Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 6, pp. 5701–5713, DOI 10.12694/scpe.v25i6.2187

DATA MINING TECHNOLOGY FOR SMART CAMPUS IN BEHAVIOR ASSOCIATION ANALYSIS OF COLLEGE STUDENTS

JUN ZHANG[1], YUNXIN KUANG[2*] AND JIAN ZHOU[3]

Abstract. The data in smart campuses is complex and massive, with insufficient utilization, and existing data processing methods have many limitations. Therefore, in order to improve the efficiency of data processing in universities and assist in student management, a data processing method integrating cluster analysis and association rule mining is proposed. The proposed method is divided into two parts. Firstly, an improved K-Means model based on information entropy and density optimization is constructed for clustering analysis of student consumption, learning, and other data; Secondly, use the improved Mapping Apriori to obtain the correlation between student grades, consumption records, and learning behavior. The clustering results on student consumption data show that the average accuracy of ED-K-Means clustering is 97.41%, which is 12.8%, 8.5% and 4.0% higher than the comparison algorithm. The result of the correlation between consumption level and achievement shows that when the amount of consumption is less than 1500 yuan, the student's achievement is directly proportional to the amount of consumption. Therefore, the proposed method can effectively mine and analyze student behavior data, which has important practical significance for intelligent management in universities.

Key words: SC; Data mining; ED-K-Means clustering algorithm; Mapping-Apriori; Association analysis

1. Introduction. As an important platform for training talents, higher education is an indispensable part of education. With society developing, colleges' and universities' management needs are constantly changing. Science and technology integrated development drives the SC construction [1]. SC effectively integrates university information resources by using information technology. It can realize reasonable resource allocation, coordinate and optimize education and teaching, student management, teacher office and other work [2]. The huge data contained in the SC can provide theoretical basis for educational operation decision-making, teachers and students management, campus convenience and other aspects. This can also enhance the quality of the campus and promote the efficient operation of the campus [3]. Among them, as an important part of higher education, college students can generate behavior data in a variety of ways in daily life. Consumption records, travel records, learning and training data are very large. How to analyze them and understand students' behavior laws is the research direction of many scholars [4]. Data mining technology can mine potential information in massive data, make reasonable use of college education and teaching information and various kinds of data, and provide reference basis for student management.

University data has the characteristics of complexity and large quantity. Although data mining technology has significant advantages in university management, its application still faces challenges such as insufficient data utilization. In addition, existing data mining technologies also face issues such as insufficient computational efficiency and poor mining results. In view of this, new data processing methods urgently need to be proposed to address the aforementioned issues. Therefore, the study first utilizes the improved K-Means algorithm based on information entropy and density optimization to cluster massive data such as student consumption and learning; Subsequently, the improved Mapping Apriori is used to explore the correlation between student grades, consumption records, and learning behavior.

The significance of the research lies in the aim of analyzing the daily behavioral characteristics of college students and their relationship with academic performance through improved data mining algorithms, providing scientific basis for university management. Through the K-Means algorithm based on information entropy and

[1]School of Artificial Intelligence, Hunan Railway Professional Technology College, Zhuzhou, 412001, China
[2*]Integration of Industry and Education, Hunan Railway Professional Technology College, Zhuzhou, 412001, China (yunxin_kuang@gmx.com)
[3]School of Artificial Intelligence, Hunan Railway Professional Technology College, Zhuzhou, 412001, China

