Reinventing Educational Data Mining with Collaborative CAD and Clustering Technology

Kaihe Yu[1*]

[1] School of Marxism, Zhejiang Yuexiu University, Shaoxing Zhejiang, 311201, China

Corresponding Author: Kaihe Yu, yukaihe888@126.com

Abstract. In the era of big data information, traditional student behavior management has increasingly revealed the shortcomings of delayed intervention and "post-positioning" governance models. Nowadays, by applying educational big data in the analysis and monitoring of students' daily behavior, managers can actively grasp the characteristics and laws of student behavior and make research judgments accordingly. With the development of college information management systems (such as student card systems), it has become easier and more convenient for people to collect and analyze students' behavioral data. Extracting useful features from these behavioral patterns is helpful in understanding the learning process of students and also reflects an important factor in the way of studying and living habits in school. Based on the study of the fuzzy C-means clustering algorithm based on the kernel function (KFCM), this paper proposes a fuzzy C-means clustering algorithm based on the hybrid kernel function (CKFCM). On this basis, this paper proposes a feature modeling method based on student behavior sequences. Based on the careful consideration of students' behavior intention and attention, the CFCCM algorithm is used to achieve precise training of students' short-term behavior characteristics in order to improve the effect of performance clustering prediction. Finally, in the experiment, this paper takes a college student campus card system as an example to conduct empirical research and test the effectiveness of the proposed hybrid algorithm on real data.

Keywords: big data analysis; kernel function; fuzzy C-means clustering; educational data mining; student performance prediction; Collaborative CAD.
DOI: https://doi.org/10.14733/cadaps.2025.S5.75-92

1 INTRODUCTION

With the advancement of society and the development of science and technology, people's lives are becoming increasingly networked and digitized, and information management systems have replaced traditional text recording methods. What follows is the linear growth of data recording the trajectory of human daily behavior[13]. In education, especially in school education, student behavior data has increasingly become the most significant indicator of teaching improvement. Educational data mining (EDM) [1], as a technology for mining potential information from massive user learning behavior data, has been widely used in scientific research, business, finance, and other fields.

The application of the campus card system in Chinese colleges and universities has penetrated into the full range of students' learning dynamics and life trajectories. This digital information is completely recorded in the form of text, which has certain practical significance for education managers when analyzing the traces of students' activities. The ultimate goal of big data analysis in the education field is to improve students' academic performance and promote the overall development of students, and each student is an independent individual with different behavioral characteristics and motivations. By exploring students' behavioral data, it is possible to study whether they plan to invest more time in learning. Because students' behaviors are intuitive, they can judge the results more directly and quickly. Don't take action when students' learning and life problems are discovered at the end of the semester [17]. On this basis, this article discusses the prediction methods of student performance from the students' campus behavior. With a view to improving student performance, these methods can pay more attention to and identify students with poor academic performance and, at the same time, enable educators to obtain early feedback and take timely intervention measures.

In order to study this problem, many scientific researchers use different techniques [15], such as statistical analysis, data mining, and questionnaire surveys, to dig out the behavioral characteristics of students through the massive data obtained from third-party platforms and make school administrators aware of the students. The potential relationship between behavioral data and academic development undoubtedly provides an important reference for the school's daily management and decision-making. But they still have limitations [22]:

1. These methods mainly focus on manually extracting statistical features from pre-stored data, leading to lag in predicting student performance and discovering student problems.

2. Due to the limited ability to represent manually extracted features, they cannot extract deeply effective information from the data and find valuable knowledge and content. They can only understand students' behavior in a simple way.

In the education field, with the increasing popularity of digital and scientific e-learning systems for different terminal devices, it is natural that the behavioral data recorded in the database system has the characteristics of diversity and mass, which makes the application value of the education big data more prominent. Educational data mining is committed to exploring unique educational technology methods from the educational environment and using these methods to understand students and their learning environment better. In terms of the application types of educational data mining, the current educational data mining includes the following three aspects [5]:

1. Establish a learner model based on various educational data and real-time analysis

Through the modeling of students' living habits, behavior motivation, and learning strategies to reveal their learning characteristics. Reference [16] proposed a student modeling framework based on login behavior data, which uses unsupervised and supervised clustering to build a learner model based on the learning environment. Experiments show that the framework can automatically identify meaningful student interactions and perform online clustering of new student behaviors. Reference [7] proposed a method aimed at modeling learning characteristics of students' changing knowledge states in the process of acquiring skills. This method uses dynamic Bayesian networks to represent the uncertainty in time series data, which is used to explain the causal relationship between student knowledge and dynamic behavior. Another reference [19] showed that smartphones and niche ways of collecting real-time records of student life and learning activities. The analysis of these data reveals many unreported correlations between behavioral characteristics and academic performance.

2. Predict the learner's future learning performance based on the learning activity process data

This is also one of the most popular applications of educational data mining. The reference [12] conducts statistical analysis on the profile of 16 courses and learners on the EDX platform, summarizes learners' learning behavior characteristics, and conducts data mining on some typical behavior characteristics. Specifically, logistic regression methods are used to analyze whether the learner can complete the learning task to make predictions. Existing studies have concluded that [27] the learning participation, login frequency, interaction quality, and emotional state of different learners are quite different. Based on students' online learning behavior and their performance in the curriculum, a combination of educational data is introduced in such research. Mining and regression analysis techniques explore the predictive framework that affects performance. Reference [14] extracts behavioral features (such as note-taking, attention, homework submission, etc.) from real student data, compares the experimental performance of diverse clustering techniques, and learns the behavioral tendencies described by students in class. Find a qualitative model for the performance prediction problem, thereby establishing an efficient early warning system.

3. Based on deep learning and recommendation systems, differentiated recommendations for learners' interests

Courses, learning methods, and personalized learning paths can be recommended to learners according to their own characteristics. Reference [11] considers both the personality of the students and the commonality of the group of students, uses cognitive diagnosis technology to model the learning state of the students and proposes a collaborative filtering test question recommendation method based on the mastery of the student's knowledge points. Reference [3] constructed an ALS model that supports collaborative recommendation in a web learning environment. The model uses the learner's cognitive level as the introduction point to extract the learning paths of similar learners to generate personalized learning path recommendations. This method can be implemented based on accurate resource recommendations, which can effectively improve the learning quality and results of learners.

