DOI: 10.23977/jaip.2023.060408
ISSN 2371-8412 Vol. 6 Num. 4

Research on Precise Ideological and Political Education Based on Improved K-means Algorithm for College Students' Portrait Construction

Qingpeng Huang*

School of Automobile, Guangdong Mechanical & Electrical Polytechnic, Guangzhou, 510550, China

**Corresponding author*

Keywords: K-means clustering algorithm; Ideological and political education; Crowd portrait; Student performance

Abstract: In order to analyze the performance of college students ' learning and life in all aspects during their stay in school, by collecting quantitative daily life behavior data such as moral, intellectual, physical, aesthetic and labor of college students, the attribute characteristics used to construct student user portraits are selected, and the data analysis model is established by using the improved K-means algorithm to select the initial center point.Based on K-means clustering technology, this paper analyzes student behavior data and student achievement data. The students ' performance is divided into six dimensions: professional performance, sports performance, competition performance, scholarship level, student cadre status, and second classroom performance. Drawing a crowd portrait that can fully show the students ' ability, accurately evaluate the students ' comprehensive ability, and realize the scientificity and feasibility of accurate ideological and political education.

1. Introduction

The Ministry of Education held a press conference on the employment progress of college graduates in 2022. It pointed out that the total number of college graduates in 2023 reached 10.2 million, an increase of 355 compared with the previous year. China's higher education is moving towards popularization. With the advent of the intelligent era, the further integration of information technology and education is profoundly changing the way of education and teaching. The '14th Five-Year Plan' clearly put forward the goal of 'deepening the reform of education evaluation in the new era and establishing and improving the education evaluation system and mechanism'. At the same time, the Central Committee of the Communist Party of China and the State Council issued the "Overall Plan for Deepening the Reform of Education Evaluation in the New Era," which pointed out that " adhere to scientific and effective, improve result evaluation, improve comprehensive evaluation, and make full use of information technology to improve the scientificity and professionalism of education evaluation." In this case, student portrait technology composed of emerging technologies such as big data sets and machine learning can become an important means to comprehensively describe students ' abilities and serve teaching evaluation. In order to achieve this goal, this paper designs and implements a student evaluation system based on user portraits. The system can realize data preprocessing, data vectorization, and complete the feature classification of each dimension of students. By constructing student portraits, the system weights the data on each dimension, and finally forms a comprehensive evaluation of students, helping teachers comprehensively and accurately understand the situation of each student. Jiang Nan and Xu Weisheng [1] analyzed the operation of students ' consumption habits based on the daily consumption data of campus card. The K-means algorithm with improved initial center point is used to determine the selection of the initial center point through the degree of intra-class density, cluster the students ' consumption habits, and use the Apriori algorithm to analyze the correlation degree of learning behavior, so as to assist the college student staff to manage the students in different categories. Xiong Zhongyang et al. [2] proposed a density-based idea to optimize the initial clustering center selection method-the maximum distance method, which has greatly improved the convergence speed and accuracy, but manually input the density coefficient, the algorithm scalability is limited ;Liu Jing [3] used the distance between the data center points and the ratio of the difference between the data inside and between the data as a function to determine the initial center point, which improved the efficiency of the clustering algorithm, but the size of the data volume will also affect the efficiency; Shadab Irfan et al. [4] used genetic algorithm to iteratively minimize the objective function and similarity function, which reduced the time complexity of the function. Zhang et al. [5] extracted user characteristics by constructing machine learning models, and studied non-directional online behavior to predict students ' academic performance.The research of Huang Gang [6] and Jiang Nan [7] is representative. The two scholars used the k-means algorithm to cluster students on the student campus consumption data set, analyzed the students' consumption habits and group characteristics, and gave a portrait description to provide a basis for the management of college students.

With the continuous deepening of information construction in colleges and universities, the resources of educational big data are becoming more and more abundant [8]. Big data technology has brought new historical opportunities for the innovative development of ideological and political work in colleges and universities in the new era [9]. Ideological and political workers in colleges and universities often face the confusion of accurately identifying students, accurately serving students, accurately predicting students, and accurately evaluating students. It is the technical advantages of big data that can respond to accurate needs, provide accurate services, and promote accurate implementation. The precise ideological and political work model has gradually attracted the attention of ideological and political workers in colleges and universities [10]. The student portrait method in ideological and political work in colleges and universities also has increasing research value and has gradually become a research hotspot [11,12]. Based on these research backgrounds, this paper starts with student portraits and explores the practical path of deep integration of big data technology and ideological and political work in colleges and universities.

