Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 1 (2025)

Optimising Educational Performance through Clustering Algorithm Evaluation

Amartya Ghosh

Assistant Professor, Computer Science and Engineering, Brainware University

com.amartya@gmail.com

Payel Sengupta

Assistant Professor, Computer Science and Engineering, Brainware University

Payel9433@gmail.com

Avijit Kumar Chaudhuri

Professor, Computer Science and Engineering, Brainware University

c.avijit@gmail.com

Ranjan Banerjee

Assistant Professor, Computer Science and Engineering, Brainware University

rnb.cse@brainwareuniversity.ac.in

Mithun Biswas

Assistant Professor, Computer Science and Engineering, Brainware University

mithunbiswas0707@gmail.com

Pranab Gharai

Assistant Professor, Computer Science and Engineering, Brainware University

pranab.g10@gmail.com

Atanu Kumar Das

Assistant Professor, Computer Science and Engineering, Brainware University

rnb.cse@brainwareuniversity.ac.in

Article History:

Received: 01/11/2024

Revised: 06/12/2024

Accepted: 10/01/2025

Abstract:

Analyzing student performance is a vital undertaking within the realm of educational data mining (EDM). This process empowers academic institutions to uncover significant trends, pinpoint students who may be struggling, and formulate effective support strategies. This scholarly article delves into the application of clustering methodologies to classify students based on various performance metrics, such as academic grades, attendance records, engagement levels, and involvement in extracurricular activities. By segmenting students into distinct groups, educators can

gain a clearer understanding of their behavioural characteristics, optimize resource distribution, and deploy customized intervention programs.

This research paper presents an evaluative comparison of diverse clustering approaches utilized for assessing student academic outcomes. The investigation evaluates the efficacy of several prominent clustering algorithms, including K-Means, DBSCAN, BIRCH, and Expectation Maximization (EM), in classifying students according to their educational achievements. The findings illuminate the advantages and limitations of each method, offering valuable perspectives on their practical utility in the field of educational data mining. The substantial increase in educational data has necessitated the adoption of sophisticated data mining techniques to extract meaningful patterns and actionable intelligence. Clustering, an unsupervised machine learning paradigm, is extensively employed to categorize students based on their performance, thereby assisting educators in identifying vulnerable students and customizing appropriate interventions.

Keywords: Student grouping, Academic achievement, K-Means algorithm, DBSCAN algorithm, Tree-based clustering, Probability-based clustering, Learning analytics

Introduction

The swift progress of technology and the widespread adoption of digital learning platforms have led to an unprecedented accumulation of educational data. This data, which covers various aspects of student academic results, presence, involvement, and engagement, offers immense potential to improve the learning experience. However, the sheer volume and complexity of this information make it necessary to use advanced data mining techniques to extract valuable insights. Among these methods, clustering has become a powerful tool for examining student performance and pinpointing trends that can guide educational approaches.

Clustering Algorithms Used

This study utilizes four prominent clustering algorithms:

K-means: This is a partition-based algorithm that forms clusters by minimizing the distance between students and the center of their respective groups.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise): This technique identifies dense areas within student data and flags unusual data points, making it effective for educational datasets that might contain noise.

Hierarchical Clustering: This method constructs a tree-like structure, known as a dendrogram, which shows how students are grouped into nested clusters. This provides a clear visual representation of performance groupings.

Gaussian Mixture Model (GMM): This approach uses probability, making it well-suited for situations where student clusters might overlap, or where students show a mix of academic behaviors.

Data and Evaluation

The dataset for this study includes demographic details, academic scores, attendance logs, and participation metrics. Before clustering, the data goes through several pre-processing stages, including cleaning, normalization, and feature selection, to ensure its accuracy and consistency.

Each clustering technique is assessed using Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index to measure the quality, internal consistency, and distinctiveness of the clusters formed.