4. Visually analyze learner knowledge information based on visualization technology

In educational data mining, visualization technology can help researchers understand educational big data more intuitively and clearly. Reference [23] proposed a new data mining research data flow called process mining. When conducting educational data analysis, this method can provide a complete visual representation of the education (evaluation) process and prove its applicability in the educational data mining environment. Reference [10] proposes a learning analysis tool for educators' guidance programs by providing feedback on students' learning activities and performance to educators.

In view of the existing problems in the current work, this paper proposes a performance prediction modeling method based on student behavior sequences based on the correlation analysis of campus behavior data. In this article, in order to explore the performance-related factors and behavior patterns in behavior data and establish more accurate EDM methods, an efficient clustering model, fuzzy c-means clustering based on kernel function (CKFCM), is constructed to achieve the goal of predicting performance based on students' behavior data. Such a proposed method will make full use of the attributes and characteristics of education data and use data mining-related algorithms as a means to make up for the shortcomings of traditional methods.

2 EDUCATIONAL BIG DATA MINING BASED ON THE FEATURE EXTRACTION OF STUDENT CAMPUS CARD DATA

In recent years, many universities have begun to use big data analysis technology to solve practical problems encountered in daily teaching management. With the increasing popularity of campus all-in-one cards in major universities, a large amount of student behavior trajectory data has been collected. This paper mainly conducts data mining analysis from the perspective of clustering and statistical characteristics and explores the internal relationship between student behavior characteristics and performance. A novel clustering algorithm is used to predict students' future academic performance and to provide timely guidance for students in learning and life [20].

2.1 Data Collection and Cleaning

The data used in this article comes from the open-source data set of a third-party platform, which is mainly used to publicly study the personal development of students and the supervision and management of the school. The content of this data set includes the behavior data of students using the campus card swiping card in two academic years of a university and the grade ranking data in the teaching management system. Specifically, it includes book borrowing data, all-in-one card data, dormitory access control data, library access control data, and student performance data. The student information (such as student ID) in the data set has been anonymized through data desensitization. Part of the data field information of the learned behavior is shown in Table 1.

Serial number of indexes	*Name of indexes*	*Description of indexes*
Student library borrowing data		
1	*id*	*Student number*
2	*borrow_time*	*Borrow time*
3	*book*	*Book name*
4	*ISBN*	*Index book number*
Student card system data		
1	*id*	*Student number*
2	*card_type*	*Consumption type*
3	*card_place*	*Consumption place*
4	*card_way*	*Consumption cway*
5	*card_time*	*Consumption time*
6	*card_amount*	*Consumption amount of money*
7	*card_balance*	*Rest amount of money*
Student dormitory access control data		
1	*id*	*Student number*
2	*dorm_time*	*In and out time*
3	*dorm_direction*	*In and out direction (0 or 1)*
Student library access control data		
1	*id*	*Student number*
2	*library_time*	*In and out time*
3	*library_number*	*Number of entrance guard*
Student achievement ranking data		
1	*id*	*Student number*
2	*College*	*Number of colleges*

3	*rank*	*Rank of scores*

Table 1: Data field information of each part of student behavior.

Among them, the consumption category field specifically includes POS machine consumption, card loss report, and card replacement. The consumption mode fields mainly include the canteen, supermarket, boiling water, bathing, library, laundry, printing center, academic affairs office, school bus, and school hospital. Data cleaning is generally for specific applications. Specifically, data cleaning is a process of streamlining the database to remove duplicate records and convert the remaining part into an acceptable standard format. In order to better protect the privacy of students, the results are converted into rankings and normalized, and data mining algorithms are used for learning.

The consumption mode field in the all-in-one card data is the key data to be analyzed, but when there are missing values in the original data, the row where the data is located should be deleted. When the borrowing time, consumption time, and entry and exit time fields are exactly the same twice, delete the row where the data is located. Repeated records will affect data analysis and may even lead to errors. If more than half of the extracted traditional features are 0 in a row, it should be interpreted as the student's lack of school behavior data due to personal reasons or other reasons, and the row should be deleted. Through data transformation and integration, the data from multiple data sources are combined, merged, and converted into a form suitable for data mining to perform data reduction. To reduce the dimensionality of features, the original attributes that are not related to the mining task of the campus card are deleted during data cleaning.

2.2 Student Behavior Feature Extraction

Based on learning and summarizing previous experience, this paper studies the behavioral data of students' campus cards and understands that there are many aspects of site evaluation that are related to student's performance ranking and behavior. First, statistical analysis methods are used to select behavioral attributes related to performance ranking. This paper manually extracts 18 features (as shown in Table 2, based on the existing machine learning algorithms, finds the most suitable parameters to form the optimal state, and predicts and evaluates the performance of students of different ranking levels. The characteristic fields of Study Habits (SH), Living Habits (LH), and Consumption Habits (CH) are as follows: The study habits of college students have been formed for a long time, and their tendencies and behaviors are not easily changed with external factors.

Studying Habits	*Living Habits*	*Consumption Habits*
borrowBookDaily	*earlyDorm*	*aveCanteen*
borrowBookTest	*lateDorm*	*aveMarket*
lateLibrary	*stayInDormHour*	*aveWater*
earlyLibrary	*ShowerWeekly*	
stayLabHour	*earlyBreakfast*	
libraryTest	*printCenterDaily*	
libraryDaily	*printCenterTest*	
posStatistics		

Table 2: Types of traditional methods for extracting student behavior characteristics.

Features 1-8, respectively, represent the number of times students borrowed books during the non-exam time period, the number of times they borrowed books during exam time, the number of times entering the library before 8:00, the number of leaving the library after 22:00, the average daily in the library. The length of stay, the number of times to go to the library during non-examination time, the number of times to go to the library during examination time, and the number of times to use the POS machine during class time are classified as learning habits. Good and regular living habits have certain benefits to academic performance. These habits are closely related to students' self-control and self-discipline abilities.

