Applied Mathematics and Nonlinear Sciences

<https://www.sciendo.com>

Analysis and Prediction of College Students' Mental Health Based on K-means Clustering Algorithm

Yanwu Liu[†]

Beijing Polytechnic, Beijing 100176, China

Submission Info

Communicated by Juan Luis García Guirao

Received June 17th 2021

Accepted September 24th 2021

Available online December 30th 2021

Abstract

Mental health is an important basic condition for the adult development of college students, and education workers gradually pay attention to the strengthening of mental health education for college students. In this paper, a psychological management system based on the K-means clustering analysis method is proposed. Based on the basic functions of the traditional system, the students' psychological data are reutilised by using the idea of data mining. By optimising the iterative process of the K-means algorithm, the valuable parts of a large number of precipitation students' psychological data are extracted. A data model is established and it provides decision guidance for managers to scientifically manage students' mental health process. The system establishes the data mining model in the process of analysis, carries on the mining to the students' psychological data in the database, analyses the different college students' mental health state characteristics and provides the corresponding solution. The part of the test drew on mental health data from 1,000 students at a school, and the results show that the system uses the k-means algorithm to divide students into 3 categories, which are 20.6%, 31.9% and 47.1% respectively, which are 1.6% different from the data test results in the ideal state.

Keywords: Data mining, psychological management system, k-means algorithm.**AMS 2010 codes:** 62M20

1 Introduction

Contemporary college students are in the most important stage of life, and they have higher cultural knowledge and wonderful life experience. But in the face of social pressure and life, employment and other difficulties, there are often problems faced by them such as inferiority, autism, extreme, withdrawn, depression and other psychological problems. Many social fields of contemporary college students' psychological research, research results show that the proportion of China's college students in need of mental health treatment is gradually increasing more and more because of the occurrence of their mental health problems which have garnered the

[†]Corresponding author.Email address: dzxyjjw@sina.com

attention of many in the society. Therefore, it is an inevitable trend to use data mining techniques to analyse the mental health data of college students to get the corresponding solutions.

In recent years, the mental health problems of college students have been of great concern in society, and scholars at home and abroad have been deeply exploring their mental health problems. At present, the psychological intervention model of college students has become the focus of many kinds of research. In our country, various education departments and universities have different measures in the aspect of college students' mental health education, and even some colleges and universities have gradually incorporated it into the training plan, set up special institutions to carry out the heart of the health care and medical institutions of many colleges and universities have opened psychological counselling centres to provide counselling for students. After providing psychological counselling, it is recorded in the archives to do work on the psychological activity status of the follow-up students during the school period. But there are still some colleges and universities in China which are lacking in mental health work, mental health education support is insufficient, equipment and institutions are not perfect, and there is a paucity of mental health-related personnel, which seems to be the current situation prevailing in colleges and universities.

Over time, concerning the number of mental health management centres, which stores mental health data, it is on the rise, and these data are a compilation of different mental health problems of different students; the collection of these problems with 'characteristics' for analysis, can effectively prevent the occurrence of similar mental health problems among students. To improve and optimise decision-making and improve the efficiency of psychological counselling for college students, it is necessary to conduct an accurate and timely analysis of information. Cluster analysis, as a non-intervention mode technology, has a good application market in the field of data mining. Cluster analysis can transform the miscellaneous and hidden data into actual decision-making activities and provide a more scientific reference basis. Therefore, it is of great significance to study the performance and application of different algorithms.

This paper constructs a cluster analysis model based on the K-means algorithm and applies the algorithm model to college students' mental health counselling work. Through long-term algorithm research, framework design and programme design, a complete mental health consultation system has been established, which provides scientific, objective and reliable data guarantee for the correct implementation of decision-making strategies. Cluster analysis is integrated into the process of college students' mental health analysis, making it more convenient for decision-makers to understand and master all aspects of students' information, to analyse and customise solutions for different types of students, and provide a good foundation for different groups of students to provide different psychological treatment methods.

