International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 17 (2025) pp. 1-14

https://doi.org/10.70917/ijcisim-2025-0239
Article

Research on teaching state space dimensionality reduction and enhancement strategies incorporating principal component analysis in open education platforms

Liurong Peng *

College of Education, Sichuan Institute of Industrial Technology, Deyang, Sichuan, 618500, China;
pengliurong0508@163.com

Abstract: In the context of the rise of online education, learners have generated large-scale learning behavior records in open education platforms providing sufficient material for educational data mining. In this paper, based on edX open education dataset, 11 features representing learners' learning status are downscaled by principal component analysis technique, learners are clustered based on learning status and using the proposed K-means clustering algorithm, and the number of clusters is determined by the aggregation coefficient method. The experimental results show that the aggregation coefficient peaks when the number of clusters is 4. Therefore, the optimal teaching state is categorized into 4 types of spaces, which are general learning space, negative learning space, interactive learning space, and active learning space. Finally, on the basis of the clustering results, corresponding teaching enhancement strategies are proposed for the strengths and weaknesses of the groups. The method proposed in the article can identify and analyze students' learning status in real time and accurately, improve teachers' mastery of students' classroom performance, and provide a new research idea and technical reserve for the development of open education.

Keywords: principal component analysis; K-means clustering; edX open education; teaching state

1. Introduction

The development of E-learning has entered a relatively mature period after 2010, forming a basic online learning logic and framework [1]. In recent years, the rapid development and widespread application of Moocs and other global learning platforms have marked that the informatization of open education platforms has entered the stage of integrated development and innovation [2]. With the popularization and gradual deepening of educational information visualization, various digital learning systems have generated a large amount of educational information, i.e., the analysis and application of educational big data has become a research hotspot [3-4].

Learning analytics is the application of “big data” in the field of education, which refers to the use of tools to measure and collect the data of students' online learning to build appropriate models for analysis [5]. In the field of open education, learning analytics is used to analyze potential problems in student learning, predict student performance, improve student engagement and retention in new ways, and provide students with high-quality and personalized learning experiences [6-7].

Principal Component Analysis is a very powerful data processing tool in learning analytics and is a method for transforming multidimensional spatial problems into low-dimensional spatial problems [8]. The essential idea is the dimensionality reduction of the data, which reduces the complexity of data analysis by reducing the number of indicators in the original dataset and forming fewer new indicators

[9]. And the mutual uncorrelation between the new indicators ensures the completeness of the original information to the greatest extent possible [10]. The idea of dimensionality reduction of principal component analysis has provided good theoretical and technical support for the evaluation of comprehensive indicators from the beginning of its creation, and now it has been widely used in a variety of fields, such as the discrete Calhoun a Lowe transform in the field of signal processing, orthogonal decomposition and singular value decomposition in the field of mechanical engineering, as well as the eigenvalue decomposition and function decomposition in the field of linear algebra, all of which are borrowed from and morphed into the principal component analysis applications [11-13].

With the popularity of online education and open education platforms, the attention of educational researchers to online learning behaviors continues to rise [14]. Hamada et al. (2011) built an automatic tool based on the Felder-Silverman learning style model in order to personalize the classroom education, and used the learner's behavioral data in the webpage to infer their learning style cues [15]. Peng (2017) constructed an intelligent analysis model of online learning behavior from a multidimensional and multilevel perspective, analyzed the correlation between learning behavior and learning effect based on learning data, and also realized personalized course recommendation based on intelligent algorithms [16]. Zhang (2018) analyzed the online learning logs of 1,088 students, and studied the correlation between students' behaviors through a variety of methods that Revealed the factors affecting learners' learning process and results, including online learning practices, resource utilization efficiency, social interactions, etc. [17]. Yan et al. (2019) analyzed the relationship between online learning behavioral characteristics and course grades, and they found that the correlation between the number of days of online learning access, the number of clicks, and users' course grades was high, while users' ages and genders had the lowest correlation with them, and through online learning behavior data teachers can identify students with learning difficulties and provide assistance in a timely manner [18]. Matcha et al. (2019) explored the temporal and temporal characteristics of learning strategies and explored their relationship with feedback from online pre-course activity data from a flipped classroom. Clustering, sequence mining and process mining were used to detect and explain learning strategies. Inferential statistical tests were used to find a positive correlation between personalized feedback and positive correlation between effective strategies [19] rate, recall, precision, and F-value metrics, verifying its feasibility for learning style identification [20]. Zarzour et al. (2020) investigated students' use of e-books based on Facebook features based on log data]. Mehenaoui et al. (2019) proposed an online learning behavior analysis model which mines learners' behavioral data and identifies learners' learning styles in online learning environments. The study tested the model's behavioral patterns of identifying accurate methods of learning and explored the differences between high and low engagement students in their learning behaviors [21]. Li et al. (2023) encapsulated students' online learning behaviors into online learning engagement behaviors, persistence behaviors, procrastination behaviors, and absenteeism behaviors, and proposed teaching improvement suggestions based on learners' online learning behaviors in terms of promoting positive behaviors and avoiding negative behaviors to promote the development of education informatization [22].

With the increasing use of large databases, the details of which are often difficult to understand, the use of principal component analysis reduces the complexity and improves the interpretability of the data [23]. Hargreaves et al. (2015) reduced a large number of stock features to a few major influencing factors and visualized them in a perceptual map by leveraging the data downscaling capabilities of principal component analysis, a technique that can be used to identify stock features of successful stocks so that users can select the best stocks [24]. Gløersen et al. (2018) extracted the “major” motor components of athletes' multistages through principal component analysis, and found that the athletes' competitive level is related to their “major” motor components and body weight. “major” motion components and body center of gravity, and principal component analysis simplified existing motion analysis methods [25]. Omuya et al. (2021) developed a hybrid filtering model based on principal component analysis and information gain for feature selection. The hybrid model reduces the data dimensionality and allows for the selection of an appropriate set of features, which assists the machine learning technique in obtaining superior classification performance [26]. Although PCA has achieved wide application in many fields, there are fewer studies on its application to learning data in open education platforms. Relevant reports in recent years have revealed that principal component analysis also has great potential for application in the field of education. Zhang et al. (2019) used principal component analysis to downscale multiple measures of learning behavior data in MOOCs, and finally obtained three principal component factors, and then the principal component logistic regression model successfully predicted the course pass rate of learners [27]. Hershcovits et al. (2019) mined the student data that changed continuously over time from an independent learning system and used principal component analysis to measure it so as to construct the trajectory of user activities, and divided the student groups by the way the trajectory changed with the practice to study the student's participation in the independent learning system [28]. Wang's (2025) study

used data mining techniques to obtain learners' learning behavior data from open education platforms and reduced the dimensionality of the data through principal component analysis to improve the computational efficiency and identification of the data, this analysis helps educators to provide better personalized learning experiences for students [29].

