Application of the intra cluster, characteristic of k-means clustering method in English score analysis in Colleges

Qin Wang

1Guilin University of Electronic Technology Beihai Campus, Beihai 536000, Guangxi, China

Abstract. As an important data mining method, clustering algorithm can find valuable information hidden in data objects through unsupervised classification method. It is widely used in data analysis, image segmentation, feature learning and other fields. It is one of the common methods in traditional machine learning algorithm. K-means clustering algorithm is a partition-based algorithm in clustering analysis, which is efficient and accurate, easy to implement, and is generally implemented in machine learning and other fields. In this paper, based on the characteristics of K-means clustering method, the scores of each part of 30 college students' English scores are used as the input data, and the cluster center is used as the influencing factor to participate in the clustering process. The clustering effect is evaluated under the average distance of the characteristics within the cluster after several iterations.

1. Introduction

As one of the most important information carriers, English has become the most widely used language in human social life in the era of economic globalization. The importance of English level for contemporary college students in our country is much necessary. To find out what kind of English teaching method is most suitable for students is the problem that college English education needs to face. The analysis of students' English score is a way to improve students' English level. The commonly used score analysis methods include correlation analysis, decision tree algorithm, association rules and cluster analysis. For example, in 2019, Liu Yifan et al.[1] used cart decision tree algorithm to analyze and study the scores of computer network courses, and generated a student score classification model based on cart tree. After ten fold cross validation, the accuracy of the model go rise to 78.2%. Finally, according to the analysis results, reasonable and feasible teaching suggestions are put forward. In 2020, Sun Fei et al.[2] used the Apriori algorithm to analyze the correlation between the high and low grades of each subject in the senior high school entrance examination. The support and confidence of the final frequent item set are analyzed to determine the correlation influence of each section of the subject, so as to provide a reference method for improving students' score.

Among all kinds of score analysis methods, cluster analysis method is one of the most widely used and most promising. Cluster analysis classifies the research objects according to their characteristics to reduce the number of research objects. It is a statistical analysis technology that divides the research objects into relatively homogeneous clusters. The advantage of this method is that its conclusion form is simple and clear, and it is easier to find the implied law.
There are many kinds of clustering analysis methods, such as k-means, options, K-means etc. In 2019, Guo Wenjuan [3] used neighborhood based k-medoids to cluster analysis the scores of 112 students in five courses in a semester. It is observed that the cluster centers after running the algorithm are far away from each other, and they are all located in the dense area of the data set, which achieves good clustering effect. However, when the number of objects to be clustered is too many, k-medoids algorithm will have a slow convergence speed and easy to produce clustering bias. However, other hierarchical clustering methods [4] have some problems, such as too much computation and easy clustering into chains. Compared with other score analysis algorithms, the k-means algorithm used in this paper has the characteristics of simple calculation, easy implementation, high efficiency and local optimal solution. As one of the most common clustering algorithms, K-means algorithm has been used in score analysis. In 2017, ye Fulan [5] analyzed the students' scores based on K-means algorithm, and found that the appropriate cluster number was 4, and the iteration number was 3, and the minimum distance between the initial centers was 36.865. In 2018, Li Zhen and others conducted K-means cluster analysis on Freshmen's entrance scores. According to the analysis results, the students were divided into four categories, and the comprehensive ranking of students was obtained. The experimental results show that the K-means cluster analysis method can objectively evaluate the learning situation of college students. In this paper, three basic parameters of K-means clustering algorithm [6] are used to analyze college English scores: the number of data subsets, the initial clustering center and the similarity measure.

2. Principle
K-means is a common clustering algorithm, its principle is: for a given sample set, the distance between samples is used to measure the similarity between data objects, the sample set is divided into K clusters (K is the user specified parameter, that is, the number of clusters expected), the center position of each cluster is determined, the distance between the points in the cluster and the center of the cluster is calculated, and then redivided cluster center. Repeat this operation continuously to make the points in the cluster as close as possible, and make the distance between clusters as large as possible, until the cluster does not change, or until the cluster center does not change. The basic data expression is as follows:

\[ E = \sum_{i=1}^{k} \sum_{p \in C_i} \left| p - m_i \right|^2 \]  

(1)

In the above formula, \( E \) is the sum of square errors of all objects in the cluster, \( p \) is the known data object, and \( m_i \) is the cluster center of cluster \( C_i \).