The colleges and mental health workers are constantly exploring on how to improve the mental health level of college students, solve their mental health problems, and ensure that college students can move towards a modernised and healthy future. The problems and concerns of campus mental health services, as well as the challenges related to them, have also attracted much attention [1], and statistics-based analysis of the mental health status of college students has been developing [2]. Chang et al. proposed a fast search and search density peak (K-CFSFDP) algorithm based on K-means and clustering, which has a good clustering effect in large-scale campus data [3]. Chun et al. used the K-means clustering algorithm to analyse the results of graduate enrolment, classify the scores of graduate enrolment, and find out the distribution characteristics of students' grades and the relationship between grades [4]. Goh et al. used the deterministic model K-means clustering method to analyse students' comprehensive performance [5]. L et al. studied the application of the K-means clustering algorithm in the analysis of college students' online entertainment consumption [6]. Xiao et al. proposed a learning ability model of college students based on fuzzy cluster analysis [7].

K-means clustering methods to solve the mental health problems of college students are also developing. Based on mental health disorders, Pontoh et al. grouped four categories by the K-means clustering method and conducted a confirmatory factor analysis to determine the relationship between parental rearing behaviour and children's mental health [8]. Chu et al. applied the distance formula and clustering criterion function commonly used in cluster analysis to conduct a detailed analysis and research on the mental health status of

college students [9]. Wu et al. used knowledge network methods such as centrality, systematic clustering and K-cluster analysis to conduct statistical analysis on the data, find out the most significant abnormal psychology of college students and evaluate the relationship between different abnormal psychology of college students [10]. Nelsen et al. used hierarchical clustering analysis and K-means clustering analysis to study the mental health status of adaptive perfectionists and maladaptive perfectionists [11].

In this paper, a psychological management system based on the K-means clustering analysis method is proposed. Based on the basic functions of the traditional system, data mining is used to make secondary use of students' psychological data, analyse the state characteristics of different college students' mental health and provide corresponding solutions.

2 Research Methods

2.1 Introduction to K-means clustering algorithm

K-means algorithm, also known as K-average or K-mean, is a classic clustering algorithm adopted by many clustering tasks. It is the distance from the data point to the prototype as the objective function for optimisation. In this algorithm, the data set is divided into different categories through iterative operation by the method of finding the limit of a function so that the evaluation index J is minimised and the generated class distances are similar.

2.2 Dividing Clustering Method Three points in the clustering of data sets:

(1) Select a certain distance as the measurement between samples

The search process of the K-means clustering algorithm is limited to a part of all possible partition Spaces. If the sample similarity between each class is very low, the K-means algorithm can often achieve better results. However, if the sample similarity between the classes is high, the phenomenon of further division of clustering may occur. Therefore, it is possible to obtain the local rather than the global minimum solution of the scoring function due to the convergence of the algorithm.

Select criterion function

The K-means clustering algorithm will be affected by the selected similarity measurement method. The commonly used similarity measurement method uses the error sum of squares criterion function to improve the clustering performance. Suppose X contains k clustering subsets $X_1, X_2, \dots, X_k$; The number of samples in each cluster subset is $n_1, n_2, \dots, n_k$; If the mean points of each cluster subset are $m_1, m_2, \dots, m_k$, then the error sum of squares criterion function formula is:

$$E = \sum_{i=1}^k \sum_{p \in X_i} p - m_i^2$$

The similarity is calculated according to the average value of objects in a cluster.

Randomly assign K clustering centre

For each sample X_i , it is assigned to the nearest class according to the principle of minimum distance.

Move the sample mean in the cluster to the centre of the cluster.

Repeat Step 2 and Step 3 until the centre of the cluster does not change

The data object set S is shown in Table 1. As a two-dimensional sample of cluster analysis, the number of clusters required is $k=2$.

$O_1(0,2)$ and $O_2(0,0)$ were selected as the initial cluster centres, namely $M_1 = O_1 = (0,2), M_2 = O_2 = (0,0)$.

