Research on Evaluation of College Students' Professional Ability Based on K-means Clustering

ChuanShi Chen[1], *Kun* Liu[2,*], and *Kun* Ma[3]

[1-3]School of Information Science and Engineering, University of Jinan Jinan 250022, PR China Nanxinzhuang Road No. 336

Abstract. This paper presents a program to evaluate students' professional ability, which transforms the students and their teaching and practice activities during school into the form of topological graph, analyzes the similarity between student nodes, calculates teaching activities and students' professional ability through Apriori algorithm, and uses the K-means algorithm to cluster student data with different parameter sets based on different measurement goals. This paper analyzes the overall professional ability of students, compares cultivating differences among students in accordance with different professional ability, and finally gives the results of the analysis to facilitate teaching managers understand the distribution of students' professional ability to develop appropriate teaching plans.

1 Introduction

Improving the quality of education and cultivating overall-developed and high-quality talents has always been the core issue of continuous exploration and study in education. For a long time, the evaluation of teachers' teaching quality has been placed more emphasis while the programs evaluating the professional ability of student groups have not been sophisticated. At present, when evaluating higher education students, the mainstream evaluation method is qualitative analysis, that is, according to a certain weight, students' intellectual, moral education and additional reduction scores are weighted. With the popularization and implementation of credit system, this method is becoming unsatisfactory to reflect the internal relations among students and is more and more difficult to accurately evaluate the students.

2 Related research results

In recent years, a lot of domestic researches on the evaluation of students' comprehensive quality have been made, providing a lot of feasible solutions.

Literature[1] describes a method used to evaluate secondary school students' professional ability based on the APH-gray clustering. The program divides the evaluation index into three levels, including 13 influencing factors such as moral education, knowledge, quality and so on. It constructs and uses matrix to find the weight vector of these influencing factors and deals with students' data in many levels using APH-gray clustering. Literature[2] uses the method of fuzzy inquiries to compare the professional ability of thirty students in their college with the weighted scores. The main basis is the students' comprehensive evaluation results and their current moral scores. Literature [3] makes use of BP neural network to make a comprehensive evaluation of physical fitness of all the students in Nanchang University. The physical and artificial neural network models of male and female college students are respectively established, and the influence of various sports items on physical fitness of college students is analyzed. Literature [4] suggested original approach for Master Program in Software Engineering competence evaluation as a combination academic competences and professional competences from European Competence model (e-CF).

The above several literatures are the mainstream solutions to the problem at present. Literatures[1-3] adopt the score data as the main parameter of evaluation but neglect the differences of data from different school years. They only make comprehensive evaluations on students, lacking of analyses from different perspectives. The literature[4] is based on the European e-Competence Framework, the author added academic competences that reflect competences of the specific subjects of the proposed Software Engineering Master program.

The program proposed in this paper uses students' score rankings to analyze the students' ability. It

* Corresponding author: liukun@ujn.edu.cn

eliminates the differences of scores made by teachers in different school years. The program mainly analyzes the distribution of students in professional ability and the differences students made with different professional ability by using several commonly used data mining algorithms such as K-means, SimRank and Apriori algorithms. The K-means algorithm was first proposed by MacQueen in 1967 [5]. Its core idea is to divide the n sample data into k classes in n-dimensional Euclidean space, and then re-sample the samples by comparing the similarity between the cluster sample and the cluster center. The SimRank algorithm was first proposed by Glen Jeh and Jennifer Widom [6] in 2002 as a model that measures the similarity between any two nodes based on graph topology information. The Apriori algorithm was first proposed by Rakesh Agrawal in 1997 [7] and recursively finds the set of items with strong association rules in the data set.

3 Methods of evaluating college students' professional ability

The overall design method of this paper is shown in Figure 1. Using Apriori algorithm, we analyze the relationship between teaching activities and the cultivating of students' professional ability. According to the differences of the analyzed professional ability, we take different scores of students' teaching activities for clustering. Finally, we analyze students' professional ability based on the clustering results.

