PERSPECTIVE

A Perspective on Student Behavior Analytics via Approximate Fast Clustering and Deep Learning

Shaoyuan Weng [1,*]

[1] Xiamen Institute of Software Technology, Xiamen 361024, China

Abstract

The rapid digitization of education has generated massive, heterogeneous student behavioral data, creating both opportunities and methodological challenges for educational data mining. This perspective paper discusses an emerging paradigm that integrates approximate fast clustering algorithms with deep learning techniques to enable scalable, high-resolution student behavior analysis. By leveraging approximate computation, clustering efficiency can be significantly improved without sacrificing analytical fidelity, while deep learning models facilitate the extraction of high-level representations from multimodal behavioral data. The synergy between these approaches could enable robust applications such as anomaly detection, performance prediction, and personalized intervention. In addition, key research directions are highlighted, including mixed-type feature modeling, interpretability, and real-time adaptive analytics, which are critical for advancing intelligent education systems.

Keywords: approximate fast clustering, deep learning, student behavior analysis, educational data mining.

Submitted: 23 February 2026
Accepted: 24 March 2026
Published: 26 March 2026

Vol. 2, **No.** 1, 2026.
10.62762/TEDM.2026.361754

***Corresponding author:**
Shaoyuan Weng
wengshaoyuany@163.com

1 Introduction

Traditional educational paradigms rely heavily on aggregated evaluation metrics [1], such as total scores and standardized assessments, which impose a flattened view of student performance. This "one-size-fits-all" framework fails to capture the diversity of learning behaviors and cognitive profiles across students. With the proliferation of campus information systems and online learning platforms [2], large-scale behavioral data (ranging from interaction logs to assessment trajectories) has become readily available.

This shift from data scarcity to data abundance has catalyzed the development of Educational Data Mining (EDM) [3], a field dedicated to extracting actionable insights from educational datasets. The core objective is no longer limited to performance evaluation but extends to understanding learning processes, diagnosing behavioral patterns, and enabling personalized education. However, the scale, heterogeneity, and dynamic nature of these data pose substantial challenges to conventional analytical methods.

Thus, this paper presents a forward-looking perspective on combining approximate fast clustering algorithms with deep learning models to address these challenges. The central idea is that efficiency-oriented clustering and representation

Citation
Weng, S. (2026). A Perspective on Student Behavior Analytics via Approximate Fast Clustering and Deep Learning. *ICCK Transactions on Educational Data Mining*, 2(1), 33–37.

learning are complementary where the clustering structures the data space while deep learning uncovers the latent semantics.

2 Scientific Significance and Application Prospects

2.1 From Aggregate Evaluation to Behavioral Intelligence

The integration of data-driven analytics into education fundamentally reshapes how student development is understood and managed [4]. Instead of relying solely on static evaluation metrics, behavioral analytics enables continuous monitoring and fine-grained diagnosis of learning processes.

By identifying patterns in learning behaviors (e.g., engagement frequency, problem-solving strategies, and time allocation), data mining techniques are able to generate personalized learning recommendations. Students can gain insight into their strengths and weaknesses while instructors can detect early warning signals and intervene proactively [5, 6].

From an institutional perspective, such analytics support evidence-based decision-making. Curriculum design can be optimized based on observed learning trajectories, and resource allocation can be aligned with student needs [7]. Ultimately, this leads to improved learning outcomes and higher satisfaction.

2.2 The Role of Scalable Algorithms

A critical bottleneck in current EDM systems is scalability. As datasets grow in volume and dimensionality, traditional clustering algorithms (e.g., k-means and hierarchical clustering) suffer from high computational complexity and limited adaptability [8].

Approximate fast clustering algorithms address this issue by introducing controlled approximations (e.g., sampling, sketching, or locality-sensitive hashing) to reduce computational overhead [9]. These methods maintain acceptable accuracy while enabling real-time or near-real-time analysis, which is essential for large-scale educational platforms.

