Empirical Evaluation of EM, K-Means, and BIRCH Algorithms for High-Dimensional Educational Datasets

Sulagna Basu[1], Swarnabha Chakraborty[2], Abhik Bhattacharya[3], Shankar Prasad Mitra[4], Kaushik Paul[5], Debmalya Mukherjee[6], Ankita Sinha[7], Ranjan Banerjee[8], Ananya Smruti Snigdha Ojha[9]

[1,2,3,4,5,6,7,8,9]**Assistant Professor**

[1,4,5,7,8,9] **Computer Science & Engineering, Brainware University**

[2,3,6] **Computational Sciences, Brainware University**

sbd.cse@brainwareuniversity.ac.in,swc.cs@brainwareuniversity.ac.in, abh.cs@brainwareuniversity.ac.in, spmitra2016@gmail.com, kkp.cse@brainwareuniversity.ac.in kkp.cse@brainwareuniversity.ac.in,dbml.mukherjee@gmail.com, asn.cse@brainwareuniversity.ac.in, rbkpccst@gmail.com, ananyaojha1006@gmail.com

Orchid Id:0009-0004-2624-8889, Orcid Id.: 0009-0007-1372-6096, Orcid Id: 0009-0003-2392-2040, Orcid Id:0009-0003-5457-5534, Orcid Id: 0009-0009-6929-7651, 0009-0006-9946-0964
Orcid Id: 0009-0002-3179-227X, Orcid Id: 0009-0003-1950-7530, ORCID ID: 0009-0005-3844-1965

DOI: https://doi.org/10.63001/tbs.2026.v21.i02.S.I(2).pp560-565

KEYWORDS
Clustering Algorithms, Expectation-Maximization (EM), K-Means Clustering, BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), Comparative Analysis, High-Dimensional Data, Dimensionality Reduction, Curseof Dimensionality Scalability, Educational Data Mining (EDM), Learning Analytics, Student Profiling

Received on:
10-03-2026

Accepted on:
09-04-2026

Published on:
06-05-2026

Abstract

The rapid digitalization of higher education has produced vast repositories of learner data, necessitating advanced analytical frameworks to identify latent student performance patterns. This research presents a comparative investigation into four distinct clustering paradigms—partition-based (K-Means), density-based (DBSCAN), hierarchical (BIRCH), and probabilistic (Expectation-Maximization)—applied to a multi-dimensional dataset of academic and engagement metrics. Utilizing internal validity indices including the Silhouette Coefficient and Normalized Information Gain, we demonstrate that the Expectation-Maximization (EM) algorithm yields the most refined student segments, achieving a Group Purity (GP) of 0.71.[1] This paper delineates the theoretical trade-offs between "hard" and "soft" clustering assignments and provides a strategic roadmap for institutions to deploy data-driven at-risk intervention systems.

Introduction:

The Strategic Imperative of EDM

The current educational landscape is defined by the "unprecedented accumulation" of digital logs from Learning Management Systems (LMS) and collaborative platforms.[1] Traditional assessment methods, which often rely on retrospective grade analysis, fail to

capture the real-time, non-linear trajectories of modern students.[1] Educational Data Mining (EDM) has emerged to fill this gap, moving beyond descriptive statistics toward prescriptive analytics.

Clustering is uniquely suited for this domain because it identifies "natural groupings" without the bias of predefined success labels. As highlighted in the 2024 EDM conference theme, "New tools, new prospects, new risks," the goal is no longer just to group students, but to do so with an awareness of algorithmic fairness and the specific needs of diverse learner cohorts, such as those from lower-income (B40) demographics.

Theoretical Background and Literature Review

The efficacy of clustering in educational settings depends on the alignment between algorithm characteristics and the data's inherent structure.

Partitioning and Density-Based Evolution

K-Means has long been the baseline for EDM due to its efficiency in high-dimensional spaces. However, research by Chen and Smith (2018) noted that K-Means' sensitivity to outliers can skew results in datasets where attendance or engagement is irregular.[1] Density-based approaches like DBSCAN offer a solution by isolating noise points and identifying clusters of arbitrary shapes, though they require meticulous parameter tuning ($epsilon$ and $MinPts$) to avoid fragmented results.

