The Implementation and Application of Data Mining in University Students' Grades

Jiang Hongyan
Computing Center of Anshan Normal University
Anshan,Liaoning China 114005
E-mail: as671315@163.com

Li Liangjun
Computing Center of Anshan Normal University
Anshan,Liaoning China 114005
E-mail: asllj2007@sina.com

Abstract—Nowadays the society is full of all kinds of information. With the rapid development of science and technology, more and more people come to realize the value of information. Much information hide in the sea of data, there are so much data around us. So data mining technology is applied widely. This paper introduces the concept of association rules in data mining and the classic Apriori algorithm. We will mine out the relations which will affect the students' marks and give the core steps of the algorithm and detailed realizing steps. At last we give the results and analysis of the running of the example. The conclusion is a practical suggestion for the computer teaching administration. Some unexpected DM knowledge may enhance the quality of education and guide the teaching methods scientifically. With the further development of DM, more and more DM knowledge will be applied to teaching and it surely will bring us an unprecedented rewards and surprise[1].

Keywords-Data mining; Association rules; Apriori algorithm; Support; Confidence

With the development of social education, efforts to improve the teaching quality of college students is the goal of each college, but the student grade is the important basis of evaluating the teaching quality. Now with the scale expands unceasingly, the number of students is growing constantly, how to find the potential rule from a lot of results data and to predict the development trend of results, which teachers' teaching targeted suggestions are put forward. The management of students teaching put have targets,to improve the quality of teaching is very important. Data mining technique is feasible and effective method to solve the problem. This paper tries to analyze correlation technique in data mining and obtains some interesting knowledge.It plays a positive role in promoting the quality of teaching. So it helps the teaching work smoothly[2,3].

I. INTRODUCTION OF APRIORI ALGORITHM

Apriori algorithm is one of the most influential algorithms of frequent itemsets in mining Boolean association rules, Use a search step by step iteration method：K- set used to search (k+1) - itemsets. First of all, find the set of frequent 1 – itemsets, as L1, Ll is used to find the set of frequent 2 - itemsets L2, L2 is used to find the set of frequent 3 - itemsets L3, until cannot find frequent k - itemsets, every search needs to scan the database at a time[4].

The Apriori algorithm consists of the following two parts:[5]

(1)Use the candidate itemsets to find frequent itemsets

How to use the frequent (k - 1) - itemsets Lk - 1 to find frequent k - itemsets Lk is the core of the Apriori algorithm, this process using Apriori nature compress space, In order to improve the speed of search database, this step is divided into steps of connect and pruning [6]:

• Connect the step:

For the Lk, the set of k- candidate itemsets are produced by Lk-1 and their connections. The set of candidate itemsets,denoted as Ck. Let L1 and L2 be itemsets in Lk-1. Mark li[j] as j of Li (For example,L1[4] as the fourth of l1).Assume that the transaction and the item sorting sequence according to the dictionary,Implementation of L k-1 X Lk-1,If their former (k - 2) in the same item, Lk - 1 elements l1 and l2 can be connected. if (ll[1]= l2[1])∧(ll[2]= l2[2]) ∧…∧(ll[k-2] = l2[k-2])∧(ll[k-1]< l2[k-1]),Connect the L1 and L2 result set is ll[1]ll[2]…ll[k-1] l2[k-1],and ll[k-1] < l2[k-1] is simple to ensure no repeat[7].

• The pruning step：

Ck is a superset of Lk,The Ck sets may be a frequent and may also be non frequent, But frequent k-itemsets are all included in the Ck. By scanning the database, calculate the support degree of itemsets in Ck,and compared with the minimum support degree,determine Lk.However, Ck may contain a lot of itemsets,so the calculation amount involved is very big.For the compression of Ck,you can use the Apriori property：if any one (k-1) - a subset of a candidate k-set is not in Lk-1,the k- candidate item sets may not be frequent,to delete the candidate k-item sets from Ck.This can reduce k- itemsets number in the Ck,reduce the times of scanning database,to improve the efficiency of the algorithm[8].

According to the above two key steps,the Apriori algorithm can be described as a specific process.the first step: we can find out frequent 1 - itemsets L1 from containing each candidate_1_itemsets (Cl). Then use Lk-1 to connect the generation of candidate Ck,and according to the nature of Apriori delete those with candidate itemsets of non frequent subsets. Next,scan the database,count up the candidate itemsets support count,compared with the minimum support count,form the frequent sets of Lk.

(2)Generating association rules from frequent itemsets

For each frequent itemsets l,generate All non empty set of l, if $\frac{\sup port\_count(l)}{\sup port\_count(s)} \geq \min\_conf$,then obtain the rules " s $\Rightarrow$ (l-s)".Among them,min_conf is the minimum confidence threshold[9].

