Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

A Study on Optimizing Student Stratification Management in English Teaching Based on K-mean Clustering Algorithm

Wenjing Huang[1, ✉]

[1] *School of Foreign Languages, Hubei Engineering University, Xiaogan, Hubei, 432000, China*

ABSTRACT

The continuous development of digital informatization has opened the era of intelligent education in the field of education. Higher education has accumulated a huge amount of data, but it is not fully utilized, and in-depth mining and analysis of these data can reveal the students' learning and life status and provide powerful support for teaching management. Therefore, the research of using clustering algorithm to build a hierarchical management model for English teaching is very necessary. Clustering algorithm provides an effective way for the analysis of students' learning behavior, and for the research needs of English teaching, this paper proposes a multi-factor improved K-means clustering algorithm and compares and verifies its clustering effect. For the problem of stratified division of student groups, firstly, the clustering index system of students' book borrowing behavior and English course learning behavior constructed is used. Then, the improved K-Means clustering algorithm is used to cluster and mine the data of each student's behavior to discover the student groups under different behaviors, so as to realize the hierarchical clustering of students in hierarchical management. Finally, for English teaching, a student stratification management model is established from three aspects: student stratification, teaching goal stratification and teaching process stratification, which provides important decision support for student stratification determination in English teaching and provides a more rationalized management model for student management workers.

Keywords: k-means algorithm, clustering algorithm, learning behavior, English teaching, stratified management model

1. Introduction

The rapid development of social economy and the deepening of the reform process of China's education system provide an important opportunity and a favorable social environment for the improvement of the quality and efficiency of college English teaching. As an important element of the new curriculum reform model under quality education, the college classroom has put forward higher

✉ Corresponding author.
E-mail address: wenjingh91@163.com (W. Huang).
Received 20 March 2024; Revised 05 June 2024; Accepted 25 October 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-048

classroom teaching requirements for college English. On the basis of the innovative English teaching mode, not only should we focus on the classroom education and teaching methods, but we should also take the students as the basis, combine with the actual development of the students, and carry out hierarchical teaching [1-3]. According to the degree of self-knowledge of each student and the teacher's understanding of the students, according to the understanding of the students will be different levels of students to do a re-division, for the students divided into levels to develop a diversified effect of the teaching measures [4-6]. In the selection of teaching materials, can be based on the current state of learning and the ability of students to accept knowledge as a selection of conditions, and to the psychological factors and emotional state of students as a prerequisite, and with the students to develop a recent and most reasonable learning goals, in order to enhance the students' interest in learning and their own confidence, which can enhance the efficiency of the teaching, but also the personalized development of the students into a new level [7-11]. Layered teaching will be divided according to the actual situation of students, different levels of students to implement different educational methods, so as to save the time of students with superior performance, for the results of the lower students to lay the foundation for English learning [12-13]. Stratified teaching is a new attempt under the educational concept of the new curriculum reform, and the stratified teaching method embodies the teaching idea of modern education and student development, and further interprets the teaching concept of teaching students according to their aptitude, and gradually explores the teaching methods suitable for the students' actuality during the teaching process, so as to improve the efficiency of college English teaching [14-16].

By teaching students in the English classroom on a hierarchical basis, we can tailor-make the most suitable learning method for each student, so that the students can raise their English knowledge to a higher level in the near future learning time, and provide a good basic condition for the development of students in the future. Literature [17] describes the theoretical basis of hierarchical teaching method in college English teaching, the environmental background analysis should be done before teaching, the basic work should be implemented during teaching, and the teaching effect test should be carried out after teaching, so as to realize the tailor-made teaching. Literature [18] explored the differences in learning motivation of students at different levels under the English layered teaching method, and found that intrinsic motivation is a key factor affecting students' learning effectiveness, so it is necessary to enhance students' awareness of the layered teaching model, to develop a free and harmonious classroom environment, and to actively cultivate students' learning interests and learning motivation. Literature [19] puts forward the construction strategy of multimedia network college English teaching mode of layered three-dimensional interaction with the goal of improving the overall efficiency of college English teaching, which strengthens the reform process of English teaching and successfully delivers high-quality English talents for the development of the society in the new era. Literature [20] emphasizes the differences of people, indicating that the layered teaching mode of English classroom can make full use of such differences to cultivate different types of talents in line with the development of society and meet the requirements of quality education.

A large number of scholars have studied the student management style in the tiered teaching mode. Literature [21] argues that layered teaching in college English teaching should include lectures, assignments and post-course evaluations, and proposes the use of convolutional neural network algorithms to analyze the different learning styles of students, to provide appropriate teaching planning for students with different levels of ability, to improve the quality of teaching in English courses, and to set up a highly efficient and credible course assessment system to further satisfy the actual needs of students. Literature [22] pointed out that clustering algorithms have the ability to mine the potential information of data objects, so the K-means clustering algorithm based on division was applied to the cluster analysis of college English grades, and its clustering effect was evaluated. Literature [23] establishes student portraits based on the knowledge level reflected by students' test scores and uses clustering technology to classify students in groups, which can be categorized into

similar student clusters and complementary student clusters along with the different teaching objectives to give full play to the group influence and improve students' English learning performance. Literature [24] shows that carrying out English teaching evaluation can identify relevant teaching problems and solve them in time, so as to improve the level of teaching and learning, and the use of fuzzy K-means clustering algorithm to transform the problem of English teaching evaluation into an objective function solving problem, which can ensure the accuracy of the students' English evaluation, and provide a guarantee of high-quality teaching and learning for students at different levels.

