International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-20

https://doi.org/10.70917/ijcisim-2026-0164
Article

Research on the Application of Multidimensional Data Mining Algorithms in Intelligent Management of Higher Education Institutions

Ye Zhang *

Editorial Department of Journal, Yancheng Teachers University, Yancheng, Jiangsu, 224007, China
yczyi39@163.com

Abstract: Based on the campus card data of students at School A's Business School, this paper conducts data mining on student behavior from three dimensions: dining consumption levels, living patterns, and diligence in studying. An improved K-means algorithm is used to classify student behavior data. Then, conduct a correlation analysis of academic performance and behavioral indicators to identify patterns between behavior and academic performance. Use these insights to guide students in improving their poor behavioral habits and enhancing their learning efficiency, thereby improving their academic performance. This paper also improves the Apriori algorithm in association rules and compares it with other algorithms through experiments to verify the feasibility and high efficiency of the improved algorithm. The improved algorithm is then applied to student behavior analysis, revealing the correlation between student behavior habits and academic performance. The main factors influencing academic performance are identified as lifestyle behavior and learning behavior, while consumption behavior has a relatively minor impact on academic performance.

Keywords: data mining; clustering analysis; Apriori algorithm; intelligent management in higher education institutions

1. Introduction

In recent years, with the continuous advancement and development of information technology, the concept of smart management has gradually emerged in the management processes of various universities [1-2]. Fully implementing and developing this concept can significantly enhance the effectiveness and quality of university management [3-4]. University management is based on information, and to facilitate the efficient processing of faculty and student information data and ensure the long-term smooth operation of management systems, a robust data mining platform is essential to guarantee the security and controllability of large volumes of information data [5]. Today, electronic information data is more comprehensive, and many universities have adopted application systems that meticulously analyze data elements to uniformly consolidate faculty and student information data, significantly enhancing their interconnectivity [6-8].

In this context, the management of student learning information data has become a critical task for universities. Lei et al. optimized the university intelligent management system using IoT and intelligent algorithms. Test results showed that the optimized system had a shorter average response time compared to the traditional system, enabling effective input, classification, and management of student information to facilitate student information grading [9]. Li developed an intelligent campus management system based on IoT technology, which collects data through facial recognition terminals and manages and analyzes it through an intelligent system, enabling the formulation of more efficient learning, teaching, and research plans for faculty and students. The satisfaction rating for the system's practicality was 8.0 [10]. Wu et al. designed a big data technology-based intelligent management platform system for industry-education collaboration using support vector machines and data mining algorithms. The results indicated that big data algorithms improved the system's classification accuracy and reduced its response

time, thereby promoting the comprehensive advancement of the industry-education collaboration system [11].

School administrative departments selected data mining algorithms for application in decision-making analysis data to achieve more efficient data management. Zhang et al. studied the application of data mining technology in the development of university information management systems. The results of applying data mining methods in university information management demonstrated that it effectively enhanced the data analysis capabilities of administrative staff and improved the management level of universities [12]. Sun utilized an improved clustering algorithm to mine student performance data, and by comparing evaluation results, validated the rationality and feasibility of using clustering algorithms for university education management. The study concluded that integrating data mining technology into university academic management is meaningful [13]. Hu et al. developed and optimized a university management system based on data mining technology to improve educational management efficiency and address issues of uneven course resource allocation. This system can collect data structures from databases and extract evaluation indicators for higher education management systems, experimental results showed that the system maintains good scheduling performance and provides better teaching services for teachers and students [14]. Natek et al. used different data mining tools to predict student success rates in course learning and studied the correlation between student characteristics and success rates [15]. Zheng utilized an improved neural network for data mining of educational management theories, conducting a comprehensive systematic evaluation of educational management theories. The results indicated that the improved neural network algorithm achieved quality evaluation prediction for educational management theories [16]. Wei et al. designed an IoT-enabled intelligent educational management system, combining data mining algorithms for real-time classroom visualization, enabling learning status statistics, and providing users with personalized learning plans. Experimental results indicate that the system can effectively complete management tasks under real-time data streams, demonstrating strong practicality [17]. The intelligent management of higher education institutions is playing an increasingly significant role in teaching and management processes, becoming an inevitable trend for efficient information management in the future. Data mining technology plays a crucial role in this process.

This paper implements intelligent management in higher education institutions through the following steps: student behavior mining, behavior classification, and analysis of the correlation between behavior and academic performance. First, comprehensive data governance is performed on the integrated campus data using R language technology. Then, addressing issues such as the high computational complexity of the iterative process of the original K-means clustering algorithm, the need for manual determination of the number of clusters, sensitivity to initial cluster centers, and susceptibility to outliers, an improved K-means algorithm is proposed that combines multiple factors based on an optimized outlier detection algorithm, the maximum-minimum distance principle, and heuristic methods. Next, the interest-based Apriori algorithm is introduced, with $Conf(\overline{X} \Rightarrow Y)$ corrections applied to the interest-based metric based on differences. Finally, the aforementioned methods and steps are applied to analyze University A as an example.

2. Application of Data Mining Algorithms in Intelligent Management of Higher Education Institutions

2.1. Research on Data Mining Algorithms

Data mining, in simple terms, is the process of extracting knowledge from large amounts of data. More specifically, data mining is the process of extracting information that is not immediately apparent, previously unknown, implicit, yet valuable from massive, noisy, ambiguous, incomplete, and random data [18].

The data mining process primarily includes:

(1) Data collection

By analyzing the research content and related data, the required target data is identified, and the necessary data is extracted from the data sources.

(2) Data cleaning

Due to the complexity of raw data, it may contain unnecessary or erroneous data, which can impact the results of data mining. Therefore, data cleaning is required to remove noise, delete unnecessary or erroneous data.

(3) Data integration

During the data mining process, the required data may come from different data sources. Combining data from multiple sources facilitates future data management and operations.

(4) Data transformation

Data transformation is the process of converting the current data format into the format required for mining research to better utilize and mine the data. Commonly used data transformation methods include data normalization, aggregation, attribute construction, and generalization.

(5) Data mining

Based on the nature of the target data for knowledge representation, appropriate data mining algorithms are selected. Intelligent methods are employed to extract data patterns and construct models, thereby uncovering latent patterns and knowledge from the data.

(6) Knowledge Representation

After the aforementioned processes, the latent knowledge extracted from the data is represented using visualization or knowledge representation techniques. This clearly presents the knowledge to users, providing them with decision support.

2.1.1. Data mining-related technologies

Data mining in application-driven fields incorporates technologies from many application areas, such as machine learning, statistics, databases, pattern recognition, information retrieval, and visualization.

Statistics mainly studies the collection, interpretation, analysis, and representation of data. Data mining and statistics are closely related. Statistical models are generally mathematical functions that use probability distributions and random variables to describe the behavior of the objects being studied.

Machine learning is the study of computers learning from data and is a technology aimed at improving computer performance. It primarily focuses on enabling computer programs to automatically learn and recognize complex patterns based on data and make intelligent judgments. Machine learning is primarily divided into supervised learning, unsupervised learning, semi-supervised learning, and active learning.

A data warehouse integrates data from multiple sources and different time periods. It primarily merges data in a multidimensional space to form data cubes. This facilitates OLAP in multidimensional databases and drives the development of multidimensional data mining.

Information retrieval is the science of searching documents. Documents can take various forms, such as text or multimedia, and may be hosted on the Web. Compared to traditional information retrieval, database systems have two key differences: the data being searched in information retrieval is unstructured. Information retrieval queries primarily search for keywords without complex structures, while database systems use SQL statements for queries. Probabilistic models are a classic method in information retrieval.

2.1.2. Commonly used data analysis tools

(1) Python Language

Python is an object-oriented, interpreted high-level computer programming language. Its object-oriented features enable advanced concepts such as multiple inheritance, operator overloading, and support for polymorphism, making it an ideal scripting tool for high-level languages like Java and C++. Python has a clear and concise syntax, a powerful and extensive library, and is an open-source, purely free software. It is also known as a glue language, as it can effectively connect modules from other scripting languages.