This article chooses the average daily water fetching times and dormitory stay times as the characteristics of living habits. Features 9-15, respectively, indicate the number of times students leave the dormitory before 8:00, the number of times they enter the dormitory after 22:00, the average staying time in the dormitory every day, the number of showers per week, the number of breakfasts eaten before 8 o'clock, and the times when they go outside the test. The number of times for the printing center without a test and the number of visits to the printing center during the test period are also recorded as living habits. College students with different academic achievements have differences in the amount of campus card consumption, which reflects the different consumer needs and consumer psychology of college students. Features 16-18, respectively, represent average daily canteen consumption, daily average supermarket consumption, and average daily water fetching times.

2.3 Analysis of Student Behavior Data

In order to better explain the behavioral data and have a more comprehensive understanding of the data in advance, this article first divides the students' performance ranking into three categories according to the normal distribution, namely good, medium, and poor. Set the label to "Type 1", which accounts for 19.84% of the total number of people, the label to "Type 2", which accounts for 60.02% of the total number of people, and the label to "Type 3", which accounts for 20.14% of the total number of people. Then, the campus behavior and performance ranking data will be analyzed, and the original behavior data will be converted into behavior characteristics related to academic performance. This article compares the differences in the number of times the three types of students go to the library, borrow books, and go to the printing center during the examination and non-examination periods. As shown in Figure 1, in two academic years, all college' students were counted on library borrowing data.

The results showed that the students in the first category borrowed an average of 52 books, the students in the second category borrowed an average of 48 books, and the students in the third category borrowed an average of 42 books. The behavior of self-study and borrowing books in the library decreases with the types of students. The first type of students have the best learning and borrowing habits, and the third type of students have relatively few of the above three behaviors, and they lack the practical ability to act in the daily campus. Due to the non-uniformity of the teaching arrangements and daily management of the colleges, there will be differences in the rules of students' work and rest. In order to avoid large data errors, this paper selects the student behavior data of the No. 1 college for independent analysis.

As shown in Figure 2, the statistics of the three meals of the three types of students in college No. 1 show that the number of lunches and dinners is much greater than the number of breakfasts, indicating that most students have irregular eating habits. The students who belong to the first category have the most breakfasts, indicating that compared to this category, students have better eating habits. Whether it is the use of books, the contribution of club activities, or the consumption of food, it illustrates the importance of self-discipline to a college student, which is also determined by the characteristics of the college student.

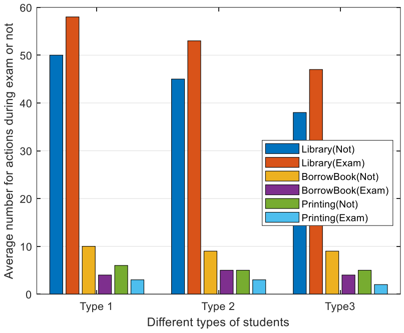

Figure 1: Comparison of the average number of behaviors of the three types of students during examination time and non-examination time.

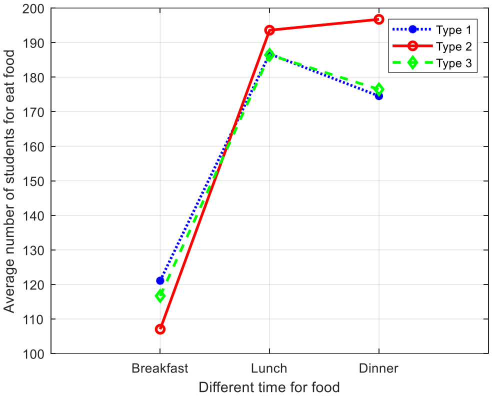

Figure 2: Statistical meal results for the three types of students in the college on a certain day.

3 FCM CLUSTERING ALGORITHM BASED ON IMPROVED KERNEL FUNCTION (CKFCM)

Cluster analysis is a very important technology in data mining technology. It can classify data without guidance to assist decision-making. Fuzzy cluster analysis is a data analysis method that combines fuzzy theory with cluster analysis technology in data mining. As the basic task of data mining, fuzzy clustering analysis provides a broad development space for data mining. Data mining based on this has played a great role in knowledge acquisition. It also has the characteristics of high efficiency, flexibility, and scalability and is suitable for high-dimensional processing. It is widely used in many fields [3]. Among the many algorithms of fuzzy clustering, the fuzzy C-means algorithm (FCM) [6] can be said to be the most widely used and most sensitive algorithm. However, because it is a local search algorithm, it faces the problem of being sensitive to initialization and easily falling into local optimal solutions.

There are many types of kernel functions. In summary, there are two main types of kernel functions: local kernel functions and global kernel functions. The Gaussian kernel function is a more local kernel function, and its interpolation ability decreases with the increase of the parameter σ; the polynomial kernel function is a more global kernel function with strong extrapolation ability, and a low order function will bring the stronger extrapolation ability [8]. In order to establish a more flexible learning model, the Gaussian kernel function and the polynomial kernel function can be combined to make full use of the respective advantages of the two types of kernel functions so that the combined kernel function has both better learning ability and better promotion ability. Therefore, this paper will use such combined kernel function as an entry point to find a clustering algorithm suitable for data mining of college scores by adjusting the parameters.

