Association Rule Mining for Selecting Proper Students to Take Part in Proper Discipline Competition: A Case Study of Zhejiang University of Finance and Economics

https://doi.org/10.3991/ijet.v13i03.8382

Xiaoling Huang, Yangbing Xu, Shuai Zhang(✉), Wenyu Zhang
Zhejiang University of Finance and Economics, Hangzhou, China
zhangshuai@zufe.edu.cn

Abstract—In recent years, the educational issues have attracted more and more researchers' and teachers' attention. On the other hand, the development of data mining technology, provides a new method to extract the useful information from the complex educational data. In order to increase the chance of students to be awarded in discipline competition, it is better to select the proper students to take part in the proper discipline competition. Therefore, in this study, we collect the information of 164 undergraduate students as a case study. All students majored in Software Engineering in Zhejiang University of Finance and Economics. The Apriori algorithm with group strategy is used to find the relationship between the students' courses scores and competition awards. According to the results of association rule mining, we find that the students with higher scores of *C# Development*, *Object-Oriented*, *Internet Web Design*, *Data Structure(C#)*, and *Basic Programming* will have a higher probability to be awarded in the competition.

Keywords—association rule mining, Apriori algorithm, R programming, discipline competition

1 Introduction

In recent years, the issue of how to extract the useful information from educational data to promote the students' overall development has attracted a lot of researchers' and teachers' attention. In addition, the development of computer science and data mining technology provides a new method to find the relationship between students' courses scores and other features.

In order to select proper students to take part in the proper discipline competition so that the probability for them to be awarded in the discipline competition becomes higher, in this article, we aim to find the relationship between the students' courses scores and discipline competition awards. Because the R Programming, as a free platform, is a powerful tool for data mining, we adopt R 3.4.0 to do association rule mining from the data of 164 undergraduate students who majored in Software Engineering in Zhejiang University of Finance and Economics.

The remainder of this article is organized as follows. In section 2, some related studies are reviewed. In section 3, we introduce the data used in case study. The partial results of association rule mining are analyzed in section 4. And the conclusion and further work are introduced in the last section.

2 Related work

In recent years, the data mining technologies, such as clustering algorithm, classification algorithm, and prediction algorithm, have been applied in many domains, including educational area. For example, Huang and Zhang [1] presented the online teaching idea through the data mining technology. Zhu [2] predicted the engineering students' performance by establishing a pre-control warning model. Sen et al. [3] used support vector machine to predict the secondary education placement-test scores. Gómez-Rey et al. [4] proposed an ordinal regression model to predict the performance of students with different grades. In our previous work, Zhang et al. [5] used the support vector machine to predict the student's graduation thesis grade. Sael et al. [6] analyzed the relationship between students' profiles and students' performance through clustering algorithm and association rule mining.

The Apriori algorithm [7] is one of the most popular algorithms for association rule mining, which has been applied in many areas. For example, Chalmers et al. [8] used Apriori algorithm to find the association between preseason scores and injuries of football players. Zhang [9] used Apriori algorithm in library personalized service field. However, because Apriori algorithm generates a large set of candidate rules, it is sometimes regarded as an inefficient algorithm. And combining group strategy with Apriori algorithm [10] is a popular direction to improve the efficiency of Apriori algorithm.

To best of our knowledge, association rule mining technology has not been used to find the relationship between students' course scores and the discipline competition awards in the literature. In this work, in order to select the proper students to take part in the proper discipline competition, we use the Apriori algorithm with group strategy to extract the useful information from 164 undergraduate students who majored in Software Engineering in Zhejiang University of Finance and Economics.

3 Data preprocessing, transformation, and association rule mining

In this section, we aim to explore the relationship between the students' courses scores and the discipline competition awards. The original data used in this case study contains the information of 164 undergraduate students who majored in Software Engineering in Zhejiang University of Finance and Economics, including 54 courses scores and the discipline competition awards the students have achieved.

3.1 Data Preprocessing and Transformation

In order to protect the students' private information, we remove their ID No.s and their names. Secondly, to increase the accuracy of association rule mining, we remove some courses that few students take part in, such as the courses of second major for Software Engineering, remove some courses that all of students got the similar scores, such as *Military Training*, *Graduation Project*, and so on, and remove the information of students that contain few attributes. According to the student handbook of Zhejiang University of Finance and Economics, the students who got high scores of English in the college entrance examination can be exempted from examination of *English(2)*, so we fill the NA value of the *English(2)* by the maximum score value of this course. Then, we fill the NA values of other courses by the average score of its corresponding course, and change the records of abandoning the examination, cancelling the examination, and cheating on examination to the value of zero. Then, we combine some similar courses into one course and assign its value with corresponding average scores. For example, we combine *English(2)*, *English(3)*, and *English(4)* into *College English*. In order to make the data fit the data type of association rule mining in R, we divide the scores into four grades: Fail (score < 60), Pass (score >=60 and score <80), Good (score >=80 and score < 90), and Excellent (score >=90).

