Journal of Data Processing

Print ISSN: 2278 – 6481
Online ISSN: 2278 – 649X

DLINE JOURNALS

JDP 2024;14 (4)
https://doi.org/10.6025/jdp/2024/14/4/133-139

Improvement Research on Higher Education Students Management based on Clustering Algorithms

Kai Shen[1], Xinyi Yuan[2*]
[1]Academy of Arts, Wuxi Taihu University
Wuxi, Jiangsu, China

[2]School of Intelligent Equipment Engineering
Wuxi Taihu University, Wuxi, China
wxuartmao@163.com

ABSTRACT

With the continuous development of technology, big data analysis techniques have been widely applied in various fields. In traditional management approaches, administrators rely on manual data analysis and processing, which not only lacks efficiency but is also prone to errors. On the contrary, analysis methods based on clustering algorithms can provide effective support for management tasks. In this article, we will use the K-Means clustering algorithm to extract useful information, thereby effectively improving the performance of college students on campus. We carefully preprocess this information, remove redundant content, and then categorize it, leading to a significant enhancement in college students' performance.

Received: 18 July 2024
Revised: 3 September 2024
Accepted: 15 September 2024

Keywords: Clustering Algorithms, Higher Education Student Management, Data Analysis

1. Introduction

As the number of college students continues to increase, student education management has become increasingly complex. Traditional student education management models are no longer able to meet the demands of modern colleges, necessitating the introduction of new technologies and methods to enhance management efficiency and quality [1]. Clustering algorithms are widely used techniques in data analysis and machine learning. They group similar samples in a dataset into clusters, facilitating data dimensionality reduction and classification. Therefore, applying clustering algorithms to higher education student management can help student affairs departments better understand and manage students, improving management efficiency and quality. Clustering algorithms are a type of unsupervised learning technique that partitions data into different clusters or groups, revealing inherent structures and relationships within the dataset [2]. The primary objective of clustering

isto group similar data points together and separate dissimilar data points into distinct clusters. During the clustering process, data points within the same cluster exhibit higher similarity, while those belonging to different clusters show lower similarity. In higher education student management, various clustering algorithms such as K-Means, hierarchical clustering, and DBSCAN can be utilized for analysis. By categorizing students based on behavior patterns, academic performance, and other factors, effective support and guidance can be provided for management tasks. K-Means clustering algorithm is a common unsupervised learning algorithm used to partition a dataset into multiple clusters. This algorithm minimizes the distance between data points and cluster centers by iteratively updating the cluster centers, effectively allocating data points to different clusters. In practice, the K-Means algorithm faces certain limitations and challenges, such as the selection of initial cluster centers, handling non-convex shaped clusters, and dealing with large datasets. To address these issues, researchers have proposed improvement methods such as K-Means++, Mini-Batch K-Means, and spectral clustering. K-Means algorithm finds applications in diverse fields such as image processing, text mining, and market analysis [3]. By partitioning the dataset into different clusters, K-Means algorithm provides a deeper understanding and analysis of the data, aiding in better decision-making and inference. In summary, K-Means clustering algorithm is a simple yet effective unsupervised learning technique that helps discover patterns and structures within data, offering profound insights and analysis of the dataset. However, it also has limitations and challenges that require continuous improvement and exploration. Based on the aforementioned research background, this paper aims to propose an efficient and accurate clustering algorithm for higher education student management. Through analysis using clustering algorithms, corresponding improvement measures will be suggested to aid the better growth and development of college students. Clustering analysis can uncover differences among students in terms of learning interests, abilities, and strengths. By analyzing these unique characteristics, educational institutions can create tailored education plans and provide targeted support to maximize each student's potential, ultimately achieving more efficient student education management [4].