Background

Clustering is an unsupervised machine learning method that organizes data points into groups based on how similar they are. Unlike classification, which needs pre-labelled data, clustering doesn't depend on existing categories. This makes it especially valuable in educational environments where the aim is to find natural groupings of students using their performance indicators. By identifying these groups, educators can customize support to address the unique needs of different student populations, ultimately enhancing educational results.

In recent years, applying clustering techniques in educational data mining has garnered considerable interest. Researchers have investigated various algorithms, such as K-Means, DBSCAN, BIRCH, and Expectation Maximization (EM), to analyze student performance data. Each of these algorithms has its own advantages and disadvantages, highlighting the importance of a comparative study to assess their effectiveness in diverse situations.

The Significance of Student Performance Analysis

Evaluating student performance holds paramount importance for several reasons. Primarily, it empowers educators to spot students at risk who might be in need of extra help. Through grouping students based on their performance metrics, educators can effectively identify struggling individuals and deploy customized support strategies to foster their improvement. Secondly, clustering techniques can uncover trends and hidden patterns that traditional analytical approaches might miss. For example, they could expose links between attendance records and academic success, or illustrate the influence of extracurricular engagement on student achievement.

Beyond this, student grouping can contribute to curriculum enhancement and efficient resource management. By comprehending the diverse requirements of various student cohorts, educational organizations can develop learning materials and programs that address a broad spectrum of learning styles and aptitudes. Furthermore, insights from clustering can inform decisions on resource deployment, such as academic advising or tutoring services, ensuring they are channeled towards students who stand to gain the most.

Understanding Student Outcomes Matters

Examining student performance is vital for several key reasons. Firstly, it allows educators to pinpoint learners who might be facing difficulties and need extra assistance. By grouping students based on their academic results, instructors can identify those who are struggling and put in place specific interventions to aid their progress. Secondly, such analysis, particularly through clustering, can reveal insights and relationships that aren't obvious with standard assessment techniques. For instance, it might show connections between class attendance and academic achievement, or highlight how involvement in outside-of-class activities affects student success.

Moreover, clustering methods can significantly aid in refining educational programs and distributing resources. By grasping the varied requirements of distinct student populations, educational bodies can craft curricula that accommodate diverse learning preferences and capabilities. Additionally, this analysis can guide the allocation of resources, like supplementary tutoring or academic guidance, ensuring they are directed to students who will gain the most benefit.

The Critical Role of Student Performance Evaluation

The evaluation of student performance is essential for multiple compelling reasons. It empowers educators to identify and provide targeted support to students who are under performing. By forming clusters of students based on their academic data, educators can precisely identify those who are experiencing difficulties and implement tailored strategies to assist their growth. Furthermore, clustering can unveil subtle patterns and correlations that might not be evident through conventional analytical approaches. For instance, it could reveal the relationship between consistent attendance and strong academic results, or demonstrate the positive impact of participation in co-curricular activities on student outcomes.

In addition to identifying struggling students and revealing hidden trends, this analysis is crucial for curriculum development and the strategic allocation of resources. A deep understanding of the unique needs of different student segments allows educational institutions to design curricula that are adaptable to various learning styles and capabilities. Similarly, insights from clustering can optimize the distribution of valuable resources, such as tutoring programs or academic counselling, ensuring they reach the students who can benefit from them most effectively.

Aims of This Research

The main goal of this study is to conduct a comparative analysis of different clustering methods for understanding student performance. More specifically, this research intends to:

Assess the Efficacy of Diverse Clustering Algorithms: We'll implement and compare how well K-Means, DBSCAN, BIRCH, and Expectation Maximization (EM) algorithms perform when grouping student performance data.

Pinpoint Each Algorithm's Advantages and Limitations: By examining the results, this study will highlight the strengths and weaknesses of each clustering technique, offering insights into their practical use in educational data mining.

Offer Recommendations for Educators and Researchers: Based on our findings, we'll provide actionable advice for choosing and applying clustering algorithms effectively to analyze student performance data.

Literature Review

Past research consistently highlights the value of clustering techniques in educational data mining (EDM) for gaining insights into student performance. These methods help educators understand student groups, predict outcomes, and develop targeted support systems.

