Employment Information Recommendation Model Based on Improved Density Clustering

Guiqin Duan [1, 2], Yuxian Wang [1, 2], Liying Guo [3], Chengsong Zou [3*]

E-mail box of the corresponding author: 190352915@qq.com

[1]School of Computer and Information Engineering, Guangdong Songshan Polytechnic, Shaoguan, Guangdong, 512126;

[2]Shaoguan Ecoculture Big Data Engineering Technology Research Center, Shaoguan, Guangdong, 512126;

[3]School of Electrical Engineering, Guangdong Songshan Polytechnic, Shaoguan, Guangdong, 512126

Abstract: Given employment difficulties and low employment quality of college students under the background of higher vocational enrollment expansion, an employment information recommendation model based on improved density clustering was designed. A weighted user similarity calculation method based on professional similarity, job intention similarity, and professional ability similarity was given, and the quantitative analysis process of professional ability was optimized. Considering that high-density samples are closely surrounded by low-density samples, a new density clustering algorithm was proposed to improve the accuracy of employment recommendations. The practice has proved that this model can effectively mine students' employment information, and reduce the iteration times of the clustering algorithm, accompanied by high information retrieval integrity and good personalized employment recommendation performance.

Keywords: cluster analysis; density clustering; selection of initial cluster center; employment recommendation; education informatization

1 Introduction

As the informatization construction of vocational education has been continuously improved, the data mining technology-based analysis of the data generated in education and teaching can not only dig into students' professional potentials but also provide objective and scientific reference for their career planning [1,2]. As a classical unsupervised data mining technology, cluster analysis has been widely used in many types of research and practices in the field of education. Guo P [3] used the improved K-means algorithm in combination with the CH index to evaluate the clustering effect of student marks. Peng L J et al. [4] used the k-means algorithm to label some data based on the analysis of students' comprehensive ability factors, and classified other data using the trained SVM model. Li H K et al. [5] used web crawler technology to crawl massive recruitment and employment data from the internet, and combined cluster analysis algorithm, multiple linear regression model, and word segmentation to extract keywords to construct an open employment information recommendation model integrating universities, students, and enterprises. For the comprehensive application of

ITEI 2022, December 23-25, Harbin, People's Republic of China

DOI 10.4108/eai.23-12-2022.2329117

algorithms, Li N [6] implemented a comprehensive evaluation system for college students based on the web, realizing standardized storage and data analysis of student information. With the data of previous graduates' graduation destinations as the actual reference, Shen D [7] used the integrated classification algorithm to predict the graduation destination for students, and performed employment prediction for students according to their actual situation and previous graduates' employment situation, so the recommendation system is of more reference value for recent graduates. Liu X X [8] analyzed the traditional collaborative filtering recommendation algorithm based on users, put forward a new hybrid algorithm, and compared the application results of the two algorithms in the employment recommendation system of higher vocational colleges, confirming that the improved hybrid algorithm is more suitable for the employment recommendation system of higher vocational colleges and has improved the accuracy of similar students' recommendation. Zhou X M and Duan H X [9] analyzed personal information, users' browsing, resume sending and other behaviors, and adopted the improved clustering algorithm to recommend matching jobs for users, which improved the efficiency of users' job search and reduced the corresponding time cost.

As revealed by the above research results, providing scientific vocational guidance for graduates through information technology has become a powerful guarantee for improving the employment quality of colleges and universities. Therein, cluster analysis technology is a conventional means of solving the employment problem of graduates. On the basis of the above literature review, a professional ability clustering algorithm based on employment similarity was put forward. Then, the weighted user similarity calculation method based on professional similarity, job intention similarity, and professional ability similarity was given. Moreover, the improved density clustering algorithm was proposed and applied to students' employment recommendations.

2 Improved Density Clustering Algorithm

The improved density clustering algorithm consists of two parts: sampled data preprocessing and clustering algorithm improvement.

(1) Sampled data preprocessing

According to the standard of post-professional ability and professional quality, the weight of professional ability feature items was determined, and each feature item of students in the school was quantified. The weighted calculation method was used to ensure the rationality of the feature weight coefficient calculation so that it could accurately and effectively reflect the importance of students' feature attributes to employment choice.

