An English Educational ability Evaluation Model Considering Big Data Improved K-means Clustering Algorithm

Xi Li[1,2*], Wanni Mo[1,2]

[1] *College of Xiangsihu, Guangxi Minzu University, Nanning, 530225, China*

[2] *Faculty of Social Sciences and Humanities, The National University of Malaysia, Bangi, 43600, Malaysia*

Keywords*:* Big data; K-means clustering algorithm; English educational ability; Least squares; Index coefficient

Abstract

As teaching methods and content develop variously, assessing English educational ability has become a crucial component of English education and a key research focus. The traditional English educational ability evaluation model has problems such as low accuracy and efficiency. To address these issues, the study proposes an English educational ability evaluation model based on a big data-improved K-means Clustering Algorithm (KCA). The study first uses big data indicator coefficients to perform cluster analysis on educational ability data. Then, based on clustering analysis, the KCA clustering algorithm is improved using the Least Squares (LS) method, and the improved method is used to construct an English educational ability evaluation model. The results indicated that the weight values and data cleaning errors of this model in the educational ability assessment were 0.29 and 5.13 %, respectively. Concurrently, enhancements were made to the KCA method, noise reduction by 36.19% and fitting distribution error by 3.59%. The average amplitude and distribution coordination of its waveform fluctuations were 1.63 % and 91.26 %, respectively. Its test results were superior to the comparison methods. This indicates that the constructed model has better clustering performance and higher classification accuracy in English educational ability evaluation. This helps to provide more objective and scientific evaluation methods for English education. The research aims to provide more reliable data to support the development of English education.

[*] Corresponding author: Xi Li. (Email: lixi_ukm@outlook.com; ORCID: 0009-0005-1784-3148).

1. Introduction

With the development of society and the popularization of education, English education is receiving increasing attention worldwide. Given the complexity and diversity of English education, effectively utilizing big data for English Education Ability (EEA) assessment presents a significant challenge. Traditional EEA evaluation methods still have problems such as high computational complexity and slow convergence speed [1-3]. To address these issues, the study utilizes big data technology and a KCA to evaluate EEA. Big data technology can process vast volumes of data and extract latent information, which ca n comprehensively understand the level and development trends of EEA from multiple dimensions [4-5]. KCA is an unsupervised learning method based on the distance between samples. By dividing samples into K non-overlapping clusters, samples within the same cluster tend to be similar, while samples between different clusters tend to be dissimilar [6-7]. In EEA evaluation, KCA is used to divide students into different groups to understand their EEA and characteristics, providing a scientific basis for educational decision-making. The study first utilizes big data technology to process a large amount of data in English education and conducts research on educational ability data through indicator coefficient clustering analysis. Then, the LS and improved KCA are fused, and the fused method is used to construct a comprehensive and accurate EEA evaluation model. The research aims to use the constructed EEA evaluation model to provide a scientific basis for English education decision-making and teaching improvement.

The first part of this study introduces the current research status of EEA evaluation and clustering algorithms. Secondly, the study first uses big data index coefficients to perform clustering analysis on the data, and then integrates LS and KCA to construct an educational ability evaluation model. Next, the performance of the constructed EEA evaluation model is verified through simulation experiments. Finally, the experimental results are summarized, and the advantages and disadvantages of the research methods are analyzed.

2. Related works

The arrival of the big data era poses unprecedented challenges and opportunities for education. As a vital skill in global communication, the assessment of English proficiency for both students and teachers is crucial. Therefore, numerous scholars have conducted extensive research on the evaluation of educational ability. Teemu T et al. constructed a multi-standard evaluation measurement model to evaluate academic and professional abilities using reliable predictive indicators. This model evaluated the personnel involved in testing through methods such as low redundancy, high reliability, and external evaluation. These results confirmed that the evaluation measurement model could effectively assess the abilities of specific personnel [8]. Familyarska L et al. designed a method to evaluate the adaptability of teachers to challenges in educational environments using information and communication technology. This method aimed to understand the collaborative communication ability of teachers by describing their information and communication mobility. These results confirmed that this method could provide an effective evaluation of teachers' adaptability in different educational environments [9]. Kartowagiran B et al. designed a cognitive and technology-based assessment learning method to investigate students' ability to combine knowledge with reality. The research was evaluated through a combination of scientific thinking models and reliability tools. The study evaluated 289 high school students using this model. The effective evaluation capability of this model reached 0.923, and the composite reliability range was 0.88-0.90. This indicated that the model had good evaluation performance [10]. Ohashi Y et al. constructed a self-evaluation standard model to analyze the self-evaluation ability of students. The study utilized four different classification methods to analyze students' self-assessment abilities, namely problem analysis, originality of thinking, collaboration and sharing, and reflection. After unifying and standardizing these four classification methods, students in the same grade were evaluated. These results confirmed that 90 % of students could complete the evaluation content, indicating that the model had high evaluation feasibility [11].

Ke X proposed a training model for accounting professional courses based on big data analysis to transform accounting education and teaching using big data. This model completed the construction of the cultivation mode by revising the curriculum outline, strengthening textbook construction, strengthening curriculum construction, reforming curriculum assessment plans, and strengthening the construction of "curriculum ideological and political theory teaching". These results confirmed that the

model could significantly enhance the practical skills of accounting students [12]. Fischer et al. proposed a behavior research method based on big data to analyze student behavior during the teaching process. This method could discover the patterns of student behavior by collecting and classifying digital traces of student behavior, thereby providing effective assistance for teachers to tailor their teaching to their individual needs. These results confirmed that the method could effectively induce and analyze student behavior, greatly improving teaching quality [13]. Li Y et al. proposed an education data mining model based on 2D and 3D clustering evaluation to improve the quality of education. This model could calculate students' regular exam scores and provide educators with their learning status based on clustering analysis of exam scores at different periods. These results confirmed that 2D and 3D clustering evaluation could greatly enhance teachers' assessment of student learning status, thereby providing effective support for teaching decision-making [14]. Wei J et al. proposed an open interactive education algorithm based on cloud computing to enhance self-learning ability and big data scheduling ability in interactive education. This method found corresponding feature data through parallel scheduling and association rule mining, and then effectively analyzed it using cloud computing. These results confirmed that this algorithm could significantly improve the clustering accuracy and feature parsing ability of feature information [15].

Overall, the analysis of KCA based on big data improvement is of great significance in the EEA evaluation. A synthesis of numerous studies reveals that the conventional educational evaluation paradigm is constrained by three primary limitations: the necessity for extensive computation, the slowness of convergence, and the inaccuracy of information classification. Therefore, based on summarizing the existing research, this paper constructs an EEA evaluation model by using big data and KCA. Compared with the traditional model, the innovation and contribution of this model lie in three aspects. First, the singular value decomposition method is used to analyze the characteristics of EEA in big data. Subsequently, a clustering analysis of educational ability data is carried out by using big data analysis, so that the data can be changed from complex to simple and easier to be understood and processed. Second, the constraint parameters utilized in exponential coefficient analysis serve to limit the generalization capacity of the educational system. Quantitative recursive analysis is employed to furnish data substantiation for decision-making processes, whereby association rules are mined to enhance the convergence and precision of the model. Third, based on quantitative recursion, the LS method is introduced to improve the KCA, which is used for cluster analysis of big data and information fusion of educational ability. The evaluation of the EEA is transformed into a solution for the LS value of the KCA clustering objective function, thus further improving the accuracy and convergence speed of the English assessment model. This study aims to improve the assessment performance of EEA, thus providing a more objective and scientific assessment method for English education.