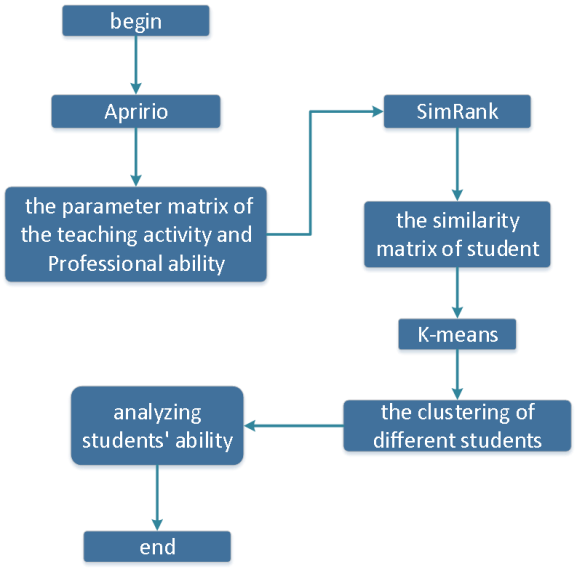

Fig 1. The overall steps of the analysis methods.

Figure 2 shows the detailed process of using data analysis for evaluating students' ability. In order to calculate the weight difference between different teaching activities, this paper proposes a SimRank algorithm which is based on weight. It calculates the node similarity according to the node weight in the topological map. The improved SimRank algorithm is shown in Formula 1.

$$S(a,b)=\begin{cases}1 & ,a=b\\ \dfrac{C}{|I(a)||I(b)|}S(i,j)\sum i\in I(a)W(i)\sum j\in I(b)W(j) & ,a\neq b\end{cases} \quad (1)$$

In Formula 1, S (a, b) represents the similarity between node a and node b; S(i, j) represents the similarity between node i and node j; I(a) and I(b) represent the adjacency point collection of node i and node j; W (i) and W (j) represent the weight of node i and node j. C∈ (0,1) is the attenuation factor, usually taking the value of (0.6,0.8).

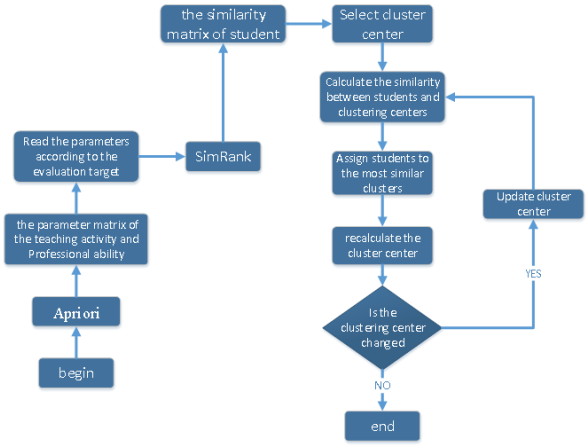

Fig 2. Detailed process of data analysis.

The process of detailed analysis is as follows:

Step 1: Record students' scores. The format is i = (K_1, K_2, K_3 ... K_n), in which K represents the achievements made by students when participating in teaching activities. The general course is divided into 5 segments; each reflects 20% students' scores. The practical course is divided into A, B, C, D, E level; respectively represents Excellent, Good, Medium, Pass and Failure.

Step 2: Abstract the relationship between teaching activity and students' ability from University of Jinan's Talent Cultivation Program. Then use Apriori algorithm to get the confidence between teaching activity Ki and students' ability A_j, and multiply them by the credits of teaching activity to get the weight $V_{k_i a_j}$ .Use Formula 2 to calculate $W_{k_i a_j}$ to make the sum of the weight is 1.

$$W_{k_i a_j}=\frac{V_{k_i a_j}}{\sum_1^n V_{k_i a_j}} \quad (2)$$

In Formula 2, $W_{k_i a_j}$ represents the weight of ith teaching activity K corresponding to the jth student's professional ability A; n represents the amount of

teaching activities corresponded by professional ability A_j.

Step 3. Use Formula 1 to calculate the similarity of the students in the cluster to generate the similarity matrix of student.