(2) R Language

R, commonly abbreviated as R, is defined as a language environment that can be freely and effectively utilized for statistical computing and graphics. It is primarily used in various data analysis fields such as statistical analysis and data mining.

(3) MATLAB

MATLAB focuses more on mathematical operations, integrating powerful features such as matrix computation, numerical analysis, modeling and simulation of nonlinear dynamic systems, and scientific data visualization into an easy-to-use graphical user interface.

(4) SAS

SAS is a data mining product developed by SAS Inc., which can be integrated with OLAP and SAS data warehouses to achieve “end-to-end” knowledge discovery from data extraction, data capture, to obtaining results. It can also use graphical modules to combine data mining units into process flow diagrams and build data mining processes based on them.

(5) Weak

Weak is an open-source data analysis software developed using the Java programming language. It is one of the most well-known data analysis software in the open-source field and one of the most widely used open-source tools in machine learning.

(6) Hadoop

Hadoop is a distributed computing platform that provides users with detailed underlying details and a transparent distributed infrastructure. HDFS (Hadoop Distributed File System) and MapReduce are its core technologies. HDFS is designed for storing massive amounts of data, while MapReduce is a distributed computing model primarily used for computation. The results obtained after data processing are stored and managed using HBase.

2.2. Improvements to the K-means clustering algorithm

In the process of intelligent management in higher education institutions, considering that the student behavior data obtained is continuous data with relatively small dimensionality and numerical values, and given the various advantages of the K-means algorithm, this paper selects the K-means algorithm for cluster analysis of student behavior. To address the limitations of the K-means algorithm, this paper proposes an improved clustering algorithm that combines multi-point optimization from four aspects. Experimental results on public datasets demonstrate that the proposed improved algorithm achieves excellent clustering performance and good computational efficiency.

2.2.1. Ideas for improving the K-means algorithm

Optimizing the K-means clustering algorithm involves considering multiple factors, primarily focusing on two aspects: parameter determination and algorithm processing steps. Specifically, this includes four key areas: determining the number of clusters, allocating sample points during iteration, detecting outliers, and selecting initial cluster centers. Finally, the improved algorithm is described in detail, and its time complexity is analyzed. First, a combination of the contour coefficient method and the elbow rule is used to determine the number of clusters. Then, for the allocation of sample points during the iteration process, a heuristic method is employed to reduce the number of calculations required to compute the distances between sample points and cluster centers. Second, for outlier detection, the CLOF algorithm is adopted to improve the efficiency of the outlier detection algorithm. First, a heuristic method is used to optimize the selected clustering algorithm, and then the optimized outlier detection algorithm is applied for detection. Finally, the initial cluster centers are gradually determined based on the outlier candidate set generated by outlier detection and the maximum-minimum distance principle.

2.2.2. Determining the number of clusters

The determination of the number of clusters has a significant impact on clustering results; however, relying on experience or dataset characteristics to determine the number of clusters lacks accuracy. There are two common methods for determining the number of clusters: the elbow rule and the contour coefficient method.

(1) Elbow Rule Analysis

The elbow rule uses the sum of squared errors (SSE) as its cost function. As the number of clusters increases, each sample point is processed more precisely, and the sample clusters become more cohesive, resulting in a decrease in the SSE value. Additionally, before reaching the optimal number of clusters, as the number of clusters increases, the increase in cluster cohesion is significant, and the decrease in SSE is also substantial. When the number of clusters reaches the optimal number, the rate of increase in the cohesion of sample clusters decreases, and the rate of decrease in the sum of squared errors undergoes a sharp change, after which it becomes relatively flat. The graph of SSE versus k takes an elbow shape, and the point where the SSE changes sharply is the optimal number of clusters. The definition of the sum of squared errors (SSE) is shown in Formula (1):

$$SSE = \sum_{i=1}^{K} \sum_{p \in c_i} \left| p - m_i \right|^2 \tag{1}$$

In this context, K represents the number of clusters, C_i denotes the i th cluster, p is a sample point within cluster C_i, and m_i is the centroid of cluster C_i.

(2) Contour coefficient method analysis

The effectiveness of clustering is generally measured based on two aspects: intra-cluster and inter-cluster. When the intra-cluster distance is minimized and the inter-cluster distance is maximized, the clustering result is considered an ideal clustering effect, i.e., it has the minimum intra-cluster cohesion and the maximum inter-cluster separation.

For any sample point X_i, its cluster is C_i. The average distance between sample point X_i and other

sample points X_j in the same cluster is called the intra-cluster cohesion of sample point X_i, denoted as a_i, defined as shown in formula (2):

$$a_i = \frac{1}{num(C_i)-1} \sum_{C_j=C_i; j\neq i; X_j} dist\left(X_i, X_j\right) \tag{2}$$

Where $num(C_i)$ denotes the number of samples in cluster C_i.

Select a cluster C that does not contain the sample point X_i, calculate the average distance between the sample point X_i and all sample points in that cluster, and perform the same process for other clusters. The smallest average distance is called the inter-cluster separation of the sample point X_i, denoted as b_i, and is defined as shown in formula (3):

$$b_i = \min_{C_q \neq C_i} \left\{ \frac{1}{num(C_q)} \sum_{C_p=C} dist(X_i, X_q) \right\} \tag{3}$$

The contour coefficient of sample point X_i, denoted as s_i, is defined as shown in formula (4):

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{4}$$

The range of values for the contour coefficient s_i is $-1 \le s_i \le 1$. The closer s_i is to -1, the greater the average distance between the sample point X_i and the remaining sample points in the same cluster is compared to the nearest other clusters, indicating that the sample point X_i should not be assigned to this cluster. The closer s_i is to 1, the smaller the average distance between the sample point X_i and the sample points in the same cluster is compared to the nearest other clusters, and the greater the difference between the sample point X_i and the sample points in other clusters, indicating that the sample point X_i is suitable for assignment to this cluster.

If the total number of sample points is n and the number of clusters is k, the average silhouette coefficient is denoted as S_k, defined as shown in Formula (5):

$$S_k = \frac{1}{n} \sum_{i=1}^{n} s_i \tag{5}$$

The average contour coefficient S_k ranges from $-1 \le S_k \le 1$. The larger the average contour coefficient, the better the clustering effect. For different k values, the k corresponding to the maximum value of the average contour coefficient is the optimal number of clusters.

2.2.3. Improvements to the K-means algorithm

(1) Sample point allocation in the iterative process

This study uses a heuristic method to reduce the computational time required for sample point allocation and improve operational efficiency. In each iteration, a simple data structure is used to store the distance between each point and its nearest cluster. In the next iteration, the distance between each point and its previously nearest cluster is first calculated. If the newly calculated distance is less than or equal to the previously stored distance, the point remains in its original cluster, and there is no need to calculate the distance between the point and other clusters, thereby saving the time required to calculate the distances between the point and other clusters.

(2) Outlier Detection

Outlier detection is a widely prevalent problem, and many data mining applications typically use outlier detection as an initial step to filter outliers and establish more representative models. Among various outlier detection methods, the Local Outlier Factor (LOF) detection method is a widely used high-precision outlier detection method [19]. In the LOF algorithm, density is calculated based on the distance between points: the farther the distance between points, the lower the density; the closer the

distance, the higher the density.

(3) Selection of Initial Cluster Centers

The main idea of the maximum-minimum distance algorithm is to set a distance threshold to determine the initial cluster centers. The process of using the maximum-minimum distance algorithm to determine the initial cluster centers is shown in Formula (6):

$$c = \max\left\{Dist(i,j) \mid i = 1,2,\ldots,n; j = 1,2,\ldots,k\right\} \tag{6}$$

Where $Dist(i,j)$ denotes the minimum distance between data object x_i and the existing initial cluster centers. n is the number of samples in the dataset excluding those already selected as initial cluster centers. k is the number of initial cluster centers already selected.

The calculation method for $Dist(i,j)$ is shown in Formula (7):

$$Dist(i,j) = \min\left\{dist(x_i, c_j) \mid xi \in D/C, c_j \in C\right\} \tag{7}$$

In this context, D denotes the dataset. C denotes the predefined initial cluster centers. $dist(x_i, c_j)$ denotes the distance between sample point x_i and initial cluster center c_j.