3. Construction of an EEA evaluation model based on big data-improved clustering algorithm

In English education, data-driven decision-making can be used to improve teaching strategies, personalized learning paths, and EEA evaluation. The study first uses big data methods for clustering analysis of indicator coefficients, and then improves KCA using LS based on clustering, and applies the improved method to the construction of the EEA evaluation model.

3.1 Educational ability data research based on cluster analysis of big data index coefficients

Effectively evaluating EEA requires clustering and analyzing the big data involved in English education. The study samples EEA data based on time series and assumes that the data used for EEA evaluation have linear Gaussian features [16-17]. At this point, the linear fitting value of EEA can be obtained, which can be represented by formula (1).

$$x_0 = \varphi_0 + \sum_{i=1}^{p} \varphi_i x_{n-i} + \sum_{j=0}^{q} \theta_j \eta_{n-j} \tag{1}$$

In formula (1), φ_i represents the inversion coefficient of the i-th data in the EEA time series. θ_j represents the average coefficient value of the j-th sliding time window in the process of EEA. p represents the time series sampling coefficient. n represents the number of data points in the time

series. q represents the number of sliding times. To obtain simplified linear fitting values, the study sets educational ability data in a 2D space. In this space, the high-dimensional phase space reconstruction method is used to reconstruct the features of the data. A matrix of the feature distribution of educational ability data can be obtained, which can be represented by formula (2).

$$Q = \begin{bmatrix} x_1 & x_{1+\tau} & \cdots & x_{1+(m-1)\tau} \\ x_2 & x_{2+\tau} & \cdots & x_{2+(m-1)\tau} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N-1} & x_{N-1+\tau} & \cdots & x_{N-1+(m-1)\tau} \end{bmatrix} \tag{2}$$

In formula (2), N represents the total amount of data collected on educational ability. τ represents the time interval during which educational ability data is sampled. m represents the numerical value of the correlation dimension of feature information. After completing the construction of linear fitting and feature distribution matrix, it is possible to analyze the characteristics of educational ability data. Fig.1 shows the process of linear fitting and feature distribution.

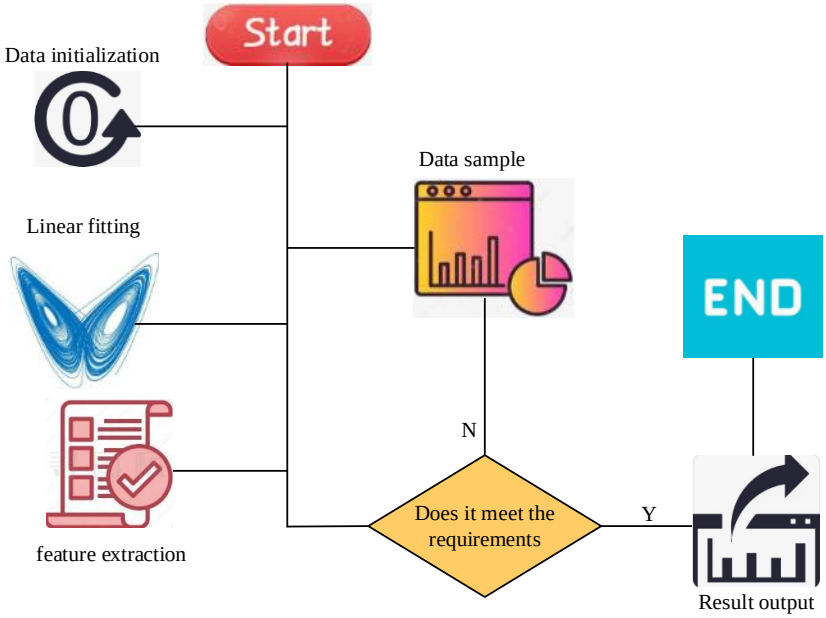


Fig.1 Flowchart of linear fitting and feature distribution.

The resulting matrix is then decomposed using the principal components of educational ability, and the singular value decomposition method will be used for feature analysis of big data EEA. At this point, a linear relationship matrix can be obtained, which can be represented by formula (3).

$$C = (c_1, c_2, \cdots, c_n) \tag{3}$$

In formula (3), c_n represents the number of elements in the linear relationship matrix. Applying singular values to big data eigenvalue analysis can obtain their attribute set descriptions. After completing the relevant attribute description, the obtained matrix can be used for orthogonality to obtain the attribute feature matrix of teaching ability. Using big data analysis to cluster educational ability data, the reconstructed educational ability matrix can be represented by formula (4).

$$X = \begin{bmatrix} a_1^T c_1 & a_2^T c_2 & \cdots & a_1^T c_m \\ a_2^T c_1 & a_2^T c_2 & \cdots & a_2^T c_m \\ \vdots & \vdots & \ddots & \vdots \\ a_N^T c_1 & a_N^T c_2 & \cdots & a_N^T c_m \end{bmatrix} \tag{4}$$

In formula (4), a_N^T represents the matrix coefficients. Through the processing of EEA big data mentioned above, the preliminary data processing of EEA has been completed. Meanwhile, the study

also investigates the data classification and processing capabilities of traditional EEA evaluation methods. The research has found that traditional data classification and processing methods suffer from low accuracy and poor classification performance [18-20]. Based on the preliminary processing of big data, the study further utilizes the constraint parameter analysis in exponential coefficient analysis. In this context, constraint parameters are introduced to regularize the model, preventing overfitting and ensuring the generalizability of the EEA evaluation, usually associated with the performance indicators of the model or system. In optimization problems, constraint parameters are used to represent constraint conditions, such as boundary constraints, resource constraints, etc. Therefore, constraint parameters are applied to EEA data analysis [21-22]. To apply constraint parameters to the evaluation of applied educational ability, it is necessary to first obtain indicator parameters in the time series, which mainly include English teacher ability, English teaching facilities, and attention to education. Fig.2 shows the evaluation process of educational ability using constraint parameters.

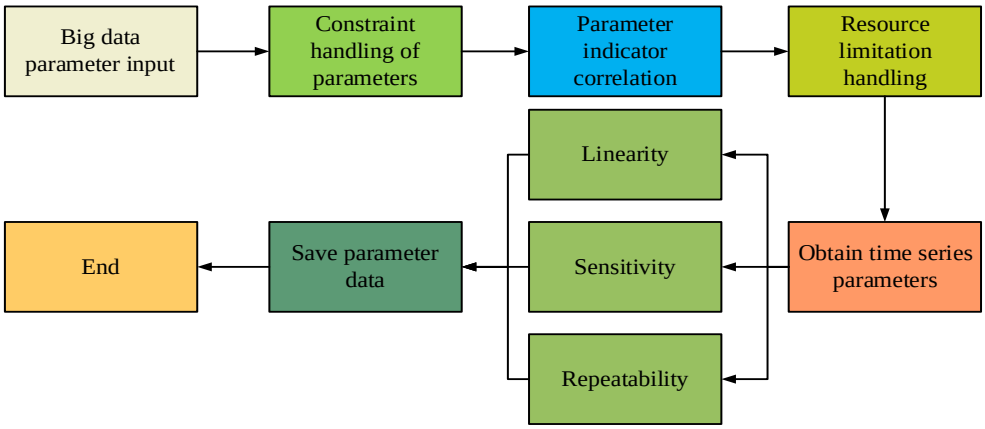


Fig.2 Flow chart of evaluating educational ability with constraint parameters.

