International Journal of Web-Based Learning and Teaching Technologies
Volume 18 • Issue 2

The Evaluation Algorithm of English Teaching Ability Based on Big Data Fuzzy K-Means Clustering

Lili Qin, Hechi University, China
Weixuan Zhong, Hechi University, China*
Hugh Davis, University of Southampton, UK

ABSTRACT

In response to the problem of inaccurate classification of big data information in traditional English teaching ability evaluation algorithms, this paper proposes an English teaching ability estimation algorithm based on big data fuzzy K-means clustering. Firstly, the article establishes a constraint parameter index analysis model. Secondly, quantitative recursive analysis is used to evaluate the capabilities of big data information models and achieve entropy feature extraction of capability constrained feature information. Finally, by integrating big data information fusion and K-means clustering algorithm, the article achieves clustering and integration of indicator parameters for English teaching ability, prepares corresponding teaching resource allocation plans, and evaluates English teaching ability. The experimental results show that using this method to evaluate English teaching ability has good information fusion analysis ability and improves the accuracy of teaching ability evaluation and the efficiency of teaching resource application.

KEYWORDS

Big Data, Data Clustering, English Teaching, Information Fusion, Teaching Ability Evaluation

INTRODUCTION

The use of information processing technology and big data analysis technology for teaching evaluation and resource information scheduling has positive and important significance in improving the quantitative management and planning ability of teaching processes (Zhen, 2021). In recent years, with information technology as the core support and "digital" and "intelligence" as the theme of industry reform, education Big data has become China's national strategy for the first time (Miao, 2021). The Internet has created a more open, free, equal, and interconnected learning space. The original simple interaction of "one-to-one, one-to-many" in the process of teaching and learning has been transformed into a complex "many-to-many" interaction, which intensifies the teaching and

DOI: 10.4018/IJWLTT.325348 *Corresponding Author

learning process (Shang & Liang, 2022). Due to the uncertain, disordered, and multi-level nature of education, the relationship between teaching and learning presents complex system characteristics, and the original simple one-way, linear thinking mode and teaching rules are difficult to explain. Researchers applied the concepts and characteristics of complex systems to the field of education and explained the complexity of learning from two levels of collective behavioural complexity and individual behavioural complexity (Li, 2022). From five aspects—organization, system level, initial value sensitivity, nonlinearity, and emergence—the complexity of individual behaviour shows three characteristics: parallelism, conditional triggering, and adaptation and evolution (Debao et al., 2021). Using complex network analysis, machine learning, simulation, natural language processing, and other new methods to reveal new laws of teaching and learning has also become a research hotspot in the field of international education (Sreedhar et al., 2017).

However, at present, the educational concepts of higher education in China, such as learning-centred education, result-oriented education, innovation and entrepreneurship education, quality education, and quality culture, have not been fully implemented, and there are still many incompatibilities with the requirements of the national action program (Zhang, 2021). Therefore, the organic combination of higher education and information technology in the big data environment has the characteristics of real-time, continuity, dynamism, and comprehensiveness compared to the traditional organic combination of higher education and information technology (Duan, 2022).

In the ancient traditional taxonomy, the classification problem mainly comes from people's cognition of things. People mainly rely on experience and domain knowledge (Peng, 2022). The classification of things is mainly in the qualitative sense, and it is difficult to achieve the quantitative sense (Buslim et al., 2021). However, it is for the classification problems, and the ancient traditional taxonomy based only on experience and field knowledge is powerless (Ravuri & Vasundra, 2020). Mathematics is introduced into taxonomy as a tool, forming a numerical taxonomy with quantitative classification significance (Borlea et al., 2021). After that, with the further increase of the difficulty of classification problems, people began to gradually introduce the related techniques of multivariate analysis into numerical taxonomy, forming the widely used cluster analysis technology today (Liu et al., 2019).

This article studies the evaluation of English teaching ability based on big data analysis. This article proposes an English teaching ability estimation method based on big data fuzzy *k*-means clustering and information fusion, which achieves clustering and integration of English teaching ability indicator parameters, prepares corresponding teaching resource allocation plans, achieves quantitative planning of English teaching ability evaluation, and achieves accurate evaluation of English teaching ability.

MATERIALS AND METHODS

Research Review

In *Big data and PISA*, Andreas Schleicher hopes that big data will become a pilot (Wu&Wen, 2022). Only in this way can we accurately formulate policies and goals for decision-making, management, reform, and implementation. Chen Shuangye talked about big data: subjectivity, superficiality, empirical data, potential, comprehensive decision-making, and further emphasized its educational nature. Zhang Junchao and others pointed out that whether as an education management department or as universities and research institutions at all levels, it is necessary to analyse and explore various educational phenomena in order to grasp the truth of education. Zhang Yannan and others believe that the new generation of information technology with the Internet as the core has become an indispensable element in education and teaching. It is not only about new environments and new means, such as network expansion of resource channels and mixed reality optimization of situational experience. It may also be new content, such as information literacy and media literacy training, which are

included in the talent training objectives of various academic stages. Even with the advancement of perceptual intelligence towards cognitive intelligence, machines may become new cognitive agents, such as the first Chinese AI student "Hua Zhibing" created by the research and development team of the Computer Science Department at Tsinghua University. At the same time, the intervention of new elements has reshaped the relationship between teachers, students, learning resources, environment, and other original teaching elements. When learners can easily obtain a large amount of information resources through the network and when remote individuals and even machines on the network can become learning partners, the focus of teaching, the functions of the classroom, the relationship between teachers and students, and even the entire teaching structure and process are affected. Flipped classrooms, learning-based classrooms, and blended learning are typical cases. Teachers not only impart knowledge, but also design activities and resources and facilitate learning. The focus of teaching is no longer limited to low-level cognitive goals such as "memory, understanding, and application," but includes high-level goals such as "analysis, evaluation, and creation." Even with the maturity of artificial intelligence technology, simple knowledge transfer, answering questions, and evaluation tasks can be replaced by machines. At this point, the "educational" function of teachers, as well as their ability to innovate, adapt, design, guide, and organize and coordinate, becomes prominent. Teaching evaluation is highly valued by schools and teachers. Firstly, it is conducive to teachers' understanding of teaching ideas and content. Teachers can change teaching methods and content based on the evaluation results, encourage teachers to use new teaching theories, and stimulate their enthusiasm and innovation in classroom teaching. Moreover, teaching evaluation can enable teachers to recognize their achievements and mistakes in teaching work and improve their teaching abilities.

