Journal of http://www.satursonpublishing.com/
Research in Multidisciplinary Methods and Applications
ISSN: 2957-3920

Volume 3, Issue 5, May 2024

Analysis of Student Size at Different Stages in Guizhou Province Based on Improved K-Means Clustering Algorithm

Li Guangzhou, Yang Zhengyun, Lu Yanan

Xingyi Normal University for Nationalities, Xingyi 562400, Guizhou, China

Corresponding to: Li Guangzhou

Abstract: In view of the shortcomings of the original K-means algorithm that it is sensitive to distant group points and the K value is difficult to determine, an improved K-means clustering algorithm is designed on this basis, and the idea of elbow method is introduced to optimize the data to determine the number of clustering clusters K and the clustering center of clusters, and the cluster center is calculated on the basis of the K-means algorithm, and the improved K-means algorithm has a significantly better effect than the K-means algorithm. An improved k-means clustering algorithm was proposed to analyze the number of students at each stage in Guizhou Province. On this basis, the influencing factors are briefly analyzed and the following countermeasures are put forward to scientifically grasp the trend of population change and optimize the supply of various types of educational resources at all levels. Based on the urban development planning layout of prefecture-level cities, strengthen the overall design of the education of the children of the floating population; Improve mechanisms for sharing high-quality educational resources, and promote the high-quality and balanced development of compulsory education.

Keywords: Student Size; K-Means Clustering; Elbow Method

1 Introduction

Since the 18th National Congress of the Communist Party of China, the popularization level of basic education has been greatly improved, the conditions for running schools have been significantly improved, the institutional system has been basically improved, and the quality of education has been steadily improved. Education is related to the development of the country and the future of the nation, and access to education has always been regarded as an important way for disadvantaged groups, especially poor children in rural areas, to increase their income and achieve class leapfrogging. Since 1978, the development of education has gone through a stage from elite to popular and then to the current popularization, and the overall level of education in China has reached a new height. Moreover, in terms of education development, General Secretary Xi Jinping put forward the scientific assertion that "a prosperous education will lead to a prosperous country, and a strong education will make the country strong".

In recent years, the number of students in school is often used to measure the value of real estate investment in a city, because in most cases, behind every student, there is at least one family settled in the city, which will generate residential and consumption needs. However, the level of school education varies from city to city, so this paper studies the distribution of education in different cities. For the distribution of education, this paper uses k-means clustering to coarse cluster education in different cities, and then uses the improved k-means clustering algorithm to optimize clustering[1].

2 Basic Theory

2.1 Selection Of Indicators

First of all, we use the number of students at different stages as the main indicator. Second, we select secondary vocational education, junior high school education, high school education, primary education, and kindergarten education. Among them, primary and junior high schools belong to nine-year compulsory

Journal of http://www.satursonpublishing.com/
Research in Multidisciplinary Methods and Applications
ISSN: 2957-3920

Volume 3, Issue 5, May 2024

education, and senior high schools belong to senior secondary education.

The number of students enrolled refers to the number of students enrolled in full-time colleges and universities, which is equivalent to the number of undergraduate students, which is an important indicator in the education plan and reflects the total scale of the development of education. That's why we chose it as an indicator.

2.2 K-Means Clustering

2.2.1 K-means clustering

The k-means clustering algorithm is an iterative clustering algorithm that follows:

(1) Set K clustering center points;

(2) Place these K points randomly to complete initialization;

(3) the Euclidean distance from each sample point in the original dataset to the center point of each cluster was calculated separately, and the sample points were classified into the class where the nearest cluster center point was located, and after a round of operation, the original dataset was divided into K classes;

(4) According to the sample points contained in K classes, the mean center of the sample points in each class was recalculated as the new clustering center point;

(5) The classification results of the sample points were updated again, and the distance between the sample points was recalculated according to the new clustering center points and K classes were divided.

