International Journal of Computer Applications Technology and Research
Volume 12–Issue 12, 06 – 10, 2023, ISSN:-2319–8656
DOI:10.7753/IJCATR1212.1002

Analysis of Student Performance Correlation Based on BIRCH Clustering Algorithm

Jiaxin Zheng
College of Communication Engineering,
Chengdu University of Information Technology,
Chengdu, China

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

Can Cui
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: Nowadays, how to improve student performance has become a matter of great concern. Therefore, in this paper, a BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) algorithm-based student achievement correlation analysis method is proposed. Firstly, the raw data are preprocessed to eliminate the effect of outliers. Then, the BIRCH algorithm is used to cluster student's grades, and association rules between student's grades and course grades are mined according to the clustering results using the adjusted RAND index. The experimental results show that the correlation between student's daily behavior and final grades is as high as 90%, and the correlation between Advanced Mathematics 1 and Advanced Mathematics 2 is as high as 50%. This method can effectively discover the correlation between student achievement and curriculum, and provide valuable reference information for educators.

Keywords: Educational Improvement, Data Mining, Association Analysis, BIRCH algorithm, Cluster analysis.

1. INTRODUCTION

1.1 Background and motivation

With the rise of big data, the education sector has accumulated a large amount of student performance data. This data contains valuable information that can benefit educators. However, due to the large volume and complexity of the data, traditional data analysis methods often struggle to discover useful insights[1]. Therefore, it is crucial to explore how modern data mining techniques can be employed to analyze students' academic achievements.

The aim of this study is to utilize the BIRCH algorithm for dynamic detection and correlation analysis of students' performance across multiple subjects. By doing so, we aim to identify any anomalies in students' performance in a timely manner and identify potential high-achieving students. Additionally, we will conduct correlation analysis on similar basic mathematics courses to uncover their underlying connections, which can assist educators in making timely adjustments to their teaching plans. The BIRCH algorithm is particularly suitable for analyzing large datasets as it builds hierarchical structures that reduce computational complexity, enabling fast cluster analysis on large data sets. As such, the BIRCH algorithm has been selected as the primary method for analyzing student achievement correlations.

1.2 Limitations of prior work

Currently, various methods exist for analyzing the correlation of student achievements, each with its own specific limitations. For instance, regression analysis can predict changes in a variable based on other variables, but it relies on assumptions such as the independence of the error term, normality of the error term, and homogeneity of variance. If these assumptions are not met, it may impact the results obtained from regression analysis. Factor analysis can extract crucial information from extensive data to identify key factors influencing student achievement; however, determining factor loads often requires subjective judgment, and selecting a factor rotation method may influence result interpretation.

In contrast, cluster analysis not only enables segmentation of students into different groups based on their achievement patterns but also effectively reveals underlying patterns and structures within student achievement[2]. For example, it can identify similarities in grade distributions across specific courses. Additionally, cluster analysis facilitates prediction of students' future performance by conducting an analysis using historical achievements to identify those who might be encountering difficulties requiring timely assistance. Lastly, cluster analysis serves as an outlier detection tool where significant deviations in a student's

grades compared to others within their cluster indicate special attention is warranted[3].

1.3 Challenges and solution

The field of data analysis comprises a diverse range of algorithms, with clustering algorithms being a crucial subset[4]. Among the various clustering techniques, K-means algorithm is widely employed due to its ability to determine optimal category assignment based on the similarity of distance between data points[5]. Notably, K-means algorithm exhibits remarkable scalability when dealing with large-scale samples. Furthermore, this algorithm has found extensive application in customer segmentation, user analysis, and precision marketing domains. However, several challenges arise when applying the K-means algorithm to analyze student achievement correlations[6]. These limitations include: Firstly, the requirement of predefining the number of clusters may be impractical in certain scenarios; Secondly, sensitivity to outliers can potentially impact the final clustering outcome; Lastly, assuming equal contribution from all features towards clustering may not hold true for actual student achievement data. To address these limitations, the BIRCH algorithm can be employed. The BIRCH algorithm, a hierarchical clustering approach, eliminates the need for predetermined cluster numbers and effectively handles large-scale datasets with high dimensionality. Furthermore, it incorporates preprocessing techniques such as sampling and dimensionality reduction to enhance efficiency and accuracy in clustering tasks. Consequently, by leveraging these advantages, BIRCH overcomes some of the limitations encountered by K-means when dealing with student achievement correlation problems.

