Research on the Application of Clustering Algorithm in the Analysis of College English Achievement

Xiaoling Cao

School of Foreign Languages, Liaodong University, Dandong, Liaoning, China

764859609@qq.com

Abstract. With the rapid development of educational informatization, teachers in colleges and universities are paying more attention to make use of big data technology in the field of education. It is worthwhile for teachers to take advantage of K-means clustering algorithm in College English achievement data analysis to process and analyze their students' college English performance, integrating and classifying the performance data of different types to have a more comprehensive understanding of students' achievement and progress in English learning. The experimental results show that K-means clustering algorithm are of great significance in data analysis and decision-making in college English teaching which can effectively identify different types of students' College English achievement distribution, and reveal the characteristics and trends of each college English achievement type. Through analyzing different types of scores, teachers will have a more comprehensive understanding of students' performance and progress in college English learning, and valuable reference information can be provided for them to reform their teaching models and develop more personalized teaching strategies based on the K-means clustering analysis of students' College English achievement evaluation results, which is conducive to improving the quality and the effectiveness of College English teaching.

Keywords: k-means; clustering algorithm; College English Achievement; Big data.

1 Introduction

With the dawn of the big data era, data analysis technology has found its way into various domains, revolutionizing the way we approach problems and gather insights. In the realm of education, performance metrics serve as a critical foundation for assessing the quality of teaching and the efficiency of student learning. However, traditional data analysis methods often only focus on simple statistical indicators such as average scores and standard deviations, which fails to fully leverage the latent value in students' performance data. With the continuous advancement of technology and the deepening of education informatization, data mining technology has shown strong vitality. [1] Data mining technology is widely utilized in various domains, particularly in

M. Yu et al. (eds.), *Proceedings of the 2024 5th International Conference on Big Data and Informatization Education (ICBDIE 2024)*, Advances in Intelligent Systems Research 182,
https://doi.org/10.2991/978-94-6463-417-4_36

education, as a technique for extracting valuable information and knowledge from extensive datasets. And K-means clustering algorithm is the in-depth data information analysis method in data mining technology [2] which can divide the data set into several categories, making the data within the same category closer to each other and the data between different categories further away from each other. In the assessment of students' College English performance, the application of K-means clustering algorithm to process, classify and analyze students' College English performance data which can help teachers deeply understand the distribution of students' scores, score differences, learning efficiencies, etc., and discover the useful information hidden in the data and have a good grasp of students' learning status and learning needs, which is great meaningful for them to formulate more scientific and excellent teaching plans and measures to further improve students' college English learning effectiveness. At the same time, college students can also understand their own learning status and score positioning from this, constantly adjust their learning strategies and methods. K-means algorithm is an very extremely classic clustering algorithm, which has been widely used for its simplicity, fast clustering speed, good results and ease of implementation.[3]

2 The Idea and Principles of K-means Clustering Algorithm

In data mining technology, cluster analysis is a multivariate statistical method that quantitatively classifies research objects based on their characteristics. [4] K-means clustering algorithm is one of the classic algorithms on the base of partitioning, which was first proposed by Mac in 1967. [5] It can divide data points into different clusters (or categories). So that data points within the same cluster are as close as possible to each other, while data points in different clusters are as far apart as possible. Specifically, it selects initial clustering centers and assigns objects to the corresponding clusters based on the minimum distance. Then, it recalculates the center of each cluster until the clustering stabilizes. Objects within the same cluster exhibit greater similarity, while those in distinct clusters exhibit lesser similarity. The ultimate goal of clustering is to make objects within the same group similar and objects in different groups distinguishable. [6] The greater the similarity of sample points within a group, the lower the similarity between groups, resulting in a better clustering effect. [7] And the K-means clustering technology enables a comprehensive exploration of the connections between examination outcomes and various influencing factors when applied in data analysis procedures.

The basic idea and principles of the K-means clustering algorithm are as follows: The algorithm commences by choosing K data points as the initial clustering centers. The value of K, which determines the number of clusters to be formed, is specified by the user as a parameter. And distances between each data point and the initial clustering centers are calculated, and each data point is allocated to the cluster whose center is closest. The center of each cluster is then based on the points in the cluster. And then the allocation and update steps are repeated until there are no further changes in the

clusters or the target function satisfies the specified condition. Being a dynamic clustering algorithm, K-means involves a continuous iterative optimization process and the precise steps of this algorithm are as follows:

1) Randomly selecting K initial clustering centers among data objects.

2) Calculating the mean (center object) of each cluster object, calculating the distances between each object and these center objects, reassigning the corresponding object to the cluster according to the minimum distance and forming a cluster with the allocated data points.

3) Re-calculating the mean (center object) of each cluster (if there are changes).

