ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Research on Modeling and Precise Profiling of English Learners' Competency Structures Based on Principal Component Analysis Algorithm in Internet Education Platforms

Jun Duan[1,*]

[1] School of Foreign Languages, Xi'an Shiyou University, Xi'an, Shaanxi, 710065, China

SUMMARY: *Against the backdrop of "Internet Plus," the connotation of English learners' learning abilities continues to evolve and adjust. This paper employs principal component analysis (PCA) to screen effective data elements of English graduates' competencies, clarify the logical relationships among various learning ability factors and their contribution rates to the competency structure model, and identify core principal components. By integrating learner profile characteristics and collecting online learning data from multiple students across four semesters of a two-year English program, the study identifies attribute features for constructing English learner profiles. A data analysis model is established using the DPCA method combined with a modified K-means algorithm for selecting initial cluster centers. Through multidimensional feature reduction and extraction of student behavioral data, an objective and detailed learner profile is depicted. Using the DPCA-K-means algorithm, English learners can be categorized into four groups: excellent learners, efficient learners, low-level learners, and high-risk learners. High-efficiency learners demonstrated strong performance in task completion, video viewing progress, and chapter test completion, though their completion rates were not the highest. They exhibited the highest video replay rates and participated most frequently in online discussions, indicating proactive and motivated learning.*

KEYWORDS: *Principal Component Analysis; K-means Algorithm; English Learner Profiling; Learning Competency Factors*

1 Introduction

In today's digital age, English education is no longer confined to traditional classroom learning models. With the development of the internet and increasing demand for education, numerous online learning platforms have emerged [1, 2]. These platforms offer students more flexible and diverse learning methods, significantly enhancing the accessibility and effectiveness of English education [3-5]. However, in a highly competitive market, the challenge for online education platforms lies in understanding and analyzing learner behavior to build precise user profiles, thereby delivering more personalized and tailored learning experiences.

User profiling involves analyzing and organizing personal information to create individualized, multidimensional descriptions of users [6, 7]. Online education platforms can construct user profiles based on fundamental details, learning habits, and educational interests [8]. First, inferring learning needs and goals from basic user information such as age, gender, education level, and occupation [9, 10]. Users of different age groups and professions may

*junduan7@163.com
https://doi.org/10.65102/is2026639
Published: 30 April, 2026

exhibit distinct learning requirements and objectives [11]. Second, understanding learning behaviors through habits like study duration, frequency, and preferred time slots [12, 13]. Some may prefer utilizing fragmented time for learning, while others may favor fixed study periods. Third, user learning preferences are tracked through interests in specific subjects or course types [14-16]. The degree of interest in a particular subject or course category influences users' willingness and motivation during the learning process [17]. By analyzing users' basic information, learning habits, and learning interests, the platform can better understand users and provide personalized learning services tailored to their needs [18, 19].

This paper analyzes the teaching advantages of micro-courses, MOOCs, and online classrooms in the context of "Internet + Education," highlighting the need to adjust the competency development requirements for English learners amid the current development of internet education platforms. It proposes a principal component analysis (PCA) algorithm, combined with data on competency elements for university English graduates, to analyze learners' competency factors. The paper describes the essence and characteristics of learner profiling, refines the PCA algorithm, and integrates clustering methods to present a precise profiling solution for English learners based on the DPCA-K-means algorithm. By extracting multidimensional features from learner data sources and categorizing learner types, it achieves accurate learner profiling.

2 The Development of English Teaching in the Context of "Internet Plus Education"

2.1 Advantages of Micro-Lessons, MOOCs, and Online Classrooms

(1) Breaking down barriers of learning time and location to grant students greater flexibility in scheduling. The "Internet + Education" model eliminates constraints on students' learning schedules, geographic locations, and content, bridging the temporal and spatial gap between teachers and learners. Currently, most online courses offered are free, significantly reducing students' learning costs. Through the internet, students can freely select any course content from any teacher in any country based on their personal interests and knowledge level.

(2) Catering to diverse learner needs with personalized options. Traditional teaching models cannot guarantee alignment between teaching schedules and content with students' actual learning demands. Pre-set teaching timetables and curricula often fail to meet individualized learning requirements. Under the "Internet + Education" model, the internet establishes a swift and practical interactive platform for students and teachers. Teachers' instructional materials can be reused by students, who can also proactively select content that interests them. Students can directly provide feedback on their personalized needs to teachers via the internet, enabling teachers to offer tailored solutions based on these requirements.

