On Studying Students' Professional Aptitude Based on the Clustering Quality Evaluation

Guiqin Duan[1,2], Yongsong Chen[1], Chensong Zou[3(✉)], and Liu Feng[4]

[1] School of Computing and Information Engineering, Guangdong Songshan Polytechnic College, Shaoguan, China
[2] Shaoguan Ecological Culture Big Data Engineering Technology Research Center, Shaoguan, China
[3] School of Electrical Engineering of Guangdong Songshan Polytechnic College, Shaoguan, China
190352915@qq.com
[4] Department of Information Engineering, Luoding Polytechnic College, Luoding, China

Abstract. When solving the problems of educational research and teaching practice, it is difficult to determine the number of clusters of the clustering algorithm, and the standard for clustering quality evaluation are diverse. Aiming at these problems, a clustering analysis model of professional ability has been designed. The model first uses affinity propagation to calculate the similarity matrix of professional ability and screen out the representative points of the cluster center by alternately updating the attractiveness and membership degree to determine the clustering upper limit k_{max} and complete the compression of cluster space. On this basis, DB, CH, Dunn and IGP indexes are used to obtain the optimal clustering disaggregation, and then the average value is taken as the final k value to achieve the clustering division of professional abilities. The study results show that the model can reasonably mine students' professional aptitude, providing a new idea for the implementation of educational reform such as students' professional ability analysis, career development planning, and hierarchical classification training.

Keywords: cluster · clustering quality evaluation · evaluation standard · data mining · professional ability

1 Introduction

Educational data mining [1] is to extract implicit and meaningful information in the educational system through data mining technology while clustering analysis technology is a common method used in educational data mining. At present, this technology is widely used in innovative application cases of educational big data, covering aspects such as improvement of teaching quality, coordination of students' dynamic development, scientific planning of resource allocation, and intelligent decision development in colleges and universities. The representative cases are as follows. Combined with the fore-end visual framework, the density-based clustering algorithm was adopted by

X. Yuan et al. (Eds.): ICEKIM 2023, AHCS 13, pp. 486–496, 2023.
https://doi.org/10.2991/978-94-6463-172-2_54

Yan Z. [2] to mine and analyze the relationship between students' academic performance and daily attendance. At the same time, the stability of academic performance was explored. Besides, adopting SQL Server 2008 Data Mining, Wang G.H. [3] implemented the clustering analysis technology on learners' behavior characteristics under the network environment, and implemented the qualitative analysis technology on the relationship between learning behaviors and learning effects, giving targeted guidance to four types of learners. Peng L.J. [4] proposed a semi-supervised learning model based on K-means and SVM algorithm and efficiently classified a large number of students to further distinguish students with different characteristics more accurately. Although the clustering analysis algorithm has solved many problems in educational research and teaching practice to a certain extent, the diversity and uncertainty of clustering division bring new challenges to clustering quality evaluation due to the setting of cluster K depending too much on the human experience. Therefore, the COMET ability model was adopted by Yang Y. [5] to extract three characteristics of professional abilities, namely, frequency, importance and relevance. He also combined density algorithm and DB index to summarize and extract typical work tasks. Duan G.Q. [6] proposed a new clustering effectiveness index to achieve the optimal clustering division and clustering evaluation of professional abilities by balancing the relationship between density within clusters and clustering separation. Guo P. [7] used the improved K-means algorithm to disperse the performance information and combined the Apriori algorithm and CH index to mine the potential relationship between courses. Last but not least, Gu X.C. [8] used the global K-means algorithm to cluster the students' performance and used the I Index as the clustering validity function to achieve the unsupervised classification of students' multiple-subject performance.

Referring to the above application cases of the clustering algorithm, this paper proposed to use affinity propagation to calculate the clustering upper limit kmax of the sample and took the k value as the optimal clustering disaggregation in case the internal evaluation indexes DB, CH, DVI and IGP take the extreme value. Finally, this paper took the professional ability. Data of students from the Big Data Technology in a university as a sample and completed the clustering analysis of the professional ability to verify the practicality of the model.