The remaining objects are grouped into the nearest cluster according to their distance from each central cluster. The O_3 :

$$d(M_1, O_3) = \sqrt{(0 - 1.5)^2 + (2 - 0)^2} = 2.5$$

$$d(M_2, O_3) = \sqrt{(0 - 1.5)^2 + (0 - 0)^2} = 1.5$$

Clearly $d(M_2, O_2) \leq d(M_1, O_3)$, so we assign O_3 to C_2

The O_4 :

$$d(m_2, O_4) = \sqrt{(0 - 5)^2 + (2 - 0)^2} = \sqrt{29}$$

$$d(M_2, O_4) = \sqrt{(0 - 5)^2 + (0 - 0)^2} = 5$$

Because $d(M_2, O_4) \leq d(M_1, O_4)$, so we assign O_4 to C_2

The O_5 :

$$d(M_1, O_5) = \sqrt{(0 - 5)^2 + (2 - 2)^2} = 5$$

$$d(M_2, O_5) = \sqrt{(0 - 5)^2 + (0 - 2)^2} = \sqrt{29}$$

Because $d(M_1, O_5) \leq d(M_2, O_5)$, so we assign O_5 to C_1

Update, get a new cluster $C_1 = \{O_1, O_5\}$ and $C_2 = \{O_2, O_3, O_4\}$

The square error criterion is calculated. The single variance is

$$E_1 = [(0-0)^2 + (2-2)^2] + [(0-5)^2 + (2-2)^2] = 25$$

$$M_1 = O_1 = (0, 2)$$

$$E_2 = 27.25$$

$$M_2 = O_2 = (0, 0)$$

The population mean variance is: $E = E_1 + E_2 = 25 + 27.25 = 52.25$

Calculate the centre of the new cluster

$$M_1 = ((0+5)/2, (2+2)/2) = (2.5, 2)$$

$$M_2 = ((0+1.5+5)/3, (0+0+0)/3) = (2.17, 0)$$

Repeat (2) and (3) to get O_1 assigned to C_1 ; O_2 is allocated to C_2 ; O_3 is allocated to C_2 ; O_4 is allocated to C_2 ; O_5 is assigned to C_1 . Update, get a new cluster $C_1 = \{O_1, O_5\}$ and $C_2 = \{O_2, O_3, O_4\}$

Center for $M_1 = (2.5, 2)$, $M_2 = (2.17, 0)$

The individual variances are

$$E_1 = [(0-2.5)^2 + (2-2)^2] + [(2.5-5)^2 + (2-2)^2] = 12.5$$

$$E_2 = 13.15$$

The overall mean error is: $E = E_1 + E_2 = 12.5 + 13.15 = 25.65$ +

After the first iteration, the average error is improved well. Since the cluster centre did not change in the two iterations, the iteration process was stopped and the algorithm was terminated.

3 Results analysis and discussion

(1) Determine classification. In this example, the optimal number of three categories is determined, and the clustering distribution table shows the frequency of each category. According to the data, 203 students are classified into cluster 1, 314 students in cluster 2, and the remaining 464 students are classified into cluster 3, among which 16 students are excluded by the system due to lack of data. As shown in Table 1.

Meanwhile, a two-dimensional distribution diagram of clustering analysis is obtained from the data, as shown in Figure 1.

As can be seen from the contour diagram, the second group of students is relatively close to the third group. Some of the interference data in the second group of students lead to duplication with the data of the third group of students, making some points fall into the third group of students, as shown in Figure 2.

The system adjusts the node accuracy from 10 up. The results show that students can be divided into three categories as a whole, as shown in Figure 3.

After adjusting, you can see an improvement over the previous side profile, as shown in Figure 4.

Table 1 Cluster distribution table

Name	Value	N	% of Combined	% of Total
Cluster 1	1	203	20.3%	20.6%
Cluster 2	2	314	31.4%	31.9%
Cluster 3	3	464	46.4%	47.1%
Combined		1000	100.00%	98.4%
Excluded		16		1.6%
Cases				
Total		984		100%

Fig. 1 Cluster analysis diagram of students' mental health status **Fig. 2** Profile of student data analysis

The final result diagram, distribution diagram, thermal energy diagram and probability diagram of cluster analysis are used to display the data results, as shown in Figure 5-9.

According to the AIC criterion, the optimal classification number is 3, and the corresponding AIC value is 8647.63.