(3) Optimizing Teaching Resources for More Rational Learning Traditional teaching models often fail to meet students' and parents' demand for high-quality educational resources due to limited access to prestigious schools or renowned teachers. Emerging internet-based teaching models address this need. Content from micro-lectures, MOOCs, and online classrooms can be accessed anytime, anywhere, with unlimited replays. This liberates premium teaching resources, enabling students to enjoy richer, more diverse learning materials. Internet information technology fulfills the needs for spatial dispersion, temporal flexibility, and matching educational supply with demand. It optimizes teaching resources, enabling students to learn more rationally.

2.2 Development of English Learners' Competencies

In the context of “Internet Plus,” the essence of English learners' autonomous learning ability primarily manifests in their capacity to achieve personalized learning goals through effective learning strategies and resource management while independently controlling their learning process.

Supported by internet technology, learners can access knowledge through diverse media such as online platforms and mobile applications, transcending temporal and spatial constraints. Autonomous learning ability extends beyond merely selecting and mastering learning content; it emphasizes the learner's capacity for self-monitoring and self-assessment throughout the learning journey. In this process, students must develop efficient resource screening and integration skills to identify high-quality content within vast information oceans, thereby formulating practical learning plans.

The technological support enabled by “Internet Plus” transforms self-directed learning from mere individualization of the learning process into the digitalization and intelligentization of learning methods. Learners can flexibly arrange their studies and receive timely feedback adjustments through online classrooms, interactive platforms, and self-directed learning tools.

3 Modeling the Competency Structure of English Learners

3.1 Principal Component Analysis

Classic principal component analysis aims to encapsulate most of the information from the original data through a small number of composite indicators derived from linear transformations of the original data. These composite indicators are mutually independent, thereby reducing computational complexity and simplifying data processing [20-22]. Let the original time series data be $X_n \times p = (Z_1, Z_2, \cdots, Z_p)$ containing n samples, each with p variables. The original data observation matrix is given by Equation (1):

$$X_{n\times p} = \begin{bmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{np} \end{bmatrix} \tag{1}$$

where $Z_i = (a_{ii}, a_{2i}, \cdots, a_{ni})^T$ for $i = 1, \cdots, p$.

Perform a linear transformation on the original data, as shown in Equation (2):

$$\begin{cases} f_1 = \mu_{11}Z_1 + \mu_{21}Z_2 + \cdots + \mu_{p1}Z_p = \mu_1^T z \\ f_2 = \mu_{12}Z_1 + \mu_{22}Z_2 + \cdots + \mu_{p2}Z_p = \mu_2^T z \\ \cdots \\ f_p = \mu_{1p}Z_1 + \mu_{2p}Z_2 + \cdots + \mu_{pp}Z_p = \mu_p^T z \end{cases} \tag{2}$$

The new composite indicator is set as $f = (f_1, f_2, \cdots, f_p)$, where f_i are mutually independent. Principal component analysis seeks to solve for μ_i such that the variance of the $k(k \le p)$ linear transformations $f_i = \mu_i^T z$ satisfies: $Var(f_i) = \mu_i^T \Sigma \mu_i$ (where Σ is the covariance matrix of z). That is, the objective function is given by Equation (3):

$$\begin{cases} \max \mathrm{Var}(f_i) = \max \mu_i^{\mathrm{T}} \Sigma \mu_i \\ \mu_i^{\mathrm{T}} \mu_i = 1 \end{cases} \tag{3}$$

Let the eigenvalues of the covariance matrix be λ_i, arranged in descending order as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Applying the Lagrange multiplier method yields $\max Var(f_i) = \lambda_1$, where λ_i is the largest eigenvalue of the covariance matrix and μ_i is the corresponding eigenvector. We define $f_1 = \mu_1^T z$ is the first principal component.

In principle, p principal components can be obtained. However, including all p components in the model defeats the purpose of simplification. Therefore, a criterion is typically used to retain only k principal components, ensuring these k components capture most of the information from the original data. A commonly used criterion is variance contribution rate. Specifically, the first k principal components are selected such that the cumulative variance contribution rate exceeds a certain threshold, such as 75%.

3.2 Principal Component Analysis of Competency Factors and Model Construction

The data originates from a third-party data company commissioned by a science and engineering university to evaluate the quality of English graduate training for the Class of 2024.

This data excludes liberal arts majors and missing values, covering 50 majors across 12 science and engineering colleges. It comprises 1,896 valid responses, with a response rate of 30.5% among science and engineering students.

Twelve elements of university English learners' competencies form the English Talent Structure Model. To further clarify the logical relationships among these elements and their contribution rates to the competency structure model, SPSS was employed for in-depth analysis of the survey data on English learner competency elements.