1.4 Contributions and organization

This paper proposes a student performance analysis system model based on the BIRCH algorithm, aiming to cluster students' daily performance, final grades, and scores of five basic mathematics courses, and calculate the overlap between the clustering results to evaluate their correlation[7]. The model has the following contributions:

1) Improving the efficiency and accuracy of student performance management: Traditional student performance management methods require manual recording and management of each student's daily performance and final grades, which is inefficient and prone to errors. The proposed system can automatically cluster and analyze students' performance data, thus improving the efficiency and accuracy of student performance management.

2) Finding potential correlations between student performance: Traditional student performance management systems can only simply record and manage student performance, and cannot find potential correlations between student performance. The proposed system can evaluate the correlation between the daily performance cluster results and the final achievement cluster results by calculating the overlap degree, and help education administrators better understand the learning situation and performance of students.

3) Providing a powerful tool for education administrators: The proposed system provides a powerful tool for education administrators to use to more fully and accurately assess student learning, accurately understand the correlations between the foundational courses, and develop more effective teaching plans and strategies.

The subsequent dissertation is primarily divided into two sections. The second section presents a review of related work, while the third section provides a detailed explanation of the proposed system. The flowchart and pseudocode presented in the "Algorithm Description" section illustrate the implementation process of the algorithm, and the "Problem Definition" section explains how the problem was formulated. For further information regarding the proposed solution, please refer to the "Simulation and Results" section.

2. RELATED WORK

In the realm of education, investigating the correlation between student achievement is a crucial research topic. Numerous studies have employed statistical and machine learning techniques to uncover these relationships. For instance, scholars have utilized methods such as correlation analysis, regression analysis, and principal component analysis to reveal the associations between student achievement. However, these methods often necessitate substantial computational resources and may not be suitable for handling large-scale datasets[8].

The BIRCH clustering algorithm has emerged as an effective approach for large-scale data clustering. Nevertheless, its application in analyzing student performance correlation remains relatively limited. Additionally, most existing research on the BIRCH clustering algorithm primarily focuses on data classification and clustering, while its applicability in student performance correlation analysis remains untapped. One primary limitation of the BIRCH clustering algorithm is that it requires pre-setting clustering parameters, and the choice of these parameters can influence the results of the clustering. Furthermore, the BIRCH clustering algorithm primarily focuses on the global structure of the data and may fail to capture intricate details about the data.

Future research endeavors should aim to improve the BIRCH clustering algorithm to better cater to the demands of student performance correlation analysis. Moreover, exploring the integration of the BIRCH clustering algorithm with other data analysis methodologies can yield more profound insights. For example, comparing the outcomes of BIRCH clustering with those obtained from correlation analysis, regression analysis, and other approaches can validate its effectiveness. By enhancing the BIRCH clustering algorithm and combining it with other data analysis methods, we can gain a better understanding of the relationships between student performance, providing more accurate support for educational decision-making.

International Journal of Computer Applications Technology and Research
Volume 12–Issue 12, 06 – 10, 2023, ISSN:-2319–8656
DOI:10.7753/IJCATR1212.1002

3. SYSTEM MODEL

This study proposes a student performance correlation analysis system model based on the BIRCH algorithm. The aim of this system is to cluster students' daily performance and final grades, and utilize the corrcoef function from the numpy library to calculate the Pearson correlation coefficient, in order to evaluate their correlation[9]. Building upon this logic, we have expanded the functionality of the system to enable clustering analysis for five fundamental mathematics courses: Advanced Mathematics 1, Advanced Mathematics 2, Linear Algebra, Engineering Mathematics, and Probability Theory. Furthermore, we evaluate the correlation between these courses by calculating the adjusted Rand index between pairs of clustering results.