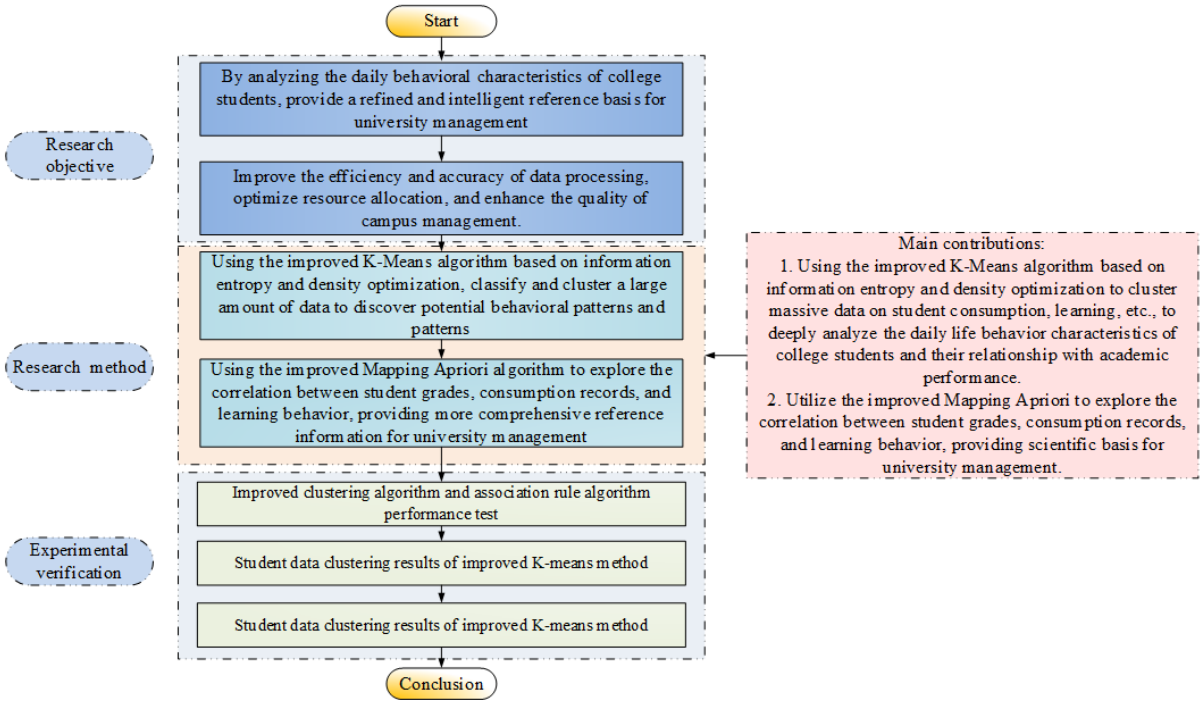


Fig. 1.1: Main Process and Contributions of the Study

density optimization, as well as the improved Mapping Apriori algorithm, it is expected to improve the efficiency and accuracy of data processing, and provide strong support for optimizing resource allocation and improving campus management quality, thereby helping universities achieve refined and intelligent management. The main process and contributions of the study are shown in Figure 1.1.

2. Related Work. As information technology develops, SC is widespread. Aiming at data digitization in the SC construction, Li W scholars studied the construction of a SC system using IoT technology. The system used face recognition technology for data collection and unified management and analysis in the background. Through the experimental test, the data of students and teachers analyzed could help colleges and universities make reasonable learning plans, and the satisfaction of system users reached 8.0 points [5]. To realize data visualization, Prandi et al. built an interactive and intelligent system, studied the sensors to collect real-time campus data, and interacted with relevant data in the system background. By testing the student experiment, the system could visualize a large amount of data in front of users, and the participation enthusiasm of student users had been greatly improved, which had certain use value [6]. Facing the difficulties encountered in SC construction, Jurva research team proposed an intelligent campus framework consisting of 5G and the Internet of Things. Through actual case analysis, the intelligent campus could effectively collect, store and analyze massive data, in which 5G technology played an important role [7]. Faced with the increasingly huge campus resources and the need to optimize the information management and decision-making methods, researchers such as Valks used Internet technology to build a SC. Through the application practice test, the results showed that the system could collect data in real time, and feedback user data changes according to the space utilization rate, which could provide a basis for decision-making for campus management [8]. Starting from the evolution and application of SC, Zhang and other scholars summarized the problems in the construction and development of the system through literature reading, case analysis and other methods. The results showed that the awareness of massive data sharing contained in the SC is insufficient, and the evaluation system formulation strategy needed to be optimized and improved [9].

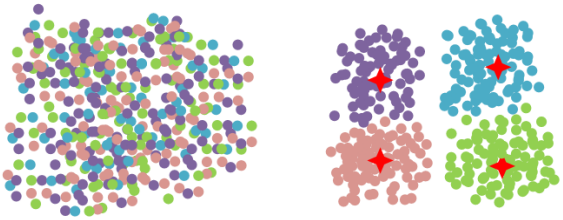
Students are crucial in higher education. It is very important to excavate and analyze student information. Fan and other scholars started with online courses and studied a course recommendation method combining attention mechanism. This method was based on students' preferences and learning records, and personalized course recommendations are made for them through data analysis. Through the data validation test, the recommendation results of this method basically met the students' expectations. And the recommendation effect was good, which could realize personalized course recommendation [10]. To analyze the factors of students' rapid adaptation to school, the research team of Roorda used meta-analysis and survey methods to analyze the role of teachers. The results showed that teachers could help students integrate into campus life more quickly, actively participate in activities, and promote students' internal will and externalization behavior to be positively correlated [11]. Bureau et al. explored the relevant factors and influencing mechanism of students' self-determined motivation through case analysis. The results show that students' self - ability is the most significant factor affecting self - determination motivation, followed by autonomy; And teachers' encouragement and support have a greater impact than parents' [12]. Li scholars analyzed the relationship between students' anxiety, depression, sleep quality and mobile phone addiction, and calculated the correlation coefficient through the effect model. Students' mobile phone addiction correlates positively with anxiety and depression, negatively with sleep quality. Students with mobile phone addiction were more prone to anxiety, depression and other emotions. And their sleep quality was not high [13]. Pérez-Pérez research team faced the problem of students' satisfaction with learning management system. The research explored by building a relationship analysis model, and used the partial least squares method to calculate the relationship coefficient. The results showed that system information quality is significantly impacted on students' satisfaction. At the same time, building a good virtual learning environment could also improve students' satisfaction. The research results could provide a reference basis to improve the system [14].

As information technology popularizes, it is found that colleges and universities have gradually attached importance to the construction of SC data and the analysis of campus data. But many data have not been in-depth studied from the brief introduction of the achievements of domestic and foreign researchers. The study analyzes the relationship between students' different behavior characteristics and grades, and provides feasible suggestions for student management by students' daily behavior data such as consumption, study, and diet rules.

3. Analysis on students' behavior data by improved clustering and association rule algorithm.

3.1. Clustering algorithm based on information entropy and density optimization. With the SC construction and development, data related have expanded rapidly. Data mining technology can sort out, analyze and summarize valuable information from massive data, which is widely used by people [15]. Through the analysis of various behavioral data generated by college students in their daily life, students' behavioral laws can be obtained. This can help the development and improvement of campus quality service, scholarship evaluation, campus security and other work [16]. Cluster analysis can classify and process similar data, which is the most commonly used technology in data analysis. The research will apply cluster technology to the analysis of college students' behavior data. The relationship between students' grades and various behaviors will be explored through association rule algorithm to provide suggestions and ideas for student management. When applying the clustering algorithm, it is essential that appropriate algorithms are chosen for different situations and requirements. K-Means clustering algorithm divides the data into different classes through continuous iteration, with simple concept and strong scalability. Figure 2 shows the iterative clustering process of K-Means algorithm when $K = 4$.