In addition, Principal Component Analysis (PCA) still has some significant limitations in practice, such as the inability to effectively deal with nonlinear data or the presence of outliers [30]. For this reason, De La Torre et al. (2003) used the Geman-McClure loss function in their study, which replaces the L2 paradigm used in PCA in the field of robust statistics, and through this alternative, the sensitivity of the model to strong noise and outliers can be significantly reduced [31]. Kwak (2013) set up an objective function determined by an arbitrary p-value of the Lp paradigm and computed the gradient of the objective function, the proposed method is easy to compute and can find the local optimal solution, and the performance of the optimized principal component analysis method has been significantly improved [32]. Li et al. (2021) introduced an ℓ2,p -paradigm regularization term in principal component analysis, the projection matrix became sparse, and then the learned rows of the sparse and orthogonal projection matrix were used for selecting features with discriminative features, which improves the convergence of the algorithm and reduces the computational complexity [33].

Aiming at the problem of large sample size and high data dimensionality in the open education platform, this paper proposes to comprehensively analyze the collected data through principal component analysis and complete the dimensionality reduction of the ponderous data under the premise of ensuring no loss of variable information. Then, the K-means algorithm is applied to the processed learning state data set, and the learners are aggregated by the clustering algorithm according to their interactive learning characteristics, and the number of clusters is determined by the aggregation coefficient method, and the learners will be aggregated into different classes according to their activity level in the learning platform. At the same time, through the clustering results of the proposed targeted teaching enhancement strategy, early detection and intervention of online learners' learning risks, for the teaching staff to provide reasonable help in advance, and to improve the online learning effect and quality.

2. Underlying conceptual and technical foundations

2.1. Basic concepts

2.1.1. Open Education Platform

MOOC, Massive Open Online Course. Among them, the first letter “M”, that is, Massive; compared with the traditional classroom lecture system with only a few dozen students, MOOC courses can easily involve tens of thousands of participants, which are included in the university is global, and the amount of course data contained is huge. The second letter “O” is Online, the age of information networks, without leaving home, you can enjoy learning resources through the Internet at any time, without the limitations of time and space. The third letter “O” is Open, which promotes the open sharing of resources, as long as a registered account can share MOOC courses. The fourth letter “C” is Course, which refers to the course resources uploaded to MOOC, which is a large-scale open course on the Internet, and it is an open course distributed on the Internet for the purpose of enhancing the dissemination of knowledge and released by individuals or organizations with the spirit of sharing and collaboration. In the Internet era, MOOC's well illustrate the core concepts of connection and sharing [34].

After the rise of the MOOC trend, well-known universities have joined mainstream online platforms to launch more open education platforms. For example, in 2013, Tsinghua University joined the edX platform, and in the same year, it created the “Xuedang Online” platform, which provides online courses for global learners, and the platform is constantly undergoing technological improvements, and is committed to creating the best Chinese MOOC platform in the world. In addition, in addition to Tsinghua University, Fudan University, Shanghai Jiao Tong University and National Taiwan University have also joined Coursera and launched their own high-quality online courses. In addition to China's well-known universities to join MOOC, network operating companies also follow closely, taking advantage of the business opportunities of MOOC and China's vast market. NetEase Open Class was the first to join online education in China and fully cooperated with Coursera in 2013. NetEase provides video hosting service for Coursera, so that Chinese users can watch the courses directly on NetEase Open Class. NetEase also opened a Chinese learning community for Coursera to help Chinese learners eliminate language barriers.

In the era of big data, when MOOC online education is sweeping the world, Chinese universities and online companies have shown their sensitivity. When MOOC was in the ascendant, top institutions of higher education in different countries (regions) joined in and brought high-quality educational resources, so MOOC was labeled as “high class” when it first entered China. Therefore, MOOCs were already

labeled as “high class” when they first entered China. Moreover, in China nowadays, the concept of lifelong learning has been accepted by the public, and the Internet has become an indispensable part of people's life, study and work in the era of big data, and the popularization of China's network has also provided a new platform for the emergence and development of open education platforms.

2.1.2. Pedagogical state space

Teaching state space is a conceptual framework that integrates mathematical concepts, which abstracts the state of a course at a certain point in time into a multidimensional space composed of multiple key variables. Traditional classroom teaching is limited by the teacher's limited energy and too many students, which makes the information transfer and feedback between the teacher and the students have certain limitations, and the teacher can not timely and accurately grasp each student's mood, attention and learning behavior changes. Traditional classroom teaching focuses on the content of the teacher's teaching and ignores the information of students' emotional changes. In the classroom, students' emotional changes to a certain extent reflect the overall classroom status and satisfaction with the quality of teaching, as well as student acceptance of the classroom, which in turn provides feedback on the overall quality of teaching. However, for the present time when open education platforms are flourishing, these dynamically changing student state variables (cognitive state, behavioral state, affective state, etc.) and environmental state variables (resource state, activity state, interaction state, etc.) in the classroom can be captured. Therefore, classroom teaching state mining based on open education platforms has become a popular research direction.

On the basis of existing research and methods, it is found that although the existing research methods have achieved good research results in their respective research tasks, it is difficult to meet the demand of classroom state mining in complex large-scale open classroom scenarios, and expression feature learning, expression recognition and classroom state space applicable to classroom scenarios need further research. Therefore, the main task of this project is to downscale the teaching state space in the open education platform, aiming to provide a theoretical basis for the subsequent cluster analysis and the development of teaching enhancement strategies.