This study combines the maximum-minimum distance idea with an improved outlier detection algorithm to select the initial cluster centers. First, an optimized outlier detection algorithm is used to detect outliers, simultaneously obtaining a set of outlier candidates and a set of outliers. The outlier set is then removed from the original dataset and temporarily stored to avoid the influence of outliers on the final clustering results. Then, based on the outlier candidate set, the first two initial cluster centers are selected by calculating the pairwise distances between all points in the outlier candidate set and identifying the two points with the greatest distance as the initial cluster centers.

2.3. Association Rule Mining Algorithms and Their Improvements

2.3.1. Association Rule Mining Algorithm

Association rule mining can search for intrinsic relationships and mutual influences between things from messy data [20]. These association rules are likely to be unexpected and can serve as effective reference data to help adjust and optimize decisions.

An association rule is an expression of the form $X \Rightarrow Y$, consisting of a antecedent X and a consequent Y. Suppose that the dataset D is the set of all items in the database, where each item T is a non-empty item set, i.e., a collection of several data items. Let X and Y be two item sets contained in the object T, i.e., $X \subset T$ and $Y \subset T$. If there exist $X \neq \varnothing$ and $Y \neq \varnothing$ such that $X \cap Y = \varnothing$, then $X \Rightarrow Y$ constitutes an association rule in the set of objects D.

2.3.2. Measurement of association rules

The metrics for association rules include two parameters: support and confidence.

The support of an association rule $X \Rightarrow Y$ refers to the percentage of items in the item dataset D that contain both features X and Y out of the total number of items. Support can be expressed as:

$$Sup(X \Rightarrow Y) = \frac{Count(X \cup Y)}{Count(D)} = P(X \cup Y) \tag{8}$$

The higher the support degree, the higher the frequency of simultaneous occurrence of X and Y. If the support degree is very low, then the frequency of simultaneous occurrence of X and Y is also very low, and the association between X and Y is also weak.

The confidence of an association rule $X \Rightarrow Y$ represents the proportion of items in the item dataset D that contain $X \cup Y$ among those that contain X. It can also be understood as the probability that an item containing X also contains Y, i.e., the probability of Y given X. The calculation method for confidence is:

$$Conf(X \Rightarrow Y) = \frac{Sup(X \Rightarrow Y)}{Sup(X)} = P(Y \mid X) \tag{9}$$

The higher the confidence value, the greater the probability that Y will occur simultaneously with

X, i.e., the stronger the directionality of X to Y. When the confidence value is 100%, it means that whenever X occurs, Y will definitely occur simultaneously.

2.3.3. Classic Algorithms for Association Rule Mining

Association rule mining consists of two steps:

(1) Mining frequent item sets from the object dataset D: Each frequent item set and each data item contained therein must appear more than the predetermined minimum support threshold.

(2) Generating association rules based on frequent item sets, where each rule must satisfy both the support and confidence threshold requirements.

The core issue in association rule mining is the discovery of frequent item sets. Therefore, the differences between various algorithms primarily lie in the number of times the database is scanned, as well as the methods used to generate and filter candidate item sets. Currently, the most commonly used association rule mining algorithms include the Apriori algorithm [21] and the FP_growth algorithm [22]. Each algorithm has its own advantages and disadvantages. The Apriori algorithm has less stringent requirements for data quality and good scalability, but it generates redundant candidate item sets during algorithm execution, requiring multiple traversals of the transaction database, resulting in low algorithm efficiency. The FP_growth algorithm compresses the transaction database by constructing a tree structure, reducing the number of times the transaction database is scanned, but increasing memory consumption.

2.3.4. Improved Association Rule Mining Algorithm

This section introduces interest-based metrics to improve the Apriori algorithm for association rule mining.

Interest reflects the degree to which users are interested in a rule, taking into account the novelty, usability, and understandability of the knowledge extracted. Interest can be used to filter rules based on mining objectives and user preferences, eliminating useless or misleading rules and addressing the limitations of the support-confidence model. Interest can be divided into subjective interest and objective interest.

(1) Subjective interest

Subjective interest primarily focuses on users' subjective needs, designed by users for specific mining objectives, and is closely related to domain knowledge. Subjective interest can filter out rules that have both support and confidence above the threshold but are not of interest or unnecessary to users, thereby selecting the specific types of rules users desire.

Template-based interest measurement is a commonly used type of subjective interest measurement. A template describes a set of rules by specifying which attributes can appear in the antecedent and consequent of the rules, and is further divided into inclusive templates and restrictive templates. If a rule matches an inclusive template, it is considered interesting; if it matches a restrictive template, it is considered uninteresting.

(2) Objective Interest

Objective interest is a statistical analysis-based approach that uses certain statistical measures in association rules to assist in rule filtering. Unlike subjective interest, objective interest does not rely on domain-specific knowledge and is a universally applicable interest metric.

1) Probability-based interest

Example of probability-based interest metrics:

$$Interest(X \Rightarrow Y) = \frac{P(XY)}{P(X)P(Y)} \tag{10}$$

This measure of interest considers the correlation between the preceding term X and the subsequent term Y from a probabilistic perspective. When $Interest(X \Rightarrow Y) = 1$, $P(XY) = P(X)P(Y)$, meaning that X and Y are independent of each other, and the occurrence of the antecedent X does not affect the occurrence of the consequent Y. In this case, even if the support and confidence of the association rule $X \Rightarrow Y$ meet the requirements, the rule has no practical significance. When the interest degree $Interest(X \Rightarrow Y) > 1$, X and Y are positively correlated, and the inequality can be transformed into $P(Y \mid X) > P(Y)$. In this case, the occurrence of X increases the probability of Y occurring, and the association rule is interesting. When the interest degree $Interest(X \Rightarrow Y) < 1$, X and Y are negatively correlated, and the occurrence of X reduces the probability of Y occurring. The

association rule $X \Rightarrow Y$ is uninteresting.

The disadvantage of this interest measure is that it cannot distinguish between the antecedent and consequent of a rule. That is, for the same item set $\{X, Y\}$, the two different rules $X \Rightarrow Y$ and $Y \Rightarrow X$ have the same interest value. In addition, this measure does not have the property that the greater the interest, the greater the correlation between X and Y, and it cannot compare the interestingness of different rules.

2) Difference-based interest degree

The literature defines a difference-based interest degree measure:

$$Interest(X \Rightarrow Y) = \frac{Conf(X \Rightarrow Y) - Sup(Y)}{\max\{Conf(X \Rightarrow Y), Sup(Y)\}} \tag{11}$$

This interest degree reflects the effect of X on Y by comparing the difference between the conditional probability of Y occurring given the occurrence of X and the probability of Y occurring on its own, thereby measuring the interestingness of the rule. The denominator of the interest degree, $\max\{Conf(X \Rightarrow Y), Sup(Y)\}$, serves as a standardization factor, controlling the range of interest degree values between $[-1,1]$ to set a minimum interest degree threshold for rule validity assessment. When $Interest(X \Rightarrow Y) > 0$, X and Y are positively correlated, and the rule $X \Rightarrow Y$ is valuable. When $Interest(X \Rightarrow Y) < 0$, X and Y are negatively correlated. When $Interest(X \Rightarrow Y) = 0$, X has no effect on Y, X and Y are independent, and the rule $X \Rightarrow Y$ is meaningless.

For the above two issues, this paper introduces $Conf(\bar{X} \Rightarrow Y)$ into the interest metric to ensure its symmetry and completeness.

3) Difference-based interest metric corrected by introducing $Conf(\bar{X} \Rightarrow Y)$:

$$Interest(X \Rightarrow Y) = \frac{Conf(X \Rightarrow Y) - Conf(\bar{X} \Rightarrow Y)}{\max\{Conf(X \Rightarrow Y), Conf(\bar{X} \Rightarrow Y)\}} \tag{12}$$

The interest level is measured by the difference in confidence as a standard for evaluating the correlation between X and Y and whether the rule is interesting. When $Interest(X \Rightarrow Y) > 0$, the probability of Y occurring when X is present is greater than the probability of Y occurring when X is not present, meaning that the presence of X has a positive effect on Y, and X and Y are positively correlated. When $Interest(X \Rightarrow Y) < 0$, X and Y are negatively correlated. When $Interest(X \Rightarrow Y) = 0$, X and Y are independent of each other.