\[ m_i = \frac{1}{|C_i|} \sum_{p \in C_i} p \]  

(2)

The purpose of K-means clustering is to minimize the sum of squared errors \( E \). The calculation steps are as follows:

1. Input data and specified K value (number of clusters).
2. K points in the data space are selected as the initial clustering center \( C_i \) (\( 1 \leq i \leq K \)), and each point represents the center of this category.
3. The distance between all input data and cluster center \( C_i \) is calculated. For the K clusters, find out all the points belonging to the cluster, and then modify the coordinates of the cluster to the midpoint of these data points.
4. If the algorithm converges, output the result. If the algorithm does not converge, repeat step 3. The iterative process of K-means clustering method can also be shown in Figure 1.
3. Experimental design
The data used in this experiment come from the final scores of 30 English majors. According to the scores of listening, speaking, reading, translation and writing, this paper analyzes the k-means algorithm. The score of each part is shown in Table 1.
Table 1. 30 students scored in each part of their final English scores.

| Number | Listening | Speaking | Reading | Translation | Writing |
|--------|-----------|----------|---------|-------------|---------|
| 1      | 19        | 18       | 18      | 19          | 17      |
| 2      | 17        | 17       | 16      | 17          | 16      |
| 3      | 15        | 15       | 20      | 20          | 17      |
| 4      | 18        | 17       | 19      | 18          | 18      |
| 5      | 16        | 16       | 17      | 18          | 20      |
| 6      | 18        | 18       | 20      | 17          | 20      |

Before the K-means clustering analysis, it is necessary to standardize the data, so that the data can be converted to the same level, and then the data can be compared. The processing method is: divide the score of each part and the total score of this part, so that the standardized value is in the range of [0,1].

The standardized data table is shown in Table 2

Table 2. Standardized scores of each part of 30 students' final English scores.

| Number | Listening | Speaking | Reading | Translation | Writing |
|--------|-----------|----------|---------|-------------|---------|
| 1      | 0.95      | 0.9      | 0.9     | 0.95        | 0.85    |
| 2      | 0.85      | 0.85     | 0.8     | 0.85        | 0.8     |
| 3      | 0.75      | 0.75     | 1       | 1           | 0.85    |
| 4      | 0.9       | 0.85     | 0.95    | 0.9         | 0.9     |
| 5      | 0.8       | 0.8      | 0.85    | 0.9         | 1       |
| 6      | 0.9       | 0.9      | 1       | 0.85        | 1       |

The k-means algorithm is used to cluster students' final English scores. Set the number of clusters to 4, divide the students with the same characteristics into a group, and get the cluster center and the number of samples in the cluster, as shown in Table 3.

Table 3. Sample data after clustering analysis by k-means algorithm.

| Cluster center | Number | Listening | Speaking | Reading | Translation | Writing |
|----------------|--------|-----------|----------|---------|-------------|---------|
| 1              | 4      | 0.942     | 0.923    | 0.945   | 0.922       | 0.905   |
| 2              | 13     | 0.89      | 0.817    | 0.901   | 0.912       | 0.831   |
| 3              | 10     | 0.793     | 0.904    | 0.888   | 0.824       | 0.761   |
| 4              | 3      | 0.631     | 0.645    | 0.864   | 0.766       | 0.75    |

4. Discussion and analysis

Visually analyze the data processed by K-means algorithm, and draw the radar chart as shown in Figure 2.
According to figure 2, it’s clearly showed the strength and comprehensive ability of each class. Then draw the cluster figure of the four clusters, as shown in Figure 3, it also can be intuitively seen the hierarchical phenomenon of the results of each cluster.

Set the score in the range of 0.9-1 as excellent, the score in the range of 0.8-0.9 as good, the score in the range of 0.7-0.8 as general, and the score in the range of 0-0.7 as poor. According to figure 3, it is easy to describe the characteristics of each cluster as follows:

Cluster 1: the number of students accounts for 13% of the total number. These students' listening, speaking, reading, translation and writing are in the excellent range, and they have a good grasp of knowledge, without weak items.