2 Affinity Propagation

The affinity propagation [9] (AP for short) is a new clustering algorithm published by Frey and Dueck on SCIENCE. As a clustering method of transmitting information between neighbors, its main idea is to regard all data points as potential cluster centers and then connect these data points to form a network. In the algorithm implementation, the cluster center to which each sample belongs is selected through the information transmission between sides in the network. The process of the AP algorithm is as follows. The Euclidean distance formula is first used to calculate the similarity between sample pairs, then the similarity matrix of all samples is obtained, and the attractiveness and membership degree of samples are updated iteratively. When the condition that the number of iterations exceeds the preset value or the representative point's updating is stopped after multiple iterations are met, the AP algorithm will be terminated. Finally, the

samples outside the class representative points are divided into corresponding clusters to complete clustering. The AP algorithm consists of four steps, which are described as follows:

Step 1: Use the Euclidean distance formula to obtain the distance between x_i and x_j in the set.

$$d(x_i, x_j) = \sqrt{\sum_{p=1}^{l} \left(x_i^p - x_j^p\right)^2} \tag{1}$$

Specifically, $i = 1,2,\ldots,N$; $j = 1,2,\ldots,N$; l means the characteristic dimension of the samples.

Step 2: Take the opposite of the Euclidean distance as the similarity value $s(i,k)$ of the samples x_i and x_k to obtain the similarity matrix.

$$s(i,k) = \begin{cases} -\|x_i - x_k\| & i \neq k \\ p(k) & i = k \end{cases} \tag{2}$$

Specifically, the larger the $p(k)$ is, the greater the probability that sample x_k is selected as the representative of the cluster center.

Step 3: Transmit information, alternately update the attractiveness r and the membership degree a, and generate the cluster center with high representativeness. Specifically, the attractiveness $r(i,k)$ means the attractiveness degree of x_k to x_i. The greater the attractiveness degree, the greater the probability that x_k will become the center of x_i cluster. The membership degree $a(i,k)$ means the membership degree of x_i to x_k. The greater the membership degree, the greater the possibility that x_i chooses x_k as the cluster center. The alternative updating method of attractiveness r and membership degree a is shown in Formulas (3) and (4):

$$r(i,k) = s(i,k) - \max_{k' \neq k}\left\{a\left(i,k'\right) + s\left(i,k'\right)\right\} \tag{3}$$

$$a(i,k) = \begin{cases} \min\left\{0, r(k,k) + \sum\limits_{i' \notin \{i,k\}} \max\left[0, r(i',k)\right]\right\} & i \neq k \\ a(k,k) = \sum\limits_{i' \neq k} \max[0, r(i',k)] & i = k \end{cases} \tag{4}$$

There are oscillations in the process of updating attractiveness and membership degree. The damping parameter λ is introduced to reduce the oscillation range, eliminate oscillation and rectify $r(i,k)$ and $a(i,k)$ in the process of iteration to make the iteration process more stable. Set the damping parameter as $\lambda \in [0,1)$ and the iteration time as t. The rectified iterative updating process is shown in Formulas (5) and (6):

$$r(i,k)^{t+1} = (1-\lambda) \times r(i,k)^{t+1} + \lambda \times r(i,k)^t \tag{5}$$

$$a(i,k)^{t+1} = (1-\lambda) \times a(i,k)^{t+1} + \lambda \times a(i,k)^t \tag{6}$$

Step 4: Define the cluster center representative. Select the sample x_k that satisfies the maximum sum of attractiveness $r(i,k)$ and membership degree $a(i,k)$ as the class representative point of the cluster where x_i belongs to:

$$k = \arg\max\{a(i,k) + r(i,k)\} \tag{7}$$

3 Clustering Quality Evaluation Indexes

The clustering quality evaluation is divided into the external evaluation and internal evaluation. External evaluation generally refers to the use of external evaluation indexes to detect the clustering results under the condition of known correct clustering division. The evaluation method adopted in this paper refers to the internal evaluation that has no connection with external information and only uses the internal attribute of samples to evaluate the clustering quality. In the evaluation process, "compact within clusters and separation among clusters" is generally taken as an important criterion for the rationality of clustering division. Common internal evaluation index formulas and characteristics are described as follows:

(1) DB Index (Davies-Bouldin Index) [10]

$$DB(k) = \frac{1}{k}\sum_{i=1}^{k} \max_{j,j\neq 1} \frac{\frac{1}{n_i}\sum_{x\in C_i} d(x, v_i) + \frac{1}{n_j}\sum_{x\in C_j} d(x, v_j)}{d(v_i, v_j)} \tag{8}$$

The DB index takes the sum of the average distance between each sample in the cluster of two neighboring clusters and the cluster center as the distance within the cluster. The distance between cluster centers of two neighboring clusters is taken as the distance between clusters. The maximum ratio of the two is taken as the similarity of the cluster, and the similarity of all clusters is averaged to the DB index of the sample set. It can be seen that the smaller the index is, the more compact the clusters of the current clustering division are. Besides, the more separated the clusters are, and the more ideal the clustering results are

(2) CH Index (Calinski-Harabasz) [11]

$$CH(k) = \frac{\sum_{i=1}^{k} n_i d^2(v_i, c) \Big/ (k-1)}{\sum_{i=1}^{k}\sum_{x\in C_i} d^2(x, v_i) \Big/ (N-k)} \tag{9}$$

To facilitate the understanding, the numerator in the CH index can be regarded as the separating capacity between clusters and the denominator as the compactness within clusters. The numerator is represented by the product of the number of samples in the cluster and the square sum of the distance between the cluster center and the sample center. The denominator is represented by the sum of the square sum distance within each cluster. It can be seen that the larger the numerator in the index is, the better the separation

effect between clusters is. The smaller the denominator is, the higher the compactness within clusters is. Therefore, the larger the index is, the better the clustering quality is

(3) *Dunn* Index (Dunn's Indices) [12]

$$DVI = \min_{1\le i\le K}\left(\min_{1\le j\le K, i\ne j}\left(\frac{d(C_i, C_j)}{\max\limits_{1\le t\le K}(\delta(C_t))}\right)\right) \tag{10}$$

Dunn index is expressed by the ratio of the distance between clusters to the distance within clusters. Specifically, the distance between clusters is the minimum distance between samples of any two different clusters. The distance within a cluster is expressed by the distance between the two farthest samples in the same cluster. It can be seen that the larger the Dunn index is, the better the clustering quality is

(4) *IGP* Index (In-Group Proportion) [13]

$$IGP(K) = \frac{1}{K}\sum_{i=1}^{K} igp(i, X) \tag{11}$$

IGP takes the ratio of the two nearest samples to the same cluster as the standard to judge the clustering quality of which the basis is that when implementing clustering division on a sample, other subjects in the same cluster of the sample shall have the highest similarity with the sample. That is:

$$igp(i, X) = \frac{\left\{j|Class_X(j) = Class_X\left(j^N\right) = i\right\}}{\{j|Class_X(j) = i\}} \tag{12}$$

where, *igp* (i, X) represents the index value of class i in data set X. j^N represents the sample with the shortest distance from sample j. $\text{Class}_X(j)$ represents the class of the j th sample in data set X. The larger the *IGP* index value is, the higher the probability that the sample and its nearest neighbor will be classified into the same class, and the better the clustering result will be.

4 Clustering Analysis Model of Professional Ability Based on Effectiveness Evaluation

The clustering analysis model of professional ability consists of two parts. First, the AP algorithm is used to calculate the upper limit value of the sample clusters, and then the optimal clustering number and their mean value are obtained by calculating each internal evaluation index. The model structure is shown in Fig. 1. The algorithm is divided into seven steps of which the process is as follows.

4.1 Algorithm Process

Step 1 Use Formulas (1) and (2) to obtain the similarity matrix of students' professional abilities.