(6) Repeat steps 4 and 5 until the clustering center to which all sample points belong does not change or satisfies a certain number of iterations. Each time a sample is assigned, the cluster's center of clustering is recalculated based on the objects existing in the cluster. This process will be repeated until a certain termination condition is met. The termination condition can be that no (or minimum) number of objects are reassigned to different clusters, no (or minimum) clustering centers are changed again, and the squared and local errors are minimal.

The K-means clustering algorithm also has some shortcomings, it has a good aggregation effect on spherical clusters, but poor performance on other forms of cluster clustering: the K value of the K-means algorithm is artificially set, and the clustering effect of different K-values is significantly different: the initial value of the K value is randomly generated, and if the initial value happens to take the poor initial centroid, the centroid may be in the local optimal when it is updated, which will lead to the termination of the program. The K-means measure the similarity of the points in the cluster by distance, and in many applications other parameters such as time are needed as clustering parameters, and appropriate improvements to the distance measurement are required.

2.2.2 Improve the K-means clustering algorithm

There are six commonly used data mining techniques: cluster analysis, decision tree, neural network, regression, association rule, and Bayesian classification [2]. Among them, cluster analysis is a method to divide data according to their internal characteristics without the help of external information, and it is also one of the tools to realize data mining. As an indispensable technology for the application of big data analysis, cluster analysis has been widely used in various scenarios in life. Among the many clustering algorithms, the K-means clustering algorithm has been studied at a deeper level and has been maturely applied in some fields, but at the same time, it has also revealed obvious shortcomings: the importance difference between clustering indicators is difficult to measure, the optimal number of clusters is difficult to determine, and the initial clustering center is difficult to select, which will affect the clustering effect to a certain extent [3]. In order to solve the above problems, this paper optimizes the traditional K-means clustering algorithm. In this paper, the SSE-based elbow method is used to determine the number of clusters, where the sum of squares of error (SSE) is defined as follows:

$$SSE = \sum_{i=1}^{k} \sum_{p \in i} \|p - q_i\|_2$$

where p represents the data objects in group i, the mean of all data objects in group i, and k represents the number of clusters. $\theta_i q_i$

With the increase of the number of clusters k, the sample division will be more fine, and the degree of aggregation of each cluster will gradually increase, so the sum of the square of the error will naturally become smaller.

When k is less than the true number of clusters, because the increase of k will greatly increase the degree of aggregation of each cluster, the decrease of SSE will be very large, and when k reaches the true number of clusters, the return of the degree of aggregation obtained by increasing k will quickly become smaller, so the decline of SSE will decrease sharply, and then tend to flatten as the value of k continues to increase, that is to say, the relationship between SSE and k is the shape of an elbow, and the k-value corresponding to this elbow is the real number of clusters of the data. The elbow method is common in many methods for selecting the best parameters, and the optimal k-value is found by finding the turning point of the inertia (the sum of squares within the cluster) [4-5].

3 Empirical Analysis

3.1 Data Sources

Table 1 is based on the 2021 Statistical Yearbook of Guizhou Province, which analyzes the number of students in each city in Guizhou Province at different stages.

Journal of http://www.satursonpublishing.com/
Research in Multidisciplinary Methods and Applications
ISSN: 2957-3920

Volume 3, Issue 5, May 2024

TABLE 1 NUMBER OF STUDENTS ENROLLED IN EACH CITY IN GUIZHOU PROVINCE AT DIFFERENT STAGES

School Stage	Secondary Vocational Education (People)	High School (People)	Junior High School (People)	Primary School (People)	Kindergarten (People)
Guiyang City	97015	90803	173824	470340	220430
Liupanshui	15298	74532	137097	343291	156045
Zunyi City	67089	154577	293349	612123	259096
Anshun City	24212	52339	124873	271887	113993
Bijie City	36609	200363	401995	848092	303250
Tongren City	37351	116253	167448	340241	144330
Qiannan	25093	93732	165591	329517	130288
Qiandongnan	38151	101219	183106	398870	168738
Qiannan	56697	81758	152623	348885	159177

3.2 EXPERIMENTAL RESULTS OF K-MEANS ALGORITHM

The data in Table 1 were divided into two groups, and the number of students in each stage of the two regions was randomly selected as the initial clustering center, and then the distance between each object and each seed clustering center was calculated, and each object was assigned to the nearest clustering center, and the clustering center and the objects assigned to them represented a cluster.