To display the mean value and standard deviation of each variable in each cluster, the PivotTable was formulated according to the system requirements. As shown in Table 2.

In the table, Mean represents the Mean difference of the attributes of the corresponding cluster, and the

Fig. 3 Student cluster node diagram **Fig. 4** Profile of Student Data Analysis (After Adjustment) **Fig. 5** Results of cluster analysis

Fig. 6 Distribution

Fig. 7 Thermal energy diagram

Fig. 8 Probability Figure 1

Fig. 9 Probability Figure 2

Table 2 Distribution table of mean difference and variance values of all variables in clustering

		Adaptation and anxiety		Introversion and extroversion		Emotional and serene alertness		Cowardice and courage	
		Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation
Cluster	1	5.79	1.672	3.99	1.144	5.08	1.325	4.48	1.523
	2	3.68	1.894	6.96	1.111	5.93	1.956	4.47	1.428
	3	4.25	2.147	6.07	1.832	5.28	1.429	3.99	1.806
Combined		4.47	2.11	5.81	1.857	5.44	1.632	4.28	1.625

standard deviation represents its variance. According to the requirements of the system, the 16PF criteria were set and divided into 1 ~ 3 (Fail), 4 ~ 7 (Good) and 8 ~ 10 (Excellent). Since the average value of the tested group in the table tends to be between 4 and 7 points, it can be concluded that the psychological state of the student group is basically normal.

According to the statistical results of mean difference and variance of male and female information, P-value is less than 0.01, indicating that there is a significant difference between male and female students. The data showed that girls were higher than boys in environmental adaptability, extroversion and learning aggressiveness, and boys were better than girls in the decisiveness of execution. From the analysis and comparison of male and female students, it can be seen that the psychological state of female students is better than that of male students in most cases, while the problems of male students are mainly reflected in the aspects of personality and behaviour habits. As shown in Table 3.

Table 3 Differences between male and female students

Secondary personality factor	Male college student (N=1302)		Female college student (N=915)		T value	P value
	Mean	Std. Deviation	Mean	Std. Deviation		
Adaptation and anxiety	5.17	1.66	4.47	1.66	8.28	0
Introversion and extroversion	5.3	1.78	5.71	1.71	4.52	0
Emotional and serene alertness	5.19	1.41	4.48	1.46	9.73	0
Cowardice and courage	5.13	1.42	4.29	1.46	11.43	0

(1) As shown in Figure 10, the abscissa 0 is the horizontal standard for adaptation and anxiety in each cluster, and the ordinate is the cluster. A class less than 0 means below average, and a class above 0 means better than average.

Fig. 10 Proportion diagram of adaptation and anxiety in each cluster

In Figure 10, the anxiety value of the third group in adapting to the environment is much higher than the average level of other groups, indicating that this group of students have a low performance in adapting to the environment. Among them, women account for a higher proportion. They often show excitement and anxiety in their own situation, and excessive anxiety will not only affect their work and study efficiency but also harm their own health. The second group is much lower than the average level. In this category of students, they generally perform very well in the ability to adapt to the environment, and their mental state will not be affected by other factors, but a small number of them will have negative attitudes towards things. Compared with the first two categories, the first category is between the two categories and represents the average value of the whole group, showing a relatively stable state.

(1) As shown in Figure 11, a comparison was made on the attribute introversion and extroversion clustering, in which the first category was more introverted, while the third category was more extroverted.

Fig. 11 The proportion of cowardice and boldness in each cluster

The results show that the college students can be divided into three categories: normal (the first category), need of attention (the second category), focus on attention and related treatment. The first group had 1831 students, accounting for 82% of the total. There are 206 students in the second category, accounting for 9.24% of the total. The third category had 180 students, accounting for 8.06% of the total. The analysis results are as follows:

The personality structure of most students is harmonious, upward and positive.

Compared with female college students, the male college students have more physiological problems than the former. Male college students are more competitive, stubborn, independent and serious and responsible. Girls, by contrast, are sensitive and dreamy, but also impatient and easily upset. Due to the obvious differences between male and female students in personality factors, it is suggested that more consideration should be given to the students' family background and actual situation, and the differences of gender and personality should be paid more attention to the mental health level of female college students when conducting mental health education.