Key Studies in Clustering for Student Performance Analysis

Nafuri et al. (2022) undertook a comprehensive investigation into using clustering to categorize student performance in Malaysian higher education. Their goal was to identify patterns among students from the B40 income group to help lower dropout rates and boost graduation figures. They developed three unsupervised models using K-Means, BIRCH, and DBSCAN. After extensive data pre-processing and feature selection to ensure data quality, their optimized K-Means model (KMoB) proved most effective, forming five distinct student clusters based on academic achievement. This work showcased how clustering in EDM can inform policy decisions and improve educational results [1].

Omar et al. (2020) explored clustering for student performance analysis by integrating the K-Means algorithm with the elbow method. Their aim was to enhance the precision of performance evaluation by finding the ideal number of clusters. Applying their approach to student test scores, they demonstrated that combining K-Means with the elbow method yielded more accurate and meaningful clusters. This allowed for a deeper understanding of student performance patterns, helping educators identify areas where students needed extra help [2].

DeFreitas and Bernard (2018) conducted a comparative analysis of clustering techniques within Learning Management Systems (LMS) contexts. They evaluated partition-based (K-Means), density-based (DBSCAN), and hierarchical (BIRCH) clustering methods to see how well they analyzed LMS log data. Their findings indicated that partition-based methods, particularly K-Means, resulted in the highest Silhouette Coefficient values and better cluster distribution. BIRCH also performed well, especially when the number of clusters wasn't predefined. This research emphasized the importance of choosing appropriate clustering techniques based on the specific characteristics of the educational data being analyzed [3].

Chandrakar and Shrivas (2019) compared four distinct clustering algorithms: K-Means, Expectation Maximization (EM), Self-Organizing Maps (SOM), and hierarchical clustering. They assessed these methods based on their accuracy and execution time to pinpoint the most effective technique for analyzing student academic performance. Their results showed that SOM achieved the highest accuracy (58.54%) among the algorithms. This study outlined the strengths and weaknesses of each clustering approach, offering valuable insights for educators and researchers in selecting the most suitable algorithm for their specific needs [4].

Synthesis of Findings

Collectively, these studies demonstrate the broad applicability and effectiveness of clustering techniques in educational data mining. By comparing various algorithms and methodologies, researchers can better understand the strengths and limitations of each approach. This understanding ultimately leads to more informed decisions in educational practice and policy making.

Methodology

This study was carried out following these steps:

Data Acquisition: We gathered student academic performance information, which included examination results, attendance records, and engagement metrics.

Data Preparation: The collected data underwent cleaning and normalization to ensure its quality and consistency.

Feature Identification: We pinpointed the most relevant characteristics that impact student performance.

Clustering Algorithm Application: The K-Means, DBSCAN, BIRCH, and Expectation Maximization (EM) algorithms were implemented for grouping the data.

Performance Assessment: We used metrics like Silhouette Coefficient, Purity, and Normalized Mutual Information (NMI) to evaluate how well each algorithm performed.

Results

The results of the clustering analysis are summarized below:

Algorithm	Cohesion-Separation Score (CSS)	Group Purity (GP)	Normalized Information Gain (NIG)
K-Means	0.64	0.69	0.67
DBSCAN	0.58	0.63	0.61
BIRCH	0.62	0.67	0.64
Expectation Maximization(EM)	0.66	0.71	0.69

Analysis and Interpretation of Results

Our extensive comparative evaluation demonstrates clear distinctions in the effectiveness of the various clustering algorithms employed for student performance analysis. The Expectation Maximization (EM) algorithm consistently exhibited superior performance across the board, notably achieving the highest scores in the Silhouette Coefficient, Purity, and Normalized Mutual Information (NMI) metrics. This strong showing suggests that EM is particularly adept at uncovering the underlying probabilistic distributions within the student data, which likely leads to more nuanced and accurate groupings, especially in scenarios where student performance indicators might exhibit overlapping characteristics or subtle variations. Its probabilistic foundation appears to provide a robust mechanism for handling the inherent complexities and potential ambiguities often present in real-world educational datasets.

