The Research of Data Mining for Quantitative Association Rules and Algorithm for Numerical Attribute

Du Junping
School of Computer Science
Beijing University of Posts and Telecommunications, Beijing, 100876

Wang Rui jie
School of Information Engineering
University of Science and Technology, Beijing, 100083

Abstract

In this paper we proposed the concept and types of the quantitative association rules. We proposed the data mining for the quantitative association rules in detail. We also researched the concept for discretization of numerical attribute, and the strategy of discretization result. We established the better algorithm for discretization of numerical attribute.

1 Introduction

Association rules reflect dependency and association of one event and others. Data association in database is representation of things' relation in real world. As a sort of structurized data organization's format, database makes use of possibility of adhered data model to describe association of data (e.g. primary key and foreign key). At the same time, association among data is very complicate, which not only means association adhered in data model mentioned above, but also means a majority of hidden association. The aim of Data mining for Association rules is finding hidden association information. Association can be divided into simple association,time series association, causal association, quantitative association and etc.

These associations are not always forecasted, but gained by analyzing data association in database. As a result, it provides business decision making with new value.

Data mining for association rules is the most common method for finding association knowledge, of which Apriori and its amelioration. For finding significant association rules, two threshold vales are needed to be given: minimum support and minimum confidence. Mined association rules must satisfy minimum support stated by users, and it denotes minimum association degree that a group of associated items need to satisfy. Mined association rules need to satisfy minimum confidence stated by users, and it reflects minimum reliability of one association rule. In this sense, the aim of data mining system is mining association rules which satisfy minimum support and minimum confidence from source database. Research and application for association rules are the most active and embedded branches of data mining. Many theories and algorithm of data miming for association rules are raised.

Discretization of Numerical Attribute is the key issue of data miming for association rules, and actually it partitions attribute field into intervals. Partition methods play a significant role in quality of data mining for the quantitative association rules. Min-confidence: if numbers of

Supported by Beijing Natural Science Foundation of China (4042012), Key Project of Beijing Educational Committee (KZ200510011009) and Project of National Science Foundation of China (60442003).

interval are exiguous, the number of group which support interval will increase, and the support degree of strong item set of inclusion interval will go up. At the same time, if support of strong item set's subset remains the same, the rule confidence of right end inclusion subset will decline. If confidence threshold vale cannot be reached, information will be lost.

Min-support: if numbers of interval partitioned are overabundant, support of interval will decline, and entire strong item set will not be created effectively.

2 Quantitative Association Rules

A. Concept of Quantitative Association Rules

Quantitative association rules are association rules including classify attribute (Boolean attribute can be deemed to special classify attribute) and quantitative attribute. Generally, quantitative association rules are multidimensional association rules issues, in which the key issue is discretization of Numerical Attribute to satisfy a certain data mining criterion. Quantitative association rules also need to satisfy $X \cap Y = \varnothing$ and constraint condition of support and confidence. In a general way, Quantitative association rules issues are different from Boolean association rules in knowledge representation, Boolean association rules seeks for frequent item sets, moreover, quantitative association rules seeks for frequent predication set. Some characters of Quantitative association rules are listed as below.

Subset of frequent item set to a certainty is frequent. This character is consistent with Boolean association rules.

Suppose X^ is generalization of X, X is specialization of X^, if X^ is frequent, then X is frequent, and support (X^) >support (X),

It is impossible that two items in frequent item set have same attribute. These characters are easily understood from analogy in previous Boolean association rules (its demonstration is ignored); whereas the time that k+1 item set are created by frequent k item set reduces greatly. When k+1 item set are created by frequent, algorithm only needs to calculate different attributes, and it is impossible that the disjoint intervals with same attribute are frequent.

B. Categories of quantitative association rules

Numerous intervals are created after Discretization of Numerical Attribute, and then these intervals are mapped to Boolean attribute directly and mined by Boolean association rules. It is a common method for data mining for the quantitative association rules, and it is characterized by complicated format of rules. In fact, users are usually interested in rules of a certain format, i.e. Reference[3] of data mining for the quantitative association rules is mostly based upon a certain rules model. So we can classify the quantitative association rules mining to 3 types according to the rules formation.

(1) Rules of mining similar to "numerical value attribute ∩ classification attribute => numerical value attribute ∩ classification attribute", e.g. sex=female∩age∈[20,30]=> wage∈[$5,$10]. This sort of rules is complicate, and they are normal quantitative association rules.

