Eng Int Syst (2024) 3: 89–93

Engineering
Intelligent Systems

A Clustering Analysis of Students' English Scores After Targeted Improvement

Xia Sun*

Department of Foreign Language, Hefei Normal University, Hefei, Anhui 230061, China

In every school, the analysis of students' scores is an important means of improving teaching outcomes. By analyzing these scores, pedagogical shortcomings can be identified and addressed so that students' results in the future are better. However, the traditional way of assessing by using scores alone does not adequately capture all the information contained in students' scores. In this study, the English scores obtained by 100 students selected for targeted improvement, were analyzed using the K-means clustering algorithm. The students were classified into four categories based on the elbow method. The first category had the highest average score for composition (22.00); the second category had the highest average score for listening (25.60) and reading (26.60); the third category had the highest average score for word selection (10.25) and reading (25.75); the average scores of students in the fourth category were all unsatisfactory. The results of the study demonstrated that the K-means algorithm can analyze the characteristics of students' English scores effectively, which can provide a sound basis for teachers to improve teaching methods and take appropriate measures to improve students' English scores.

Keywords: clustering analysis, K-means clustering algorithm, score analysis

1. INTRODUCTION

In this study, students' academic performance was analyzed after mid-term, final and practice examinations, enabling teachers to take specific measures to improve students' scores. Traditional score assessment involves using the total score to place students into categories, and only the students know their scores and ranking. For example, a score below 60 points represents a fail, 60–70 represents a pass, 70–90 represents good, and 90–100 represents excellent. However, this assessment method is not accurate and lacks in-depth and detailed analysis. For instance, there is only a small difference between 79 and 80, but these two students cannot be appropriately classified for targeted teaching. Pedagogical methods cannot be improved and targeted teaching cannot be performed to address the students' specific learning needs if teachers do not know each student's particular area of weakness. An overall result does not provide this sort of information. The analysis of students' scores can provide teachers with more comprehensive information on students' strengths and areas needing improvement, and teachers can adjust their teaching methods accordingly. Clustering analysis is an important method used to analyze data, and includes fields such as data mining (Zhou, 2022). This rapidly-developing method is widely used in statistics (Jones et al., 2021), image processing (Lei et al., 2018), document classification (Sardar and Anrisa, 2018), market segmentation (Hung et al., 2019) etc. Clustering analysis involves dividing a set of physical or abstract objects into multiple categories comprising similar objects, and objects in different categories are different (Jing et al., 2021). Clustering analysis can classify individuals and determine the characteristics of each category objectively and logically. The K-means algorithm is an iterative clustering algorithm. Wang et al. (2021) pointed out that K-means algorithm is a useful clustering

*Corresponding address: No. 1688, Lianhua Road, Jingkai District, Hefei, Anhui 230061, China. Email: xiazhi7286@126.com

analysis algorithm, while Xu et al. (2018) believed that the genetic algorithm and the K-means algorithm can increase the efficiency and accuracy of clustering. Chen (2022) found that the K-means algorithm can balance the parallel computing capability, accuracy and required iterations, and used case analysis to verify the effectiveness of the K-means algorithm. Zhu et al. (2022) used the K-means clustering algorithm to determine the change of water content under different temperatures, and further determined the irrigation strategy. The paper used the K-means algorithm to analyze students' English scores after targeted tutoring and discussed whether the K-means clustering algorithm was feasible for score analysis, thereby helping educators better understand students' strengths and weaknesses in English and conduct targeted tutoring.

2. K-MEANS CLUSTERING ALGORITHM

Clustering is used to divide the data into different categories thereby allowing it to be compared and differences to be identified. The data in each cluster should be as similar as possible, while each cluster should be as different from the others as possible. The final number of clusters in cluster analysis is completely independent of external conditions, and is completely determined by the attributes of data objects and their distribution characteristics. The K-means clustering algorithm is a partition-based algorithm applied to clustering analysis, and also an unsupervised learning algorithm (Li et al., 2021). The basic idea of the K-means algorithm is to cluster k-points and group the most similar data objects under the same category. Through the iterative method, the value of every cluster center is updated gradually until the best clustering result is obtained. The final clustering results should enable the data objects with strong similarity to be placed into the same cluster, while the data objects with large differences should be divided into different clusters.

The specific flow of the K-means clustering algorithm is as follows.

