Improving the Quality and Education Systems Through Integration's Approach of Data Mining Clustering in E-Learning

Agung Triayudi, Iskandar Fitri, Wahyu Oktri Widyarto, and Sumiati

Abstract Educational Data Mining (EDM) is a discipline developed by focusing on improving independent and adaptive learning methods to find hidden education patterns. In this area, heterogeneous data is known to continue to develop in a big-data paradigm. Several specific data mining techniques are required to extract information with an adaptive value from the available educational data. Therefore, this study aims to present a grouping approach related to partitioning students into a different group or cluster based on the students' behavior during lessons. Then, the architecture related to the e-learning system will be personalized to detect and provide suitable teaching methods and content according to each student's learning ability so that students can improve their quality and learning ability. The grouping methods that can be done in this educational data mining include K-Means, K-Medoids, Agglomerative Hierarchical Cluster Trees, Noise-Based Application Density-Based Spatial Clustering, and Fast Search and Density Peak Findings through Heat Diffusion (CFSFDP-HD) Shows the average compute time with different student count benchmarks: 600, 1200, 1800, 2400, 3000, 3600. Then, it has been found that the CFSFDP-HD method has strong results compared to other methods.

Keywords Educational data mining · Big data · Clustering · Profile learning · e-Learning

A. Triayudi (✉) · I. Fitri
Department of ICT, Universitas Nasional, Jakarta, Indonesia
e-mail: agungtriayudi@civitas.unas.ac.id

W. O. Widyarto
Industrial Engineering Department, Universitas Serang Raya, Serang, Indonesia

Sumiati
Informatics Department, Universitas Serang Raya, Serang, Indonesia

T. Triwiyanto et al. (eds.), *Proceedings of the 2nd International Conference on Electronics, Biomedical Engineering, and Health Informatics*, Lecture Notes in Electrical Engineering 898, https://doi.org/10.1007/978-981-19-1804-9_3

1 Introduction

The Educational Data Mining method is concerned with modern and up-to-date adaptive studies and developments, where available instruments are used to analyze and visualize hidden patterns of educational datasets. [1]. The educational data also depends on the geographical conditions, most of which are structured, semi-structured, and unstructured [2]. EDM appears as an intrinsic data analysis research field also has a function in extracting information related to hidden pattern predictions to obtain information related to the educational data set. EDM is considered as an effort to implement educational data mining methods that can generate new patterns also analyze the big data artificially with efficient results [3, 4].

Nowadays, technology such as Artificial Intelligence (AI), Internet of Things, Sensor, and various social networks are integrated into education systems to improve teaching and learning activities [5]. At an educational level, the amount of data or information available can provide new insights in prediction or decision-making efforts that can improve learning skills by both teachers and learners. It is why educational data mining plays a significant role to improve the quality and education systems, through: (1) the discovery of new perspectives related to experimental data, (2) the discovery of of knowledge along with associations from various fields, and (3) the improvement of the quality of individual-based education. Along with the development of increasingly advanced and fast-paced communication technology, much use of sensors and intelligent devices are implemented into the education systems to observe the patterns and behaviors from all elements involved in the education system, where it has a variety of information related to thoughts and events that are present in the form of semi-structured or unstructured. Thus, in a virtual-based education system, there is a need for improvement efforts by implementing all techniques available in data mining to meet the needs of an educational institution effectively [6, 7]. There are many models and data mining met hods that can be implemented into an education field, where these methods can be categorized as classification, clustering, relationship mining, and neural networks. *Grouping* is the primary unattended method that is useful for partitioning data sets into a separate cluster based on estimates of intrinsic characteristics or similarities. This method can also be categorized by partition-based, model-based, density-based and hierarchical. The grouping methods that can be done in this educational data mining include K-Means, K-Medoids, Agglomerative Hierarchical Cluster Trees, Noise-Based Application Density-Based Spatial Clustering, and Fast Search and Density Peak Findings through Heat Diffusion (CFSFDP-HD) [8, 9].

2 Literature Review

2.1 Techniques of Education Data Mining Clustering

EDM serves as a research area involving a series of collaborations between psychological and computational methods [10]. Research related to educational data mining has also been done to find techniques in analyzing learning related to big data to know the adaptive learning systems (ALS) model [11]. This ALS model seeks to empower teachers to always quickly understand the needs and adjustments that need to be made to keep up with the changes that continue to occur related to this educational unit. Other research also states that the exploration that has been done on various studies and data sets related to the field of EDM can be used in recognizing students who have a risk of failure in learning, distinguishing adaptation methods of groups or individual students, increasing the number of graduations, and improving the quality and quality of educational institutions [12, 13].

Related to the various benefits that can be obtained from the implementation results, K-means, as one of the partition-based grouping algorithms commonly applied to EDM, can understand student behavior based on available data, as well as consider the impact of self-characteristics on the performance of students, so that it will present clear types of classifications [14, 15].

