Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 3126347, 10 pages
https://doi.org/10.1155/2021/3126347

Research Article

A College Student Behavior Analysis and Management Method Based on Machine Learning Technology

Xiaoying Shen[1] **and Chao Yuan**[2,3]

[1]*Wuxi Vocational College of Science and Technology, No. 8 Xinxi Road, Wuxi, Jiangsu 214000, China*
[2]*School of Design, Jiangnan University, 1800 Lihu Avenue, Wuxi, Jiangsu 214122, China*
[3]*College of Economics and Management, Nanjing University of Aeronautics and Astronautics, 29 Jiangjun Avenue, Nanjing, 211100 Nanjing, China*

Correspondence should be addressed to Chao Yuan; circle@jiangnan.edu.cn

Received 30 June 2021; Revised 16 July 2021; Accepted 20 July 2021; Published 2 August 2021

Academic Editor: Shan Zhong

A digital campus will generate a large amount of student-related data. How to analyze and apply these data has become the key to improving the management level of students. The analysis of student behavior data can not only assist schools in early warning of dangerous events and strengthen school safety but also can use real data to describe student behavior, thereby providing quantitative data support for scholarship and grant evaluation. This paper takes a university student as the research object, collects various data in the digital campus platform, and uses an adaptive K-means algorithm in the machine learning algorithm to cluster the data. Analyze the behavior of college students from the clustering results, so as to provide a basis for the education management and learning ability improvement of college students. Specifically, the student's study, life, and consumption data are selected as the data to describe the student's behavior at school. This data is input into the adaptive K-means algorithm to obtain different types of student consumption habits, living habits, and learning habits. Through the analysis results, it can be found that the problem of the group of students with low financial ability, the problem of too long online time for students, and the number of books borrowed are too low. According to the characteristics of these problems, teachers and schools are provided with targeted management suggestions. The analysis of student behavior based on machine learning technology provides a reference for the formulation of students' school management policies and provides teachers with information on students' personality characteristics, which is conducive to improving teachers' teaching effects. In short, the management of the results of student behavior analysis can provide a basis for the school to formulate reasonable management policies, thereby promoting precision management and scientific decision-making.

1. Introduction

The establishment of a digital campus has improved the efficiency of university management and has also brought great convenience to students, faculty, and staff. The digital management system can collect a large amount of data, which plays an important role in the management of the school. As for the daily management of students, if we can learn more about students, we can implement more effective programs for different students, so that we can teach students in accordance with their aptitude and improve the education level of the school. The traditional analysis and management of student behavior mostly relies on the personal experience of the manager and lacks the individualized cognition of the learner. At the same time, it cannot in-depth guide students' learning behaviors, provide personalized learning situations, and promote learning optimization. Analyzing student life and learning behavior based on intelligent technology are of great significance to the investigation of potential abnormal students and the prediction of students' future development. The key to understanding students is the data collected in the digital campus for students' study, life, and consumption. At present, many schools have established corresponding all-in-one card systems, which make students' daily campus life more convenient. Students can use the campus card to consume in canteens, supermarkets, etc., or use the card to borrow books in the library, etc. These operations will generate a large amount of student behavior data. How to use these data to discover the information contained in it is a problem that needs to be solved. Based on the machine learning technology, this paper conducts cluster analysis on campus all-in-one card data and analyzes the behavior of students.

In recent years, there has been a lot of research on student behavior analysis. Reference [1] measures student behavior based on entropy measurement. The study defined two behavioral characteristics of orderliness and diligence and analyzed the correlation between the regularity of campus life and academic performance. Reference [2] proposed a pre-class student performance prediction method based on multiexample multilabel learning. The idea of this method is to use students' behavior in completed courses to predict their difficulties in learning new courses. The results of this study are convenient for teachers to track and understand the learning situation of each student. Reference [3] proposes an education measurement system to characterize educational behavior by collecting campus Wi-Fi network data. The results show that the system can obtain information about the relationship between punctuality, distraction, and academic performance. Reference [4] uses an improved recurrent neural network to simulate the student's answering process according to the student's answer records and the content of each exercise to predict the student's future performance. Based on MOOCs learner behavior data, reference [5] established a prediction model based on clustering algorithm and neural network to mine the learning rules in the learning process. The predicted results can provide personalized guidance for each learner. Reference [6] proposes a classification system to analyze the behavior of students in the teaching system and to find students with poor performance early. Reference [7] predicts their course performance based on the relevant data generated by students during the online course learning process. Reference [8] input multimodel data and semester information into the linear mixed effects model to predict the future performance of students. Reference [9] found that an improved random forest method can be used to predict the grades of freshmen and existing courses. The application of these technologies brings hope to student degree planning, lecturer intervention, and personalized advice. Annapol State University in India has developed a product [10, 11], which is used to monitor student activity areas. This product analyzes students' participation in organizing activities based on the records of students swiping their ID cards. Through this software, information about students who participate in activities with low frequency can be detected.