Therefore, the construction of teaching evaluation systems in colleges and universities is the trend of the current education and teaching reform and development, and a scientific evaluation system is of great significance to improve teaching quality and promote teaching management. However, most of the traditional teaching evaluation adopts questionnaire and qualitative survey methods, which are highly subjective and make the subject of the evaluation relatively limited; the evaluation content cannot reflect the teaching content over time. Therefore, it is necessary to establish a dynamic evaluation system that can change as teaching changes.

Education Big Data and Cluster Analysis

Big data has the characteristics of 5V: Volume, Velocity, Variability, Value, and Veracity. Big data is also a technology platform, through which a series of tasks such as collecting, preprocessing, and managing big data can be completed. Education big data refers to the data collected by educational administrative departments, schools, internship and training enterprises, virtual learning communities, social organizations, and other educational related institutions in all educational and learning activities (Huo, 2022). The generated data has obvious hierarchical, temporal, and situational characteristics. It is the behavioral data formed by students in various learning processes, mainly formed on online learning platforms and student-related information management systems, teaching groups, campus networks, etc. (Jamel & Akay, 2019). Narrowly defined education big data mainly refers to learning-related data from learners on online education platforms. Broadly speaking, educational big data refers to all behavioral data sourced from everyone in daily educational activities (Khan et al., 2019). The education big data in this article mainly has a narrow understanding of its meaning.

The application of cluster analysis is very extensive. In business, it can effectively help marketers and managers understand and master consumer consumption patterns and habits, thereby analyzing future sales of products and consumers (Hai et al., 2018). Cluster analysis can scientifically predict the procurement situation to truly target business activities (Gao, 2021). In medicine, cluster analysis can effectively help medical institutions and hospitals quickly understand and master the incidence rate and cure status of various patients in different periods (Zeng, 2022). In educational and teaching activities, cluster analysis can help quickly and effectively grasp children's learning situation. Due to the overlap of various clustering analysis algorithms in their respective characteristics, it is difficult

to find a clear classification plan to provide a concise classification of clustering analysis methods (Razavi et al., 2021).

The current commonly used classification methods are mainly based on the idea of clustering for partitioning.

- Partitioning based clustering algorithm. Given a dataset containing n data objects, a partitioning method is used to construct k partitions of the data. Each partition represents a class, and $k \leq n$.
- A hierarchical clustering algorithm. This method can be further divided into "bottom-up" hierarchical methods and "top-down" hierarchical methods based on different situations of the problem (Pandey & Shukla, 2023). The basic idea of layering method: When the method is divided into layers, it can be divided based on distance, density, and connectivity, or can be extended to subspaces for layering (Heil et al., 2019).
- Density based clustering algorithm. The vast majority of partitioning methods cluster based on the distance between objects, which can only discover spherical classes. However, they encounter difficulties in discovering arbitrarily shaped classes, resulting in density-based clustering methods. The main idea of density-based clustering methods is to continue clustering as long as the density (number of objects or data points) of adjacent regions exceeds a certain threshold.
- Model based clustering algorithm. There are two main methods based on models: statistical methods and neural network methods. Most clustering methods use statistical methods, which use probability parameters to help determine concepts or clusters. Each cluster obtained is usually represented by a probability description. COB-WEB is a commonly used and simple incremental conceptual clustering method. Its input object is described by symbolic quantity, and a hierarchical clustering is created in the form of classification tree. A layer in a classification tree forms a partition. Another version of COBWEB, CLASSIT, can perform incremental clustering on continuously valued attributes. Neither of these methods is suitable for clustering large databases. The neural network clustering method describes each cluster as an instance, which serves as the prototype of the cluster, and then assigns new objects to the most similar cluster based on a certain metric. The main methods include competitive learning and self-organizing feature mapping (Maylawati et al., 2020).

K-Means Clustering Algorithm

K-means clustering is a relatively mature method in cluster analysis. Due to its simplicity and efficiency, it has become the most widely used clustering algorithm among all. Given a set of data points and the required number of clusters k, k is specified by the user. The k-means algorithm repeatedly divides data into k clusters based on a certain distance function. The traditional k-means clustering algorithm uses hard classification, where each sample can only be divided into one cluster. In fuzzy k-means clustering, each sample has a certain degree of membership, indicating the degree to which it belongs to each cluster. The algorithm flow is shown in Table 1.

Since the results of one clustering may not be interpretable, this study uses a cyclic analysis process first, and then clustering analysis is performed. After the results are obtained, they are compared with the actual situation. If the cluster is too small or cannot be explained, it is deleted, cluster analysis is performed again, and the analysis is stopped until the final index of acceptable interpretation results is obtained. The data analysis process is as shown in Figure 1.