K-Means has high clustering efficiency and good scalability. However, for the initial cluster center, the random selection of K-Means may make the cluster center in the same cluster, and the validity of the clustering results will be greatly reduced [17]. At the same time, the size of the cluster (i.e., K value) cannot be accurately calculated, which is usually determined by the user's experience. It can affect the clustering accuracy greatly. In view of the shortcomings of K-Means, the K-Means of Information Entropy and Density Optimization (ED-K-Means) is proposed. The optimization idea is to use the information entropy of sample data attributes to assign the Euclidean distance, and determine the cluster center more reasonably from the assigned value. Information entropy is one of the methods to eliminate information uncertainty, which can be used to measure the amount of information. In the process of using information entropy to assign values to each data, the sample data set



(a) Initial data set (b) Cluster results

Fig. 3.1: K-Means iterative clustering process

contains multiple dimensions, and the attribute matrix in the database needs to be constructed. Formula (3.1) the is expression.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix} \tag{3.1}$$

m is students' number in the sample data. n is the behavior data number generated by students in formula (3.1). The number of times that students appear behavior j is expressed by i when calculating the weight of each behavior attribute of students in the data set. In the calculation process, the data needs to be limited to the range of [0,1]. Formula (3.2) shows the data process.

$$Q_{i,j} = \frac{a_{i,j}}{\sum_{j=1}^{m} a_{i,j}} \tag{3.2}$$

In formula (3.2), $Q_{i,j}$ represents the proportion value of behavior attribute. Formula (3.3) is the expression of behavior i information entropy.

$$I_i = -\sum_{j=1}^{m} \log Q_{i.j} * Q_{i.j} \tag{3.3}$$

In formula (3.3), when $Q_{i,j} = 0$, $a_{i.j}$ is all equal, and there is a maximum value of I_i. In the data set, to distinguish the differences between behavior attributes, the difference coefficient is defined as h_i. Formula (3.4) is the calculation.

$$h_i = 1 - I_i \tag{3.4}$$

Formula (3.4) shows that the smaller the information entropy I_i is, the larger the h_i is, the more important the behavior attribute is, and the greater the clustering effect is. On the contrary, the clustering effect is smaller. When $I_i = 1$, $h_i = 0$, which means behavior attribute has no clustering effect. The difference coefficient is used to assign values to different behavior attributes, and Formula (3.5) is the weight expression.

$$w_i = \frac{h_i}{\sum_{i=1}^{n} h_i} \tag{3.5}$$

After the assignment, the similarity between data objects is calculated according to the Euclidean distance formula. Formula (3.6) is the changed Euclidean distance.

$$d_w(x_i, x_j) = \sqrt{\sum_{p=1}^{n} w_p (x_{ip} - x_{jp})^2} \tag{3.6}$$

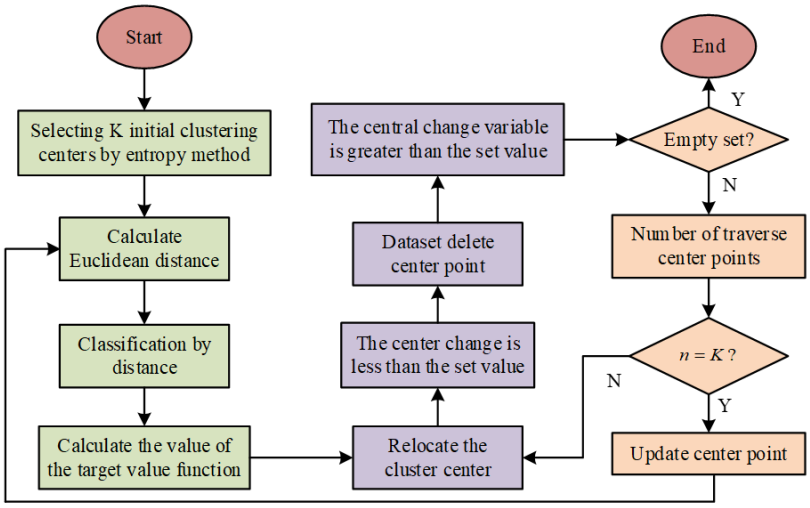


Fig. 3.2: Operation process of ED-K-Means algorithm

w_p refers to the weight of attribute p in formula (3.6). Through similarity calculation, the attribute that corresponds to weight is appropriately reduced or enlarged. This makes the clustering effect of attributes with large weights greater, and the clustering effect of attributes with small weights smaller. If clustering objective function is the standard deviation, the expression is Formula (3.7).

$$\sigma_i = \sqrt{\frac{\sum d_w\left(a_i, c_i\right)}{|C_i| - 1}} \tag{3.7}$$

c_i represents the centroid of data objects of the same category, and $|C_i|$ represents the number of data objects contained in c_i in formula (3.7). The σ_i is smaller, the data similarity objects in the cluster is greater, the data objects are denser, and the clustering effect of selecting the centroid in the cluster is better. Therefore, when the initial clustering center of the optimization algorithm is selected, the Euclidean distance between two data points is weighted to measure the similarity of data points. A more accurate initial clustering center can be obtained by ordering the value of σ_i of the objective value function. The running process of ED-K-Means algorithm is shown in Figure 3.2. K initial clustering centers are selected by entropy method, and the remaining data are clustered by calculating the assigned Euclidean distance. The clustering center is repositioned by the target value function σ_i and iterated repeatedly until the clustering requirements are met.