The Probabilistic Shift

Recent studies, including those discussed in 2024–2025 journals, argue that student behavior is rarely binary. Instead, students often exhibit a "mix of academic behaviors," making probabilistic models like the Gaussian Mixture Model (GMM) more appropriate.[1] These models utilize the Expectation-Maximization (EM) algorithm to assign "soft" membership probabilities, providing a more nuanced representation of students who may be transitioning between different performance tiers.

Mathematical Architectures of Clustering Paradigms

To ensure analytical rigor, this study evaluates models across three major mathematical foundations.

Partitioning: K-Means WCSS Optimization

The K-Means algorithm partitions observations by minimizing the within-cluster sum of squares (WCSS):

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2$$

Where $||x_i^{(j)} - c_j||^2$ represents the Euclidean distance from a data point x_i to its respective centroid c_j.[1] While computationally fast, its performance relies heavily on a priori knowledge of k, often determined through the "elbow method" or Silhouette analysis.

Probabilistic: Expectation-Maximization for GMM

The GMM assumes that data points are generated from a mixture of K Gaussian

distributions, each with a mean μ, covariance Σ, and a mixing coefficient π [1]:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

The EM algorithm iteratively alternates between calculating the posterior probability of membership (E-step) and updating the distribution parameters to maximize the likelihood function (M-step).[2]

Scaling: BIRCH Clustering Feature Trees

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is designed for massive datasets. It constructs a Clustering Feature (CF) tree, where each CF is a triple (N, LS, SS) summarizing local data density.[1] This allows for a "one-time scan" of the dataset, making it the most scalable option for national-level educational portals.

Methodology and Comparative Analysis

Data Preparation and Pre-processing

The dataset used in this study integrates demographic data, examination results, attendance records, and engagement indices.[1] To ensure feature parity, Z-score standardization was applied to all numerical attributes:

$$z = \frac{x - \mu}{\sigma}$$

This prevents variables with larger ranges (e.g., test scores) from dominating distance calculations.[1] Missing values were addressed using mean imputation strategies to preserve the integrity of longitudinal records.[1]

Evaluation Metrics

We utilize three internal validity indices to objectively rank algorithm performance:

1. **Silhouette Score:** Measures the ratio of cluster cohesion to separation (range -1 to 1).
2. **Davies-Bouldin Index (DBI):** A lower score indicates better separation between clusters.
3. **Calinski-Harabasz Index:** Evaluates cluster compactness and separation based on the variance ratio.

Data Analysis and Results

The comparative evaluation demonstrates that probabilistic modeling significantly outperforms traditional partitioning in the student performance domain.[1]

Comparative Performance Metrics

Algorithm	Cohesion-Separation Score (CSS)	Group Purity (GP)	Normalized Information Gain (NIG)

Expectation-Maximization (EM)	**0.66**	**0.71**	**0.69**
K-Means	0.64	0.69	0.67
BIRCH	0.62	0.67	0.64
DBSCAN	0.58	0.63	0.61

Source: [1]

The EM algorithm's superior Group Purity (0.71) confirms its ability to handle "overlapping student groups" and "subtle variations" in academic indicators that other algorithms might miss.[1]

Discussion of Algorithm Efficacy

- **EM Robustness:** By inferring "soft assignments," EM provides a statistically sound representation of students who do not fit into rigid categories.[1] This is critical for identifying "borderline" at-risk students.
- **K-Means Reliability:** Ranking second in CSS (0.64), K-Means remains a "highly competitive contender" for institutions with well-defined student cohorts (k).[1]
- **BIRCH Scalability:** While slightly less accurate (CSS 0.62), BIRCH is identified as the ideal solution for "big data" educational scenarios where speed is a primary constraint.
- **DBSCAN Sensitivity:** Recording the lowest scores (CSS 0.58), DBSCAN proved sensitive to the varying data densities of different student segments, making it less suitable for general performance clustering without extensive manual tuning.[1]

Strategic Implications: From Clusters to Interventions

The results of this analysis allow institutions to move beyond simple grouping toward targeted pedagogical support.

1. **Identifying At-Risk Profiles:** High NIG (0.69) for the EM algorithm indicates that the resulting clusters are

highly informative.[1] This allows for "proactive identification" of students whose engagement probability is declining.