Similar to association rule mining algorithms based on the support-confidence model, association rule mining algorithms that introduce interest degree measurement also consist of two main steps. The first step is to use the Apriori algorithm to mine all frequent item sets based on the support threshold. The second step is to introduce interest degree judgment when generating strong association rules, and rules with both confidence and interest degrees greater than their minimum thresholds are judged as strong association rules. The specific algorithm flow is shown in Figure 1.

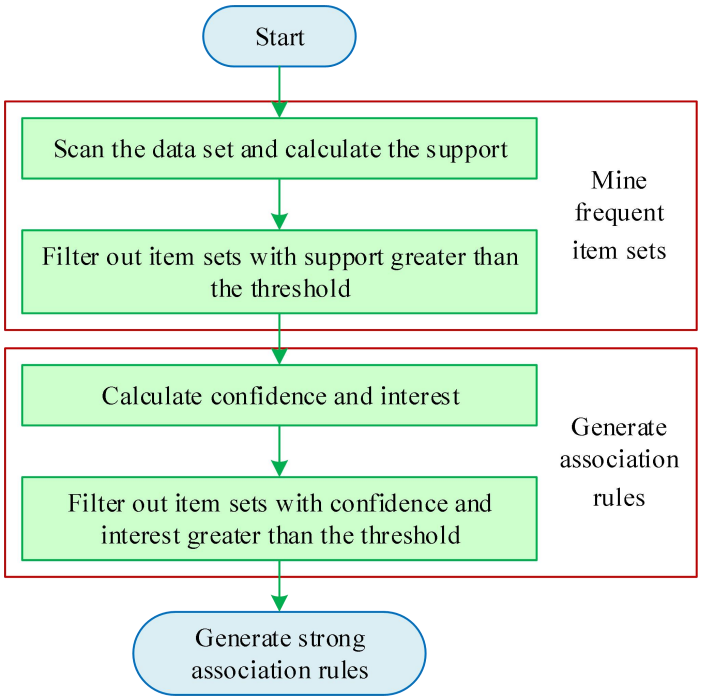


Figure 1 The correlation rule mining algorithm process is introduced

3. Multi-dimensional data correlation analysis of students

This chapter takes 400 students from the Business School of University A as the research object and conducts a correlation analysis between the students' behavioral habits and their academic performance to achieve intelligent management. First, R language technology is used to perform comprehensive data governance on the integrated campus data. Then, an improved K-means clustering algorithm is used to perform cluster analysis on student behavior. Finally, an improved Apriori association rule algorithm is used to perform a correlation analysis between student behavioral habits and academic performance.

3.1. Academic Performance Analysis

3.1.1. The role of histograms in performance analysis

Histograms can perfectly illustrate the regularity of data changes and provide a stable and intuitive representation of the characteristic distribution of the data being analyzed. Histograms are primarily used to analyze data changes by describing the frequency distribution of continuous variables. Kernel density plots are suitable for observing the distribution of continuous variables and serve as a nonparametric method for estimating the probability density function of a random variable.

Using score values as continuous variables, statistical analyses were conducted on students' Business Management, College English, and Advanced Mathematics grades. The histograms and kernel density plots for the three subjects are shown in Figures 2–4. First, the different distribution patterns of the three subjects can be observed overall, while kernel density plots can more accurately determine whether the distribution is normal. As shown in Figure 2, the histogram of business management scores resembles a small mountain, with peaks and valleys, and the mountain is particularly steep, with scores between 90 and 95 being the most prominent. The histogram for university English scores is relatively flat, with scores between 80 and 85 being the most prominent. The number of high and low scores decreases gradually, and the kernel density curve is also relatively smooth. The histogram and kernel density curve for higher mathematics scores show uneven distribution, with two distinct peaks, and the distribution in the figure appears somewhat random.