Cluster 2: the number of students accounts for 43% of the total number. These students have a high level of reading and translation, and they are in the excellent range, while listening, speaking and writing are in the good range, and their overall performance is excellent, but their should pay attention to improve their oral English.
Cluster 3: the number of students accounts for 33% of the total number. Their oral English is in the excellent range, and their reading and translation are also relatively good, while their listening and writing are relatively general. Their should pay attention to improve their listening and writing.

Cluster 4: the number of students accounts for 10% of the total number, the overall score deviation, only the reading level is in a good level, while the translation and writing are in a general level, listening is in a poor level, this part of the students should pay attention to all aspects to improve their English level.

Table 4. Sum of the square of the four clustering results and the average distance within the cluster.

| Cluster (cluster center) | Sum of squares of average distance within the cluster |
|--------------------------|-----------------------------------------------------|
| Cluster1(0.942,0.923,0.945,0.922,0.905) | 0.23 |
| Cluster2(0.89,0.817,0.901,0.912,0.831) | 0.31 |
| Cluster3(0.793,0.904,0.888,0.824,0.761) | 0.64 |
| Cluster4(0.631,0.645,0.864,0.766,0.75) | 0.84 |

The sum of squares of the average distances within the four clusters is calculated, and the results are shown in Table 4. By comparing the sum of squares of the intra cluster distance in Table 4, it can see that the sum of squares of the average distance of cluster 1 is the smallest. In the clustering results, the distance between the data is closer, and the distribution of the data is closer, that is, the distribution of the students in cluster 1 is more stable while their grades are excellent. The students of Cluster 4 have larger sum of square of average distance within the cluster, and the data distribution is loose and unstable. However, the maximum sum of square of average distance within the cluster using k-means clustering method is only 0.84, which shows that the aggregation effect is excellent when the number of aggregation objects is 30.

5. Conclusions

By using the characteristics of K-means clustering method to analyze the students in different clusters, schools can clearly understand the strengths and weaknesses of each student, and make them more objective to find a more suitable learning method. According to the clustering results, different types of students are trained according to different teaching methods, which not only improves the teaching efficiency, but also saves teaching resources to a certain extent. The method of dividing students according to each part of the students' English score avoids the single defect brought by only using the total score to evaluate the students' English achievement. It can more scientifically and pertinently reflect the students' English learning situation, and can provide meaningful guidance for the school to provide students with follow-up teaching programs. It can be seen that K-means clustering method is of great significance to the analysis of College English scores. It shows that the algorithm can provide a reference for the classification and analysis of students with different levels and characteristics, realizing different teaching for each student, helping to improve the quality of College English education, and cultivating students' English level with appropriate resources.

Acknowledgement

This work is supported by the Undergraduate Education Reform Project of Guangxi Higher Education in 2020 (Grant No. 2020JGA182); the second batch of Production-learning-Researching Cooperative Education Project of the Ministry of Education in 2019 (Grant No. 201902079011), the Teaching Reform Project of Guilin University of Electronic Technology (Grant No. JGBBH202001), and the Basic Scientific Research Ability Improvement Project for Young and Middle-aged Teachers of Guangxi’s University (Grant No. 2021KY0184).
References

[1] Liu Yifan, Zhan Jing, fan Xue. Analysis of students' achievement in computer network course based on cart decision tree [J]. Computer education, 2019(12):76-79+84.

[2] Sun Fei, Ren Fudong, Shi Shengyong, Chen Jia, Zou lulu. Research on the application of Apriori algorithm in the analysis of high school entrance examination results [J]. Journal of Beihua Institute of aerospace technology, 2020,30 (04): 26-28

[3] Guo Wenjuan. Application of neighborhood based K-medoids clustering algorithm in student performance evaluation [J]. Gansu science and technology, 2019,35 (22): 27-29

[4] Zhang Wenbin, Ming Yong, Chu Weiwei, Huang Zhezhi. Research on shopping basket compression based on clustering algorithm [J]. Computer technology and development, 2018,28 (01): 169-173 + 177

[5] Ye Fulan. Analysis of student achievement based on K-means algorithm -- taking information management and information system major of Fuzhou University of Foreign Studies and trade as an example [J]. Journal of Guiyang University (NATURAL SCIENCE EDITION), 2017,12 (03): 17-20

[6] Wang Qian, Wang Cheng, Feng Zhenyuan, ye Jinfeng. Review of K-means clustering algorithm [J]. Electronic design engineering, 2012,20 (07): 21-24