II. APRIORI CORRELATION ALGORITHM IN THE APPLICATION OF STUDENT GRADE

1.Data preprocessing

(1)Data integration

Data integration is to combine the data from multiple data sources. In this study, using database technology to generate basic database 1 of student performance analysis from the multi database files. Randomly select some students of computer course （basic computer,Visual Basic the program design,the program design）grades,as shown in Table I：

K1 for Computer Basic grade,K2 for Visual Basic the program design grade,K3 is a multimedia courseware making grades.

TABLE I. STUDENT GRADE ANALYSIS OF BASIC DATA TABLE

xh	k1	k2	k3
040101	85	88	76
040103	86	95	87
040107	72	68	70
040113	85	94	80
040115	94	91	86
040206	68	45	35
040213	67	48	59
040215	95	97	72
040218	98	95	94
040229	97	94	98
...	...	...	...

(2)Data conversion

Data conversion is mainly carried out standardized operation of data.

During the mining association analysis to student grade, we need logical data type,so we should convert the student score data into Boolean representation,Because of mining is the excellent relations between various disciplines,so more than 90 points field value is "1",there is the item in the transaction, as a "0", the item does not exist in the transaction.

Table I can be converted to a logical data table for mining, as shown in Table Ⅱ：

TABLE II. STUDENT GRADE ANALYSIS OF LOGIC AND DATA TABLE

xh	k1	k2	k3
040101	0	0	0
040103	0	1	0
040107	0	0	0
040113	0	1	0
040115	1	1	0
040206	0	0	0
040213	0	0	0
040215	1	1	0
040218	1	1	1
040229	1	1	1
...	...	...	...

2.The specific steps are as follows：

(1)Set up transaction data tables DM_df_AllStudents.Because only consider good grades,so the record that can not meet the needs of mining tasks should be deleted.

(2)Calling a stored procedure sp_find_ candidate_1_itemsets,counting the occurrences of each,in the frequent 1- itemsets data table,this table only two fields are Item and SupCount,field of Item for the character data type,to hold the project name as K1, K2 and other fields,SupCount is a numerical type,used to store the number of records containing this project in the transaction data table.

(3)Call a stored procedure sp_find_frequent_1_itemsets,calculate the support of each item, delete support count that is less than the minimum support degree records in the table frequent_1_itemsets. Get the frequent 1 – itemsets.

(4)In the subsequent all frequent itemsets divided into two steps.The first step : produce the candidate items,The second step : generate frequent item sets. Specific process is：First of all, conduct the natural connection by frequent_ (n-1) itemsets (n>=2) ,generate candidate n_itemsets,store in frequent_ n_itemsets According to the ascending order.

The second,obtain the support degree of candidate n_itemsets by Scanning DM_df_AllStudents data table, record the data in table field frequent_n_itemsets SupCount.

Finally, delete the records whose support count is less than the minimum support min_sup. Until find out all the frequent itemsets. If discover that a certain number of candidate itemsets is zero, stop the operation. Finally, output the frequency set of all projects. The number of scanning the transaction data table depends on the maximum length of frequent itemsets.

(5)Calculate the confidence of each the final frequent item non-empty set, remove the records which is less than the minimum confidence threshold. Finally generate rules,storage in the data table DM_df_ Association Rules.

Core steps of the algorithm[10]：

(1)L1=find_frequent_1-itemsets(D);

(2)for (k=2; Lk-l≠Φ; k++) do{

(3) Ck=Apriori_gen (Lk-1,min_sup) ;

// the new candidate set

(4) for each transaction t∈D do{

(5) Ct=subset (Ck,t);

// the candidate set in transaction t

(6) for each candidate c∈Ct do

(7) c.count++;

(8) }

(9) Lk={c ∈Ck|c.count≥min_sup}

(10) }

(11)return L=∪kLk;

procedure Apriori_gen(Lk-1:frequent(k-1)-itemsets; min_sup: minimum support threshold)

(1) for each itemset l1∈Lk-1

(2) for each itemset l2∈Lk-1

(3) if(ll[1]= l2[1])∧(ll[2]= l2[2])∧…∧(ll[k-2] = l2[k-2])∧(ll[k-1]< l2[k-1]) then{

(4) c=l1 X l2

//join step: the candidate item sets

(5) if has_infrequent subset(c,Lk-1) then //prun step

(6) delete c; pruning

(7) else add c to Ck;

(8) }

(9) return Ck;

procedure has_infrequent_subset(c: candidate k-itemset; Lk-l: frequent(k-1)-itemset)

(1) for each (k-1)-subset s of c

(2) if s∉Lk-1 then

(3) return TRUE;

(4) return FALSE;

The analysis of the results of the operation an example

Assume that the transaction number is 11

TABLE III. ASSOCIATION RULE MINING DATA TABLE

xh	k1	k2	k3
040103	0	1	0
040113	0	1	0
040115	1	1	0
040215	1	1	0
040218	1	1	1
040229	1	1	1
040307	0	1	0
040310	0	0	1
040401	1	1	1
040405	1	1	1
040410	1	1	1

For excellent courses, association rules mining in the table Ⅲ,need to give the support and confidence.Assume that the minimum support degree is 30%、When the confidence level is 50%.