4) Repeating steps 2) and 3) until there is no change in each cluster. [8]

3 College English Achievement Analysis Based on K-means Clustering Algorithm

The participants in this experiment were sophomores at Liaodong University. The experimental group consisted of 126 students from 3 classes majoring in Computer Science. In the process of analyzing students' college English performance data, the K-means clustering algorithm is utilized. K-means algorithm is renowned for its speed, efficiency, high clustering accuracy, and strong interpretability, making it well-suited for meeting the demands of data analysis. Six indicators are selected as research variables during the process of the analysis of college English performance including college students' achievements in the final examination, assignment, classroom performance, basic knowledge, speaking and listening.

3.1 Data preprocessing

Data preprocessing is a crucial prerequisite for effective data mining using the K-means clustering algorithm to analyze students' college English achievement data. Data preprocessing is to discover and correct errors that can be recognized in the data file, including checking data consistency, handling outliers and missing values, etc. to avoid adverse effects on the clustering results.[9] Without proper preprocessing, the clustering algorithm may not produce accurate and meaningful results. The collected and organized data on college English performance in this research primarily comprises the college English scores of 126 students majoring in Computer Science for one semester in 2020. These scores include individual components such as student ID, final examination scores, assignment scores, class evaluation scores, vocabulary scores, speaking scores, and listening scores in students' college English achievement which are all presented in Table I. And additionally, Table II displays comprehensive details on the highest score achieved, the lowest score, and the average score for each of these components.

Table 1. basic data sheet of analysis of student's College English achievment

Student No.	Examination	Assignments	Evaluation	Vocabulary	Listening	Speaking
202001	81	85	90	94	81	85
202002	43	60	65	69	46	50
202003	65	70	75	65	72	70
202004	68	70	75	63	75	75
202005	68	75	80	73	83	85
202006	73	75	80	78	81	85
202007	76	80	85	73	86	90
202008	66	75	80	71	81	85
202009	69	85	85	80	83	85
202010	67	75	80	65	75	80
202011	95	100	95	99	100	95
202012	70	75	85	82	92	90
…	…	…	…	…	…	…

Table 2. The highest, lowest and average score of students' College English

Numbers(126)	Lowest Score	Highest Score	Everage Score
Examination	42	95	74.75
Assignment	60	100	82.22
Evaluation	65	95	83.69
Vocabulary	42	100	81.02
Listening	39	100	74.37
Speaking	50	95	80.20

3.2 Important Functions in Evaluating the Quality of Clustering Result

In clustering analysis, several important functions play a crucial role in obtaining better results. Not only similarity measurement functions, but also appropriate clustering criterion functions are essential. Cluster analysis is a classification method based on the similarity between data objects. To ensure accurate and reasonable classification, it is crucial to accurately calculate the similarity between objects. The similarity between sample objects is typically calculated using distance functions. Common distance functions used in cluster analysis include Euclidean distance, Manhattan distance, and Minkowski distance. The k-means algorithm primarily employs the Euclidean distance calculation method. The Euclidean distance between two n-dimensional vectors is calculated as shown in formula (1):

$$\mathrm{d} = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2} \tag{1}$$

Based on accurately calculating the similarity between objects, clusters to which sample objects belong are divided. To determine the cluster to which a sample object belongs, the Euclidean distance between the object and each cluster center is calculated. The minimum Euclidean distance is selected as the cluster to which it belongs, and its calculation formula is shown in formula (2). Additionally, the cluster center is determined using the formula in (3).

$$\mathrm{d}(x_i, c_m) = \sqrt{(x_i - c_m)^2} \tag{2}$$

In which, i=1, 2, …n, m=1, 2, …k:

$$c_i = \frac{1}{n_i}\sum_{x \epsilon c_i} x \tag{3}$$

In which, i=1, 2, …n, m=1, 2, …k, and ni represents the number of sample data in cluster Ci.

The clustering criterion function serves as an important basis for evaluating the quality of clustering results. In the traditional k-means algorithm, the sum of squared errors criterion is employed as the evaluation function. Assuming that there are k clusters C1, C2, ..., Ck in the data set D, and each cluster has a corresponding cluster center c1, c2, ..., ck, the sum of squared errors criterion formula is as follows:

$$\mathrm{E} = \sum_{i=1}^{k} \sum_{x \in C_i} |x - c_i|^2 \tag{4}$$

Where E is the sum of the squared errors between all data and the cluster center, and x is the data in cluster Ci. This criterion aims to minimize the sum of the squared distance between each sample point and its cluster center, thereby optimizing the clustering effect. By minimizing the sum of squared errors criterion, the k-means algorithm can divide data points into clusters with smaller internal distances, achieving effective clustering.