3.2. Improvement of college students' behavior association analysis algorithm. The constructed K-Means can be used for clustering analysis of student consumption, learning, and other data. Subsequently, the improved Apriori algorithm is used to explore the correlation between student grades, consumption records, learning behavior, and other student behaviors. The Apriori algorithm has high efficiency and wide application in association rule mining. Through frequent itemset generation and pruning, it can effectively discover association relationships in the dataset, making it particularly suitable for processing large-scale data. In addition, the attribute relationships between various types of data in universities are complex and diverse. The Apriori algorithm can use raw data analysis to extract potential rules and information between data, calculate their association rules, and help users quickly grasp information and make effective decisions [18]. Therefore, the study utilizes the Apriori algorithm to explore the deep correlation between student grades, consumption records, and learning behavior, providing a basis for improving student management and optimizing resource allocation. Apriori algorithm's core idea is to mine the set of frequent items from the target database by means of layer-by-layer iteration and search, analyze the association rules between frequent item sets, and find out the association rules between the target data. Assume that there is a transaction database that is T, the

transaction sets A and B meet $A \subseteq T, B \subseteq T$, and $A \cap B = \emptyset$. The percentage which is the union of A and B in the transaction database in all transactions is called support, and the calculation formula is (3.8).

$$Sup\,(A \Rightarrow B) = \frac{count\,(A \Rightarrow B)}{T} \tag{3.8}$$

In the transaction database, for item set P, the number ratio of the item occurrences set to the total is called absolute support. Then the absolute support expression of item set P in database T is shown in formula (3.9).

$$Sup\,(P) = \frac{count\,(P)}{T} \tag{3.9}$$

In formula (3.9), $count\,(P)$ represents the number of occurrences of item set P in database T. By comparing the size of support and minimum support, it can determine whether the item set is frequent. For item set P, if $Sup\,(P) \geq \min \sup$ is met, item set P is frequent. The minimum support is a threshold set by yourself. It can be adjusted for the actual situation. Through the minimum support, all frequent item sets can be found. For association rule $A \Rightarrow B$, it is also necessary to determine whether it meets the needs of users. The confidence level of rule $A \Rightarrow B$ is set as the number ratio of A and B union in the transaction database to the number of A in the transaction database. Formula (3.10) is the specific confidence calculation.

$$Conf\,(A \Rightarrow B) = \frac{count\,(A \Rightarrow B)}{count\,(A)} \tag{3.10}$$

The minimum confidence threshold can be used to measure whether the rule is reliable. When the support and confidence of rule $A \Rightarrow B$ are not less than the threshold, the rule is a strong association rule. The original Apriori algorithm scans the target object database many times, which will cause heavy burden on the algorithm operation. Reducing the scans number is the most direct and effective way. Mapping Apriori uses the "mapping" principle to save and compress the object data structure in the database, which can significantly reduce Apriori algorithm calculation support complexity and calculations as much as possible. In Mapping-Apriori algorithm, it is assumed that the expression of the set of all items and the target object library is Formula (3.11).

$$\begin{cases} F = \{F_1, F_2, \ldots, F_l\} \\ D = \{T_1, T_2, \ldots, T_l\} \end{cases} \tag{3.11}$$

Among them, the identification of $T \subset F$ is T_{ad}. According to the mapping principle, the storage structure of target object library D can be designed as Figure 3.3.

In Figure 3.3, the values corresponding to keys F_k $(1 \leq k \leq l)$ and F_{sum} are one-dimensional arrays. Formula (3.12) is the definition formula of d_{ab}.

$$d_{ab} = \begin{cases} 0, F_a \notin T_b \\ 1, F_a \in T_b \end{cases}, 1 \leq a \leq |F|, 1 \leq b \leq y, y = |D| \tag{3.12}$$

$|D|$ represents the number of things in formula (3.12). s_j represents the number of item sets contained in the target database. Formula (3.13) is the definition.

$$S_b = \sum_{a=1}^{l} d_{ab}, l = |F| \tag{3.13}$$

$|F|$ is the number of item sets in formula (3.13). The value of RC represents the number of duplicate things in the transaction database. Different situations need to be discussed. If there are no duplicate things in the transaction database, the value of RC is 1. Formula (3.14) is expression.

$$RC_a = 1, \quad if \quad T_{sb} \neq T_{tb}\,(s \neq t)\ , b = 1, 2, \ldots y; a, s, t = 1, 2, \ldots, l \tag{3.14}$$

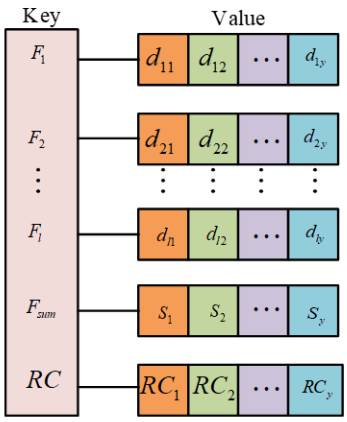


Fig. 3.3: Storage structure diagram of target thing library

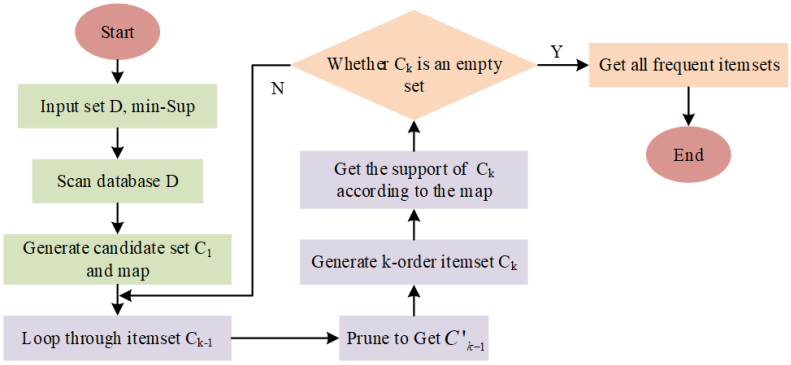


Fig. 3.4: Mapping-Apriori algorithm flow chart

If one thing is repeated, a value is assigned to RC, and Formula (3.15) is the specific expression.