2.2. Technical basis

2.2.1. Principal Component Analysis

In order to balance the advantages and disadvantages of multivariate and large samples, the number of variables should be reduced as much as possible without losing the information of the variables when comprehensively analyzing the collected data. However, the correlation between variables cannot be lifted singly from time to time, so it is necessary to synthesize all kinds of information existing in each variable with a comprehensive index less than the number of variables in the original collected data to complete the dimensionality reduction of the complicated data, and Principal Component Analysis (PCA) belongs to the method of this dimensionality reduction.

The specific implementation steps of principal component analysis are as follows:

Step 1, Assume that from m n -dimensional data, the original data is arranged into a matrix X of n rows and m columns, i.e.:

$$X = [x_1, x_2, x_3, ..., x_n]^T \tag{1}$$

where $x_i (i = 1, 2, ..., m)$ is the row vector of $1 \times m$.

Step 2, zero-mean x_i , i.e., each value in the vector is subtracted from the mean of this row to find the covariance matrix C :

$$C = \begin{pmatrix} \operatorname{cov}(x_1, x_1) & \cdots & \operatorname{cov}(x_1, x_n) \\ \vdots & \ddots & \vdots \\ \operatorname{cov}(x_n, x_1) & \cdots & \operatorname{cov}(x_n, x_n) \end{pmatrix} \tag{2}$$

where the covariance of the two vectors can be expressed as:

$$\operatorname{cov}(x_i, x_k) = E[(x_i - \bar{x}_l)(x_k - \bar{x}_k)] \tag{3}$$

Step 3, since the covariance matrix C is a real symmetric matrix whose eigenvectors corresponding to different eigenvalues must be orthogonal and can be diagonalized, the eigenvalues λ of the

covariance matrix C and its corresponding eigenvectors q can be found by performing a diagonalization computation of C, that is:

$$Q^T CQ = \Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \tag{4}$$

where $Q = [q_1, q_2, ..., q_n]$ is the matrix consisting of the eigenvectors corresponding to the eigenvalues. Where λ_1 to λ_n, ordered from largest to smallest, i.e., $\lambda_1 > \lambda_2 > ... > \lambda_n$.

Step 4, take the matrix composed of the eigenvectors, take the first j eigenvectors to form the projection matrix P according to the need, and use the P matrix to obtain the projection of X on it Y, where Y is the original matrix X after downscaling it to the j dimension:

$$Y = PX \tag{5}$$

2.2.2. Cluster analysis

Clustering is the process of dividing a collection of data objects into clusters composed of multiple groups of similar samples, a cluster is a collection of data objects, which can be divided by clustering so that objects in the same cluster are similar to each other, and objects in different clusters are different from each other. Cluster analysis is the basis for further processing data and analyzing data, and is often applied to various fields as a means of data preprocessing. In this study, it is proposed to use K-means clustering algorithm to divide the learners according to the degree of active learning state, so that the classification results in the teaching state space will be more accurate.

The specific implementation steps of K-means clustering algorithm are as follows [35]:

Let $X = \{x_i\}$, $i = 1, 2, ..., n$ is the set of points of dimension n clustered into k clusters $C = \{c_k, k = 1, 2, ..., k\}$. The K-means algorithm finds a partition that minimizes the squared error between the empirical mean of the cluster and the points in the cluster. Let μ_k be the mean of the cluster c_k, then the squared error between μ_k and the other points in the cluster will be defined as:

$$J(c_k) = \sum_{x_i \in c_k} \|x_i - \mu_k\|^2 \tag{6}$$

The goal of the K-means algorithm is to minimize the sum of the squared errors of all k clusters, i.e.:

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in c_k} \|x_i - \mu_k\|^2 \tag{7}$$

Minimizing this objective function is known to be an NP-hard problem, and K-means is a greedy algorithm that can only converge to a locally optimal solution.The K-means algorithm starts clustering into k classes with an initial partition and assigns patterns to the clusters in order to reduce the squared error. Since the squared error always decreases as the number of clusters k increases, it can only be minimized for a fixed number of clusters.The main steps of the K-means algorithm are as follows:

(1) Randomly select k data objects from the dataset as the initial clustering centers;

(2) Assign each data object to the most “similar” cluster;

(3) Recalculate the centers of the clusters;

(4) Determine whether the objective function converges, if so, the algorithm ends; otherwise, return to the second step.

3. Teaching state space construction and research design

3.1. Teaching state feature selection

3.1.1. Classification of teaching status

Driven by the Internet and computer technology, open education platforms have become an increasingly popular means of learning. Compared with the learning behaviors generated in the traditional classroom, open education platforms have their outstanding features, which are not bound by

space, more abundant in learning resources, and more intelligent and efficient in learning. More crucially, the behavior generated by online open education in the interaction between learners and the Internet is easy to collect, mine and analyze, which can provide data and theoretical basis for the improvement of the level of open education in the future.

The teaching state can be mainly categorized as follows:

(1) Watching course videos

Watching course videos is the main means for learners to learn in MOOC, and through the learning of video resources, learners can learn the content of the course intuitively by following the instructor.

(2) Submit assignments and quizzes

In the learning platform, as the course progresses, the teacher will assign some homework for learners to answer. Assignments can not only check the learning effect of learners in time, but also broaden their horizons and make them summarize and reflect. With the increase of learning content, teachers hope to understand the learning effect of learners through tests and exams, and at the same time give students the opportunity to check and fill in the gaps through tests and exams.

(3) Participate in discussion forums

After learning the course video, learners may still have doubts about certain issues of the course, and then they can go to the corresponding discussion forum to search for problems and see if there are other learners who have encountered the same problems. You can also post your own questions to seek help from other learners, or answer questions raised by other learners in the forum to help everyone make progress together.

(4) Other learning status

Learners in the MOOC course, in addition to video resources for learning, the teaching staff will also provide learners with other forms of learning resources, such as courseware, bibliography, sample code and other content, which can be used as a supplement to video learning.

3.1.2. Feature selection

In this chapter's study based on MOOC education data, in order to analyze and predict learners more accurately, 11 features that can represent the characteristics of their education status are selected from the behavioral logs produced by learners during the MOOC learning process, and the specific names and descriptions are shown in Table 1. Among them, the 11 state features characterize the learners' state characteristics from different dimensions such as the number of activities, the situation of watching videos, the situation of assignments, and the situation of forum discussions.

Table 1. Selection of Educational Status Characteristics.