(1) k samples are randomly selected from dataset X as initial cluster centers.

(2) Based on equation (1), the Euclidean distance is used to calculate the similarity between the remaining samples in the space:

$$d(x, c_i) = \sqrt{\sum_{j=1}^{m} (x_j - c_{ij})^2}, \tag{1}$$

where X represents data object, c_i represents the i-th cluster center, m represents the dimension of the data object, and x_j and c_{ij} represents the j-th attribute value of x and c_i.

(3) According to the calculated distance, the samples are categorized to cluster center $c_i c_i$ with the shortest distance, and the next clustering is performed again according to equation (2):

$$c_t = \frac{\sum_{x \in s_t} x}{|s_t|}, \tag{2}$$

where c_t is the center of the t-th cluster center and $|s_t|$ is the number of data objects in the t-th cluster.

(4) Steps (2) and (3) are repeated until there is no obvious change in the classified value; otherwise, the iteration continues (Kang et al., 2020).

3. CASE ANALYSIS

3.1 Data Preprocessing

(1) Data selection

The data selected for this experiment were the final English scores of students who passed the test after two weeks of consistent and targeted teaching of listening, word selection, reading, and writing. During the two-week targeted learning program, the teaching of listening consisted of teachers explaining the test paper and helping students to correct errors. The students received extensive training in listening for half an hour to one hour and completed a test paper every day. For the targeted teaching of word selection, teachers explained English grammar and meanings of English words and helped students memorize them. For the targeted teaching of reading, students were given at least three pieces of reading comprehension texts every day and given hints about the paragraph that contained the answer to a question, and students were asked to mark the sentences related to the questions. For the targeted teaching of composition, teachers asked students to write three to five essays every week, marked the work and corrected mistakes, and explained some basic sentences for students to memorize, students were also asked to recite 20–30 English words and these were checked by the teacher every day. For this research, the data comprised the English scores of 100 students from the Department of Foreign Language of Hefei Normal University after targeted teaching. The total score possible was 100 points.

(2) Data processing

The students' total English score after the targeted improvement program was divided into four parts: listening, word selection, reading, and composition based on the teacher's targeted tutoring. The scores for these four parts were recorded, cleaned, and arranged in Excel. Data cleaning process involved the removal of redundancies, logic errors, and other specific data, processing missing values, and converting data formats (Yu et al., 2018). The data we obtained may contain some invalid information due to various factors in the collection process, and any irrelevant information was deleted. The results of data preprocessing are shown in Table 1. The total score for the listening component was 30, for word selection it was 15, for reading it was 30, and for composition it was 25.

Table 1 Distribution of English scores.

Number	Listening (30 points)	Word selection (15 points)	Reading (30 points)	Composition (25 points)
1	16	10	25	15
2	18	8	18	19
3	18	3	15	17
4	18	11	25	20
5	20	10	25	18
6	20	5	17	16
7	23	6	26	17
8	16	9	20	21
9	20	10	28	16
……	……	……	……	……
100	28	4	25	15

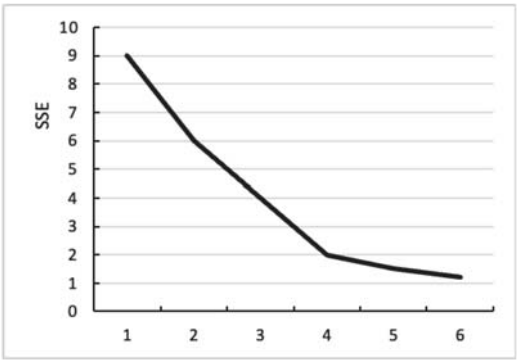

Figure 1 k value selection curves.

Table 2 SSE calculation values.

K value	SSE
1	9.02
2	6.11
3	4.09
4	1.97
5	1.52
6	1.26

3.2 Determination of *k*-Value of Cluster Number

The value of k is crucial to the final results of the experiment, but it is difficult to determine the value of k in the K-means algorithm (Mohadab et al., 2020). Small k values may result in a large difference between the data objects in the same cluster, and large k values may result in only a small difference between the various clusters. In this paper, the elbow method (Aggarwal and Sharma, 2019) was used to determine the k-value of the clustering number. When the k value was less than the actual number of clusters, the value of the sum of squares of errors (SSE) decreased greatly. When the k value reached the true number of clusters, the reduction of the SSE would weaken, and then it would show a gentle trend as the value of k continued to increase. As a curve whose SSE value changed with the k value, the k value corresponding to the elbow point in the curve was close to the optimal cluster number.