In order to realize hierarchical management in English teaching from a practical point of view, this paper adopts a clustering algorithm to study student behavior in English teaching. Aiming at the problems that the number of clusters in the traditional K-means algorithm needs to be determined by human beings, is sensitive to the initial clustering center, is susceptible to the influence of isolated points, and is complicated to compute in the iterative process, the algorithm improvement method is investigated, and based on the optimized outlier detection algorithm and the idea of maximum-minimum distance as well as heuristics, a kind of improved K-means algorithm combining multi-point factors is proposed. After verifying the clustering effect of the improved algorithm, the acquired dataset was first preprocessed, and the behavioral description index system was constructed, and finally the improved clustering algorithm was used to cluster and analyze the students' behavioral data in terms of students' book borrowing behaviors and learning behaviors of English courses. On this basis, a hierarchical management model for students in English teaching was finally established.

2. Student behavior analysis based on k-mean clustering algorithm

2.1. Clustering method based on K-mean algorithm

Cluster analysis, as a more intuitive, unsupervised learning method for discovering the hidden laws behind massive data, needs an indicator to describe the relationship between objects, and the cluster similarity metric is such an indicator, which is based on the characteristics of the data objects to portray the similarity metric function between different trajectories. The difference between different categories of trajectory clustering methods lies in the difference in the calculation of the similarity metric. The clustering similarity metric clusters data with spatial attribute information based on the geometric spatial distances of different categories.

The key to the common K-means algorithm is clustering, which is based on the calculation of the distance and dissimilarity of each point to a prototype of the centroid of an object (also called a cluster) that is randomly generated for the first time and then iteratively generated over time.

Cluster analysis usually uses Euclidean distance to characterize the similarity between data objects; the farther the distance, the lower the similarity, and the less likely they are to be clustered in the same cluster. Conversely, the closer the distance, the higher the similarity, the easier it is to be clustered into the same cluster.

The distance formulas used in the K-means algorithm are generally as follows:

The formula is the distance between two elements in Euclidean space, which is also commonly known as the distance between two points formula, the traditional K-means algorithm are measured by Euclidean distance:

$$d(X,Y)=\sqrt{\left(x_1-y_1\right)^2+\left(x_2-y_2\right)^2+\cdots+\left(x_n-y_n\right)^2}=\sqrt{\sum_{i=1}^{n}\left(x_i-y_i\right)^2} \tag{1}$$

Manhattan distance, also known as cab distance:

$$d(X,Y)=\sum|x_i-y_i| \tag{2}$$

Minkowski distance, which is a general form of Euclidean distance, Manhattan distance, where p is a variable parameter:

$$d\left(X,Y\right)=\sqrt[p]{\left|x_1-y_1\right|^p+\left|x_2-y_2\right|^p+\cdots+\left|x_n-y_n\right|^p}=\left(\sum_{u=1}^{n}\left|x_u-y_u\right|^p\right)^{\frac{1}{p}} \tag{3}$$

The most intuitive explanation of the Chebyshev distance is that in chess, the king can move straight, sideways, or diagonally, i.e., the king can move to any of the 8 neighboring squares in one move. The minimum number of moves required for the king to move from square (x_1,x_2) to square (y_1,y_2):

$$d(X,Y)=\lim_{p\to\infty}\left(\sum_{i=1}^{n}\left|x_i-y_i\right|^p\right)^{1/p}=\max\left(\left|x_i-y_i\right|\right) \tag{4}$$

The Mahalanobis distance is used to represent the covariance distance of the data. It is based on the Euclidean distance, which takes into account the correlation between the data features two-by-two, i.e., independent of the measurement scale:

$$d_M(x,y)=\sqrt{(x-y)^T\sum^{-1}(x-y)} \tag{5}$$

The degree of dissimilarity is a mapping of two elements X and Y to the set of real numbers R. The quantization of the mapped real numbers indicates the degree of dissimilarity of the two elements. As a simple example, a cat and a dog have a lower degree of dissimilarity than a cat and a flower. To put it in mathematical terms, let's say that $X=\{x_1,x_2,x_3,..,x_n\}$ and $Y=\{y_1,y_2,y_3,..,y_n\}$, where X and Y are two element terms, each with n measurable characteristic attributes, then the dissimilarity of X and Y is defined as: $d=(X,Y)=f(X,Y)\to R$, where R is the real number field.

The core steps in the execution of the K-means algorithm are as follows:

1) Start by randomly selecting k initial points as the center of mass.

2) Assign the points in the data set to a cluster, i.e., find the closest center of mass for each discrete point and assign it to the cluster corresponding to that center of mass.

3) Move the center of mass of each cluster by the average of all points in that cluster.

4) Repeat steps 2) and 3) until the assignment of each point to the corresponding cluster remains unchanged.