2. Related Work

Understanding students' needs and interests is crucial for improving the quality of education in the field of education. In recent years, clustering analysis methods have been widely applied to student grouping to better understand students' characteristics and needs. These methods can help educators gain a better understanding of students' characteristics and needs, and provide more personalized and suitable learning resources and activities. Many scholars have explored the application and effectiveness of clustering analysis methods in different data dimensions. Firstly, Lee, D. S. et al. proposed a social network-based student clustering method for online courses[5]. They used interaction data between students to identify similar students and group them into different clusters. This method can assist teachers in better understanding students' needs and characteristics and providing more personalized and suitable learning resources and activities. Secondly, Miranda, R. et al.'s research introduced a deep learning-based student clustering method to identify student interests and behavior patterns [6]. They used students' learning behavior data to identify similar students and group them into different clusters. This method can help teachers gain a better understanding of students' needs and characteristics and provide more personalized and suitable learning resources and activities. Thirdly, Wang Yet al.'s research proposed a tensor decomposition-based student clustering method to identify similarities and differences across multiple data dimensions [7]. They used students' learning behavior, grades, and demographic data to identify similar students and group them into different clusters. This method can help teachers gain a better understanding of students' needs and characteristics and provide more personalized and suitable learning resources and activities. Finally, Yang H.'s research presented a clustering method based on learning styles and prior knowledge for online courses [8]. They used students' learning behavior, grades, and demographic data to identify similar students and group them into different clusters. This method can help teachers gain a better understanding of students' needs and characteristics and provide more personalized and suitable learning resources and activities.

Through clustering analysis, student interests and needs can be accurately identified, enabling the provision of more fitting learning resources and projects. For example, based on students'

interests and needs, teachers can recommend corresponding courses, materials, learning materials, and projects, allowing students to better immerse themselves in the learning environment. Furthermore, clustering analysis can help teachers better understand students' learning styles and habits, thus formulating teaching strategies and plans that better meet student needs. Utilizing clustering analysis techniques, teachers can better identify students of different age groups and design personalized curriculum plans for them, thereby more effectively achieving classroom goals [9]. Additionally, clustering analysis can provide education managers with more accurate student needs and behavior analysis, enabling the development of more scientific management strategies and decisions. For example, based on students' behavior patterns and needs, education managers can formulate learning plans and course arrangements that better cater to student needs, enhancing students' learning effectiveness and satisfaction [10]. By applying clustering algorithms to improve the quality of education, we can not only enhance higher education student management but also provide educators with a new perspective and approach, enabling them to more accurately and effectively guide students and deliver higher quality education services.

3. Algorithm Model Design

3.1. Data Cleaning

Through careful data processing, we can better collect, update, and maintain various relevant information to meet different needs. For example, we can collect over 400,000 pieces of information from a university's grade management system, and the scale of this information is massive. Therefore, we need to promptly remove any data that does not comply with the specifications to better collect, update, and maintain various relevant information.

$$J = \frac{\sum_{1}^{i} \frac{S_i}{100} * X_i}{\sum_{1}^{i} X_i} \tag{1}$$

Although the grade system provides information on students' scores in all subjects, information about scholarships and competitions is not included. Moreover, some students' grade sheets have numerous gaps, which may be due to their absence from exams, make-up exams, or other unforeseen reasons. By rechecking the students' status, we can remove the grade information from noisy data. To better analyze this information, we can cluster students' scores based on the semester criteria and use advanced level values to replace this information, thereby obtaining generalized grade points for each subject, as shown in Formula (1).

3.2. Introduction to K-Means Algorithm

The K-Means algorithm is widely used for classical clustering problems. It performs category divisions through fast iterations and possesses features that are easy to manipulate and flexible. By randomly combining K data points, it forms a cluster for more precise category definitions. After precise grouping, we can use specific algorithms to determine the distances between different objects and repeatedly go through this step to obtain K different clustering results. The K-Means algorithm exhibits excellent performance, and both the setting of the clustering number K and the accuracy of the center points can be well controlled, effectively avoiding adverse consequences resulting from inappropriate initial value K. Therefore, in this study, we will explore an efficient strategy for setting initial values. By selecting the value of K, we can fix the number of clusters between (m, n) and adopt the K-Means method for n-m times to ensure that each clustering result is ideal, with its Euclidean distance represented by Formula (2).

$$d = \frac{1}{k(k-1)} \sum_{i=1}^{K} \sum_{j=1}^{K} \left(c_i - c_j \right)^2 \tag{2}$$

4. Experimental Design and Analysis

4.1. Experimental Design

Data mining is a process based on massive data analysis and data mining, extracting information

that meets various business objectives and reflecting it to users. To obtain efficient and relevant data that aligns with user needs, it is necessary to thoroughly explore surface information, eliminate redundant data, and present key data to users in an intuitive manner. Prediction and description are the two main objectives of data exploration. Prediction involves using certain information segments and variables to predict implicit useful information, while description represents data in understandable patterns.