Index	Name	Explanation
X1	Number of visits to the course	The number of times one visits the course during the MOOC learning process
X2	Duration of the visiting course	The duration of accessing courses during the MOOC learning process
X3	Visit other modules of the course	The number of times other modules of the course are accessed during the MOOC learning process
X4	The number of visits to the assignment	The number of times one accesses assignments during the MOOC learning process
X5	Number of visits to the forum	The number of times the corresponding forum of the MOOC learning course is visited
X6	Number of visits to wiki	The number of visits to Wikipedia during the MOOC learning process
X7	Total number of activities	The total number of learning activities during the MOOC learning process
X8	Active days	There are days of activities during the MOOC learning process

X9	The number of times the web page is closed	The number of times a web page is closed during the MOOC learning process
X10	Number of video views	The number of times one watches teaching videos during the MOOC learning process
X11	Video viewing time	The number of times web pages are closed during the MOOC learning process

3.2. Study design

The teaching state mentioned in the study mainly refers to the state of learners' participation in e-learning recorded in the database on the edX open education platform, which mainly includes: the number of times of accessing the course (X1), the time of accessing the course (X2), and the access to other modules of the course (X3). The clustering research process of teaching state space is shown in Figure 1:

(1) The study first collects and preprocesses data on the open dataset, and divides the dataset into four modules: course information, basic learner information, learning behavior information and learning outcome information.

(2) Since the study chose to use the level of teaching status for clustering, the learner behavioral input factors were extracted here using PCA with factor analysis. PCA was used to determine the initial factor loadings and factor analysis was used to determine the extracted factors.

(3) The final scores of the learning status level factors were selected based on the purpose of clustering and used to make the variables in the cluster analysis, using hierarchical clustering to determine the k-value and then K-means for cluster analysis.

(4) Finally, an exploration of the factors influencing the learning status of different groups of learners was conducted.

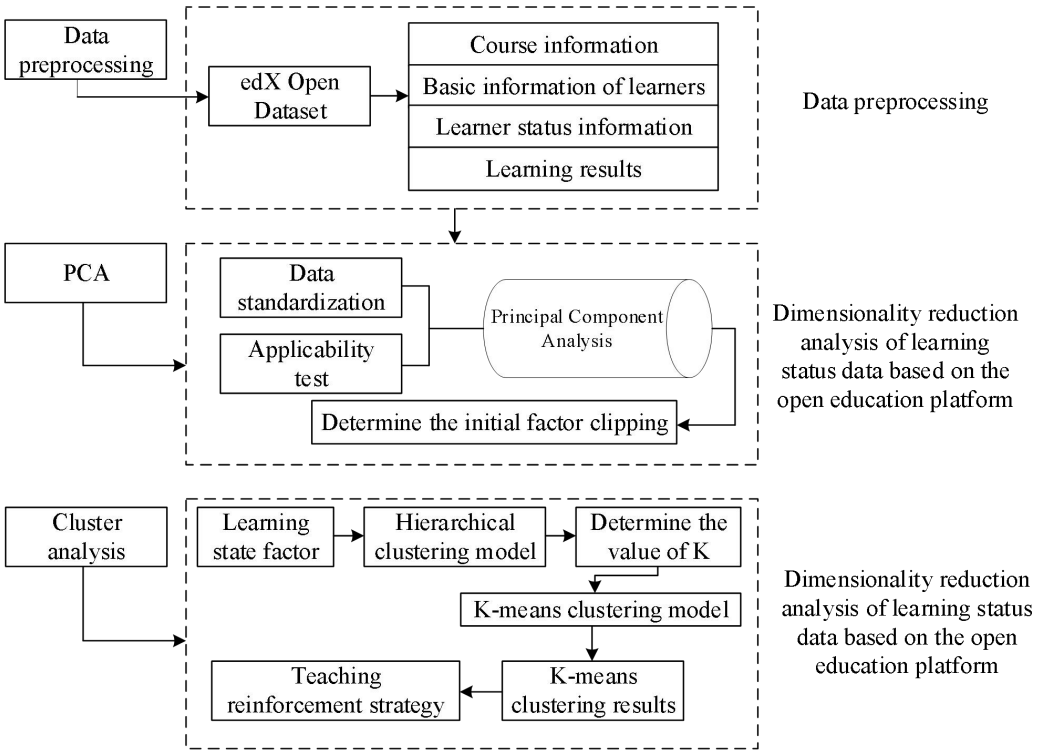

Figure 1. Research process of clustering in the teaching state space.

3.3. Data sets and pre-processing

3.3.1. edX dataset

The selection of research objects is one of the key aspects of conducting data mining. As a non-profit open education learning platform, edX has been playing an important role in promoting data openness and education and teaching research. The platform has a teaching status data set of 16 courses and more than 600,000 participants, with a large number of detailed information and statistical data, rich learning objects, and a stable learning process. The social science class has the highest participation rate, and the humanities and social sciences class has the lowest participation rate, so the dataset selected in this paper is the social science courses on the edX open education platform.

The open data source comes from the requirements of scientific research and academic innovation, and the data selected for the study comes from the dataset released by edX in 2019, which contains the data of five courses offered by Tsinghua University on the edX platform in three semesters. Considering that the data generated by students enrolled in open courses in the edX platform are protected by the Family Educational Rights and Privacy Act for the privacy of student records, the dataset has been de-identified. Open data from learners on the edX Learning Platform are desensitized by the platform's system to protect learner privacy information through a series of data processing. The processed dataset is relatively reduced, and the statistics of some data items may be affected, but there is little difference in the overall data analysis.

3.3.2. Data pre-processing

In this data, the incomplete_flag column data is processed first, a value of 1 means that the data is internally inconsistent and the data is less reliable, so this part of the data is chosen to be deleted. The learner's performance data is normalized with a range from 0-1, and learners exceeding this range are classified as abnormal data for outlier processing. And in the dataset, some of the data values were NA, which indicated that the student had created an edX account before the corresponding student registration question, so the values were populated based on the previous account. In addition, some of the duplicates in the data were removed using data processing techniques.

Since there are still a large number of missing values in the learner data in this data, the forward filling method was chosen to remove the overall learner data that still had missing values after filling, and the processing and deletion of duplicate value data was carried out, and the final test yielded 52,327 status data for 835 learners.