The solution process of the K-means algorithm is generally given a dataset $X=\{X_1,X_2,X_3,.....,X_N\}$, where N is the number of samples, $X_i(i=1,2,...,N)$ refers to the i th tuple in the dataset, the i th data object is denoted as $X_i=\{X_{i1},X_{i2},.....,X_{im}\}$, and m is the number of features. The N tuples of X are divided into $K(K<N)$ clusters $U_j(j=1,2,...,K)$ based on similarity, which is initially given as a $N\times K$ binomial distribution matrix, and u_{ip} indicates that the i th sample is classified into the p th cluster. $Z=Z_1,Z_2,.....,Z_k$ is the k cluster center vector, and $Z_p=Z_{p1},Z_{p2},.....,Z_{pm}$ is the pth cluster center.

where the K clustering subgroups satisfy:

1) j st clustering grouping $|U_j|>0(j=1,2,\ldots,K)$.

2) $U_1\cup U_2\cup...\cup U_K=U$.

3) $U_1\cap U_2\cap...\cap U_K=\Phi$.

That is, the groupings obtained from the final clustering are independent and unrelated to each other, and their concatenated set is the original dataset at the very beginning.

The objective function of K-means algorithm can be expressed as:

$$P(U,Z)=\sum_{p=1}^{k}\sum_{i=1}^{n}u_{ip}\sum_{j=1}^{m}(x_{ij}-z_{pj})^2 \tag{6}$$

$$s.t\sum_{p=1}^{k}u_{ip}=1 \tag{7}$$

4) K-means algorithm for solving cluster center matrix.

Solve the partial derivatives to get Eq:

$$\frac{\partial P(U,Z)}{\partial z_{pj}}=-2\sum_{i=1}^{n}u_{ip}(x_{ij}-z_{pj}) \tag{8}$$

Let Eq. 0, the computational expression for the cluster center can be obtained as:

$$\sum_{i=1}^{n}u_{ip}(x_{ij}-z_{pj})=0\Rightarrow\sum_{i=1}^{n}u_{ip}x_{ij}=\sum_{i=1}^{n}u_{ip}z_{pj}\Rightarrow z_{pj}=\frac{\sum_{i=1}^{n}u_{ip}x_{ij}}{\sum_{i=1}^{n}u_{ip}} \tag{9}$$

After solving to obtain the cluster center matrix Z, the corresponding sample points are classified to the cluster centers with the highest similarity by calculating the distances between the sample points and the K cluster centers and comparing these distances horizontally to obtain the assignment matrix:

$$u_{ip}=\begin{cases}1,\sum_{j=1}^{m}(x_{ij}-z_{pj})^2\le\sum_{j=1}^{m}(x_{ij}-z_{tj})^2,\text{for } 1\le t\le k\\0,\text{otherwise}\end{cases} \tag{10}$$

2.2. Improvement of K-means Algorithm for English Teaching and Learning

2.2.1. Multifactor Improved K-means Clustering Algorithm. Considering that the acquired data of students' behavior in English teaching are continuous data, and the data dimensions and values are relatively small, as well as the multiple advantages of the K-means algorithm, this paper chooses the K-means algorithm for clustering analysis of students' behavior. Aiming at the problems of K-means algorithm, it is improved from four aspects, and an improved clustering algorithm combined with multi-point optimization is proposed to improve the clustering effect and operational efficiency.

The K-means clustering algorithm is optimized and improved from multiple factors, mainly from the two aspects of parameter determination of the algorithm and processing steps of the algorithm, specifically including the determination of the number of clusters, the allocation of sample points in the iterative process, the detection of outlying points, and the selection of initial clustering centers, and finally, the improved algorithm is described in its entirety, and the time complexity is analyzed. First, the combination of contour coefficient method and elbow rule is used for the determination of the number of clusters. Then for the allocation of sample points in the iterative process, a heuristic method is used to reduce the number of calculations of the distance between the sample points and the center of each cluster. Second, for the detection of outlier points, in order to improve the efficiency of the outlier detection algorithm, the CLOF algorithm is used, and the selected clustering algorithm is optimized using heuristics first, and then the optimized outlier detection algorithm is used for detection. Finally, the initial clustering center is determined step by step based on the candidate set of outlier points generated by outlier point detection and the maximum minimum distance idea.

Determination of the number of clusters:

The determination of the number of clusters has a great impact on the clustering results, but the determination of the number of clusters based on experience or dataset characteristics lacks accuracy. Contour coefficient method is a common method to determine the number of clusters, in short, the

clustering effectiveness measure is generally based on the intra-cluster and inter-cluster two aspects of the measure, when the intra-cluster distance to achieve the minimum and the inter-cluster distance to achieve the maximum of the clustering results is the ideal clustering effect, that is, with the minimum degree of cohesion of the intra-clusters and the maximum degree of separation of the inter-clusters. The contour coefficient method considers not only intra-cluster cohesion but also inter-cluster separation, and determines the quality of clustering by calculating the overall contour coefficient of the clusters.