3.1 Improved kernel Based Fuzzy C-Means Cluster Algorithm (CKFCM)

Here the Gaussian kernel function and the linear kernel function are linearly combined [21]:

$$K_{mix} = \lambda K_{poly} + (1-\lambda)K_{RBF} = \lambda\left(a\left(x \cdot y\right)+b\right)^d + (1-\lambda)\exp\left(-\frac{\|x-y\|^2}{2\delta^2}\right), \lambda \in [0,1] \tag{1}$$

where $a > 0, b \geq 0$, and d is a positive integer, indicating the order of the polynomial. $\delta > 0$ is the width parameter of the function, and $\|x-y\|^2$ represents the Euclidean distance between objects x and y. Based on the FCM algorithm of the improved kernel function, the objective function is established:

$$J_m^{\phi}(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c}\left(u_{ik}\right)^m \|\phi(x_k)-\phi(v_i)\|^2, m \in (1,\infty) \tag{2}$$

where $\phi(\cdot)$ is a non-linear transformation function, $\phi(x_k), \phi(v_i)$ respectively represent the image of the sample and the cluster center in the feature space, and m is the weighted exponent problem:

$$\min J_m^{\phi}(U^*,V^*) = \sum_{k=1}^{n}\sum_{i=1}^{c}\left(u_{ik}\right)^m \|\phi(x_k)-\phi(v_i)\|^2 \tag{3}$$

Make the following condition subject to:

$$\sum_{i=1}^{c}\left(u_{ik}\right) = 1, 1 \leq k \leq n, 1 \leq i \leq c, c \in \left\{2,3,...,n-1\right\} \tag{4}$$

Theorem 1

Let $X = \left\{x_1, x_2, ..., x_n\right\}$, and $x_i \in R^p$ is a given data set. Set $c \in \left\{2,3,...,n-1\right\}$ and $m \in \left\{1,\infty\right\}$, assuming that all $1 \leq k \leq n$ and $1 \leq i \leq c$ have $\|\phi(x_k)-\phi(v_i)\| \neq 0$. Then the condition is established only if the following equation:

$$u_{ik} = \frac{\left(K_{mix}(x_k,x_k)+K_{mix}(v_i,v_i)-2K_{mix}(x_k,v_i)\right)^{\frac{1}{1-m}}}{\left[\sum_{j=1}^{c}\left(K_{mix}(x_k,x_k)+K_{mix}(v_j,v_j)-2K_{mix}(x_k,v_j)\right)\right]^{\frac{1}{1-m}}}, 1 \leq i \leq c, 1 \leq k \leq n \tag{5}$$

$$\phi(v_i)=\frac{\sum_{k=1}^{n}(u_{ik})^m\phi(x_k)}{\sum_{k=1}^{n}(u_{ik})^m},1\le i\le c \tag{6}$$

Make the equation (3) reach the minimum value of equation (2). The definition of each parameter is $\phi(x_k)$, and $\phi(v_i)$ respectively represent the image of the sample and the cluster center in the feature space.

Proof of theorem 1

The following conditions are known: $c\in\{2,3,...,n-1\}$, $m\in\{1,\infty\}$, $1\le k\le n$, and $1\le i\le c$. Assuming that v_i is constant, the problem becomes the minimization problem of J_m^ϕ on u_{ik} under constraint equation (4). With the introduction of Lagrangian multipliers, this problem is equivalent to the following unconstrained minimization problem:

$$L(U,\lambda)=\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m\left\|\phi(x_k)-\phi(v_i)\right\|^2-\sum_{k=1}^{n}\lambda_k\left(\sum_{i=1}^{c}u_{ik}-1\right) \tag{7}$$

The necessary conditions for this question are:

$$\frac{\partial L(U,\lambda)}{\partial u_{ik}}=\left[m(u_{ik})^{m-1}\left\|\phi(x_k)-\phi(v_i)\right\|^2-\lambda_k\right]=0 \tag{8}$$

$$\frac{\partial L(U,\lambda)}{\partial \lambda_k}=\sum_{i=1}^{c}u_{ik}-1=0,1\le k\le n \tag{9}$$

From equation (8):

$$u_{ik}=\left(\frac{\lambda_k}{m\left\|\phi(x_k)-\phi(v_i)\right\|^2}\right)^{\frac{1}{m-1}} \tag{10}$$

Substituting equation (10) into equation (9), we can get:

$$\left(\frac{\lambda_k}{m}\right)^{\frac{1}{m-1}}=\frac{1}{\sum_{i=1}^{c}\left(\frac{1}{\left\|\phi(x_k)-\phi(v_i)\right\|^2}\right)^{\frac{1}{m-1}}} \tag{11}$$

Substituting equation (11) into equation (10), we can get:

$$u_{ik}=\frac{1}{\sum_{j=1}^{c}\left(\frac{\|\phi(x_k)-\phi(v_i)\|^2}{\|\phi(x_k)-\phi(v_j)\|^2}\right)^{\frac{1}{m-1}}} \tag{12}$$

Calculated by the following equation:

$$\begin{aligned}&\|\phi(x_k)-\phi(v_i)\|^2\\&=K_{mix}(x_k,x_k)+K_{mix}(v_i,v_i)-2K_{mix}(x_k,v_i)\\&=(\phi(x_k)-\phi(v_i))^T(\phi(x_k)-\phi(v_i))=\phi(x_k)^T\phi(x_k)+\phi(v_i)^T\phi(v_i)-2\phi(x_k)^T\phi(v_i)\end{aligned} \tag{13}$$

Substituting equation (13) into equation (12), equation (5) can be obtained.

In order to prove that equation (6) is true, suppose u_{ik} does not change, and let $g_i=\phi(v_i)$. Then this problem becomes an unconstrained minimization problem, and its necessary condition is $\frac{\partial J_m^\phi(U,V)}{\partial g_i}=0$, so we get:

$$\sum_{k=1}^{n}(u_{ik})^m\frac{\partial}{\partial g_i}\|\phi(x_k)-g_i\|^2=0,$$

$$\frac{\partial J_m^\phi(U,V)}{\partial g_i}=-\sum_{k=1}^{n}2(u_{ik})^m\|\phi(x_k)-g_i\|=0,$$

then $$\phi(v_i)=g_i=\frac{\sum_{k=1}^{n}(u_{ik})^m\phi(x_k)}{\sum_{k=1}^{n}(u_{ik})^m},1\le i\le c.$$

The certificate is complete.

3.1 Convergence of CKFCM Algorithm

Theorem 1 gives the necessary conditions for (U,V) to reach a local minimum of J_m^ϕ. It is unclear whether the algorithm converges to a local minimum of J_m^ϕ. The following proves the convergence of the FCM clustering algorithm based on the improved kernel function.