Most of the above studies have completed the analysis of student behavior based on machine learning technologies [12–14]. The current popular machine learning technologies contain the tradition machine learning methods [15, 16] and some advanced machine leaning technologies [17–19]. These advanced machine leaning technologies have been used in many practical applications, such as medicine [20], industry [21], and basic theory research [22–24]. In terms of performance, the accuracy of behavior analysis based on deep learning is indeed higher, but its computational complexity and hardware performance requirements are higher. The time complexity of behavior analysis based on machine learning is relatively low, and the algorithm implementation is simple. Therefore, this article chooses student behavior analysis based on machine learning. In machine learning, K-means clustering algorithm [25, 26] is widely used in student behavior analysis research. However, traditional K-means is very sensitive to outliers, and a small number of outliers will have a great impact on the final clustering results. On the other hand, the K-means algorithm still has the problem that the K value cannot be adapted. In response to these problems, this paper uses an adaptive K-means algorithm to improve the efficiency and clustering accuracy of the algorithm.

The main work of this paper is as follows:

(1) Collect student consumption, life, and learning data through the campus all-in-one card system and integrate data from different institutions to form a comprehensive data set for student behavior analysis

(2) An adaptive K-means clustering algorithm is used for student behavior analysis. The algorithm introduces the elbow rule to optimize the data and find points far away from the cluster, thereby effectively improving the clustering performance. In addition, the algorithm also introduces the idea of self-adaptation and automatically adjusts the value of k based on the sum of squared errors. The adaptive K value is more suitable for objective reality

(3) Based on the above model, the analysis results of students' consumption, life, and learning are obtained. The analysis results can be used to improve the management level of the school and truly teach students in accordance with their aptitude

2. Related Work

2.1. Student Behavior Analysis and Management. The complexity of individual college students makes it impossible for school administrators to understand students' dynamics in real time. For some students with abnormal behaviors, the students around them may be inconvenient or embarrassed to inform the management staff of the specific situation, which causes a certain lag in the work of the student management staff. In order to understand the behavior and habits of students in real time, the management of students will be transformed from passive to active. The daily behavior data of students needs to be displayed intuitively on the student behavior analysis system. On the premise of ensuring the privacy and safety of students, the data of the system mainly comes from the data collected by the school's digital systems, and the machine learning technology is used to analyze the data. The goal of student behavior analysis is shown in Figure 1.

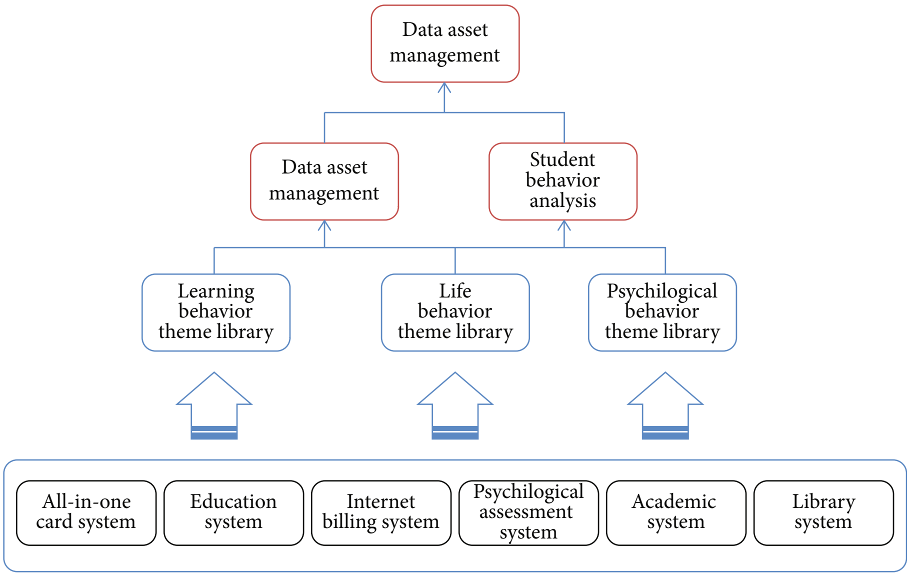
Data asset management
Data asset management
Student behavior analysis
Learning behavior theme library
Life behavior theme library
Psychological behavior theme library
All-in-one card system
Education system
Internet billing system
Psychological assessment system
Academic system
Library system

Figure 1: Student behavior analysis and management goals.

As shown in Figure 1, first, the historical data stored in the school's digital management systems need to be integrated and stored. Second, establish a data analysis model based on the dimensions of student behavior data analysis. Third, establish a student behavior data analysis system to achieve student management goals based on data services. Finally, a real-time monitoring system for student behavior is established to monitor abnormal students in real time.

2.2. Student Behavior Analysis Process Based on Machine Learning. The process of applying machine learning to student behavior analysis is shown in Figure 2. As shown in Figure 2, it is first necessary to collect and preprocess student behavior data. Collection is by exporting system data such as the school's teaching system and campus card. The format and structure of the exported data vary greatly. Various data needs to be integrated to obtain comprehensive data. Second, because there will be a lot of noise data in the integrated data, the data needs to be preprocessed. The preprocessed data often has problems such as high dimensionality, so it is necessary to perform feature extraction on the data. Third, conduct behavior analysis model training based on the training set. Fourth, input the test data into the trained analysis model to obtain the analysis result. Finally, perform related management and application based on the analysis results.