$$RC_a = RC_a + 1, \quad if \quad T_{sb} = T_{tb}\,(s \neq t) \quad , b = 1, 2, \ldots y; a, s, t = 1, 2, \ldots, l \tag{3.15}$$

By assigning a value to RC, the map size can be reduced effectively. The Maping-Apriori algorithm uses the mapping data structure to reduce the complexity of computing item set support, and also pre-prunes frequent item sets to reduce a large number of comparisons. The execution process of the Mapping-Apriori algorithm is shown in Figure 3.4. The target transaction database D is entered with minimum support. C_1 and mapping can be obtained by scanning the database. The $k-1$ order frequent item set C_{k-1} can be obtained by cycle operation. C'_{k-1} can be obtained by pruning C_{k-1}, and the k order candidate set C_k can be generated. Calculate the support of C_k according to the array of corresponding item sets obtained from the mapping, and compare it with the threshold to obtain all frequent item sets.

subsectionPreprocessing of college students' behavior data

As information technology develops, the SC system has produced massive data, but there are inevitably incomplete and low-quality data in it. The reasons for generating these data are various, mainly including collection errors, inconsistent data format, and the lack of data that cannot be obtained [19]. These data cannot be directly analyzed, and need to be processed into relevant data that can be directly analyzed. The preprocessing technology came into being. Data preprocessing is the premise of data mining analysis. The data

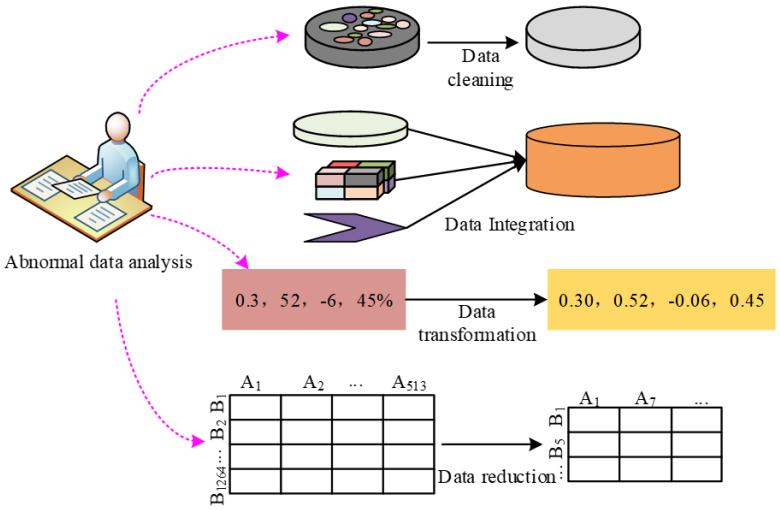


Fig. 3.5: Data preprocessing method flow

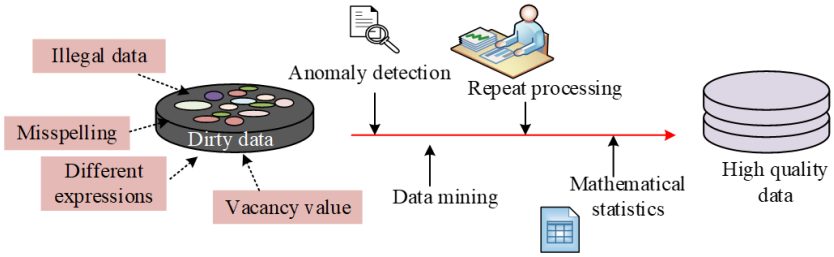


Fig. 3.6: Data cleaning principle

after data preprocessing is directly related to the analysis results. Data preprocessing methods consists of data cleaning, integration, transformation and reduction. Figure 3.5 is the specific process.

Data cleaning is to find and correct identifiable errors in data files. It will check data consistency, process invalid data and missing values. After getting the data, you should first check which data is unreasonable and its basic situation, and then clean it through the common methods of data governance. Common data cleaning methods include manual cleaning of missing value data, that is, replacing the missing value with the average value, maximum and minimum value or probability estimate. For noise data processing, it is usually to delete isolated points isolated from other data. The disadvantage of this method is that it may delete valuable data. Inconsistent data is mainly corrected by referring to paper records. Figure 3.6 shows the principle of data cleaning.

Data integration is because the data to be integrated is obtained from the databases of multiple application systems. Because its different formats, attributes and characteristics need to be collected into a database for analysis, data from different data sources need to be sorted and consolidated into data storage with consistent characteristics. Good data integration can reduce the result data set redundancy and inconsistency, and improve the accuracy and efficiency of its subsequent mining process. Data transformation refers to the normalization of data to achieve the purpose of mining. Data transformation mainly involves smoothing, data aggregation, data generalization, and data normalization.