Iterate with R, as shown in Table 2,

TABLE 2 PREFECTURE-LEVEL CITIES IN GUIZHOU PROVINCE ARE DIVIDED INTO TWO GROUPS

GY	LBS	ASS	TRS	QXN	QDN	QN	ZYS	BJS
1	1	1	1	1	1	1	2	2

The results showed that Guiyang City (GY), Liupanshui (LBS), Anshun City (ASS), Tongren City (TRS), Southwest Guizhou City (QXN), Southeast Guizhou City (QDN) and South Guizhou (QN) were clustered into one category, and Zunyi City (ZYS) and Bijie City (BJS) were clustered into one category.

Iteratively update the cluster center point and its sample area until it is unable to update. The location of the cluster center is very important when initializing. In the K-means algorithm, because different initial positions may lead to different results, it is generally required that the clustering centers be spread out as much as possible. The clustering is not visible from Table 2

Whether the centers are spread out as much as possible, then use the k-means clustering algorithm to run Figure 1 with R, and plot the graph as follows:

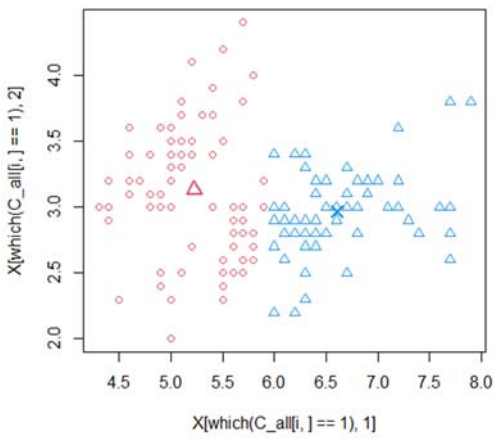

FIGURE 1 DATA FOR K-MEANS CLUSTER ANALYSIS

It can be seen that the clustering centers in Figure 1 are spread out as much as possible, and because different initial positions may lead to different results, the clustering effect in Figure 1 is not the best.

Therefore, the improved k-means clustering algorithm elbow method is used to determine the initial center. Organize your data

The result of the subsequent run is as follows:

Journal of http://www.satursonpublishing.com/
Research in Multidisciplinary Methods and Applications
ISSN: 2957-3920
Volume 3, Issue 5, May 2024

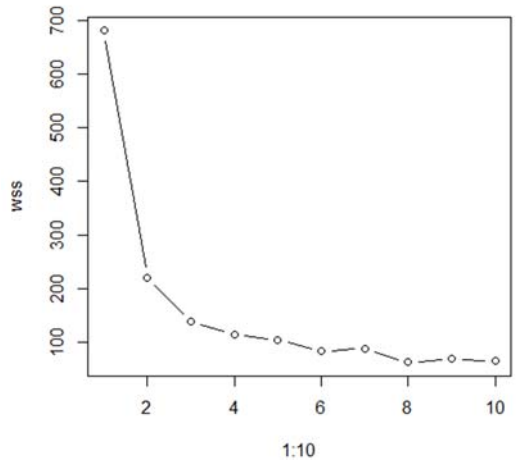

FIGURE 2 THE ELBOW METHOD DETERMINES THE OPTIMAL K-VALUE METHOD

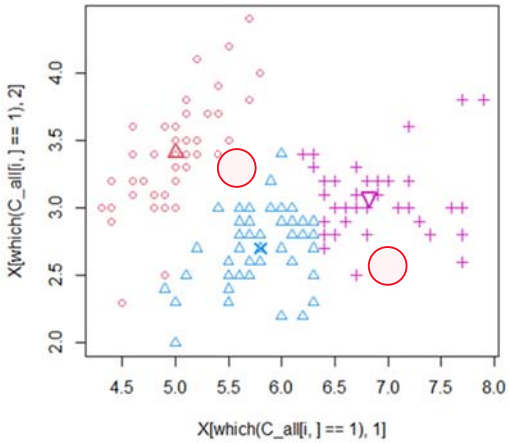

FIGURE 3: THE NUMBER OF CLUSTERS IS A RESULT OF 3 CATEGORIES

From Figure 2, the optimal k value method is determined by the elbow method to determine the k value of 3, so that the clustering results can be divided into three categories, and there is a result plot in Figure 3.