The total variance explained is shown in Table 1. The test results indicate: KMO value of 0.960, Bartlett's sphericity test P-value of 0.000, and chi-square value of 25347.231. Factor analysis, an extension of principal component analysis, tends to employ rotation techniques to identify and interpret latent factors, representing variables as linear combinations of these factors. Principal component analysis, however, focuses on reducing existing variables to a smaller set of new variables, expressing principal components as linear combinations of the original variables. Therefore, principal component analysis is more suitable for this study. Through PCA, it is evident that the first four principal components account for 87.019% of the total variance. This indicates that these four components sufficiently represent the information of the original 12 elements. Given that the cumulative contribution rate of principal components exceeds 85%, and based on core literacy theory, fewer principal components are preferable to identify the core principal components.

Table 1: The total variance of the explanation

Constituent	Initial eigenvalue			Extract the sum of squares and load GP components		
	Total	Variance %	Cumulative%	Total	Variance %	Cumulative%
1	8.651	75.365	75.365	8.651	75.365	75.365
2	0.911	7.621	80.215	0.911	7.621	80.215
3	0.512	4.509	84.332	0.512	4.509	84.332
4	0.365	2.653	87.019	0.365	2.653	87.019

The principal component matrix after data standardization is shown in Table 2. The 12 competency elements for English learners include: knowledge and skills, problem analysis ability, solution development ability, research ability, modern tools usage ability, social application ability, sustainable development ability, professional ethics ability, individual and teamwork ability, communication ability, lifelong learning ability, and English professional ability.

Table 2: The main component matrix after standardization

		Y1	Y2	Y3	Y4
Knowledge ability	X1	0.856	-0.315	0.121	0.095
Problem analysis ability	X2	0.812	-0.235	0.187	0.012
Develop solution capabilities	X3	0.736	-0.312	-0.052	-0.082
Research ability	X4	0.814	-0.254	0.061	0.006
Use modern tools	X5	0.873	-0.138	0.023	-0.061
Social ability	X6	0.804	-0.052	-0.214	0.023
Sustainable development ability	X7	0.817	0.268	-0.336	0.321
Occupational specification	X8	0.859	0.347	-0.233	-0.074
Personal and team ability	X9	0.872	0.339	0.036	-0.201
Communication ability	X10	0.854	0.452	0.095	0.232
Lifelong learning ability	X11	0.822	0.216	0.214	0.331
English professional ability	X12	0.836	0.232	-0.263	-0.255

The four principal components are denoted as Y1, Y2, Y3, and Y4. First, the original data were standardized using SPSS, followed by principal component analysis to obtain the principal component matrix. Table 2 presents the linear combinations of the four principal components.

The essence of English learners' abilities can be explained through four key components: English Proficiency (P), English Literacy (Q), Knowledge Acquisition Ability (K), and Career Development Ability (C). The relationships among these four components are intricate and interconnected, forming twelve distinct clusters of interactive relationships.

4 Precise Profiling of English Learners

4.1 Implications and Characteristics of Learner Profiles

Learner profiling is the application of user profiling in the education sector. User profiling essentially involves extracting the most comprehensive and detailed picture of user information possible through data mining and refinement. This is achieved by generating sets of digital tags to characterize and predict user behavior, thereby helping businesses solve the problem of converting data into value. Currently, user profiling is widely applied across numerous fields and industries—including commerce, healthcare, management, and intelligence—to enhance user experience, deliver precision services, and provide references for management decisions.

Characteristics of learner profiling are as follows:

(1) Educational Value of Data Sources

Compared to general user profiling data sources, learner profiling data emphasizes “educational contextuality” and “value density.” To achieve this, researchers employ three primary methods—social surveys, web data collection, and perception technology—to gather multimodal data throughout the learner's entire process.

Social surveys, represented by interviews, observations, and questionnaires, capture

learners' psychological traits, higher-order abilities, and competency indicators. Examples include using learning style and motivation scales to collect learner typology data.

Online data collection captures network data that traces learners' behavioral trajectories. This involves using web crawlers to obtain explicit data such as learners' voluntarily provided personal information, comments, and ratings. Log mining and platform database collection techniques capture learning behavior logs, including resource access sequences and session durations.

Sensor technologies, leveraging smart devices or wearables, capture interactive data generated during learning—such as text, images, audio, video, visual, and tactile inputs. For instance, EEG experiments analyze brainwave patterns to map learner attention and interest profiles.

(2) Methodological Integration in Modeling

Unlike traditional learner analysis methods, learner profiling modeling integrates both quantitative and qualitative approaches. Quantitative methods involve detailed statistical analysis and computation of learner data to precisely identify learning needs. Qualitative methods analyze educational contexts and learner mental states to abstract and generalize user characteristics, thereby assigning semantic meaning to generated tags.