3. Student Behavior Analysis Based on Adaptive *K*-Means

3.1. Behavior Analysis Framework. As shown in Figure 3, first, integrate the data collected by the campus card, educational administration system, etc. These data are mainly composed of students' life, consumption, and learning data. Second, because there is noisy data in the integrated data, it is necessary to perform preprocessing such as cleaning the data. Third, perform feature extraction on the preprocessed data and extract the main features for subsequent processing. The feature extraction method used in this paper is principal component analysis (PCA) [27, 28]. Fourth, divide the feature data set into a training set and a test set. The training set is used to train the behavior analysis model. Fifth, the test set is input to the analysis model to obtain the analysis result. Finally, the analysis results are managed and applied.

3.2. Behavior Analysis Model. The analysis model used in Figure 3 is an adaptive *K*-means algorithm. The idea of the algorithm used is to optimize the sample data set X through the elbow rule to determine outliers. When the algorithm is implemented, the sample data set that eliminates outliers is used, and after the algorithm is completed, the final outlier is determined according to the similarity between the outliers and each cluster. Based on the adaptive idea, after each iteration is completed, the value of k is automatically adjusted according to the cluster evaluation index error of each cluster until the error range is met.

3.2.1. The Elbow Rule Detects Outliers. The similarity determination of traditional K-means algorithm is based on Euclidean distance. Outliers will affect the estimation of k value, thereby increasing the time complexity of the algorithm. Use the elbow method to effectively detect outliers in the data set to optimize the algorithm. The specific implementation is as follows:

Let the data set be $X = \{x_i \mid i = 1, 2, \cdots, m\}$ and m be the number of samples. Each sample has n features $(n > 0)$; the sample is divided into different categories $C = \{c_1, c_2, \cdots, c_k\}$, k is the number of clusters. Initially, all samples in X are regarded as one class, and the initial class center is

$$\mu_j = \frac{1}{N(C_j)} \sum_{x_i \in C_j} x_i, \qquad (1)$$

where C represents the sample set contained in the j-th cluster and $N(C_j)$ represents the number of samples in the

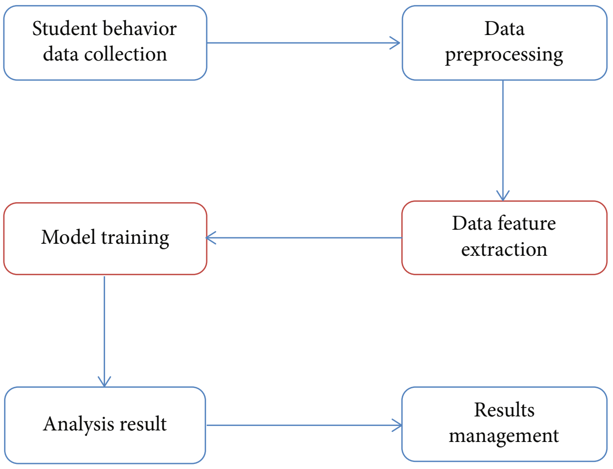

Figure 2: Flow chart of student behavior analysis based on machine learning.

j-th cluster. Calculate the Euclidean distance set D from each sample in X to the cluster center μ_j; the specific formula is as follows:

$$d\left(x_i, \mu_j\right) = \sqrt{\left(x_i - \mu_j\right)^2}, \tag{2}$$

where x_i represents the i-th sample, $i \in [1, m]$, $D = \{d_1, d_2, \cdots, d_m\}$.

According to the elbow rule, sort the data in D from small to large and get an x-d two-dimensional line graph, where d is the distance from the sample data point to the center point and x is the sample corresponding to d. As d increases, the value corresponding to the position where the distortion improvement effect increases the most is the elbow. Therefore, the elbow meets the following conditions:

$$\Delta d = \max\left(d_i - d_{i-1}\right), i \in [1, m]. \tag{3}$$

Divide at the elbow and temporarily define data whose distance is greater than the corresponding distance of the elbow point as an outlier. Assuming there are w samples $(w \leq m)$ after eliminating outliers, there are $(m\text{-}w)$ outliers. Store the nonoutliers in the sample set X' and renumber the samples to get $X' = \{x_1, x_2, \cdots, x_w\}$. The outliers are stored in the data set Y and renumbered to obtain $Y' = \{x_{w+1}, x_{w+2}, \cdots, x_m\}$. The use of X' sample set in the implementation of the algorithm can eliminate the influence of outliers to a certain extent.

After the algorithm is implemented, cluster $C = \{c_1, c_2, \cdots, c_z\}$ is obtained, z is the number of clusters, and the maximum distance from the sample in the cluster to the center of the cluster in the j-th cluster is $\max d_j$.

$$\max d_j = \max\left\{d_{j1}, d_{j2}, \cdots\right\}. \tag{4}$$

According to Equation (2), calculate the distance $\{(d_{(w+1)1}, d_{(w+1)2}, \cdots, d_{(w+1)z}), (d_{(w+2)1}, d_{(w+2)2}, \cdots, d_{(w+2)z}), \cdots, (d_{m1}, d_{m2}, \cdots, d_{mz})\}$ from the sample in Y to the center of each cluster. If there is a cluster $(a \in (0, z))$, make the sample $x_b (b \in [w+1, m])$ in the set Y satisfy that the distance from x_b to the center of cluster a is less than the maximum distance from the samples in the cluster to the center of the cluster, namely, $(d_{b1} < \max d_1) \| (d_{b2} < \max d_2) \| \cdots \| (d_{ba} < \max d_a)$. Then, divide x_b into the nearest cluster among a clusters.

$$j(x_b) = \min\left(d_{b1}, d_{b2}, \cdots, d_{ba}\right). \tag{5}$$

If there is no such cluster, the sample is defined as an outlier. Until all the samples in Y are traversed, the final outliers can be divided.

