Application of Data Mining in University Teaching and Management

Shaorong Feng

School of Information Science and Engineering, Xiamen University, FujianXiamen 361005, China

shaorong@xmu.edu.cn

Keywords: Data mining; Association rules; Student mark management; Teaching; Clustering

Abstract. The association rule is an important pattern in data mining. It is widely applied to the field of finance and commerce, but it is not so in university teaching and management process, and some factors that influence the effect of teaching and management are consequently ignored. Based on the research of the algorithm on the mining association rule, a method is proposed for clustering association rule. An example is given to explain the application of the proposed algorithm on analyzing students' scores. The results show that method is more reasonable and scientific than the traditional method in dependence analysis between the courses, which provides a scientific basis for university management and decision making.

Introduction

The theme "Education Data Mining"(EDM) [1] began to appear in 2005 and is discusses for numerous times at international seminars about American Association for Artificial Intelligence, (AAAI), Artificial Intelligence Education Development Council (AIED) and Intelligent Tutor System (ITS). The first international academic conference on education data mining was also held in Canada in 2008, and the 5th advanced data mining and application conference also first added the theme "application of data mining in education". The application of educational data mining technology includes: the teaching and educational administration management, student management, teacher and personnel management, enrollment and employment, equipment resource management and so on. Foreign educational data mining technology began early; data mining technology has been well applied in school management, and achieved fruitful results, to improve the quality of teaching management and teaching level of the tool. Ha and other detailed description of the possibility of the application of Web mining in the network distance education, and shows the prospect of application of Web mining in the network distance education [2], this article attracted the attention of people to the study. The application of data mining method is the core content of EDM research. Zaiane [3] uses data mining methods to assess the learning process, to help network learners to learn, the article is the most cited in the current EDM research articles. Romero and Ventura[4] through many aspects of EDM education tools, data sources, EDM method and other relevant documents on the EDM conducted in-depth description, they will according to the task of the education system in the network data mining method is divided into statistics, visualization and Web mining two. As a result of their work, they become the authoritative information on the study of foreign EDM.

At present, although our country is engaged in data mining researchers mostly concentrated in universities, but the technology in school management has not been fully applied in many universities, although the use of educational information management system, information management system of employment, but only on the function of data query, classification and statistics on a simple system in the accumulation of a large number of information, but cannot be used, these data can only reflect the data itself, but did not reflect a deeper level of more valuable information, how to utilize the data again, the existing historical data into available knowledge, so as to improve the university the management level and the quality of education, many universities are considering the problem. Data mining technology can solve the above problems in a certain extent, the use of advanced data mining analysis of multi-level multi angle, to the teaching of data technology, the analysis result can assist teaching and management to make scientific decisions, improve the level of teaching management, improve teaching quality, save the cost of running a

school. As a result, the application of data mining technology in the field of education management is getting more and more attention, and its application is mainly focused on the following aspects:

(1) Mine data of students' score, find key subjects affecting students' overall performance and improve scores of key subjects through teaching and management of these key subjects so as to indirectly improve the learning achievements of other subjects;
(2) Mine data for students in choosing subjects and find factors affecting students in choosing subjects so as provide basis for formulating scientific and reasonable plans to cultivate students and guide students in choose subjects at the same time;
(3) Mine students' career data and analyze key factors affecting students' career as well as the association between students' scores and career so as to provide decision-making basis for career guidance.
(4) Study the association between teachers' quality and teaching quality to provide decision-making and guidance for teaching management department, make pre-alarm to students' scores, urge students to learn and improve teaching quality.

In this paper, based on data mining technology, through the student achievement data, curriculum data mining, in order to find the curriculum and students' test scores, the correlation between curriculum and curriculum. To make education more scientific, to guide the relevant curriculum, to provide a scientific basis for curriculum arrangement.

Data Mining Relevant Technologies

Association Rule Mining Technology. Association rule mining is to find the interesting correlation or relevance between massive data. Association rule mining is an important branch of studying association rule mining.

The basic model of association rules is: Let $I=\{i_1, i_2, \ldots, i_m\}$ be a set of different items of m. Given a transaction database D, in which each transaction T is a collection of items in the I, namely $T \subseteq I$, T has a unique identifier TID. If the item sets $A \subseteq I$ and $T \subseteq A$, then the transaction T contains the item set A. An association rule is like the implication of $A \rightarrow B$, including $A \subset I$, $B \subset I$ and $A \cap B = \Phi$. If there are s% transactions in the transaction database including A and B, then we say that the support for association rules is s. If the transaction database D which contains the A transaction in c% also contains the B, then we say that the confidence of the association rules is c.

