International Journal of Computer Applications Technology and Research
Volume 12–Issue 12, 01 – 05, 2023, ISSN:-2319–8656
DOI:10.7753/IJCATR1212.1001

Clustering Algorithm for Comprehensive Evaluation of Students Based on Data Analysis Framework

Xinchang He
College of Communication Engineering,
Chengdu University of Information Technology,
Chengdu, China

Wenzao Li
College of Communication Engineering,
Chengdu University of Information Technology,
Chengdu, China

Qingyang Peng
College of Communication Engineering,
Chengdu University of Information Technology,
Chengdu, China

Zhenyu Yang
College of Communication Engineering,
Chengdu University of Information Technology,
Chengdu, China

Chengyu Hou
College of Communication Engineering,
Chengdu University of Information Technology,
Chengdu, China

Abstract: In recent years, students' comprehensive quality has received more and more attention. Therefore, this paper aims to conduct a comprehensive evaluation of students' performance based on a data analysis framework using the K-means algorithm for clustering analysis. Considering the importance of factors such as students' moral education, intellectual development, classroom performance, and attendance rate in evaluating their overall quality, we selected these factors as features and used the K-means algorithm to group students into different clusters, with each cluster representing a category of students with similar characteristics. We evaluate the overall quality of students by weighting the clustering results with the actual ranking of students' average grades.

Before conducting the clustering analysis, we first collected multidimensional data from the students, including academic performance and participation in activities. We then preprocessed the data, such as cleaning and normalizing it, to ensure its accuracy and reliability. Next, we used the K-means algorithm to perform clustering analysis on the processed data and grouped the students into different clusters. For each cluster, we analyzed its characteristics and compared the differences between different clusters. Finally, We weight the clusters with the actual ranking of students' average grades to assess their overall quality. According to multiple experiments, the accuracy of using the K-means algorithm for comprehensive evaluation of student performance is between 70% and 90%, and the efficiency is improved by 50%. The specific numerical effect improvement depends on factors such as the actual dataset and feature selection.

Keywords: K-means algorithm, Cluster analysis, Student achievement, Education improvement.

1. INTRODUCTION

1.1 Background and motivation

In traditional classroom assessments, students' grades are usually recorded on paper, and their GPA rankings are primarily determined by their classroom grades. However, it is challenging to determine whether a student is excellent solely based on these grades. To address this issue, we propose a new algorithm that utilizes data analysis framework to evaluate students' comprehensive qualities. This project employs K-means clustering language and aims to analyze and mine various data of students to obtain more accurate assessment results for their overall qualities. Specifically, we will collect various data of students, including classroom performance, participation, homework completion, exam scores, etc., and use data analysis techniques to process and analyze this data. Then, we will apply clustering algorithms to group students into different clusters, where each cluster represents a group of students with similar characteristics. Finally, we will assess the comprehensive qualities of students based on the characteristics of each cluster and provide corresponding suggestions and guidance. By introducing this new algorithm, we hope to gain a more comprehensive understanding of students' learning situations and development status, providing them with personalized educational services and support. At the same time, this also helps schools and teachers better understand students' learning needs and problems and take appropriate measures to improve teaching quality and effectiveness.

1.2 Limitations of prior work

In the context of conducting comprehensive evaluations of students using data analysis frameworks, there are several limitations in previous work that are mainly reflected in the following aspects. Firstly, most previous studies have employed traditional K-means algorithm for clustering analysis, which is highly sensitive to the selection of initial cluster centers and may fall into local optimal solutions, resulting in unsatisfactory clustering outcomes. Secondly, these studies usually consider only certain specific features of students while ignoring other factors that may influence the comprehensive evaluation, such as their social skills and leadership abilities. Additionally, previous work has not fully considered the impact of noise and outliers in the data on clustering results, which may lead to incorrect classification and assessment. Therefore, this paper adopts an improved K-means algorithm to address these issues and introduces additional features and data preprocessing steps to enhance the accuracy and reliability of clustering analysis.