Data reduction refers to a process of minimizing the data and maintaining its integrity as much as possible.

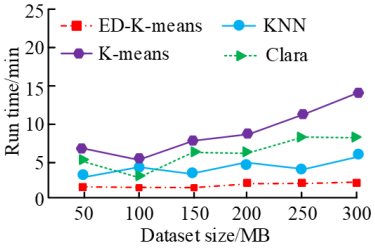


Fig. 4.1: Running time results of four algorithms

Although large databases can be used for data mining, the selected mining algorithm may not be applicable due to its high dimension and large amount of data. The large amount of calculation will cause heavy load on the equipment. It is necessary to reduce large data sets to small data sets as much as possible. Without changing the mining analysis results, reducing the data size and data reduction can improve the mining efficiency.

4. Analysis on data performance and application of mining technology in college students' behavior association analysis.

4.1. Improved clustering algorithm and association rule algorithm performance test. To analyze the proposed ED-K-Means' performance, a comparative experiment was conducted with other clustering algorithms. Clara (Clustering LARge Applications), traditional K-Means and KNN (K-Nearest Neighbor) were selected. The test data was a self-made database composed of college students' consumption behavior, score records, book borrowing, etc. The data collection process involves multiple steps. Firstly, by collaborating with relevant departments on campus, we obtained student consumption records, academic performance data, and library borrowing records. These data cover students of different grades, majors, and genders, and have high comprehensiveness and representativeness. During the data cleaning and preprocessing process, missing and outliers were removed to ensure data quality and consistency. The final dataset includes multiple behavioral characteristics and academic performance, which can comprehensively reflect the overall behavioral patterns of college students and provide a reliable data foundation for the performance evaluation of clustering algorithms. The results obtained from the running time analysis are in Figure 4.1.

The runtime curves of the four algorithms show an upward trend with the increase of the dataset in Figure 3.6. Among them, K-Means has the longest running time, consuming more than 5min, the longest running time is 15.08min, and the average running time is 10.36min. The maximum running time of Clara is 8.97 min, the average time is 6.22 min, and the running speed is accelerated. The average operation time of KNN is 5.11min, and the longest time is 6.18min. The operation speed and stability are further improved. ED-K-Means has the best performance in running time and stability, with an average running time of 1.96 minutes. Compared with KNN, Clara and K-Means, its running efficiency has been improved by 81.1%, 68.5% and 61.6% respectively. This shows that the idea of information entropy combined with density optimization can improve the algorithm efficiency, and has better performance.

To test the improved Mapping-Apriori algorithm effectiveness, the performance test is carried out with Apriori and Frequent Pattern Growth (FP-Growth) algorithm under the same conditions. Figure 4.2 shows the loss curve.

The change trend of the loss curve of the three algorithms is similar, and the decline speed and fluctuation range of the loss value are different. Among them, Apriori's loss value decreases slowest and gradually converged to within 40 after 150 iterations, and the loss value basically fluctuated between 30-45 after 200 iterations. The decline speed of FP-Growth loss curve was accelerated, the loss value dropped to within 30 after 150 training, and then fluctuated between 20-30, and the stability is improved. The loss curve of Mapping-Apriori decreases the fastest, and the loss value is within 20 after 80 training. With training times increasing, the fluctuation range of the loss value is between 10-15, and the algorithm has the best stability. It shows that Apriori improvement by mapping can improve the iterative algorithm efficiency and stability.

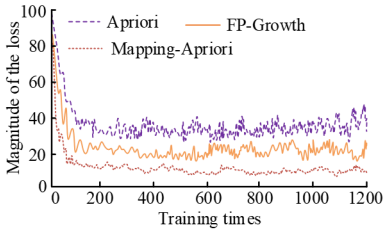


Fig. 4.2: Loss curve results of three algorithms

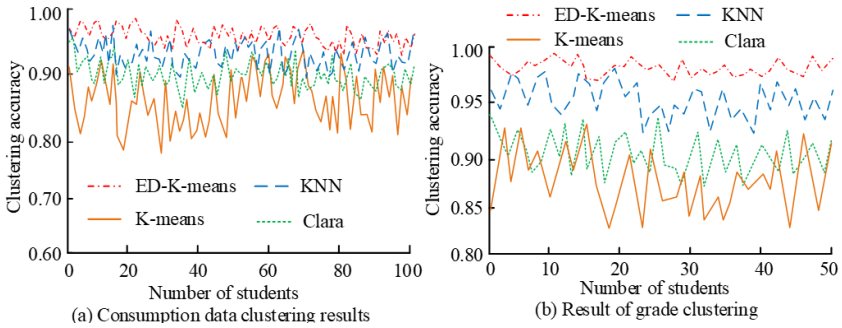


Fig. 4.3: Cluster accuracy results

4.2. Student data clustering results of improved K-means method. The consumption behavior and performance data of students were collected in a certain major in the SC database. Several clustering algorithms are applied for clustering analysis of relevant data, and the clustering accuracy is used for evaluation and analysis. Figure 4.3 shows the clustering accuracy results obtained from the two data.

From Figure 4.3(a), the clustering accuracy of K-Means is poor, with an average accuracy of 86.33%. The stability of the algorithm is also poor, and the accuracy curve fluctuates greatly, with the maximum difference of 16.9%. Clara's clustering accuracy has increased, with an average accuracy of 89.82%. Its stability has improved, with a fluctuation of 10.7%. The accuracy curve of KNN fluctuates by 8.6%, and the stability of the algorithm is further strengthened. The average clustering accuracy is 93.65%, and the clustering effect is good. The clustering accuracy of ED-K-Means is mostly above 95%, with the smallest fluctuation range and the largest difference of 5.9%. The average accuracy of clustering is 97.41%. Compared with the comparison algorithm, the accuracy is improved by 12.8%, 8.5% and 4.0%.