Theorem 2

Let $\phi(v_i)^{(l)}, u_{ik}^{(l)} (l = 0,1,2,...)$ be the sequence calculated based on the improved kernel function FCM clustering algorithm. If $m > 1$, and for all $k = 1,2,...,n$, $i = 1,2,...,c$ and $l = 0,1,2,...$, there is $\|\phi(x_k) - \phi(v_i)\| \neq 0$, then for all $l = 0,1,2,...$, there is:

$$J_m^\phi\left(U^{(l+1)}, V^{(l+1)}\right) \leq J_m^\phi\left(U^{(l)}, V^{(l)}\right) \tag{14}$$

That is, J_m^ϕ is convergent.

Proof of the convergent

(i) This part proves that equation (5) is the necessary and sufficient condition for $J_m^\phi\left(U, V\right)$ to reach the minimum value.

Let $\phi(U) = J_m^\phi\left(U, V\right)$, where V is constant, and for all $k = 1,2,...,n$, $i = 1,2,...,c$, there is $\|\phi(x_k) - \phi(v_i)\| \neq 0$, the necessary condition for $\phi(U)$ to reach a local minimum is equation (5), which has been proved in Theorem 1. The following proves its sufficiency.

Introduce the Hessian matrix $\boldsymbol{H}(U)$ of $\phi(U)$ of Lagrange multiplier. According to equation (7), there are:

$$h_{st,ik}(U) = \frac{\partial}{\partial u_{st}}\left[\frac{\partial \phi(U)}{\partial u_{ik}}\right] = \begin{cases} m(m-1)(u_{st})^{m-1}\|\phi(x_t) - \phi(x_s)\|, s = i, t = k \\ 0, otherwise \end{cases} \tag{15}$$

where u_{st} is calculated by equation (5). Therefore, $\boldsymbol{H}(U) = \left[h_{st,ik}(U)\right]$ is a diagonal matrix. Since $m > 1$ and for all $k = 1,2,...,n$, $i = 1,2,...,c$, there is $\|\phi(x_k) - \phi(v_i)\| \neq 0$, and then there is $m(m-1)(u_{st})^{m-1}\|\phi(x_t) - \phi(x_s)\| > 0$. Therefore, the Hessian matrix $\boldsymbol{H}(U)$ is positive definite, which shows that equation (5) is a sufficient condition for minimizing $\phi(U)$.

(ii) This section proves that equation (6) is the necessary and sufficient condition for $J_m^\phi\left(U, V\right)$ to reach the minimum value.

Let $\varphi(V) = J_m^\phi\left(U, V\right)$, where U is invariant, and for all $k = 1,2,...,n$, $i = 1,2,...,c$ have $\|\phi(x_k) - \phi(v_i)\| \neq 0$. The necessary condition for $\varphi(V)$ to reach a local minimum is that equation (6) has been proved in Theorem 1. The following proves its sufficiency. Let $\phi(v_i) = g_i$, according to equation (13), there are:

$$\frac{\partial}{\partial u_{st}}\left[\frac{\partial \varphi(U)}{\partial g_i}\right] = \begin{cases} \sum_{k=1}^{n} 2(u_{jk})^m, j = i \\ 0, otherwise \end{cases} \tag{16}$$

where $\phi(v_i) = g_i$ is calculated by equation (6). Therefore, the Hessian matrix is a diagonal matrix. Since $m > 1$, $\sum_{k=1}^{n} 2(u_{jk})^m > 0$. Therefore, the Hessian matrix is positive definite, which shows that equation (6) is a sufficient condition for minimizing $\varphi(V)$.

(iii) This section proves that equation (14) holds.

Since $U^{(l)}$ is calculated by equation (5) when V is constant, according to the proof of (i):

$$J_m^{\phi}\left(U^{(l+1)}, V^{(l)}\right) \le J_m^{\phi}\left(U^{(l)}, V^{(l)}\right) \tag{17}$$

Since $V^{(l)}$ is calculated by equation (6) when U is unchanged, according to the proof of (ii):

$$J_m^{\phi}\left(U^{(l+1)}, V^{(l+1)}\right) \le J_m^{\phi}\left(U^{(l+1)}, V^{(l)}\right) \tag{18}$$

Combining equation (17) and equation (18) can give equation (14). The certificate is complete.

3.2 CKFCM Algorithm Steps

The steps of FCM clustering algorithm based on improved kernel function are as follows:

Step 1 Given a data set $X = \{x_1, x_2, ..., x_n\}, x_i \in R^p$. Set $c \in \{2, 3, ..., n-1\}$ and $m \in (1, \infty)$, set the iteration stop threshold ε, initialize the cluster prototype matrix $V^{(0)}$, and set the iteration counter $b = 0$;

Step 2 Calculate the kernel matrix $\boldsymbol{K}(x_i, x_j)$, $i, j = 1, 2, ..., n$;

Step 3 Calculate the membership function of each sample in the feature space according to equation (2);

Step 4 According to equation (6), it can be derived:

$$K_{mix}(x_k, v_i) = \phi(x_k) \cdot \phi(v_i) = \frac{\sum_{j=1}^{n} u_{ij}^m K_{mix}(x_i, x_j)}{\sum_{j=1}^{n} u_{ij}^m} \tag{19}$$

$$K_{mix}(v_i, v_i) = \phi(v_i) \cdot \phi(v_i) = \frac{\sum_{j=1}^{n}\sum_{l=1}^{n} u_{ij}^m u_{il}^m K_{mix}(x_j, x_l)}{\left(\sum_{j=1}^{n} u_{ij}^m\right)^2} \tag{20}$$

Calculate the new kernel matrix $\boldsymbol{K}(x_k, v_i)$ and $\boldsymbol{K}(v_i, v_i)$ from equations (19) and (20), and update the membership degree $U^{(b+1)}$ according to equation (5);

Step 5 If $\left\|U^{(b+1)}-U^{(b)}\right\|<\varepsilon$ or the number of iterations is equal to the predetermined number, stop, otherwise, go to Step 4.