For any sample point X_i, which belongs to cluster C_i. the average distance between sample point X_i and other sample points X_j in the same cluster is called the intra-cluster cohesion of sample point X_i, denoted as a_i, which is defined as in Eq:

$$a_i = \frac{1}{num(C_i)-1} \sum_{C_j=C_i, X_j \neq X_i} dist(X_i, X_j) \tag{11}$$

Select a certain cluster C that does not contain sample point X_i, calculate the average distance between sample point X_i and all sample points in that cluster for other clusters with the same treatment, the smallest average distance is called the inter-cluster separation of sample point X_i, denoted as b_i, which is defined as in Eq:

$$b_i = \min_{C \neq C_i} \{ \frac{1}{num(C_q)} \sum_{C_q=C} dist(X_i, X_q) \} \tag{12}$$

The contour coefficient s_i has a value in the range of $-1 \leq s_i \leq 1$. The closer s_i to -1, the greater the average distance between sample point X_i and the remaining sample points within the same cluster is than the nearest other cluster, and sample point X_i should not be assigned to this cluster. s_i The closer to 1, it indicates that the average distance between sample point X_i and the sample points within the same cluster is less than the nearest other cluster, sample point X_i is more different from the sample points in other clusters, and sample point X_i is suitable to be assigned to this cluster.

If the total number of sample points is n and the number of clusters is k, the average profile coefficient is denoted as S_k, which is defined as:

$$S_k = \frac{1}{n} \sum_{i=1}^{n} s_i \tag{13}$$

For different values of k, k is the optimal number of clusters for the largest value of the average profile coefficient.

In practice, in general, the higher the average profile coefficient, the better the clustering effect, but sometimes when the number of clusters take a smaller value such as $k=2$, this time the average profile coefficient of the value of the highest, and the clustering results of the SSE value is very large, when the value of k increases such as $k=4$, this time the average profile coefficient of the value is also very high, only slightly lower than $k=2$, and the clustering results of the SSE value will be a lot smaller, therefore, under this circumstance $k=4$ is the optimal number of clusters. In order to determine the optimal number of clusters more accurately, this study combines the elbow rule and the contour coefficient method, determines the number of clusters for the first time by the elbow rule, evaluates the quality of clusters by using the contour coefficient method, and then determines the final optimal number of clusters by comprehensively taking into account the values of contour coefficients and SSE values.

The main idea of maximum-minimum distance algorithm is to set the distance threshold, according to which the initial clustering center is determined. The maximum minimum distance algorithm idea is used to determine the initial clustering center as:

$$c = \max\left\{Dist(i,j) \mid i = 1,2,...,n;\ j = 1,2,...,k\right\} \tag{14}$$

$Dist(i,j)$ is calculated as:

$$Dist(i,j) = \min\left\{dist(x_i, c_j) \mid xi \in D/C, c_j \in C\right\} \tag{15}$$

2.2.2. Effectiveness analysis of improved k-mean clustering algorithm. In order to verify the effectiveness of the improved K-means clustering algorithm proposed in this paper, it is compared with five representative clustering algorithms, namely, traditional K-means, DBSCAN, GMM, HAC, and FCM, and evaluated with SC value, CHI value, and DBI value. All the algorithms are performed under the same experimental conditions, and the hardware configuration of the experiments is Inter(R)Core(TM)i9, RAM 16GB, and the running environment is Python 3.8.

The results of the comparison of SC values of different algorithms are shown in Fig. 1.The contour coefficient (SC) is a measure of how good the clustering effect is, and its value ranges from -1 to 1. The closer the SC value is to 1, the better the clustering effect is, i.e., the samples within the same cluster have high similarity, and the samples between different clusters have low similarity. From the figure, it can be clearly seen that the SC value of the improved K-means algorithm of this paper is distributed between 0.30 and 0.93, and the traditional K-means algorithm is between 0.20 and 0.83, and the SC value of the improved K-means algorithm of this paper has a significant improvement compared with the traditional K-means algorithm. This is, and DBSCAN, GMM, HAC and FCM clustering algorithms still have a significant advantage, especially when the number of clusters is 2~3 clusters, the SC value of this paper's algorithm is 0.831 and 0.925 respectively.

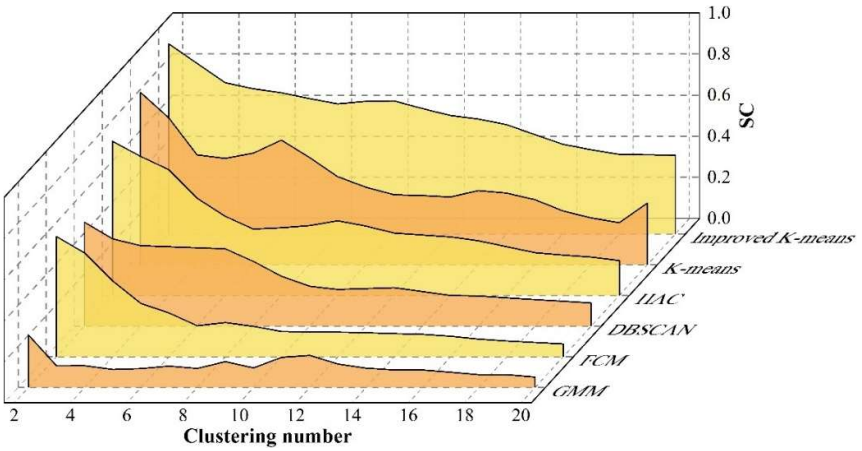

Fig. 1. The SC values of different algorithms are compared

The Kalinsky-Harabas Index (CHI) evaluates the effectiveness of clustering by comparing the intra-cluster scatter and inter-cluster scatter.The larger the CHI value, the better the clustering, i.e., the closer the data points within the clusters and the further apart the clusters.The CHI value provides a quantitative method to evaluate the performance on different student behavioral pattern distinctions.