(2) Clustering algorithm improvement

In the link of initial center selection, firstly, the density of each sample was obtained by using a self-defined density formula, and the high-density samples were taken as the candidate representative points [10] of the cluster center, thus generating a candidate representative point set. In the set, the one with the minimum sum of the distance from other candidate representative points is selected as the first initial cluster center, and then the product maximum method was used to complete the initial cluster center selection, so as to obtain the initial cluster center set, that is, $Z=\{z1,z2,\ldots,zk\}$. In the iterative calculation of the cluster center, the initial clustering was completed according to set Z, and the distance matrix between each sample in the cluster and the cluster mean center was calculated. To reduce the deviation between the cluster center obtained by the mean value method and the actual cluster center, the sample closest to the cluster mean was taken as the temporary center of the cluster, and a temporary cluster center set was generated, that is, Z={z1', z2',…, zk'}. Then, the relevant samples were divided into their own clusters by the minimum distance method and repeated, and the iterative calculation of cluster centers was repeated until the convergence of the criterion function, thus finishing the clustering process. The definition, formula, and algorithm description of the improved density clustering algorithm are as follows.

2.1 Basic concepts and formulas

X is set as a data sample containing n student samples, X={x1,x2,……,xn}, the number of feature attributes of each student sample is p, and xi={xi1, xi2,..., xip}. X is divided into k clusters, X={C1, C2,…,Ck}, in which |Ci| denotes the number of samples in cluster i, zk is the center of the cluster k, and the set constituted by multiple cluster centers is Z, namely, Z={z1,z2,…,zk}.

Definition 1. The employment similarity (distance between samples) between any two students is:

$$d\left(x_i,x_j\right)=f_{Pro}\left(x_i,x_j\right)*w_{Pro}+f_{Job}\left(x_i,x_j\right)*w_{Job}+f_{Abi}\left(x_i,x_j\right)*w_{Abi} \tag{1}$$

Where fPro(xi,xj) is the professional similarity between students xi and xj, which can be solved through Formula 2. fJob(xi,xj) stands for the job intention similarity between xi and xj, which can be calculated as per Formula 3. fAbi(xi,xj) represents the professional ability similarity between xi and xj, which can be obtained through Formula 4. Given certain differences in the importance of professional similarity, job intention similarity, and professional ability similarity [11], their weights are denoted by wPro=0.2, wJob=0.2, and wAbi=0.6, respectively, and wPro +wJob +wAbi=1.

$$f_{Pro}\left(x_i,x_j\right)=\begin{cases}1 & \text{if }\left(x_i\text{ and }x_j\text{ have the same major}\right)\\ 0.8 & \text{if }\left(x_i\text{ and }x_j\text{ belong to the same post group}\right)\\ 0.5 & \text{if }\left(x_i\text{ and }x_j\text{ study in the same department}\right)\\ 0 & \text{if }\left(x_i\text{ and }x_j\text{ are from different department}\right)\end{cases} \tag{2}$$

$$f_{Job}\left(x_i,x_j\right)=\begin{cases}1 & \text{if }\left(x_i\text{ and }x_j\text{ work at the same post}\right)\\ 0.8 & \text{if }\left(x_i\text{ and }x_j\text{ are in the same professional post group}\right)\\ 0.5 & \text{if }\left(x_i\text{ and }x_j\text{ job tasks are overlapped}\right)\\ 0 & \text{if }\left(x_i\text{ and }x_j\text{ job tasks are not overlapped}\right)\end{cases} \tag{3}$$

$$f_{Abi}\left(x_i,x_j\right)=\sum_{l=1}^{p}\left(1-\frac{\left|x_i^l-x_j^l\right|}{100}\right) \tag{4}$$

Where i=1,2,...,n; j=1,2,...,n; l=1,2,…,p, | xi1- xj1| stands for the similarity between xi and xj in the occupational characteristic l. If the value is 1, the two students are completely consistent in the professional ability l, but if it is 0, the two students are totally different in this professional ability.

Definition 2. The distance sum of any student sample xi is defined as the sum of the distance of this sample to different samples in the dataset.

$$\mathtt{distSum}\left(x_i\right)=\sum_{j=1}^{n} d\left(x_i, x_j\right) \tag{5}$$

Definition 3. The density of sample xi is:

$$\mathtt{density}\left(x_i\right)=\sum_{j=1,\, x_i\neq x_j}^{n}\frac{\mathtt{distSum}\left(x_j\right)}{d\left(x_i, x_j\right)} \tag{6}$$

In this study, density was defined based on the following idea: From the positional relationship, when a sample xi is tightly surrounded by other samples, this sample has a relatively small distance sum from other samples. When sample xi presents a dispersed positional relationship with other samples, the distance sum between this sample and other samples is relatively large. In the expression of density, the distance between samples xis and xj serves as the denominator, the distance sum from xj to all samples is the numerator, and the cumulative sum of their distance ratio denotes the degree to which sample xi is surrounded by other samples, i.e., the density of xi. With the sample xi as an example, when the numerator in the cumulative sum of Equation (6) is large, the cumulative distance sum of other samples is also large in addition to xi. In the case of a small denominator, the cumulative distance sum of xi from other samples is small. Therefore, a greater numerator and a smaller denominator represent the greater value of the expression, the larger the density of xi being surrounded by other samples, that is to say, the greater the relative density of xi, and the stronger its representativeness as the cluster center.