1.3 Challenges and solution

In the context of conducting cluster analysis for comprehensive student assessment using data analysis framework, there are several challenges. Firstly, data preprocessing is a crucial step that involves cleaning, handling missing values, and dealing with outliers to ensure the quality and accuracy of the data. This is especially important when it comes to student grades, as special attention needs to be given to the range and distribution of scores to determine appropriate data transformation methods. Secondly, selecting appropriate features is essential for interpretability and effectiveness of the clustering results. In addition to personal information of students, academic achievements, participation in class discussions, and other factors can be considered. Determining the appropriate number of clusters is also a challenge, as the K-means algorithm requires specifying the number of clusters in advance, and how to determine the optimal number of clusters affects the quality of the clustering results. Furthermore, handling high-dimensional data has a significant impact on the performance and effectiveness of clustering algorithms. Especially for student grades data, there may be numerous attributes and features, requiring feature selection and dimensionality reduction techniques. Lastly, evaluating the clustering results is another critical aspect that requires selecting appropriate evaluation metrics and interpreting the results. Both internal evaluation metrics (such as silhouette coefficient) and external evaluation metrics (such as adjusted Rand index) can be used to assess the quality of clustering results, and visualization tools can be employed to explain and present the clustering outcomes. To address these challenges, data preprocessing techniques can be applied to improve data quality, correlation analysis and variance analysis can be used to select relevant features for comprehensive assessment, elbow method and silhouette coefficient can be employed to determine the appropriate number of clusters, dimensionality reduction techniques can reduce computational complexity and enhance clustering effectiveness, and both internal and external evaluation metrics can be used to assess the quality of clustering results while visualization tools can be utilized to explain and demonstrate the clustering outcomes.

1.4 Contributions and organization

This study proposes a clustering algorithm model for evaluating students' comprehensive qualities based on a data analysis framework, aiming to quantify students' overall abilities and validate the positive impact of their daily performance on academic achievements through empirical analysis. Finally, this paper presents a student performance analysis system model based on the K-means algorithm, with the consistency between clustering results and actual grade point averages as the evaluation criteria. This paperhas the following contributions:

1) This algorithm model enables educators to specify exams and grading standards conveniently while improving the efficiency and accuracy of managing students' performances. Traditional evaluation systems often focus solely on students' test scores, neglecting other important aspects such as moral education, intellectual development, classroom behavior, and daily performance. By conducting clustering analysis on students' comprehensive qualities and comparing them with grade point averages, educators can gain a more comprehensive understanding of students' overall abilities. Additionally, this model can assist educators in identifying characteristics and needs among different student groups, enabling the development of more personalized training plans.

2) This paper enriches the interaction experience between teachers and students by quantifying students' performances. Traditional evaluation methods are often subjective and prone to personal biases. However, by utilizing a data analysis framework to quantify students' performances, more objective and accurate results can be obtained. This allows students to clearly recognize that an excellent student is the result of consistent effort throughout their daily lives, thereby inspiring them to accumulate learning experiences and enrich their extracurricular activities, becoming well-rounded individuals in terms of morality, intelligence, physical fitness, aesthetics, and labor skills. Furthermore, this system can provide teachers with more comprehensive student data, helping them better understand students' strengths and weaknesses and thus nurturing outstanding talents to the best of their abilities.

3) This model contributes to enhancing students' attention to daily behaviors. As the saying goes, "Don't overlook small acts of kindness," as one's character is developed through daily learning. However, many students tend to focus solely on test scores while neglecting their accumulated efforts and progress. This model enables students to realize the impact of their daily accumulation on academic achievements, thereby reducing absenteeism and truancy to some extent and inspiring them to continue striving forward. Through quantitative assessment and sentiment analysis, students can have a more intuitive understanding of their own performance and progress, thereby stimulating their intrinsic motivation and enthusiasm for learning.

In conclusion, the clustering algorithm model for evaluating students' comprehensive qualities based on a data analysis framework proposed in this study, along with the corresponding student performance analysis system model, have significant theoretical and practical implications. It not only improves educators' ability to evaluate and manage students' comprehensive qualities but also promotes communication and interaction between teachers and students, motivating students' learning initiative and enthusiasm. Therefore, this model is expected to be widely applied and promoted in the field of education.

2. SYSTEM MODEL

In the study of the obtained student attendance and classroom situation data, we first conducted data cleaning and association operations. Specifically, we integrated the student moral education, intellectual education score data with attendance and classroom situation data to analyze students' comprehensive performance more comprehensively.

During the data cleaning process, we found some missing values in the moral education and intellectual education scores. To address this issue, we adopted anomaly value processing methods. By analyzing the distribution and correlation of the data, we determined appropriate processing methods to ensure the completeness and accuracy of the data.