3.3 Clustering analysis based on K-means method

k-means algorithm program is operated and the distribution of sample numbers in each cluster is reasonable when the number of clusters is set to 5, and it is possible to form large clusters with distinct features and it is easy to distinguish. Finally, 126 records are clustered into 5 clusters, and the number of cases in each cluster is shown in Table III:

Table 3. Basic data sheet of every cluster's mean

Cluster	1	36.000
	2	4.000
	3	11.000
	4	31.000
	5	44.000
Valid		126.000
Missing		.000

Table III presented the clustering results, where 126 students were grouped into five distinct clusters based on their college English achievements at the end of the semester. Cluster 1 comprised of 36 students, representing 28.6% of the total population. Cluster 2 comprised of 4 students, accounting for 3.2% of the total population. Cluster 3 comprised of 11 students, accounting for 8.7% of the total. Cluster 4 comprised of 31 students, accounting for 24.6% of the total and cluster 5 comprised of 44 students, accounting for 34.9% of the total population. The distribution proportion of students' college English achievement in the cluster center is shown in the Figure 1 below:

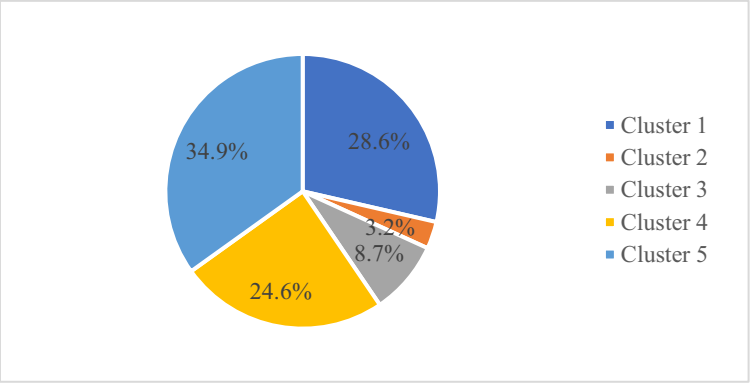

Fig. 1. The distribution proportion of students' college English acheivement

After applying the K-means clustering method and running the K-means algorithm, the final clustering centers were formed, and the students' college English achievements were classified into five levels: excellent, good, medium, pass, and fail, and the cluster mean data is displayed in Table IV.

Table 4. the cluster mean data of College English Scores

Cluster	1	2	3	4	5
Examination	77	46	62	87	70
Assignments	84	60	70	94	78
Evaluation	85	65	75	93	80
Vocabulary	86	61	66	94	73
Listening	77	43	54	88	71
Speaking	82	55	63	91	77

Based on the experimental results presented in Table IV and Table 2, Cluster 4 represents the highest-level group, comprising individuals who have achieved scores exceeding 87 points, Cluster 1 denotes the higher-level group, consisting of individuals with scores surpassing 77 points. Cluster 5 is the medium-level group with scores above 70 points, and Cluster 2 and Cluster 3 is the low-level group with scores below 75 points and 61 points in the final examination, assignments, classroom performance, basic knowledge, speaking and listening. Approximately 24.6% of the students exhibited excellent performance, while 28.6% showed good proficiency in college English. About

34.9% of the students achieved average scores, and 11.9% demonstrated poor performance in their college English learning.

According to Table III and Table IV, the values of each item in the mean of cluster 4 are the highest among the five clusters, indicating that cluster 4 is composed of excellent students and their college English final examination scores are very high, with an average score of 87 at the end of the semester. Students in cluster 4 achieved high scores above 91 in assignments, evaluation, vocabulary and listening passed. All of them have a serious learning attitude, solid English foundation and excellent performance in College English learning.

Cluster 1 is composed of 36 students and their college English achievement are better among the five clusters. Except the average score of their final examination and listening is 77, their assignments scores, class evaluation scores, vocabulary scores, speaking scores and listening scores are higher. They have better performance in college English learning. They can achieve better achievement in college English through their hard-working with the help of the teacher.

Cluster 5 presented that the majority of students recorded a medium level. There are 44 students accounting for 34.9% of the total number of students and both of them college English achievements are at the average level. These students are more scattered with lower listening and speaking abilities. However they can actively participate in the activities in college English classroom with the good attitude towards college English learning and have basically completed all learning tasks, but their scores in final examination, speaking and listening are lower.

Cluster 3 comprises 11 students. The final examination scores of the students in this group are low. Some students have poor listening abilities and have failed the listening test. Students in this group can participate in classroom interactions and complete assignments. The scores for vocabulary and speaking are generally low, but their listening scores are particularly poor. Students in this category have a weak foundation in college English. The scores of their college English final examination are generally low, and their listening skills are poor.