Simultaneously, due to the interplay between individual and social realities in education, learner profiling must demonstrate a hierarchical integration of personal and group profiles. First, individual-based profiling identifies unique learner characteristics to comprehensively understand personal learning needs. Building upon this foundation, it enables functions like individual assessment, personalized resource recommendations, and user behavior prediction through relevant technologies. Second, group learner profiles enable segmentation of learners and uncover patterns of collective behavioral characteristics within specific educational contexts, providing decision support for teachers and educational administrators.

(3) Multidimensionality of Profiling Models

Learner profiling essentially constitutes a set of feature tags, with the tag library determined by the profiling model. To achieve objectives like learner trait analysis, group identification, and evaluation, learners can be comprehensively characterized across five dimensions—general learner traits, knowledge-ability structure, implicit psychology, explicit learning behaviors, and learning interactions—across diverse scenarios.

Specifically:

(a) General learner characteristics encompass basic information that outlines individual learner profiles.

(b) The knowledge and competency dimension reflects learners' mastery of subject knowledge, ability levels, and skill proficiency, typically modeled based on assessment data.

(c) The implicit psychological dimension represents learners' mental states, including learning styles, interests, emotions, and their evolving trends.

(d) The learning behavior dimension encompasses behavioral representations generated during the learning process, such as task performance and information dissemination.

(e) The learning interaction dimension covers learners' interactions with peers, instructors, and resources.

4.2 Student Profiling Based on the DPCA-K-means Algorithm

(1) Fundamental Concept of the K-means Algorithm

The K-means algorithm is an unsupervised learning method that clusters data points based on their mean values. Its key principle involves randomly selecting cluster centers at different positions while ensuring sufficient distance between these centers. The core concept is as follows: First, select the number of clusters k. Then, randomly choose samples as initial cluster

centers. For each sample, examine its distance to each current cluster center and assign it to the nearest cluster. After completing one round of sample assignments, update the center of each cluster. Repeat this process until no further changes occur in the cluster centers, indicating convergence. The k clusters exhibit the following characteristics: Data points within each cluster are as compact as possible, while points between clusters are as distant as possible. The algorithm converges by minimizing the sum of squared errors objective function. The clustering process and its outcome are fundamentally influenced by the initial center selection; the performance of the mean-centered clustering algorithm hinges on the rational choice of cluster centers. The sum of squared errors is defined as in Equation (4):

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

In the equation, X represents the data points in the dataset. C_i denotes the centroid.

TSS is a strict coordinate descent process that employs Euclidean distance as the metric function between variables. At each iteration, it finds the optimal solution by approximating the direction toward a variable C_i. Taking the partial derivative of the objective function TSS and setting it equal to zero yields Equation (5):

$$C_i = \frac{1}{k} \sum X \tag{5}$$

In the formula, k denotes the number of elements in the cluster containing C_i.

The mean of the current cluster represents the optimal solution (minimum value) for the current direction. Similar to each iteration in K-means, this ensures that the TSS value decreases after each iteration and eventually converges. However, since TSS is a non-convex function, it cannot guarantee finding the global optimum but only ensures a local optimum. To prevent K-means from settling on local optima during data processing, the DPCA approach is employed. This involves reducing the dimensionality of the data, followed by reprojecting it to preserve all feature information. The weight of each feature is recorded, and the K-means algorithm is executed multiple times. The final result is selected based on the smallest TSS value.

(2) Fundamental Concept of DPCA-K-means Algorithm

To eliminate differences in dimensionality and magnitude among features, data is normalized to obtain a standardized matrix. Assume the dataset contains n data points $X = \{X_1, X_2, \cdots, X_n\}$ divided into k clusters $\{X_1, X_2, \cdots, X_i\}$. with k centroids $C = \{c_1, c_2, \cdots, c_i\}$. The k clusters must satisfy the following condition: $X_1 \cup X_2 \cup X_k = X$. $C_k \in X_k; X_i \neq \Phi$, where $k = 1, 2, \cdots, n$.

To eliminate the effects of dimensions and scales, the data is first normalized to create a dimensionless dataset, facilitating comparison and weighting of metrics with different units or magnitudes. The normalization process is described by Equation (6):

$$x_{ij} = \frac{x_{ij} - \overline{x}_i}{\sigma_i} \tag{6}$$

In the equation, x_i represents the source data for the j th sample of the i th system feature. $\overline{x}_i$ and σ_i denote the mean and standard deviation of the i th feature system,

respectively.

