Rule): P \rightarrow c$

Support(R) is the support threshold. Confidence(R) is the confidence threshold. The unique support threshold of each attribute needs to be calculated. The formula is:

$$ATSupport(R)_{Ai}=Support(R)*(1+1/n-W_{Ai}) \quad (1)$$

$ATSupport(R)_{Ai}$: support threshold of attribute A_i

Support(R): basic support threshold

W_{Ai}: weight of attribute Ai

n: number of attributes

The process of rule generation is as follows.

Firstly, all the data in T are scanned. The support of each attribute value is compared with $ATSupport(R)_{Ai}$. If the support of an attribute value is greater than $ATSupport(R)_{Ai}$, this attribute value is chosen and added to a set NF-List. In the set, these attribute values are sorted by the support in descending order.

Now we scanned the T again and create a new FP tree. The new FP tree is a prefix tree about the set NF-list. In the set T, each record is read sequentially. For each record, operations are as follows. According to the order of attribute values in NF-List, we judge whether each attribute value is in the current record. If an attribute value appears in the record, it is inserted into the new FP tree. When inserting the last attribute value, the class symbol and class count of this record are appended to the node.

Finally, classification association rules come from the above new FP-tree. The new FP-tree is usually traversed from bottom to top to produce rules. From the last attribute value to the first one in NF-List, each attribute value is projected in the new FP tree to produce a set of branches that contain this attribute value. In each set of branches, frequent patterns are searched. After the processing about branches of an attribute value is finished, all nodes of this attribute value and their class symbol are merged into their parent nodes. Repeat the process until all the rules are found.

The following is an example to illustrate the improved algorithm.

The sample data set T is shown in Table I.

TABLE I. THE SAMPLE DATA

Tid	A	B	C	D	Class
1	a_1	b_1	c_1	d_1	A
2	a_1	b_2	c_1	d_2	B
3	a_2	b_2	c_2	d_3	A
4	a_1	b_2	c_3	d_3	A
5	a_1	b_2	c_1	d_3	C
6	a_2	b_2	c_1	d_3	B
7	a_1	b_3	c_3	d_2	C
8	a_2	b_1	c_3	d_2	A
9	a_2	b_3	c_2	d_1	A
10	a_1	b_1	c_1	d_2	C

There are four attributes in sample data. Those are A,B,C and D. The weights are 0.125, 0.25, 0.375 and 0.25. The support threshold is four. The confidence threshold is 30%.

According to Formula 1, the support thresholds of these four attributes are calculated as follows.

$ATSupport(R)_A=4*(1+1/4-0.125)=4.5\approx5$

$ATSupport(R)_B=4*(1+1/4-0.25)=4$

$ATSupport(R)_C=4*(1+1/4-0.375)=3.5\approx4$

$ATSupport(R)_D=4*(1+1/4-0.25)=4$

The values of attribute A are a_1 and a_2. The support of a_1 is six. The support of a_2 is four. Because the support of a_1 is greater than $ATSupport(R)_A$, a_1 belongs to the set NF-List. Similarly, all the other attribute values that meet the requirements are added to the set NF-List. They are sorted by support in descending order in the NF-List.

NF-List={a_1,b_2,c_1,d_2,d_3}

According to the order of attribute values in the NF-List, records in the set T are sequentially scanned to build the new FP tree. For the first record, attribute values a_1 and c_1 are inserted in the tree as nodes in turn. The class symbol A and the class count A=1 are attached to the node c_1. In the second record, effective attribute values are a_1,b_2,c_1,d_2. They share prefix a_1 with (a_1,c_1). So in the new branch starting from a_1, the node of b_2,c_1,d_2 are inserted. By this way, all branches of the tree are generated. The final tree is shown in Figure I.

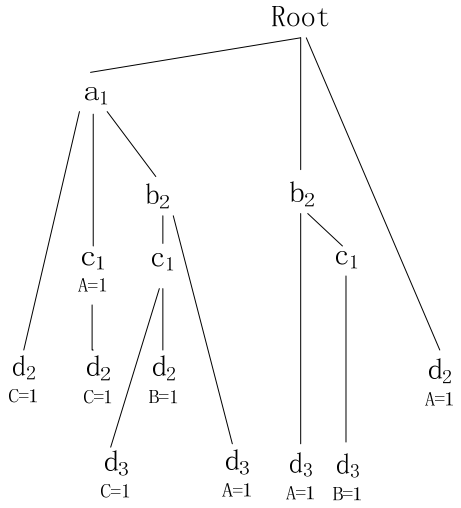

FIGURE I. THE NEW FP TREE

Next, frequent patterns will be derived from the new FP tree. From the last attribute value d_3 to the first attribute value a_1, the subset of data about each attribute is obtained in turn. Each subset must contain the current attribute value, but not attribute values that have been scanned. In this example, there are five subset:1) contains d_3; 2)contains d_2 but not d_3;3)contains c_1 but not d_2 and d_3;4)contains b_2 but not c_1,d_2 and d_3;5)only contains a_1.