Definition 4. The average density of the sample set is defined below:

$$\mathtt{avgDensity}(X)=\frac{\sum_{i=1}^{n}\mathtt{density}(x_i)}{n} \tag{7}$$

Definition 5. The set of candidate representative points is defined as the set of samples with the density α-times higher than the average density of the sample set

$$H=\{h_i\} \tag{8}$$

Where xi, xj∈Ct, t=1,2,…k

Definition 6. The distance matrix between candidate representative points is defined as follows:

$$\mathrm{HDist}=\begin{bmatrix} 0 & d(h_1,h_2) & \dots & d(h_1,h_j) \\ d(h_2,h_1) & 0 & \dots & d(h_2,h_j) \\ \dots & \dots & 0 & \dots \\ d(h_j,h_1) & d(h_j,h_2) & \dots & 0 \end{bmatrix} \tag{9}$$

where j denotes the number of elements in the set H.

Definition 7. The distance matrix between a sample and the mean center of this cluster

$$\mathrm{distMean}(m)=\begin{bmatrix} d(x_1,\mathrm{mean}(C_m)) \\ d(x_2,\mathrm{mean}(C_m)) \\ \dots \\ d(x_{|C_m|},\mathrm{mean}(C_m)) \end{bmatrix} \tag{10}$$

Where m=1,2,...,k, Cm stands for the sample set of the cluster m.

Definition 8. After cluster updating, the sample xi the closest to the mean value in the cluster serves as the cluster center and meets the following condition:

$$d(x_i,\mathrm{mean}(C_m))=\min(\mathrm{distMean}(m)) \tag{11}$$

Definition 9. The error sum of squares (E) of clustering is defined as below:

$$E=\sum_{i=1}^{k}\sum_{j=1}^{m}\left|x_{ij}-z_i\right|^2 \tag{12}$$

Where xij is the sample j in cluster i, and zi is the center of cluster i.

2.2 Algorithm description

Step 1. Calculate the employment similarity between any two students via Equations (1)– (4).

Step 2. Calculate the density of student samples using Equations (5) and (6).

Step 3. Obtain the candidate representative point set H according to Equations (7) and (8), where the parameter α is 1.0.

Step 4. Calculate the distance matrix between candidate representative points using Equations (1) and (9), and select the sample with the minimum distance sum from other candidate representative points in H as the first cluster center z1 and store it in the set Z.

Step 5. Select the candidate representative point z2 most distant from z1 in set H and store it in set Z.

Step 6. Select a representative point meeting max(d(hi, z1)×d(hi, z2)) in the set H as z3 and store it in set Z.

Step 7. Operate Step 6 repeatedly until |Z|=k, where the number of clusters (k) is the number of professional posts that previous graduates have taken.

Step 8. Calculate the distance between each sample in X and each candidate point in the set Z through Equation (1) and classify it into the cluster with the minimum distance.

Step 9. Calculate the distance matrix of each sample to the mean center of the cluster through Equation (10), and take the sample closest to the mean value in the cluster as the new cluster center according to Equation (11).

Step 10. Repeat Steps 8 and 9 and update the cluster center set Z.

Classify samples in X into the nearest cluster according to distance, and calculate and judge whether E converges through Equation (12). If yes, the algorithm terminates, and if not, turn to Step 8 and update the cluster center once again.

3 Application of Employment Information Recommendation Model

The model proposed in this study consists of two parts: cluster division of previous graduates and employment recommendation of fresh graduates. In the former part, initial clustering was performed using the improved density clustering algorithm, followed by the cluster division of operating posts of previous graduates and the calculation of the mean value of the post-group center (cluster center). In the latter part, the similarity between the data of fresh graduates and the post-group center was calculated and employment recommendation was completed. The model structure is shown in Figure 1.