2.3 Deep Learning for Multimodal Representation

Student behavior data is inherently multimodal, encompassing numerical scores, categorical attributes, temporal sequences, and even textual or visual information [10]. Deep learning models, particularly neural networks with hierarchical architectures, are well-suited for capturing such complexity. By automatically learning high-level feature representations, deep models eliminate the need for manual feature engineering and enable the discovery of latent behavioral patterns. When combined with clustering, these representations can significantly enhance the quality of student segmentation and downstream predictive tasks.

3 Research Landscape

3.1 Clustering-Based Educational Data Mining

Clustering has been widely used in EDM to group students based on behavioral or performance similarities. Early systems demonstrated that clustering can improve analytical accuracy and support personalized services [11]. For example, clustering-based frameworks have been applied to identify learning patterns in programming education and detect early signs of academic difficulty.

In addition, improvements to fuzzy clustering methods incorporated statistical distance measures to better handle data variability while density-based approaches enabled the discovery of arbitrarily shaped clusters in complex datasets [12]. However, most of these approaches rely on a single clustering paradigm, limiting their ability to capture multi-level or heterogeneous structures in student data.

3.2 Hybrid and Ensemble Approaches

To overcome the limitations of single-method approaches, researchers have explored hybrid and ensemble techniques. These methods combine multiple algorithms (e.g., clustering with classification or graph-based learning) to improve robustness and generalization. For example, graph-based ensemble learning has been used to integrate multiple predictive models [13]; the hybrid classification frameworks combining support vector machines and nearest neighbor methods [14]; and feature selection and weighting strategies integrated with clustering for improved segmentation [15].

These approaches demonstrate that algorithmic integration can uncover deeper insights into student behavior. Nevertheless, they often lack scalability and struggle with high-dimensional, multimodal data.

3.3 Deep Learning in Educational Analytics

Recent advances in deep learning have introduced powerful tools for modeling complex educational data [16]. Neural network architectures, including graph convolutional networks and attention-based models, have been applied to student performance prediction and recommendation systems. These models excel at capturing nonlinear relationships and long-range dependencies. For instance, attention mechanisms can dynamically assign importance to different behavioral factors, while graph-based models can exploit relational structures among students or knowledge concepts.

Despite their strengths, deep learning models are often criticized for limited interpretability and high computational demands, which hinder their deployment in real-world educational settings.

4 A Unified Framework: Approximate Clustering Meets Deep Learning

4.1 Conceptual Architecture

In this perspective paper, a unified analytical framework is proposed that integrates approximate fast clustering with deep representation learning, organized into three interconnected layers: (1) Approximate Fast Clustering Layer. This layer performs efficient partitioning of large-scale student behavioral data through approximate computation strategies. It provides a coarse-grained structural organization of the data space, enabling the discovery of inherent group patterns while significantly reducing computational overhead; (2) Deep Representation Learning Layer. Built upon the clustered data, this layer employs deep learning models to extract high-level feature representations from multimodal inputs. It captures latent behavioral semantics and complex nonlinear relationships that are difficult to model using traditional feature engineering approaches; (3) Task-Oriented Analysis Layer. The final layer focuses on downstream analytical tasks, such as anomaly detection and performance prediction. It integrates both the structural information derived from clustering and the semantic representations learned by deep models to support accurate and actionable decision-making.

Overall, this hierarchical design follows the principle of "structure first, semantics next, decision last", ensuring a coherent progression from data organization to knowledge extraction and ultimately to intelligent analysis.

4.2 Key Advantages

The proposed framework can provide several notable advantages:

The first advantage is scalability. Approximate clustering methods effectively reduce computational complexity, making the framework suitable for large-scale and high-dimensional educational datasets.

The second advantage is expressiveness. Deep learning models provide strong nonlinear modeling capabilities, enabling the extraction of rich and informative representations from heterogeneous data sources.

The third advantage is flexibility. The modular architecture supports a wide range of downstream tasks, allowing seamless adaptation to different analytical objectives in educational settings.

The fourth advantage is robustness. By integrating clustering and deep learning, the framework mitigates the limitations of individual methods, resulting in more stable and reliable analytical performance.