To describe the parameter indicators more directly, mathematical differentiation is used, and the constraint parameters can be represented by formula (5).

$$x_n = h(z(t_0 + n\Delta t)) + \omega_n \tag{5}$$

In formula (5), $h(\cdot)$ represents the evaluation of the value of a multivariate function. ω_n represents the measurement function value in the process of EEA. In the high-dimensional feature distribution space, it is necessary to obtain a solution vector for capability evaluation to complete correlation fusion. There are multiple variables in the process of EEA evaluation, which can affect the accuracy of the evaluation process. Therefore, based on the preliminary processing, this study constrains the generalization ability of educational ability to ensure that the educational ability evaluation model has good convergence performance [23-24]. The constraint condition at this point can be represented by formula (6).

$$\psi_x(\omega) = \ln \Phi_x(\omega) = -\frac{1}{2}\omega^2\sigma^2 \tag{6}$$

In formula (6), σ^2 represents variance. ω represents the convergence coefficient. Φ_x represents the generalization value of educational ability in the constraint. After completing the generalization process, through a large amount of data preprocessing analysis, there are still certain shortcomings in the mining of association rules in the processing of big data indicator coefficients. Quantitative recursive analysis is a method used to analyze time series data. It analyzes the recursive graph of time series data qualitatively and quantitatively to reveal the complex dynamic characteristics in the data. Therefore, quantitative recursive analysis is introduced, which can provide data support for decision-making through mining association rules. This can be used to construct an objective function for educational ability evaluation, represented by formula (7).

$$F(a) = (SP,\ AB,\ LB,\ IH,\ DP) \tag{7}$$

In formula (7), SP represents the quality of teaching by the teacher. AB represents the student's grades. LB represents learning behavior. IH represents participation. DP represents the frequency

of interaction. After obtaining the objective function, the probability density functional of EEA prediction estimation can be expressed using formula (8).

$$u_c(t) = Kx_c(t) \tag{8}$$

In formula (8), AA $u_c(t)$ is the probability density functional. K represents the number of clusters. $x_c(t)$ represents the element value corresponding to the x-th element in the data at time t. In a high-dimensional feature space, by integrating big data information, exponential relationships in educational capability big data information can be found. According to the exponential relationship, quantification can be carried out to find the extracted values of quantitative recursive features in the process of EEA, which can be represented by formula (9).

$$x_a = a_0 + \sum_{i=1}^{K} a_i x_{n-i} + \sum_{j=0}^{K} b_{n-j} \tag{9}$$

In formula (9), a represents the sampling amplitude during the initial EEA evaluation process. x_{n-i} represents the time series value of a scalar during the evaluation process. b_{n-j} represents the attenuation value in the EEA evaluation. By utilizing big data for effective analysis of EEA data information, reliable input data can be provided for the evaluation of educational capabilities. To visualize EEA, the study subdivides it into evaluation target, indicator, and element layers. Table 1 shows the system structure of EEA.

Table 1 The system structure of EEA

Target layer	Indicator layer	Element layer
Educational ability	Teachers	Teaching level
		Teaching attitude
	Classroom teaching	Teaching level
		Teaching content
		Teaching method
		Teaching attitude
	Practice	Teaching content
		Teaching equipment
	Teaching management	Management level
		Managing

3.2 Design of EEA evaluation model based on LS and improved KCA

Based on the cluster analysis mentioned above, the study has identified the key elements of EEA assessment. This section will delve into the design of an EEA assessment model that integrates these elements. Through clustering analysis of EEA data, the elements that can be used for EEA evaluation can be divided into teachers, students, and teaching resources. These three elements are essential in EEA evaluation, and they also contain numerous factor factors, which can play a decisive role in EEA [25-26]. To more intuitively reflect the significance of the three elements in EEA, an analysis model integrating the three elements is designed in Fig.3.

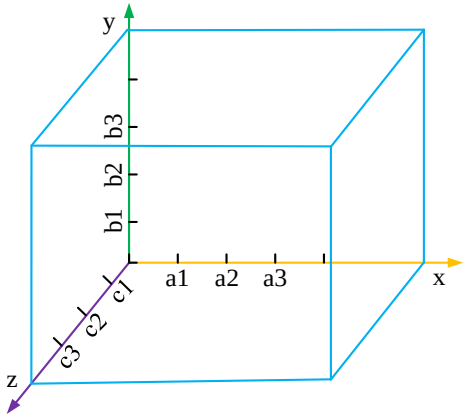


Fig.3 Analysis model diagram of educational ability integrating three elements.

Based on the analysis in Fig.3, the study can provide a deeper understanding of the evaluation of educational abilities. In the coordinate system, the x-axis is teaching resource elements, the y-axis represents student elements, and the z-axis means teacher elements. Each point on these three coordinate axes may correspond to a factor, which together form a cube. The volume of this cube represents the overall strength of educational ability. A large cube area means strong educational ability. On the contrary, if the area is small, the educational ability is relatively weak [27-28]. They study further introduces the LS to improve KCA, aiming to optimize the determination of cluster centers. KCA is a commonly used unsupervised learning algorithm whose principle is to group data points according to feature similarity to form different clusters. The distance between data points within each cluster and the mean within the cluster is minimized, while the distance between different clusters is maximized. The algorithm exhibits high computational efficiency, rendering it suitable for processing large amounts of data. It is therefore widely used in data mining and other fields. However, KCA is difficult to determine the number of central values. LS method is a mathematical optimization technique, which can find the best function matching of data by minimizing the sum of squares of errors. It can make up for the defect of KCA. Therefore, the evaluation of EEA is transformed into the solution of the LS value of the KCA objective function, which minimizes the squared error between the predicted value and the actual value to find the optimal solution. Fig.4 shows the system diagram of EEA.

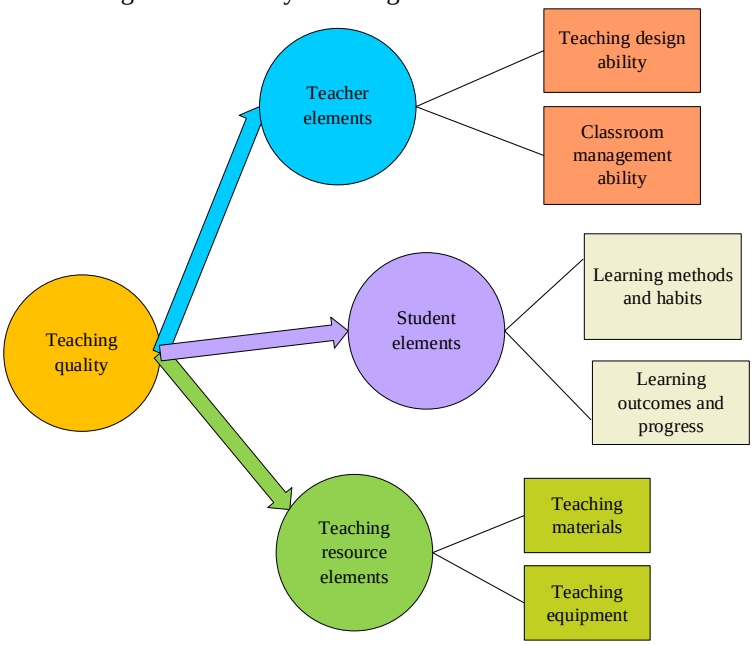


Fig.4 System diagram of EEA.