After the data preprocessing and transformation, we obtain the data containing the information of 159 students with 25 courses grades and the discipline competition awards they have obtained. Table 1 shows the partial data that are used for association rule mining.

Table 1. Partial Normalized Data in the Case Study

Awarded in A Discipline Competition	Awarded in New Talent Competition	Awarded in Service Outsourcing Competition	Awarded in Programming Competition	Basic Accountancy	Computer Composition and Architecture	Basic Programming	Object-Oriented
No	No	No	No	Fail	Pass	Pass	Pass
No	No	No	No	Pass	Pass	Pass	Pass
No	No	No	No	Pass	Pass	Pass	Pass
Yes	No	No	Yes	Pass	Pass	Excellent	Excellent
No	No	No	No	Good	Good	Pass	Pass
No	No	No	No	Pass	Pass	Good	Good
No	No	No	No	Pass	Pass	Good	Pass
Yes	Yes	Yes	No	Pass	Pass	Pass	Good
No	No	No	No	Pass	Pass	Good	Pass
Yes	Yes	Yes	No	Pass	Pass	Pass	Pass
No	No	No	No	Pass	Pass	Pass	Good
No	No	No	No	Pass	Pass	Pass	Pass
No	No	No	No	Pass	Pass	Good	Pass
Yes	Yes	No	No	Good	Good	Pass	Good
No	No	No	No	Fail	Pass	Pass	Fail

3.2 Association Rule Mining

In this article, we used R 3.4.0 and the Apriori algorithm with group strategy for further association rule mining. Firstly, we count the frequency of each item after finding the frequent 1-itemsets, then we divide the items into several groups based on their frequency and label the groups according to the items frequency. Therefore, we can find the frequent k-itemsets based on some groups, whose labels values are bigger than k, rather than the whole database. The pseudo code of Apriori algorithm with group strategy is shown in Algorithm 1. The association rule has the form of A=>B, where $A \in I$, $B \in I$, $A \cap B = \varnothing$, and I is the set of items. The support and confidence are two key parameters for association rule mining. The support of the association rule A=>B is the probability that both A and B occur in all observations. Thus, the support of the association rule A=>B is equal to the support of the association rule B=>A. The confidence of association rule A=>B is the probability that both A and B occur in the observations that contain the item A. The mathematic equation of support and confidence in association rule mining are showed in equation (1) and equation (2), respectively [11]. In this article, the item A is the set of course grades and the item B is the discipline competition awards.

$$\text{Support}(A=>B) = P(A \cap B), \tag{1}$$

$$\text{Confidence}(A=>B) = P(B|A). \tag{2}$$

According to the definition of support and confidence, the rule with too small support or confidence is meaningless, because it has a high probability occurred by accident. After a series of test, we set the threshold of support and confidence as 0.07 and 0.8, respectively. In addition, the lift is another index of association rule, which is calculated in equation (3) [11]. According to the equation (3), there is no correlation between A and B, when the lift of association rule A=>B equals one. Thus, we extract the subset from the rules that the lift is bigger than one. Finally, we sort the subset by the lift decreasing.

$$\text{Lift}(A=>B) = P(B|A) / P(B). \tag{3}$$

Before doing the experiments of association rule mining, we firstly observe the item frequency of students' courses grades. Fig. 1 shows the top 10 frequent items, and the most frequent items shown indicate that more than 70% students got a pass grade in *College English*.

Algorithm 1. The Pseudo Code of Apriori Algorithm

```
Input: The cleaned data and the thresholds of parameters
Output: Some Rules
L1 = find_frequent_1-itemsets(D);
group_strategy();
for (k=2;Lk-1 ≠Φ ;k++) {
    Ck = Apriori_process(Lk-1 ,min_sup);
    for each transaction t ∈ D {
        Ct = subset(Ck,t);
        for each candidate c ∈ Ct
        c.count++;
    }
    Lk ={c ∈ Ck|c.count≥min_sup}
}
return L = k ∪ Lk;
```