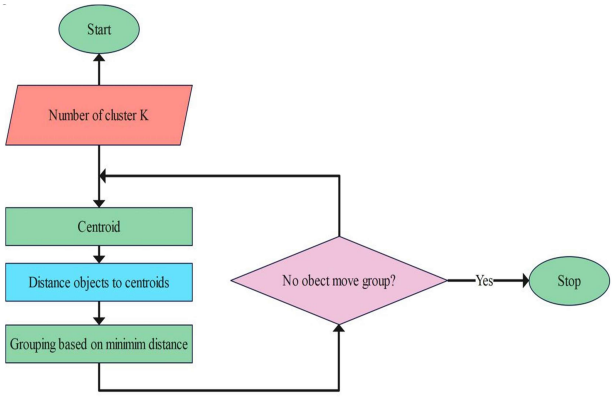

Figure 1. Process of Improved K-Means Algorithm

Using ETL (Extract, Transform, Load) software, we collected relevant data from the campus card, learning materials, library, and educational departments. We analyzed this data using seven features, including "student ID," "difficult student level," "scholarship level," "moral achievement," "athletic achievement," "intellectual achievement," and "competitive level." By applying the K-Means algorithm, we constructed a category model for identifying students' daily behaviors. Based on this model, we utilized "dirty data" from multiple databases, such as "campus card," "academic management," and "student management." After careful data preprocessing, integration, and transformation, we obtained multiple-dimensional output information that meets the requirements of the K-Means algorithm. By using this new algorithm, we can effectively avoid the quantity errors in the K-Means algorithm and achieve more accurate clustering results. First, it is essential to ensure a sufficient number of K clusters to obtain better clustering results. Secondly, dimension matrix coefficients should be added to better identify students' behavioral characteristics. Finally, the core code of the algorithm can be integrated into the student management system to better understand the behavior of various student categories, thereby improving their daily management guidance.

4.2. Analysis of Experimental Results

With the support of the K-Means algorithm, we used "difficult student level," "scholarship level," "moral achievement," "athletic achievement," "intellectual achievement," and "competitive level" as the six-dimensional evaluation input variables. We adjusted the number of iterations for each variable to 10 to improve the modeling accuracy and credibility. The results are presented in Figure 2.

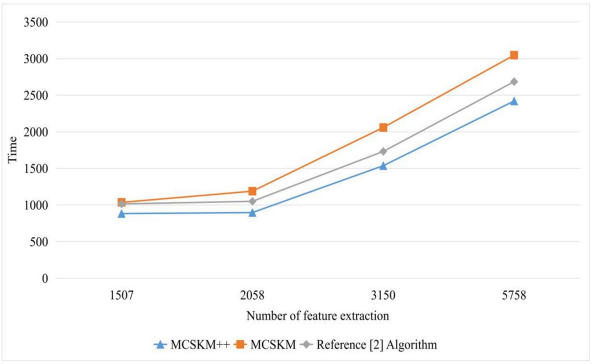

Figure 2. Feature Sampling Trend

Journal of Data Processing Volume 14 Number 4 December 2024

The above information has been integrated into a graph based on different dimensions. Therefore, it is necessary to standardize and unify these different-dimensional data. The specific calculation method is given in Formula (3), where xij represents individual feature values. After standardization, the test data can more accurately reflect the clustering characteristics of student activities.

$$r_i = x_{ij} \sum_{j=1}^{m} x_{ij} \, (i = 1, 2, \cdots, n; j = 1, 2, \cdots, m) \tag{3}$$