For EEA evaluation, by replacing it with the LS value of the KCA objective function, the evaluation results can be quantified into specific numerical values for better comparison and evaluation of EEA. When using LS for evaluation quantification, it is necessary to find the relationship between evaluation elements and objectives. Considering the complexity and operability of relationships, multiple linear variables are used in this study to balance their connections. In the evaluation of educational ability, both the evaluation elements and objectives are subjectively influenced by teachers and students, which can increase evaluation errors [29-30]. To reduce this part of the error, LS is applied to cluster data processing, and multiple linear regression equations are found in the processing using LS. Through this equation, the sum of squares with the minimized error is obtained, thus enabling the identification of the optimal model for the data, which in turn allows for more precise prediction and analysis, which can be represented by formula (10).

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} + \varepsilon_i \tag{10}$$

In formula (10), y_i represents the target value of the i-th measurement data in the data. x_{ki} represents the evaluation element value. b_k represents the connection between goals and elements. ε_i is the residual value of the i-th measurement data. In the regression equation, the evaluation of factors is used to determine the evaluation objectives. Fig.5 shows the relationship between the evaluation factors and the objectives.

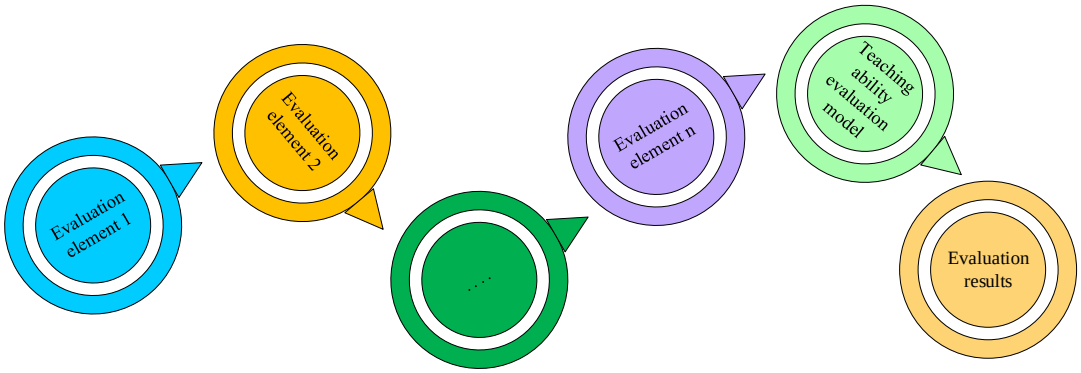


Fig.5 The relationship between elements and goals in educational ability evaluation.

In Fig.5, the EEA evaluation consists of three parts: evaluation elements, results, and model. The input evaluation elements are measurable quantities, and the evaluation model can associate the input and output to establish effective evaluation performance. At this point, the evaluation of EEA can be converted into LS solutions that can be calculated using formula (11).

$$z(t) = x(t) + iy(t) \tag{11}$$

In formula (11), $x(t)$ is the real part value of the time series in educational ability big data. $y(t)$ represents the imaginary part value of the time series. By transforming the evaluation of educational ability and utilizing the random processing ability of LS, corresponding k-class subsets can be obtained. Based on this, the distribution utilization rate of EEA can be solved, the distribution utilization rate is English teaching resources, and the utilization rate of teaching resources reflects the EEA, which can be represented by formula (12).

$$U_{util} = \gamma \bar{X} \tag{12}$$

In formula (12), γ represents the utilization coefficient. $\bar{X}$ represents the resource value. The fusion of LS and KCA can establish a principal component feature quantity for EEA evaluation, which can be used to solve the similarity of educational resources . By using similarity, each point of the same data cluster can be more closely connected, the boundary between different data clusters can be clearer, and the clustering quality can be improved, as represented by formula (13).

$$Sim_1(d_i, d_{1j}) = \frac{\sum_{k=1}^{M} W_i k \times W_{1j} k}{\sqrt{\sum_{k=1}^{M} W_i^2 k} \cdot \sqrt{\sum_{k=1}^{M} W_{1jk}^2 k}} \tag{13}$$

In formula (13), d_i represents the vector value of the prior feature. d_{1j} is the cluster center vector value corresponding to the first layer of educational ability big data. M means cluster value. W represents the distribution of resources. By combining the fusion values of the linear features mentioned above, clustering and integration of EEA indicator parameters can be completed. At this point, the fused information can be obtained, which can be represented by formula (14).

$$P(\omega / x) = P(\omega x) / P(x) \tag{14}$$

In formula (14), $P(\omega x)$ is the total amount of information. $P(x)$ is the pre-fusion information. By integrating information from indicator parameters, educational capacity can be more comprehensively evaluated and effective allocation of resources can be achieved. This method can improve the accuracy and reliability of evaluation. The combination of multiple indicator parameters can provide a more comprehensive and accurate representation of an individual's educational ability. In EEA evaluation, a model is established by calculating the parameter set of LS. The method sets the residual squared as the minimum value and solves the parameter set using the reconstructed matrix after clustering to obtain a more accurate and reliable evaluation model, which better guides educational practice. The parameter set at this point can be represented by formula (15).

$$B = (X'X)^{-1} X'Y \tag{15}$$

In formula (15), X' represents the parameter values of the reconstruction matrix. Y is the parameter value of the vertical axis of the matrix after clustering reconstruction. After completing the parameter set technique, an educational ability evaluation model can be obtained, which can be represented by formula (16).

$$J = (b_0, b_1, b_2, \cdots, b_k) \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} \tag{16}$$

Based on the above research, Fig.6 shows an educational ability evaluation model based on LS and improved KCA.

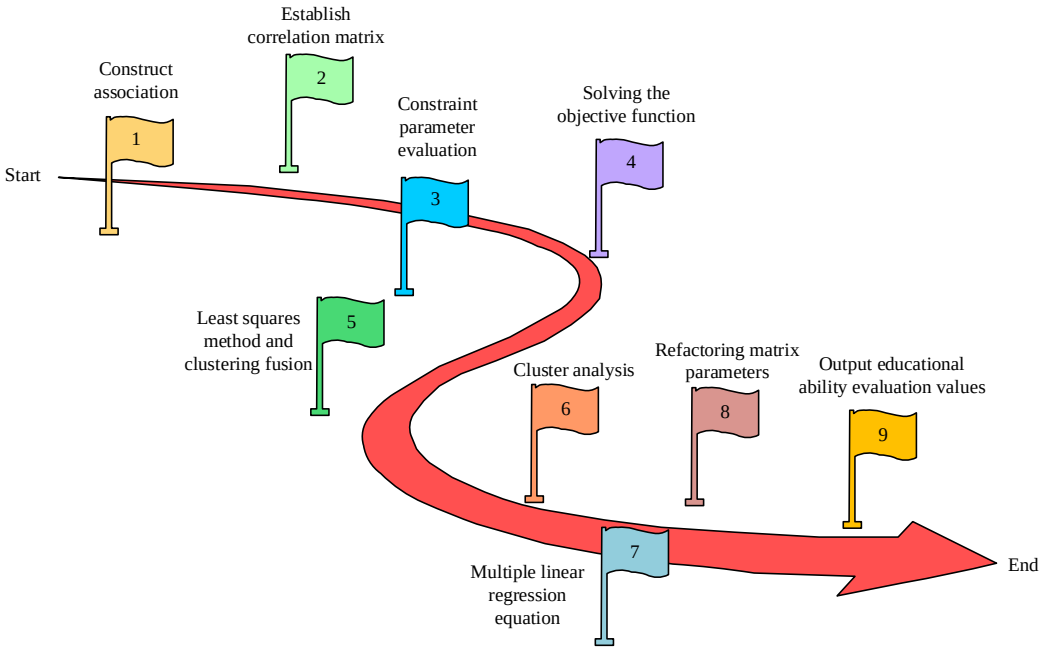


Fig.6 Educational ability evaluation model process based on LS method and improved KCA.