The problem of mining association rules is that the support degree and confidence level are greater than the minimum support and minimum confidence association rules specified by the user. It can be divided into two steps: The first step is to identify all the frequent item sets, which support not less than the minimum support program. The second step is to generate the confidence of the rule which is not less than the minimum confidence. The first step of the work is the most complex, because it requires a large number of I/O operations. Second steps in the formation of the rule is relatively easy. At present, most researches focus on the first step.

Apriori algorithm is a typical mining method of association rules. It uses layer by layer iterative method based on candidate generation to find frequent item sets. The main theoretical basis for the R.Agrawal project in space theory -- a subset of frequent item sets and frequent item sets is a superset of non-frequent item sets is non-frequent item sets. The core idea is: If the evaluation value of a regular neutron set is lower than the preset threshold value, then it will be discarded [5-9]. The specific process of Apriori algorithm is as follows:

Apriori Algorithm:

Input: D: Transaction database, min_sup: Minimum support count threshold.

Output L: Frequent item sets in D.

Method:

(1) L_1=find_frequent_1-itemsets(D)
(2) for (k=2; $L_{k-1} \neq \Phi$; k++) {
(3) C_k=apriori_gen(L_{k-1})
(4) for each transaction $t \in D$ {

(5) C_t=subset(C_k, t);
(6) for each candidates $c \in C_t$
(7) $c.count$++;
(8) }
(9) L_k={ $c \in C_k | c.count \geqslant min\_sup$ }
(10) }
(11) return $L = \cup_k L_k$;
procedure aprior_gen(L_{k-1}: frequent(k-1)-itemsets)
(1) for each itemsets$l_1 \in L_{k-1}$
(2) for each itemsets$l_2 \in L_{k-1}$
(3) if ($l_1[1]=l_2[1]$)$\wedge$($l_1[2]=l_2[2]$)$\wedge \ldots \wedge$($l_1[k\text{-}2]=l_2[k\text{-}2]$)$\wedge$($l_1[k\text{-}1]<l_2[k\text{-}1]$) then {
(4) $c=l_1 \triangleright\triangleleft l_2$; // Connection step: generating candidate
(5) if has_infrequent_subset(c, L_{k-1}) then
(6) delete c;
(7) else add c to C_k;
(8) }
(9) return C_k;
procedure has_infrequent_subset(c: candidates k-itemset, L_{k-1}: frequent(k-1)-itemsets)
(1) for each (k-1)-subset s of c
(2) if $s \notin L_{k-1}$ then
(3) return TRUE;
(4) return FALSE;

Clustering Technique. A cluster is a group of related sets according to some similarity function or similarity criterion is divided into several categories, so that individual differences in the same cluster is minimized, and the individual differences among different categories to maximize. Mainly related to the following important aspects.

(1) Determination of clustering object

Not what data are suitable for clustering, only after the pretreatment data, with a certain representation of the data before it is suitable as the object of clustering mining.

(2) The determination of "clustering criterion"

The determination of "similarity" or "similarity" standard. Emphasis is on the criteria and basis for division. And specific industry and specific data sets, specific tasks and other factors. "Clustering criterion" is the key point of cluster research.

(3) Clustering analysis of the results of association rules

In order to generate the clustering structure which is convenient for analysis from all the association rules, it is needed to define the distance between the rules. It can define the distance between the rules from different angles. According to the characteristics such as the rule itself (such as support and confidence) of structural differences between the rules (or the differences in the structure and the consequent) to define rules for the distance between. Because of the limitation of the traditional Euclidean distance, this paper proposes to use the correlation between the rules to define the distance between the rules, which will be a class of rules with high structure correlation.

With rule r_1: $X_1 \rightarrow Y_1$ and rule r_2: $X_2 \rightarrow Y_2$, the distance $Distance_{set}(X_i, Y_j)$ between the attribute set X_i and the attribute set Y_j is defined as follows:

$$Distance_{set}(X_i, Y_j) = 1 - \left| \frac{X_i \bigcap Y_j}{X_i \bigcup Y_j} \right| \tag{1}$$

The distance between rule r_1 and rule r_2 is defined as:

$$Distance_{ru1e}(r_1, r_2) = \lambda_1 \alpha \times Distance_{set}(X_1 \cup Y_1, X_2 \cup Y_2) + \lambda_2 \times Distance_{set}(X_1, X_2) + \lambda_3 \times Distance_{set}(Y_1, Y_2) \tag{2}$$

Among them, λ_1, λ_2, λ_3 is a non-negative real number, and $\lambda_1+\lambda_2+\lambda_3=1$, according to the specific value of the three parts of the user specified preferences, for example, $\lambda_1=0$, $\lambda_2=2/3$, $\lambda_3=1/3$, said the

rules emphasize positive correlation between users.