Cluster 2 is composed of only 4 students with poor performance in their college English studies. Students in this category, whose final examination scores, assignment scores, classroom evaluation scores, vocabulary test scores, speaking scores, and listening scores were very low across all the five clusters, with the minimum score among them being only 39 points. This group of students should focus on strengthening their usual tasks in their college English learning. College English teachers should take into account the influencing factors in their future teaching to improve these students' low achievement in their college English learning.

3.4 Implications of the experimental results

The experimental findings indicate that K-means clustering algorithm is highly effective in identifying distinct categories of students' English performance data distributions, and reveal the characteristics and trends of each category of student’s college English performance data distribution. These findings will provide valuable reference information for teachers' future teaching and the development of more personalized

teaching strategies. For students with excellent grades, teachers can encourage them to continue to maintain their enthusiasm for college English learning, and appropriately increase the difficulties and challenges of their learning, cultivating their autonomous learning ability and critical thinking; for those students with good grades, teachers can help them discover problems in college English learning, improve their learning effectiveness, and further enhance their college English achievement. For those students with moderate grades, teachers can give more guidance and attention, helping them strengthen their English learning, improve their learning methods to make their college English learning scores improved. For students with poor grades, teachers should give more attention and support to help them overcome English learning difficulties, and take targeted teaching measures to provide them with additional guidance and support, helping them enhance their learning confidence, thereby improving their college English learning achievement. At the same time, K-means clustering algorithm can help students understand their learning strengths, weaknesses, and preferences. It can also guide them in making informed decisions about their study strategies, resources, and course selection. The algorithm can enhance students' self-efficacy and motivation to learn, allowing them to adapt more quickly to changes in their learning environment, and providing valuable insights for their improvement.

4 Conclusions

In today's highly developed education informatization, college teachers are increasingly emphasizing the application of big data technology in the field of education to meet the needs of talent cultivation in the age of information. Clustering algorithm is an unsupervised learning method that can divide objects in a data set into different groups or clusters according to similarity or relevance. Clustering algorithms can effectively be applied in various fields and the k-means algorithm is very useful for the clustering analysis of students' College English achievement. The results of the clustering analysis of college English achievement based on the k-means algorithm can better show the distribution of different types of students' college English achievement and the characteristics and trends of each type, providing valuable reference information for college English teachers to better evaluate their students' performance in college English learning such as students' final examination, assignment, classroom evaluation, vocabulary, speaking and listening in college English and develop more targeted teaching plans and measures to improve students' learning effectiveness and scores. The application of k-means clustering algorithm in the analysis of college English achievement has important value in improving the quality of college English teaching and the effectiveness of students' college English learning, and contributing to the overall improvement of modern college English education quality and effectiveness. In the age when big data prevails, the deep integration of modern information technology and education has become an inevitable trend in the transformation and upgrading of higher education. With the advancement of information technology, the utilization of big data analysis techniques for personalized teaching has emerged as a pivotal topic in contemporary educational research. The application of the K-means clustering algorithm in

analyzing students' college English achievement enables us to unlock the full potential of learning process data, presenting an advanced approach to achieving personalized teaching in the era of big data.

Acknowledgment

This paper is supported by Social Science Foundation Project of Liaoning Province in 2021: Research on Thoreau's Transcendental Poetics from the Perspective of Culture, Project No.: L21BWW009.

References

1. Jain A K, Myrthy M N, Flynn P J. Data clustering: A survey[J].acm computing survey, 1999(31) :264-323.
2. Han J W, Micheline Kamber. Data Mining: Concepts and Techniques [M]. 2nd Edition. Beijing: China Machine Press, 2007.
3. Wang S, Liu C, Xing S J. Review on K-means Clustering Algorithm [J]. Journal of East China Jiaotong University, 2022, 39(5): 119-126.
4. Wang S F, Applied Statistics [M]. 2nd Edition. Beijing: Peking University Press, 2011, 232-237.
5. MAC Q J. Some methods for classification and analysis of multivariate observations[M]. Berkeley: Berkeley Symposium on Mathematical Statistics and Probability, 1967.
6. PANG N T, MICHAEL S, VIPIN K. Introduction to data mining[M]. Beijing: The People′s Posts and Telecommunications Press, 2011.
7. JI Q, SUN Y F, HU Y L, et al. Review of clustering with deep learning[J]. Journal of Beijing University of Technology, 2021, 47(8):912-924.
8. Zeng J X, Wang B B, Chen Z L. Linear Detection Algorithm of Stochastic Hough Transform based on Distance Constraints [M]. Journal of Nanchang Hangkong University (Natural Science Edition), 2011, 9.
9. GUO P, CAI C. Data Mining and Anaysis of Students' Score Based on Clustering and Association Algorithm [J]. Computer Engineering and Applications, 2019, 55 (17) :169-179.