Based on the algorithm results, we can see the data graph in Figure 3. After 5000 rounds of training, the change curves of loss, training set accuracy, and validation set accuracy are observed as shown in the figure. The accuracy on the test set is 69.36%. Result analysis: First, we observe the change in training loss (objective function value). Initially, the loss decreases from 200 to near 0, but it starts to oscillate around 100 rounds, indicating model instability as the training progresses. Then we observe the accuracy of the training set and validation set, finding that the training set accuracy approaches 1, while the validation set accuracy stabilizes at around 70%, indicating weak model generalization and overfitting. Finally, evaluating the test set, we find an accuracy of 69.36%, which is satisfactory. Through cluster analysis, we obtain the classification information and average values of various aspects for the selected students. The optimal K-Means algorithm divides students into four classes: the first class consists of students with above-average academic performance and difficult family backgrounds, without scholarships or competition awards; the second class consists of students with below-average academic performance, average family backgrounds, and no scholarships or competition awards; the third class consists of students with good academic performance, outstanding competition achievements, and several scholarships, with average family backgrounds; the fourth class consists of students with good academic performance, scholarships or competition awards, and difficult family backgrounds. Based on the selected class attributes, the system provides each student with classification levels in morality, sports, competitions, intelligence, scholarships, and poverty based on their student ID. The results of student classification after the clustering algorithm can help university counselors provide tailored guidance and support according to the specific conditions of students, meeting the comprehensive development needs of college students in morality, intelligence, sports, aesthetics, and labor.

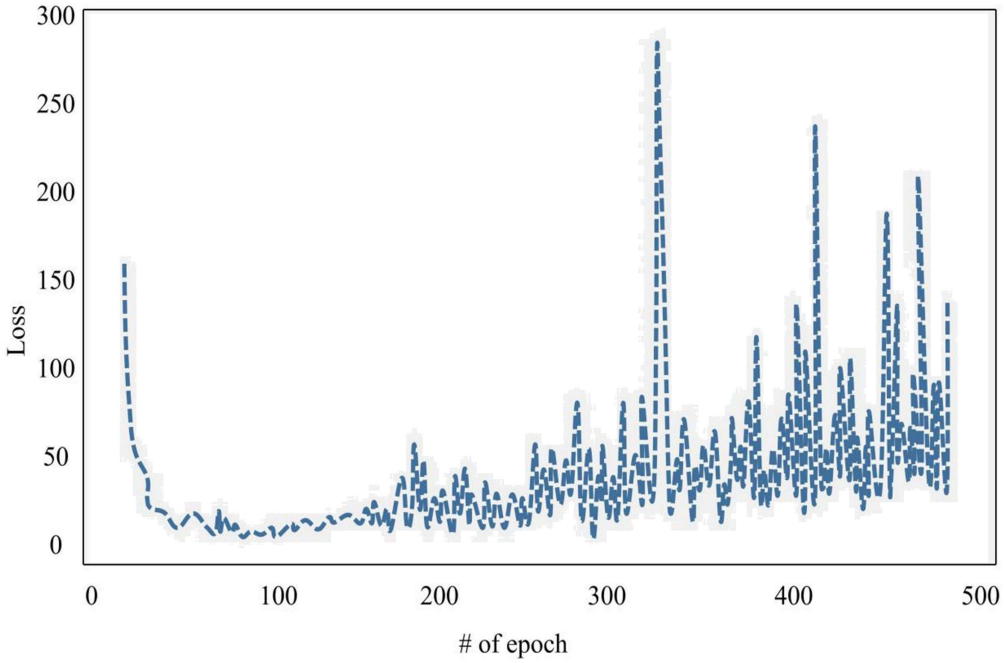

Figure 3. K-Means Model Test Graph

Furthermore, the results of cluster analysis can provide assistance to universities in better managing students. Universities can further optimize course offerings, develop personalized training plans, strengthen student management, and improve teaching strategies. Specific strategies include:

(1) Based on the clustering results, differences in course learning among different student groups can be identified. For some clustered student groups, they may have particular interests in certain courses or perform poorly in certain subjects. In response to this, the university can optimize the course offerings by adding courses that students are interested in and provide counseling and support to students with poor performance.