Based on the standardized data matrix, a covariance matrix is constructed to reflect the degree of correlation among the standardized data. Its definition is given by Equation (7):

$$COV = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{x})(X_i - \overline{x})^T \tag{7}$$

In the formula, n represents the number of data points. X_i denotes any single data point. $\overline{x}$ represents the mean value of all data points.

The sample mean is calculated as in Equation (8):

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{8}$$

Calculate the eigenvalues, principal component contribution rates, and cumulative variance contribution rates based on the covariance matrix R. Arrange the eigenvalues $\lambda_i (i = 1, 2, \cdots, n)$ in descending order, i.e., $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_i \geq 0$. The eigenvalues represent the variance of each principal component, reflecting the influence of each component. Define the weight of the partition in terms of the total information as in Equation (9), and define the sum of all partition weights as in Equation (10):

$$W_i = \sqrt{\frac{\lambda_i}{\sum_{i=1}^{n}\lambda_i^2}} \tag{9}$$

$$W = \sqrt{\sum_{i=1}^{k}\lambda_i^2} \tag{10}$$

In the formula, W_i represents the weight of each feature relative to all features. W denotes the total sum of the weights of all feature values.

In this paper, the weight factor W is set to 1, meaning the cumulative proportion of feature values sums to 100%. After reweighting the data, the new sample dataset is formed as $D = \{X_1', X_2', \cdots, X_n'\}$. New features are obtained by multiplying the weight factor by each attribute, as shown in Equation (11):

$$d_i = (W_1 X_{i1} + W_2 X_{i2} + \cdots + W_m X_{im}) / m \tag{11}$$

Euclidean distance is used to measure the similarity between data points and cluster centers. The dataset $X = \{X_1, X_2, \cdots, X_n\}$ consists of n data points. Its feature dimensions are formed by m element values $X_i = \{x_{i1}, x_{i2}, \cdots, x_m\}$, creating a m-dimensional feature space. Each feature within this feature set constitutes the numerical value for each dimension. In the m-dimensional space, the two data matrices each form one point. Calculating the distance between these two points yields the Euclidean distance. This is described as follows:

n data point $X = \{X_1, X_2, \cdots, X_n\}$, $X_i = \{x_{i1}, x_{i2}, \cdots, x_{in}\}$, where the distance $D(X_i X_j)$

between the two data points $X_i = \{x_{i1}, x_{i2}, \cdots, x_{in}\}$ and $X_j = \{x_{j1}, x_{j2}, \cdots, x_{jm}\}$ is calculated as in Equation (12):

$$D(X_i X_j) = \sqrt{\sum x_i x_j \mathrm{e} X (X_i - X_j)^2} \tag{12}$$

In the formula, m denotes the feature dimension of the data point.

The first initial point is selected as the data point closest to the mean, denoted as C_1. When selecting the second initial point, define the distance WD_i from each partitioned mean point m_i to the first initial point C_i. Select the data point with the largest WD_i as the second initial cluster center, denoted as C_2 Here, the distance WD_i from m_i to the first initial point C_1 is defined as the weighted Euclidean distance, as shown in Equation (13):

$$WD_i = W_i \sqrt{(c_{i1} - m_{j1})^2 + \cdots + (c_{i1} - m_{jk})^2} \tag{13}$$

The description of the minimum distance selection is as shown in Equation (14):

$$MinD = \min\{D(X_i X_{j_d})\} \tag{14}$$

The algorithm convergence criteria are error smoothing and minimum value.

4.3 Experiments and Results Analysis

4.3.1 Experimental Environment and Data Sources

This study employs Python 3.8 as the experimental platform, running on a Windows 10 operating system with a CoreTM i7-9750H CPU. Experimental data was sourced from the SuperStar Learning Platform, collecting online learning data from 330 students across four semesters over two years in an English course at a science and engineering university. The data comprises two components: student basic attribute data and learning behavior data. Basic attribute data includes student ID, enrollment year, and gender. Learning behavior data encompassed: task completion rate, course video progress, average video replay ratio, chapter test progress, number of completed tasks, video viewing duration, discussion participation frequency, chapter review frequency, average chapter test score, overall grade, and grade level.

4.3.2 DPCA-K-means Multidimensional Feature Extraction

(1) Linear Feature Transformation Based on DPCA

By combining the collected learner data and calculating the cumulative contribution rate of each feature dimension, it is evident that 12 principal components can represent over 95% of the original data's information. Therefore, the 12 transformed features are extracted as the linear features of the dataset.

(2) Nonlinear Feature Learning Based on K-means

After multiple comparative experiments, this study employs a Gaussian-Bernoulli restricted Boltzmann machine with 18 hidden layer neurons for feature learning. With a learning rate of 0.01 and 50 iterations, the network converges with a mean squared error of 17.24.