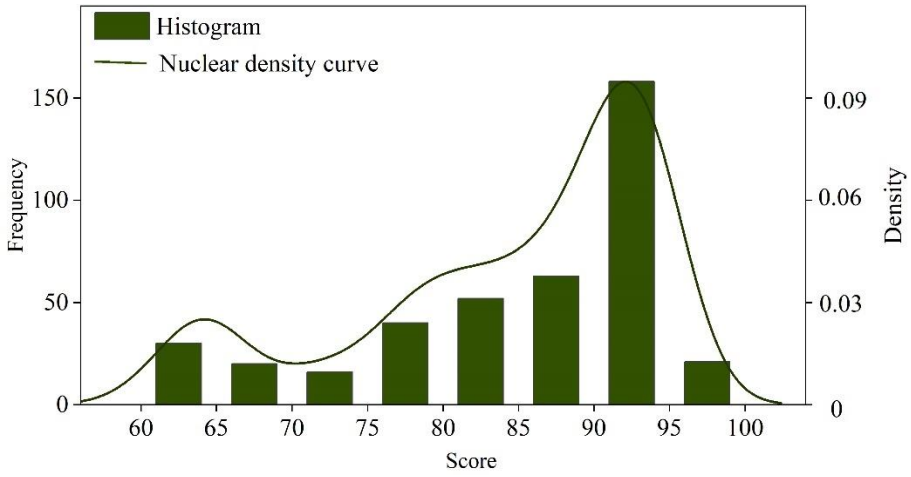


Figure 2 The overall trend chart of business management scores

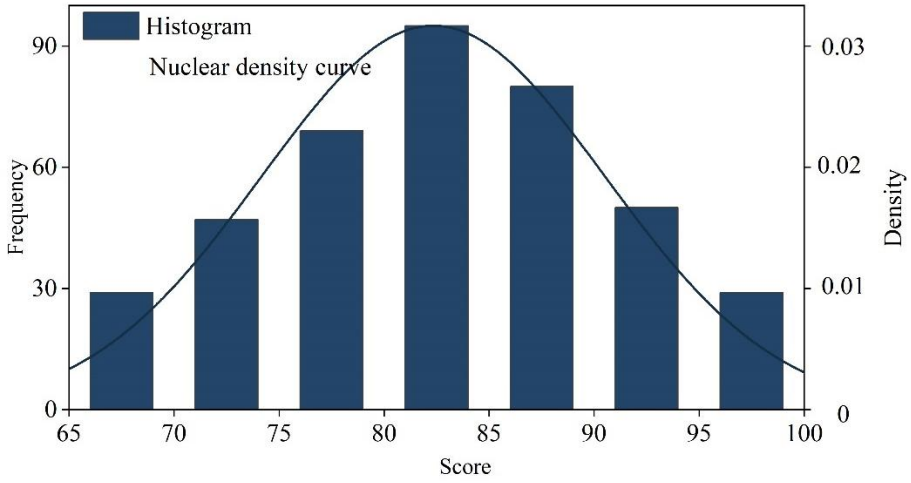


Figure 3 The overall trend chart of college English scores

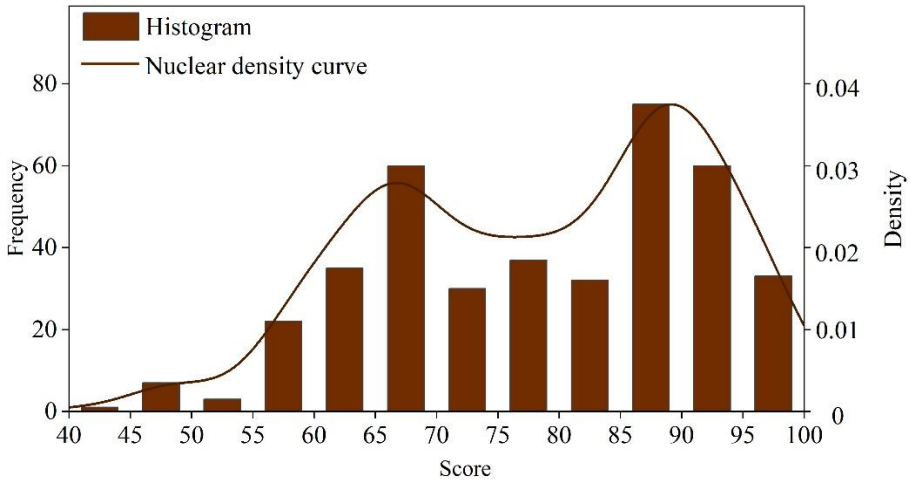


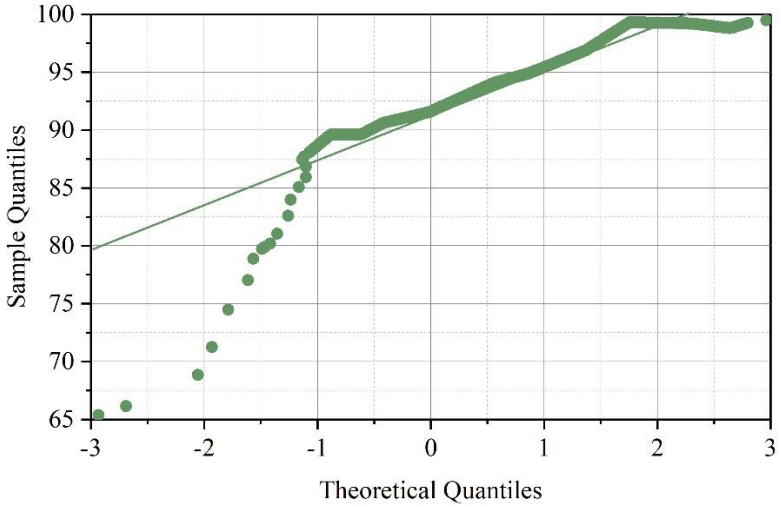
Figure 4 The overall trend chart of higher math scores

Based on the preliminary findings outlined above, we can proceed with the subsequent comparative analysis. The distribution of students across different score ranges in business management is uneven, creating significant score gaps. The disparity in the number of students across different score ranges is too large, which may also be related to the relatively easy nature of the questions, as there are a relatively large number of high scorers above 90 points. In contrast, university English scores present a more ideal assessment outcome, with fewer students in the high and low score ranges. Scores generally show a gradual increase, with the largest proportion of students scoring around 80 points, accounting for one-quarter of the total student population. The score range of 80 and above decreases step by step,

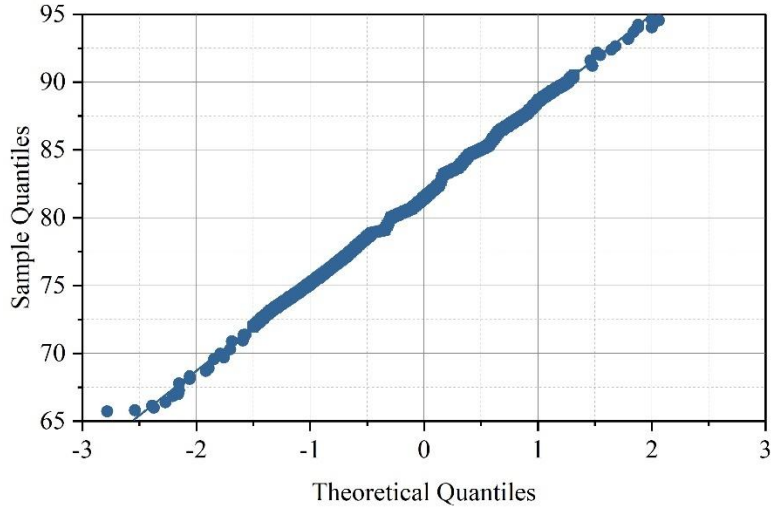
achieving a better teaching effect. And for the results of the investigation of advanced mathematics, it is the most inferior of the three subjects, the figure accounts for a large proportion of the total number of people is the two extremes of high and low scores, the number of scores in the middle of 70 to 85 is small, and the number of scores between 85 and 90 is the largest, this situation is generally related to the proposition teacher, the proposition teacher has not been able to effectively avoid the problem of the difficulty of the test questions, so the teaching effect is poor, for students with good learning ability, this is an opportunity to open the gap with students with poor grades, However, for students with poor grades, the number of difficult problems is relatively large due to the weak receptivity of this part of the students, resulting in a large number of low scores, so the course grades do not conform to the normal distribution. By running the R language environment, the generated student course performance distribution profile map can help teachers accurately grasp whether the teaching process is in line with personalized education.

3.1.2. The role of QQ images in performance analysis

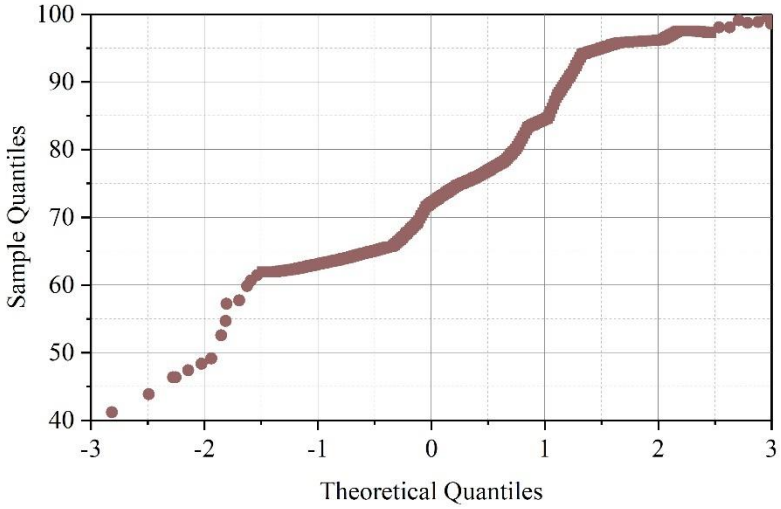
Determining whether the data follows a normal distribution is of utmost importance. A QQ plot can assist in analyzing whether the distribution of the data aligns with a normal distribution. When the points on the QQ plot are clustered around the line $y = \sigma x + \mu$, it indicates that the data is close to a normal distribution. In this case, σ and μ represent the standard deviation of the slope and the mean of the intercept of the line, respectively. Whether the data conforms to a normal distribution can be determined by observing how close the points around the line are to the line itself; the closer they are, the better the fit to a normal distribution. In this study, the author created an executable script in R language using student data, generating the QQ plots shown in Figure 5, where (a) to (c) represent the normal distribution plots for Business Management, College English, and Calculus scores, respectively. Since the University English scores are the closest to the linear fit, University English is the course that best fits the normal distribution.



(a) Business management performance normal distribution



(b) College English performance normal distribution



(c) Higher mathematics achievement normal distribution

Figure 5 Progressive normal distribution

3.2. Cluster analysis of student behavior based on the K-means algorithm

This chapter's experiment utilizes behavioral data from 400 business school students over a single semester from August 2023 to January 2024 for a K-means clustering analysis.

The behavioral data recorded by the campus ID card system, including shopping, dining, medical services, water dispensing, library book borrowing, and access control, represent the actual and specific activities of college students on campus. By combining this behavioral data with the actual conditions on campus, a practical evaluation indicator system can be constructed to cluster and classify student behavior. This study primarily focuses on data mining and analysis of a dataset integrated from campus ID card data, library book borrowing data, and academic performance data. The behavioral data of college students on campus is divided into three evaluation indicator systems: dining consumption level, regularity, and diligence, to describe and analyze the behavior of students on campus. By conducting cluster analysis based on these indicators, student models can be abstracted, helping student affairs managers gain a deeper understanding of students, accurately grasp their circumstances, and provide decision-making support for the implementation of student management work.

3.2.1. Clustering of dining consumption levels

This experiment first extracted data on the dining consumption levels of college students on campus from the organized campus ID card data set and conducted a cluster analysis experiment. Using the improved K-means algorithm in the R language tool to perform cluster experiments on the data, the sum

of squared errors within clusters was used as the indicator for evaluating the optimal number of clusters, and the optimal number of clusters is shown in Figure 6. When K = 3, the slope change is not obvious.