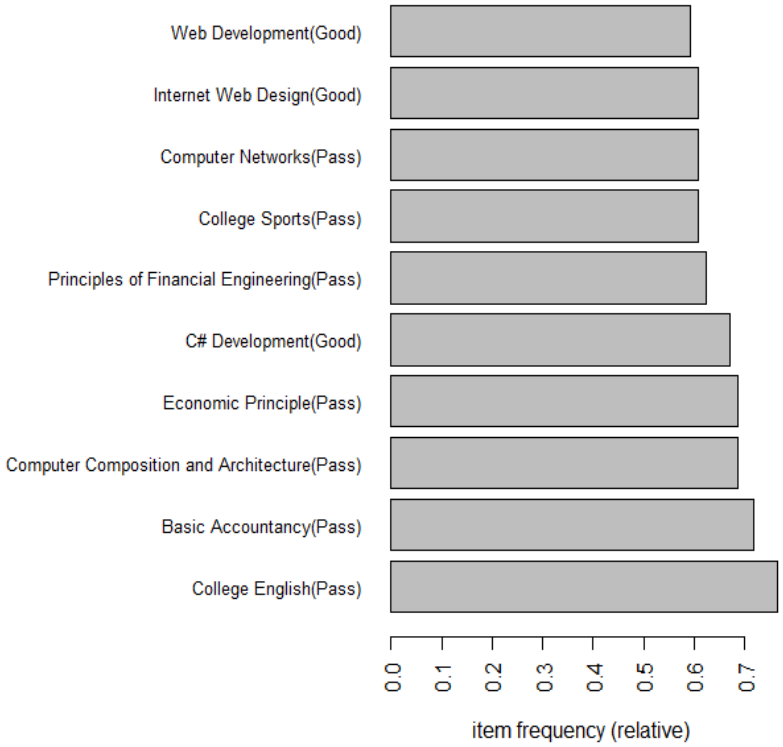

Fig. 1. The top 10 most frequent 1-itemsets

4 Result and analysis

Figs. 2-5 are four scatter plots of association rules of students being awarded in discipline competitions. Each point in these figures represents one association rule, and its color represents the value of lift. The X axis represents the value of support, and the Y axis represents the value of confidence. As shown in Fig. 2 and Fig. 4, association rules are grouped with the low levels of support and lift, the support value of most rules are lower than 0.25, and there are few rules whose supports are higher than 0.4. Fig. 3 and Fig. 5 show that the confidences of most association rules are bigger than 0.85, and their lifts are smaller than 2, which are probably caused by the fact that most of students are not awarded in New Talent Competition and Programming Competition. According to Fig. 3 and Fig. 5, there are a few rules with high lift but low support, which means that it's difficult for students to be awarded in the New Talent Competition or Programming Competition, and only a few students with high scores of curriculums have the chances to be awarded.

Figs. 6-9 are four connection graphs of association rules of students being awarded in discipline competitions. Because there are over thousands of rules for each competition, we choose the top 10 rules of students being awarded in competitions and the top 10 rules of students not being awarded in competitions as the illustrative examples. If the number of rules of students being awarded in discipline competitions is less than 10, we will select all of them, and so do the rules of students not being awarded in discipline competitions. The circles in these graphs represent the association rules. The circle color represents the lift of the rule, and the deeper the color the higher the lift. The circle size represents the support, and the bigger the size the higher the support. As shown in Figs. 6-9, the rules of students being awarded in discipline competitions have smaller supports but higher lifts than the rule of students being not awarded in discipline competitions. This phenomenon is probably caused by the fact that most students did not take part in the discipline competition or were not awarded in a discipline competition. According to Figs. 6-9, the students with high scores of curriculum, especially the curriculum related to programming, have more chances to be awarded in discipline competitions than other students. And as shown in Figs. 6-9, some students who were awarded in discipline competition only got a Pass grade in the curriculum related to finance. This phenomenon is probably caused by the fact that these students have spent too much time on discipline competitions but spent too less time on the curriculum related to finance.