4 SIMULATION EXPERIMENT AND RESULT ANALYSIS

4.1 Experimental Data Set

Performance prediction is one of the classic problems of educational data mining. The behavior information recorded by the campus card is inextricably related to academic performance data. Sequence-based performance cluster prediction (SPC) is to predict the current performance of students by clustering the recent sequence behaviors of students (such as entering and leaving the library, fetching water, and going to the cafeteria) [25]. For a particular student, when the school only provides a very limited record of student behavior, whether the study can judge the student's performance based on their behavioral intentions.

Suppose $[x_1, x_2, ..., x_i, ..., x_t]$ is a student behavior sequence, where x_i is a specific behavior indicator in m campus card terminal devices. Let $R=\left(r_{i,j}\right)_{n,m}$ represent the relationship items between n students and m campus card terminal devices, and each item $r_{i,j}$ represents that student i has swiped the card on device j. The performance prediction model M is established by clustering the given behavior sequence $x=[x_1, x_2, ..., x_{t-1}, x_t],(1 \le t \le m)$ into different levels (i.e., good, medium, and poor). Compared with the traditional performance prediction task, this paper focuses on how to make timely predictions by improving the representative features of the FCM clustering algorithm of the kernel function [24]. The formal definition of the task is:

Input: student set n , campus card equipment set m , student behavior sequence $[x_1, x_2, ..., x_i, ..., x_t]$ every week.

Output: The mapping function maps the input sequence to the grade cluster level $M: x \rightarrow R$.

Since all the records in the data set are obtained after data desensitization of the "raw data records", there are some duplicate or abnormal records, so data cleaning is carried out to make it more suitable for the work of this paper.

In order to reduce the impact of inaccurate data, the experiment further extracted the behavior records of 9,341 students for 29 weeks (March to June 2018 and April to June 2019). This data set is a subset of student id, behavior, and academic performance. That is, it contains variables related to students' study and life attributes at school, including entering and leaving the library, fetching water, and 13 other behaviors. In order to further model according to the card swiping sequence, the student's behaviors are sorted according to the time when the students visit the card reader, and then the behavior sequence $[x_1, x_2, ..., x_{t-1}, x_t]$ is formed, where x_i represents a kind of card reading. It is located in different buildings, such as the front door of a library or laboratory.

#Students	*#Devices*	*#Sequences*	*Avg. length*
9341	*16*	*125956*	*51.63*

Table 3: Statistical results of the campus card data set.

Table 3 describes the statistical information of the data set. There are four reasons for choosing these data sets: First, these behavioral data are not directly related to academic performance, so the relationship between the two aspects can be explored. Secondly, these behaviors are not compelling, so they can objectively reflect the lifestyle of students without experimental bias. For example, it can be seen from Table 3, that most college students in China live and study on campus and are not prone to emergencies. Therefore, the data set used has sufficient coverage to verify the results. Finally, analyzing academic performance is not only conducive to the management of teachers' daily activities but also provides important information support and forward-looking services for education and teaching.

4.2 Experimental Results and Analysis

In order to verify the performance of the CKFCM algorithm proposed in this paper in the student's academic performance prediction task, it was compared with the following baseline methods (logistic regression [2], naive Bayes [18], decision tree [21] and random forest [26]) Comparison: Table 4 shows the results of the comparison experiment on accuracy and recall between the performance prediction clustering algorithm based on behavior sequence proposed in this paper and several other benchmark experiment algorithms.

Predicting Algorithms	*Accuracy*	*Improved Accuracy(%)*	*Recall*	*Improved Recall(%)*
Logistic Regression	*58.62*	*42.00*	*44.25*	*71.80*
Naïve Bayesian	*58.79*	*41.59*	*43.33*	*75.44*
Decision Tree	*59.13*	*40.77*	*41.57*	*82.87*
Random Forest	*59.44*	*40.04*	*42.35*	*79.50*
CKFCM	*83.24*	-	*76.02*	-

Table 4: Comparison results of the CKFCM algorithm proposed in this paper with existing algorithms in academic performance prediction.

The following observations and conclusions can be drawn from it:

1. It can be observed that the CKFCM method based on big data analysis and clustering proposed in this paper is more consistent and significant, which shows that the FCM algorithm is good at processing sequence information in multiple sequences and effectively solving clustering tasks.

2. By considering the students' sequence behavior and main purpose, the CKFCM method proposed in this paper can outperform all baselines. The relative performance of CKFCM is 83.24% and 76.02% in terms of accuracy and recall rate, respectively.

3. Obviously, other data mining models have almost no difference in experimental results, and these models are not satisfactory in accuracy and recall. One of the possible reasons is that these models have their own limitations when dealing with various types of data, so they cannot effectively deal with the characteristics of behavior sequences.

4. Therefore, it is feasible to combine the kernel function technique with the FCM algorithm since it can learn the potential relationship between student behaviors and can determine which sequential behavior features are more important for multi-cluster performance prediction tasks. When studying educational issues, researchers should not only focus on whether the results of the model are good or bad but also have more in-depth discussions.

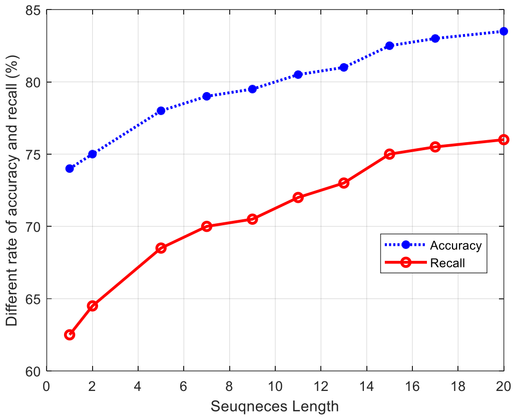

Figure 3: Different sequence lengths on the accuracy and recall of the SPC model.