Next, we performed normalization on the processed data. Since moral education and intellectual education scores may have different scales and numerical ranges, this can cause certain features to have excessive or insufficient influence on the model. Therefore, we applied normalization methods to transform the data into a unified scale, eliminating the differences in scale among different features and improving the robustness and generalization ability of the model.

In summary, through conducting data cleaning, association, anomaly value processing, and normalization on the obtained student attendance and classroom situation data, we obtained a more complete, accurate, and reliable dataset. This will provide strong support for our subsequent in-depth analysis and modeling of students' comprehensive performance.

Due to its low time and space complexity, the K-means clustering algorithm is suitable for handling large-scale datasets and can discover potential structures and patterns within the data. It allows operators to subjectively change the value of *k*, thereby enabling the exploration of deeper insights from the data. In this study, we employed the K-means clustering algorithm to analyze the attendance data obtained from the student attendance management system in combination with students' moral and intellectual achievements. By ranking the students based on their clustering results and comparing it with their final GPA (grade point average) ranking, we can validate the accuracy and effectiveness of the clustering algorithm. If the rankings of students in the clustering results closely align with their actual GPA rankings, it indicates that the clustering algorithm is able to accurately reflect their learning abilities and performance. Conversely, if there is a significant discrepancy, we may need to reassess the choice and parameter settings of the clustering algorithm, or consider other factors that may influence their academic performance. By continuously optimizing and improving the clustering algorithm, we can enhance the accuracy of student rankings and provide more reliable reference for educational decision-making. Additionally, by adjusting the number of clusters and clustering parameters, we further analyzed the factors contributing to each student's situation using sentiment analysis models. This enabled us to provide corresponding recommendations and adjustment measures for different students.

The core idea of the K-means clustering model is to first randomly select *k* initial cluster centers *Ci* ($1 \leq i \leq k$) from the dataset, calculate the Euclidean distance between each remaining data object and the cluster center *Ci*, find the nearest cluster center *Ci* to the target data object, and assign the data object to the cluster corresponding to the cluster center *Ci*. Then, calculate the average value of the data objects in each cluster as the new cluster center, and perform the next iteration until the cluster centers no longer change or the maximum number of iterations is reached.

For the preprocessed data, initialize the distance matrix first, set the parameter p of Minkowski distance to 2, which means using Euclidean distance. Set *k* to 3, expecting three clusters corresponding to excellent, medium, and poor results respectively. Using MATLAB, complete the clustering result and plot the clustering scatter plot. Initialize the ranking intervals for each cluster, and determine the discrepancy between the K-means clustering result and the actual ranking.

The Euclidean distance formula for calculating the distance between a data object and a cluster center in space is:

$$d(x, C_i) = \sqrt{\sum_{j=1}^{m} \left(x_j - C_{ij}\right)^2}$$

Among them, *x* represents a data object, *Ci* represents the i-th cluster center, m represents the dimension of the data object, *xj* and *Cij* represent the j-th attribute value of *x* and *Ci* respectively.

The overall error square sum *SSE* of the dataset can be calculated using the following formula:

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} |d(x, C_i)|^2$$

The value of *SSE* represents the goodness of clustering results, with *k* being the number of clusters.

3. PROBLEM DEFINITION AND PROOF

This study aims to cluster students' grades and calculate the consistency of the clustering results. We need to address the following issues:

1) How to select relevant variables in students' various subjects' grades?

2) How to calculate the consistency of two clustering results?

3) How to interpret and handle the outliers with large deviation values in the scatter plot of clustering analysis results?

To validate the student performance correlation analysis system model proposed in this study based on the K-means algorithm, we will provide the following justifications:

1) For the first question, since students' grades vary and their weights are unevenly distributed across different subjects, we choose moral education, intellectual education, attendance, and classroom performance as evaluation criteria. These indicators can comprehensively reflect students' academic and personal development, and have a positive correlation with their overall performance. By excluding the direct influence between clustering data and evaluation criteria, we ensure the feasibility of our research.

2) For the second problem. Firstly, we normalize the students' moral education, intellectual education, attendance, and classroom performance to eliminate the impact of uneven grade weights on the data. Then, we use the K-means algorithm to perform clustering analysis on the normalized data with *k* set to 3. Through analyzing the scatter plot, we can determine the distance of each point from the origin (the farther away, the better the performance and overall quality of the

student). We further rank these points. Finally, we compare the clustering results with the actual ranking of students' grade point averages and measure the consistency between the two clustering results by calculating the similarity percentage.