(2) Rules of mining similar to "numerical value attribute ∩ classification => classification attribute", e.g. oppucation=bussiniessman∩age∈[35,50]=>goa board=Yes. This sort of rules' left end generally denotes one subset of database. Any attributes in database, and any value combination can be used as left end of rules, the right end denotes one predefined category, and it is quite similar to classification rules. The typical classification methods are decision tree、genetic algorithms、neural network、Bayesian classification and etc. The efficiency of resolving classification problems by association rules is very high, especially to large data set, classical classification algorithm is difficult to gain mining result.

(3) Rules of mining similar to “classification attribute=> numerical value attribute ∩ classification attribute”, e.g. sex=females =>salary∈[$1000,$2000]. It is opposite to the second sort of rules, yet to some problems, we will not expatiate anymore here.

C. Common steps of quantitative association rules

Mining quantitative association rules by Boolean association rules algorithm and its theory, such as support、 confidence、frequent item set and other concepts are most effective approach of resolving quantitative association problems.

One typical method is ameliorating Aprior algorithm to adapt to data mining for quantitative association rules, and mainly divided into five steps that shown as Fig.1.

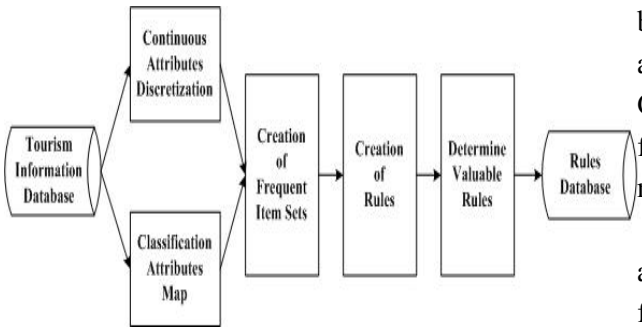

Fig.1 Common steps of quantitative association rules

(1) Pointed to each numerical value attribute, Reference [4] of proper discretization algorithm is selected and the number of intervals is determined. The difficulty of this step is selecting proper discretization algorithm. At first, discretization algorithm has not unified criterion, we should select one or several sort of algorithm according to distribution characters of data. For instance, this text will discuss that equi-depth partitioning method is not ideal for highly skewed data discretization, but it works well on partition of equally distributed data. In the next place, numbers of intervals partitioned, i.e. granularity of partition, too many or too little will lead to information lost.

(2) The value is mapped to continuous integers to classification attribute. To those numerical value attributes which need not partitioning intervals, such as some numerical value attributes with little value, and the value can be mapped to continuous integers when they are sorted by size. If numerical value attributes are discretized to intervals, intervals will be mapped to continuous integers according to discretized order of intervals. The operation to these continuous integral value is equal to operation to data set will be mined, and these continuous integers are transparent to algorithm. All classification attributes and continuous attributes are mapped to Boolean attribute, and all of such attributes compose item sets.

(3) Creation of frequent item sets. This step is the same to the steps of frequent item set created by Aprior algorithm. Aprior amelioration algorithm is also applicable here. Based on Character 3, new item sets will not be created from adjacent sections, in this way they can be merged by definite merging criterion.

(4) Application for frequent set creation association rules. If both ABCD and AB are frequent sets, the rule AB=>CD is true or not needs to be determined by result of calculating conf=supp (ABCD)/supp (AB) exceeds minimum confidence or not.

(5) The valuable association rules are determined as output. For helping users to mine valuable rules and decline redundancy of rules, the association rules based on interest measure has been established in Boolean data mining for association rules, and extensively applied in data mining for the quantitative association rules.

3 Concept and strategy for discretization

Algorithm for discretization of numerical attribute will face the same problems during resolving classification problems, however, problems of data mining for association rules is different from classification. From the machine

learning point of view, data mining for association rules is unsupervised learning category, but data mining for classification model is supervised category. In other words, discretization algorithm that suits classification does not necessarily suit data mining for association rules; hence it is necessary to research new discretization methods combined with characters of data mining for association rules. Combined with characters of data mining for association rules, methods for numerical value discretization mainly are equi-width partition, equi-depth partition and distance-based partition.

A. Equi-depth Partition Method

During numerical value discretization, equi-depth partition is the most common discretization method, commonly adequate for data set with low association among attributes. For the data set with tight association, application of equi-depth partition is difficult to mine ideal result. Equi-depth partition method tends to partition adjacent value with commonness and high support into different intervals. When data distribution reach peak value near to a certain point, the mechanical method like equi-depth partition cannot reflect characters of data itself, hence it does not work well on highly skewed data discretization. For example, during tourist data mining, all kinds of discretization methods suit for processing scenic spot attribute and other data with low association, such as tourists' age、numbers of scenic spots.

R is one relation, A is one attribute of it, T is one tuple, C is one condition of R, T(A) denotes A attribute value for tuple T.