The proportion of each attribute clustering figure shows that the adaptability to the environment provides the easiest way to make students produce greater psychological effects when compared with previous attributes (such as whether to inside and outside, calm, timid and bold, etc.), which also played a role and also suggests that the counsellor at the same time attaches great importance to the freshman's adaptation to both personalities combined with the other three factors of influence.

We conducted statistics on the students who participated in the questionnaire survey, and the specific data are shown in Table 4.

Table 4 Statistical analysis of college student personality questionnaire (UPI)

Type	Level	The number of cases	The percentage
Good (Class C)	20 points in the following	1899	80.11%
General (Type B)	Choose "Yes" from 20 to 24 or 8, 16, and 26	260	10.20%
Key objects (type A)	More than 25 points	171	9.00%
Invalid	L=4 (5,10,15,50)	16	0.69%

It can be seen from the above table that our speculated result is similar to the conclusion of the questionnaire and the result of the UPI questionnaire, which proves that the cluster analysis in the system is feasible.

4 Conclusion

In this paper, a psychological management system based on the K-means clustering analysis method is proposed. Based on the basic functions of the traditional system, the idea of data mining is used to make secondary use of students' psychological data. Through the optimisation of the iterative process of the K-means algorithm, the characteristics of different mental health states of college students are analysed. In the test part, the mental health data of 1000 students from a school is extracted. The results show that the system uses the K-means algorithm to divide students into 3 categories, which are 20.6%, 31.9% and 47.1% respectively, which is 1.6% different from the test results of the ideal state

References

- [1] Juan H, Cao W, Liu X. A Dissolved Oxygen Prediction Method Based on K-Means Clustering and the ELM Neural Network: A Case Study of the Changdang Lake, China[J]. *Applied Engineering in Agriculture*, 2017, 33(4):461-469.
- [2] Feng Y, Zhao S, Liu H. Analysis of Network Coverage Optimization Based on feedback K-means Clustering and Artificial Fish Swarm Algorithm[J]. *IEEE Access*, 2020, PP(99):1-1.
- [3] Mc A, Suo J B. Analysis and research on mental health of college students based on cognitive computing[J]. *Cognitive Systems Research*, 2019, 56:151-158.
- [4] Xu L, J Lü. Bayberry image segmentation based on homomorphic filtering and K-means clustering algorithm[J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2015, 31(14):202-208.
- [5] Brockmeier A J, Principe J C. Learning Recurrent Waveforms Within EEGs[J]. *IEEE Transactions on Biomedical Engineering*, 2015, 63(1):43-54.
- [6] Zhang T, Ma F. Improved rough k-means clustering algorithm based on weighted distance measure with Gaussian function[J]. *International Journal of Computer Mathematics*, 2017, 94(1-4):663-675.

- [7] Li C, Lian S, Jia J, et al. Risk assessment of water pollution sources based on an integrated k-means clustering and set pair analysis method in the region of Shiyan, China[J]. *Science of the Total Environment*, 2016, 557-558(Jul.1):307-316.
- [8] Liu H, Xu Q, Xu H, et al. Judgment Method of Vehicle Lateral Stability Based on K Means Clustering Analysis[J]. *Hunan Daxue Xuebao/Journal of Hunan University Natural Sciences*, 2018, 45(8):48-53.
- [9] Preeti, Arora, Deepali, et al. Analysis of K-Means and K-Medoids Algorithm For Big Data - ScienceDirect[J]. *Procedia Computer Science*, 2016, 78:507-512.
- [10] Aidara S, Sagna Y. BSDEs driven by two mutually independent fractional Brownian motions with stochastic Lipschitz coefficients[J]. *Applied Mathematics and Nonlinear Sciences*, 2019, 4(1):151-162.
- [11] El-Borhamy M, Mosalam N. On the existence of periodic solution and the transition to chaos of Rayleigh-Duffing equation with application of gyro dynamic[J]. *Applied Mathematics and Nonlinear Sciences*, 2020, 5(1):93-108.

This page is intentionally left blank