4.3 Application Scenarios

The proposed framework can be effectively applied to multiple educational data mining tasks, including (1) anomaly detection. Identifying students with atypical learning behaviors, enabling early warning mechanisms and timely pedagogical interventions; (2) performance prediction. Forecasting academic outcomes based on behavioral trajectories and latent feature representations, supporting data-driven evaluation and guidance; (3) personalized recommendation. Providing tailored learning resources and strategies by leveraging both cluster membership and learned behavioral features, thereby enhancing individualized learning experiences.

5 Challenges and Future Directions

5.1 Mixed-Type Data Modeling

Educational datasets are inherently heterogeneous, typically comprising continuous, categorical, and ordinal variables [17]. Traditional similarity measures and representation techniques are often tailored to a single data type, limiting their effectiveness in such complex settings. Developing unified modeling strategies that can accurately capture relationships across mixed-type features remains a fundamental challenge. Future work should explore hybrid distance metrics, embedding-based representations, and modality-aware learning mechanisms to better handle data heterogeneity.

5.2 Interpretability and Trust

In educational contexts, analytical outcomes often inform high-stakes decisions [18], such as academic interventions and performance evaluation. Therefore, model interpretability and transparency are of paramount importance [19]. While deep learning and clustering methods offer strong predictive capabilities, they are frequently criticized for their "black-box" nature. Future research should focus on integrating explainable artificial intelligence techniques into deep clustering frameworks, enabling the extraction of human-understandable rules, feature attributions, and decision rationales to enhance trust and accountability.

5.3 Real-Time and Adaptive Analytics

Learning environments are dynamic, with student behaviors continuously evolving over time [20]. Static models trained on historical data may quickly become outdated and fail to reflect current learning states. This necessitates the development of real-time and adaptive analytical frameworks capable of processing streaming data. Promising directions include incremental learning, online clustering, and continual learning mechanisms, which allow models to update their knowledge without retraining from scratch, thereby maintaining both efficiency and relevance.

5.4 Integration with Pedagogical Theory

Although data-driven approaches have demonstrated significant potential, their effectiveness can be limited when they operate independently of established educational theories. Bridging the gap between computational models and pedagogical principles is essential for ensuring both the validity and practical applicability of analytical outcomes. Future research should emphasize theory-informed modeling, where insights from learning sciences—such as cognitive processes, motivation, and knowledge acquisition—are explicitly incorporated into algorithm design and interpretation.

6 Conclusion

The convergence of approximate fast clustering algorithms and deep learning represents a promising direction for advancing student behavior analytics. By addressing both scalability and representational complexity, this integrated approach enables more accurate, efficient, and actionable insights into learning processes.

As educational systems continue to evolve toward intelligence and personalization, the development of robust, interpretable, and adaptive analytical frameworks will be essential. This perspective highlights not only the current state of the field but also the critical pathways for future research, aiming to bridge the gap between data-driven modeling and real-world educational impact.

Data Availability Statement

Not applicable.

Funding

This work was supported by the Fujian Provincial Young and Middle-aged Teachers' Educational Research Project (Science and Technology Category), China under Grant JAT241390.

Conflicts of Interest

The author declares no conflicts of interest.

AI Use Statement

The author declares that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Gallagher, C. J. (2003). Reconciling a tradition of testing with a new learning paradigm. *Educational Psychology Review*, *15*(1), 83–99. [CrossRef]

[2] Liu, M., & Yu, D. (2023). Towards intelligent E-learning systems. *Education and Information Technologies*, *28*(7), 7845–7876. [CrossRef]

[3] Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H. Y., & Hussain, A. (2023). Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies*, *28*(1), 905–971. [CrossRef]

[4] Sajja, R., Sermet, Y., Cwiertny, D., & Demir, I. (2025). Integrating AI and learning analytics for data-driven pedagogical decisions and personalized interventions in education. *Technology, Knowledge and Learning*, 1–31. [CrossRef]