2. Data Processing and Key Technologies based on K-means Algorithm

2.1 Basic Ideas of Improved K-means Clustering Algorithm

The K-means algorithm is an unsupervised learning algorithm for clustering data points by means. The key of the algorithm is to ensure that the clustering centers at different locations are randomly selected when the centers are far away from each other. The core idea is: select the number of clusters k, randomly select the sample as the initial centroid, examine the distance between each sample and the current centroid in turn, and select the nearest cluster for classification. After a round of investigation of all samples, the centroid of each cluster is updated separately, and the above process is repeated until the centroid no longer changes and the algorithm ends. The k clusters have the following characteristics: the data points in each cluster are as compact as possible, and the data points between clusters are as far away as possible [13]. The final convergence condition of the algorithm adopts the objective function of minimizing the sum of squared errors. The clustering process and its results are generally affected by the selection of the basic center point. The key to the performance of the mean clustering algorithm lies in the reasonable selection of the clustering center. Where the sum of squared errors is defined as Formula (1):

$$TSS = \sum_{i=1}^{k} \sum \mathrm{x} \in c_i \parallel \mathrm{X} - \mathrm{C_i} \parallel^2, \tag{1}$$

In the formula: X is the data point in the data set; ci is the center of mass.

TSS is a strict coordinate descent process, using Euclidean distance as a measure function between variables. Each approach; the direction of a variable Ci finds the optimal solution (2):

$$C_{\mathrm{i}} = \frac{1}{k} \sum X, \tag{2}$$

k denotes the number of elements in the cluster of Ci.

The mean value of the current clustering is the optimal solution (minimum value) of the current direction. Like each iteration process of K-means, it ensures that the TSS value becomes smaller after each iteration and finally converges. However, because TSS is a non-convex function, TSS does not guarantee to find the global optimal solution, but only to ensure the local optimal solution. In order to prevent the local optimum of the K-means algorithm from using the specific method in the process of processing the data, that is, the data is re-projected after the dimensionality reduction of the data, all the feature information is retained, the weight of each feature is recorded, and the K-means algorithm is performed many times. Finally, the minimum TSS is selected as the final result.

As a classical clustering algorithm, K-Means is mainly realized by iterative process. Now the data set is divided into different categories, the algorithm has simple heterosexual scalability. The advantages of the K-Means algorithm first randomly selects K from the sample set S. A sample is used as the initial clustering center. For K-means algorithm, the algorithm is simple and efficient. But algorithm clustering number. The selection of K time and center point lacks clear standard definition, and most of them are randomly given, which easily affects the results of the algorithm. Based on this, this paper proposes a method to solve the selection of initial value K. In the selection of K value, the range of clustering is limited according to the actual situation, that is, assuming that the range of clustering number K is (m, n), the traditional algorithm of n-m times K-means is carried out, and the optimal clustering number is selected from multiple clustering as the best clustering tree.

The specific process of the optimized K-means algorithm is shown in Figure 1.

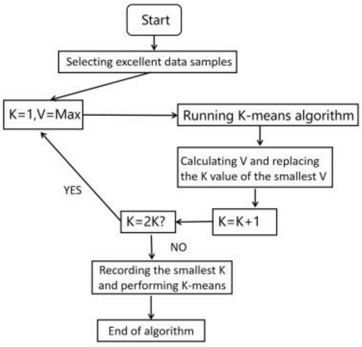
Start
Selecting excellent data samples
K=1,V=Max
Running K-means algorithm
Calculating V and replacing the K value of the smallest V
K=K+1
K=2K?
YES
NO
Recording the smallest K and performing K-means
End of algorithm

Figure 1: Improved K-means algorithm flow

2.2 Data Preprocessing of K-means Algorithm

The author analyzes the comprehensive evaluation data of the class with the author. The data comes from the comprehensive quality evaluation table of students in the spring of 2023. The original document is ' The comprehensive evaluation table of 2023 students in Automobile College.xls ', with a total of 700 students, all participating in the evaluation, and the evaluation results are real and effective. Through the improvement of K-means algorithm, based on the data preprocessing in the previous section, six dimensions of ' professional performance, sports performance, competition performance, scholarship level, student cadre status, and second classroom performance ' are selected as input variables. The maximum number of iterations is 10, and the data format after preprocessing is shown. The comprehensive evaluation score table contains all kinds of scores, academic performance and other attributes, and all the scores are percentage system, the minimum score unit is 0, no order of magnitude difference. In order to compare with the existing evaluation quantitative result data, the existing data dimension table is divided into different categories and valued according to the given weight coefficient. The integrated data are shown in chart 2, where K1 represents professional achievement, K2 represents sports achievement, K3 represents competition, K4 represents second classroom achievement, K5 represents scholarship level and K6 represents student cadre score.

The above data are integrated data based on different dimensions. Therefore, it is necessary to unify the dimensions of these different dimensions. The specific calculation is Equation (3).

$$\mathrm{r}_i = x_{ij}\sum_{j=1}^{m} x_{ij}\,(i=1,2,...,n;\ j=1,2,...,m) \tag{3}$$

Where Xij is the individual element value. The data to be tested after dimensional unification can be more true.