3.2.2. Selection of Adaptive k Value. Appropriate selection of k value needs to be evaluated based on clustering evaluation index. This article uses the sum of squares of errors SSE within the cluster. The calculation formula of this indicator is as follows:

$$E = \sum_{j=1}^{k} \sum_{x \in c_j} \left|x - \mu_j\right|^2, \tag{6}$$

where k represents the number of clusters, x represents samples, μ_j represents the cluster center of the j-th cluster, and C_j represents the set of samples contained in the j-th cluster. E describes the tightness of each cluster sample to a certain extent; the smaller the E, the better the clustering effect. According to the SSE, the sum of squared errors Je in each cluster is obtained. The calculation formula of Je is as follows:

$$Je_j = \sum_{x \in C_j} \frac{1}{N\left(C_j\right)} \left|x_i - \mu_j\right|^2, \tag{7}$$

where x_i is the sample in the j-th cluster, $N(C_j)$ is the number of samples in the j-th cluster, and μ_j is the sample in the j-th cluster. The smaller Je_j means the better the clustering effect of the j-th cluster.

Initially, set Je and the threshold N of the minimum number of samples in the cluster. After each cluster is divided, the number of samples $(N_1, N_2, \cdots, N_k)$ in each cluster can be obtained, and the $(Je_1, Je_2, \cdots, Je_k)$ of each cluster can be calculated according to Equation (7). Then, calculate the error ΔJe_j of the j-th cluster clustering evaluation index and the difference ΔN_j between the number of samples in the cluster and the initial value. The specific formula is as follows:

$$\Delta Je_j = Je_j - b, \tag{8}$$

$$\Delta N_j = N_j - N. \tag{9}$$

Combine Equation (8) and Equation (9) to get the change

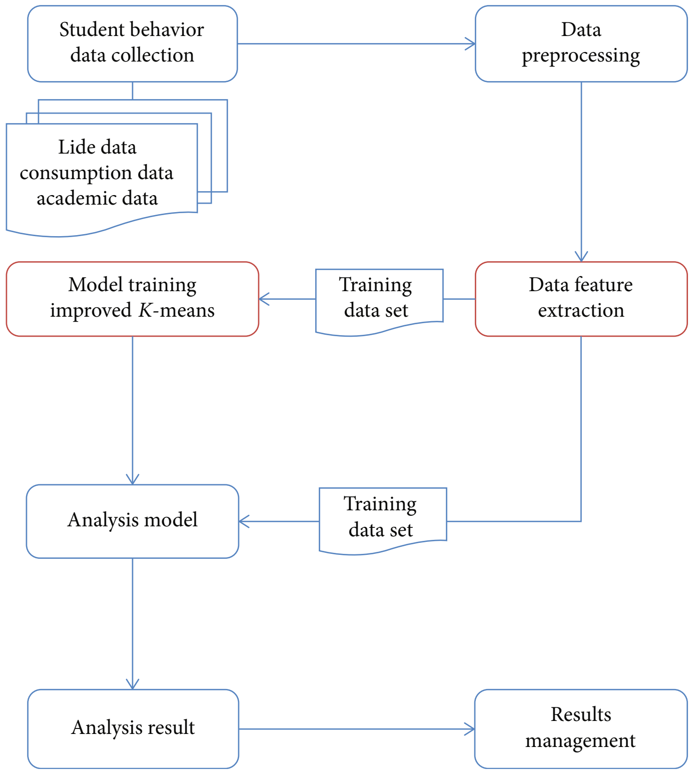

Figure 3: Architecture diagram of student behavior analysis.

of k value in the j-th cluster

$$\Delta k_j = \frac{\mathrm{sgn}\ (\Delta N_j) - 1}{2} + \frac{\mathrm{sgn}\ (\Delta N_j) + 1}{2} * \Pi(\log_w(\Delta Je_j + 1))\theta(\Delta Je_j), \tag{10}$$

where w is the number of samples in the data setX', sgn () is the symbolic function, $\theta()$ is the unit step function, and the symbol $\Pi()$ is rounded up.

If $N_j < N$, then ΔN_j is negative, sgn $(\Delta N_j) = -1$, so(sgn $(\Delta N_j) + 1)/2 = 0$, (sgn (ΔN_j)-1$)/2 = -1$. $\Delta k_j = -1$ represents that when the number of samples in the j-th cluster is less than the initial value, delete the cluster center of the cluster.