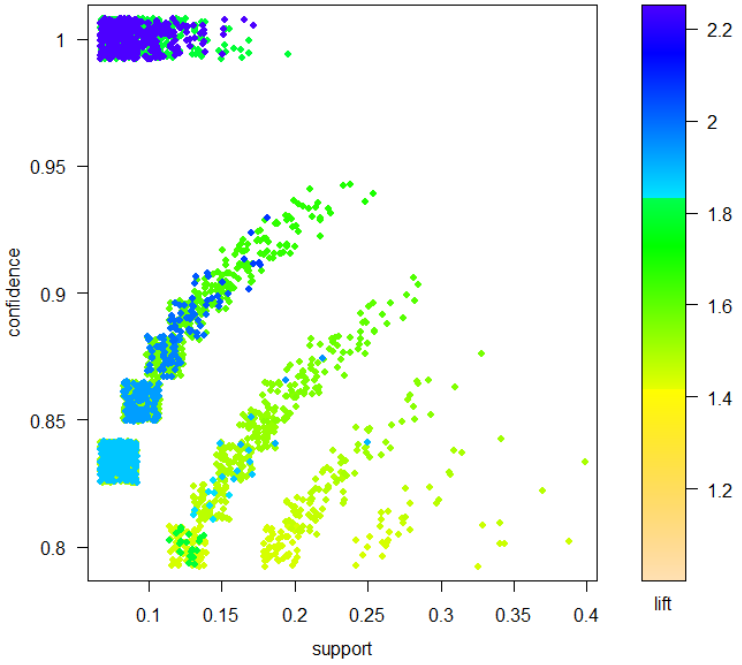

Fig. 2. Rules of students being awarded in a Discipline Competition

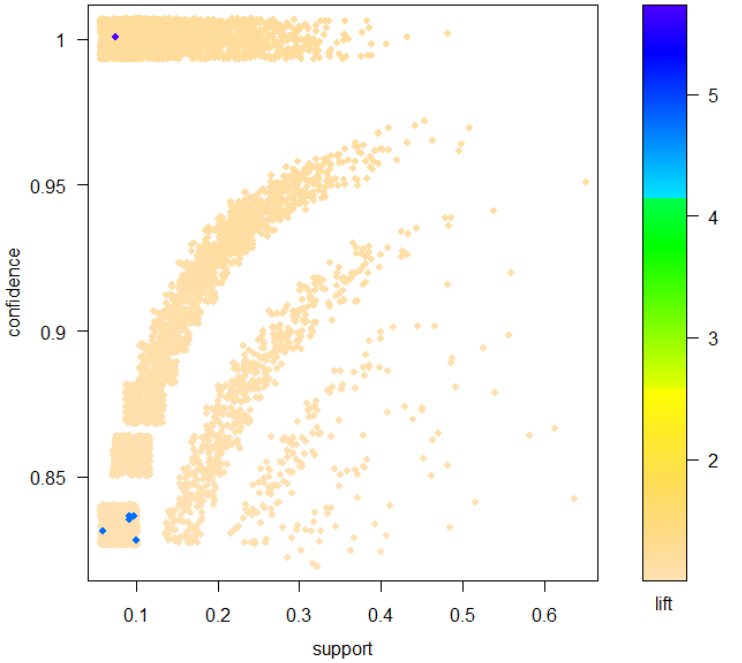

Fig. 3. Rules of students being awarded in New Talent Competition

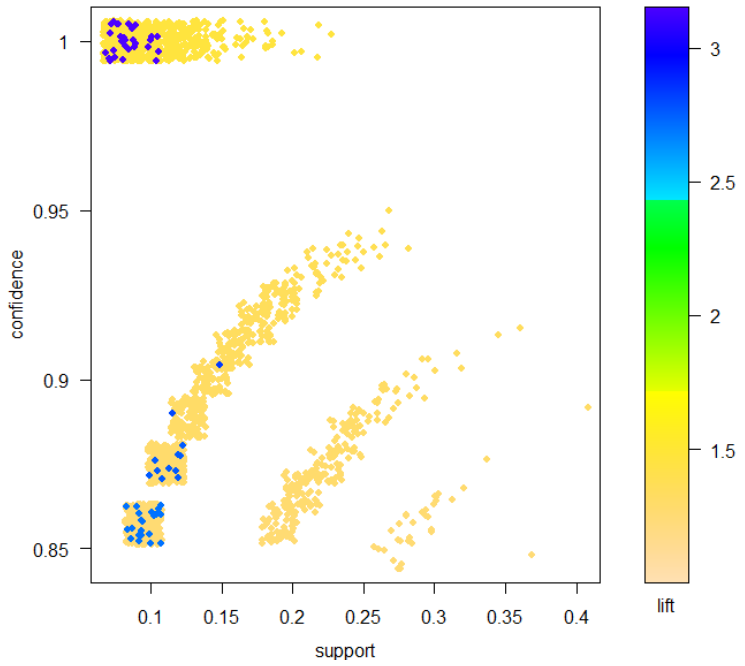

Fig. 4. Rules of students being awarded in Service Outsourcing Competition

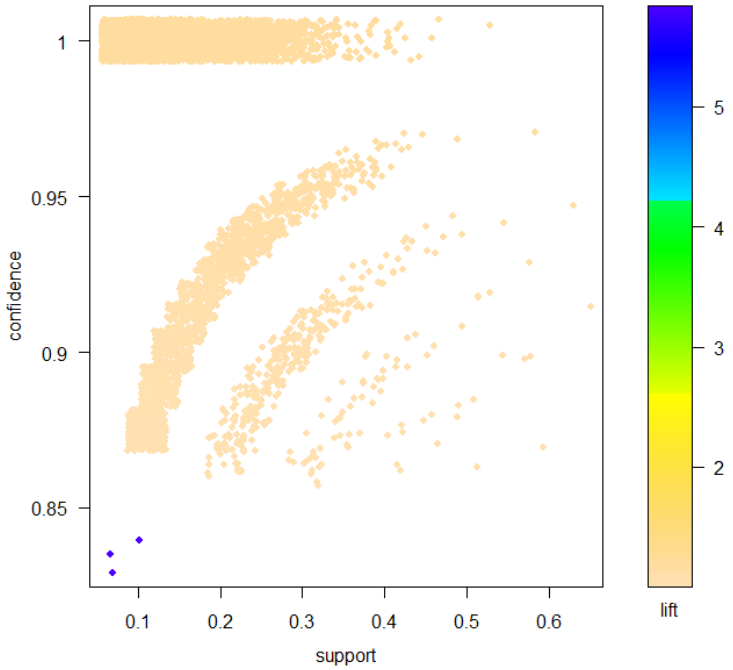

Fig. 5. Rules of students being awarded in Programming Competition

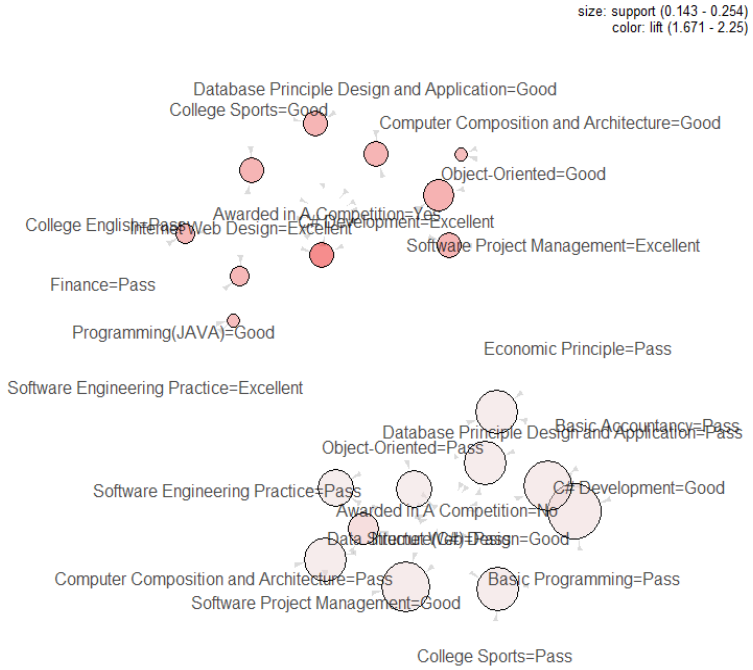

Fig. 6. Graph for rules of students being awarded in a Discipline Competition

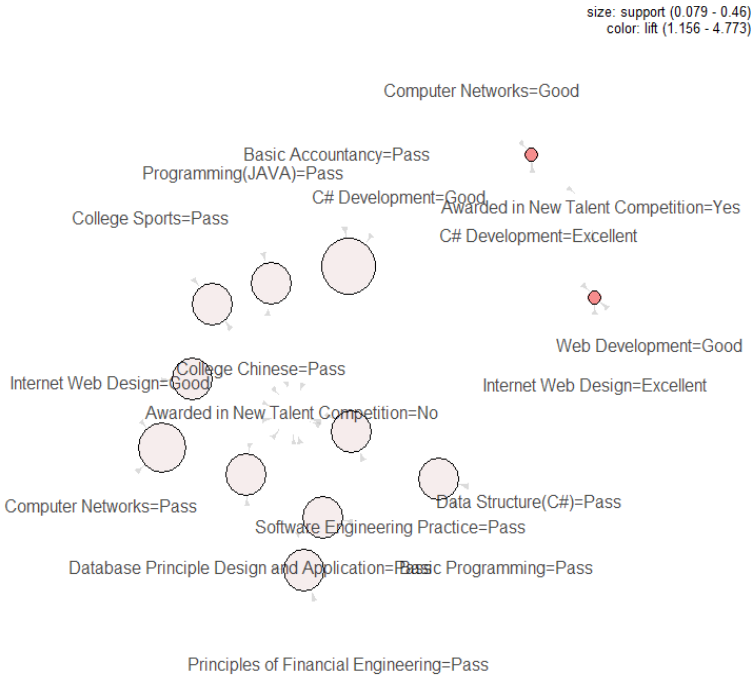

Fig. 7. Graph for rules of students being awarded in New Talent Competition

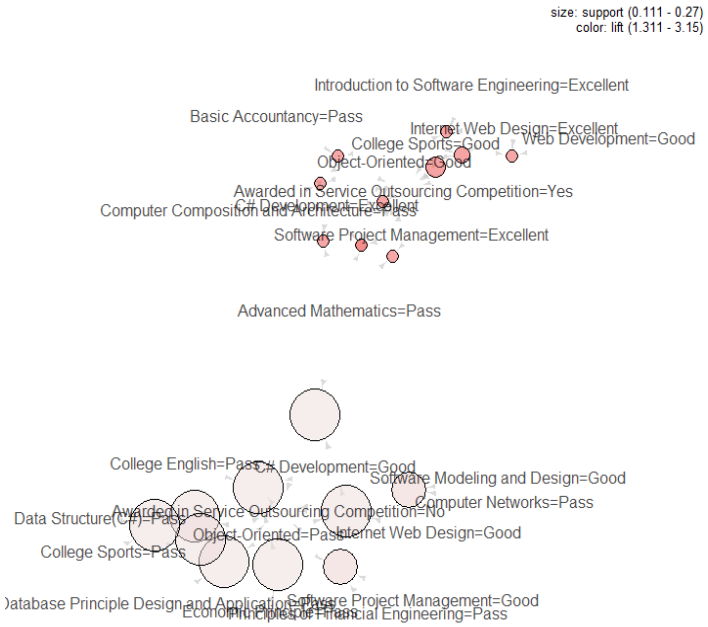