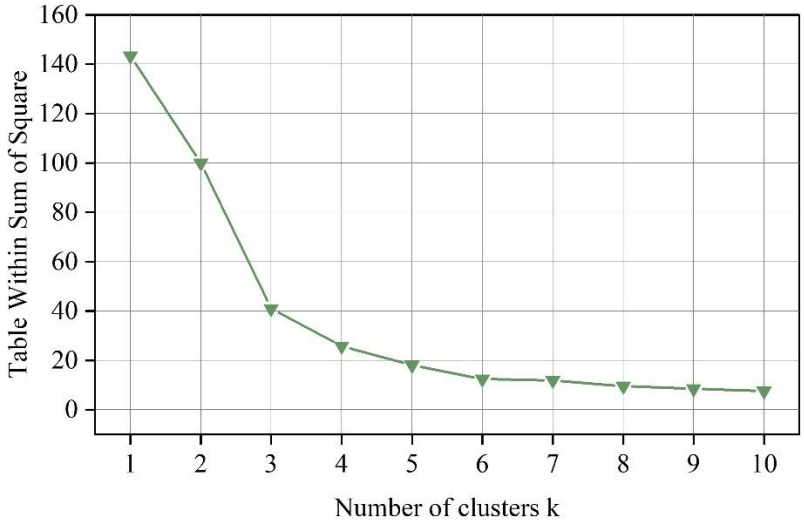


Figure 6 Optimal clustering number diagram

With K = 3 as the optimal number of clusters, the clustering results for dining consumption levels are shown in Figure 7.

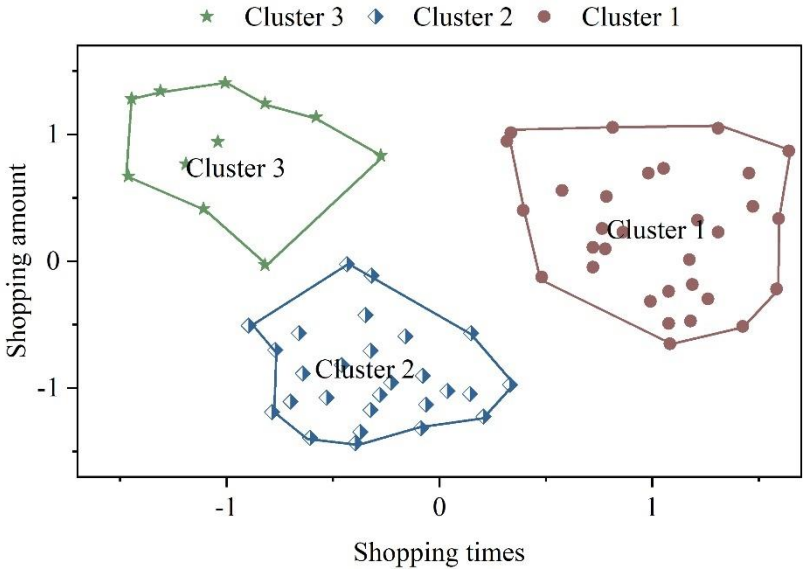


Figure 7 Consumption level clustering results

Through experimentation, the results of meal consumption clustering were obtained. The final clustered results are shown in Table 1. Among students in Category 1, the frequency of on-campus meal consumption transactions was relatively high, with the lowest average transaction amount per transaction, aligning with the overall consumption level of most students in this major. It can be observed that the basic living expenses of this category of students primarily occur within the campus, with meal consumption patterns being relatively regular. Among students in Category 2, the frequency of on-campus dining consumption transactions was moderate, with the highest average transaction amount. Their overall consumption level accounted for a relatively large proportion among students in this major. It can be seen that most of their basic living expenses occur on campus, with a small portion of off-campus consumption, and their dining consumption is relatively regular. Among students in Category 3, the frequency of on-campus dining consumption transactions was low, with an average transaction amount at the midpoint, and they accounted for a small proportion among students in this major. It can be seen that the majority of basic living expenses for this category of students are likely to occur off-campus. Teachers should pay close attention to this category of students, verify whether they are studying and living on campus normally, and guide them to develop safe and healthy on-campus dining consumption habits.

Table 1 Consumer clustering results

Categories	Consumption number	Consumption amount	Student category label	Student ratio
1	469	2366	Regular type	46.32%
2	268	1523	Moderate type	35.11%
3	125	627	Thrifty type	18.57%

3.2.2. Regular clustering

Secondly, cluster analysis was performed on regular data such as meal consumption levels and shower water usage in student behavior data. The optimal number of clusters is shown in Figure 8. When K = 4, there is no significant change in slope.

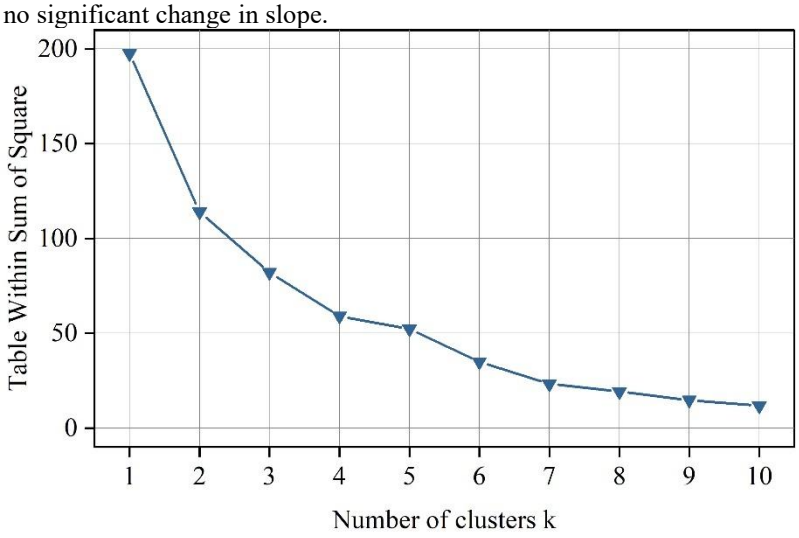


Figure 8 Optimal clustering number diagram

With K=4 as the optimal number of clusters, the regular clustering results are shown in Figure 9.

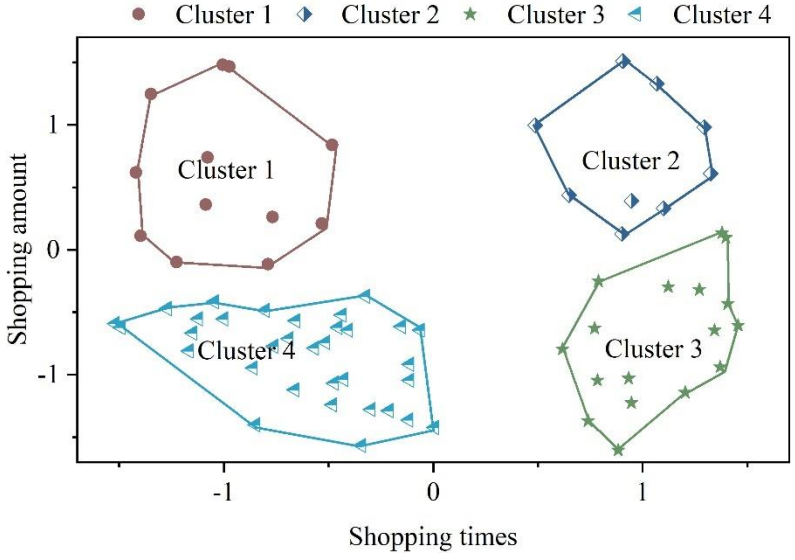


Figure 9 Regular clustering results

The final regular clustering results are shown in Table 2. Among Category 1 students, the frequency of swiping cards for on-campus dining was the lowest, while the frequency of showering and boiling water was the highest. It can be seen that these students pay more attention to personal hygiene and are more likely to engage in physical activities on campus. However, their low on-campus dining frequency may be due to frequent off-campus dining or ordering takeout, so it is important to guide these students to develop healthy eating habits. Among students in Category 2, the frequency of on-campus dining card swipes was the highest, while the frequency of showering and water dispenser use was at an intermediate

level. It can be seen that students in this category primarily dine at the on-campus cafeteria, demonstrating regularity and good living habits, and are considered a better category among peers in this major. Among students in Category 3, the frequency of on-campus dining card swipes was relatively low, while the frequency of showering and water dispenser use was at an intermediate level, making them the mainstream among students in this major. It can be seen that this category of students frequently engage in off-campus activities or order takeout. Teachers should pay special attention to this category of students, guiding them to develop safe and healthy on-campus learning and living habits to eliminate safety risks. Among Category 4 students, the frequency of on-campus dining card swipes is relatively high, while the frequency of showering and water dispenser use is at the lowest level. It can be seen that these students tend to be more reclusive and less attentive to personal hygiene. Teachers should communicate more with these students, verify their attendance records, and guide them to develop good hygiene and study habits.