4. Performance analysis of EEA evaluation model based on improved mean clustering algorithm

To verify the performance of the EEA evaluation model based on improved KCA, the study compared KCA and Novel Global KCA Clustering Algorithm (NGKCA) with improved KCA. The specific performance of the EEA evaluation model was validated by comparing three methods.

4.1 Data processing performance analysis of EEA evaluation model

To verify the data processing ability of the constructed EEA evaluation model in educational ability big data, simulation experiments are conducted using Matlab to evaluate the analytical performance of big data in EEA. Simulation experiments were conducted using Matlab. The dataset, comprising student scores, was imported and analyzed. The database consists of student tables, teacher tables, class schedules, and grade tables. The fields in the student table include student number, name, gender, date of birth, and class. The fields in the teacher table include teacher number, name, date of birth, and department. The score table includes student number, course number, and grade. The curriculum table includes course number, course name, and teacher number. According to the content of the study, the EEA of students is evaluated by using the data set of students' English achievement. The study sets the threshold for EEA to 1.5, the correlation parameter to 0.6, a frequency of 500Hz, and an adaptive step size of 0.95. Then, KCA, NGKCA, and improved KCA are compared. Using the weight of student grades as an indicator, three methods are compared and analyzed for their processing weights in EEA evaluation of student performance data. Among them, NGKCA is a new clustering algorithm. Firstly, Normalized Cuts and Spectral Clustering (NJW) algorithm is used for column reduction and normalization of large data sets. Then the temporary cluster center set is obtained by using the particle swarm optimization algorithm to reduce the dimension of the data set again. Finally, KCA is used to cluster the optimal center set to get the cluster center point, and particle swarm optimization algorithm is used to adjust the cluster center point to get the cluster division. Fig.7 shows the comparison results of weight values for three methods.

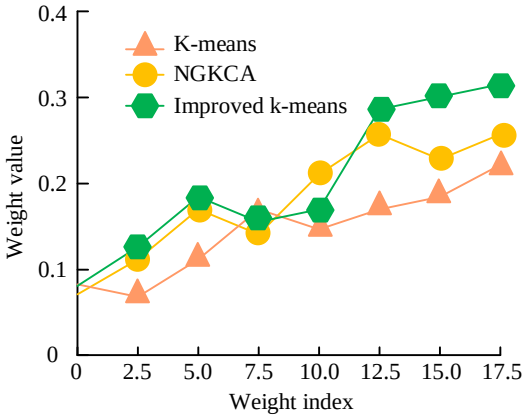
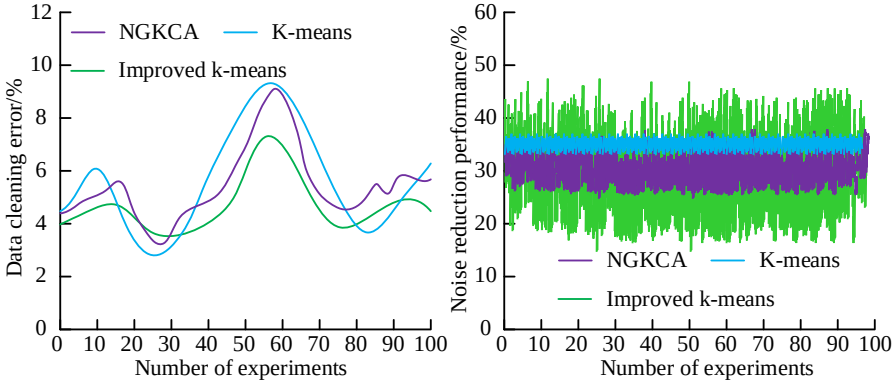


Fig.7 Comparison results of weight values of three methods.

In Fig.7, as the weight index increases, the assigned weights from the three methods exhibit distinct trends. In the processing of educational ability data, weight values were used to determine the importance level of different indicators or data. The weight value of the improved KCA was 0.29, NGKCA was 0.22, and KCA was 0.19. This indicated that improved KCA had better evaluation performance in the analysis of student performance data. To verify the cleaning and noise reduction performance of the evaluation model in educational ability data, the cleaning and noise reduction results after data clustering were used as validation indicators for performance verification. Fig.8 shows the comparison results of data cleaning and noise reduction using three methods after data clustering.



(a) Comparison of data cleaning results after clustering education ability data

(b) Comparison of noise reduction results after clustering of educational ability data

Fig.8 Comparison of data cleaning and noise reduction results of three methods after data clustering.

In Fig.8 (a), as the experiments increased, the cleaning error of educational ability data changed. The data cleaning errors of the three methods were all below 10 %, with the lowest error being the improved KCA, followed by NGKCA and KCA. The data cleaning errors of these three were 5.13 %, 7.26 %, and 7.83 %, respectively. In Fig.8 (b), in the denoising of educational ability data, the improved KCA had a stable denoising ability of 36.19 %, which was stable and efficient throughout the entire process. Although the NGKCA and KCA algorithms occasionally achieved higher peak denoising values than the improved KCA, their performance fluctuated considerably and was overall less stable. Their minimum values were significantly lower than the minimum values of the improved KCA. The noise reduction performance of NGKCA and KCA was 31.85 % and 29.79 %, respectively. This indicated that improved KCA had a more stable cleaning and noise reduction ability in big data processing of educational capabilities. To verify the performance of data clustering fitting in EEA, the study also used the above three methods to verify the clustering fitting performance. Fig.9 shows the comparison results of clustering fitting distribution errors for three methods.

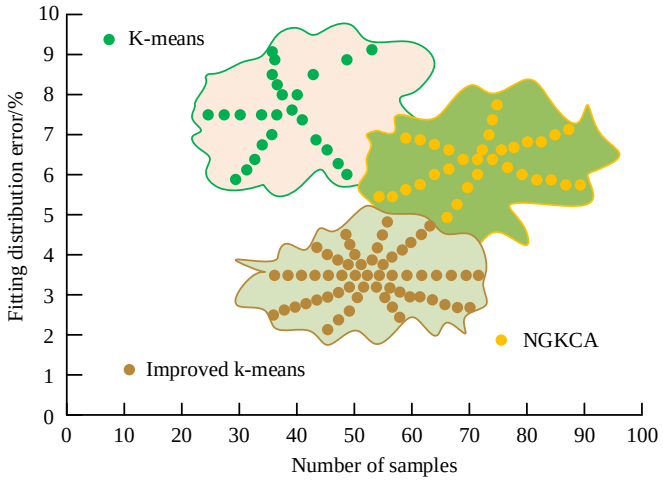
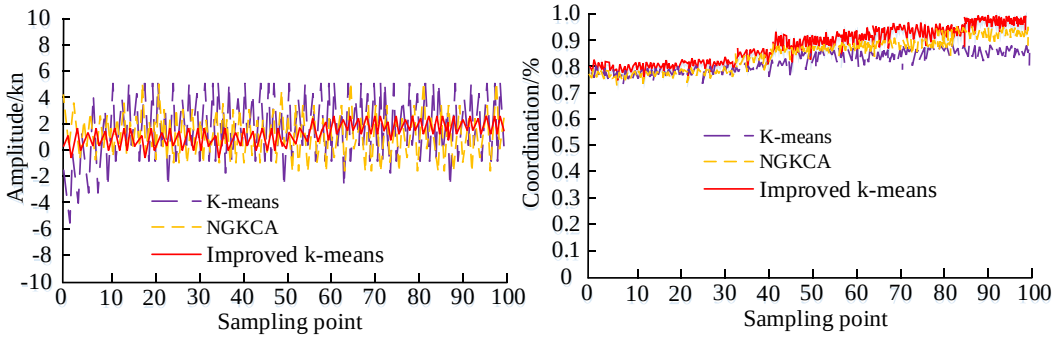


Fig.9 Comparison results of clustering fitting distribution errors of three methods.