3.1 Model structure

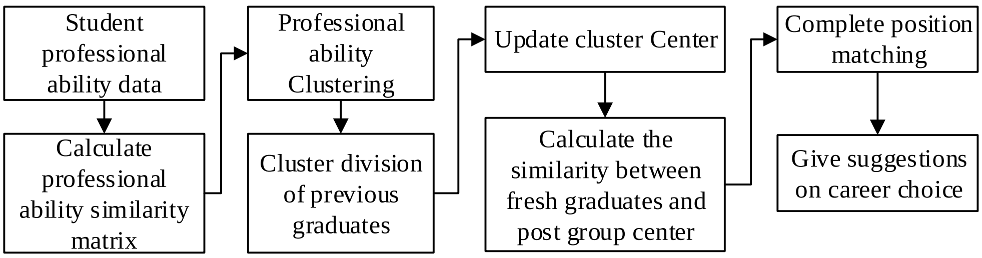


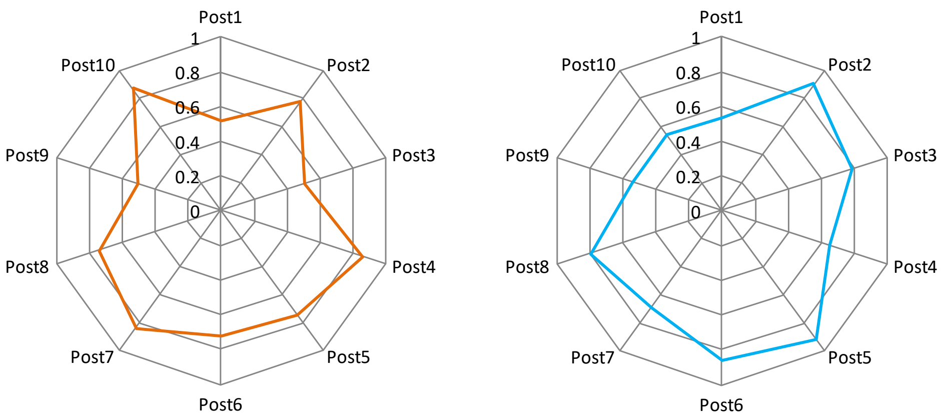
Figure 1. Talent Recommendation Model Based on Professional Ability Clustering Algorithm

3.2 Position matching and talent recommendation

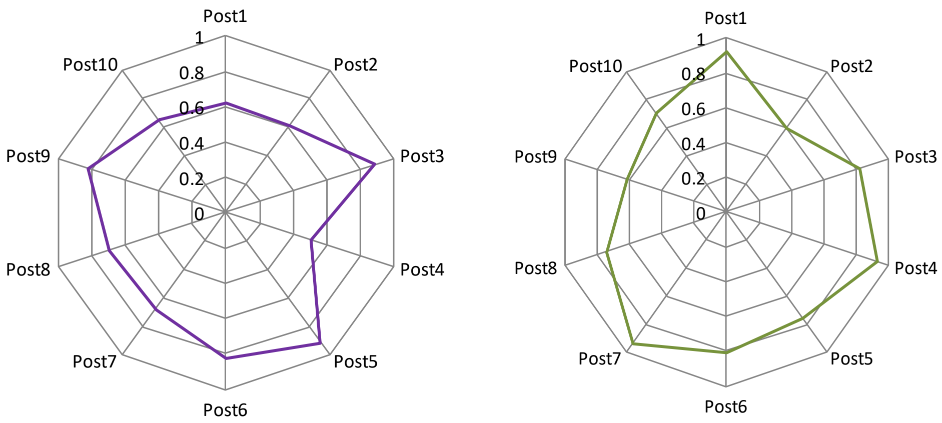
Figure 2,Table 1 show the career recommendation results obtained by using the talent recommendation model. Figure 2 shows the radar chart of position-matching degrees of graduates No.1 to No.4. It can be seen that the positions that match their professional abilities most are respectively position 4, position 5, position 5, and position 7. Table 1 lists the ten positions with the highest matching degree between 16 graduates and their professional abilities,from the perspective of enterprise recruitment, taking the position-ability matching degree≥0.9 as an example, the number of students meeting the professional ability requirements of post 1 to post 4 is: post 1={4, 12, 14}, post 2={2, 7}, post 3={7, 10, 15, 16}, and post 4={4, 8, 13}. It can be seen that compared with other students, graduate No.4 is more probable to be chosen by relevant enterprises in the employment recommendation, with a wider range of independent choices of employment and relatively evident employment advantages.