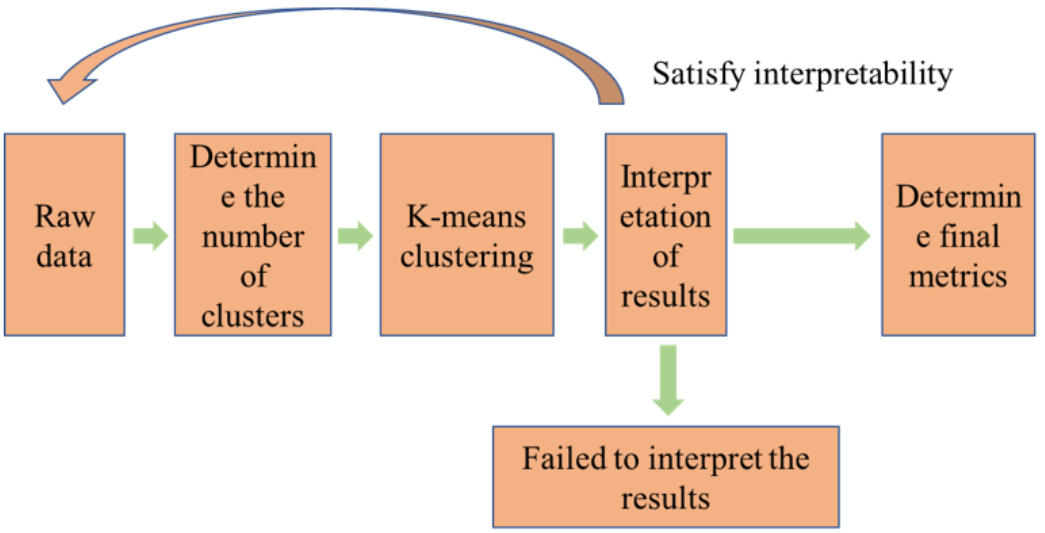
This algorithm has several advantages. It belongs to an unsupervised machine learning algorithm, so no training set is required; the algorithm only divides classes by calculating distance, the principle is simple, the algorithm complexity is low, and it is easy to implement; and the results of the division are highly interpretable.

The algorithm also has several disadvantages. The number of clusters k is an input parameter, so the choice of k is very important. If k is not known in advance and the features to be classified are

Table 1. The flow of k-means clustering algorithm

Input: Training
Procedure: Functions *k*-means (D,k,maxIter)
1:k samples are randomly selected from D as the initial "cluster centre" vector: $\mu^{(1)}$, $\mu^{(2)}$,…, $\mu^{(k)}$
2: repeat
3: order $C_i = \varnothing_{(1\le i\le k)}$
4:forj =1,2.....m
5:Calculate the Euclidean distance of sample $x^{(j)}$ from each "cluster centre" vector $\mu^{(i)}{}_{(1\le i\le k)}$
6: Determine the marker $x^{(j)}$ based on the nearest Cluster Centre vector
7: Divide the sample $x^{(j)}$ into the corresponding cluster: $C_{\lambda j} = C\lambda j \cup \left\{x^{(j)}\right\}$
8: endfor
9: fori=1, 2…, ki=1,2…, kdo
10: Calculates the new Cluster Centre vector
11:if$(\mu^{(i)})^t$=$\mu^{(i)}$then
12: Updates the current Cluster Centre vector $\mu^{(i)}$ to $(\mu^{(i)})^t$
13:clse
14: Keeps the current mean vector unchanged
15:endif
16:endfor
17:else
18: until Cluster Centre does not need to be updated
Update Output: Clustering C=C_1, C_2…, CK

Figure 1. Data analysis process



not obvious, or if the selection of *k* is improper, then the clustering result may be sub-standard. The first *k* cluster centres are randomly determined. When the amount of data to be partitioned is large, it may converge to a local minimum. So, the quality of the results obtained by randomly selecting *k* cluster centres for the first time will directly affect the efficiency of the algorithm. The algorithm can only classify numerical data. Among the current improved algorithms, the *k*-medoids algorithm has a significant impact. In this algorithm, each class is represented by a data point close to the center of the class. This makes the central point less susceptible to extreme data, enhancing the robustness of the algorithm. Combining the *k*-means method with other technologies can also greatly improve the clustering ability of the *k*-means method.

Big Data Analysis of Constraint Parameters for Evaluating English Teaching Ability

In order to achieve accurate evaluation of English teaching ability, it is first necessary to construct an information sampling model for the constraint parameters of English teaching ability. When establishing a teaching ability evaluation model, key indicators should be determined and weighted to obtain more reliable and accurate evaluation results. Key indicators can be selected through methods such as principal component analysis and factor analysis and weighted based on their importance. When selecting indicators, in addition to conventional indicators such as teaching quality and student grades, factors such as teacher passion, professional knowledge, and classroom management need to be considered. Before analyzing educational big data, data preprocessing and cleaning are necessary to improve data quality and credibility. Data mining, text analysis, and other methods can be used to preprocess and clean the original data. The specific methods include: 1) deleting duplicate records; 2) filling in missing data; 3) checking the outlier; 4) completing normalization; 5) conducting word segmentation processing; and 6) removing stop words. Combining non-linear information fusion methods and time series analysis methods, statistical analysis of English teaching abilities can be conducted. The constraint index parameters of English teaching ability are a set of nonlinear time series. A high-dimensional feature distribution space can be constructed to represent the distribution model of parameter indicators for English proficiency assessment. The main indicators that constrain English teaching ability include teacher level, investment in teaching facilities, and policy relevance level. The information flow model of constructing a differential equation to express the constraint parameters of English teaching ability is as follows:

$$x_n = x\left(t_0 + n\Delta t\right) = h\left[z\left(t_0 + n\Delta t\right)\right] + \omega_n \tag{1}$$

In the formula, $h(.)$ is the multivariate value function for evaluating English teaching ability. ω_n is the evaluation error measurement function. The solution vector of English teaching ability evaluation is calculated using a correlation fusion method in a high-dimensional feature distribution space and obtaining the feature training subset S_i (i=1, 2, ..., L) for teaching ability evaluation.