If $N_j > N$, then ΔN_j is a positive number, sgn $(\Delta N_j) = 1$, (sgn $(\Delta N_j) + 1)/2 = 1$, (sgn (ΔN_j)-1$)/2 = 0$. The discussion is divided into the following two situations:

(1) When $Je_j > b$, then Je_j is a positive number, $\theta(Je_j) = 0$, and $\Delta k_j = \Pi(\log_m(\Delta Je_j + 1))$, so $\Delta k_j > 0$. It is necessary to add a new cluster center near the j-th cluster center to reduce the error evaluation index within the cluster. Generally, $0 < \log_m(\Delta Je_j + 1) < 1$, then$\Delta k_j = 1$. Only when the error $\Delta(Je_j)$ is particularly large, $\log_m(\Delta Je_j + 1)$ will be greater than 1

(2) When $Je_j < b$, then Je_j is negative, $\theta(Je_j) = 0$, and $\Delta k_j = 1$. $\Delta k_j > 0$ requires a new cluster center near the j-th cluster center to reduce the error evaluation index within the cluster. Generally, $0 < \log_m(\Delta Je_j + 1) < 1$, then $\Delta k_j = 0$. When the cluster evaluation index in the j-th cluster is less than the set initial value, the cluster center of the cluster is neither deleted nor added. The sample closest to the j-th cluster center is set as the new cluster center to reduce the clustering evaluation index within the cluster. After traversing each cluster, according to Equation (10), the amount of change in k value$(\Delta k_1, \Delta k_2, \cdots, \Delta k_k)$ is obtained. The updated k' is

$$k' = k + \sum_{j=1}^{k} \Delta k_j. \tag{11}$$

If $k' = k$, terminate the loop; if $k' \neq k$, continue the loop. The flow of the algorithm is shown in Figure 4.

The specific implementation steps of the algorithm are as follows:

Step 1. Data optimization processing. Find outliers according to the elbow rule and store them in sample set Y. The new sample data set after optimization is $X' = \{x_1, x_2, \cdots, x_w\}$.

Step 2. Set the threshold range of the evaluation index Je in a single cluster and the minimum number of samples N in the cluster.

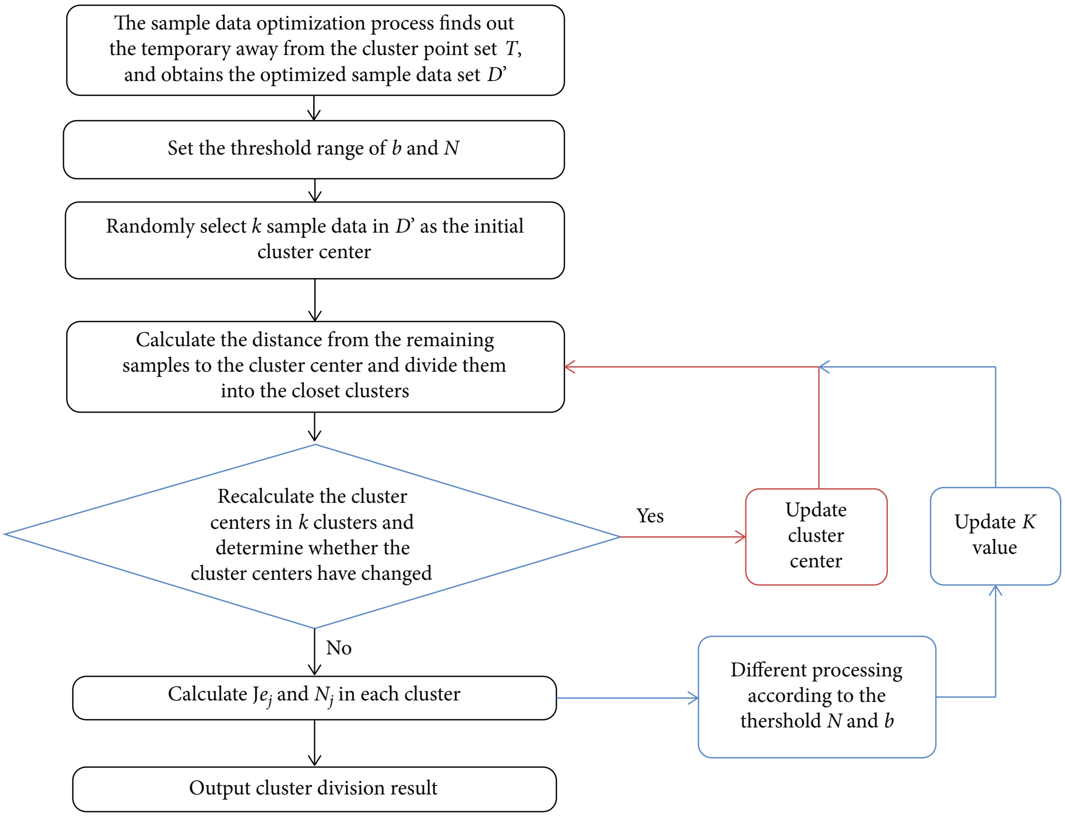
The sample data optimization process finds out the temporary away from the cluster point set T, and obtains the optimized sample data set D'
Set the threshold range of b and N
Randomly select k sample data in D' as the initial cluster center
Calculate the distance from the remaining samples to the cluster center and divide them into the closet clusters
Recalculate the cluster centers in k clusters and determine whether the cluster centers have changed
Yes
Update cluster center
Update K value
No
Calculate Je_j and N_j in each cluster
Different processing according to the thershold N and b
Output cluster division result

Figure 4: Adaptive K-means algorithm flow.