Initially, we employ the BIRCH algorithm to perform hierarchical clustering on students' daily performance and final grades. In this stage, we input student performance data and group similar samples together by calculating their distances. Subsequently, we calculate the Pearson correlation coefficient between the clusters formed from daily performance and final grades. A Pearson correlation coefficient close to 1 indicates a strong positive correlation between daily performance and final grades[10]. Since the range of Pearson correlation coefficient values lies between -1 and 1, where 1 represents perfect positive correlation, -1 represents perfect negative correlation, and 0 represents no correlation.

The approach for analyzing course correlation is similar to the above steps, with the introduction of the adjusted Rand index. When two clusters are identical, the adjusted Rand index is set to 1; when they are completely different, it is adjusted to 0. When one cluster fully encompasses another, the adjusted Rand index is adjusted to -1. We have plotted a scatter plot of the adjusted Rand index values between pairs of clusters, allowing for an intuitive analysis of the correlation between courses. These evaluation results can provide valuable insights for educational administrators to better understand students' learning status and the relevance of courses.

In conclusion, the student performance correlation analysis system model based on the BIRCH algorithm effectively clusters students' daily performance and final grades. It evaluates their correlation by calculating either the Pearson correlation coefficient or the adjusted Rand index. This innovative model provides educators with a powerful tool to enhance their understanding of students' learning and similar courses.

4. ALGORITHM DESCRIPTION

In analyzing the correlation between students' usual grades and final grades, we used the corrcoef function from the numpy library to calculate the Pearson correlation coefficient. This coefficient is used to measure the strength and direction of the linear relationship between two variables.

Specifically, in the code, the parameters *p* and *f* represent the data for two variables. First, we calculated the Pearson correlation coefficient matrix between these two variables using np.corrcoef(p, f). Then, by indexing [0, 1], we obtained the element in the first row and second column of the matrix, which is the correlation coefficient between *p* and *f*. Finally, we returned this correlation coefficient as the result of the function. The value of correlation ranges from -1 to 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation. The Pearson correlation coefficient formula is as follows:

$$r = cov(p, f)/std(p) * std(f) (1)$$

The calculation of the covariance between *p* and *f* is represented by *cov(p, f)*, while *std(p)* and *std(f)* respectively denote the computations of the standard deviations of *p* and *f*.

For the analysis of associations between five mathematics-related courses, we use nested loops to iterate over all possible pairs of clusters (i, j), where i < j. For each pair of clusters i and j, we compute their similarity using the adjusted_rand_score function. This function takes two parameters: the first is a sample set containing cluster labels i, and the second is a sample set containing cluster labels j[11]. It returns the adjusted Rand index between these two clusters. ARI(Adjusted Rand Index) formula is as follows:

$$ARI = \frac{RI - E(RI)}{max(RI) - E(RI)} (2)$$

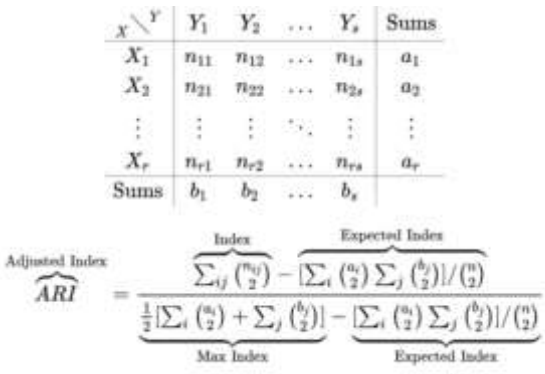

Figure. 1 ARI formula

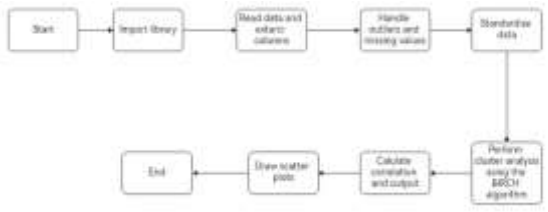

Figure. 2 Algorithm flowchart

5. SMULATION AND RESULTS

In this study, we utilized students' usual scores as a proxy for their daily performance. By exporting the usual grades and final