3. Results and Discussion

It can be seen from Figure 2 that with the increase of k value, CHS gradually decreases, and the decrease of CHS begins to slow down when k = 4, which indicates that the optimal clustering number is 4. In order to reduce the experimental error, the Improve-Kmeans algorithm is run 10 times in this paper. The CHS and contour coefficients obtained by each clustering are shown in Figure 2. Therefore, the student group portrait obtained from the experimental results is selected for analysis.

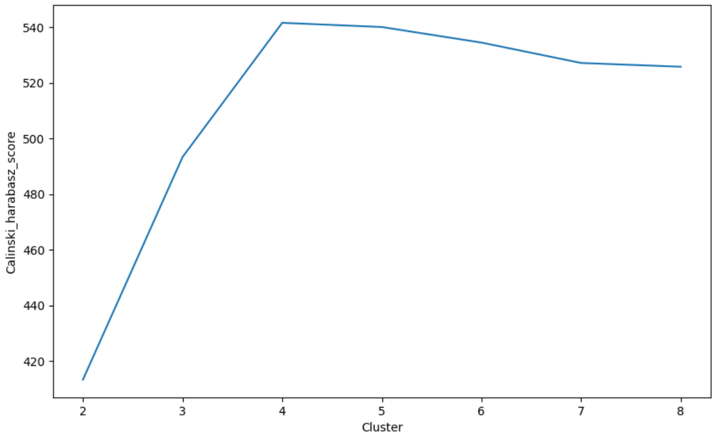

Figure 2: CHS of clustering results under different k values

Because the data is taken from the six dimensions of the students in the school, the optimized K-means algorithm divides the students into four categories, and uses the intelligent algorithm to reduce the dimension and classify. Among them, the first type of students have excellent performance, outstanding performance in extracurricular activities, and have enjoyed scholarships ; the second type of students have good grades, have certain extracurricular activities, and enjoy medium scholarships ; the third category of students have medium grades and some of them enjoy scholarships ; the fourth category of students is poor and not particularly active. Using cluster centers to represent each cluster is the most widely used cluster representation method. The Euclidean distance formula is used to calculate the distance between classes and analyze the differences between classes. In this clustering, four clustering centers are formed. Corresponding to excellent, good, medium and poor four levels. The relevant data are shown in Table 1 and Table 2.

Table 1: Euclidean distance between 4 classes

Cluster	Euclidean distance between classes			
	k1:Excellent	k2:Good	k3:Medium	k4:Poor
k1: Excellent	0	0.545	0.561	0.784
k2: Good	0.545	0	0.408	0.569
k3: Medium	0.561	0.408	0	0.512
k4: Poor	0.784	0.569	0.512	0

Table 2: Number and proportion of students by category

Cluster	Number	Percentage
k1: Excellent	115	16%
k2: Good	148	22%
k3: Medium	205	30%
k4: Poor	232	32%

It can be seen from Table 1 and Table 2 that 700 students in this class are clustered into four categories. The centers of each category are 115,148,205 and 232 students respectively. The four clusters correspond to four levels of good median difference. The number of excellent students in the comprehensive quality evaluation is 263, accounting for 38 %, the number of middle students in the evaluation is 148, accounting for 22 %, and the number of poor students in the evaluation is 205, accounting for 30 %. It can be seen from Table 1 that the distance between the cluster centers of the grades ' poor ' and ' medium ' and ' excellent ' and ' good ' is 0.784 and 0.561 respectively, indicating that the traditional concept of poor students, medium students and excellent and good survival are at a large distance, and the task of improving the quality of ' poor ' students and ' medium ' students is arduous. Because the sample used is the data of students in domestic higher vocational colleges, the overall Euclidean distance does not fit the normal distribution of ' large in the middle and small at both ends'. This is because the evaluation methods of higher vocational college students are not characterized by excellent academic performance. The data obtained show that the students and the middle students are more active in extracurricular sports activities and as student cadres. Middle school students also have a strong sense of collective consciousness, strong sense of activity innovation, and team spirit. If they want to improve their comprehensive quality and enter the excellent series, they need to enhance self-discipline and overcome their loose habits. Therefore, in the management of students, it is necessary to master the source of students, family environment, living habits, etc., and adopt the way of strengthening responsibility education with parents to improve the quality of all aspects.