Clustering Algorithm of Association Rules:

Input: correlation coefficient, λ_1、λ_2、λ_3, rule set R, neighborhood radius *eps*

Output: clustering results C_1, C_2, ..., C_k

Method:

(1) while (R!=*null*) {
(2) m=0;
(3) for each rule $r_m \in R$ do {
(4) C_k.*add*(r_m);
(5) R.*remove*(r_m); m++;
(6) n=0;
(7) for each rule $r \in R$ do {
(8) *distance*=$Distance_{rule}(r_m, r_n)$;
(9) if (*distance*$\leqslant$*eps*) then {
(10) C_k.*add*(r_n);
(11) R.*remove*(r_n); n++;
(12) }
(13) }
(14) }
(15)}

Data mining is a deep analysis of data information; it is very useful to apply the technology of data mining in the teaching evaluation. It can be a comprehensive analysis of the inherent relationship between the examination results and various factors. For example, through the analysis of the school student achievement related database system, data mining tools can answer questions such as "what factors may have an impact on the student's academic performance" and so on. This is the traditional evaluation method cannot have. Through the analysis of data mining, the evaluation results can bring unprecedented harvest and surprise to the teaching.

Data Mining Process Model

Data mining process model of the main reference CRISP-DM model, CRISP-DM (standard process cross-industry data for mining), Cross industry data mining process standards". This process model in 1999 by the EU institutions drafted by the development in recent years, CRISP-DM model (Knowledge Discovery in KDD in Data) occupies a leading position in the process of the model, more than half of the data mining process model reference CRISP-DM model[10]. Data mining can be used as a decision support process of scientific analysis method, analysis of mining technology used in the examination of the students will be in the data, through the analysis of historical data mining, school teaching and management, to produce the corresponding rules, dig out the factors affecting student achievement. For the school to make a scientific teaching plan, and constantly improve the curriculum to provide a reference, and can answer "related courses is reasonable order" and other similar problems, and these problems such as the use of the traditional method of teaching evaluation is not known. Through the analysis of the results of the data can not only get the conclusion of the above problems, but also get other useful results. According to the idea of the CRISP-DM reference model, the implementation of data mining process[11], as shown in Fig. 1.

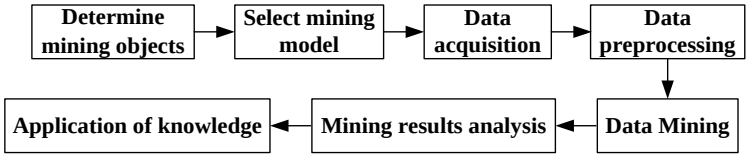

Figure 1. Data mining implementation flow chart

(1) Select mining object

The first step in data mining is to determine the object and purpose of the mining. Different mining objects can lead to different results and errors, which cannot reach the goal of data mining, which is a prerequisite for the success of data mining.

(2) Select mining model

There are many kinds of data mining methods, each method has its own characteristics, different mining methods for different problems. Select a suitable mining method, in order to ensure the solution to the problem.

(3) Data collection

Data collection is the collection of relevant historical data. Data mining is to discover the hidden information among the data through the analysis of historical data, so as to analyze the relationship between the data and predict the future development of the data. The workload of data acquisition is larger, and the time is more. Some data can be directly obtained through the questionnaire survey; the results of the analysis system data can be imported from the previous student achievement database.

(4) Data preprocessing

In view of the selected data model, the collected data are preprocessed, so that the data can be fit for the selected data model, which provides convenience for the subsequent data mining. As for the absent student achievement in the database of student achievement.

(5) Data mining

Data mining is the core of the whole process of mining, and it is the concrete realization of the data mining。Data mining algorithm to mine the data preprocessing. Data can be classified, and then establish a classification model, which can improve the efficiency of mining in mining. According to the mining algorithm of the model, the development tool is selected to implement the algorithm to the specific development system, and the data mining is completed.

(6) Result analysis.

The data generated by the system is analyzed, and the data is converted into a familiar form of data representation.

(7) Application of knowledge

To apply the knowledge gained from the analysis to the specific links, to solve the specific problems in the work and study. If the data is applied to the teacher's teaching plan, it can improve the teaching strategies, and guide the teaching to improve the quality of teaching.

Research and Analysis of Data Mining in Universities Teaching and Management

This paper will combine the Apriori algorithm and clustering technology to data mining of student achievement data, and realize the analysis of curriculum correlation.