The core index to determine the value of k was the sum of squared errors (SSE). The specific formula is:

$$\mathrm{SSE} = \sum_{i=1}^{k} \sum_{x \in c_i} |d\,(x, c_i)|^2, \tag{3}$$

where k represents the number of clusters and x represents the data object in current cluster $c_i c_i$. The value range of k was set as [1,6], and the corresponding calculation results of SSE are shown in Table 1.

As seen in Table 1 and Figure 1, from $k = 1$ to $k = 4$, the SSE value decreased significantly as the division of the data became more and more refined; however, from $k = 4$ to $k = 6$, it was seen that the change of SSE values slowed down significantly, and the overall curve showed a gentle trend. There was an obvious elbow point in the curve when $k = 4$; so, according to the elbow method, it was concluded that $k = 4$ was the most appropriate number of clusters.

3.3 Clustering Analysis of English Scores

Students' English scores after the targeted improvement program were classified and analyzed by the k-means algorithm after data preprocessing. According to the elbow method, the best clustering number was 4, i.e., the best classification result of students' English scores. An analysis of the characteristics of every category and the differences between categories can help students understand their weaknesses and strengths in English and gives teachers valuable information that can inform their teaching methods so as to improve students' academic performance.

The data in Table 3 is the cluster center of every category, i.e., the average English score of students.

Table 3 Clustering analysis results of English scores.

Category	Listening	Word selection	Reading	Composition
1	19.00	8.00	19.33	22.00
2	25.60	6.00	26.60	15.20
3	18.50	10.25	25.75	17.55
4	18.67	5.33	16.67	17.83

Table 4 Evaluation and interpretation of k-means algorithm clustering.

Category	Number of members	Evaluation and interpretation
1	36	This category had the largest number of students; these students performed well in the composition task, but were average in the other tasks.
2	31	This category had a large number of students. These students were good at listening and reading but not good at composition.
3	25	These students performed well in reading, and their average score for word selection was the highest among the four categories.
4	8	The least number of students were in this category; they did not perform well in any of the four tasks.

The cluster analysis results in Tables 3 and 4 show that the students in the first category scored well in the composition task, but their performance on the other three tasks was only average. Hence, these three areas required targeted teaching. The students in the second category scored well in listening and reading, but they had the lowest average score for word selection (6.00), which was only 40% of the total score; hence, they needed targeted tutoring on word selection. The students in the third category scored well in word selection and reading, but their average score for listening was 18.5, just reaching the pass line, so they needed targeted tutoring on listening. Students in the fourth category scored poorly in all four tasks, barely achieving a pass in listening and writing. Therefore, they needed targeted tutoring on word selection and reading as priorities, and then listening and writing.

The data presented in Table 3 shows English scores of the 100 students for word selection were obviously not very good, since the highest score was 10.25 (68.33% of a total 15 points) achieved by the third category, just reaching the pass line. Therefore, these 100 students were not good at word selection. Teachers can first give all the students targeted tutoring on word selection and arrange more practice exercises. In terms of listening, except for the second category whose average score accounted for 85.33% of the total score, the listening scores of the other three categories were around the pass line, accounting for about 60% of the total score. Therefore, teachers need to strengthen targeted tutoring on listening for the students in the other three categories. Secondly, in terms of reading, although the second and third categories scored well in reading, reaching about 80% of the total score (30), the lowest average score was 16.67, only 55.57% of the total score. Therefore, teachers should give more targeted tutoring in reading to the first and fourth categories, and concentrate on different areas with the second and third categories. Finally, in terms of composition, scores of the four categories were all above the pass line, and the highest average score was 22.00 achieved by the first category, which was 88% of the total score (25 points). However, except for the first category, the rest were hovering on the pass line, so teachers can provide extension activities for the first category and concentrate on improving the writing skills of students in the other three categories. The results of the cluster analysis showed that after the tutoring targeting the four components of the English course, students' scores improved to varying degrees. Next, teachers could stream the students according to their abilities to avoid the waste of resources and time and improve students' English scores faster.