4. Analysis of PCA downscaling and clustering results

4.1. PCA downscaling analysis

Through data preprocessing, the streamlined 52,327 data were obtained, but due to the large number of data dimensions and the fact that most of the learning states could not be quantified exactly, this paper is based on PCA and factor analysis methods for dimensionality reduction, and cluster analysis of learners with superior performance than the original variables when the loss of information reaches the minimum.

(1) Data standardization and applicability test

The software used for principal component analysis in this paper is SPSS version 21, and the original data were subjected to Z-score (standardization) to generate processed data. The standardized data were subjected to KMO and Bartlett's spherical test, i.e. factor analysis applicability test. The test results are shown in Table 2, which shows that the KMO value is equal to 0.805>0.6, and the sample data have quantitative correlation between the indicators, which is suitable for factor analysis. The approximate chi-square value of Bartlett's spherical test is 152053.27 (significance level p is 0.0001<0.001), and the result rejects the null hypothesis that there is a certain correlation between the variables of the sample data, which is suitable for PCA analysis.

Table 2. Results of the Applicability Test of Factor Analysis.

KMO and Bartlett tests		
The measurement of the suitability of KMO sampling		0.805
Bartlett sphericity test	Approximate chi-square	152053.27
	Degree of freedom	31
	Significance	0.0001

(2) Principal component analysis

Through the gravel diagram can determine the number of principal factors, gravel diagram can show each component feature root "steep slope potential energy" change process, the larger the magnitude of the difference indicates that its corresponding component is more important, the stronger the ability to explain, generally the first few features fall in the magnitude of the difference is larger, and the more the more after the more gentle. The results of the gravel map of the common factor characteristics are shown in Figure 2, in which the component number of the coordinate represents the number of factors, and the vertical coordinate is the characteristic value of the factor, and generally the factor is selected to have an eigenvalue greater than 1 and the steeper part of the gravel map. As can be seen from the figure, the first two factors have larger eigenvalues (both greater than 1) and steeper connecting lines, which can be determined as the main factor.

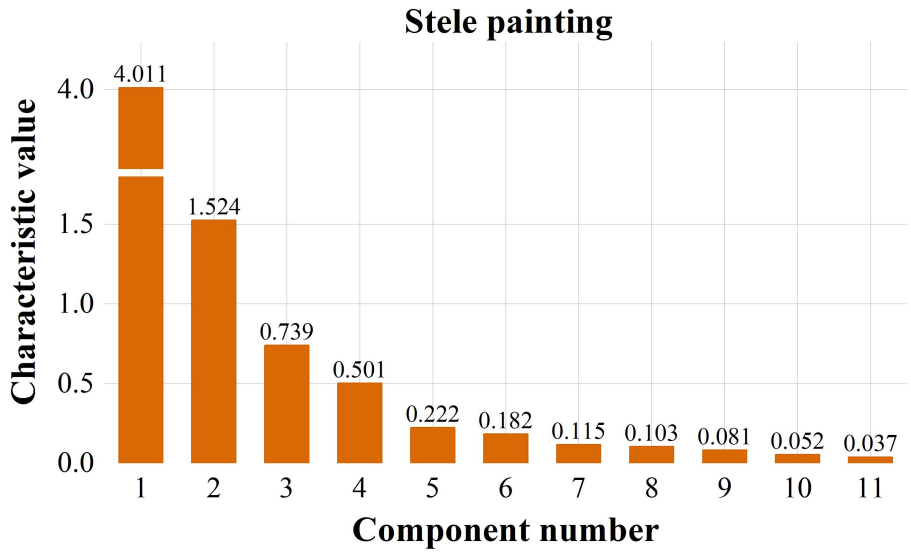

Figure 2. Common factor feature lithotripsy map.

The main purpose of principal component analysis is for indicator dimensionality reduction, but at the same time, it is also important to ensure that the loss of information is minimized as much as possible. Therefore, on top of using the gravel plot to determine the basis of the principal factors, the cumulative variance contribution rate of 80% is also required under ideal conditions. However, this is not a strict criterion, and in the actual practice of the data, a cumulative variance value of 70% or more is sufficient.

The total variance interpretation is shown in Table 3, the first two principal components explain 73.15% of the total variance, and 73.15% of the original 11 indicators of teaching status can be represented in the two principal components extracted. The first principal component explains 53.01% of the information of the teaching status in the 11 indicators, and the second principal component explains 20.14% of the information. The weights of the two principal components are 53.01/73.15 = 72.47% and 20.14/73.15 = 27.53%. Therefore, it was finally determined that two principal components could be extracted, which were set as Z1 and Z2.

Table 3. Results of the Applicability Test of Factor Analysis.

	Initial eigenvalue			Extract the sum of squares of the loads		
1	Total	%	Cumulative %	Total	%	Cumulative %
2	4.011	53.01	53.01	4.011	53.01	53.01
3	1.524	20.14	73.15	1.524	20.14	73.15
4	0.739	9.77	82.91			
5	0.501	6.62	89.53			
6	0.222	2.93	92.47			
7	0.182	2.41	94.87			

8	0.115	1.52	96.39			
9	0.103	1.36	97.75			
10	0.081	1.07	98.82			
11	0.052	0.69	99.51			

(3) Principal component naming

In this paper, based on the two initial loadings obtained from PCA, the parameters in the two public factors Z1 and Z2 were further determined using factor analysis and these two public factors were extracted.

From the factor analysis model, the first common factor consists of the number of visits to the course (X1), the length of visits to the course (X2), the number of visits to assignments (X3), the number of visits to assignments (X4), the total number of activities (X7), the number of times of watching the video (X10), and the time of watching the video (X11), which are a total of seven state features. These seven state features all reflect the number of learner behavioral events, and the more learning events there are, the higher the learner's learning state usually is. Here, the learning behavior events in the first public factor are used to characterize the learner's learning state level, i.e., Z1 is named as the student learning state factor.

The second common factor consists of four state characteristics, namely, the number of visits to the forum (X5), the number of visits to the wiki (X6), the number of days of activity (X8), and the number of times the webpage was closed (X9), which mainly represent information about the educational environment of the learner before he/she enters the course or after he/she finishes the course. Therefore, the second common factor Z2 is named as the educational environment state factor.