The results of comparing the CHI values of different algorithms are shown in Fig. 2, the CHI values of the improved K-means algorithm range from 285 to 1164, and the traditional K-means algorithm ranges from 196 to 1151, the improved algorithm of this paper also improves on the CHI value and is the best among all the algorithms. It is noteworthy that DBSCAN algorithm performs better on CHI values compared to SC values, DBI values are distributed between 138~824 and evaluated better than

HAC algorithm.

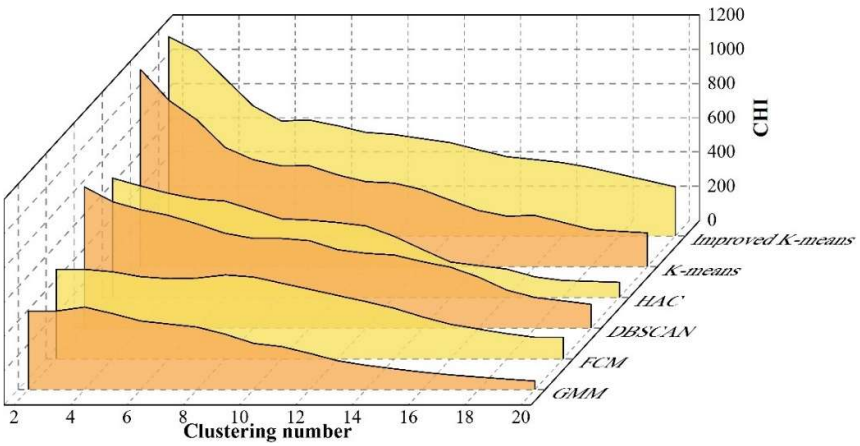

Fig. 2. The CHI values of different algorithms are compared

The Davidson Boulding Index (DBI) is used to measure the similarity within clusters versus the separation between clusters.The smaller the DBI value, the better the clustering effect, i.e., the high similarity of the samples within the same cluster and the high difference of the samples between different clusters. By analyzing the DBI value, the efficacy in distinguishing different students' behavioral patterns can be further evaluated to provide a reference for optimizing the accuracy of the clustering algorithm. The DBI comparison results of different algorithms are shown in Figure 3, from the DBI value, the improved K-means algorithm is still the optimal result among the comparison algorithms, and its DBI value is distributed between 0.8 and 2.2. The DBSCAN algorithm (1.4~2.5) also performs better in terms of DBI values.

In summary, from the comparison of SC, CHI and DBI values, it can be seen that the improved K-means algorithm in this paper performs well. Compared with the traditional algorithm can have better clustering effect, which is very helpful for the analysis of behavioral patterns.

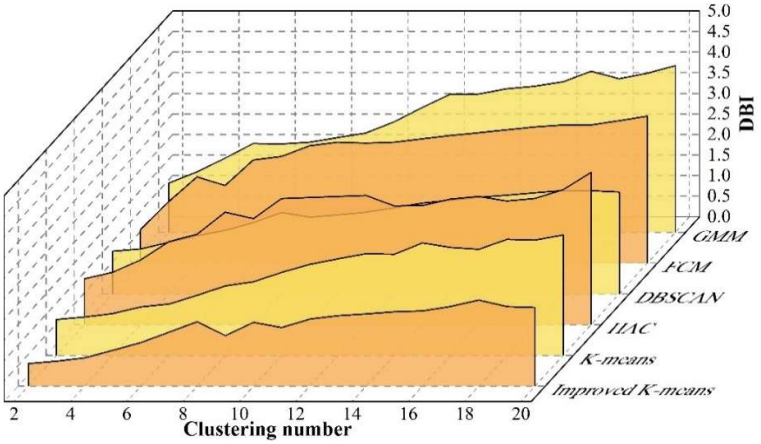

Fig. 3. The CHI values of different algorithms are compared

3. Cluster Analysis of Student Behavior Toward English Language Teaching

The data used in this experiment came from the data stored in the student data management system of University H for the academic year 2020 to 2021. The original data included the behavioral data of 43,453 students enrolled in the university during the period from October 1, 2020 to November 30, 2021, with a total of 267,979,953 entries. After these data were cleaned, integrated, transformed and

selected through data preprocessing techniques, the indicator values of each student behavior were generated in accordance with the constructed system of each student behavior indicator, and the data of student behavior indicators were clustered and analyzed separately using the improved K-Means clustering algorithm based on the improved K-Means clustering algorithm.

3.1. Construction of Learning Behavior Indicators

In the construction of the student behavior indicator system, this paper is based on a variety of behavioral data such as student consumption behavior, book borrowing behavior and study course behavior that already exist in the student behavior data warehouse.

The indicator system of book borrowing behavior is shown in Table 1, and the book borrowing behavior of students in school reflects the reading and learning situation of students in school, which is also one of the important factors considered in the construction of the digital portrait of students. The book borrowing indicators are divided into the total number of books borrowed (1~80), the total number of books borrowed (1~230), and the average number of days of single borrowing (0.1~360). These indicators can be clustered and analyzed to understand the students' book borrowing habits and the proportion of book borrowing among school students.