Let $x_{n+1}=\mu x_n (1-x_n)$ be the consensus solution of a statistical information model for English teaching ability evaluation, satisfying the initial value feature decomposition condition $U=\{u(t)| u(t)\in X, \| u \|\leq d, t\in I\}$, where, $(I)_{i \in N}=\{x_1, x_2, \cdots, x_m\}$. For a group of multivariate statistical characteristic distribution sequence $x(n)$ of English teaching ability assessment, the data information flow model of English teaching ability assessment based on the previous statistical measurements is as follows:

$$\begin{aligned} c_{1x}\left(\tau\right) &= E\left\{x\left(n\right)\right\} = 0 \\ c_{2x}\left(\tau\right) &= E\left\{x\left(n\right)x\left(n+\tau\right)\right\} = r\left(\tau\right) \\ c_{kx}\left(\tau_1, \tau_2, \cdots \tau_{k-1}\right) &\equiv 0, k \geq 3 \end{aligned} \tag{2}$$

When q=2, the evaluation of English teaching ability satisfies the (2+1) dimensional continuous functional condition for the level of teaching staff and the distribution of teaching resources—that is, the evaluation of English teaching ability has a convergence solution, and the constraint conditions are:

$$\Psi\left(w\right)=\ln\Phi_{x}\left(w\right)=-\frac{1}{2}w^{2}\sigma^{2} \tag{3}$$

According to the constructed data information flow model of English teaching ability assessment, a group of scalar sampling sequence components are constructed into a big data distribution model to provide an accurate data input basis for English teaching ability assessment.

Quantitative Recursive Analysis of Teaching Ability Evaluation

The quantitative recursive analysis method is used to analyze the big data information model for evaluating English teaching ability, and the control objective function for predicting and estimating English teaching ability is constructed as follows:

$$\max_{x_{a,b,d,p}}\sum_{a\in A}\sum_{b\in B}\sum_{d\in D}\sum_{p\in P}x_{a,b,d,p}V_{p} \tag{4}$$

$$s.t.\quad\sum_{a\in A}\sum_{b\in B}\sum_{d\in D}\sum_{p\in P}x_{a,b,d,p}R_{p}^{bw}\leq K_{b}^{bw}\left(S\right),b\in B \tag{5}$$

The grey model is used to make a quantitative recursive assessment of the level of English teaching ability. Assuming that the historical data of the distribution of English teaching ability is expressed as $\{x_i\}^{N}_{i=1}$, and the initial value of the disturbance characteristics is fixed, the probability density functional theory of the prediction estimation of English teaching ability is obtained as follows:

$$u_{c}\left(t\right)=Kx_{c}\left(t\right) \tag{6}$$

The quantitative recursive analysis method is used to obtain the output index distribution of English teaching ability assessment. The k nearest neighbor sample value of big data information flow is:

$$P_{1J}=\sum_{d_{j}\in KNN}Sim\left(x,d_{i}\right)y\left(d_{i},C_{j}\right) \tag{7}$$

The big data information fusion method is adopted to construct the inter domain classification objective function of English teaching ability assessment distribution big data information flow, that is, the big cluster analysis objective function is:

$$J_{m}\left(U,V\right)=\sum_{k=1}^{n}\sum_{i=1}^{n}\mu_{ik}^{m}\left(d_{ik}\right)^{2} \tag{8}$$

The exponential correlation distribution sequence $\{x_i\}^{N}_{i=1}$ of the English teaching ability evaluation studied is quantitatively analysed and combined with the k-value optimization method to obtain the quantitative recursive feature extraction results of teaching ability evaluation as follows:

$$x_n = a_0 + \sum_{i=1}^{M_{AR}} a_i x_{n-1} + \sum_{i=1}^{M_{AR}} b_i \eta_{n-1} \tag{9}$$

In the formula, a_0 is the sampling amplitude of the initial English teaching ability teaching evaluation, x_{n-i} is a scalar time series, and bj is the oscillation attenuation value of English teaching ability evaluation.

Optimization and Realization of English Teaching Ability Evaluation Model

In order to improve the quantitative evaluation of English teaching level, an English teaching ability estimation method based on big data fuzzy *k*-means clustering and information fusion is proposed. The English teaching ability assessment problem is translated into solving the *k*-means clustering objective function as a least squares estimation problem. The least squares problem is to obtain the consistent estimate of the resource constraint vector β for evaluating English teaching ability, so that $\|Y- X\beta\|$ reaches minimum. Among them, $\|\ .\ \|$ is the F-norm in the Euclidean norm, and the entropy feature extraction value of English teaching ability constraint feature information obtained is:

$$P_{loss} = 1 - \frac{1-p_0}{\rho} = \frac{p_0 + \rho - 1}{\rho} = \sum_{n=1}^{N} p_{K,n} \tag{10}$$

Given that d_i is the perturbed eigenvector of teaching ability evaluation, the estimation formula of English teaching ability is transformed into the least squares solution as:

$$z(t) = x(t) + iy(t) = a(t)e^{i\theta(t)} + n(t) \tag{11}$$

In the formula, $x(t)$ is the real part of the time series for evaluating the distribution of big data and $y(t)$ is the imaginary part of one constraint index sequence.

Using the surrogate data method to randomize the amplitude of English teaching ability, $x'(k)$ is obtained. The perturbation functional is applied to the empirical distribution data of teaching ability evaluation in class k to obtain the subset set of class k. The utilization rate of English teaching resource distribution can be expressed as:

$$U_{\text{util}} = \gamma \overline{X} \tag{12}$$

A hierarchical tree is constructed, big data analysis methods are used to establish principal component feature quantities for English teaching ability evaluation, and a fuzzy closeness filling method is used to solve the similarity of teaching resource distribution:

$$Sim_1\left(d_i, d_{1j}\right) = \frac{\sum_{k=1}^{M} W_{ik} \times_{1jk}}{\sqrt{\sum_{k=1}^{M} W_{ik}^2} \sqrt{\sum_{k=1}^{M} W_{1jk}^2}} \tag{13}$$

In the formula, d_i is the prior distribution feature vector for evaluating English teaching ability and d_{1j} is the *k*-means clustering center vector of the first layer of big data.