The network outputs were recorded, yielding 18 new features learned by K-means. These were concatenated with the 12 linear features obtained from DPCA transformation, producing a 258×21 feature matrix after multidimensional feature extraction. A three-dimensional scatter

plot of partial raw data is shown in Figure 1. It is evident that due to the diversity and individuality of learner behaviors, the categorical features in the original dataset are not distinctly discernible.

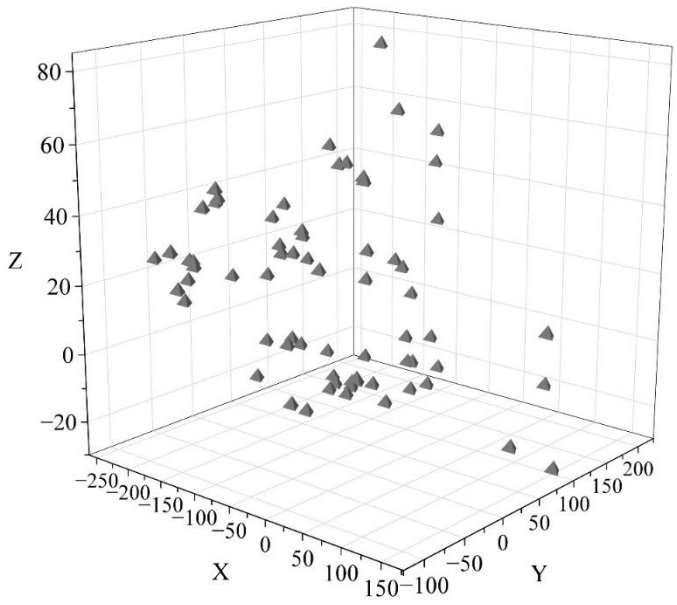

Figure 1: The three-dimensional divergence of some original data

Figure 2 shows the three-dimensional scatter plot of a portion of the data after multidimensional feature extraction using the DPCA-K-means algorithm. Following multidimensional feature extraction, distinct categorical features become apparent. However, due to the distribution characteristics of the original data, it is challenging to determine initial cluster centers, making partition-based clustering algorithms unsuitable. Therefore, this paper employs the K-means algorithm based on principal component analysis.

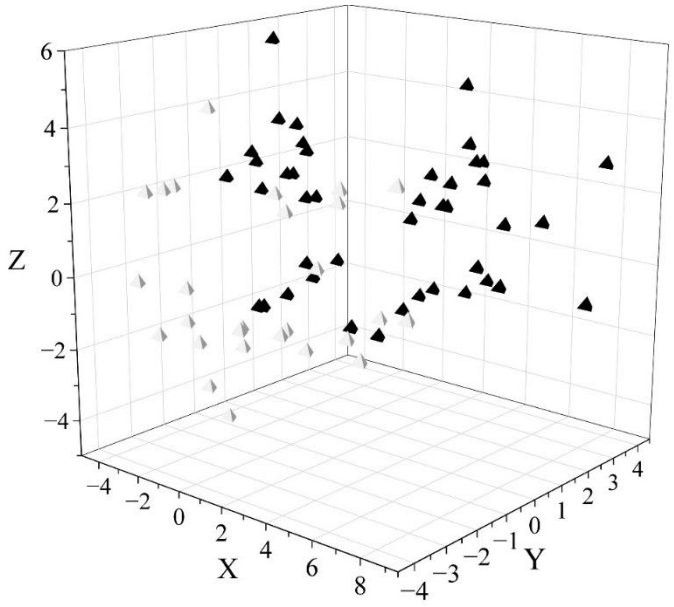

Figure 2: The algorithm extracts some of the data three dimensional scattered points

4.3.3 Learner Profile Analysis

Based on the clustering results generated by the DPCA-K-means algorithm, four learner categories were constructed within the dataset. The distribution of learners across each category is shown in Table 3. Cluster 2 accounts for 52% of the total learners.

Table 3: Statistics on the number of learners

		Number
The clustering results generated by the DPCA-K-means algorithm	Cluster 1	93
	Cluster 2	173
	Cluster 3	38
	Cluster 4	26

Figure 3 illustrates the average distribution of learning behaviors across eight characteristics for each learner cluster, including task completion rate, video viewing progress, and average video review ratio. (Full score: 10 points)

It can be observed that Cluster 1 learners demonstrated the strongest performance in task completion (8.3), video viewing progress (7.5), and chapter test completion (8.3), while also achieving the highest average chapter test scores. This reflects that learners in this category spend more time on coursework and possess solid foundational knowledge. However, their video replay ratio and discussion participation frequency are slightly lower than those of Cluster 2 learners, indicating weaker critical thinking and proactive learning abilities, resulting in lower overall scores compared to Cluster 1 learners. This category is defined as excellent learners.