Fig. 8. Graph for rules of students being awarded in Service Outsourcing Competition

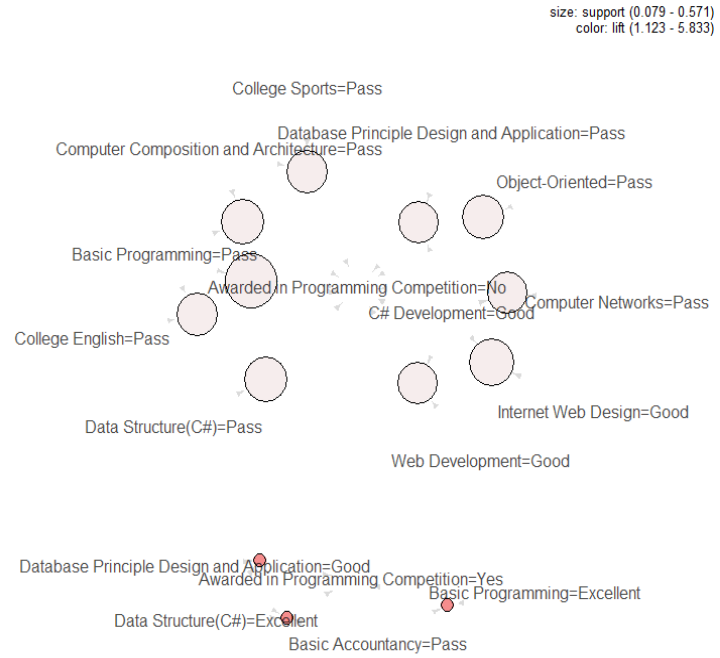

Fig. 9. Graph for rules of students being awarded in Programming Competition

The partial results of relationship between the courses grades and discipline competition awards are showed in Tables 2-5. The association rule mining reveals the relationship between items and objects. For example, the first rule in Table 2 identifies that selecting the student with excellent grade of *C# Development* and excellent grade of *Internet Web Design* to take part in a discipline competition has a high probability to be awarded. This rule can be explained with that the *C# Development* and *Internet Web Design* courses not only train students' development capability, but also train students' programming capability. The rules in Tables 3 and 4 reveal that a student with excellent grade of *C# Development* or good grade of *Object-Oriented* has more chances to be awarded in New Talent or Service Outsourcing Competitions. The rules in Table 5 identify that the student with the excellent grade of *Data-Structure(C#)* or *Basic Programming* courses tend to be awarded in Programming Competition. This can be explained with that these courses are the footstone of programing. According to the rules shown in Tables 2-5, the students with higher grade of course have a higher probability to be awarded in the discipline competitions than others.