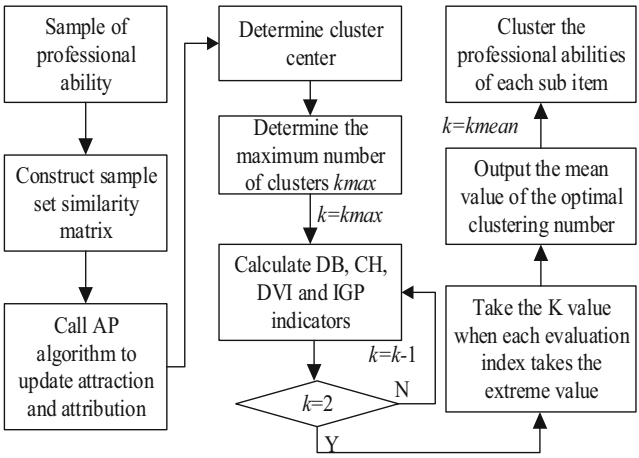

Fig. 1. Cluster model of professional ability

Step 2 Use Formulas (3) and (4) to alternately update the attractiveness r and membership a to obtain the representative of the cluster center.

Step 3 Set the damping parameter λ and the total iterative parameter t. Use Formulas (5) and (6) to reduce the vibration amplitude and correct the attractiveness and membership degree.

Step 4 Take the candidate point that meets Formula (7) as the representative point of the cluster center, repeat step 3 in this link, correct the attractiveness and membership degree of all samples again, screen out the representative point of the cluster center, divide other samples into the corresponding clusters, obtain the upper limit value k_{max} of the clusters, and set $k = k_{max}$.

Step 5 Use Formulas (8), (9), (10) and (11) in turn to obtain *DB*, *CH*, *Dunn* and *IGP* indexes.

Step 6 Let $k = k$-1, execute Step 5 circularly, calculate each internal evaluation index when $k = \{2, 3,\ldots, k_{max}\}$, and according to the characteristics of each index, take the k value as the optimal cluster when the index takes the extreme value.

Step 7 Take the average value of the optimal clustering number of each index as the final k value to complete the clustering division of sub-divisional professional abilities.

4.2 Application of the Algorithm

1) Sample data of the professional abilities.

In combination with modern education and teaching theory and the quality, knowledge and skills requirements of the major, the project team designed the professional ability level standard according to the post requirements of the Big Data Technology major (Table 1). 30 professional abilities from the first semester to the fourth semester of 75 students in the Big Data Technology major in 2020 grade of the school were collected and processed in advance. Limited to space, only some sample data (Table 2) were listed.

Table 1. Big Data Technology Professional ability level standard

Level	Assessment of task completion effect					
	High quality	High efficiency	Successful completion	Resolution of sudden problems	Need for help	Need for guidance
5	√	√	√	√	×	×
4	√	√	√	×	×	×
3	×	×	√	×	×	×
2	×	×	√	×	×	√
1	×	×	√	×	√	√

Table 2. Sample data of professional ability of big data technology students

Short name for the professional ability	Attribute value of professional abilities of 75 students						
A1	4.8	4.9	4.9	4.8	5.0	…	5.0
A2	4.4	4.5	4.4	4.4	4.5	…	4.7
A3	4.2	4.3	3.8	3.9	3.8	…	4.2
…	…	…	…	…	…	…	…
B7	3.6	3.8	4.2	4.5	3.8	…	4.5
…	…	…	…	…	…	…	…
C13	3.8	3.4	3.3	3.5	3.3	…	4.6

2) The output of clustering results.

Using the model, this link clusters the samples in Table 2 from three aspects: quality, knowledge and skills. Specifically, the damping parameter of the AP algorithm was $\lambda = 0.90$, the total number of iterations was $t = 1000$, and the operating environment was Matlab 2020B. The optimal clustering number and the average value of each internal evaluation index were shown in Table 3. The clustering division of the professional ability of each sub-item was shown in Fig. 2. The detailed division results were shown in Table 4.

