Advances in Social Science, Education and Humanities Research, volume 322
2nd International Seminar on Education Research and Social Science (ISERSS 2019)

Using Apriori Algorithm on Students' Performance Data for Association Rules Mining

Xiaodong Wu*, Yuzhu Zeng
Faculty of Mathematics and Computer Science, Quanzhou Normal University
Fujian Provincial Key Laboratory of Data Intensive Computing
Key Laboratory of Intelligent Computing and Information Processing, Fujian Province University
Quanzhou, China

***Abstract*—With the development of information technology, many colleges and universities have established student information management system. The long-term operation of the student information management system will generate big data for colleges and universities. Moreover, there exists valuable information in the huge amount of data. Hence, it is necessary to use the data mining method to mine the massive data and get some valuable reference information so as to improve the teaching and management of students. In this paper, the Apriori algorithm is used to mine association rules of 34 courses of 100 students majoring in computer science and technology, so as to find out the correlation between courses and the factors that lead to the high or low grades of courses. R is used to conduct the experiment to discover the association rules, and the association rules are analyzed and discussed. The results of data mining on students' achievements in this work are expected to provide a reference for improving the teaching quality of computer science and technology courses.**

Keywords—data mining; apriori algorithm; association rules; students' achievement analysis

I. INTRODUCTION

In recent decades, with the wide application of Information Technology, many colleges and universities have established their student management system. With the continuous operation of the system, the amount of student information data keep increasing, and eventually a large amount of data will be accumulated. However, most of those data have not got well exploitation and utilization. It is essential that we should make full use of those data to get some meaningful information, such as to provide some valuable reference for improving the teaching quality and assist the students to improve their academic performance.

There are increasing research interests in utilizing data mining method on the educational field in recent years. Baker and Yacef [1] referred to utilizing data mining methods to educational data as Educational Data Mining (EDM). Raheela Asif et al. [2] used the data mining methods to study the performance of undergraduate students and predict the students' achievement. Altaher and BaRukab [3] used the Adaptive Neuro-Fuzzy Inference System to predict students' academic achievements by the previous exams results of the students. Ahmed et al. [4] investigated the factors that affect students' performance in order to improve the education quality through the use of data classifier methods. Hamsa et al. [5] used the Decision Tree and Fuzzy Genetic Algorithm classification methods to develop the students' academic achievement prediction model. In [6], Elbadrawy et al. proposed a regression-based methods and a matrix factorization-based methods, which were extensively used in the e-commerce recommender systems, to forecast students' achievement in the future courses as well as their in-class assessments. In [7], Harwati et al. used the K-means Cluster algorithm to classify students based on their demographic.

In this study, we utilize the apriori algorithm, a classic algorithm of associate rule mining, on the students' achievement data to reveal the hidden relationships between the achievements of students in different subjects in colleges and universities, so as to provide a reference for teachers to improve teaching quality and students to improve their achievements.

II. ASSOCIATION ANALYSIS AND APRIORI ALGORITHM

A. Association Analysis

The association analysis method is useful when discovering relationships that are hidden in big datasets. The relationships can be represented in the form of frequent itemsets, i.e., sets of items frequently presented in many transactions, or in the form of association rules, which indicate the relationships between two itemsets. An association rule can be expressed as $A \Longrightarrow B$, where A and B are the subsets of a given itemset T. For example, we can use {programming language = failure} $\Longrightarrow$ {data structure = failure} to show an association rule that if a student fails in the course programming language, he will probably fail in data structure.

Three measures are used to evaluate the association rules in this work, i.e., support, confidence and lift.

1) Support

The support of an association rule $A \Longrightarrow B$ is the percentage of transactions that contain $A \cup B$, i.e., the union of A and B. The support of rule $A \Longrightarrow B$ can be calculated as the probability [8],

$$\text{support}(A \Longrightarrow B) = P(A \cup B). \quad (1)$$

2) Confidence

Let D denotes a set of transactions, each of which is an itemset. The confidence of a rule $A \Longrightarrow B$ in a transaction set D

indicates how frequently items in the subset B appear in the transactions that contain A. The confidence of rule $A \Longrightarrow B$ can be calculated as the conditional probability [8],

$$\text{confidence}(A \Longrightarrow B) = P(B|A) \quad (2)$$

3) Lift

The lift between A and B reflects the correlation between A and B [8]. It can be calculated by (3):

$$\text{lift}(A,B) = \frac{P(A \cup B)}{P(A)P(B)} \quad (3)$$

B. Apriori Algorithm

Apriori [9] is a classical algorithm for discovering frequent item sets and mining association rules over transactional datasets. A *frequent itemset* is a set of items that frequently appear together in a transaction data set. Frequent itemset discovering is essential to the association rule mining.