Table 1. The index system of book lending behavior

Index name	Coding	Range of values	Describe
Book borrowing	BA1	1~80	Students borrow the total number of books in one year
Total number of books borrowed	BA2	1~230	The total number of books borrowed by students in one year
Average number of days	BA3	0.1~360	Student's total number of days/Drawings

The indicator system of English teaching course behavior is shown in Table 2. The indicator system of students' course behavior consists of the number of attendance (0~36), theoretical test scores (0~120) and comprehensive test scores (0~100), and the cluster analysis of these indicators is used to understand the English learning habits of school students and the degree of enjoyment of English learning by different groups of students.

Table 2. The index system of English teaching behavior

Index name	Coding	Range of values	Describe
Attendance frequency	SA1	0~40	The number of hours of attendance
Theoretical achievement	SA2	0~120	Theoretical test results
Comprehensive achievement	SA3	0~100	Comprehensive results of theoretical examination and course examination

3.2. Cluster Analysis of Students' Book Borrowing Behavior

The clustering results of students' book borrowing behavior are shown in Table 3, the improved K-Means clustering algorithm in this paper divides students into 4 classes, in which the largest proportion of the number of people is class cluster 1, accounting for 46.98% of the total number of readers, and all the indicators of the cluster heart of this class are lower than the average value. The smallest percentage of the number of people is the category Cluster 4, accounting for 6.16% of the total number. A comprehensive observation of the total number of books borrowed, the total number of books borrowed, and the average number of days of single borrowing indicators of the cluster hearts reveals that the students of class Cluster 1 have a low number of books borrowed and their borrowing habits are more regular. The average number of days of single borrowing of students in class cluster 4 is as high as 144.85 days, but the total number of books borrowed and the total number of times of book borrowing are the lowest among all the class clusters. Thus, the readers of this cluster have the least amount of borrowing and the borrowing cycle is unstable. Class clusters 2 and 3 accounted for 8.91% and 37.95% of the total number of readers, respectively. It can be inferred from the comprehensive indicators of students' book borrowing behavior that students of cluster 2 borrow books frequently and their borrowing habits are more regular. Cluster 3 students have low borrowing volume and high single borrowing time.

Table 3. Student book lending behavior clustering results

Cluster type	Student ratio(%)	BA1	BA2	BA3	Total duration	Tags
1	46.98	3.42	5.92	22.35	35	The borrowing is low and the borrowing habits are more regular
2	8.91	15.78	35.20	41.03	224	The borrowing is the most, the borrowing habits are more regular
3	37.95	3.31	5.75	63.32	45	The borrowing is low and the single time is high
4	6.16	2.40	3.81	144.85	36	The minimum amount of borrowing, and
Mean	—	6.23	12.67	67.89	85	—

3.3. Behavioral Cluster Analysis of English Teaching Courses

The results of English teaching course behavior clustering are shown in Table 4, the clustering situation of English teaching behavior is divided into five types, of which, the class cluster 5 students account for the largest proportion of 34.28%, the class cluster cluster heart in the three indicators were 15.04, 92.13 and 87.21, which are greater than the average value of the overall student behavior indicators. Comprehensive observation of the indicators of the cluster heart can be found that the students in this category of clusters are among those who have more autonomy in English learning behavior and good academic performance. The percentages of students in class clusters 1 to 4 are 20.01%, 9.59%, 16.24% and 19.88%, respectively, and the number of students in class cluster 1 is relatively higher. Class Cluster 1 has the highest mean value on SA2 and SA3 indicators, which are 116.32 and 95.12, respectively. Combining the indicators of Cluster Heart, Class Cluster 1 students are very autonomous in their English learning behaviors and have excellent academic performance. Class Cluster 2 is a lazy English learning and very poor academic performance students, its SA2 and SA3 indicators mean value of 30.15 and 28.84 are not up to the passing level. Class Cluster 3 and Class Cluster 4 are students with unstable English learning behavior and average academic performance as well as more autonomous English learning behavior and good academic performance, respectively.

Table 4. Results of the English teaching course behavior clustering

Cluster type	Student ratio(%)	SA1	SA2	SA3	Tags
1	20.01	20.61	116.32	95.12	English learning behavior is very autonomous and the study is excellent
2	9.59	3.45	30.15	28.84	Study is lazy, study is poor
3	16.24	8.08	75.85	71.63	English learning behavior is unstable and the learning results are common
4	19.88	12.23	63.18	60.32	English learning behavior is less autonomous and has passed
5	34.28	15.04	92.13	87.21	Learning behavior is more autonomous and learning is good
Mean	—	11.882	75.53	68.62	—

4. The construction of student hierarchical management model for English teaching and learning

The model of student hierarchical management in English language teaching is shown in Figure 4. Hierarchical progressive teaching refers to a teaching strategy that is implemented in classroom teaching that is compatible with the learning motivation of students at all levels, and that focuses on the improvement of students at different levels. Specifically, under the premise of the classroom system, according to the characteristics of the students, such as intellectual and non-intellectual

characteristics, the existing knowledge structure of students and the degree of existing learning ability to learn mathematics, etc., and then the teacher targeted design of the teaching objectives as well as the corresponding teaching methods and means, so as to comprehensively improve the overall quality of the students in the teaching mode. According to Babanski's "optimization of the teaching process" idea, according to the actual situation of the students to improve, develop effective teaching methods, and strive to do in the limited time and energy to complete the teaching content at the same time, make the performance index of teaching greatly improved. In teaching, the hierarchical progression of teaching adopts the organizational form of combining the three forms of teaching: class, group and individual, so as to give full play to the advantages of these three forms of teaching.