Table 3. Optimal clustering number of professional ability

	DB	CH	DVI	IGP	means
Quality	4	4	3	5	4
Knowledge	3	3	3	3	3
Skill	3	4	3	6	4

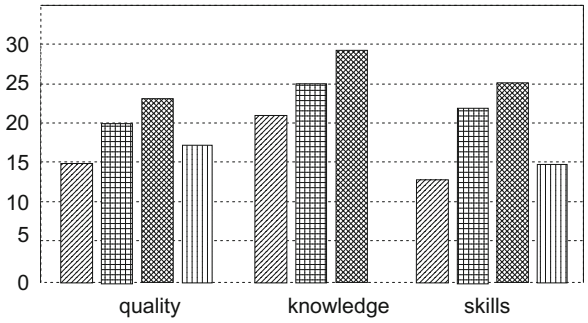

Fig. 2. Optimal division of professional ability

Table 4. Classification results of professional abilities

	quality	knowledge	skills
Class I	7,13,14,19,26,30,32,34,38,42,46,57,59,66,68	2,7,11,13,14,20,24,26,30,32,34,37,38,41,42,46,50,56,57,59,68	7,11,13,14,19,23,30,32,42,46,56,66,68
Class II	2,8,12,15,20,24,36,39,41,47,48,55,56,58,63,64,70,72,73,75	1,6,8,10,15,16,18,19,27,28,36,39,40,45,47,48,52,58,60,62,63,64,66,69,73	2,8,10,15,16,22,26,34,36,38,39,44,48,49,58,61,63,64,65,69,72, 73
Class III	3,4,6,10,16,17,18,21,22,23,25,27,29,35,37,40,44,50,60,61,65,67,71	3,4,5,9,12,17,21,22,23,25,29,31,33,35,43,44,49,51,53,54,55,61,65,67,70,71,72,74,75	4,5,6,9,12,17,20,21,24,25,27,28,29,31,37,40,51,52,55,57,60,71,74,70,75
Class IV	1,5,9,11,28,31,33,43,45,49,51,52,53,54,62,69,74	----------	1,3,18,33,35,41,43,45,47,50,53,54,59,62,67

5 Analysis of Professional Aptitude

5.1 Analysis of "Skill" Clustering Result

It can be seen from Table 4 that "skill" samples were divided into four classes. Based on the sample data analysis of the learning situation in Table 2, the results are as follows:

There were 13 students of Class I in total with outstanding C1, C6, C7 and C9 characteristics, all of which were above 4.5, indicating that students of this class have good abilities in big data processing, calculation, analysis and product data visualization.

Therefore, students of this class shall take the position of big data analyst as their career development direction.

There were 22 students in Class II. Compared with the students of Class I, their C1, C6, C7 and C9 attribute values were slightly lower, all lower than 4.3. However, their other skill attributes covered a relatively wide range, especially the characteristic values of C1, C4, C7, C9 and C10, which were all above 4.1, indicating that students of this class were relatively familiar with data extraction, real-time data collection and database migration and had certain data analysis abilities. Therefore, students of this class shall take ETL data processing engineer or big data analysis as their career development direction.

There were 25 students of Class III of which the common feature was that their basic analysis algorithm design ability and application ability was lower than those of the first two classes. Their C7 attribute values were between 3.7 and 4.0, and the C11, C12, and C13 characteristic values were between 3.8 and 4.2, indicating that their WEB application development and Java project development abilities were good. Therefore, students of this class shall continue to learn in depth in Java software development, RESTful application development, Web fore-end development, and micro-service development, and take Java development engineers as target posts。

There were 15 students in Class IV with a professional potential advantage not obvious, of which 13 skill characteristics were lower than 3.7. Compared with other attributes, their C3 and C8 values were slightly higher, between 3.5 and 3.7. Students of this class shall improve their installation ability of the operating system and building and maintenance ability of big data platform, and pay attention to professional practice to further improve their professional skills.

5.2 Analysis of "Knowledge" Clustering Result

The samples of the "knowledge" type in Table 4 were divided into four classes. Based on the analysis of the sample data of the learning situation in Table 2, it can be seen that:

There were 21 students in Class I, with B1–B7 characteristic values above 4.3. Based on the above "skill" clustering results, it can be found that 11 people of Class I of skill clustering overlapped with the current Class. In addition, 13 people of Class I of quality clustering were also classified in this class. It can be seen that students of this class not only have solid basic knowledge but are also good at self-study ability, self-cognition ability and professional skills. They shall invest more time and energy in professional courses, lay a solid foundation, further improve the knowledge system and choose "upgrading from a junior college to university" to continue their studies.