Table 2. The Partial Results of the Association Rules of Students Being Awarded in a Discipline Competition

Rules	Support	Confidence	Lift
{*C# Development* = Excellent, *Internet Web Design* = Excellent} => {*Awarded in A Discipline Competition* = Yes}	0.174603	1	2.25
{*Object-Oriented* = Good, *C# Development* = Excellent} => { *Awarded in A Discipline Competition* = Yes}	0.190476	0.923077	2.076923
{*C# Development* = Excellent, *Software Project Management* = Excellent} => { *Awarded in A Discipline Competition* = Yes}	0.174603	0.916667	2.0625
{*Object-Oriented* = Pass, *Data Structure(C#)* = Pass, *Software Project Management* = Good} => { *Awarded in A Discipline Competition* = No}	0.190476	1	1.8
{*Basic Programming* = Pass, *Database Principle Design and Application* = Pass, *C# Development* = Good} => { *Awarded in A Discipline Competition* = No}	0.253968	0.941176	1.694118
{*Data Structure(C#)* = Pass, *Software Project Management* = Good, *Internet Web Design* = Good} => { *Awarded in A Discipline Competition* = No}	0.238095	0.9375	1.6875

Table 3. The Partial Results of the Association Rules of Students Being Awarded in New Talent Competition

Rules	Support	Confidence	Lift
{*C# Development* = Excellent, *Computer Networks* = Good} => {*Awarded in New Talent Competition* = Yes}	0.079365	0.833333	4.772727
{*C# Development* = Excellent, *Internet Web Design* = Excellent, *Web Development* = Good} => { *Awarded in New Talent Competition* = Yes}	0.079365	0.833333	4.772727
{*Software Engineering Practice* = Pass, *College Chinese* = Pass} => { *Awarded in New Talent Competition* = No}	0.333333	1	1.211538
{*Basic Accountancy* = Pass, *C# Development* = Good} => { *Awarded in New Talent Competition* = No}	0.460317	0.966667	1.171154
{*Internet Web Design* = Good, *Computer Networks* = Pass} => { *Awarded in New Talent Competition* = No}	0.396825	0.961538	1.164941
{*Basic Programming* = Pass, *Data Structure(C#)* = Pass} => { *Awarded in New Talent Competition* = No}	0.333333	0.954545	1.156469

Table 4. The Partial Results of the Association Rules of Students Being Awarded in Service Outsourcing Competition

Rules	Support	Confidence	Lift
{*C# Development* = Excellent, *Software Project Management* = Excellent, *Advanced Mathematics* = Pass} => {*Awarded in Service Outsourcing Competition* = Yes}	0.111111	1	3.15
{*Object-Oriented* = Good, *College Sports* = Good} => { *Awarded in Service Outsourcing Competition* = Yes}	0.142857	0.9	2.835
{*Object-Oriented* = Good, *Internet Web Design* = Excellent, *College Sports* = Good} => { *Awarded in Service Outsourcing Competition* = Yes}	0.126984	0.888889	2.8
{*Object-Oriented* = Pass, *Software Project Management* = Good, *Principles of Financial Engineering* = Pass} => { *Awarded in Service Outsourcing Competition* = No}	0.206349	1	1.465116
{*C# Development* = Good, *Software Modeling and Design* = Good, *Computer Networks* = Pass} => { *Awarded in Service Outsourcing Competition* = No}	0.206349	1	1.465116
{*Object-Oriented* = Pass, *Data Structure(C#)* = Pass} => { *Awarded in Service Outsourcing Competition* = No}	0.269841	0.894737	1.310894

Table 5. The Partial Results of the Association Rules of Students Being Awarded in Programming Competition

Rules	Support	Confidence	Lift
{*Database Principle Design and Application* = Good, *Data Structure(C#)* = Excellent} => {*Awarded in Programming Competition* = Yes}	0.079365	0.833333	5.833333
{*Basic Accountancy* = Pass, *Basic Programming* = Excellent} => { *Awarded in Programming Competition* = Yes}	0.079365	0.833333	5.833333
{*Basic Accountancy* = Pass, *Database Principle Design and Application* = Good, *Data Structure(C#)* = Excellent} => { *Awarded in Programming Competition* = Yes}	0.079365	0.833333	5.833333
{*Data Structure(C#)* = Pass} => { *Awarded in Programming Competition* = No}	0.460317	1	1.166667
{*Object-Oriented* = Pass} => { *Awarded in Programming Competition* = No}	0.444444	1	1.166667
{*Basic Programming* = Pass} => { *Awarded in Programming Competition* = No}	0.571429	0.972973	1.135135

5 Conclusion and further work

In this article, we aim to find the relationship between the courses scores and discipline competition awards. The rules will help to increase the probability for students to be awarded in discipline competitions. We extracted the useful information from 164 undergraduate students majored in Software Engineering in Zhejiang University of Finance and Economics to model the association rule mining. The raw data is preprocessed and transformed into the format that is suitable for Apriori algorithm with group strategy in R Programming. Then we obtained the data that contains 159 students with 25 courses grades and the corresponding discipline competition awards. Finally, according to the result obtained by the association rule mining, which are showed in Tables 2-5, partially, we found that the *C# Development*, *Internet Web Design*, *Object-Oriented*, *Data Structure(C#)*, and *Basic Programming* courses make a great influence on the probability for students to be awarded in discipline competitions.