4. CONCLUSION

The paper discussed the K-means algorithm and applied it to the analysis of students' English scores. After analyzing 100 students' English scores after targeted improvement, they were divided into four categories. The highest average score of the first category was 22.00 for composition. The highest average score of the second category was 25.60 for listening and 26.60 for reading. The highest average score of the third category was 10.25 for word selection and 25.75 for reading. Students in the fourth category did not obtain satisfactory scores. The results suggested that the K-means algorithm is able to address the shortcomings of the traditional evaluation methods. Instead of evaluating only the total score, the K-means algorithm evaluated all the components affecting students' total English score and revealed students' strengths and/or weaknesses that contributed to the English test scores. This method can help teachers to understand students' acquisition of knowledge in all aspects of English objectively, provide a sound basis for teachers to design targeted strategies, and enable education to be more scientific, logical, and targeted.

REFERENCES

1. Aggarwal, D., & Sharma, D. (2019). Application of Clustering for Student Result Analysis. *International Journal of Recent Technology and Engineering (IJRTE)*, *7*(6c).

2. Chen, Z., Li, G., He, J., Yang, Z., & Wang, J. (2022). A new parallel adaptive structural reliability analysis method based on importance sampling and K-medoids clustering. *Reliability Engineering & System Safety*, *218*, 108124.1-108124.14.
3. Hung, P.D., Ngoc, N.D., & Hanh, T.D. (2019). K-means clustering using RA case study of market segmentation. *Proceedings of the 2019 5th International Conference on E-Business and Applications. New York: ACM*, 100–104.
4. Jing, J., Ke, S., Li, T., & Wang, T. (2021). Energy method of geophysical logging lithology based on K-means dynamic clustering analysis. *Environmental Technology & Innovation*, *23*(JAN.1), 101534.
5. Jones, B.G., Streeter, A.J., Baker, A., Moyeed, R., & Creanor, S. (2021). Bayesian statistics in the design and analysis of cluster randomised controlled trials and their reporting quality: a methodological systematic review. *Systematic Reviews*, *10*(1), 1–14.
6. Kang, H., Chen, Y., & Zhao, H. (2020). K-means Algorithm Description and the Application of Cluster Analysis in Heavy Truck Vehicle Fault. *2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC)*.
7. Lei, T., Jia, X., Zhang, Y., He, L., Meng, H., & Nandi, A.K. (2018). Significantly Fast and Robust Fuzzy C-Means Clustering Algorithm Based on Morphological Reconstruction and Membership Filtering. *IEEE Transactions on Fuzzy Systems*, *26*(5), 3027–3041.
8. Li, Y., Yang, Z., & Han, K. (2021). K-Means Parallel Algorithm of Big Data Clustering Based on Mapreduce PCAM Method. *Engineering Intelligent Systems*, *29*(6), 411–418.
9. Mohadab, M.E., Bouikhalene, B., & Safi, S. (2020). Automatic CV processing for scientific research using data mining algorithm. *Journal of King Saud University - Computer and Information Sciences*, *32*(5), 561–567.
10. Sardar, T.H., & Anrisa, A. (2018). An analysis of MapReduce efficiency in document clustering using parallel K-means algorithm. *Future Computing and Informatics Journal*, *3*(2), 200–209.
11. Shi, H., & Xu, M. (2018). A data classification method using genetic algorithm and K-means algorithm with optimizing initial cluster center. *2018 IEEE International Conference on Computer and Communication Engineering Technology*, 224–228.
12. Wang, W., Ma, Q., Liu, Y., Yao, N., Liu, J., Wang, Z., & Li, H. (2021). Clustering analysis method of power grid company based on K-means. *Journal of Physics: Conference Series*, *1883*(1), 012072.
13. Yu, N., Li, Z., & Yu, Z. (2018). Survey on Encoding Schemes for Genomic Data Representation and Feature Learning—From Signal Processing to Machine Learning. *Big Data Mining and Analytics*, (3), 191–210.
14. Zhou, K. (2022). Research on Online Creativity Education for College Students Using Data Mining. *Engineering Intelligent Systems*, *30*(6), 447–452.
15. Zhu, K., Zhao, Y., Ma, Y., Zhang, Q., Kang, Z., & Hu, X. (2022). Drip irrigation strategy for tomatoes grown in greenhouse on the basis of fuzzy Borda and K-means analysis method. *Agricultural Water Management*, *267*(C).