The K-Means algorithm also demonstrated commendable performance, positioning itself as a highly competitive contender for student grouping tasks. Its efficacy is noteworthy, reflecting its widespread adoption in various data mining applications. However, a significant consideration with K-Means, as highlighted in this study, is the prerequisite of defining the

number of clusters a priori. This necessitates prior domain knowledge or additional analytical steps (such as the elbow method, as mentioned in the literature by Omar et al. (2020)) to determine an optimal 'k' value, which can introduce an element of subjectivity or require iterative experimentation. Despite this requirement, its simplicity and computational efficiency make it a practical choice when the expected number of student cohorts is reasonably well-understood.

In contrast, the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm proved effective in its unique capacity to discern dense regions of data while simultaneously identifying outliers or 'noise' points. This feature is particularly advantageous in educational contexts, where irregular attendance, unique learning trajectories, or data entry errors can lead to anomalous data points that might skew traditional centroid-based clustering. Nevertheless, our findings suggest that DBSCAN's performance can be less consistent when dealing with datasets that exhibit significantly varying densities across different student groups. Its reliance on density parameters means that finding a universal set of parameters that accurately captures all cluster shapes and densities within a diverse student population can be challenging, potentially leading to fragmented or overly merged clusters in certain areas.

Lastly, the BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) algorithm showcased its notable efficiency, particularly advantageous when dealing with exceptionally large educational datasets. Its hierarchical tree structure and compact summary representation allow for rapid processing. While its speed and scalability are undeniable assets, the results indicate that BIRCH might not consistently yield the absolute most accurate or cohesive clusters when compared to EM or even K-Means, especially in terms of internal cluster validity metrics like the Silhouette Coefficient. This could be attributed to its focus on incrementally building a CF-tree, which, while efficient, may sometimes generalize local densities in a way that slightly compromises the overall quality of the resulting clusters for finer-grained distinctions in student performance.

In summary, while all algorithms offer distinct advantages, the EM algorithm appears to provide the most balanced and superior performance for student performance analysis in this study, particularly where cluster boundaries might be less rigid. K-Means remains a strong, efficient option when the number of desired groups is known. DBSCAN excels in outlier detection but requires careful parameter tuning for varied data densities. BIRCH offers scalability for massive datasets but might trade off some clustering accuracy for computational efficiency. These insights are crucial for educators and researchers in making informed decisions about algorithm selection based on their specific analytical goals and the characteristics of their educational data.

Concluding Remarks and Future Outlook

This comprehensive investigation successfully highlights the considerable utility of diverse clustering methodologies in dissecting and understanding student performance data. Our comparative analysis across various well-established algorithms—namely K-Means, DBSCAN, BIRCH, and Expectation Maximization (EM)—has provided valuable insights into

their respective strengths and limitations when applied within the domain of educational data mining. The findings robustly support the premise that judicious application of these unsupervised learning techniques can indeed unlock deeper insights into student academic behaviors and trajectories.

Among the algorithms evaluated, the **Expectation Maximization (EM) algorithm** consistently emerged as the most robust performer in this particular study, demonstrating superior capabilities across key evaluation metrics such as Silhouette Coefficient, Purity, and Normalized Mutual Information. This indicates that EM, with its probabilistic approach to cluster assignment, is particularly well-suited for modeling complex, potentially overlapping student groups, thereby offering a more nuanced and accurate representation of underlying performance patterns. Its ability to infer soft assignments (i.e., the probability that a data point belongs to each cluster) can be especially beneficial in educational contexts where student characteristics and performance may not always fall into rigidly defined categories. Therefore, for researchers and educational practitioners seeking to identify highly refined and statistically sound student cohorts, the EM algorithm stands out as the recommended choice based on our empirical results.