4.2. Results of K-mean cluster analysis

In this study, the data records left by learners in the edX open education platform are studied by clustering learners into different groups using the method of K-mean cluster analysis. That is, the data indicators are compared and analyzed, and learners are clustered into one class if their behavioral characteristics are similar, and vice versa, they are divided into different classes. In this case, the hierarchical clustering method was chosen to determine the number of clusters, and then the K-Means algorithm was used to cluster the learners.

This study uses SPSS to complete the hierarchical clustering, the specific clustering method selected intergroup linkage method, the distance measurement interval is selected as the square distance measurement interval, and the "number of categories" as the horizontal coordinate, "aggregation coefficient" as the vertical coordinate to draw a line graph, specifically as shown in Figure 3. As shown in Figure 3. Analysis shows that the folding line tends to slow down when the number of categories is 4~7, so the number of clustering categories of the teaching state space is considered to be taken in this interval, and after many attempts, the number of categories of the teaching state space in this study is finally set at 4.

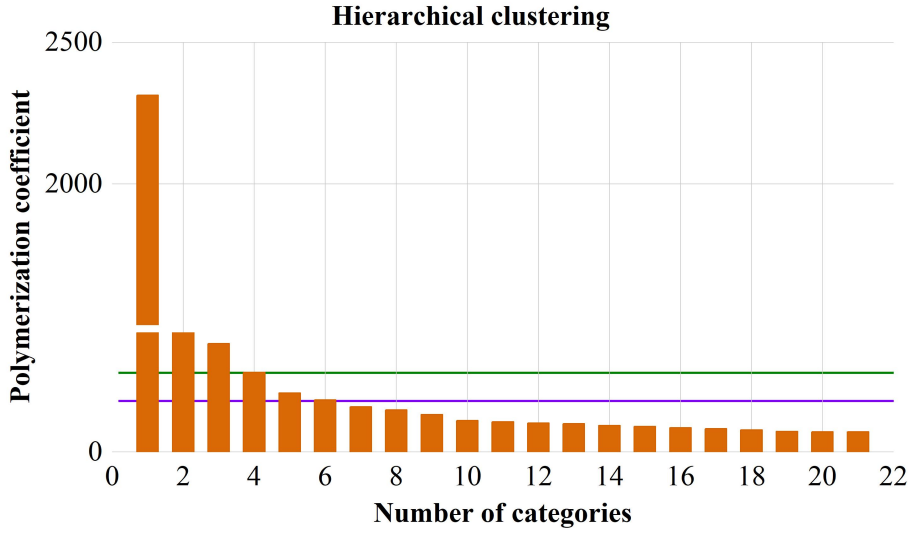

Figure 3. Hierarchical clustering results.

According to the specific idea of K-Means algorithm, firstly, N learners from 835 learners in edX dataset are arbitrarily selected as the initial center node of the clustering, and then the distance between the remaining samples and the center node is repeatedly counted, and then based on this, the corresponding sample object is divided, and finally, if it meets the corresponding conditions, then the calculation is stopped. In this study, SPSS was chosen to cluster all the learning behavior samples, and the visualization results of K-Means clustering are shown in Figure 4.The number of categories of K-Means algorithm can be established as 4, and the results show that 835 learners are distributed in Cluster I as 182, Cluster II accounts for 334, Cluster III accounts for 241, and Cluster IV accounts for 78.

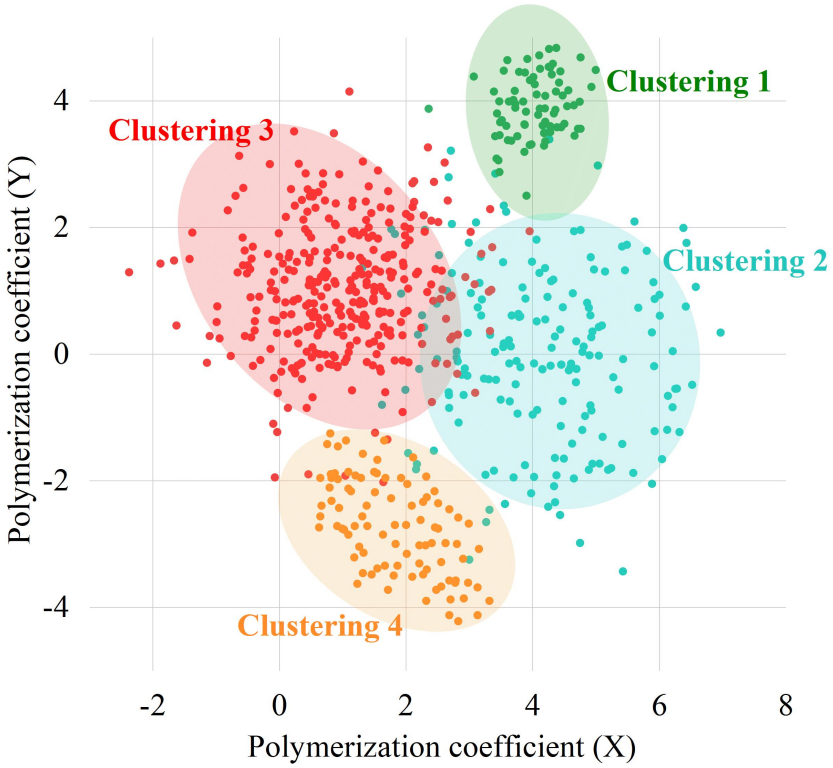

Figure 4. K-Means clustering visualization results.

In order to confirm the reasonableness of the clustering results, the quality of clustering needs to be tested, i.e., to confirm the existence of significant differences among the four categories of learning groups in each behavioral indicator. The analysis of each behavioral variable was found to be consistent with normal distribution, so the study was grouped according to the categories of the learning groups, and the results of one-way ANOVA were shown in Table 4. It can be seen that the asymptotic significance of the four categories of learning groups in each behavioral indicator is less than 0.05, that is, there is a significant difference, indicating that the clustering quality is good. In addition to this, according to the size of the resultant F-value, it is also possible to determine which category of factors contributes the most to the clustering. The analysis shows that the length of access to the course (X2, F=138.304) is a significant influence on the spatial clustering of learners' learning status, followed by the number of access to assignments (X4, F=104.382).