Table 2 Regular clustering results

Categories	Consumption number	Shower frequency	Open water frequency	Student ratio
1	87	29	151	24%
2	454	18	126	20%
3	127	19	118	38%
4	284	15	61	18%

3.2.3. Diligence Clustering

Since most of the drinking water dispensers in this school are located in the teaching building, this paper uses the number of times water is dispensed as an indicator of the number of times students enter the teaching building to study, and combines this with the number of books borrowed as indicators of diligence for clustering. The number of optimal clusters is determined using the sum of squared errors within clusters as an evaluation criterion, as shown in Figure 10. When K=3, there is no significant change in slope.

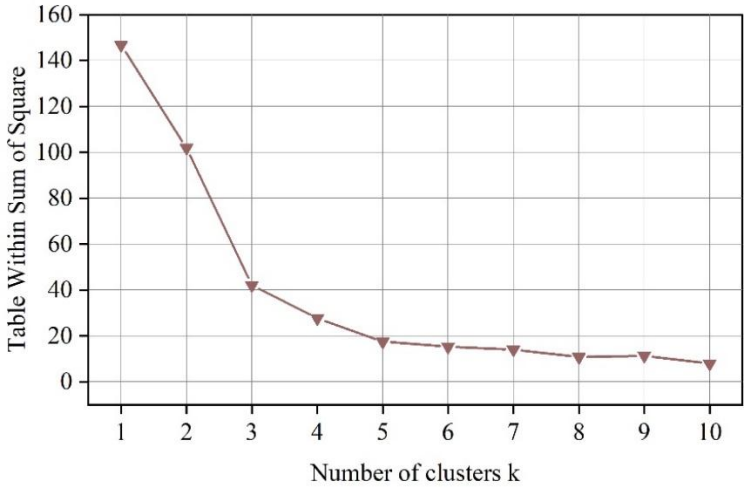


Figure 10 Optimal clustering number diagram

With K = 3 as the optimal number of clusters, the diligence clustering results are shown in Figure 11.

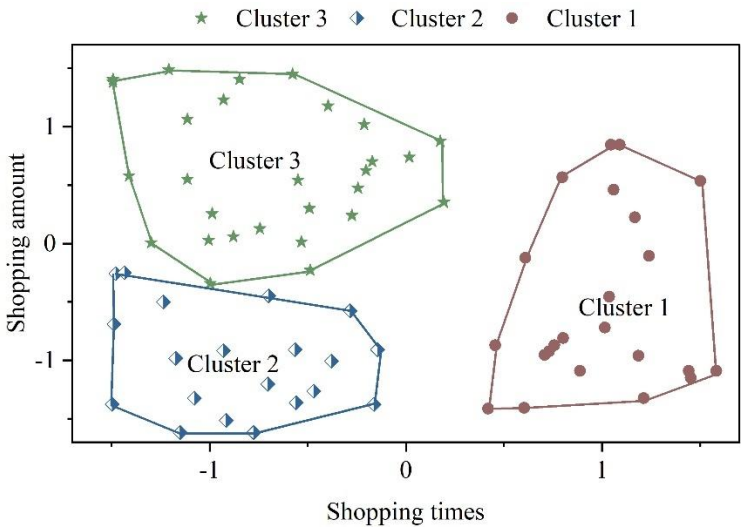


Figure 11 Diligence clustering results

The final results of the diligence clustering are shown in Table 3. Students in Category 1 had an average frequency of swiping their cards to access water in the academic building and borrowed the highest number of books this semester, accounting for the highest proportion of students in this major. It can be seen that these students love reading and are relatively good at using library resources. Students in Category 2 had the highest frequency of swiping their cards to access water in the academic building and borrowed an average number of books this semester. It can be seen that these students frequent the teaching building more often, prefer to study independently, and effectively utilize learning resources. Among students in Category 3, the frequency of swiping cards to access water in the teaching building is the lowest, and the number of books borrowed this semester is at the lowest level. It can be seen that these students appear in learning venues less frequently, with a small number of students not visiting the library and generally not actively utilizing these learning resources for study. Library staff can organize reading activities for this group of students, survey their reading needs, and better manage library operations to make reasonable decisions.

Table 3 Diligence clustering results

Categories	Open water frequency	Book number	Student ratio
1	84	13	38%
2	113	6	32%
3	50	4	30%

3.3. Analysis of the correlation between student behavior habits and academic performance

3.3.1. Experimental Analysis of the Improved Apriori Algorithm

This section uses an improved Apriori algorithm to analyze the correlation between student behavior habits and academic performance. To verify the efficiency of the improved Apriori algorithm, a comparative experiment was conducted between the original Apriori algorithm, the transaction compression-based Apriori algorithm (Apriori_C), the hash-based Apriori algorithm (Apriori_H), and the improved Apriori algorithm proposed in this paper. Using student behavioral data and academic performance data, association rule mining was performed with the same two threshold parameters: a minimum support threshold of 0.2 and a minimum confidence threshold of 0.4. The required time was compared, and the results are shown in Table 4. As can be seen from the figure, the improved algorithm in this paper takes less time, consuming only 1.23 seconds.

Table 4 The time-consuming comparison of the four algorithms

Algorithm name	Apriori	Apriori_C	Apriori_H	Ours
Time consuming	4.87s	3.96s	2.55s	1.23

To validate the efficiency of the improved Apriori algorithm across different datasets and to comprehensively verify the algorithm's effectiveness, three publicly available datasets were employed: Groceries, Movies, and Adult. The Groceries dataset consists of real “shopping basket” transaction records sourced from the open-source software RGui, containing 150 items and 30 feature attributes. The Movies dataset pertains to movie genres, with 62,300 records and 8 feature attributes. The Adult dataset is one of the datasets in the UCI database, containing 47,856 records and 11 feature attributes, including 6 discrete variables such as occupation and education level, and 5 numerical continuous variables such as age and income. In this experiment, 6 discrete variables were used. The same minimum support threshold of 0.2 and minimum confidence threshold of 0.4 were set, and the results are shown in Table 5. The improved Apriori algorithm achieved the highest execution speed.

Table 5 Time consumption of four algorithms on the three datasets

Algorithm name	Apriori	Apriori_C	Apriori_H	Ours
Groceries	73.63s	66.52s	5.96s	3.98s
Movies	15.63s	14.63s	12.36s	4.63s
Adult	9.63s	8.41s	6.74s	5.26s

3.3.2. Improving the Application and Analysis of the Apriori Algorithm

After verifying that the improved Apriori algorithm achieves optimal execution speed, this section uses the algorithm to conduct an association analysis between student behavior habits and academic performance. The results of the association rule mining between student behavior and academic performance are shown in Table 6.

Among the extracted association rules, two post-rules are associated with excellent academic performance: Rule 1: Regular type, more books borrowed => excellent; Rule 2: Semi-regular type, more books borrowed => excellent. From Rules 1 and 2, it can be observed that students with excellent academic performance tend to have more structured lifestyles, spend more time studying in the library, and frequently borrow books.