Step 3. Randomly select k samples from w samples as the initial cluster center $1 < k \leq w$.

Step 4. Calculate the Euclidean distance of each remaining sample to the cluster center according to Equation (2) and divide each sample into the cluster closest to it.

Step 5. Recalculate the new cluster center of each cluster according to Equation (1).

Step 6. If the new cluster center is the same as the original center or less than a certain threshold, the iteration is terminated. If the new cluster center changes, continue to repeat Steps 4 and 5 until convergence.

Step 7. Calculate the Je_j of each cluster and the number of samples N in each cluster. Compare Je_j and N_j with the initial threshold range and calculate Δk_j according to Equation (10). After traversing all the clusters, calculate the new k value k' according to Equation (11). If any center is deleted or added, return to Step 4 until there are no new or deleted cluster centers.

Step 8. Determine the final divided outliers according to the similarity between the samples in the sample set Y and the cluster centers of each cluster.

Step 9. Output the final cluster partition $C = \{c_1, c_2, \cdots . c_z\}$.

4. Experiment and Analysis

4.1. Experimental Background. The data used in this article comes from data in a university's digital system database. The consumption data comes from the campus all-in-one card system, which mainly includes canteen consumption and school supermarket consumption. The life data comes from the all-in-one card system, which mainly includes exercise clock-in and time spent online. The learning data comes from the educational administration system and the book borrowing system. A total of 2017 students were selected as a sample. The extracted raw data mainly contains 89,682 pieces of consumption data, 49,860 pieces of life data, and 15,629 pieces of learning data. After the integration and preprocessing of the data, sample data of 3356 students were obtained as experimental data. Table 1 is the definition of students' consumption habits, Table 2 is the definition of students' living habits, and Table 3 is the definition of students' learning habits.

The hardware configuration information used in this experiment is as follows: CPU is Intel Core i7, graphics memory is GTX960M 4G, and memory is 16G. The operating system is Windows 10 64-bit, and the development language is MATLAB.

4.2. Experimental Results and Analysis

4.2.1. Analysis of Consumption Data. The results of cluster analysis of consumption data based on the algorithm used in this article are shown in Table 4.

Table 1: Student consumption habit indicators.

Index	Ranges	Index description
Average monthly consumption	0-2000	The total consumption per student per semester divided by the number of semester months
Average monthly consumption frequency	0-500	The sum of consumption times per student per semester divided by the number of semester months
Peak monthly consumption	0-unknown	Peak consumption of each student in the months of each semester

Table 2: Student living habit indicators.

Index	Ranges	Index description
Dining habits	0-90	The average number of days the student eats regularly per month in the semester
Work and rest	0-30	The average number of days of regular work and rest per month for students in each semester
Internet habits	0-600	Average online time per month per semester by students
Exercise habits	0-60	The average number of student exercises per month per semester

Table 3: Student study habit indicators.

Index	Ranges	Index description
Attend class	0-1	Number of student attendance in class
Book borrowing	0-unknown	Number of books borrowed by students per semester
	0-unknown	Number of times the student enters the library per semester

Table 4: Cluster analysis results of student consumption data.

Number	Student ratio	Monthly consumption	Monthly consumption	Peak monthly consumption
1	10.78	150.23	586.56	652.14
2	25.16	136.52	1091.82	1356.60
3	34.51	112.10	829.75	928.03
4	16.45	77.79	678.50	788.50
5	13.10	45.97	467.92	600.31
Mean		104.52	730.91	865.12

It can be seen from Table 4 that the most suitable k value obtained based on the K-means clustering algorithm used in this article is 5, which shows that 5 types of consumption habits can be obtained according to the consumption data of students. Each habit corresponds to different types of students. The characteristics of each type of student are as follows:

(1) Type 1 students have lower monthly consumption levels and single-month peak consumption, but they consume more times. This type belongs to the group with lower consumption levels. Such students have poor family economic conditions and live frugal lives. It is recommended that school administrators pay attention to the living conditions of such students, and the identification and funding of poor students can consider choosing from such students

(2) The average monthly consumption of type 2 students is the highest, the number of consumption is high, and the consumption peak is also high, indicating that this type of student is a high consumption group in the cafeteria

(3) The monthly consumption level of type 3 students is above average, and the consumption frequency is higher each month, and the consumption amount is stable. It shows that such students often eat in the school cafeteria, and their consumption is stable, which is in line with the normal eating rules of most students in school

(4) The monthly consumption level of type 4 students is in the middle, and the average monthly consumption is not high. However, the maximum consumption in a single month is relatively high, and the number of consumptions is also relatively small. Such students eat irregularly in the cafeteria and usually like to eat out of school or order takeaways

(5) The average monthly consumption of type 5 students is relatively low, the number of times is relatively small, and the maximum consumption is not high. Such students do not consume frequently in the cafeteria and are more likely to eat outside of school and consume more outside of school. School administrators should pay attention to the food safety and personal safety of such students

5. Life Data Analysis

Table 5 shows the results of clustering analysis of life data based on the algorithm used in this paper.