However, there are still some limitations in this article. For example, because the Software Engineering is a relatively new major in Zhejiang University of Finance and Economics, we only collect the information of 164 students. The limited data set affects the accuracy of association rule mining. In this article, we only analyze some representative rules, but the remained rules will be analyzed in our further work. In addition, we can combine other data mining technologies with association rule mining to find the meaningful relationship with higher accuracy and efficiency.

6 Acknowledgment

This work was supported by National Natural Science Foundation of China (No.51475410), National Social Science Foundation of China (No.17BGL047), and Zhejiang Provincial Higher Education Reform Program of China (No. JG20160109).

7 References

[1] Hung, J.L., Zhang, K. (2008). Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching. MERLOT Journal of Online Learning and Teaching, 4(4): 426–437.

[2] Zhu, K. (2014). Research based on data mining of an early warning technology for predicting engineering students' performance. World Transactions on Engineering and Technology Education, 12(3): 572–575.

[3] Şen, B., Uçar, E., Delen, D. (2012). Predicting and analyzing secondary education placement-test scores: A data mining approach. Expert Systems with Applications, 39(10): 9468–9476. https://doi.org/10.1016/j.eswa.2012.02.112

[4] Gómez-Rey, P., Fernández-Navarro, F., Barberà, E. (2015). Ordinal regression by a gravitational model in the field of educational data mining. Expert Systems, 33(2): 161–175. https://doi.org/10.1111/exsy.12138

[5] Zhang, W.Y., Zhang, S.X., Zhang, S. (2015). Predicting the graduation thesis grade using SVM. International Journal of Intelligent Information Processing, 5(3): 60–66.

[6] Sael, N., Marzak, A., Behja, H. (2013). Multilevel clustering and association rule mining for learners' profiles analysis. International Journal of Computer Science Issues, 10(3): 188–194.

[7] Agrawal, R., Imieliński, T., Swami, A. (1993). Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., ACM SIGMOD Record, May 26–28, 22(2), pp. 207–216. https://doi.org/10.1145/170035.170072

[8] Chalmers, S., Fuller, J.T., Debenedictis, T.A., Townsley, S., Kynagh, M., Gleeson, C., et al. (2017). Asymmetry during preseason functional movement screen testing is associated with injury during a junior Australian football season. Journal of Science and Medicine in Sport, 20(7): 653–657. https://doi.org/10.1016/j.jsams.2016.12.076

[9] Zhang, C.Y. (2017). Research on library personalized service based on Apriori algorithm. Agro Food Industry Hi-tech, 28(1): 2555–2559.

[10] Si, G.Z., Liu, Y. (2013). Application and research of grouping Apriori algorithm in library system (in Chinese). Microprocessors: 35–38.

[11] Agrawal, R., Srikant, R. (1994). Fast algorithms for mining association rules. 20th VLDB Conference, Santiago, Chile, September 12-15, pp. 487–499.

8 Authors

Xiaoling Huang is currently a teaching secretary at the School of International Education, Zhejiang University of Finance and Economics, China. She received her M.A. degree in education from Zhejiang University, China, in 2014. Her research interests include data mining, supply chain management, and E-government.

Yangbing Xu is currently a postgraduate student and working towards his M.S. degree at Zhejiang University of Finance and Economics, China. His current research interest is data mining and bibliometrics.

Shuai Zhang is currently a full-time professor at the School of Information, Zhejiang University of Finance and Economics, China. He received his Ph.D degree in mechanical engineering from Zhejiang University, China in March 2005. He has published more than 30 papers in international journals in the recent ten years, covering supply chain management, business intelligence, data mining, E-government, and manufacturing informatization.

Wenyu Zhang is currently a full-time professor at the School of Information, Zhejiang University of Finance and Economics, China. He received his B.S. degree from Zhejiang University, China in 1989, and his Ph.D. degree from Nanyang Technological University, Singapore in 2002. He has published more than 40 papers in international journals and more than 20 papers in international conference proceedings in the recent ten years, covering a wide range of digital information management, especially supply chain management, digital library, concurrent engineering, distributed manufacturing, business intelligence, business analytics, data mining, multi-agent technology, and semantic Web.

Article submitted 07 February 2018. Final acceptance 23 February 2018. Final version published as submitted by the author.