By combining the linear correlation feature fusion method, the clustering and integration of indicator parameters for English teaching ability evaluation are achieved, and the output teaching resource information fusion expression is obtained as follows:

$$P\left(w \mid x\right) = P\left(x \mid w\right) / P\left(x\right) \tag{14}$$

If the quantitative recursive feature $(N(i) \bmod L) <$m, the probability density feature of teaching resource distribution $p(i)= \dfrac{N\left(i\right)}{L}$, the English teaching ability evaluation big data stream $X(i)$ are divided into $p(i)$ submatrix X_{ij} with a size of Nij × m. By clustering and integrating indicator parameters, corresponding teaching resource allocation plans were formulated, thereby achieving optimization of English teaching ability evaluation.

RESULT ANALYSIS AND DISCUSSION

Discussion of Simulation Experiment

This study uses Matlab software for big data analysis of English teaching ability evaluation. The computer hardware is configured with an Intel Core i9 processor and 16GB of memory. This study collected teaching data from English teachers from multiple schools, including student exam scores, teaching evaluations, etc. Through data cleaning and pre-processing, missing values and outliers were removed and the data were normalized and standardized. Using Matlab simulation analysis to test the big data analysis performance of English teaching ability evaluation and using statistical analysis to sample data for English teaching ability evaluation, the decision threshold value for teaching ability evaluation is D_x=2 and the correlation parameter for English teaching resource distribution is set as $\xi_{c_1}^{d_2}$=3/5, $\xi_{c_2}^{d_2}$=2/5, $\xi_{c_3}^{d_2}$=2/5, max$g_{c1}(d_2)$ =6/5, max $g_{c2}(d_2)$ =3/8, max $g_{c3}(d_2)$ =1/10. The sampling frequency f_0=600 Hz. Adaptive initial step size is ρ= 0.97. The correlation coefficient of the distribution of teaching resource characteristics is B=1.14. Based on the above parameter settings, the big data reconstruction of the constraint parameters for English teaching ability evaluation was carried out, and the time-domain waveform of the big data distribution was obtained as shown in Figure 2.

Taking the big data statistical results of the index parameters of English teaching ability assessment in Figure 2 as the research object, cluster analysis and information fusion processing are carried out to achieve teaching ability assessment. Figure 3 shows the test results of evaluation accuracy and other indicators. Analysis shows that the accuracy of using this method for teaching ability assessment is high and the utilization rate of teaching resources is good. By analogy, we can also use this method in the teaching evaluation of other disciplines and deliver more outstanding talents to society. At the same time, it can also allow teachers to save time in tedious teaching evaluation of their own teaching, thereby improving the quality of their classrooms and allowing students to experience higher-quality teaching.

Experimental Results

First, after pre-processing the school statistical data, the number of clusters is determined. Due to the different methods of teachers using the evaluation platform, many of the functions of the platform are not used by teachers. Therefore, in order to ensure the results of clustering, this study compared the different results of the silhouette coefficient method on the number of clusters. The results of the first clustering analysis of all data are shown in Figure 4.

Figure 2. Time-domain waveform of big data time-domain distribution

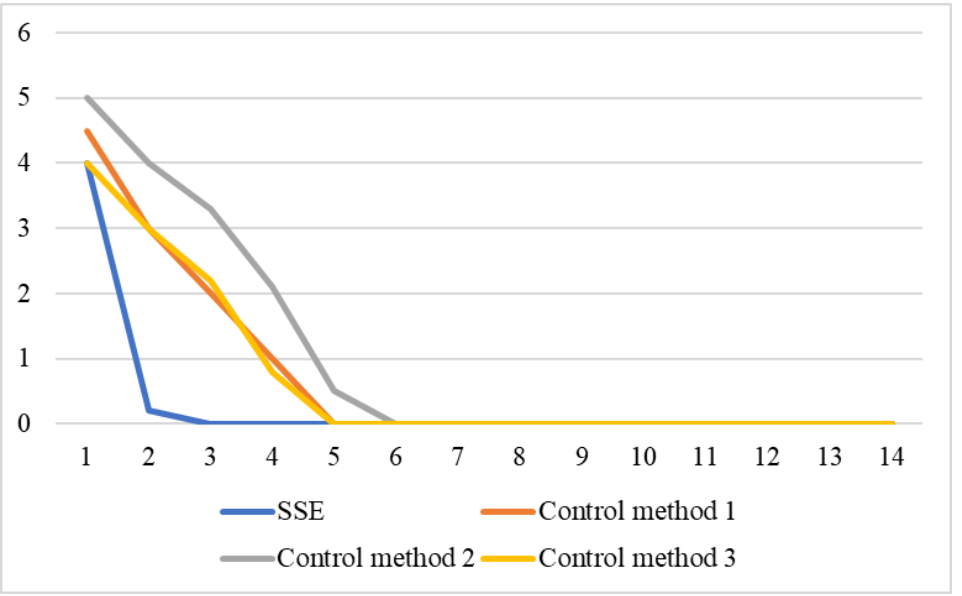
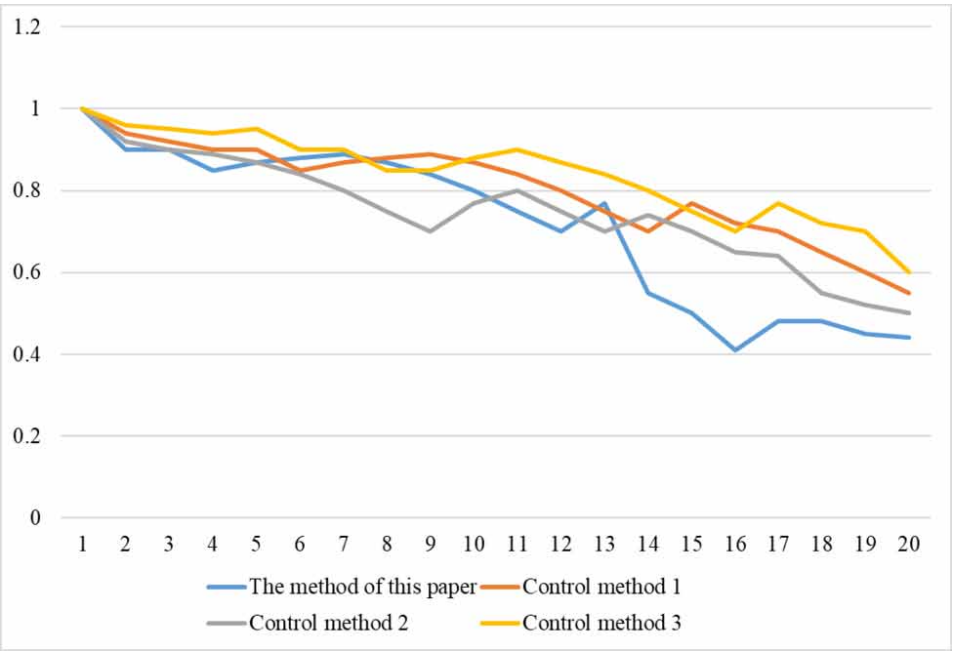