Figure 4.3(b) shows the clustering results of 50 students' performance data. It can be seen that the accuracy of several algorithms is above 80%, and ED-K-Means clustering has the best accuracy and stability. The average clustering accuracy of K-Means, Clara, KNN and ED-K-Means is 87.18%, 91.27%, 94.88% and 97.83% respectively. Compared with the comparison algorithm, the accuracy of ED-K-Means is 10.65%, 6.56% and 2.95% higher. According to Figure 10, ED-K-Means has the best clustering accuracy and stability in the face of different clustering data. The clustering results of students' monthly consumption level using ED-K-Means are shown below.

When the total monthly consumption of a student is more than 600 yuan in Table 4.1, the student is at a high consumption level and belongs to non-poor students. Students whose consumption amount is between 200 and 600 belong to poor students, and the consumption amount is less than 200, which indicates that the student's consumption is extremely low and belongs to extremely poor students. According to the clustering results, the total consumption of most students is below 600. The school can provide decision-making basis for the assessment of poor students based on relevant information.

Table 4.1: Student monthly consumption level clustering results

Consumption level	*Cluster center range*	*Cluster center value*	*Percentage of students*
High	Total amount>600 yuan	629.88	16.67%
Medium	200-600 yuan	447.29	57.19%
Low	Total amount <200 yuan	179.45	26.14%

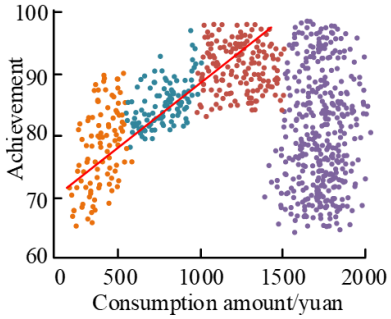


Fig. 4.4: The correlation between student achievement and consumption

Table 4.2: Association result

Result	*Consumption level*	*Consumption habits*	*Academic record*	*Learning behavior*	*Conf*
1	High	Regular	A(>85)	Regular	0.153
2	High	Regular	B(75-84.5)	Regular	0.162
3	Medium	Regular	A(>85)	Regular	0.417
4	Medium	Irregular	C(60-74.5)	Regular	0.118
5	Medium	Regular	A(>85)	Regular	0.474
6	Medium	Irregular	D(<60)	Irregular	0.122
7	Low	Regular	A(>85)	Regular	0.453
8	Low	Irregular	D(<60)	Irregular	0.136

4.3. Results of college students' behavior correlation analysis. The research establishes a relevant database through cluster analysis of students' consumption records, grades and learning behaviors. And Maping-Apriori is used to explore the correlation between student performance and consumption behavior rules, consumption level and learning behavior. Figure 4.3 shows the correlation between student achievement and consumption.

From Figure 4.4, when the amount of consumption is less than 1500 yuan, there is a certain positive correlation between the level of consumption and student performance. At the initial stage, with the increase of the amount of consumption, the student's performance improved, but after the amount of consumption reached 1500, the positive correlation disappeared. The minimum support is set to 0.15 and the confidence is set to 0.85. Table 4.2 shows the correlation results.

The consumption rules in Table 4.2 mainly refer to the consumption records of students' three meals a day, and the learning behaviors mainly refer to the records of access to the library and dormitory. In Table 3, there is a certain correlation between students' grades and their diet rules, library and dormitory access rules. Students with good grades have relatively regular life. This is because students who insist on eating breakfast on time generally have better self-discipline and can arrange time scientifically. At the same time, students who insist on eating breakfast on time can also improve their mental state in class at ordinary times. Among them,

most of the students with moderate consumption level, regular consumption and regular learning behavior are "A" and "B" in their academic achievements; The students with irregular consumption and learning behavior are mostly "C" and "D" in their academic performance. Through correlation analysis, colleges and universities can observe the characteristics of students' daily behavior in school, help and guide students to develop good living and learning habits, and promote students' healthy growth and smooth completion of their studies.

5. Conclusion. The overall operation effect is closely corresponding to the management of students' behavior. Strengthening the management of students' behavior will improve the overall operation efficiency. In face of massive data in SC, the improved ED-K-Means method is proposed to cluster relevant data, and the optimized Mapping-Apriori algorithm is used for association analysis. Through the performance test, the average ED-K-Means running time was 1.96 minutes. Compared with the comparison algorithm, the running efficiency was improved by 81.1%, 68.5% and 61.6%. The loss curve of Mapping-Apriori decreased the fastest, the fluctuation range of loss value was the smallest, and the algorithm stability was the best. Through cluster analysis, the results showed that the average accuracy of ED-K-Means clustering was 97.41% for student consumption data. In terms of student achievement data, the accuracy of the algorithm was as high as 97.83%, which was better than the comparison algorithm. At the same time, the algorithm divided different consumption levels and performance grades according to the total monthly consumption. The correlation results obtained from Mapping-Apriori show that students with good grades live relatively regularly. Through the analysis of relevant data, colleges and universities can help poor students to carry out the assessment work smoothly, and guide students to cultivate good habits. The research mainly uses structured data in the SC system. The unstructured data has not yet been analyzed and processed. In the future, unstructured data such as pictures and videos will be analyzed to improve the reliability of the results.