Table 4. Results of one-way Analysis of Variance.

Status indicator	Clustering		Error		F	Sig.
	Equal square	DF	Equal square	DF		
X1	0.971	3	0.914	834	4.285	0.0270
X2	0.819	3	0.786	834	138.304	0.0004
X3	0.712	3	0.915	834	28.442	0.0009
X4	0.799	3	0.862	834	104.382	0.0013
X5	0.532	3	0.533	834	13.741	0.0024

X6	0.492	3	0.108	834	0.992	0.0272
X7	0.391	3	0.241	834	33.350	0.0008
X8	0.784	3	0.640	834	72.458	0.0003
X9	0.735	3	0.451	834	93.487	0.0001
X10	0.228	3	0.287	834	29.33	0.0001
X11	0.803	3	0.771	834	78.212	0.0004

Through cluster analysis, learners can be clustered into four categories by the learning status indicators extracted from the edX open education platform, and the results show that there are obvious differences between each category of learning groups. And each category of learners is defined as follows:

(1) Learners in category 1 perform well in general, have a longer access time to the course, and their interactive performance in the forum is comparable to that of category 4, but the number of active days in this category is low, so this category is defined as “general learners”.

(2) The overall performance of the learning group in category 2 is poor, and the observation of the length of accessing courses, watching videos and time are all lower, which is specifically analyzed to be due to the lower average test length of this group. Moreover, the procrastination of this group is high, so this group is defined as “negative learners”.

(3) The overall performance of Category 3 learners is comparable to that of Category 4, with a good number of hours and visits to the course, and more active completion of tasks. However, in terms of interaction, this group achieved the highest discussion results and actively replied to posts in the forum, so this group was defined as “interactive learners”.

(4) Learners in category 4 achieved excellent results in all teaching states, had the highest efficacy in homework tests among the four categories, participated more actively in topic interactions, and had low procrastination in completing tasks, thus defining this group as “active learners”.

4.3. Instructional Reinforcement Strategies

For the four types of learners obtained from the cluster analysis in the open education platform, the following teaching enhancement strategies are proposed:

(1) The proportion of positive emotions of “active learners” is the highest among the four types of learners, indicating their affirmation and support for open courses. Combined with the analysis of the online learning behavior and topics of interest of this type of learners, it is found that they are highly motivated to learn and also express their affirmation of the teacher's lecturing methods in the forum area, but there are still a small amount of neutral emotions in this type of learners. However, this group of learners still has a small amount of neutral emotions. Therefore, teachers can guide students to maintain their positive emotions, but they also need to combine work and rest and pay attention to learning styles.

(2) Although “negative learners” have negative emotions, the proportion of their positive emotions still reaches 60%, indicating that this group still has a certain degree of enthusiasm for learning. Analyzing the online learning attitudes of this group of learners, it is found that their learning time is short and their procrastination is high. Therefore, teachers should pay more attention to this group of learners, and because of their low grades, teachers can adjust the teaching progress according to their concerns about the course content to realize the tailored teaching for this group of learners.

(3) The emotional distribution of “interactive learners” is similar to that of “active learners”, and the analysis of their online learning behaviors reveals that this group of learners has the highest frequency of posting in the MOOC comment area and discussion forum. Therefore, teachers can take corresponding measures, such as liking their posts or guiding them according to the content of their posts, to give full play to the activity of this group.

(4) Although “general learners” do not have any negative emotions, the proportion of their positive emotions is the lowest among the four groups, while the proportion of neutral emotions is the highest. Analyzing their learning attitudes on the open education platform, we can see that this group has lower test scores and longer delays in completing tasks, but they indicated in the comment section that the course content is easy to understand and learn, which indicates that they are highly malleable. Therefore, teachers should pay attention to this type of learning group and provide more guidance to promote the positive transformation of neutral emotions, thus mobilizing learning motivation.

5. Conclusion

In this paper, from the perspective of the large amount of learning status data generated in the learning process of open education platform, for the problem of poor learning effect of open education platform, through the data analysis means such as principal component analysis and cluster analysis, the learning status data are mined and analyzed, and the results of the obtained cluster analysis are used to put forward targeted teaching and learning enhancement strategies The research in this paper is not only important for perfecting the analysis of the learning status, the construction and the development of in open education platform plays an important role, and at the same time, through the monitoring of learners' learning status, the learning risk of online learners can be found in advance, for the teaching staff to provide reasonable help in advance, and to enhance the online learning effect and quality. The main contents of this paper are summarized as follows:

On the basis of the preprocessing of learners' learning status data in edX open education platform, PCA dimensionality reduction technology is used to reduce the dimensionality of the 11 learning status features extracted. Then, the K-means algorithm was used to cluster the learners, and the number of clusters was experimented from 2 to 5 respectively, and the profile coefficient method was used to make comparisons, and the results showed that the coefficient of aggregation reached the peak when the number of clusters was 4, that is, the choice of clustering the learners into 4 classes. Specifically, active learners, negative learners, interactive learners and active learners. Finally, targeted instructional reinforcement strategies are proposed for the clustering results.

However, as far as the definition of behavioral indicators is divided. Although this study refers to the state analysis division indexes in the existing university MOOC research when constructing the teaching state space, and also combines the real learning behaviors of learners in the edX open education platform, the final state indexes are still to be considered, and the learning groups divided by richer teaching state indexes are more representative. For this reason, subsequent research will expand and extend the methodology of this paper, carry out real-time collection and analysis of other dimensions of classroom performance data, and try to combine with the learning status data to carry out multi-dimensional, multi-scale comprehensive research and judgment, in order to make full use of the advantages of a variety of cutting-edge intelligent technologies for the deep mining of teaching behaviors and measurement and diagnosis, to accelerate the construction of smart classrooms, smart schools, and to help the realization of lifelong digital education. The following is a summary of the results of the research and evaluation.