Discovering frequent item sets is a two-step process: join and prune. In the join step, the frequent (k-1)-itemsets, which contain k-1 items, are joined to each other so as to generate k-itemsets containing k items, denoted by c_k. Then, in the prune step, the k-itemsets in c_k which are not frequent are excluded, such that the frequent k-itemsets are obtained. After the frequent itemsets are determined, association rules can be obtained by generating non-empty subsets of each frequent itemsets and selecting those satisfying the minimum support and confidence.

III. Data and Methodology

A. Data preparation

The raw data used in this work is the achievements of 34 courses for 100 students who are majoring in computer science and technology and had been enrolled in September, 2014. The original score data are shown in table I.

TABLE I. THE RAW SCORE DATA

No.	College Chinese	College English1	Specialty English	Linear Algebra	Discrete Math	…
1	81	68	62	60	65	…
2	78	60	68	61	62	…
3	95	81	73	47	70	…
4	76	45	46	61	55	…
5	94	66	63	62	45	…
6	86	73	67	60	60	…
7	91	91	61	60	50	…
8	78	41	64	83	54	…
9	89	80	77	83	61	…
10	75	72	66	82	67	…
…	…	…	…	…	…	…

As can be seen from Table I, teachers of different subjects may have different preferences for giving a mark. For instance, achievement of Discrete Mathematics is generally lower than that of College Chinese. For this reason, it is unreasonable to determine the grades (usually denoted as five levels of A, B, C, D and E) of the students simply by the scope of their marks. Therefore, in this work, the grades of the students are determined by their rank rather than score. For each subject, no matter what the actual score is, the score ranking ranges from 1 to 20 are classified as grade A. and the score in ranges [21, 40], [41, 60], [61, 80], [81, 100] are respectively classified as grade B, C, D and E. According to our strategy, the corresponding grades converted from the raw data in Table I are shown in Table II.

TABLE II. THE CONVERTED GRADE DATA

No.	College Chinese	College English1	Specialty English	Linear Algebra	Discrete Math	…
1	D	D	D	D	C	…
2	E	E	B	D	C	…
3	A	B	A	E	B	…
4	E	E	E	D	E	…
5	B	D	C	D	E	…
6	D	C	B	D	E	…
7	B	A	D	D	E	…
8	E	E	C	A	E	…
9	C	B	A	A	D	…
10	E	D	C	A	B	…
…	…	…	…	…	…	…

Note that a student with a score 78 in College Chinese can only obtain a grade E. However, he can obtain a grade B in Discrete Math even with a score 70. This is because the overall achievement of Discrete Mathematics is lower than that of College Chinese. Although a student has scored 78 in the College Chinese exam, his rank may be lower. On the contrary, if a student only got 70 marks in Discrete Mathematics, he could get a higher grade, that is B.

B. Methodology

After the data preprocessing, R [10] is used to conduct the experiment and apriori algorithm from package arules [11] is applied to perform the association rule mining.

Firstly, the grades as shown in Table II, rather than the raw scores, in 34 courses for 100 students are put together into a dataframe, e.g., gradeFrame. Then, before the apriori performing association rules mining in R, the dataframe is converted into the form of transaction, e.g., trans, using the following instruction:

```
trans <- as(gradeFrame, "transactions")
```

Once the transaction data sets are generated, the association rules can be simply obtained by calling the function apriori in R:

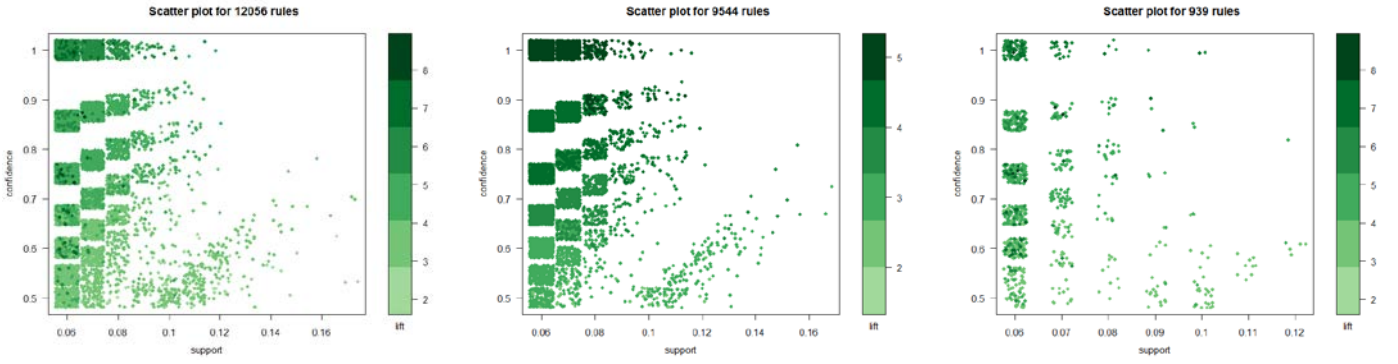
```
rules <- apriori(trans, parameter = list(support = min_sup, confidence = min_conf, minlen=len))
```

where min_sup and min_conf are the predefined minimum support and minimum confidence, respectively. In this work, the parameters min_sup and min_conf are set to 0.06 and 0.5, respectively. Since Apriori only creates rules with one item in the right hand side and the parameter len is the minimal number of items per itemset with default value 1. To avoid the left hand side of the rules being empty sets len is set to 2.

IV. Experimental Results and Discussion

In the experiment, Apriori algorithm is applied on the students' grade data to process the association rule mining. 12056 rules are generated and the overall support-confidence scatter plot is shown in Fig. 1(a). In order to study the rules that

lead to high and low grades, the subset rulesA and rulesE with grade A and E on the right hand side are then filtered out from the 12056 rules. Correspondingly, 9544 and 939 rules are obtained as shown in Fig. 1(b-c). Next, we will focus on the analysis of these filtered rules with a view to try to enhance students' strengths and avoid poor grades.



(a) (b) (c)

Fig. 1. The support-confidence scatter plot of the association rules. (a) All 12056 association rules. (b) 9544 rules with grade A on the right hand side. (c) 939 rules with grade E on the right hand side.

In order to investigate the higher-grade rules, the rules A (the set of rules with grade A on the right hand side) is sorted in descending order of support, confidence and lift. Table III shows 10 typical association rules with higher confidence, support and lift. As can be seen from Table III, there is a strong correlation between some English-related courses, such as the English course of four semesters and the Specialty English. In addition, Good grades in English are also helpful for some professional course, i.e., Compiling Principle, to achieve good performance. Moreover, there is also a strong correlation between hardware-related courses or between courses and their prerequisites, such as Embedded Systems and Interface, Operating System and Compiling Principle. Further, the foundation of Discrete Mathematics is also important for some basic professional courses, i.e., Programming Language and Introduction of Computer Science, to achieve good marks. in order to achieve good results, we should lay a good foundation in English and some professional basic courses in the freshman stage, in order to achieve good results, students majoring computer science should lay a good foundation in English courses and some major basic courses in the freshman stage.

TABLE III. 10 TYPICAL RULES WITH GRADE A ON RIGHT SIDE

No.	rules_lhs	rules_rhs	support	confidence	lift	count
1	{CollegeEnglish2=A}	{CollegeEnglish3=A}	0.16	0.800	3.333	16
2	{CollegeEnglish1=A}	{CollegeEnglish3=A}	0.15	0.750	3.125	15
3	{EmbeddedSystems=A}	{Interface=A}	0.14	0.700	3.333	14
4	{CollegeEnglish4=A}	{SpecialtyEnglish=A}	0.15	0.682	3.099	15
5	{OperatingSystem=A}	{CompilingPrinciple=A}	0.14	0.667	2.778	14
6	{ObjectOriented=A}	{SoftwareEngineer=A}	0.13	0.650	3.095	13
7	{CollegeEnglish1=A}	{SpecialtyEnglish=A}	0.13	0.650	2.955	13
8	{DiscreteMath=A}	{ProgrammingLanguage=A}	0.13	0.650	2.708	13
9	{CollegeEnglish2=A}	{CompilingPrinciple=A}	0.13	0.650	2.708	13
10	{DiscreteMath=A}	{IntroductionOfCS=A}	0.13	0.650	2.500	13