Table 5: Cluster analysis results of student life data.

Number	Student ratio	Dining	Work and rest	Internet	Exercise
1	38.82	79.80	21.02	381.66	45.93
2	9.86	28.67	7.16	548.50	6.93
3	51.32	40.48	18.85	420.24	16.40
Mean		49.65	15.68	450.13	23.09

Table 6: Cluster analysis results of student learning data.

Number	Student ratio	Class attendance	Number of books borrowed	Number of visits to the library
1	11.98	99.58	10.37	24.75
2	53.36	94.20	7.80	21.44
3	24.02	90.08	8.22	15.03
4	10.64	96.57	5.67	10.47
Mean		95.11	8.02	17.92

For the cluster analysis of life data, three categories were gathered. According to the information in Table 5, the following inference can be drawn.

(1) Type 1 students have regular schedules and meals every month. They spend a long time online and often participate in physical exercises. Such students have strong self-discipline, good physical fitness, and good living habits

(2) Type 2 students overslept more frequently each month. They eat irregularly in the cafeteria, spend a long time online, and exercise less frequently. Such students have poor physical fitness. School administrators should pay attention to the learning and class conditions of such students and whether they often skip classes

(3) Type 3 students often get up early every month, but they have irregular meals in the cafeteria, spend more time online, and do less physical exercise. Such students should not have a healthy habit of eating breakfast and do not like to exercise. School administrators should urge such students to change their unhealthy living habits and pay attention to their health

6. Learning Data Analysis

The results of cluster analysis of academic data based on the algorithm used in this paper are shown in Table 6.

Based on the K-means algorithm used in this article to analyze the learning data, students are divided into 4 categories. Based on the information shown in Table 6, the characteristics of each type of student are as follows:

(1) Type 1 students' classroom attendance, the number of books borrowed in the library, and the number of library entrances are all high. Such students study hard and have good study habits

(2) Type 2 students have a higher class attendance rate, but the number of books borrowed is not many, and they enter the library more often. Such students are active in class and often read and study in the study room and library. However, the small amount of books they borrow in the library indicates that such students are accustomed to reading books in the library

(3) Type 3 students have a low class attendance rate, a small amount of books borrowed, and a small number of in and out of the library. It shows that such students often skip classes and do not study hard enough. They are students who do not like to learn. School administrators should focus on the learning situation of such students and promptly urge them to form good learning habits

(4) Type 4 students have an average class attendance rate, fewer books borrowed, and fewer trips to the library. This kind of students just go to class often, and they do not have much time to study after class, and the degree of study hard is not high. This type of overstudy is a type that does not pay much attention to study at ordinary times and makes surprise review before the exam. It is recommended that such students develop regular study habits and arrange their study time reasonably and appropriately

7. Conclusion

In order to assist the school to improve the management level of students, this paper uses an adaptive K-means clustering algorithm to analyze the behavior characteristics of students. By collecting the digital system data of the campus, students' consumption, life, and learning data can be obtained. After the data is preprocessed, PCA is used for feature extraction to obtain feature data and perform model training. Input the test set into the trained model to get the clustering result. Finally, based on the analysis of the clustering results, the characteristics of all kinds of students are obtained. According to the results of consumption data analysis, five types of consumer groups were obtained. For groups with low consumption levels, more consideration can be given to poor student evaluation and work-study programs. According to the analysis results of life data, three groups are obtained. Teachers should pay special attention to groups that eat irregularly, spend a long time online, and do not exercise regularly. According to the analysis results of the learning data, 4 groups of groups are obtained. Teachers need to pay more attention to students with average attendance rate, low book reading volume, and small number of library entrances. In the management process of the school, students with different characteristics can be taught in accordance with their aptitude, thereby improving the management quality of the school. There are still some shortcomings in this article; for example, there are some limitations in the description of student behavior characteristics. In the analysis of student behavior characteristics, due to the limited data sources in the digital campus platform, it is not possible to fully reflect the behavior characteristics of students in school. Later, as the application of the school's digital campus is perfected, more campus businesses will be transferred from traditional offline to online, and more comprehensive student data can be collected.

Data Availability

The labeled dataset used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported in part by the National Social Science Fund Project under Grant 17BGL102, the Excellent Project of Jiangsu Province Social Science Union under Grant 15SYC-043, the Soft Science Research of Wuxi Science and Technology Association under Grant KX15-B-01, and the Fundamental Research Funds for the Central Universities under Grant 2015ZX18.

References

[1] Y. Cao, J. Gao, D. Lian et al., "Orderness predicts academic performance: behavioral analysis on campus lifestyle," *Journal of the Royal Society Interface*, vol. 15, no. 146, article 20180210, 2017.

[2] Y. Ma, C. Cui, X. Nie, G. Yang, K. Shaheed, and Y. Yin, "Precourse student performance prediction with multi-instance multi-label learning," *Science China Information Sciences*, vol. 62, no. 2, pp. 200–205, 2019.