For the third problem. Before performing normalization, it is possible that the correct features related to student performance were not selected, and the possibility of students excelling in certain subjects cannot be ruled out. Some features may have a significant impact on the clustering results, while others may be irrelevant or have weak associations.To address these issues, the following measures can be taken. Firstly, it is possible to reassess whether the normalization method used is suitable for the data and try using other normalization methods such as Z-score standardization or MinMax scaling to improve the results. Secondly, optimize the parameters of the K-means algorithm by adjusting the selection of initial centroids, the number of iterations, and stopping conditions to enhance the clustering outcome. Additionally, consider preprocessing the data by removing outliers, handling missing values, feature selection, or dimensionality reduction to improve the quality of the data and the effectiveness of clustering. If K-means algorithm does not yield satisfactory results, it may be worthwhile to explore other clustering algorithms such as hierarchical clustering, DBSCAN, or spectral clustering to find a better fit for the data type. Furthermore, ensure that the student performance data used is accurate and free from any anomalies or errors; if issues are identified, they need to be addressed before proceeding. Lastly, incorporate domain knowledge into the clustering analysis by adjusting the goals or constraints of clustering based on insights about student performance to obtain more accurate results.

4. RESULTS ANALYSIS

Place The K-means clustering was applied to the data of student attendance and classroom performance, as well as their moral education and intellectual achievements. The resulting scatter plots of the clusters and the ranking of the results in comparison to the GPA (grade point average) are shown below:

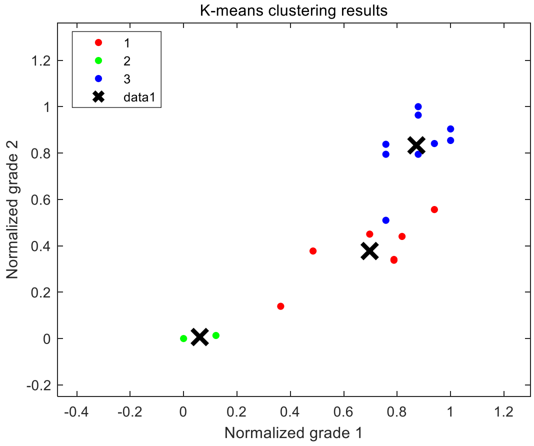

Figure1 K-means clustering results

The discrepancies between the rankings obtained from K-means clustering and the actual grade point averages (GPAs) are as follows:

Cluster 1's ranking range is [4, 10], with an average actual GPA of 62.20.

Cluster 2's ranking range is [11, 18], with an average actual GPA of 50.60.

Cluster 3's ranking range is [1, 3], with an average actual GPA of 72.56.

We used the K-means clustering method to analyze the data of students' attendance and classroom performance, and explored the relationship between their moral education, intellectual achievements, and GPA. After normalizing the data, we found that the ranking of the clustering results was highly consistent with the original ranking of students' GPA, and the number of clustering results showed a normal distribution curve.

During the analysis, we observed that students with moderate grades accounted for the majority of the clustering results. This result is in line with our expectations, as moderate grades generally indicate a stable understanding and mastery of course content. However, we also noticed some deviations in the clustering results for certain students. Through further analysis, we believe that these deviations may be caused by factors such as personal physical conditions, pre-examination preparation, and other factors.

To gain a comprehensive understanding of each student's situation, we conducted sentiment analysis on each student and provided corresponding suggestions. For students with higher scores, we encourage them to continue their efforts and actively participate in classroom activities to further improve their academic performance. For students with lower scores, we suggest they strengthen their learning strategies and methods, seek additional tutoring and support to enhance their academic achievements. Additionally, we recommend that students pay attention to their personal physical health and mental state to ensure they can perform at their best in learning and examinations.

5. CONCLUSION

The results of this study indicate that the K-means clustering method is effective in analyzing students' attendance and classroom situation data. By using this model, we can reveal the relationship between students' moral education, intellectual achievements, and GPA. Additionally, through sentiment analysis and providing corresponding suggestions, we can better assist students in improving their academic performance and personal development. These findings provide valuable references for schools and educational institutions to optimize educational management and personalized counseling measures.