References

1. Chang, V. (2016). Review and discussion: E-learning for academia and industry. International Journal of Information Management, 36(3), 476-485.
2. Castillo, N. M., Lee, J., Wagner, D. A., & Zahra, F. T. (2015). MOOCS for development: Trends, challenges, and opportunities. International Technologies & International Development, 11(6), 35-42.
3. Ang, K. L. M., Ge, F. L., & Seng, K. P. (2020). Big educational data & analytics: Survey, architecture and challenges. IEEE access, 8, 116392-116414.
4. Munshi, A. A., & Alhindi, A. (2021). Big data platform for educational analytics. Ieee Access, 9, 52883-52890.
5. Veeramanickam, M. R. M., & Ramesh, P. (2022). Analysis on quality of learning in e-Learning platforms. Advances in Engineering Software, 172, 103168.
6. Li, K. C. (2018). The Evolution of Open Learning: A Review of the Transition from Pre-e-Learning to the Era of e-Learning. Knowledge Management & E-Learning, 10(4), 408-425.
7. Nazempour, R., & Darabi, H. (2023). Personalized learning in virtual learning environments using students' behavior analysis. Education Sciences, 13(5), 457.
8. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.
9. David, C. C., & Jacobs, D. J. (2013). Principal component analysis: a method for determining the essential dynamics of proteins. In Protein dynamics: Methods and protocols (pp. 193-226). Totowa, NJ: Humana Press.
10. Journée, M., Nesterov, Y., Richtárik, P., & Sepulchre, R. (2010). Generalized power method for sparse principal component analysis. Journal of Machine Learning Research, 11(2).
11. Kong, X., Hu, C., & Duan, Z. (2017). Generalized principal component analysis. In Principal Component Analysis Networks and Algorithms (pp. 185-233). Singapore: Springer Singapore.
12. Elsamanty, M., Ibrahim, A., & Salman, W. S. (2023). Principal component analysis approach for detecting faults in rotary machines based on vibrational and electrical fused data. Mechanical systems and signal processing, 200, 110559.
13. Shang, H. L. (2014). A survey of functional principal component analysis. AStA Advances in Statistical Analysis, 98(2), 121-142.

14. Panigrahi, R., Srivastava, P. R., & Sharma, D. (2018). Online learning: Adoption, continuance, and learning outcome—A review of literature. International Journal of Information Management, 43, 1-14.
15. Hamada, A. K., Rashad, M. Z., & Darwesh, M. G. (2011). Behavior analysis in a learning environment to identify the suitable learning style. International Journal of Computer Science and Information Technology, 3(2), 48-59.
16. Peng, W. (2017). Research on online learning behavior analysis model in big data environment. Eurasia Journal of Mathematics, Science and Technology Education, 13(8), 5675-5684.
17. Zhang, J. H., Zhang, Y. X., Zou, Q., & Huang, S. (2018). What learning analytics tells us: Group behavior analysis and individual learning diagnosis based on long-term and large-scale data. Journal of Educational Technology & Society, 21(2), 245-258.
18. Yan, N., & Au, O. T. S. (2019). Online learning behavior analysis based on machine learning. Asian association of open universities journal, 14(2), 97-106.
19. Matcha, W., Gašević, D., Uzir, N. A. A., Jovanović, J., & Pardo, A. (2019, March). Analytics of learning strategies: Associations with academic performance and feedback. In Proceedings of the 9th international conference on learning analytics & knowledge (pp. 461-470).
20. Mehenaoui, Z., Lafifi, Y., & Zemmouri, L. (2022). Learning behavior analysis to identify learner's learning style based on machine learning techniques. Journal of Universal Computer Science, 28(11), 1193.
21. Zarzour, H., Bendjaballah, S., & Harirche, H. (2020). Exploring the behavioral patterns of students learning with a Facebook-based e-book approach. Computers & Education, 156, 103957.
22. Li, Z., & Liu, Y. (2023). Analysis of the current situation of the research on the influencing factors of online learning behavior and suggestions for teaching improvement. Sustainability, 15(3), 2119.
23. Hasan, B. M. S., & Abdulazeez, A. M. (2021). A review of principal component analysis algorithm for dimensionality reduction. Journal of Soft Computing and Data Mining, 2(1), 20-30.
24. Hargreaves, C. A., & Mani, C. K. (2015). The Selection of winning stocks using principal component analysis. American Journal of Marketing Research, 1(3), 183-188.
25. Gløersen, Ø., Myklebust, H., Hallén, J., & Federolf, P. (2018). Technique analysis in elite athletes using principal component analysis. Journal of sports sciences, 36(2), 229-237.
26. Omuya, E. O., Okeyo, G. O., & Kimwele, M. W. (2021). Feature selection for classification using principal component analysis and information gain. Expert Systems with Applications, 174, 114765.
27. Zhang, W., Qin, S., Yi, B., & Tian, P. (2019). Study on learning effect prediction models based on principal component analysis in MOOCs. Cluster Computing, 22(Suppl 6), 15347-15356.
28. Hershcovits, H., Vilenchik, D., & Gal, K. (2019). Modeling engagement in self-directed learning systems using principal component analysis. IEEE Transactions on Learning Technologies, 13(1), 164-171.
29. Wang, S. (2025, January). Analysis of Learning Behaviour of Open Education Learners based on Principal Component Analysis and K-means Clustering Algorithm. In 2025 International Conference on Intelligent Systems and Computational Networks (ICISCN) (pp. 1-7). IEEE.
30. Tang, E. (2021). Quantum principal component analysis only achieves an exponential speedup because of its state preparation assumptions. Physical Review Letters, 127(6), 060503.
31. De La Torre, F., & Black, M. J. (2003). A framework for robust subspace learning. International Journal of Computer Vision, 54(1), 117-142.
32. Kwak, N. (2013). Principal component analysis by Lp-norm maximization. IEEE Transactions on Cybernetics, 44(5), 594-609.
33. Li, Z., Nie, F., Bian, J., Wu, D., & Li, X. (2021). Sparse PCA via ℓ2,p-Norm Regularization for Unsupervised Feature Selection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(4), 5322-5328.
34. Babaeva M A. (2019). Online course (MOOC) "Concepts of Modern Natural Science" on the National Platform of Open Education: experience in teaching students. Journal of Physics: Conference Series, 1348, 012003 -012003.
35. Priyambada Satrio Adi, Er Mahendrawathi, Yahya Bernardo Nugroho & Usagawa Tsuyoshi. (2021). Profile-Based Cluster Evolution Analysis: Identification of Migration Patterns for Understanding Student Learning Behavior. IEEE ACCESS, 9, 101718-101728.