TABLE IV. 10 TYPICAL RULES WITH GRADE E ON RIGHT SIDE

No.	rules_lhs	rules_rhs	support	confidence	lift	count
1	{DigitLogic=E}	{DatabasePrinciples=E}	0.10	0.833	4.167	10
2	{Interface=E}	{CompilingPrinciple=E}	0.12	0.800	4.000	12
3	{DigitLogic=E}	{SoftwareEngineer=E}	0.09	0.750	3.947	9
4	{DataStructure=E}	{DatabasePrinciples=E}	0.08	0.727	3.636	8
5	{DataStructure=E}	{Interface=E}	0.07	0.636	4.242	7
6	{DataStructure=E}	{SoftwareEngineer=E}	0.07	0.636	3.349	7
7	{ComputerComposition=E}	{SingleChip=E}	0.07	0.636	3.349	7
8	{DataStructure=E}	{AssemblyLanguage=E}	0.07	0.636	3.182	7
9	{DataStructure=E}	{Algorithm=E}	0.07	0.636	3.182	7
10	{ComputerComposition=E}	{CompilingPrinciple=E}	0.07	0.636	3.182	7

Similarly, the association rules with higher support, confidence and lift are selected from rulesE (the set of rules with grade E on the right hand side) and presented in Table IV. As shown in Table IV, there is a strong correlation between Digit Logic and Database Principles, Interface and Compiling Principle, Computer Composition and Single Chip, Computer

Composition and Compiling Principle. Specially, Students with poor grades should attach importance to Data Structure. Since students with poor data structure scores are likely to get low grades in many professional courses such as Database Principles, Interface, Software Engineer, Assembly Language and Algorithm.

V. Conclusion

The wide use of information technology in colleges and universities has brought a lot of data. Data mining technology should be adopted in order to discover the hidden rules and valuable information from the big data. In this work, we use the apriori algorithm to analyze the performance data of 100 students majoring in computer science and mine the association rules in order to discover the correlation between different courses. The methods and results of this work are expected to provide reference for the study and teaching of the college courses.

Acknowledgment

This work is supported in part by the Natural Science Foundation of Fujian Province, China (Grant No. 2015J01663), by the Key Project of Quanzhou City Science and Technology Program (No.2015Z121, 2014Z134), by the 13th five-year planning project of Fujian education science (FJJKCG16-366), and by the Quanzhou Normal University Scientific Research Initiative Foundation.

References

[1] Baker R S, Yacef K. The State of Educational Data Mining in 2009: A Review and Future Visions[C]. educational data mining, 2009, 1(1): 3-17.

[2] Asif R, Merceron A, Ali S A, et al. Analyzing undergraduate students' performance using educational data mining[J]. Computers in Education, 2017: 177-194.

[3] Altaher A and BaRukab O. Prediction of Student's Academic Performance Based on Adaptive Neuro-Fuzzy Inference [J]. IJCSNS International Journal of Computer Science and Network Security, 2017, 17(1)

[4] Ahmed A M, Rizaner A, Ulusoy A H, et al. Using data Mining to Predict Instructor Performance[J]. Procedia Computer Science, 2016: 137-142.

[5] Hamsa H, Indiradevi S, Kizhakkethottam J J, et al. Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm[J]. Procedia Technology, 2016: 326-332.

[6] Elbadrawy A, Polyzou A, Ren Z, et al. Predicting Student Performance Using Personalized Analytics [J]. IEEE Computer, 2016, 49(4): 61-69.

[7] Harwati, Alfiani A P, Wulandari F A. Mapping Student's Performance Based on Data Mining Approach (A Case Study) [J]. Agriculture & Agricultural Science Procedia, 2015, 3:173-177.

[8] Han J . Data Mining: Concepts and Techniques[M]. Morgan Kaufmann Publishers Inc. 2005.

[9] Agrawal R , Srikant R . Fast algorithms for mining association rules[M]// Readings in database systems (3rd ed.). Morgan Kaufmann Publishers Inc. 1996.

[10] The R Project for Statistical Computing. https://www.r-project.org/.

[11] GitHub - mhahsler/arules: Mining Association Rules and Frequent Itemsets with R. https://github.com/mhahsler/arules.