There are two post-rules with good academic performance: Rule 3: Regular type, borrowing a relatively large number of books => good; Rule 4: Relatively regular type, borrowing a relatively large number of books => good. From these two rules, it can be seen that students with good academic performance are relatively less diligent in certain aspects compared to those with excellent academic performance, but their academic performance is maintained due to their relatively higher level of effort.

There are five rules with post-conditions for average grades, numbered 5 to 9. From these five rules, it can be seen that the primary cause of poor grades is insufficient effort in studying, with irregular lifestyles being a secondary cause, and consumption behavior having little impact on grades. Such students should focus on cultivating good study habits and learn from students with excellent grades.

There are no post-rules for poor academic performance, but there are pre-rules for poor academic performance. For example, Rule 10: Poor => Few books borrowed, and Rule 11: Poor => Irregular lifestyle. These rules indicate that students with poor academic performance generally have disordered lifestyles and lack diligence in their studies. Additionally, based on the clustering results, students with irregular lifestyles often have higher expenses. Counselors and other administrators should pay close attention to the campus life of students with poor academic performance, whether they are under excessive pressure, or exhibit negative or decadent emotions, and promptly understand and assist in adjusting their circumstances. For individual students with multiple failed courses without special circumstances, academic warnings or other measures should be imposed. For students with other abnormal circumstances, special handling should be provided.

Table 6 Student behavior and performance association rule mining results

N	Rule	Support	Confidence
1	Regular, borrowed books=>Excellence	0.469	0.87
2	More regular, more books=>Excellence	0.417	0.861
3	Regular type, more than the number of books=>Good	0.52	0.721
4	More regular, more than the number of books=>Good	0.49	0.724
5	Regular, the book is medium=>Medium	0.348	0.769
6	Less regular, the book is medium=>Medium	0.428	0.91

7	Moderate, less regular, the number of books is medium=>Medium	0.243	0.785
8	The campus is low and low, and the number of books is medium=>Medium	0.215	0.773
9	Frugal, regular, and the number of books=>Medium	0.421	0.785
10	Poor=>there are fewer books	0.559	0.921
11	Poor=>irregular	0.542	0.881
12	Poor=>the number of books is low and irregular	0.541	0.877

Based on the above analysis, the main factors affecting students' academic performance are lifestyle behaviors and learning behaviors, while consumption behaviors have a relatively minor impact on academic performance. However, cluster analysis also shows that lifestyle behaviors have a significant impact on consumption behaviors, and irregular lifestyle habits can lead to higher consumption. Therefore, managers should pay attention to students' lifestyle and learning situations.

4. Conclusion

This study utilizes R language technology for data mining of student behavior. An improved K-means clustering algorithm is employed to categorize student behavior. Additionally, an enhanced Apriori association rule algorithm is applied to analyze the correlation between student behavioral habits and academic performance. After implementing intelligent management for 400 students at School A's Business School, the following conclusions were drawn:

(1) Students' university English grades are relatively satisfactory, with scores generally showing a gradual increase. The proportion of students scoring around 80 points is the largest compared to other score ranges.

(2) A descriptive analysis of students' behaviors was conducted using one-card system data. The improved K-means clustering algorithm was applied to classify student behavior data. It was found that lifestyle behaviors significantly influence consumption behaviors, and irregular lifestyle habits can lead to higher consumption.

(3) Through association rule mining using an improved Apriori algorithm, the primary factors influencing students' academic performance were identified as lifestyle behaviors and learning behaviors, while consumption behaviors had a relatively minor impact on academic performance.

Based on the above analysis results, school administrators can promptly understand student conditions and improve management levels and efficiency.

REFERENCES

[1] Zhang, Y. (2021, April). Application of computer information processing technology in teaching management information system of colleges and universities. In Journal of Physics: Conference Series (Vol. 1852, No. 4, p. 042089). IOP Publishing.

[2] Liu, Y. (2015). Analysis on the effective integration of information technology and personnel management in colleges and universities. Creative Education, 6(08), 785.

[3] Zheng, G. (2024). Research on the Integration of Institutional Management and Emotional Management in College Student Management in the Era of Artificial Intelligence. Adult and Higher Education, 6(1), 34-43.

[4] Yongtao, Z. (2019, October). Research on the application of artificial intelligence technology in scientific research management in colleges and universities. In 2019 12th International Conference on Intelligent Computation Technology and Automation (ICICTA) (pp. 438-441). IEEE.

[5] Huebner, R. A. (2013). A Survey of Educational Data-Mining Research. Research in higher education journal, 19.

[6] Donner, E. K. (2023). Research data management systems and the organization of universities and research institutes: A systematic literature review. Journal of Librarianship and Information Science, 55(2), 261-281.
[7] Levina, E. Y., Mustafina, G. M., Nigmetzyanova, V. M., Galiyev, R. M., Chalkina, N. A., Ashmarina, S. I., & Yeremeyeva, T. S. (2015). Improving the information system of university management. Rev. Eur. Stud., 7, 109.
[8] Zheng, C., & Zhou, W. (2021, April). Research on information construction and management of education management based on data mining. In Journal of Physics: Conference Series (Vol. 1881, No. 4, p. 042073). IOP Publishing.
[9] Lei, Q., Li, Y., & Yan, S. (2022). Design and Optimization of University Management Information System Based on Internet of Things and Intelligent Computing Model. Journal of Sensors, 2022(1), 1049535.
[10] Li, W. (2021). Design of smart campus management system based on internet of things technology. Journal of Intelligent & Fuzzy Systems, 40(2), 3159-3168.
[11] Wu, M., Hao, X., Lv, Y., & Hu, Z. (2022). Design of intelligent management platform for industry–education cooperation of vocational education by data mining. Applied Sciences, 12(14), 6836.
[12] Zhang, M., Fan, J., Sharma, A., & Kukkar, A. (2022). Data mining applications in university information management system development. Journal of Intelligent Systems, 31(1), 207-220.
[13] Sun, H. (2019). Study on application of data mining technology in university computer network educational administration management system. Journal of Intelligent & Fuzzy Systems, 37(3), 3311-3318.
[14] Hu, J., & Li, H. (2021). Composition and optimization of higher education management system based on data mining technology. Scientific programming, 2021(1), 5631685.
[15] Natek, S., & Zwilling, M. (2014). Student data mining solution–knowledge management system related to higher education institutions. Expert systems with applications, 41(14), 6400-6407.
[16] Zheng, D. (2021). Research on the evaluation model of educational management theory based on data mining from the perspective of neural network. Wireless Communications and Mobile Computing, 2021(1), 7260806.
[17] Wei, W., & Jin, Y. (2023). A novel Internet of Things-supported intelligent education management system implemented via collaboration of knowledge and data. MBE: Mathematical Biosciences and Engineering, 20, 13457-73.
[18] Yanwei Zhao,Xiangyun Kong,Wei Zheng & Shahbaz Ahmad. (2024). Emotion generation method in online physical education teaching based on data mining of teacher-student interactions. PeerJ. Computer science,10,e1814-e1814.
[19] Kim Minseok,Jung Seunghwan,Kim Baekcheon,Kim Jinyong,Kim Eunkyeong,Kim Jonggeun & Kim Sungshin. (2022). Fault Detection Method via k-Nearest Neighbor Normalization and Weight Local Outlier Factor for Circulating Fluidized Bed Boiler with Multimode Process. Energies,15(17),6146-6146.
[20] Liu Lihong. (2022). Integration and Recommendation of Multimedia Network-Assisted English Instructional Resources Based on Association Rules Mining. Mobile Information Systems,2022,

[21] Zheng Tang,Zhengwei Jiang,Ying Li,Haowei Yuan,Jiayu Han & Chao Chen. (2024). Research on the Association Analysis of Online Learning Behaviors Based on the Apriori Algorithm. Frontiers in Computing and Intelligent Systems,9(2),18-22.
[22] Praveen Kumar B.,Padmavathy T.,Muthunagai S.U. & Paulraj D.. (2024). An optimized fuzzy based FP-growth algorithm for mining temporal data. Journal of Intelligent & Fuzzy Systems,46(1),41-51.