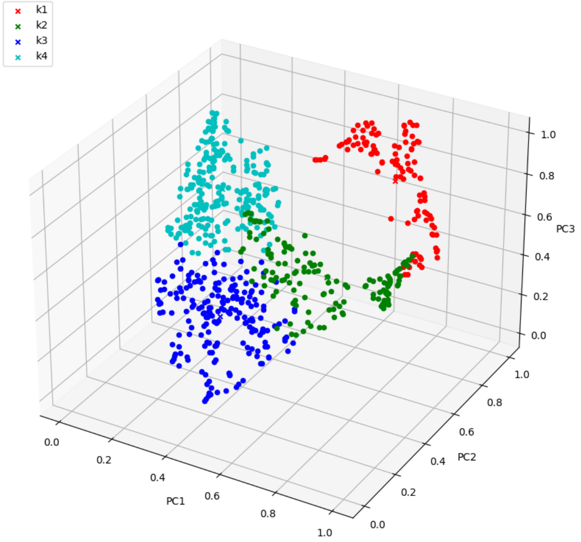

Figure 3: 3D cluster display of important features

As shown in Figure 3, in order to visually observe the effect of clustering, the first three features with the greatest impact are selected, and cluster analysis is visualized. Among them, red, green, purple and blue represent clustering 1, 2, 3 and 4 respectively, which correspond to four types of students in different schools. PC1, PC2 and PC3 represent the proportion of weighted effective data to total data. Through Figure 2, it can be clearly seen that there is a coincidence between the behaviors of students in categories 1, 2, 3 and 4, indicating that the behavior habits of such students are highly similar, only slightly different in grades and scholarship levels. The algorithm can basically distinguish students at different levels.

4. Conclusion

In this paper, students ' campus performance is selected for research. Through data analysis and mining of behavioral characteristics such as professional performance, sports activities and student cadre performance, clustering algorithm is used to establish a category model of students ' online performance, so as to realize the performance data of students ' learning life and classify students. Based on the data of educational administration system and student management system, this study preprocesses the 'dirty data' in the data system, and obtains the dimensional input data satisfying the K-means algorithm through data cleaning, integration and transformation. For the traditional K-means algorithm, the selection of clustering K and center points is easy to cause the deviation of the algorithm results. According to the actual situation, the range of clustering is limited to obtain the best clustering K value and center point. Finally, the student performance model is established through the core code of the algorithm, and the behavior characteristics of different types of students are analyzed to guide the daily management of students.

Acknowledgements

This paper is the initial result of the school-based project ' Research on the Essentials, Evolution Path and Framework Construction of Digital Literacy of Teachers in Vocational Colleges (Project Number: YJYB20230001) of Guangdong Mechanical & Electrical Polytechnic.

References

[1] Jiang N, Xu W. S. Analysis of student consumption and learning behavior based on campus card data. Microcomputer application, 2015, 31 (2):35-38.

[2] Xiong Z. Y, Chen R. T, Zhang Y. F. An effective K-means clustering center initialization method. Computer Application Research, 2011, 28 (11):4188-4190.

[3] Liu J. Ideological and political education management research based on improved K-means clustering, 2016 International Conference on Robots & Intelligent System. Zhang Jia Jie, China. IEEE, 2016: 213-216.

[4] Irfan S, Dwivedi G, Ghosh S. Optimization of K-means clustering using genetic algorithm, 2017 International Conference on Computing and Communication Technologies for Smart Nation. Gurgaon. IEEE, 2017: 156-161.

[5] Zhang J. , Yan M. Y. , Sun Z. Q. , et al. College student management based on big data Innovative research . China Education Informatization, 2020 (13):78-80.

[6] Huang G. , Liu R. , Liu H. F. , et al. Crowd portrait analysis based on campus card data. Computer and Digital Engineering, 2018, 46 (9):1881-1886.

[7] Jiang N., Xu W. S. Analysis of student consumption and learning behavior based on campus card data. Microcomputer Applications, 2015, 31 (2):35-38.

[8] Natek S, Zwilling M. Student Data Mining Solution - - Knowledge Management System for Universities. Expert System and Application, 2014, 41 (14): 6400-6407.

[9] Li X. M., Wu H. S. Design and implementation of college student informa- tion management system based on three-layer architecture. Computer Era, 2018(10):95-98.

[10] Rossi R, Gastaldi M, Orsini F. How to drive passenger airport ex- perience: a decision support system based on user profile. IET Intelli- gent Transport Systems, 2018, 12(4):301-308.

[11] Zhang L. J. Xu C. Q., Wang L. L., et al. Campus user portrait model based on multi-dimensional energy c-onsumption analysis. Renew- able Energy, 2021, 39(8):1078-1086.

[12] Green H S, Li X Q, De Pra M, et al. A rapid method for the detection of extra virgin olive oil adulteration using UHPLC-CAD profiling of triacylg- lycerols and PCA. Food Control, 2020, 107: 106773.

[13] Li X. M., Wu H. S. Design and implementation of college student information management system based on three-layer architecture. Computer Era, 2018(10):95-98.