Research on Application of Mining Association Rules in Education Evaluation

Hao Jiang Jie Cai

(School of Computer Science & Engineering, Southeast University, Nanjing, 211100)

Abstract

The paper details the algorithms and the whole process of mining association rules, and conducts a mining experiment on the database of postgraduate information. First, according to the actual situation of the postgraduate database, it adopts Allied Bus Transfer Algorithm to realize the data extraction in data pre-processing phase, then mines the association rules by related algorithms from two aspects, students' comprehensive quality and teaching quality evaluation.

Keywords: Data mining; Association rules; Education evaluation

1. Introduction

Nowadays in most universities, accumulated data has become larger and larger after years of teaching. For example, only grade table in my alma mater's information database has nearly millions of records. A new technology which can find valuable teaching information more efficiently in massive data is urgently needed. Faced with this demand, the paper applies mining association rules[1] to make a beneficial attempt by combining with the actual situation of postgraduate database in my university.

Mining association rules is one of the hottest research fields in data mining. The concept of association rules was firstly proposed by R.Agrawal and R.Srikant. It shows a kind of relationship between a group of objects in the database and is applied for analyzing supermarket shopping data to get the customers' preferences at the beginning. Then the supermarket can put related goods focused on display to increase its sales volume according to these information. Research of association rules can help to discover association properties between different attributes in the database, and conclude that if causality, constraint or dependence existed among them. These relationships have significant value for enterprises.

2. Process and algorithm of mining association rules

2.1 Process of mining association rules

Task of mining association rules is adopting certain methods to dig out all the strong rules in the transaction database by the given minimum support and minimum confidence. It means that finding all rules such as $X \to Y$ which meet $Sup(X \cup Y) >= \text{minsup}$ and $Conf(X \to Y) >= \text{minsup}$.

Process of mining association rules can be showed in Figure 1. First, Pre- processing should be done to the data set, and then mine the association rules to the processed data. Pre-processing phase usually includes data extraction, data cleaning, data discretization, data transformation and so on. Frequent itemsets can be calculated by the minimum support and then association rules can be calculated by the minimum confidence. According to some evaluation criteria, such as interest measure, some unreasonable or wrong rules can be deleted. At last, we need to analyze rules by combining with practical problems to draw valid conclusions after all rules have been produced.

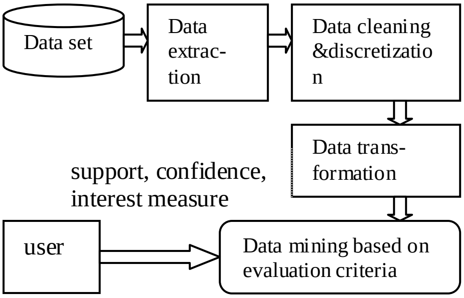

Fig. 1: Mining association rules

2.2 Apriori algorithm for association rules

Apriori algorithm[2] which was submitted in the early 90s by R.Agrawal is one of the most classic algorithms for mining the frequent itemsets of boolean association rules. It is based on priori knowledge of frequent itemsets and adopts iterative searching solution to explore frequent (k+1)-itemsets by frequent k-itemsets.

3. Process of mining association rules application in education evaluation

The database selected by this paper is from postgraduate information management system of my alma mater. It records related information over the years. The number of tables in this database is around 300, mainly includes the following aspects: school roll of postgraduate students, scholarship, training information, exam result, paper, course, tutor and so on. Process of data pre-processing and mining association rules is as below.

3.1 Data extraction

One of the complexities in mining postgraduate information is how to select the appropriate data for mining. This paper limits the scope of mining to school roll, exam result, paper, course, tutor and so on. Mining goals can be divided into two categories. One is students' comprehensive quality, including GPA, papers published, and basic information of students, the other is teaching quality evaluation, including results and related information of all courses, teachers' information and so on. There are two methods to determine the scope of the tables which are suitable for mining.

3.1.1 Query method on foreign key relationship

Query method on foreign key relationship is a method which determines the scope of mining by foreign keys between tables in the database. Foreign key relationship can be queried through the system tables in the database. However, this method has several non-ignorable shortcomings. First, it is only applicable to the database whose foreign key relationship is very strong. Furthermore, the mining rules obtained by this method cannot guarantee that covering all the tables wanted. Therefore, the paper puts forward a more effective method to determine the scope of tables, and names it Allied Bus Transfer Algorithm.