In Fig.9, although there is some overlap in the clustering distribution of different methods, there are still certain differences in the overall clustering fit. The improved KCA clustering fitting distribution is the most uniform and effective, with a fitting distribution error of 3.59 %. NGKCA is less effective than KCA, with distribution errors of 6.08 % and 7.51 %, respectively. This indicates that improved KCA has better performance in reducing data complexity and understanding data structures in EEA, which is beneficial for enhancing model evaluation ability. To further verify the performance of EEA model in data analysis, a study is conducted to verify and analyze the waveform and waveform coordination of data in the perspective of big data to reflect its performance in processing data. Fig.10 shows the comparison results of waveform distribution and waveform distribution coordination in the big data field of view.



(a) Comparison results of big data horizon waveform distribution

(b) Comparison results of waveform distribution coordination in big data field of view

Fig.10 Comparison results of big data view waveform distribution and waveform distribution coordination.

In Fig.10 (a), in the distribution of big data field waveforms, the amplitude fluctuation was more pronounced with the increase of sampling points. The improved KCA had the smallest average amplitude of waveform fluctuations among the three methods, with an amplitude of 1.63. The average amplitude of waveform fluctuations between NGKCA and KCA was 0.89 and 0.76. In Fig.10 (b), for the coordination of waveform distribution, the difference between the three methods was not significant when the sampling points were below 50. However, when the sampling points exceeded 50, the coordination of waveform distribution changed significantly. The best coordination effect of waveform distribution was achieved by improved KCA, followed by NGKCA and KCA, which were 91.26 %, 87.23 %, and 81.05 %, respectively. This indicated that in the perspective of big data on educational ability, the improved KCA used for model construction had better coordination performance, which can provide reliable data support for the evaluation of educational ability. The results demonstrate that the performance of the improved KCA is markedly superior to that of KCA and NGKCA in the comparative experiment with data cleaning and

noise reduction, cluster fitting, waveform distribution, and waveform distribution coordination in the context of big data. Compared with other algorithms, the method that fuses the LS method with the improved KCA is more novel and effective.

4.2 Application performance analysis of EEA evaluation model

To verify the application performance of the EEA evaluation model, simulation validation was conducted on the dataset. Using the accuracy of capability assessment and resource utilization as indicators, experiments were conducted using three methods: KCA, NGKCA, and improved KCA. Fig.11 shows the comparison results of evaluation accuracy and utilization of three methods.

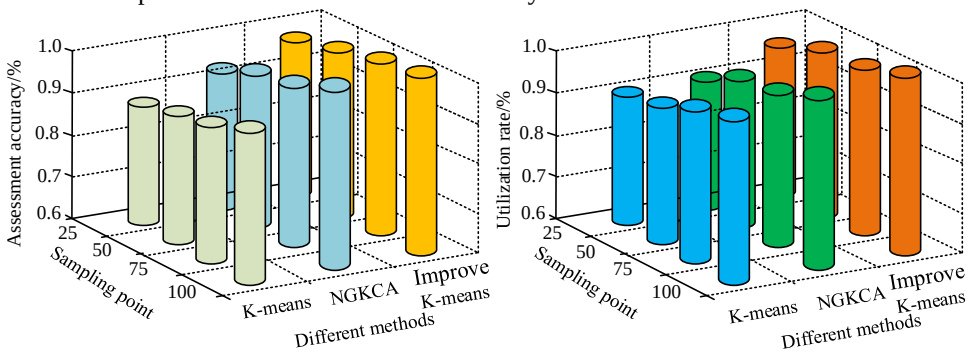


Fig.11 Comparison results of evaluation accuracy and utilization rate of three methods.

In Fig.11 (a), as the number of sampling points increased, the accuracy of these three methods did not change much. This indicated that the influence of sampling points on ability assessment was not significant in EEA. The most accurate evaluation among them was the improved KCA, with an evaluation accuracy of 94.81 %. The evaluation accuracy of NGKCA and KCA were 89.06 % and 83.76%, respectively. In Fig.11 (b), there were also certain differences in resource utilization rates among these three methods in EEA. For resource utilization, the improved KCA was 93.59 %, and the NGKCA and KCA were 88.69 % and 87.92 %, respectively. This indicated that the evaluation model in EEA had higher efficiency and stability in evaluating educational ability. To verify the clustering ability in EEA, the clustering ability of three methods was compared using the clustering center point as the standard. Fig.12 shows the clustering comparison results of three methods for evaluating educational abilities.

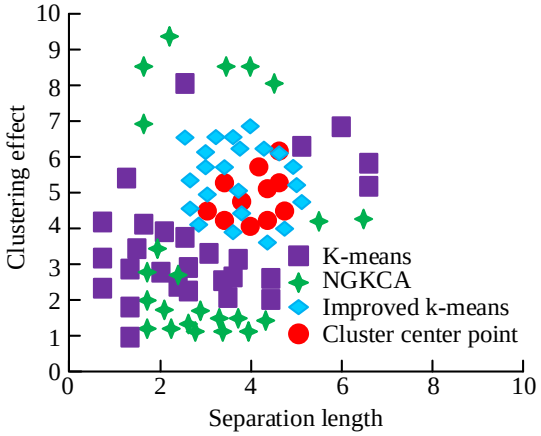


Fig.12 Cluster comparison results of three methods on the evaluation process of educational ability.

In Fig.12, in the educational ability evaluation, the cluster centers were all distributed at the center. The improved KCA had the closest distance to the cluster center point, followed by KCA, and the farthest distance was NGKCA. Optimizing clustering parameters could provide more accurate and reliable evaluation results for EEA, guide educational practice, and improve educational methods. To verify the difference between the calculated values of EEA model and the actual values of educational ability, the study compared the calculated values of this model with the actual values in the training and

testing sets, respectively. Fig.13 shows the comparison between the calculated and actual values of this evaluation model.