[5] Fan, Z., Gou, J., & Weng, S. (2025). Complementary CatBoost based on residual error for student performance prediction. *Pattern Recognition*, *161*, 111265. [CrossRef]

[6] Fan, Z., Gou, J., & Weng, S. (2024). A feature importance-based multi-layer catboost for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, *36*(11), 5495–5507. [CrossRef]

[7] Song, C., Shin, S. Y., & Shin, K. S. (2024). Implementing the dynamic feedback-driven learning optimization framework: a machine learning approach to personalize educational pathways. *Applied Sciences, 14*(2), 916. [CrossRef]

[8] Ran, X., Xi, Y., Lu, Y., Wang, X., & Lu, Z. (2023). Comprehensive survey on hierarchical clustering algorithms and the recent developments. *Artificial Intelligence Review, 56*(8), 8219-8264. [CrossRef]

[9] Manohar, M. D., Shen, Z., Blelloch, G., Dhulipala, L., Gu, Y., Simhadri, H. V., & Sun, Y. (2024, March). Parlayann: Scalable and deterministic parallel graph-based approximate nearest neighbor search algorithms. In *Proceedings of the 29th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming* (pp. 270-285). [CrossRef]

[10] Chango, W., Lara, J. A., Cerezo, R., & Romero, C. (2022). A review on data fusion in multimodal learning analytics and educational data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *12*(4), e1458. [CrossRef]

[11] Nabli, H., Ben Djemaa, R., & Amous Ben Amor, I. (2024). Improved clustering-based hybrid recommendation system to offer personalized cloud services. *Cluster Computing*, *27*(3), 2845–2874. [CrossRef]

[12] Alasalı, T., & Ortakcı, Y. (2024). Clustering techniques in data mining: A survey of methods, challenges, and applications. *Computer Science*, *9*(1), 32–50. [CrossRef]

[13] Cui, C., Yao, Y., Zhang, C., Ma, H., Ma, Y., Ren, Z., ... & Ko, J. (2024). DGEKT: a dual graph ensemble learning method for knowledge tracing. *ACM Transactions on Information Systems*, *42*(3), 1-24. [CrossRef]

[14] Ramesh, L., Radhika, S., & Jothi, S. (2023). Hybrid support vector machine and K-nearest neighbor-based software testing for educational assistant. *Concurrency and Computation: Practice and Experience*, *35*(1), e7433. [CrossRef]

[15] Galis, F., Onchis, D. M., & Istin, C. (2025). Refining Filter Global Feature Weighting for Fully Unsupervised Clustering. *Applied Sciences*, *15*(16), 9072. [CrossRef]

[16] Lin, Y., Chen, H., Xia, W., Lin, F., Wang, Z., & Liu, Y. (2025). A Comprehensive Survey on Deep Learning Techniques in Educational Data Mining. *Data Science and Engineering*, *10*, 564–590. [CrossRef]

[17] Ikegwu, A. C., Nweke, H. F., & Anikwe, C. V. (2024). Recent trends in computational intelligence for educational big data analysis. *Iran Journal of Computer Science*, *7*(1), 103–129. [CrossRef]

[18] French, S., Dickerson, A., & Mulder, R. A. (2024). A review of the benefits and drawbacks of high-stakes final examinations in higher education. *Higher Education*, *88*(3), 893–918. [CrossRef]

[19] Brahmaji, K. K. P. (2024). Explainable AI in data analytics: Enhancing transparency and trust in complex machine learning models. *International Journal of Computer Engineering and Technology*, *15*(5), 1054–1061. [CrossRef]

[20] Zhou, H. (2025). Exploring the dynamic teaching-learning relationship in interactive learning environments. *Interactive learning environments*, *33*(7), 4363-4393. [CrossRef]

Shaoyuan Weng received the master's degree from the College of Computer Science and Technology, Huaqiao University, Xiamen, China, in 2021. She is currently a lecturer with the Xiamen Institute of Software Technology. Her main research interests include in the areas of analysis of data mining and collaborative task. (Email: wengshaoyuany@163.com)