Figure 3. Performance test comparison



The digital clustering is only divided into two categories, while the result of the silhouette coefficient method is 11 categories. The silhouette coefficient can better reflect the relationship between data; therefore, 11 is used for the determination of the *k* value for the first data clustering. From Figure 4 and Figure 5, we can see that the cluster is distributed, indicating it can accurately predict the evaluation results, reflecting the true one. We can propose corresponding solutions for

Figure 4. Results of the number of clusters

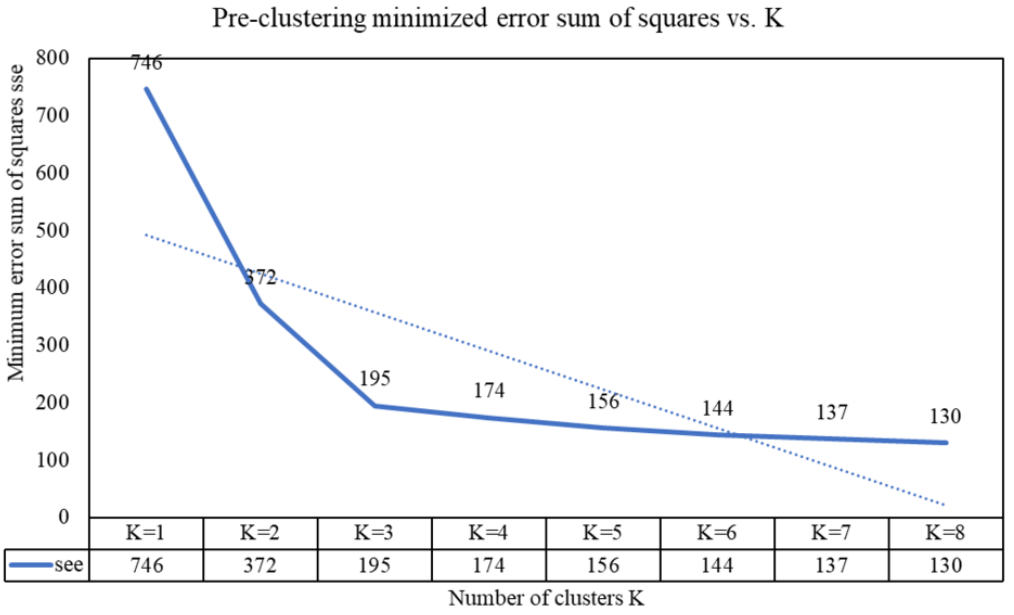
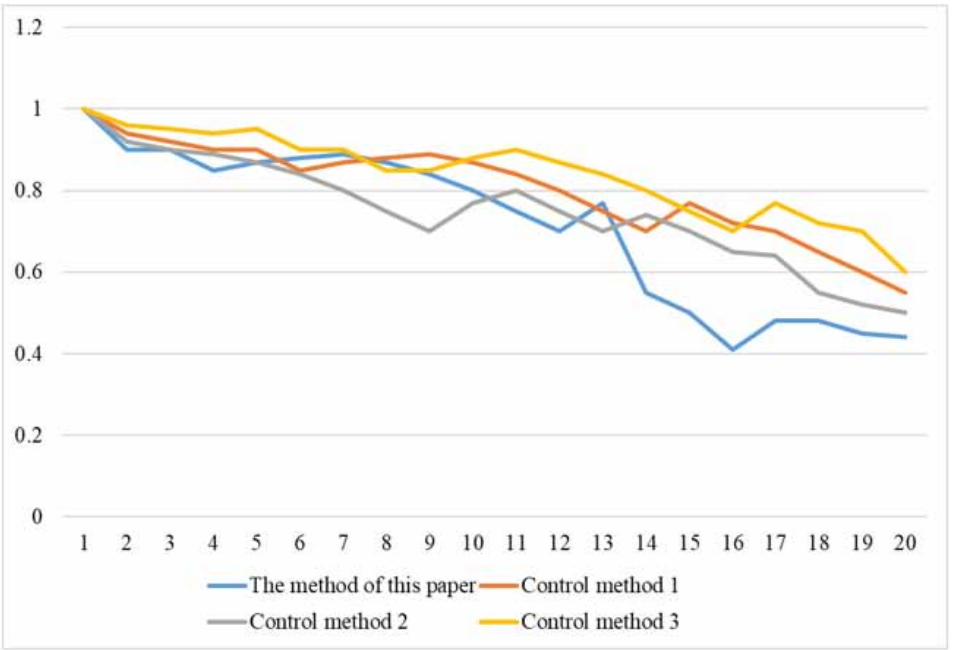


Figure 5. The number of clusters



teachers according to the ability of different teachers. At the same time, it can also provide a better teaching experience to the students.