Fundings. The research is supported by: Vocational Education Teaching Reform Research Project of Hunan Province in 2020;Research on the construction of process learning evaluation system of online courses in higher vocational colleges under the background of big data; No.: ZJGB2020019.

REFERENCES

[1] Z. Y. Dong, Y. Zhang, C. Yip, S. Swift, and K. Beswick, "SC: Definition, Framework, Technologies, and Services," *IET Smart Cities*, vol. 2, no. 1, pp. 43–54, 2020.
[2] M. A. Razzaq, J. A. Mahar, M. A. Qureshi, and Z. Abidin, "SC System Using Internet of Things: Simulation and Assessment of Vertical Scalability," *Indian J. Sci. Technol*, vol. 13, no. 28, pp. 2902–2910, 2020.
[3] A. Adnan, A. Aiyub, and A. Roziq, "SC Online Lecture Model as an Alternative Education Equity Strategy in the Era of Super Smart Society 5.0," *International Journal of Educational Review, Law and Social Sciences (IJERLAS)*, vol. 2, no. 1, pp. 207–210, 2022.
[4] H. Lutfie, "Effectiveness of Marketing Technology Website Quality on Company Performance and the Impact on SC Student Satisfaction," *Jurnal Aplikasi Manajemen*, vol. 18, no. 1, pp. 181–188, 2020.
[5] W. Li, "Design of SC Management System Based on Internet of Things Technology," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 2, pp. 3159–3168, 2021.
[6] C. Prandi, L. Monti, C. Ceccarini, and P. Salomoni, "SC: Fostering the Community Awareness Through an Intelligent Environment," *Mobile Networks and Applications*, vol. 25, pp. 945–952, 2020.
[7] R. Jurva, M. Matinmikko-Blue, V. Niemelä, and S. Nenonen, "Architecture and Operational Model for SC Digital Infrastructure," *Wireless Personal Communications*, vol. 113, pp. 1437–1454, 2020.
[8] B. Valks, M. H. Arkesteijn, A. Koutamanis, and A. C. Heijer, "Towards a SC: Supporting Campus Decisions with Internet of Things Applications," *Building Research & Information*, vol. 49, no. 1, pp. 1–20, 2021.
[9] X. Zhang, J. Shen, P. Wu, and D. Sun, "Research on the Application of Big Data Mining in the Construction of SC," *Open Access Library Journal*, vol. 8, no. 11, pp. 1–10, 2021.
[10] J. Fan, Y. Jiang, Y. Liu, et al., "Interpretable MOOC Recommendation: A Multi-Attention Network for Personalized Learning Behavior Analysis," *Internet Research*, vol. 32, no. 2, pp. 588–605, 2022.
[11] D. L. Roorda, M. Zee, and H. M. Y. Koomen, "Don't Forget Student-Teacher Dependency! A Meta-Analysis on Associations with Students' School Adjustment and the Moderating Role of Student and Teacher Characteristics," *Attachment & Human Development*, vol. 23, no. 5, pp. 490–503, 2021.
[12] J. S. Bureau, J. L. Howard, J. X. Y. Chong, and F. Guay, "Pathways to Student Motivation: A Meta-Analysis of Antecedents of Autonomous and Controlled Motivations," *Review of Educational Research*, vol. 92, no. 1, pp. 46–72, 2022.
[13] Y. Li, G. Li, L. Liu, and H. Wu, "Correlations Between Mobile Phone Addiction and Anxiety, Depression, Impulsivity, and Poor Sleep Quality Among College Students: A Systematic Review and Meta-Analysis," *Journal of Behavioral Addictions*, vol. 9, no. 3, pp. 551–571, 2020.

[14] M. Pérez-Pérez, A. M. Serrano-Bedia, and G. García-Piqueres, "An Analysis of Factors Affecting Students Perceptions of Learning Outcomes with Moodle," *Journal of Further and Higher Education*, vol. 44, no. 8, pp. 1114–1129, 2020.
[15] O. Doğan and E. C. Tırpan, "Process Mining Methodology for Digital Processes Under SC Concept," *Bilecik Şeyh Edebali Üniversitesi Fen Bilimleri Dergisi*, vol. 9, no. 2, pp. 1006–1018.
[16] K. Ansong-Gyimah, "Students' Perceptions and Continuous Intention to Use E-Learning Systems: The Case of Google Classroom," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 15, no. 11, pp. 236–244, 2020.
[17] L. Huiwei, Y. Biyu, Z. Xiaoqi, F. Xueying, W. Qinlin, and L. Xuan, "Correlation Analysis of Mood Disorders and Behavioral Patterns Among Patients with Essential Hypertension in Community Settings," *Journal of New Medicine*, vol. 51, no. 3, pp. 180–183, 2020.
[18] Y. Xiao, X. Ren, P. Zhang, and A. Ketlhoafetse, "The Effect of Service Quality on Foreign Participants' Satisfaction and Behavioral Intention with the 2016 Shanghai International Marathon," *International Journal of Sports Marketing and Sponsorship*, vol. 21, no. 1, pp. 91–105, 2020.
[19] N. Manirochana, "Relationships Between Service Quality, Service Marketing Mix, and Behavioral Intention: Consumers' Perspectives on Short-Term Accommodation Service for Tourism in Thailand," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 8, pp. 2991–2999, 2021.

Edited by: Achyut Shankar
Special issue on: Machine Learning for Smart Systems: Smart Building, Smart Campus, and Smart City
Received: Mar 24, 2023
Accepted: Aug 31, 2024