The different learning possibilities of students at all levels and the different "zone of nearest development", to determine the objectives, content, methods, methods and evaluation of teaching, targeted and effective teaching. To achieve timely feedback on the teaching process, teaching and learning situation in a timely manner stratified evaluation, timely correction, adjustment and establishment of new teaching objectives.

Student stratification, according to the existing level of development of students and learning basis of some students similar to a class or layer, an administrative class of all students into a number of levels. This paper is based on the K-mean algorithm to cluster analysis of student behavior thus realizing the stratification of students.

Teaching goal stratification is to consider the students' basic knowledge, skill development and ideological education together to establish different goals, which can be set as long-term learning goals, medium-term goals, recent goals, or goals for a unit, a week, or even a lesson, and gradually improve and stratify.

The teaching process is layered, that is, the starting point of layered teaching is to better tailor the teaching to the students, after the establishment of different periods of teaching objectives, in the teaching of different levels of different objectives of teaching, thus creating our teaching methods and means to be different.

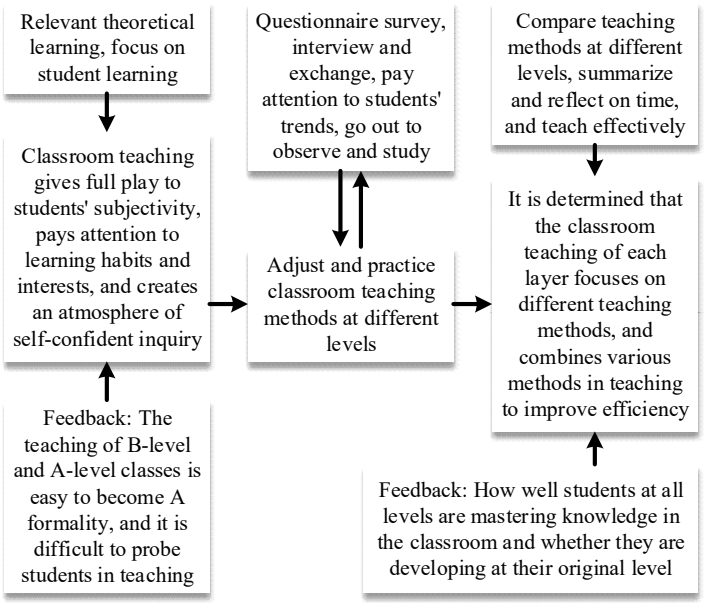

Fig. 4. The students stratified management mode of English teaching

5. Conclusion

K-means algorithm has been widely used and recognized in behavioral data analysis, in order to make full use of the massive data of English teaching and promote the development of informationization of

English teaching, this paper improves the K-means algorithm for cluster analysis of student behavioral data of English teaching. The SC values of the improved K-means algorithm are distributed between 0.30 and 0.93, the CHI values are distributed between 285 and 1164, and the DBI values are distributed between 0.8 and 2.2. Compared with the traditional K-means algorithm and other comparative algorithms the results on the three indicators are the best, having better clustering effect.

Before the clustering analysis of students' learning behavior in English teaching, the indicator system was constructed from two aspects: book borrowing behavior and course learning behavior. Then the improved K-means algorithm was utilized for cluster analysis. The research results of students' cluster analysis are as follows:

(1) The results of the clustering of students' book borrowing behavior are divided into four classes, in which the largest number of people is class cluster 1, accounting for 46.98% of the total number of readers, and the smallest number of people is class cluster 4, accounting for 6.16% of the total number of readers. The students of category cluster 1 have low book borrowing and their borrowing habits are more regular. The students of cluster 2 borrow books frequently and their borrowing habits are more regular. The students of cluster 3 have low borrowing and high single borrowing time. The average single borrowing days of students in Class Cluster 4 is as high as 144.85 days, but the total number of books borrowed and the total number of times books are borrowed are the lowest among all the class clusters, and the readers in this cluster have the least amount of books borrowed and the borrowing cycle is unstable.

(2) The clustering results of English teaching course behaviors are divided into five types, and the percentages of students in class clusters 1 to 4 are 20.01%, 9.59%, 16.24%, 19.88%, and 34.28%, respectively. Class Cluster 1 students are very autonomous in their English learning behavior and have excellent academic performance. Class Cluster 2 is lazy English learning and very poor academic performance students. Class Cluster 3 and Class Cluster 4 are students with erratic English learning behavior and average academic performance as well as more autonomous English learning behavior and good academic performance, respectively. The students in Class Cluster 5 have 15.04, 92.13 and 87.21 in the three indicators respectively, and the students in this class cluster belong to the category of students with more autonomous English learning behaviors and good academic performance.

In view of the above research, this paper establishes a student hierarchical management model for the actual situation of English teaching, which promotes the development of hierarchical progressive teaching from the three aspects of student stratification, teaching goal stratification and teaching process stratification.