The pre-clustering results are shown in Figure 6.

Through the fuzzy *k*-means clustering method, the students' evaluation of the teacher can be accurately obtained, which avoids the subjectivity of the results, can comprehensively evaluate the English teacher's teaching ability more objectively, and urges the teacher to improve his English classroom. The first clustering result only distinguishes some singular values and does not truly cluster the results, so singular values are placed in unexplainable results, and the remaining variables continue to be clustered, as shown in Figure 7.

Figure 6. Pre-clustering result

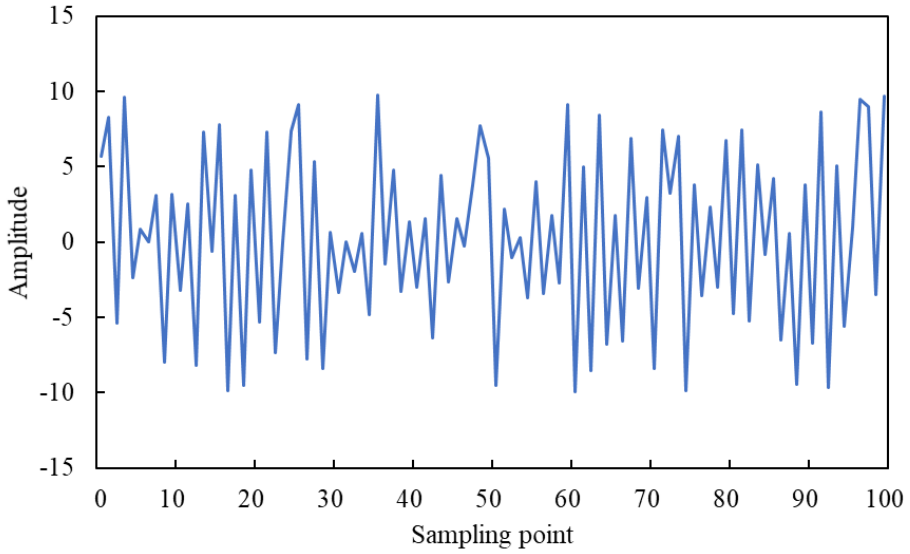
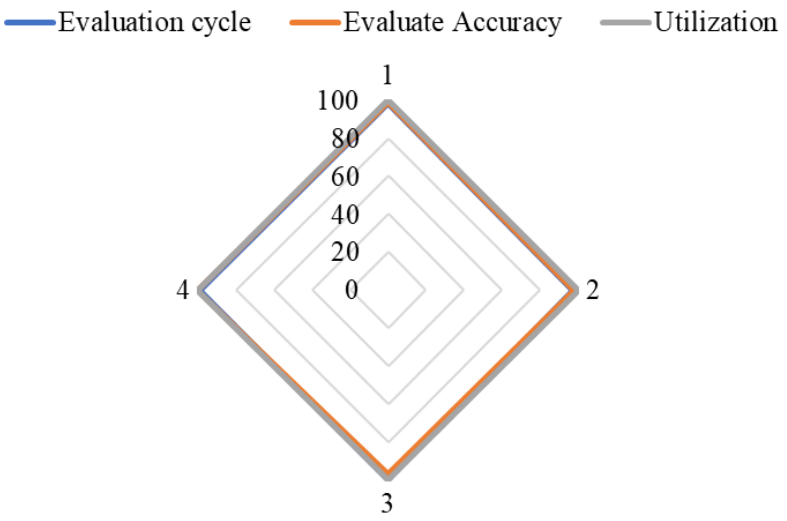


Figure 7. The first clustering result



After three cluster analyses, 80 indicators were reduced to 31, and a total of 5 dimensions were extracted as English teaching indicators. Simulation analysis actually has certain measurement errors, which can make the results less robust. Its evaluation requires multiple regressions to achieve the best results. Additionally, there may be some human factors involved in collecting data.

CONCLUSION

This article studies an optimization model for evaluating English teaching ability and proposes an English teaching ability estimation method based on big data fuzzy *k*-means clustering. The quantitative recursive analysis method is used to analyze the big data information model for evaluating English teaching ability, and the entropy feature extraction of English teaching ability constraint feature information is achieved. This study combined big data information fusion and *k*-means clustering algorithm to cluster and integrate the indicators of English teaching ability, and based on this, developed corresponding teaching resource allocation plans to achieve English teaching ability evaluation. Research has shown that the accuracy of the method used in this article for evaluating English teaching ability is good, and it improves the efficiency of utilizing English teaching resources. In the future, it is necessary to further verify the results through quantitative analysis of questionnaires and develop operational indicators to improve teachers' teaching academic ability.

ACKNOWLEDGMENT

The authors would like to show sincere thanks to those who have contributed to this research.

DATA AVAILABILITY

The figures and tables used to support the findings of this study are included in the article.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

FUNDING STATEMENT

The authors would like to extend thanks for financial support from the Inner Mongolia Education Department project "College Foreign Language Teaching and Reform Based on POA in Local Universities Under the Background of 'Golden Curriculum' – A Case Study of Chifeng University" (Grant No: NJSY22161).