Step 4. Select K pieces of data randomly from the data set as the cluster center of the initial K pieces of clusters, denoted as C_1, C_2, C_3 ... C_k.

Step 5. Assign student i to the most similar cluster C_j ($1<j<p$).

Step 6. Calculate the average similarity of every record and among records in the new group, and select the one with the highest average similarity as the new cluster center.

Step 7. If the cluster center no longer changes, then clustering ends, otherwise return to Step 5 to continue calculating.

Step 8. Data Analysis.

(1) Analysis of overall ability of students: Calculate the professional scores of students in each cluster using Formula 3, and analyze students' advantaged ability in each cluster. Formula 3 is as follows:

$$G_{aj} = \sum_i^n G_{k_i} * W_{k_i a_j} \quad (2)$$

In the formula, Gaj represents the evaluation score of the professional ability A_j embodied by the j_{th} class of students; G_k represents the average score of the teaching activity k in the cluster, $W_{k_i a_j}$ represents the weight of teaching activity K_i corresponding to the professional ability A_j in the weight matrix; N is the number of teaching activities corresponding to professional ability A_j.

(2) Analysis on the differences of students' professional ability: choose different teaching activities as parameters to cluster students, and compare the differences of students in different professional directions.

4 Experimental analysis

4.1. Data Acquisition

From the Students Cultivation Database of University of Jinan, we collected 10106 pieces of course scores of 147 students who was made up of graduates of 2016 and 2017, and students majoring in Network Engineering. In the process of data preprocessing, the information of students who failed to go up to the next score or to pass the exams were excluded. Finally, 139 records were recorded, each record corresponding to about 20 courses that students took during their studies in school.

4.2. Analysis on the Support Matrix of Courses for Professional Ability

Apriori algorithm is used to mine the correlation information between teaching activities and professional ability. The support degree is 1 and the confidence level is 0.02. Finally, the weight matrix of teaching activity for students' professional ability is shown in Table 1.

In Table 1, A_1 ~ A_5 respectively represent five kinds of students' professional ability: Network Protocol Analysis, Design and Development of Network Application System, Network Engineering Planning, Network System Management, and Network System Security Guarantee. The following is the same.

Table 1. Weight Matrix of teaching activity-students' professional ability.

Course	A1	A2	A3	A4	A5
Fundamentals Programming		0.07			
Object-Oriented		0.08			
Data Structure		0.12			
Java Programming		0.08			
Principles of Computer	0.25		0.14	0.13	0.12
Principles of Database		0.09			0.12
Applied cryptography		0.11			
Network protocols	0.19		0.1	0.09	0.09

4.3. Clustering Analysis on College Students' Professional Ability

4.3.1 Analysis on the overall ability of student groups

All teaching activities are clustered and the weight is set as 1. Then we analyzed the overall professional ability of students. Table 2 shows the corresponding professional ability and evaluation scores of all kinds of students. The evaluation scores are calculated according to Formula 3.

Table 2. The proportion of all categories of students and ability scores.

Category	Professional Direction	Advantage Ability	Ratio	score
A	The Design of Network	A1	14%	78.29
B		A2	16%	83.45

C	Networking Project	A3	22%	82.2
D	The Management and Security of Network	A4	29%	80.41
E		A5	19%	76.79

According to the results of the clustering, among the sample students, Class B students get the score of 83.45, which is the most prominent, accounting for 16% of all students. These students are better at network application system design and development. Secondly the performance of Class C students is also prominent with a score of 82.2, accounting for 22% of all students. These students' advantaged ability is network engineering planning. Among the student groups, Class E students, who performed relatively poorly, scored 76.79. They are suitable for cyber-security work, accounting for 19% of all sample students.

4.3.2 The difference of students' professional ability cultivation

According to the parameter weight provided in step 3.2, using different teaching activities as clustering factors to cluster the data multiple times; we can analyze the difference of student groups in different professional ability, as shown in Figure 2.