Cluster 2 learners constitute the largest proportion, accounting for 52% of all learners. This group performed well in task completion (7.8), video viewing progress (7.3), and chapter test completion (8.1), though their completion rates were not the highest. They exhibited the highest video replay ratio (8.9), indicating repeated viewing of key knowledge points and difficult concepts. They exhibited the highest participation in online discussions, demonstrating proactive and motivated learning. Ultimately, they achieved the highest overall scores and can be defined as high-efficiency learners.

Cluster 3 learners showed the lowest completion rates across all content areas but preferred taking chapter tests, achieving relatively high scores on them. Considering chapter test scores contribute to the course's regular assessment, these learners are driven by grades rather than fully leveraging their initiative. They are defined as low-level learners.

Cluster 4 learners did not have the lowest task completion rates but achieved the lowest overall scores. These learners face challenges in the learning process and lack targeted practice, making them a group requiring special attention in teaching. They are defined as high-risk learners.

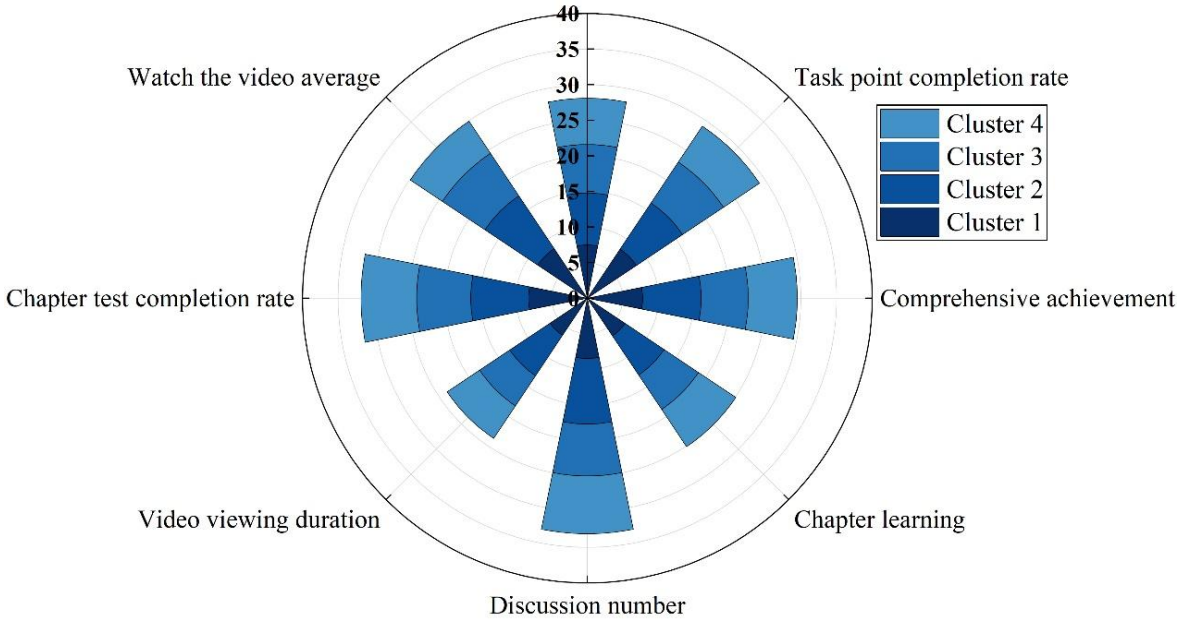

Figure 3: The average separation of the learning behavior of each learner class

5 Conclusion

This paper employs principal component analysis (PCA) to examine the competency factors of English learners and constructs a competency structure model for them. By integrating PCA with the K-means clustering algorithm, it creates precise profiles of English learners.

Through PCA, the four principal components—English major proficiency, English literacy, knowledge acquisition ability, and career development ability—accounted for 87.019% of the total variance, illustrating the competency structure of university English learners.

The multidimensional feature extraction based on the DPCA-K-means algorithm enables the analysis of English learners into four categories: excellent learners, efficient learners, low-level learners, and high-risk learners. The development of “Internet+Education” English teaching can be tailored to the specific characteristics of different learners.

References

[1] Wen, J., Wei, X., He, T., & Zhang, S. (2020). Regression Analysis on the Influencing Factors of the Acceptance of Online Education Platform among College Students. Ingénierie des Systèmes d'Information, 25(5).