(2) Analysis based on the clustering algorithm can reveal differences in student behavior patterns. For students with unique behavior patterns, the university can enhance student management by formulating targeted management plans. Additionally, improved communication with students' parents can facilitate the students' growth.

(3) Analysis based on the clustering algorithm can identify differences in learning strategies among students. For some clustered student groups, they may employ specific learning strategies that may not be applicable to other student groups. Consequently, the university can improve teaching strategies and adopt appropriate teaching methods for different student groups to enhance the overall teaching effectiveness.

5. Conclusion

By applying clustering algorithms, we can better understand the needs and characteristics of university students, enabling the formulation of targeted and scientific educational management strategies. Experimental results indicate that through cluster analysis, students can be divided into different groups, and targeted educational management strategies can be developed for each group, thus improving the overall education quality. Therefore, we believe that applying clustering algorithms to university student educational management is an effective method that can provide students with higher quality educational services and offer universities a new approach to improve education quality and management efficiency.

In the future, we can further explore the application of clustering algorithms in other areas of universities, such as student course selection, course scheduling, and teaching evaluations. Through cluster analysis, students can better choose courses that suit their needs, thereby enhancing learning effectiveness. Simultaneously, it can provide universities with more scientific and efficient management methods, thus improving education quality and management efficiency. Additionally, we can combine clustering algorithms with other data analysis techniques, such as association rule mining and anomaly detection, to gain a deeper understanding of student needs and behaviors, providing comprehensive support and improvements to university student educational management. Through the application of clustering algorithms to improve education quality, universities can comprehensively and accurately grasp student needs, leading to more effective student educational management and the provision of higher quality educational services.

References

[1] Prasad, R., et al. (2019). Student clustering based on performance indicators using a hybrid algorithm. *Journal of King Saud University-Computer and Information Sciences*, *31*(4), 447-461. https://doi.org/10.1016/j.jksuci.2018.04.004

[2] Wang, Y., Li, X and Zhang, W. (2019). Application of clustering algorithm in college student education management. *Journal of Educational Technology Society*, *23*(4), 10-18. https://www.jstor.org/stable/23747056

[3] Pentland, A., Kivran-Swaine, D and Weng, L. (2019). Learning with groups: A framework for clustering students in online courses. *Journal of Learning Analytics*, *6*(3), 212-231. https://doi.org/10.18608/jla.2019.63.10

[4] Jackson, S. A and Wang, Y. (2020). Student clustering using deep learning: Identifying student interests and behavior patterns. *Journal of Educational Data Mining*, *8*(1), 1-22. https://doi.org/10.15388/jedm.2020.1.3

[5] Lee, D. S., Park, J and Yu, B. (2021). Student clustering using tensor factorization: Accounting for multiple data dimensions. *Journal of Learning Analytics*, *8*(4), 303-325. https://doi.org/10.18608/jla.2021.84.5

[6] Miranda, R and Bagdasar, A. (2019). Student clustering in online courses based on learning styles and prior knowledge. *Journal of Distance Education*, *33*(2), 147-164. https://www.jde.edu/doi/abs/10.22541/jde.19.03.004

[7] Wang, Y., Li, J., Yang, B., et al. (2022). Stream-data-clustering based adaptive alarm threshold setting approaches for industrial processes with multiple operating conditions. *ISA Transactions*, *129*(Pt B), 594-608. https://doi.org/10.1016/j.isatra.2022.01.010

[8] Yang, H and Zhang, W. (2022). Data mining in college student education management information system. *International Journal of Embedded Systems*, *(3)*, 15. https://www.inderscience.com/journal/ijes

[9] Chu, Y and Yin, X. (2021). Data analysis of college students' mental health based on clustering analysis algorithm. *Complexity*, *2021*, 1-10. https://doi.org/10.1155/2021/6792017

[10] Tang, Q., Zhao, Y., Wei, Y., et al. (2021). Research on the mental health of college students based on fuzzy clustering algorithm. *Security and Communication Networks*, *2021*(3), 1-8. https://doi.org/10.1155/2021/2324632