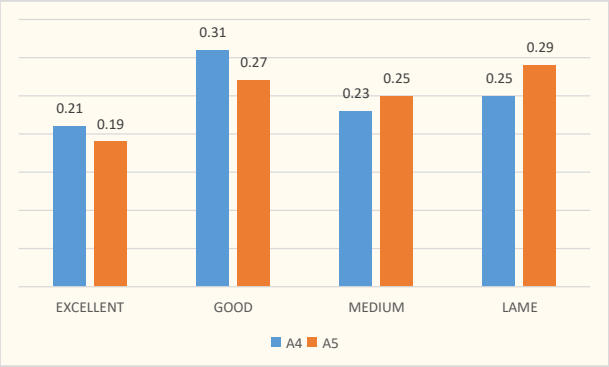

Fig 3. The difference of students' professional ability cultivation.

As can be seen from figure 3, the student population shows a trend of normal distribution in two kinds of professional abilities: A_4 and A_5. But compared with students' ability distribution in A4, the students who achieved Excellent in A_5 were 2% less than the former students and 4% more who achieved Lame. Combined with the information given in Table 2, we can judge that the student population performed better in network management.

5 Conclusion

This paper analyzes the professional ability of 139 students majoring in Network Engineering of University of Jinan, and explores the correlation among students' data by means of relevancy analysis, similarity analysis and cluster analysis. Firstly, this paper analyzes the correlation between the teaching activity and students' professional ability provided in the Talent Cultivation Program of University of Jinan to obtain the correlation matrix and the weight. Then, as the professional ability is different, this paper reads different teaching activities' results or scores as a parameter to analyze the similarity among students so as to form a student similar matrix. Finally, this paper uses K-means algorithm to cluster students' data and analyzes the differences of professional ability performed by students in different clusters. This paper analyzes the students' professional ability from two aspects. One is to analyze the overall situation of students, and to use the rank of all students' teaching activities as clustering parameter. According to the results of data clustering, the advantaged ability of the cluster is analyzed; according to the empirical formula, the scores of advantaged curriculum of the cluster are calculated. Thus the student ability distribution and the gap between the clustering students are analyzed. The second is to analyze students' certain professional ability. According to the teaching activity provided by the weight matrix of professional ability, the students' data is clustered. Based on the comparative analysis of the differences in the distribution of students' professional ability, the students' professional ability is concluded.

There are still many shortcomings in this study. First, the data type is relatively simple. Further studies can consider including the awards students received in scientific research competitions and the social practice students took to conduct more comprehensive analyses. Second, the K-means algorithm is easy to fall into the local optimal solution when performing data processing. This study does not make too much optimization on the algorithm itself. The characteristics of the students' clustering are not obvious in the actual operation, and it is necessary to analyze the data multiple times to take the optimal solution. Future studies can try to optimize the K-means algorithm to improve work efficiency.

References

1. H. Yong . The design and implement of middle school student's evaluation system with improved AHP [D].Shandong University. (2013)
2. W. Hong-xia , L. Xue-qin, C. Wen-liu. Evaluation of Training Quality of Innovative and

Entrepreneurial Talents Based on Fuzzy Comprehensiv Evaluation[J]. Journal of North China Polytechnic University, **01**, 125-129. (2017)

3. W. Nai-bo, Guo-mei H,W. Lei, C. Shu-yun. Research on physical fitness assessment of undergraduates based on artificial neural network model [J]. journal of nanchang university, **40**,5: 506-510.(2016)
4. B. Misnevsa, V. Jusas, J. Luis F. Alemán, N. Kafadarova.Remote Evaluation of Software Engineering Competences [J]. Procedia Computer Science, **104**, 20-26.(2017)
5. MacQueen J. 1967. Some Methods for Classification and Analysis of Multivariate Observations[C].In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, University of California Press, 281-297.
6. G. Jeh, J. Widom.SimRank:a measure of struc-tural-context similarity[C]. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA :ACM, 538-543. (2002)
7. R. Srikant. R. Agrawal.Mining generalized association rules[J].Future Generation Computer Systems, **13**, 2-3, 161-180.(1997)