International Journal of Computer Applications Technology and Research
Volume 12–Issue 12, 06 – 10, 2023, ISSN:-2319–8656
DOI:10.7753/IJCATR1212.1002

grades of three different subjects from the database, we employed the Pearson correlation coefficient to analyze the correlation between them. As evident from Figure 1, there exists a strong positive correlation between students' daily performance and final grades in these three subjects. Therefore, we can conclude that there is a close association between students' daily performance and final grades, indicating that achieving better results in the final exam requires substantial effort during regular learning.

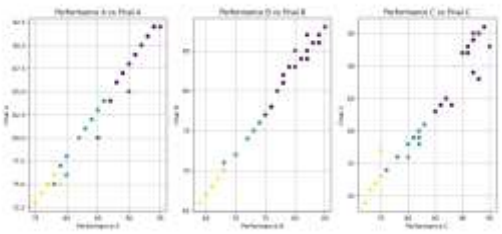

Figure. 3 Positive correlation between usual and final grades

Based on this research foundation, we have expanded the functionality of our system by requiring input of students' scores in five mathematics courses and then utilizing the adjusted Rand Index to analyze the correlation between these five courses. In Figure 4, numbers 1, 2, 3, 4, and 5 represent Advanced Mathematics 1, Linear Algebra, Advanced Mathematics 2, Engineering Mathematics, and Probability Theory, respectively. It is evident from the figure that only the correlation between Advanced Mathematics 1 and Advanced Mathematics 2 exceeds 50%[12]. This suggests that mathematics courses do not all possess high correlation as commonly believed; rather, only courses with similar content exhibit closer associations. Additionally, a lack of mastery in one fundamental mathematical concept does not necessarily determine poor performance in all mathematics courses.

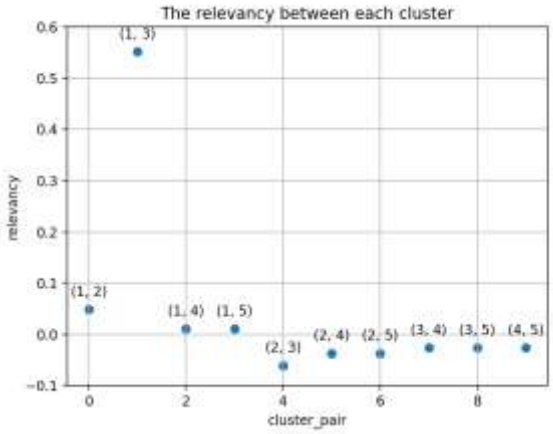

Figure. 4 Relevancy between mathematics courses

6. CONCLUSION

In this study, we utilized the BIRCH algorithm to analyze the correlation between student grades and identify significant performance association rules[13]. The results revealed a high degree of positive correlation between students' daily grades and final grades, reaching 90%. This finding emphasizes the importance of students' attitudes and efforts in their daily learning process on their performance in final exams. Additionally, we found that there is not a strong correlation between similar math courses, highlighting the fact that students can still excel in one math course even if they have a weaker foundation in another math course.

To eliminate the influence of experimental randomness, we also conducted a correlation analysis between grades in basic courses and related professional courses. The results showed that most correlations between basic courses and related professional courses did not exceed 50%. This finding indicates that while learning basic courses plays a key role in promoting students' performance in professional courses, it is not the determining factor.