The first is a subset that contains d_3. It is a collection of all branches about d_3.

(a_1,b_2,c_1,d_3):C (a_1,b_2,d_3):A (b_2,d_3):A (b_2,c_1,d_3):B

The Apriori algorithm is applied in this subset. The frequent modes are (b_2,d_3). The rules about the frequent modes are as follows.

(b_2,d_3) → A, (b_2,d_3) →B, (b_2,d_3) →C

The confidence of the first rule is 50%.It is greater than the confidence threshold. The confidence of the other two rules is smaller than the threshold. So in this subset, only the rule (b_2,d_3) →A is classification association rule.

After getting all classification association rules about d_3, the new FP tree is reduced in Figure II.

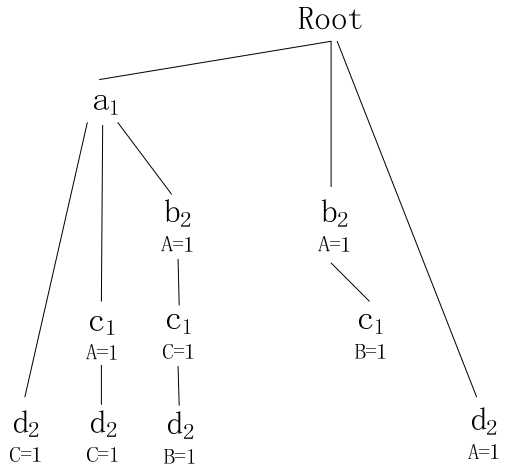

FIGURE II. THE REDUCED FP TREE

On the basis of the reduced tree, the following is the subset about attribute value d2.

(a1,d2) (a1,c1,d2) (a1,b2,c1,d2) (d2)

In this subset, there is only one association rule. It is (d2) →C.

Repeating the process for the subsets about c1,b2,a1, no effective rules are produced. So from the sample data T, the final classification association rules are only the following two.

(b2,d3) →A, (d2) →C

C. Algorithm Analysis

The improved CMAR algorithm introduces the weight of each attribute. The weight can express the importance of attribute. With the weight, different attributes are no longer judged by the same support threshold, but by different unique support threshold. In traditional CMAR algorithm, the final rules are often too common. But new algorithm leads to more rules relating to important attributes because the unique support threshold of an important attribute is often lower. However, there are still some problems in the improved CMAR algorithm.

Firstly, the characteristics of the sample data may affect the accuracy of classification. In the following cases, the accuracy of classification will decrease. 1) The number of attributes is very small, but the number of attribute value is large. 2) The number of classification is large. 3) The number of sample data records is large. These three points should be avoided as much as possible.

Next, because the weight of each attribute plays an important role in the algorithm, how to decide the weight of each attribute is very important. In practice, the value of weight is often determined by characteristics of business and users' experience. It may be better if the weights of attribute are calculated on the basis of some objective. It is the focus of future research.

Finally, the support threshold and the confidence threshold determine the classification result directly. Too large or too small thresholds may lead to unreasonable rules. The appropriate combination of support threshold and confidence threshold is very important for getting the optimal association rules. But there is no effective way to implement it.

IV. CONCLUSION

In this paper, we analyze and improve the CMAR algorithm. The main point of improvement is to give each attribute a weight to describe its importance. With the weight, the basic support threshold is revised to the unique support threshold of each attribute. When generating frequent patterns, the support of an attribute value is compared with the unique support threshold rather than the basic support threshold. After this improvement, the final set of rules will contain more rules about important attributes. But there are disadvantages in the new algorithm. Especially, the way of determining all weights will need further research in the future.

REFERENCES

[1] Jiawei Han, Micheline Kamber, Data Mining:Concepts and Techniques, Second Edition,China Machine Press,2007

[2] Mehmed Kantardzic, Data Mining Consepts, Models, Methods, and Algorithms, Tsinghua University Press,2003

[3] Pang-Ning Tan,Michael Steinbach,Vipin Kumar, Introduction to Data Mining,Posts & Telecom Press,2006

[4] DU Yongsheng, A Algorithm for Data Mining Associative Classification Rules Based on Hierarchical Frequent Pattern Tree, Journal of Jining University, Vo. 32, No.6, 2011

[5] TONG Yu jun, LI Yu, CHEN Wen shi, LIU Hong shen, Improved Class Association Rule Mining Algorithm, Journal of Liaoning University of Technology(Natural Science Edition), Vol.31, No.5,2011

[6] ZHAO Chuan shen, HE Shun gang, YANG Ji hong, CHEN Li xia, Data Stream Classification Algorithm Based on Multiple Class-association Rules, Com puter Engineering, Vol.36, No.9,2010