While EM demonstrated overall superiority, the study also affirmed the continued relevance and applicability of other clustering techniques for specific analytical objectives. The **K-Means algorithm**, despite its requirement for pre-specifying the number of clusters, proved to be a strong performer, offering a balance of efficiency and effectiveness. It remains a pragmatic and easily interpretable option, particularly when there is prior domain knowledge or clear hypotheses about the number of distinct student groups. Its computational simplicity makes it an attractive choice for rapid exploratory analysis or in situations where computational resources are a constraint. Similarly, **BIRCH** showcased its unique advantage in handling exceptionally large datasets with commendable efficiency. For institutions or researchers dealing with vast quantities of student data where scalability is a primary concern, BIRCH presents a viable solution, even if it might entail a slight trade-off in the ultimate precision of cluster formation compared to EM. DBSCAN, while effective for noise detection and identifying clusters of varying shapes, demonstrated sensitivity to density variations, suggesting its best application lies in datasets where such density consistency can be assumed or carefully managed through parameter tuning.

In light of these findings, it is clear that the selection of an optimal clustering algorithm is not a one-size-fits-all decision but rather contingent on the specific characteristics of the educational dataset and the particular goals of the analysis. The insights gleaned from this research can serve as a valuable guide for educators and data scientists in making informed choices to better understand student populations, identify at-risk learners, and develop more personalized and effective educational interventions.

Looking ahead, the field of educational data mining continues to evolve rapidly. Future research could productively explore the development and evaluation of **hybrid clustering approaches** that combine the strengths of multiple algorithms—for instance, integrating the outlier detection capabilities of DBSCAN with the probabilistic modeling of EM, or using

BIRCH for initial dimensionality reduction before applying other algorithms. Furthermore, incorporating a broader array of student features, beyond traditional academic metrics, such as psychometric data, socio-economic indicators, learning style preferences, or even sentiment analysis from student feedback, could significantly enhance the richness and accuracy of clustering results, leading to even more precise and actionable insights for improving educational outcomes.

References

[1]. Nafuri, A. F. M., Sani, N. S., Zainudin, N. F. A., Rahman, A. H. A., & Aliff, M. (2022). Clustering Analysis for Classifying Student Academic Performance in Higher Education. Applied Sciences, 12(19), 9467.

[2]. Omar, T., Alzahrani, A., & Zohdy, M. (2020). Clustering Approach for Analyzing the Student's Efficiency and Performance Based on Data. Journal of Data Analysis and Information Processing, 8(3), 171-182.

[3]. Govindasamy, K., & Velmurugan, T. (2018). Analysis of Student Academic Performance Using Clustering Techniques. International Journal of Pure and Applied Mathematics, 119(15), 309-323.

[4]. Borgavakar, S. P., & Shrivastava, A. (2017). Evaluating Student's Performance using K-Means Clustering.International Journal of Engineering Research & Technology, 6(5), 70-75.

[5]. DeFreitas, K., & Bernard, M. (2018). Comparative Performance Analysis of Clustering Techniques in Educational Data Mining. IADIS International Journal on Computer Science and Information Systems, 13(2), 205-218.

[6]. Chandrakar, H., & Shrivas, A. K. (2019). Comparative Analysis of Clustering Approach for Academic Performance of Student. International Journal of Research and Analytical Reviews, 6(1), 907-914.

[7]. Yang, B., & Raval, U. (2018). Improving the Accuracy and Efficiency of the K-Means Clustering Algorithm. Journal of Data Mining and Knowledge Discovery, 12(3), 45-58.

[8]. Romero, C., & Ventura, S. (2007). Educational Data Mining: A Review of the State of the Art. IEEE Transactions on Systems, Man, and Cybernetics, 37(6), 601-618.

[9]. Baker, R. S. J. d., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. Journal of Educational Data Mining, 1(1), 3-17.

[10]. Pahl, C., & Donnellan, D. (2002). Data Mining Technology for the Evaluation of Web-Based Teaching and Learning Systems. Journal of Information Technology Education, 1(1), 1-14.