Student achievement as an important indicator of the results of the assessment study. It is not only the detection and evaluation of students' learning effects and teachers' teaching effects, but also can be feedback to the teaching activities, serve the teaching strategies, and provide valuable information for educational research. However, the commonly used performance analysis method generally can only get information, such as mean, variance, significant difference test, reliability, validity and so on. But in the actual teaching, such as students in the learning of a course, which is a door or a few leading courses has the greatest influence, including what factors outside the teaching on student achievement has great impact and other valuable information is often not informed. Therefore, for students' in-depth data mining analysis, find out the influence of various potential factors of the students, will promote the school to carry out more targeted personalized education, and create a new teaching management environment, to further promote the teaching quality and improve the level of management.

Data sources involved in the data source from the academic department of the relevant forms of data, include: The student information table (college, professional, class, name, student number), course information table (course code, course name), performance information table (student

number, year, semester course code, course name, course grades), schedule (year, semester).

In this paper, the Information Institute of computer science and technology professional 150 student achievement information as a data source, data mining analysis, preprocessing, resulting in a suitable mining analysis of the student achievement table data source. The specific process is as follows:

(1) In this analysis, the course of the students' performance in the course is analyzed, also can change the conditions in the algorithm, find out the influence of the internal factors and the degree of correlation between the course.

Input: students score less than 60 points, the minimum support degree =0.05. Student achievement information table (part), as shown in Table 1.

Table 1 Studentscore information table

Name	Introduction to Computer	College English	Higher Mathematics	College Physics	C# programming	Computer English	C Programming	Data structure
Student 1	86	77	85	90	79	82	63	84
Student 2	71	60	75	75	87	60	56	79
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
Student 150	49	62	88	74	78	90	73	49

Table 2 Association rule graph

Rule	Support	Confidence
Higher Mathematics→College Physics	12	0.45
College Physics→Higher Mathematics	12	0.60
Higher Mathematics→C Programming	7	0.30
C Programming→Higher Mathematics	9	0.55
Higher Mathematics→College English	7	0.35
College English→Higher Mathematics	9	0.33
Introduction to Computer→Higher Mathematics	8	0.41
Data structure→C Programming	9	0.51
C Programming→C# programming	10	0.36
College English→Computer English	8	0.45

The results of 60 points below the score set to "0", by the Apriori algorithm, seeking *k*-frequent sets. This test to the end of the generation of 2- items. L_2={(Higher Mathematics, University Physics), (higher mathematics, C programming language), (Higher Mathematics, College English), (Introduction to Computer, Higher Mathematics) (data structure, C programming), (College English, computer English), (C language programming, C# programming)} by the corresponding L_2 rules generated, as shown in table 2.

If the minimum confidence level is 0.35, first, 2, 4, 5, 7, 8, 9, 10, and the rule for strong association rules. It can be known from the mining results that the higher probability of the university physics, College English and C language program design is higher in the case of the higher mathematics course. Through the above results, we first analysis, higher mathematics, College English, college physics are public basic course, the course has certain continuity at the same time, math, English and physics and high school students basic course, indicating that students in high school basic courses based on solid or not in the university is very important if some students, college entrance examination in mathematics, English and physics are relatively poor, learning these courses will affect the University period. Second, due to the requirements of the algorithm in the C programming language is higher, so if the student's math performance is not good, it will affect the C language program design. If the students' College English performance is

poor, then his computer English performance will be poor, which shows that college English learning has a great impact on computer English learning. C language program design results also affect the performance of C# program design, but also shows that the programming language has a great link, from the results show that the C language program design is the leading course of C# programming.

(2) Change the conditions, the relation between mining courses more than 85 courses, the minimum support of =0.05, the score of 85 points or more records were set to "0", the association rules generated 2- set and corresponding, we can find mutual correlation between the internal factors and other curriculums.

From the above analysis we can get the following enlightenment, there is a certain relationship between some courses, there is a certain course. Some of the results of the course can directly affect the results of other courses. The school through these potential rules and guidance on students' learning and teachers' teaching, help schools to strengthen the revision of some basic course teaching plan through the analysis of the results suggest that school administrators; help teachers to understand the students' learning rules, in the daily teaching methods for different students of different needle implementation, improve the quality of teaching; students targeted in their daily learning according to their actual situation, improve the learning efficiency.