The experimental influence obtained in this section is based on the hypothesis: when a student produces various campus behaviors in order to achieve a better performance prediction effect, his/her behavior often follows certain patterns and main intentions. However, when students perform only a few behaviors, the experiment can hardly explore the orderliness of their daily activities, and it is difficult to capture their main purpose. Therefore, this experiment compares sequence behaviors of different lengths to investigate whether the model is good at modeling long sequence behaviors. As shown in Figure 3, the horizontal axis represents the sequence length in weeks. It can be found that when the length of the behavior sequence increases from 1 to 20, the performance of the model gets better and better, and both exceed the accuracy of 60%. This shows that if the CKFCM method-based SPC model captures more target features on the basis of the existing sequence behavior features, it may achieve better predictions. Another reason may be that students are very likely to develop regular life and study habits.

Periods	*Accuracy*	*Recall*	*#Sequences*	*Avg. length*
Weekly	*84.21%*	*82.63%*	*126127*	*51.63*
Monthly	*72.42%*	*58.63%*	*44856*	*170.24*

Table 5: Different cycles on the accuracy and recall of the CKFCM method-based SPC model.

However, as shown in Table 5, it can be found that when the sequence length is divided by month, the evaluation result of the model drops (only 72.42%). The reason may be that when a sequence is too long, the administrator does not assess the status of the students in a timely manner and intervene in their learning, resulting in a decrease in accuracy. Therefore, compared with standard academic evaluation or personal static information, short-term continuous behavior modeling can more sensitively monitor students' daily activities and can more appropriately reflect their life and learning conditions.

As shown in Figure 4, the experimental results show that whether the feature dimensions varied, the training quantities will affect the selection of students' behavioral features in the method evaluation of accuracy and recall. If the hidden unit and epoch increase, the accuracy and recall rate will increase significantly.

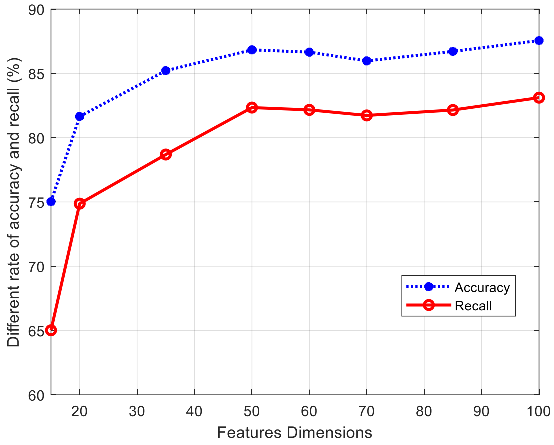

Figure 4: Different feature dimensions on the accuracy and recall of the SPC model.

It can be found that as the feature dimension is set to increase from 5 to 100, the performance improvement of the model in terms of accuracy exceeds 75% at the maximum, and the performance improvement in terms of recall rate exceeds 60%. It can also be found that as the total number of clusters increases, the curves of the two indicators show a flat trend. The lower performance in lower feature dimensions may be due to the fact that the clustering centers are not accurate enough to represent students' behavior characteristics. The more feature dimensions that are set for calculating the clustering process, the more comprehensive the students may understand. However, the representation ability of the model is limited, and the accuracy of models with more than 50-dimensional features tends to be stable. At this time, some repetitive features may appear in the algorithm training process. Therefore, setting appropriate feature dimensions is helpful in establishing student behavior models and helps educators understand student behavior and explore the information contained in the data.

5 CONCLUSIONS

This paper proposes a data mining system that can predict the probability of success of students' learning based on their behavior on campus. The behavioral data comes from a huge dataset that collects data on students' use of their campus smart cards. The main work of this paper includes:

1. This paper proves the feasibility and convergence of the fuzzy C-means algorithm based on a mixed kernel function. Clustering students' scores using such an algorithm is faster, more scientific, and more practical, and it has very good application value.

2. The problem of student behavior and academic performance is studied, and the student's performance prediction task is regarded as a short-term sequence modeling problem.

In order to extend the method in this article to real-time scenarios, the campus card data is analyzed and applied to complete the work. Compared with the benchmark algorithm, it proves that the depth of information on student behavior is more representative than traditional behavior characteristics. The effectiveness of the clustering algorithm is verified. This paper still has many shortcomings. There are still many areas to be improved, and many issues need to be studied further. The application of data mining technology in the education field is still in the exploratory stage, and there is still a lot of information worth mining in student behavior data waiting for people with lofty ideals to discover. Collaborative CAD and advanced data analytics in education have the potential to revolutionize teaching and learning by providing educators with timely, data-driven insights and more personalized and effective learning experiences for students. However, the successful execution of this vision will depend on a careful balance between innovation and addressing the various challenges that come with it.

Kaihe Yu, https://orcid.org/0009-0008-9507-1536

REFERENCES

[1] Ahuja, R.; Jha, A.; Maurya, R.: et al. Analysis of educational data mining, Harmony Search and Nature Inspired Optimization Algorithms, Springer, Singapore, 2019, 897-907. https://doi.org/10.1007/978-981-13-0761-4_85

[2] Aldowah, H.; Al-Samarraie, H.; Fauzy, W. M.: Educational data mining and learning analytics for 21st century higher education: A review and synthesis, Telematics and Informatics, 37, 2019, 13-49. https://doi.org/10.1016/j.tele.2019.01.007

[3] Arshad, A.; Riaz, S.; Jiao, L.: et al. Semi-supervised deep fuzzy c-mean clustering for software fault prediction, IEEE Access, 6, 2018, 25675-25685. https://doi.org/10.1109/ACCESS.2018.2835304

[4] Arshad, A.; Riaz, S.; Jiao, L.: Semi-supervised deep fuzzy C-mean clustering for imbalanced multi-class classification, IEEE Access, 7, 2019, 28100-28112. https://doi.org/10.1109/ACCESS.2019.2901860

[5] Asif, R.; Merceron, A.; Ali, S. A.: et al. Analyzing undergraduate students' performance using educational data mining, Computers & Education, 113, 2017, 177-194. https://doi.org/10.1016/j.compedu.2017.05.007