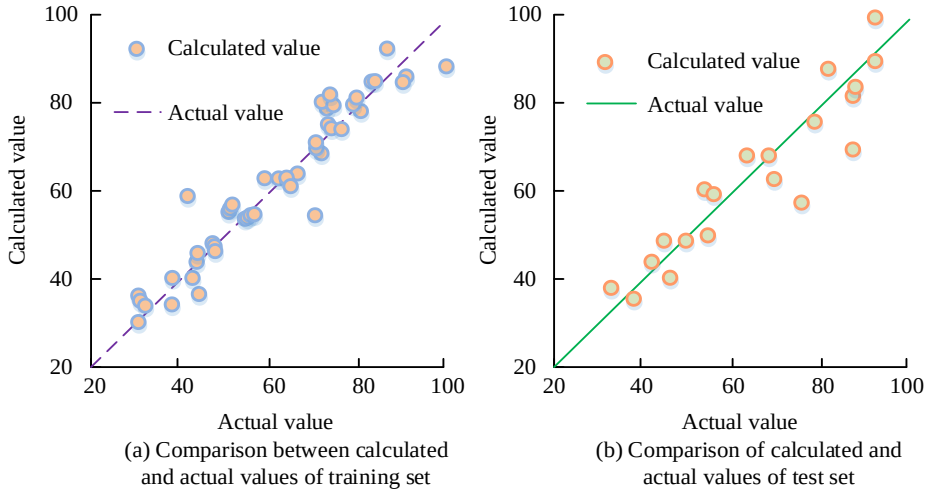


Fig.13 Comparison of the calculated value of the evaluation model with the actual value.

In Fig.13 (a), a large number of calculated values are distributed around the actual values, and the distribution is relatively uniform. This indicates that there is not much difference between the calculated and actual values of educational ability in this model in the training set. In Fig.13 (b), the comparison results of these two values in the test set are slightly worse. However, the calculated values are still distributed around the actual values, which also indicates that the educational ability evaluation model constructed in the study has good performance. To further verify the performance of the educational ability evaluation model, based on different evaluation indicators of educational ability, effective comparisons are made between the calculated values of the model and the actual values. Table 2 shows the comparison results.

Table 2 Comparison of the calculated value of the evaluation model with the actual value

Categorize	Comparative indicators	Validation group							
		1	2	3	4	5	6	7	8
Measurement data	Teaching level	81	79	79	76	80	71	67	81
	Teaching attitude	76	81	80	80	81	76	74	79
	Teaching level	82	77	78	77	76	79	65	80
	Teaching content	80	79	80	79	71	68	70	77
	Teaching methods	79	81	80	72	76	71	77	76
	Teaching attitude	82	75	81	73	80	70	69	81
	Teaching content	79	79	76	78	69	75	63	74
	Teaching equipment	83	75	73	69	78	79	71	77

	Management level	80	76	76	71	73	69	70	70
	Management attitude	79	78	80	70	77	70	68	73
	Actual value	78	80	81	76	79	74	72	78
Comparisons	Calculated value	80.1	78.0	78.3	74.5	76.1	72.8	69.4	76.8
	Error/ %	2.62	2.56	3.44	2.01	3.81	1.65	3.75	1.56

According to the comparative analysis in Table 2, the errors between the calculated values of the evaluation model and the actual values in the 8 validation groups were 2.62 %, 2.56 %, 3.44 %, 2.01 %, 3.81 %, 1.65 %, 3.75 %, and 1.56 %, respectively. All error values were less than 5%. By comparing specific educational indicators, the constructed model could quickly and effectively complete the educational ability evaluation and had good applicability.

The study only evaluates the educational ability of English, and there is data singularity when evaluating, which may have a certain impact on the evaluation. In the future, more interdisciplinary data can be collected for evaluating the performance analysis of models. For example, the proposed model can be applied to the evaluation of computer educational ability. First, the questionnaire survey method is used to collect students' evaluation of teachers' classroom teaching. The evaluation of the questionnaire is divided into five grades: excellent, good, intermediate, pass, and fail. The evaluation direction mainly includes ten aspects: 1. Strict classroom discipline, including good classroom management, both imparting classroom knowledge and conducting moral education; 2. Very familiar with the teaching content, with a good spirit and vitality during lectures, and carefully preparing the teaching plan; 3. Carefully correct homework and answer students' questions seriously; 4. The lecture content is clear and concise, with prominent knowledge points; 5. The classroom atmosphere is lively and the teaching methods are humorous, which enhances students' enthusiasm for learning; 6. Be good at understanding, easy to comprehend, and logically clear when answering questions; 7. The teaching content is novel and innovative; 8. Teach students according to their aptitude, provide diverse educational methods, and cultivate students' personality, creativity, and practical abilities; 9. Make full use of multimedia teaching resources; 10. Be good at inspiring students and guiding students to think. Second, the results of the questionnaire survey are collected for statistics. Then, big data technology combined with KCA is used for evaluation.

5. Conclusion

To address the limitations of low efficiency and high error in traditional EEA evaluation models, this research proposes a novel model based on using big data technology combined with KCA to construct an EEA evaluation model. This model fully utilized big data index coefficients and LS to optimize KCA, aiming to better process large-scale data in the evaluation model and improve the accuracy and efficiency of the evaluation. These results confirmed that in the application performance analysis of the EEA evaluation model, its accuracy and resource utilization rate were 94.81 % and 93.59 %, respectively. In the 8 validation groups, the errors between the calculated values of this evaluation model and the actual values were 2.62 %, 2.56 %, 3.44 %, 2.01 %, 3.81 %, 1.65 %, 3.75 %, and 1.56 %, respectively. The difference between the calculated and actual values of this evaluation model in the simulation experiment was less than 5%. This indicated that the improved KCA evaluation model had better clustering performance and higher classification accuracy in EEA evaluation, providing a more objective and scientific evaluation method for English education. Although this research has achieved good results, there are still certain shortcomings. The limitations of this study lie in the following aspects. Firstly, the focus of the study is solely on the English discipline, which may restrict the universal applicability of the evaluation model. Secondly, the data sources used in the evaluation process are relatively limited, which may affect the comprehensiveness and accuracy of the evaluation results. To enhance the generalization ability of the model and the accuracy of the evaluation, future research could consider collecting more interdisciplinary

data to achieve a more comprehensive analysis of the model's performance. Furthermore, although this study proposes an evaluation method combining questionnaire surveys and big data technology, it does not elaborate on how these technologies are specifically applied in the evaluation process and how they improve the accuracy and effectiveness of the evaluation. Therefore, future work could further explore and optimize these evaluation techniques to ensure the reliability and practicality of the evaluation results.

References

[1] X. Zhao and Y. Jiang, "Synchronously improving multi-user English translation ability by using AI," *Int. J. Artif. Intell. Tools*, vol. 31, no. 4, pp. 7–22, Nov. 2022, doi: 10.1142/S0218213022400073.

[2] H. Mo and X. Yan, "Problems and countermeasures of cultivating autonomous English learning ability in vocational colleges based on informationization," *J. Contemp. Educ. Res.*, vol. 5, no. 6, pp. 1–7, Oct. 2021, doi: 10.26689/jcer.v5i6.2205.

[3] N. Nadia, J. Z. Zamihu, and B. Adu, "Correlation between the students vocabulary mastery and their ability in translating English text into Indonesian at MAS Al-Barokah Kolese," Edu. J., vol. 12, no. 5, pp. 101–118, Nov. 2022, doi: 10.55340/e2j.v8i2.1079.

[4] L. Yu, X. Wu, and Y. Yang, "An online education data classification model based on Tr_MAdaBoost algorithm," *Chin. J. Electron.*, vol. 28, no. 1, pp. 21–28, Jan. 2021, doi: 10.1049/cje.2018.06.006.

[5] X. Bai and J. Li, "Intelligent platform for real-time page view statistics using educational big data digital resource sharing," *J. Intell. Fuzzy Syst.*, vol. 40, no. 2, pp. 2851–2860, Feb. 2022, doi: 10.3233/JIFS-189325.