References

[1] Dong, X. (2016). Research on the teaching methods of college English. Creative Education, 7(9), 1233-1236.

[2] Gu, S. (2016, April). Layered Teaching method of English in Vocational Colleges. In 6th International Conference on Electronic, Mechanical, Information and Management Society (pp. 274-278). Atlantis Press.

[3] Li, J., & Zeng, Z. (2023). The Effect of Layered Teaching Method of Reading Tasks on Middle School Students' English Learning Motivation. Journal of Education, Humanities and Social Sciences, 24, 170-176.

[4] Zheng, L. (2022). Clustering Algorithm in English Language Learning Pattern Matching under Big Data Framework. Computational Intelligence and Neuroscience, 2022(1), 1380046.

[5] Liu, S. (2022). Teaching Management System and Algorithm Implementation Based on Big Data Fuzzy K-Means Clustering. Mathematical Problems in Engineering, 2022(1), 8970414.

[6] Zhang, C. (2022, July). The Application of Hierarchical Teaching Mode Based on Hybrid Criterion Fuzzy Algorithm in Higher Vocational English Education. In EAI International Conference, BigIoT-EDU (pp. 424-430). Cham: Springer Nature Switzerland.

[7] Li, Y., & Jiang, C. (2023). Graded teaching Mode in English Teaching in Higher Vocational Colleges. Interdisciplinary Humanities and Communication Studies, 1(1).

[8] Guo, X. (2023, March). Exploration and Prospects of Layered Participatory Teaching to Teaching and Learning in Comprehensive English Courses Based on the BOPPPS Model. In 3rd International Conference on Education Studies: Experience and Innovation (ICESEI 2022) (pp. 135-142). Athena Publishing.

[9] Sui, D., & Li, F. (2017, September). Study on the Layering Teaching Model in College English. In 2nd International Conference on Judicial, Administrative and Humanitarian Problems of State Structures and Economic Subjects (JAHP 2017) (pp. 75-78). Atlantis Press.

[10] Wang, Y. (2021, August). Research on the innovation of teaching mode of the university English hierarchical listening and speaking under the "Internet+" era based on the analysis of big data. In Journal of Physics: Conference Series (Vol. 1992, No. 2, p. 022125). IOP Publishing.

[11] Huang, Y., & Wu, B. (2020). Developing critical thinking skills in stratified college English courses: Experiences of teachers in a large university in China. Creative Education, 11(07), 1042.

[12] Fei, D. E. N. G. (2020). The optimization research on college English classroom teaching under the network environment—based on the feedback report of college English stratified teaching in SASU. Sino-US English Teaching, 17(10), 297-300.

[13] Wang, N. (2017, December). Research on the Implementation of Stratified Teaching of Public English in Vocational Colleges. In 2017 World Conference on Management Science and Human Social Development (MSHSD 2017) (pp. 220-224). Atlantis Press.

[14] Chen, Y. (2022, December). Study on Implementation of Stratified Teaching in Applied Undergraduate English Teaching. In 2022 6th International Seminar on Education, Management and Social Sciences (ISEMSS 2022) (pp. 1324-1331). Atlantis Press.

[15] Qian, D. (2023). Innovative Approaches to Dynamic Stratified Teaching in Public English Education at Vocational Colleges Using SPOCs. Frontiers in Educational Research, 6(29).

[16] Guimei, L. (2021). Research on the Optimization Strategies of Higher Vocational English Stratification Teaching Mode Based on the Promotion of Students' Professional Competence and Quality. Frontiers in Educational Research, 4(5).

[17] Li, Z. (2016, May). The application of hierarchical teaching mode in College English Teaching. In 2016 International Conference on Economy, Management and Education Technology (pp. 1-5). Atlantis Press.

[18] Yang, M. (2018). Learning Motivation among University English-language Majors under Hierarchical Teaching Model. Educational Sciences: Theory & Practice, 18(6).

[19] He, B. (2020, April). Empirical Research on College English Teaching Mode of Hierarchical and Three-Dimensional Interaction. In 5th International Conference on Social Sciences and Economic Development (ICSSED 2020) (pp. 131-134). Atlantis Press.

[20] Han, X. (2019). Deepening the Reform of College English Classroom Teaching-Exploration of Hierarchical Teaching in Class——Take Qingdao Huanghai University for Example. Review of Educational Theory, 2(2), 12-16.

[21] Zhang, R., & Huang, H. (2022). Application of Convolutional Neural Network-Based Hierarchical Teaching Method in College English Teaching and Examination Reform. Mathematical Problems in Engineering, 2022(1), 3378599.

[22] Li, L., Luo, X., & Chen, H. (2015). Clustering students for group-based learning in foreign language learning. International Journal of Cognitive Informatics and Natural Intelligence (IJCINI), 9(2), 55-72.

[23] Wang, Q. (2021, June). Application of the intra cluster, characteristic of k-means clustering method in english score analysis in colleges. In Journal of Physics: Conference Series (Vol. 1941, No. 1, p. 012001). IOP Publishing.

[24] Li, H. (2022). Application of Fuzzy K-Means Clustering Algorithm in the Innovation of English Teaching Evaluation Method. Wireless Communications and Mobile Computing, 2022(1), 7711386.