3.1.2 Allied Bus Transfer Algorithm

Allied Bus Transfer Algorithm is guided by a phenomenon that people always look for the least transfers when they travel by bus. According to this, we may transform the problem of querying associated tables to the solution of bus transferring. Each table in the database can be treated as a bus line, and every attribute in the table can be saw as a bus station. For instance, If we have two attributes needed to be found association tables, then the problem can be converted into: given two bus station, how to get the transfer way from origin to destination.

Assume that there are M tables in the database of postgraduate students, the number of attributes in each table is L on the average and total number is N. The tables which are suitable for mining are screened by construction of reachable matrix[3]. Based on analyzing the feature of postgraduate information database in my alma mater, this paper applies table-based construction method to construct reachable matrix. Obviously, all values in the diagonal line of reachable matrix are 1 because each attribute is reachable to itself. Algorithm 1 is table-based reachable matrix construction algorithm, the time complexity is $O(M*L^2)$.

Algorithm 1: Table-based reachable matrix construction algorithm

(a) Initial the reachable matrix V whose scale is $N*N$, set all of its elements to 0;

(b) Scan each table of the database, $L_1, L_2 \cdots L_M$, combine all possible two attributes in pairs and set corresponding elements in matrix V to 1 in turn. So an adjacency matrix is obtained after scanning all the tables and then we can convert it to a weighted connected undirected graph;

(c) Based on this graph and the matrix, each pair of attributes' shortest transfer path can be calculated by Floyd-Warshall Algorithm after n times iterations. Then we can calculate the final reachable matrix. Core idea of Floyd-Warshall Algorithm is described as follows:

(Ⅰ)Array d[i][j] records the current shortest distance between i and j;

(Ⅱ)For all the intermediate k from 1 to n, update the shortest distance between any two points to min(d[i][j], d[i][k]+d[k][j]).

After the construction of reachable matrix we can know which tables needed to associate funny attributes at least. According to this conclusion we can extract suitable tables to realize data extraction. At last, the paper obtains 7 tables which are appropriate to be mined, as following: school roll table, student papers table, student exam results table, course basic information table, course arrangement table, teachers table, outstanding master or PhD papers winner information table.

3.2 Data cleaning & discretization

Data in the student information database usually has the following features: lack of data, data exception and a mass of continuous data.

1. For the lack of data, this paper adopts the following methods: first, we define the data loss rate m=number of losing data records / total number. Then if 0.5<=m<1.0, it shows the data loss is quite serious and this attribute is not suitable for mining, directly give up the attribute. If 0.25<=m<0.5, it indicates the scale of data loss is large, abandon the records of data. And if 0<m<0.25, it expresses that data loss is ignorable, now we can apply average method which accumulates the average value of all data in the database to replace lost data.

2. In the case of data exception, we can use the following scheme: first of all, define the data exception rate e=number of exceptional data records / total number. This paper adopts methods of replace or delete to deal with these exceptional data, similar to the dispose of lack of data.

3. For those abundant continuous data, this paper adopts the following process mode:

Continuous data in the postgraduate information database is mainly numeric data. We can classify the grade of the continuous data and put the data into respective classes to realize data discretization[4]. Take the character of postgraduate exam results which are more centralized into account, the traditional numbering which classifies as A,B,C,D is not reasonable. Therefore, this paper applies k-means algorithm[5] which is one of the clustering algorithms to cluster the students' exam results. The initial clustering centers of the traditional k-means algorithm are selected randomly and this may have some unexpected impact on the final clustering results. This paper puts forward some improvement in the choice of initial clustering centers through combining with the character of postgraduate exam results. Improved k-means algorithm is shown in algorithm 2.

Algorithm 2: Improved k-means algorithm

(a) Scan exam results table in the database, store the students' results into an array and record the Max and Min scores of exam results. Assume that we need to cluster the results into k classes;

(b) Calculate the initial clustering center , select k clustering centers as following: Min+0.5*(Max-Min)/k ,Min+1.5*(Max-Min)/k,…,Min+(k-0.5)*(Max-Min)/k;

(c) Compare all data with clustering centers, calculate the distance of each point to the clustering center by formula of distance between two points. Each point will be grouped into recent clustering;

(d) Recalculate the clustering center, take the average value of all data as new clustering center, If it is equal to last clustering center, the algorithm ends, or turn to step (c).

Table 1 shows the clustering results of postgraduate exam results by improved k-means algorithm. The clustering of other attributes in the database such as age is similar.