[6] W. U. Sen and W. Z. Yu, "Clustering algorithm for imbalanced data based on nearest neighbor," *Int. J. Eng. Sci.*, vol. 42, no. 9, pp. 1209–1219, Sep. 2021, doi: 10.13374/j.issn2095-9389.2019.10.09.003.

[7] K. Ahlawat, A. Chug, and A. P. Singh, "A novel hybrid sampling algorithm for solving class imbalance problem in big data," *Adv. Data Sci. Adapt. Anal.*, vol. 13, no. 2, pp. 1–18, Aug. 2021, doi: 10.1142/S2424922X21500054.

[8] T. Teemu and K. Yulia, "Measuring spatial ability for talent identification, educational assessment, and support: Evidence from adolescents with high achievement in science, arts, and sports," *Psychol. Russ.*, vol. 14, no. 2, pp. 59–85, Jun. 2021, doi: 10.11621/pir2021.0205.

[9] L. Familyarska and L. Klots, "Methodological development of educator's information and communication mobility as adaptive ability of educational and developing e-environment self-organization," *ScienceRise: Pedagog. Educ.*, vol. 3, no. 36, pp. 16–21, May 2020, doi: 10.15587/2519-4984.2020.200468.

[10] B. Kartowagiran and E. Rohaeti, "A critical thinking assessment model integrated with science process skills on chemistry for senior high school," *Eur. J. Educ. Res.*, vol. 10, no. 1, pp. 258–298, Jan. 2021, doi: 10.12973/eu-jer.10.1.285.

[11] Y. Ohashi, M. Katsumata, K. Nakamura, H. Hashiura, and H. Yamachi, "Knowledge sharing and teaching ability improvement of teaching assistants and student assistants in a first-year computer programming course," *J. JSEE*, vol. 67, no. 1, pp. 154–159, Feb. 2019, doi: 10.4307/jsee.67.1_54.

[12] X. Ke, "Research on the reconstruction of accounting major curriculums in the age of big data," *in Proc. 4th Int. Conf. Modern Manage., Educ. Technol. Soc. Sci.*, vol. 4, no. 6, Oct. 2019, pp. 670–674, doi: 10.2991/mmetss-19.2019.135.

[13] C. Fischer, Z. A. Pardos, R. S. Baker, J. J. Williams, and M. Warschauer, "Mining big data in education: Affordances and challenges," *Rev. Res. Educ.*, vol. 44, no. 1, pp. 130–160, Mar. 2020, doi: 10.3102/0091732X20903304.

[14] Y. Li, J. Gou, and Z. Fan, "Educational data mining for students' performance based on fuzzy C-means clustering," *J. Ind. Eng. Chem.*, vol. 19, no. 11, pp. 8245–8250, Sep. 2019, doi: 10.1049/joe.2019.0938.

[15] J. Wei and L. Mo, "Open interactive education algorithm based on cloud computing and big data," *Int. J. Internet Protocol Technol.*, vol. 13, no. 3, pp. 151–157, Jan. 2020, doi: 10.1504/IJIPT.2020.10028242.

[16] S. Manikandan, P. Dhanalakshmi, and S. Priya, "Intelligent and deep learning collaborative method for E-learning educational platform using TensorFlow," *Turk. J. Math.*, vol. 12, no. 10, pp. 2669–2676, Apr. 2021, doi: 10.17762/turcomat.v12i10.4881.

[17] S. Wang, J. G. Yabes, and C. H. Chang, "Hybrid density-and partition-based clustering algorithm for data with mixed-type variables," *J. Data Inf. Sci.*, vol. 19, no. 1, pp. 15–36, May 2021, doi: 10.6339/21-JDS996.

[18] M. Gheisari et al., "Data mining techniques for web mining: A survey," *Artif. Intell. Appl.*, vol. 1, no. 1, pp. 3–10, Oct. 2023, doi: 10.47852/bonviewAIA2202290.

[19] M. T. Eser and D. Çobanoglu Aktan, "Educational data mining: The analysis of the factors affecting science instruction by clustering analysis," *Int. J. Educ. Methodol.*, vol. 7, no. 3, pp. 487–500, Aug. 2021, doi: 10.12973/IJEM.7.3.487.

[20] N. Anand, R. Sehgal, S. Anand, and A. Kaushik, "Feature selection on educational data using Boruta algorithm," *Int. J. Comput. Intell. Stud.*, vol. 10, no. 1, pp. 27–35, Jan. 2021, doi: 10.1504/IJCISTUDIES.2021.113826.

[21] O. O. Oladipupo and O. O. Olugbara, "Evaluation of data analytics based clustering algorithms for knowledge mining in a student engagement data," *Intell. Data Anal.*, vol. 23, no. 5, pp. 1055–1071, Oct. 2019, doi: 10.3233/IDA-184254.

[22] A. Alazzam and B. Al-Omar, "Educational data modelling using curve fitting and average uniform algorithm," *Int. J. Grid Util. Comput.*, vol. 10, no. 1, pp. 3–9, Jan. 2019, doi: 10.1504/IJGUC.2019.10018227.

[23] R. Shang, B. Arab, and I. Zada, "Analysis of simple K-mean and parallel K-mean clustering for software products and organizational performance using education sector dataset," *Sci. Program.*, vol. 21, no. 3, pp. 1–20, May 2021, doi: 10.1155/2021/9988318.

[24] O. E. Aissaoui, Y. E. A. E. Madani, and L. Oughdir, "Toward a hybrid machine learning approach for extracting and clustering learners' behaviours in adaptive educational system," *Int. J. Comput. Sci. Math.*, vol. 12, no. 2, pp. 117–137, Nov. 2020, doi: 10.1504/IJCSM.2020.111113.

[25] W. Ding and J. Li, "Higher order extended dynamic mode decomposition based on the structured total least squares," *SIAM J. Sci. Comput.*, vol. 45, no. 2, pp. 985–1011, Jan. 2023, doi: 10.1137/21M1463665.

[26] T. L. James, J. Zhang, H. Li, J. L. Ziegelmayer, and E. D. Villacis-Calderon, "The moderating effect of technology overload on the ability of online learning to meet students' basic psychological needs," *Inf. Technol. People*, vol. 35, no. 4, pp. 1364–1382, Jul.

2022, doi: 10.1108/ITP-03-2021-0225.
[27] A. Chavoshi and H. Hamidi, “Social, individual, technological and pedagogical factors influencing mobile learning acceptance in higher education: A case from Iran,” *Telematics Informat.*, vol. 38, no. 5, pp. 133–165, May 2019, doi: 10.1016/j.tele.2018.09.007.
[28] R. Ferreras-Garcia, J. Sales-Zaguirre, and E. Serradell-López, “Developing entrepreneurial competencies in higher education: A structural model approach,” Educ. + Train., vol. 63, no. 5, pp. 720–743, Mar. 2021, doi: 10.1108/ET-09-2020-0257.
[29] V. A. Lakshmi, M. J. Meena, and S. P. S. Ibrahim, “Big data analytics framework to analyse student's performance,” *Int. J. Comput. Complex. Intell. Algorithms,* vol. 1, no. 2, pp. 154–164, Jan. 2019, doi: 10.1504/IJCCIA.2019.103745.
[30] Z. T. Kosztyán, A. Telcs, and J. Abonyi, “A multi-block clustering algorithm for high dimensional binarized sparse data,” *Expert Syst. Appl.*, vol. 191, no. 12, pp. 1–12, Apr. 2022, doi: 10.1016/j.eswa.2021.116219.