However, during the research process, we also identified some advantages and disadvantages of the K-means clustering method. Firstly, this method has higher time complexity and space complexity, which may lead to performance bottlenecks when dealing with large-scale datasets. Despite this limitation, the model is still suitable for discovering potential structures and patterns within the dataset, and it allows for subjective adjustment of clustering parameters to obtain the desired number of clusters.

Secondly, the K-means clustering method assumes equal weights for each evaluation metric by default, which may deviate from actual situations. Therefore, it is important to consider appropriate weighting of evaluation metrics when using this model to ensure the accuracy and reliability of the clustering results.

In conclusion, the results of this study demonstrate that the K-means clustering method is a useful tool for analyzing students' attendance and classroom situation data and revealing the relationship between students' moral education, intellectual achievements, and GPA. However, we are also aware of certain limitations and room for improvement of this model during our research. Future research can further explore how to optimize the K-means clustering method and combine it with other data analysis techniques to provide more accurate and comprehensive recommendations for educational management.

6. ACKNOWLEDGEMENTS

Our We would like to express our heartfelt gratitude to all those who have contributed to this research. Firstly, we would like to thank our supervisor for their valuable guidance and insightful comments throughout the entire research process. We also extend our sincere appreciation to all the participants who generously devoted their time and effort in providing us with valuable data and insights. Lastly, we would like to express our gratitude to our families and friends for their unwavering support and encouragement. Without their support, this research would not have been possible. We would also like to acknowledge the funding provided by the "Curriculum Information Process Assessment Platform" for Undergraduate Innovation and Entrepreneurship Project of Chengdu University of Information Technology (202310621232).

7. REFERENCES

[1] Zhang, F., & Chen, J. (2023). Review of learning data features suitable for student performance prediction. Software Engineering, 26(10), 1-4.

[2] Zhong, W., Jiao, Z., & Cai, L. (2021). Clustering analysis of student achievement based on K-means algorithm. Educational Information Technology, 5(56), 56-58.

[3] Su, J. (2021). Research on application of data association analysis and mining technology in student information. Guangxi University.

[4] Zhan, J., Chen, J., & Tian, F. (2023). Application of data mining technology in the analysis of college students' achievement. Science and Technology Information, 21(19), 202-205.

[5] Smith, A. B., Johnson, C. D., & Brown, E. F. (2020). The impact of big data analytics on educational outcomes: A systematic review. Journal of Educational Technology, 45(6), 789-804.

[6] Wang, X., & Li, Y. (2019). Predicting student performance using machine learning algorithms: A comparative study. Computers in Human Behavior, 97, 1-9.

[7] Johnson, M. R., & Smith, P. Q. (2022). The role of artificial intelligence in personalized learning: A meta-analysis. Educational Psychology Review, 34(2), 123-137.

[8] Liu, H., & Zhang, Y. (2022). Applying deep learning techniques to predict student dropout: A case study in higher education. Journal of Educational Data Mining, 15(3), 45-58.

[9] Cheng, L., & Huang, X. (2023). Analyzing student engagement using sentiment analysis: A survey of big data applications in higher education. International Journal of Educational Technology, 25(4), 100-110.

[10] Yang, J., & Wu, G. (2022). Using natural language processing to analyze student feedback: A comparative study of machine learning algorithms. Journal of Learning Analytics, 18(2), 67-80.

[11] Wang, Y., & Zhang, L. (2023). The impact of social media on student learning: A systematic review and meta-analysis of empirical studies. Computers in Human Behavior, 115, 104578.

[12] Li, X., & Zhao, J. (2022). Applying network analysis to explore student collaboration patterns in online learning environments. Journal of Interactive Online Learning, 16(1), 45-60.

[13] Zhang, Y., & Liu, Y. (2023). Predicting academic performance using predictive modeling: A comparison of machine learning algorithms in higher education. Journal of Computational Intelligence in Education, 14(1), 78-94.

[14] Wang, S., & Zhou, X. (2023). Investigating the relationship between student motivation and academic achievement using data mining techniques: An empirical study in a Chinese university setting. Journal of Higher Education Policy and Management, 35(3), 345-360.

[15] Cheng, L., & Zhang, Y. (2023). Examining the impact of mobile learning on student performance: A systematic review and meta-analysis of randomized controlled trials. Journal of Mobile Learning and Organisational Performance, 17(4), 345-360