[2] Malkawi, N., Rababah, M. A., Al Dalaeen, I., Ta'amneh, I. M., El Omari, A., Alkhaldi, A. A., & Rabab'ah, K. (2023). Impediments of using e-learning platforms for teaching English: A case study in Jordan. International Journal of Emerging Technologies in Learning (Online), 18(5), 95.

[3] Pu, C. (2022). Design and Application of English Online Learning Platform. Wireless Communications and Mobile Computing, 2022(1), 9693192.

[4] Li, J. (2021). Design, implementation, and evaluation of online English learning platforms. Wireless Communications and Mobile Computing, 2021(1), 5549782.

[5] Hu, H., Wang, X., Zhai, Y., & Hu, J. (2021). Evaluation of factors affecting student participation in peer-assisted English learning based on online education platform. International Journal of Emerging Technologies in Learning (iJET), 16(11), 72-87.

[6] Kaur, R., Gupta, D., Madhukar, M., Singh, A., Abdelhaq, M., Alsaqour, R., ... & Goyal, N. (2022). E-Learning environment based intelligent profiling system for enhancing user adaptation. Electronics, 11(20), 3354.

[7] Adib, J., Ait Abdelouahid, R., Marzak, A., & Moutachaouik, H. (2022). Ontological user profile for E-orientation platforms. Procedia computer science, 198, 417-422.

[8] Vilenchik, D. (2020). An Unsupervised Approach to User Characterization in Online Learning and Social Platforms. In Mathematics (Education) in the Information Age (pp. 15-36). Cham: Springer International Publishing.

[9] Kirange, S., Sawai, D., & Director, I. M. (2021). A comparative study of e-learning platforms and associated online activities. The Online Journal of Distance Education and e-Learning, 9(2), 194-199.

[10] Decuypere, M. (2019). Open Education platforms: Theoretical ideas, digital operations and the figure of the open learner. European Educational Research Journal, 18(4), 439-460.

[11] El Mabrouk, M., Gaou, S., & Rtili, M. K. (2017). Towards an intelligent hybrid recommendation system for e-learning platforms using data mining. International Journal of Emerging Technologies in Learning (Online), 12(6), 52.

[12] El Haddioui, I., & Khaldi, M. (2012). Learner behavior analysis on an online learning platform. International Journal of Emerging Technologies in Learning (iJET), 7(2), 22-25.

[13] Ouadoud, M., Chkouri, M. Y., Nejjari, A., & El Kadiri, K. E. (2016, October). Studying and comparing the free e-learning platforms. In 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt) (pp. 581-586). IEEE.

[14] Xiao, J., Wang, M., Jiang, B., & Li, J. (2018). A personalized recommendation system with combinational algorithm for online learning. Journal of ambient intelligence and humanized computing, 9(3), 667-677.

[15] Rajae, Z., & Said, A. (2022). Trace Learners Clustering to Improve Learning Object Recommendation in Online Education Platforms. International Journal of Advanced Computer Science and Applications, 13(6).

[16] Ying, Y. (2020, March). Characteristics and challenges of Chinese e-learning platforms in Indonesia. In Journal of Physics: Conference Series (Vol. 1477, No. 4, p. 042014). IOP Publishing.

[17] Amane, M., Aissaoui, K., & Berrada, M. (2021, January). A multi-agent and content-based course recommender system for university e-learning platforms. In International Conference on Digital Technologies and Applications (pp. 663-672). Cham: Springer International Publishing.

[18] Zhou, L., Xue, S., & Li, R. (2022). Extending the Technology Acceptance Model to explore students' intention to use an online education platform at a University in China. Sage Open, 12(1), 21582440221085259.

[19] Madani, Y., Erritali, M., Bengourram, J., & Sailhan, F. (2019). Social collaborative filtering approach for recommending courses in an E-learning platform. Procedia Computer Science, 151, 1164-1169.

[20] Astrid Mellbin, Udaya Rongala & Fredrik Bengtsson. (2025). Protocol for extracting and evaluating activity distributions in rat electrocorticograms with principal component analysis and k-nearest neighbor. STAR protocols,6(3),104041.

[21] Nicholas J. Connors, Christopher Monaghan, Björn Benneke & Lisa Dang. (2025). Uniform Reanalysis of JWST MIRI 15 μm Exoplanet Eclipse Observations Using Frame-normalized Principal Component Analysis. The Astrophysical Journal Letters, 989(1), L11-L11.

[22] A. A. Dhorde, T. P. Raut, A. G. Dhorde & N. Gautam. (2025). Assessing Basin Characteristics in the Mula-Mutha Watershed: Using Sub-Basin Morphometry and Principal Component Analysis. Geography and Natural Resources, 46(1),105-116.