REFERENCES

Alguliyev, R. M., Aliguliyev, R. M., & Sukhostat, L. V. (2021). Parallel batch k-means for Big data clustering. *Computers & Industrial Engineering*, *152*, 107023. doi:10.1016/j.cie.2020.107023

Borlea, I. D., Precup, R. E., Borlea, A. B., & Iercan, D. (2021). A unified form of fuzzy c-means and k-means algorithms and its partitional implementation. *Knowledge-Based Systems*, *214*, 106731. doi:10.1016/j.knosys.2020.106731

Buslim, N., Iswara, R. P., & Agustian, F. (2021). The modeling of "Mustahiq" data using k-means clustering algorithm and big data analysis (Case Study: LAZ). *Jurnal Teknik Informatika*, *13*(2), 213–230. doi:10.15408/jti.v13i2.19610

Debao, D., Yinxia, M., & Min, Z. (2021). Analysis of big data job requirements based on k-means text clustering in China. *PLoS One*, *16*(8), e0255419. doi:10.1371/journal.pone.0255419 PMID:34351951

Duan, L. (2022). Evaluation of English proficiency based on big data clustering algorithm. *Wireless Communications and Mobile Computing*, *2022*, 1–9. Advance online publication. doi:10.1155/2022/5718681

Gao, P., Li, J., & Liu, S. (2021). An introduction to key technology in artificial intelligence and big data driven e-learning and e-education. *Mobile Networks and Applications*, *26*(5), 2123–2126. doi:10.1007/s11036-021-01777-7

Hai, M., Zhang, Y., & Li, H. (2018). A performance comparison of big data processing platform based on parallel clustering algorithms. *Procedia Computer Science*, *139*, 127–135. doi:10.1016/j.procs.2018.10.228

Heil, J., Häring, V., Marschner, B., & Stumpe, B. (2019). Advantages of fuzzy k-means over k-means clustering in the classification of diffuse reflectance soil spectra: A case study with West African soils. *Geoderma*, *337*, 11–21. doi:10.1016/j.geoderma.2018.09.004

Huo, R. (2022). Reform and practice of college japanese test mode using big data analysis. *Mobile Information Systems*, *2022*, 1–9. Advance online publication. doi:10.1155/2022/2730477

Jamel, A. A., & Akay, B. (2019). A Survey and systematic categorization of parallel k-means and fuzzy-c-means algorithms. *Computer Systems Science and Engineering*, *34*(5), 259–281. doi:10.32604/csse.2019.34.259

Khan, I., Luo, Z., Huang, J. Z., & Shahzad, W. (2019). Variable weighting in fuzzy k-means clustering to determine the number of clusters. *IEEE Transactions on Knowledge and Data Engineering*, *32*(9), 1838–1853. doi:10.1109/TKDE.2019.2911582

Li, H. (2022). Application of fuzzy-means clustering algorithm in the innovation of English teaching evaluation method. *Wireless Communications and Mobile Computing*, *2022*, 1–9. Advance online publication. doi:10.1155/2022/7711386

Liu, B., He, S., He, D., Zhang, Y., & Guizani, M. (2019). A spark-based parallel fuzzy c-means segmentation algorithm for agricultural image big data. *IEEE Access : Practical Innovations, Open Solutions*, *7*, 42169–42180. doi:10.1109/ACCESS.2019.2907573

Maylawati, D. S. A., Priatna, T., Sugilar, H., & Ramdhani, M. A. (2020). Data science for digital culture improvement in higher education using k-means clustering and text analytics. *Iranian Journal of Electrical and Computer Engineering*, *10*(5), 4569. Advance online publication. doi:10.11591/ijece.v10i5.pp4569-4580

Miao, Y. (2021). Mobile information system of English teaching ability based on big data fuzzy k-means clustering. *Mobile Information Systems*, *2021*, 1–8. doi:10.1155/2021/9375664

Pandey, K. K., & Shukla, D. (2023). Min max kurtosis distance based improved initial centroid selection approach of k-means clustering for big data mining on gene expression data. *Evolving Systems*, *14*(2), 207–244. doi:10.1007/s12530-022-09447-z

Peng, C. (2022). An application of English reading mobile teaching model based on k-means algorithm. *Mobile Information Systems*, *2022*, 1–9. Advance online publication. doi:10.1155/2022/3153845

Ravuri, V., & Vasundra, S. (2020). Moth-flame optimization-bat optimization: Map-reduce framework for big data clustering using the moth-flame bat optimization and sparse fuzzy c-means. *Big Data*, *8*(3), 203–217. doi:10.1089/big.2019.0125 PMID:32429686

Razavi, S. M., Kahani, M., & Paydar, S. (2021). Big data fuzzy c-means algorithm based on bee colony optimization using an Apache Hbase. *Journal of Big Data*, *8*(1), 1–22. doi:10.1186/s40537-021-00450-w

Shang, J., & Liang, C. (2022). Application of clustering algorithm in English proficiency evaluation under the framework of big data. *Mathematical Problems in Engineering*, *2022*, 1–11. Advance online publication. doi:10.1155/2022/2463926

Sreedhar, C., Kasiviswanath, N., & Chenna Reddy, P. (2017). Clustering large datasets using k-means modified inter and intra clustering (KM-I2C) in Hadoop. *Journal of Big Data*, *4*(1), 27. doi:10.1186/s40537-017-0087-2

Wu, T., & Wen, M. (2022). An English teaching ability assessment method based on fuzzy mean-shift clustering. *Scientific Programming*, *2022*, 1–11. Advance online publication. doi:10.1155/2022/4219249

Zeng, G. (2022). Analysis of learning ability of ideological and political course based on BP neural network and improved-means cluster algorithm. *Journal of Sensors*, *2022*, 1–11. Advance online publication. doi:10.1155/2022/4397555

Zhang, T. (2021). Design of English learning effectiveness evaluation system based on k-means clustering algorithm. *Mobile Information Systems*, *2021*, 1–9. doi:10.1155/2021/5937742

Zhen, C. (2021). Using big data fuzzy k-means clustering and information fusion algorithm in English teaching ability evaluation. *Complexity*, *2021*, 1–9. doi:10.1155/2021/5554444

Hugh Davis is a professor in the Department of Electronics and Computer Science at the University of Southampton. He is a renowned expert in the field of computer science, with a particular focus on data management and analysis.