Table 1: Matching Degree between Graduates and Enterprise Posts

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Post1	0.5	0.5	0.6	0.9	0.7	0.5	0.7	0.5	0.8	0.5	0.6	0.9	0.8	0.9	0.7	0.8
Post2	0.7	0.9	0.6	0.6	0.6	0.8	0.9	0.6	0.8	0.7	0.8	0.7	0.5	0.5	0.8	0.8
Post3	0.5	0.7	0.8	0.8	0.7	0.8	0.9	0.6	0.6	0.9	0.7	0.6	0.8	0.8	0.9	0.9
Post4	0.8	0.6	0.5	0.9	0.5	0.6	0.5	0.9	0.6	0.7	0.8	0.6	0.9	0.8	0.6	0.7
Post5	0.7	0.9	0.9	0.7	0.8	0.6	0.5	0.9	0.6	0.9	0.8	0.8	0.5	0.5	0.7	0.6
Post6	0.7	0.8	0.8	0.8	0.5	0.7	0.5	0.6	0.7	0.5	0.6	0.6	0.9	0.5	0.9	0.6
Post7	0.8	0.6	0.6	0.9	0.9	0.5	0.5	0.8	0.5	0.9	0.9	0.8	0.7	0.6	0.6	0.5
Post8	0.7	0.8	0.7	0.7	0.7	0.7	0.8	0.5	0.8	0.7	0.5	0.7	0.6	0.7	0.6	0.7
Post9	0.5	0.5	0.8	0.6	0.5	0.8	0.5	0.9	0.5	0.6	0.9	0.8	0.7	0.7	0.5	0.6
Post10	0.8	0.5	0.6	0.7	0.6	0.8	0.5	0.8	0.8	0.8	0.6	0.6	0.6	0.6	0.9	0.5



(a) (b)



(c) (d)

Figure 2. Radar Map of Post Matching Degree

4 Conclusion

Under the background of enrollment expansion in higher vocational colleges and combining the actual talent demand of employers, the data mining technology and machine learning technology are applied to educational science, and the professional ability clustering and collaborative filtering recommendation technology are integrated and transplanted into the theoretical research and practical exploration of college students' employment, entrepreneurship, and career development planning, which breaks through the traditional global recommendation mode. The graduate-post matching degree is enhanced by mining and analyzing the influence of students' own characteristic attributes on employment recommendation. The method proposed in this study can provide a useful reference for the reform of industrial talent supply mode in higher vocational colleges and lay a foundation for further refined and deepening research of scientific researchers in the field of enterprise recruitment and talent recommendation.

Acknowledgments. This paper is supported by the following fund projects: Guangdong Education Science Planning Project (2022GXJK490), Scientific Research Projects of the Department of Education of Guangdong Province (2021KTSCX227,2020KTSCX243,2021ZDZX4109), and projects of Shaoguan Science and Technology Bureau (200811224533986, 2107114531595).

References

[1] Geng Y J, Zhang Z G. FCM clustering algorithm based on improved kernel function and its application in college students' performance data mining [J]. College Mathematics, 2020, 36 (06): 23–28.

[2] Zhang G Y. Students' score analysis based on clustering method [J]. Computer Knowledge and Technology, 2019, 15 (09): 1–2.
[3] Guo P, Cai P. Data mining and analysis of students' score based on clustering and association algorithm [J]. Computer Engineering and Applications, 2019, 55 (17): 169–179.
[4] Peng L J, Wu Q C, Li S M, Zhou X X, Xiao C T. Students' comprehensive assessment and classification based on k-mean and SVM algorithm [J]. Digital Technology & Application, 2020, 38 (10): 88–91.
[5] Li H K, Cao H, Wang X L, Liu Y. Research on employment recommendation model for college graduates based on big data analysis [J]. China Metallurgical Education, 2019, (03): 93–97.
[6] Li N. Design and implementation of college students' comprehensive assessment system based on WEB [D]. Jilin: Jilin University, 2016.
[7] Shen D. Research and design of college graduates' destination information management and recommendation system [D]. Donghua University, 2019.
[8] Liu X X. Algorithm of employment recommendation system in higher vocational colleges [J]. Computer & Network, 2020, 46 (23): 68–71.
[9] Zhou X M, Duan H X. The design and implementation of an employment recommendation system based on Django [J]. Computer Knowledge and Technology, 2021, 17 (27):75–77.
[10] Duan G Q. In-cluster mean minimum distance clustering algorithm based on improved density [J]. Intelligent Computer and Applications, 2021, 11 (12): 82–86.
[11] Shi P, Yao W M, Wang X. Weighted Slope One optimization integrating user fuzzy clustering and similarity [J]. Computer and Modernization, 2021 (01): 70–75.