[3] M. Zhou, M. Ma, Y. Zhang, K. SuiA, D. Pei, and T. Moscibroda, "EDUM: classroom education measurements via large-scale WiFi networks," in *ACM International Joint Conference on Pervasive & Ubiquitous Computing*, Heidelberg, Germany, 2016.

[4] Y. Su, Q. Liu, Q. Liu et al., "Exercise-enhanced sequential modeling for student performance prediction," in *Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA, 2018.

[5] Y. Zhang and W. Jiang, "Score prediction model of MOOCs learners based on neural network," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 13, no. 10, pp. 171–182, 2018.

[6] K. Casey and D. Azcona, "Utilizing student activity patterns to predict performance," *International Journal of Educational Technology in Higher Education*, vol. 14, no. 1, p. 4, 2017.

[7] J. W. You, "Identifying significant indicators using LMS data to predict course achievement in online learning," *The Internet and Higher Education*, vol. 29, pp. 23–30, 2016.

[8] D. D. Mitri, M. Scheffel, H. Drachsler, D. Börner, S. Ternier, and M. Specht, "Learning pulse: a machine learning approach for predicting performance in self-regulated learning using multimodal data," in *The Seventh International Learning Analytics & Knowledge Conference*, Vancouver, British Columbia, Canada, 2017.

[9] M. Sweeney, H. Rangwala, J. Lester, and A. Johri, "Next-term student performance prediction: a recommender systems approach," *Journal of Educational Data Mining*, vol. 8, no. 1, pp. 22–51, 2016.

[10] G. Li, "Big data related technologies, challenges and future prospects," *Information Technology & Tourism*, vol. 15, no. 3, pp. 283–285, 2015.

[11] P. Anand, "Big data is a big deal," *Journal of Petroleum Technology*, vol. 65, no. 4, pp. 18–21, 2015.

[12] C. H. Miller, M. D. Sacchet, and I. H. Gotlib, "Support vector machines and affective science," *Emotion Review*, vol. 12, no. 4, pp. 297–308, 2020.

[13] S. Kim and C. Kim, "Influence diagnostics in support vector machines," *Journal of the Korean Statistical Society*, vol. 49, no. 3, pp. 757–778, 2020.

[14] J. A. Cook and S. Siddiqui, "Random forests and selected samples," *Bulletin of Economic Research*, vol. 72, no. 3, pp. 272–287, 2020.

[15] S. Gil-Begue, C. Bielza, and P. Larrañaga, "Multi-dimensional Bayesian network classifiers: a survey," *Artificial Intelligence Review*, vol. 54, no. 1, pp. 519–559, 2021.

[16] J. Barr, M. Littman, and M. desJardins, "Decision trees," *ACM Inroads*, vol. 10, no. 3, p. 56, 2019.

[17] Y. P. Zhang, S. H. Wang, K. J. Xia, Y. Z. Jiang, and Q. J. Qian, "Alzheimer's disease multiclass diagnosis via multimodal neuroimaging embedding feature selection and fusion," *Information Fusion*, vol. 66, pp. 170–183, 2021.

[18] Y. Jiang, X. Gu, D. Wu et al., "A novel negative-transfer-resistant fuzzy clustering model with a shared cross-domain transfer latent space and its application to brain CT image segmentation," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 1, pp. 1–52, 2020.

[19] Y. Jiang, Y. Zhang, C. Lin, D. Wu, and C.-T. Lin, "EEG-based driver drowsiness estimation using an online multi-view and transfer TSK fuzzy system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1752–1764, 2021.

[20] Y. Zhang, Y. Jiang, L. Qi, M. Z. A. Bhuiyan, and P. Qian, "Epilepsy diagnosis using multi-view & multi-medoid entropy-based clustering with privacy protection," *ACM Transactions on Internet Technology*, vol. 21, no. 2, pp. 1–21, 2021.

[21] X. Wang, A. Bao, Y. Cheng, and Q. Yu, "Multipath ensemble convolutional neural network," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 2, pp. 298–306, 2021.

[22] G. Dzhezyan and H. Cecotti, "Symmetrical filters in convolutional neural networks," *International Journal of Machine Learning and Cybernetics*, vol. 12, no. 7, pp. 2027–2039, 2021.

[23] M. E. Valle and R. A. Lobo, "Hypercomplex-valued recurrent correlation neural networks," *Neurocomputing*, vol. 432, pp. 111–123, 2021.

[24] M. Alameh, Y. Abbass, A. Ibrahim, G. Moser, and M. Valle, "Touch modality classification using recurrent neural networks," *IEEE Sensors Journal*, vol. 21, no. 8, pp. 9983–9993, 2021.

[25] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pp. 281–297, Berkeley, California, USA, 1967.

[26] Z. Zhang, Q. Feng, J. Huang, Y. Guo, J. Xu, and J. Wang, "A local search algorithm for k-means with outliers," *Neurocomputing*, vol. 450, pp. 230–241, 2021.

[27] K. Sando and H. Hino, "Modal principal component analysis," *Neural Computation*, vol. 32, no. 10, pp. 1–35, 2020.

[28] A. Charpentier, S. Mussard, and T. Ouraga, "Principal component analysis: a generalized Gini approach," *European Journal of Operational Research*, vol. 194, no. 1, pp. 236–249, 2021.