It can be seen from the clustering effect in Figure 3 that the clustering effect is better than that in Figure 1, and the clustering centers are scattered, and the regions are classified on the basis of the K-means clustering algorithm, and the classification results are shown in Table 3:

TABLE 3: EACH PREFECTURE LEVEL CITY IN GUIZHOU PROVINCE IS DIVIDED INTO THREE GROUPS

GY	ZYS	BJS	LBS	ASS	TRS	QXN	QDN	QN
1	1	2	3	3	3	3	3	3

4 CONCLUSIONS AND RECOMMENDATIONS

1. The optimal K value was determined by the elbow step method, and an improved K-means algorithm was proposed on the basis of the K-means algorithm for classification, and the number of educated people in Guizhou Province was divided into three categories, namely Guiyang City and Zunyi City. Bijie City is divided into one category; Liupanshui, Anshun City, Tongren City, Qiannan Prefecture, Qiandongnan Prefecture, and Qianxinan Prefecture are divided into one category. Guiyang City and Zunyi City have fewer education people, because the economic development is better than other regions, so the educational resources are also relatively superior to other regions, on the contrary, Bijie City has more school students, but the economy is equivalent to backwardness, so the educational resources are relatively backward compared to other regions.

2. This study provides a point for the balanced distribution of educational resources among all classes in Guizhou Province and promotes social equity: Balanced allocation of educational resources can reduce the gap between social classes and promote social equity and the stability of social development. Improving the quality of education: Balanced allocation of educational resources can improve the quality of education because every student has access to the same educational resources and opportunities to better reach their potential. Cultivating all-round talents: Balanced allocation of educational resources can cultivate all-round talents, because every student has the opportunity to receive a diversified education, including culture, art, sports and other education. Promote national development: A balanced allocation of educational resources can contribute to the development of a country, as every student has the opportunity to receive a high-quality education and thus contribute to the development of the country.

FUNDING

This article is the final paper of the Guizhou Provincial Youth Science and Technology Talent Growth Project, "Improvement and Application of K-means Clustering Algorithm", Qianjiaohe KY[2020]214.

ABOUT THE AUTHOR

Li Guangzhou, born in Xingyi, Guizhou, in 1989.03, is a graduate student majoring in statistics.

Yang Zhengyun, born in Zhenning, Guizhou Province, in 1991.07, graduate student in statistics, research direction: applied statistics.

Lv Yanan, born in Huaibei, Anhui Province, was born in 1991.01, graduate student majoring in statistics, research direction: applied statistics.

REFERENCES

[1]Zeng Ruming. Research on the improvement and application of K-means clustering algorithm[D].China West Normal University,2022.DOI:10.27859/d.cnki.gxhsf.2022.000453.

[2]Gao Xin. Research on an improved K-means clustering algorithm and a new clustering effectiveness index[D].Anhui University,2020.DOI:10.26917/d.cnki.ganhu.2020.000960.

[3]WU Guangjian,ZHANG Jianlin,YUAN Ding. Research on automatic method of obtaining K value based on elbow method based on -means[J].Software,2019,40(05):167-170.)

[4]Wang Jianren, Ma Xin, Duan Ganglong. Improved K-means clustering k-value selection algorithm[J].Computer Engineering and Applications,2019,55(08):27-33.)

[5]Hu Yuping, Zhang Yue, Chen Deyun. Analysis of population and resource allocation in compulsory education stage in Beijing[J].China National Conditions and National Strength,2022(12):59-64.DOI:10.13561/j.cnki.zggqgl.2022.12.013.