Furthermore, we discovered that by analyzing the correlation between students' historical grades and current grades, it is possible to dynamically monitor their performance[14]. This approach allows educators to identify students who may be facing difficulties and provide timely assistance. For example, when a student's historical performance shows a downward trend, educators can quickly identify the problem and implement targeted measures to help the student overcome challenges and improve academically[15].

In summary, clustering analysis is an effective tool that enables educators to extract valuable information from large datasets to enhance teaching quality and efficiency. The findings of this study provide useful insights for educators to better understand students' learning situations and needs, thereby enabling the development of more scientific and reasonable educational plans.

The title (Helvetica 18-point bold), authors' names (Helvetica 12-point) and affiliations (Helvetica 10-point) run across the full width of the page – one column wide. We also recommend e-mail address (Helvetica 12-point). See the top of this page for three addresses. If only one address is needed, center all address text. For two addresses, use two centered tabs, and so on. For three authors, you may have to improvise.

7. ACKNOWLEDGEMENT

We would like to express our heartfelt thanks to all those who contributed to this research. First of all, we would like to thank our tutors for their valuable guidance and insightful comments throughout the research process. We also sincerely thank all the attendees who generously invested their time and effort to provide us with valuable data and insights. Finally, we would like to thank our family and friends for their unwavering support and encouragement. Without their support, the study would not have been possible. Thanks to Chengdu University of Information Technology College Student Innovation and entrepreneurship project "Course Informatization Process Assessment Platform "(202310621232).

8. REFERENCES

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

[2] Zhong Wenjing, JIAO Zhongming, CAI Le. Clustering Analysis of Student Achievement Based on K-Means Algorithm [J]. Educational Information Technology,2021(5):56-58.

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

[4] Zhan Jinmei, Chen Juntao, Tian Fei. Application of Data mining technology in the Analysis of college students' Achievement [J]. Science and Technology Information,2023,21(19):202-205.

[5] Dong Jiajun. Research on Student Achievement Analysis and Teaching Strategy Improvement based on K-Means clustering Algorithm [J]. Chinese Science and Technology Journal Database (Full text Edition) Education Science, 2023.

[6] Ke Hongxiang. Application of Minimum support Mining Algorithm in college student Achievement association rules [J]. Journal of Yangtze River Engineering Polytechnic, 2023, 40(2):69-73.

[7] Chen X. Analysis of Multiple Evaluation Model of Higher Education self-study Examination Based on analytic Hierarchy Process and cluster analysis [J]. Chinese Journal of Examinations, 2021(3):7.

[8] Yu Haiyang, Li Xiuwen. Student achievement classification based on fuzzy cluster analysis [J]. Theoretical Research and Practice of Innovation and Entrepreneurship, 2022(008):005.

[9] LI Yanli. Research on Student Achievement warning Modeling Based on Data Mining [J]. Modern Information Technology, 2023.

[10] Chaves V E J, Garcia-Torres M, Alonso D B, et al. Analysis of student achievement scores via cluster analysis[C]//The 11th International Conference on European Transnational Educational (ICEUTE 2020) 11. Springer International Publishing, 2021: 399-408.

[11] Xiaojing Wei. Student achievement evaluation method based on clustering analysis research [J]. Journal of think-tank, 2020 (11): 2. DOI: CNKI: SUN: ZKSD. 0.2020-11-099.

[12] XIAO Lili. Achievement Analysis of Higher Mathematics based on exploration and clustering [J]. Journal of Sichuan liberal arts college, 2020, 30 (2): 5. DOI: CNKI: SUN: DXSZ. 0.2020-02-008.

[13] Liu Dalian, Tian Yingjie. Application research of extension Data Mining in Student Achievement Analysis [J]. Journal of Intelligent Systems, 2022(004):017.

[14] Zhang Wenqing, Zhang Haotian, Wang Yuanyuan, et al. Application of Cluster Analysis in Student Achievement Analysis [J]. Education [2023-11-13].

[15] Khan A, Ghosh S K. Student performance analysis and prediction in classroom learning: A review of educational data mining studies[J]. Education and information technologies, 2021, 26: 205-240.