Calling a stored procedure sp_find_ candidate_1_itemsets and sp_find_frequent_1_itemsets,determination of frequent 1- itemsets L1 {k1,k2,k3},When for subsequent all the frequent itemsets by frequent_ (n - 1) _ itemsets (n > = 2), the natural connection,produce the candidate itemsets of n,until find all the frequent itemsets.In the case of the candidate C2 {k1,k2} 、 {k1,k3} and {k2,k3},thus draw a conclusion that frequent two itemsets L2 is {k1, k2}、{k1,k3} and {k2,k3}, we can obtain the candidate set C3 as {k1,k2,k3} by L2. Finally we learned that C3 is a frequent set L3,and for the maximum frequent set.At this time, scan is end. From the above analysis we can learn that through effective pruning,the efficiency of frequent items is greatly improved.

Calculate the confidence of ultimately frequent item sets for each nonempty set,delete less than the minimum confidence threshold records,finally produce association rules,deposit in the data table DM_df_ Association Rules.

(1)K1, k2 is good at the same time,K3 has more than 71% of the best possible;

(2)K1, k3 is good at the same time,K2 has more than 100% of the best possible;

(3)K2, k3 is good at the same time,K1 has more than 100% of the best possible;

(4)K1 is good,K2、K3 has more than 71% of the best possible;

(5)K2 is good,K1、K3 has more than 50% of the best possible;

(6)K3 is good,K1、K2 has more than 83% of the best possible;

From the above results,we can get the following potential association： There is some relationship between some courses,they influence each other,some courses will have a direct impact on the grade of some courses. Through the study, it has further proved the effective and practical of association rules in the course of correlation analysis. This will provide students with some help of relevant decisions in the course of study. The analysis result will have an important reference value to improve the teaching work and the students' teaching management in the future[11,12].

III. CONCLUSION

On student grade data mining,to obtain useful information method is the analysis of the results,it will be a beneficial attempt and important applications.It is a powerful necessary complement to performance management. Now more and more researchers are engaged in the study of these aspects,I believe in the near future,data mining technology of the application in the management of colleges and universities will be more and more widely, more and more perfect.

ACKNOWLEDGMENT

We gratefully acknowledge the supports by the National Natural Science Foundation of China (No.60773218)

About the author：Jiang Hongyan（1973－）,female,Liaoning anshan ,graduated from liaoning normal university in 1996 ,associate professor ,agraduate student.

Li Liangjun（1967-），Male， professor， Ph D. of Northeastern University.

REFERENCE

[1] [1] Yu cheng, "The application of data mining in students' grades analysis",CaiZhi,2011,No.14,P81

[2] [2] Lv Fenghua,"The application of data mining process in teaching ", Journal of the jinhua professional technology institute, 2003.9,No.3,P36～P38

[3] [3] Wang Yu,Liu Zuoyi,"Research Progress Funded by National Natural Science Foundation of China on Managerial Data Mining",Chinese Journal of Management,Nov.2011,Vol.9,No.11.P1674~1678.

[4] [4]Zhu Yanli,Gao Guohong, "Apriori algorithm research and its application in analysis of student performances'grades",FuJian Computer,2010,No.1,P147

[5] [5] Fang Weiwei, "Market Basket Analysis Based on Association Rules",Journal of Sichuan University of Science&Engineering,2010.8,No.4,P431

[6] [6].Chen Wenwei,Huang Jincai,Zhao Xinyu,Data Mining Technology [M],Beijing industrial university press ,2002

[7] [7] written by Jiawei Han,Micheline Kamber,translated by FanMing,Meng Xiaofeng,The Concept of Ddata Mining and Technology[M],Beijing : Mechanical industry press ,2007.3,Version 1 ,P151

[8] [8].Cui Xuewen, " The application Association rule mining algorithm Aprior in students'grades analysis", Journal of Hebei North University,2011,No.1, P44~P47

[9] [9] LinZhi, "The application of association rules analysis in management of students'grading", Journal of Ningbo Polytechnic,2010,No.2 ,P64~67

[10] [10].Shao Fengjing,Yu Zhongqing, Principle and Algorithm for Data Mining[M], China water conservancy and hydropower press,2003,P94

[11] [11] ZhangXao, "The application of association rule mining in students'grades analysis", Journal of Yili Normal University,2013,No.4,P7-~P43

[12] [12]Li Xueyan,"Research and application of Data Mining in the Management of Colleges Achievement",Computer & Digital Engineering,2011,Vol.39,No.7,P147~149.