[6] Bakhshinategh, B.; Zaiane, O. R.; ElAtia, S.; et al.: Educational data mining applications and tasks: A survey of the last 10 years, Education and Information Technologies, 23(1), 2018, 537-553. https://doi.org/10.1007/s10639-017-9616-z

[7] Bharara, S.; Sabitha, S.; Bansal, A.: Application of learning analytics using clustering data Mining for Students' disposition analysis, Education and Information Technologies, 23(2), 2018, 957-984. https://doi.org/10.1007/s10639-017-9645-7

[8] Chadha, A.: Efficient clustering algorithms in educational data mining, //Handbook of Research on Knowledge Management for Contemporary Business Environments, IGI Global, 2018, 279-312. https://doi.org/10.4018/978-1-5225-3725-0.ch015

[9] Charitopoulos, A.; Rangoussi, M.; Koulouriotis, D.: Educational data mining and data analysis for optimal learning content management: Applied in moodle for undergraduate engineering studies, //2017 IEEE Global Engineering Education Conference (EDUCON), IEEE, 2017, 990-998. https://doi.org/10.1109/EDUCON.2017.7942969

[10] Choubin, B.; Solaimani, K.; Roshan, M. H.; et al.: Watershed classification by remote sensing indices: A fuzzy c-means clustering approach, Journal of Mountain Science, 14(10), 2017, 2053-2063. https://doi.org/10.1007/s11629-017-4357-4

[11] Chowdhary, C. L.; Acharjya, D. P.: Clustering algorithm in possibilistic exponential fuzzy c-mean segmenting medical images, //Journal of Biomimetics, Biomaterials and Biomedical Engineering. Trans Tech Publications Ltd, 30, 2017, 12-23. https://doi.org/10.4028/www.scientific.net/JBBBE.30.12

[12] Chowdhary, C. L.; Mittal, M.; Pattanaik, P. A.; et al.: An efficient segmentation and classification system in medical images using intuitionist possibilistic fuzzy C-mean clustering and fuzzy SVM algorithm, Sensors, 20(14), 2020, 3903. https://doi.org/10.3390/s20143903

[13] Dutt, A.; Ismail, M. A.; Herawan, T.: A systematic review on educational data mining, IEEE Access, 5, 2017, 15991-16005. https://doi.org/10.1109/ACCESS.2017.2654247

[14] Faradonbeh, R. S.; Haghshenas, S. S.; Taheri, A.: et al. Application of self-organizing map and fuzzy c-mean techniques for rockburst clustering in deep underground projects, Neural Computing and Applications, 32(12), 2020, 8545-8559. https://doi.org/10.1007/s00521-019-04353-z

[15] Fernández, D. B.; Luján-Mora, S.: Comparison of applications for educational data mining in Engineering Education, //2017 IEEE World Engineering Education Conference (EDUNINE). IEEE, 2017, 81-85. https://doi.org/10.1109/EDUNINE.2017.7918187

[16] Hung, H. C.; Liu, I. F.; Liang, C. T.: et al. Applying educational data mining to explore students' learning patterns in the flipped learning approach for coding education, Symmetry, 12(2), 2020, 213. https://doi.org/10.3390/sym12020213

[17] Hussain, S.; Atallah, R.; Kamsin, A.: et al. Classification, clustering and association rule mining in educational datasets using data mining tools: A case study, Computer Science Online Conference. Springer, Cham, 2018, 196-211. https://doi.org/10.1007/978-3-319-91192-2_21

[18] Kaur, H.; Bathla, E. G.: Student performance prediction using educational data mining techniques, International Journal on Future Revolution in Computer Science & Communication Engineering, 4(12), 2018, 93–97-93–97.

[19] Kausar, S.; Huahu, X.; Hussain, I.; et al.: Integration of data mining clustering approach in the personalized E-learning system, IEEE Access, 6, 2018, 72724-72734. https://doi.org/10.1109/ACCESS.2018.2882240

[20] Majdi, A.; Beiki, M.: Applying evolutionary optimization algorithms for improving fuzzy C-mean clustering performance to predict the deformation modulus of rock mass, International Journal of Rock Mechanics and Mining Sciences, 113, 2019, 172-182. https://doi.org/10.1016/j.ijrmms.2018.10.030

[21] Manjarres, A. V.; Sandoval, L. G. M.; Suárez, M. S.: Data mining techniques applied in educational environments: Literature review, Digital Education Review, 2018 (33), 235-266. https://doi.org/10.1344/der.2018.33.235-266

[22] Mishra, A.; Bansal, R.; Singh, S. N.: Educational data mining and learning analysis, //2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence. IEEE, 2017, 491-494. https://doi.org/10.1109/CONFLUENCE.2017.7943201

[23] Nawaz, M.; Yan, H.: Saliency detection via multiple-morphological and superpixel based fast fuzzy C-mean clustering network, Expert Systems with Applications, 161, 2020, 113654. https://doi.org/10.1016/j.eswa.2020.113654

[24] Ramaphosa, K. I. M.; Zuva, T.; Kwuimi, R.: Educational data mining to improve learner performance in Gauteng primary schools, //2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD). IEEE, 2018, 1-6. https://doi.org/10.1109/ICABCD.2018.8465478

[25] Salloum, S. A.; Alshurideh, M.; Elnagar, A.: et al. Mining in educational data: review and future directions, //Joint European-US Workshop on Applications of Invariance in Computer Vision, Springer, Cham, 2020, 92-102. https://doi.org/10.1007/978-3-030-44289-7_9

[26] Tang, H.; Xing, W.; Pei, B.: Time really matters: Understanding the temporal dimension of online learning using educational data mining, Journal of Educational Computing Research, 57(5), 2019, 1326-1347. https://doi.org/10.1177/0735633118784705

[27] Zhao, Q.; Shao, S.; Lu, L.: et al. A new PV array fault diagnosis method using fuzzy C-mean clustering and fuzzy membership algorithm, Energies, 11(1), 2018, 238. https://doi.org/10.3390/en11010238