(3) Clustering of Association Rules

We can use the association rule clustering algorithm which is given in this paper, and the association rules can be grouped according to our different needs. And cluster analysis of the results of the association rules, and finally get a set of rules to meet the requirements. We classify the association rules which we care about, and get the association rules of law and value. If the correlation coefficient is: λ_1=0, λ_2=1, λ_3=0, and the smallest cluster is set to 0, is the association rules according to the requirements of the cluster, and the same attribute set. Put the association rules in the same class together. It reflects the meaning of: the same kind of course exactly what classes have an impact. If the correlation coefficient is: λ_1=0, λ_2=0, λ_3=1, is required for clustering association rules according to the set. If the smallest cluster is set to 0, is the association rules in the same class together. It reflects the meaning of: the same course, what is the impact of what classes.

Conclusion

This paper makes full use of the association mining algorithm and clustering method, and makes full use of the data of students' performance, draws the relationship between curriculum and curriculum, and analyzes the order of the curriculum. The concrete realization of the analysis of the data in the database of student achievement, the mining results are used to guide students to learn at the same time, put forward according to the results of the undergraduate teaching personnel training program to make corresponding adjustment, change the original students' blindness, association and mutual influence of curriculum classification, strengthen students learning purpose the.

At present, with the various colleges and universities are expanding the enrollment scale and improve the quality of teaching, through the university enrollment expansion, using various forms of education to provide practical and diverse learning opportunities to cope with the diversified development of the society and the demand for technology professionals for the public high school, the quality of students between the increasingly fierce competition. It can be considered, which school to get more excellent students, which school will be able to achieve faster development. Therefore, school leaders need to more accurately understand the operation of the school, the students and the school graduates by information and social acceptance, so as to more rational use of limited resources to schools, formulate the corresponding improvement measures, in order to achieve the school health rapid development needs. In the process of using the association rule discovery technology to excavate the university student information database, we found that some factors (or rules) that cannot be paid attention to. Choose a different database mining, using different thresholds (support, confidence and interest), association rules are different, specific to each school according to the actual situation and the task and aim to tap their own different set, analyzed and summarized, to dig out the repeated data mining, draw meaningful the information. In

addition, it is very important to guide the mining work and evaluate the mining results by the people who are familiar with the business.

In the study of association rule mining results, found the rules still have some errors, the reason, we believe that in the attribute field data set selection, there are many factors not taken into account, the information content is not comprehensive, the data set may not be the best data set, this point to further study in the future.

It should be explained that data mining is just a tool, it can discover some of the potential law, but will not tell you why, also cannot guarantee every decision is correct according to the mining rule. The success of data mining must have a practical understanding of the relevant business areas, need a thorough analysis of the mining results, only to understand the data to make a detailed analysis according to the results, find out the most reasonable explanation as the decision-making reference for managers.

Acknowledgement

Fund Project: Supported by the Major Program of National Social Science Foundation of China(Grant No. 13&ZD148) (Major project of 2013 National Social Science Fund)

References

[1] Educational Date Mining[DB/OL]. <http://www.educationaldatamining.org>

[2] Sung Ho Ha, Sung Min Bae, Sang Chan Park. Web Mining for Distane Education[C]. Proceedings of the 2000 IEEE International Conference on Management of Innovation and Technology, 2000. ICMIT 2000, Vol. 2: 715-719.

[3] Osmar R. Zaïane. Web Usage Mining for a Better Web-based Learning Environment[C]. Proceedings of conference on advanced technology for education, Banff, Alberta, 2001: 60-64.

[4] C. Romero, S. Ventura. Educational Data Mining: A Survey from 1995 to 2005[J]. Expert Systems with Applications, 2007, (33):135-146.

[5] R. Agrawal, R. Srikant. Fast Algorithms for Mining Association Rules[J]. Proc. Int. Conf. Very Large Databases, Santiago, Chile, 1994.9

[6] G. V. Kass. An exploratory technique for investigating large quantities of categorical data[J]. Applied Statistics, 1980, 29: 119-127.

[7] J. C. Schlimmer, D. Fisher. A case study of incremental concept induction[C]. Proceedings of the Conference on Artificial Intelligence (AAAI'86), Philadelphia, PA, 1986, page 496-501.

[8] Park J.s, Chen M, Yu P. S. An Effiective Hash Based Algorithm for Mining Association Rules[C]. ACM SIGMOD International Conference Management of Data, 1995: 813-825.

[9] HanJ, Pei, YinY. Mining Patterns without Candidate Generation[c]. SIGMOD, Dallas, Tx, May 2000: 1-12.

[10] A.W.Kamakura. M.Wedel. D.F.Rosa. A.J.Mazzon. Cross-selling through database marketing: a mixed data factor analyzer for data augmentation and prediction[J]. International Journal of Research in Marketing, 2003, 20(1): 45-65.

[11] Kuangjihong. The research and application of data mining in Analysis for students' achievement[D]. Jilin University, 2012.