Bi=[Xi,Yi] of which i=1，2，... ，m, Xi≤Yi≤ Yi+1；

If A is one numerical value attribute, and its interval is [Xl,Ym]. Interval of A is partitioned into a series of disjoint fields, viz.

Bi=[Xi,Yi] of which i=1，2，... ，m,Xi≤Yi≤ Yi+1；

Bi is entitled "one bucket of A", and number of tuples in T∈R and T(A)∈Bi is entitled "size of bucket", noted as u. if size of tow buckets are same, and they are entitled "equi-depth bung".

Algorithm shows as below:

Input: numerical attribute A, number of bucket n

Output: discretized interval

(1) Numerical attribute in order;

(2) Scan database, Stat the database item number N;

(3) Get the depth of bucket h=N/h;

(4) Scan A in order one by one. According to Definition 2, get number i and i+h in sequence, the make the output of discretized interval [li,vi+h].

B. Equi-width Partition Method

The equi-width partition method is the simplest discretization method which is used to distributing equably data. Because this algorithm need only one time to scan the database, it has a high efficiency. The equi-width partition method simply partition from mathematics' point. It doesn't consider the characteristic of the data distributing. As this method is more directly and adapt to prophase dispose of data attribute, it always unites the clustering method and get good discretization result. For example, in the tourism data mining, it used to discretize tourists' age and income, etc.

Algorithm shows as below:

Input: numerical attribute A, number of bucket n

Output: discretized interval

(1) Scan database, Get the max(A) and min(A);

(2) Try the width w of interval, w =(max(A)-min(A))/n;

(3) Make the output （li=min(A)+(i -1)* w, vi=min(A)+i* w ） of discretized interval[l1,v1],(l 2,v2),…,[l n,vn].

C. Classification Partition Method

Both of equi-width partition and equi-depth partition methods are considering neither the

data distribution, nor the experts' proposal to the attribute discretization. They only partition the data attribute according to the geometry and mathematics. The classification partition method can take attribute discretization according to data distribution feature based on the field experts' proposal. To solve the quantitative association rules problem, we can use the classification partition method to partition each attribute item to proper category which the equi-depth partition method can not solve. This can show data distribution status. The aim of classification is to partition similar data to same category.

For example, in the tourism information data mining technology, field expert hold up the classification model. The process of quantitive data discretization is partitioning one group data to different class according to the classification model. The result of classification should make the data distance in same types much shorter and in different types much longer.

The process of data classification partition is: Firstly field expert take the characteristic value A[1…k] of the partitioned k intervals, then scan every data in sequence, classify data to the category which is nearest to data number, update characteristic value of this category.

Algorithm shows as below:

Input: Numerical attribute A, Interval characteristic value A[1…k]

Output: Discretized k intervals

do

(1) Read one data, Calculate the distance between it and every characteristic value;

(2) Partition data to category which is nearest to data number;

(3) update characteristic value to average value of its interval;

until all data partitions finished.

5 Conclusion

In this paper we proposed the algorithm for discretization of numerical attribute in quantitative association rules mining. Through the analysis of the above 3 discretization algorithms, we defined divergent density as evaluation criterion of discretization effect.

Definition: If discrete category[a,b],c is the number belong to the discrete category, then$\rho=c/|b-a|\times 100\%$ is the density of this discrete category. We call the density average value of all discrete category which is $\varpi = (\sum_{i=1}^{n} \rho_i)\frac{1}{n}$ as discrete density. If this discrete density is larger, the effect of partition is better.

By this method, we calculate the discrete densities of above 3 discretization algorithms: The discrete density of equi-width partition is 10%. The discrete density of equi-depth partition is 40.84%. The discrete density of classification partition is 100%. To this data group, the effect of classification partition is the best. But because the cost of classification partition algorithm is much higher, we should take different discretization methods according to data feature.

References

[1] R .A. grawal, T. Im ielinski, and A . SwamiMining association rules between sets of ite ms in large data bases，SIGMOD'93，Washington，1993：20 7－216.

[2] Jong Soo Park, Ming-Syan Chen, and Philip S.Yu，An efective hash based algorithm for mining association rules，Kdd-94.

[3] Hilderman Robert, Hamilton Howard. Knowledge Discovery and Interestingness Measures: A Survey. Technical Report CS 99 - 04, University of Regina, Saskatchewan, Canada, 1999.

[4] Zhang Chaohui, Chen Yucang, Zhang Qian, An Algorithm for Mining Quantitative Association Rules, Journal of Software, Beijing, China, 1998(11)：801－805.

[5] Liu Tongming, etc. The Data Mining Technology and Applications, National Defence Industry Press, Beijing, China, 2001