Table 1. Clustering results of exam results by improved k-means algorithm

	clustering center	clustering number
1	73.26	2544
2	78.07	9220
3	81.41	12439
4	85.35	5

3.3 Data transformation

In data transformation phase we can convert the data which has been cleaned and discretized to target data for mining, including the determine of attributes for mining, the establishment of the temporary tables and the views. Ahead of this, we have determined the mining goals into two categories: students' comprehensive quality and teaching quality evaluation.

For the mining of students' comprehensive quality, just consider two tables: school roll table and student paper table.

We can obtain some appropriate information for mining from two tables, such as gender, enrollment way, students' undergraduate schools, GPA and paper collection situation, then make data discretization to these attributes. For example, paper collection situation can be discretized as R represents the general meeting, S represents EI and T means SCI.

After the treatment above, we have established two temporary tables in the database, One is for school roll information and another is for student paper information. We can associate them by student id to obtain the view of students' comprehensive quality, which can be used for mining directly. Establishing the view of teaching quality evaluation is similar.

3.4 Data mining

According to the views of students' comprehensive quality and teaching quality evaluation, we treat the maximum frequent itemsets and frequent closed itemsets as final itemsets to generate association rules. We choose to set support degree as 0.03 and confidence degree as 0.2 after extensive attempts. However the strong association rules mined by these metrics may not be funny, too. Thus this paper introduces two new metrics which are interest measure and positive or negative rules[6] to screen rules, named IL (Interest_Lift) analysis. Only those rules meet greater than the interest threshold and lift threshold at the same time can be included into the final results. We find set interest measure as 0.01 and lift measure as 5 is appropriate.

At last, we mine the corresponding association rules through programming based on Apriori algorithm.

3.5 Mining results and analysis

The number of rules is reduced after the screening of interest measure and lift measure. For students' comprehensive quality, goals focused on are mainly whether student exam results and paper published are associated with students' basic information. After extensive attempts， a few rules which are representative and appear most frequently are listed in Table 2:

Table 2. Mining results analysis of students' comprehensive quality

	rules	conclusion
1	recommended student -> excellent undergraduate school, academic master, outstanding results	Recommended student usually have excellent performance.
2	PhD->SCI papers published	PhDs are mainstay.
3	male, outstanding results -> take an exam for PhD	Male Student who learns well prefers to be a PhD.
4	outstanding results-> recommended student, academic master, papers from general meeting	Exam results are not necessarily associated with the level of papers.

For teaching quality evaluation, goals focused on are whether exam results and course information are linked with teachers. The mining results and analysis are listed in Table 3:

Table 3. Mining results analysis of teaching quality evaluation

	rules	conclusion
1	degree course, medium credit , professor->great exam results	Students may pay more attention to degree courses.
2	degree course, agedness->great exam results	Aged teachers usually have richer teaching experience.
3	medium credit, few class hours->great exam results, associate professor	Students may prefer the course with fewer class hours and learn well.

According to the rules mined above, we can choose some interested attributes, such as recommended student, PhD, students who have outstanding exam results and so on. List the lift measure of all other attributes which may impact on these interested attributes and sort them, we can also draw some conclusions. For instance, outstanding master papers winners are associated closely with tutors that their tutors are usually doctoral supervisors. And there is a funny phenomenon that male students are more possible to win the outstanding master papers comparing to female students.

4. Conclusion

Association rules is one of the most significant applications of data mining. This paper introduces the related knowledge of mining association rules, and applies its theory to the database of postgraduate information. It demonstrates overall process of data mining, including : put forward a new data exaction method named Allied Bus Transfer Algorithm, and then carry on the data pre-processing for the database of postgraduate information, application of interest measure and lift measure reduces some dirt-cheap rules successfully. Finally, The paper analyzes and evaluates the mining results.

5. References

[1] Jiawei Han. Data Mining: Concepts and Techniques[M]. Beijing: Mechanical Industry Press, 2000.

[2] Liping Chen, Research of Data Mining Algorithm Based on Association Rules[D]. Jiangnan University. 2008.

[3] Jian Zeng. Research on Association Rules Mining Algorithm base on compressed matrix and Its Application[D]. Hunan University, 2009.

[4] Usama M. Fayyad, Keki B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning[C]. Proceedings of the 13th International Joint Conference on Artificial Intelligence, 1993:1022-1027.

[5] Zheng Huang. Extensions to the k-means Algorithm for Clustering Large Data Sets with Categorical Values[J]. Data Mining and Knowledge Discovery, 1998, 2(3): 283–304.

[6] A. Savasere, E. Omiecinski, S. Navathe. Mining for strong negative association rules in large databases of customer transactions[C]. Proceed-

ings of 14th International Conference on Data Engineering, 1998:494-502.