There were 25 students in Class II, who had a good grasp of basic knowledge. Their mastery of basic algorithm analysis, program debugging and other knowledge was slightly lower than that of Class I students. The other six characteristic values were generally between 3.6 and 4.2. Some students had a slight phenomenon of "partial branch in learning". Students of this class shall complement their weaknesses and improve their knowledge structure while strengthening their preponderant disciplines.

There were 29 students of Category III, whose main feature was that the average value of 7 knowledge characteristics is lower than 3.3. Compared with students of Class II, their professional knowledge was slightly weak. Especially, their understanding of

key technologies such as basic algorithm analysis, program debugging and big data analysis was insufficient. In the short term, they shall focus on achieving self-motivation and improving their self-learning ability, and timely reflect and summarize to establish a correct learning concept.

6 Conclusion

Using affinity propagation and the DB, CH, DVI and IGP internal evaluation indicators, this paper designed the clustering model of students' professional abilities. It provided the optimal clustering number and the optimal clustering division from three professional abilities of quality, knowledge and skill. In addition, it mined the distribution status of students' professional abilities and completed the data modeling and quantitative and qualitative analysis of students' professional abilities. This paper also resolved the problem that the k value of the clustering algorithm in the innovative application of education big data was too dependent on artificial experience, and the clustering results were easily affected by the objective environment. In practical application, this model can provide students with objective learning guidance and professional development advice to offer new thinking to the implementation of educational reform such as teachers' satisfying teaching and targeted guidance.

Acknowledgment. Education Science Planning Project of Guangdong Province (2022GXJK 490); Scientific Research Projects of Guangdong Provincial Education Department (2021KTSCX227, 2021ZDZX4109, 2020ZDZX3119); Science and Technology Bureau Projects of Shaoguan City (200811224533986, 210718114531595); Scientific Research Platform of Guangdong Songshan Polytechnic College (2021xjkypt05); Planning project of Guangdong Higher Education Association (2022GQN39).

References

1. Zhou Q., Mou C., Yang D. Research progress on educational data mining: Asurvey [J]. Journal of Software,2015,26(11):3026−3042.
2. Yan Z. Research on visualization analysis of learning information data based on density clustering algorithm[J]. Modern Computer,2022,28(06):43–47.
3. Wang G.H., Fu G.S. The cluster analysis of online learners' behavior characteristics from the perspective of data mining[J]. Modern Distance Education Research,2018(04):106–112.
4. Peng L.J., Wu Q.C., Li S.M., et al. Research on comprehensive evaluation and classification of students based on k-means clustering and SVM algorithms[J].Digital Technology & Application,2020,38(10):88–91.
5. Yang Y., Zou C.S. Cluster analysis of occupational competency in modern apprenticeship based on comet[J]. Journal of Guangdong Polytechnic Normal
6. Duan G.Q., Zou C.S. Occupational competence evaluation model based on affinity propagation clustering[J]. Computer and Modernization,2022(05):21–27.
7. Guo P., Cai C. Data mining and analysis of students' score based on clustering and association algorithm [J].Computer Engineering and Applications,2019,55(17):169–179.

8. Gu Xi.C., Xu F.X., Yang Y., et al. Analysis of university students' grades based on global k-means algorithm [J]. Journal of Changchun University of Science and Technology (Natural Science Edition), 2019, 42(05):93–97.
9. Frey B.J., Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814):972–976
10. Davies DL, Bouldin DW. A cluster separation measure[J]. IEEE transactions on pattern analysis and machine intelligence,1979,2(2):224–227.
11. T. Caliński, J Harabasz. A dendrite method for cluster analysis[J]. Communications in Statistics, 1974,3(1):1–27.
12. J.C.Dunnâ. Well-separated clusters and optimal fuzzy partitions[J]. Journal of Cybernetics, 1974,4(1):95–104.
13. Lin T.Y. Data mining and machine-oriented modeling: a granular computing approach [J].Applied Intelligence,2000,13(2):113–124.