2. **Curriculum Refinement:** Clustering reveals how different student segments respond to varying instructional styles, enabling "curriculum optimization" for diverse cognitive aptitudes.[1]
3. **Resource Allocation:** By understanding the specific needs of each cohort, institutions can judiciously channel tutoring, advising, and mentorship resources toward those most likely to benefit.

Ethical Considerations: Bias and Transparency

As machine learning mediates educational outcomes, ethical governance is essential. Algorithmic bias, rooted in historical disparities, can lead to exclusionary support patterns if not carefully monitored. This research advocates for "ethics-by-design" principles, ensuring transparency in how clusters are formed and providing pathways for human oversight in any intervention recommendation.[3]

Conclusion and Future Outlook

This study confirms that the Expectation-Maximization (EM) algorithm provides the most effective framework for segmenting student populations in educational analytics.[1] While K-Means offers computational simplicity, the probabilistic nature of EM captures the nuance required for high-stakes at-risk detection.[1]

Future research should explore **hybrid clustering approaches**, such as utilizing DBSCAN for initial noise removal followed by EM for refined cluster generation.[1] Furthermore, incorporating multi-modal features—including sentiment analysis from learner forums and psychometric data—could significantly enhance the predictive accuracy and descriptive power of these models, fostering a more equitable educational environment.

References

[1] Banerjee, R., et al. (2024). Analyzing Student Achievement with Clustering Techniques in Educational Analytics. Brainware University.

[2] Chicco, D., et al. (2025). Evaluating Clustering with Silhouette Score and Davies-Bouldin Index. *Panamerican Mathematical Journal*.

[3] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. Yin, H., et al. (2024).

[4] A Rapid Review of Clustering Algorithms. *arXiv:2401.07389*. Gupta, S., & Sharma, R. (2019). Comparative Analysis of Probabilistic and Partition-Based Clustering in EDM.

[5] Zhang, T., et al. (1996). BIRCH: An efficient data clustering method for very large databases. *ACM SIGMOD Record*. Chen, Y., et al. (2023). Systematic Review of Cluster-Based Student Performance Prediction. *Applied Sciences*.

[6] Nafuri, A. F. M., et al. (2022). Clustering Analysis for Classifying Student Academic

Performance. *Applied Sciences*.

[7] Ghosh, A., et al. (2025). Optimising Educational Performance through Clustering Algorithm Evaluation. Educational Data Mining Conference. (2024). New tools, new prospects, new risks: EDM in the age of generative AI.

[8] Banerjee, R., Nath, S., Mitra, S. P., & Biswas, M. (2025). A Theoretical Study on Computational Linguistics and Conversational Agents: From Scripted Interactions to Contextualized Cognition. European Journal of Clinical Pharmacy.

[9] Banerjee, R., Basu, D., Mukherjee, D., & Nath, S. (2025). From Data Silos to Data Sovereignty: Building a Decentralized Data Ecosystem for Society 5.0. ResearchGate Publication.

[10] Banerjee, R., et al. (2025). Study on Computational Intelligence for Planetary Resilience: A Framework for Sustainable Biosphere Stewardship. European Journal of Clinical Pharmacy.

[11] Banerjee, R., Sengupta, P., Mitra, S. P., Chaudhuri, A. K., Mukherjee, D., Mondal, A., Gharai, P., Ghosh, A., & Ojha, A. S. S. (2025). SEO Poisoning: An In-Depth Analysis. *European Journal of Clinical Pharmacy*, 7(1), 506-512.

[12] Banerjee, R., Mukherjee, D., Basu, D., & Modak, P. K. (2025). The Mind In The Machine: Unveiling The Power Of Cyber Hypnosis. ResearchGate Publication.

[13] Banerjee, R. (2023). An Empirical Evaluation of k-Means Clustering Technique and Comparison. Conference Paper/ResearchGate.

[14] Mondal, A., Banerjee, R., Mukherjee, D., Sengupta, P., & Mitra, S. P. (2025). A Study on the Theoretical Underpinnings of Modern AI. *Journal of Applied Bioanalysis*, 11(4s), 21-27.