2.2 Personalized E-Learning System and Data Mining Clustering Approach

The architecture of e-learning systems focuses on individual needs and makes predictions of references or interests from users. Considering e-learning not only focus on teachers or students to meet virtually but also creates other electronic resources. In Fig. 1, the following is the architecture of a personalized e-learning system, where the steps in PESA are described as follows [16, 17]:

- Student's Profile
- Data Mining Clustering Approach
- Recommendations
- Database
- Virtual Systems
- Considering the big data set of academic databases implemented by educational data mining methods in this study, a virtual display containing all the needs of an institution is presented to students in the form of electronic documents [18].

Key Steps Involved in The Recommended Educational Data Mining Approach. The data mining approach recommended in this study (CFSFDP-HD) was implemented using the tools available in data mining to analyze students' patterns and behaviors and simulate them into educational data. The simulated educational

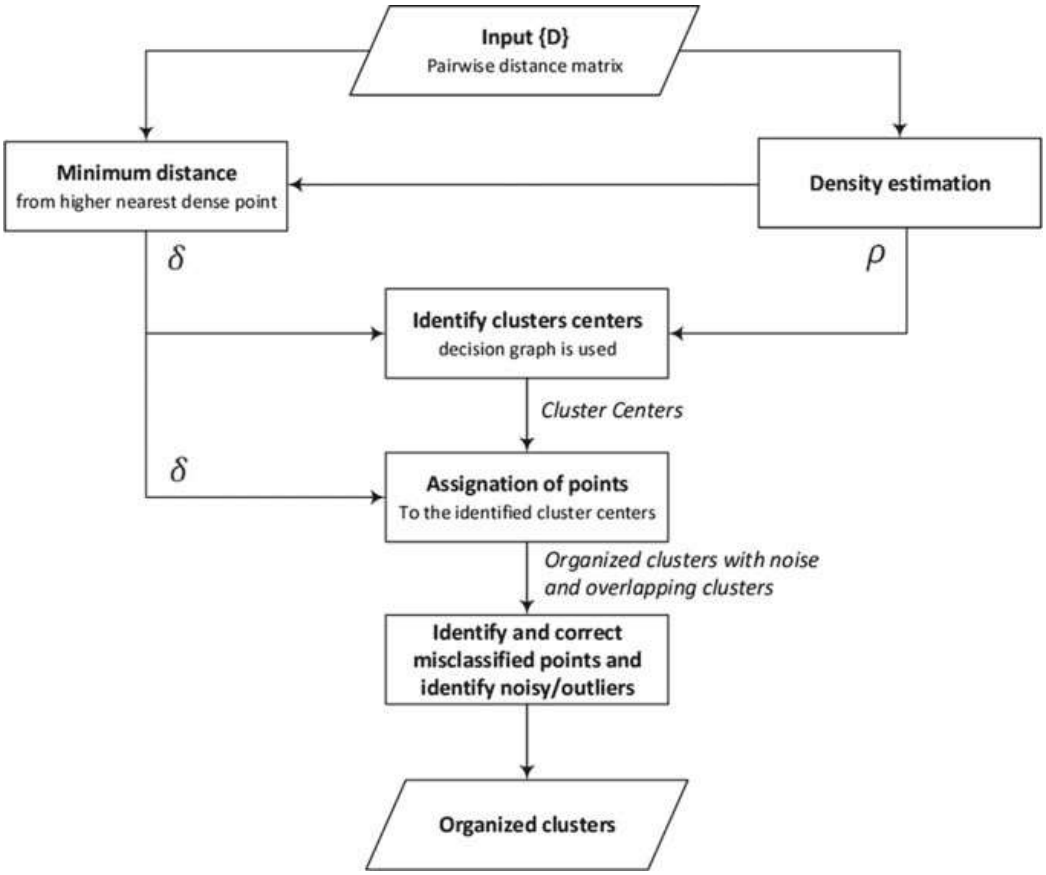

Fig. 1 Key steps involved in the recommended educational data mining approach [17]

data contains: (1) the grades obtained by each student and (2) attendance in the classroom. The indicators on the grades obtained by each student consist of: (1) two quizzes, (2) four assignments, (3) midterm exam, and (4) final semester exam.

3 Result and Discussion

In this study, there was a dataset containing 600 characteristics of students from different classes and sessions, which were then simulated using the approach of CFSFDP-HD to partition them into groups according to the similarity of characteristics and other indicators based on the scores obtained (quizzes, assignments, midterm exams, final semester exams) and attendence in the classroom (Table 1).

This effort is necessary to design suitable teaching methods according to each student's characteristic needs so that an effective learning process will be realized and improve the quality and performance of each individual's learning. In the image below, the decision graph is defined as follows:

Table 1 Dataset

Subject	Students	Datasets	Posts	Times
Computer security	280	Submission (assignment)	189	90
		Course modul (forum)	443	120
Knowledge management	320	Discussion (forum)	108	25
		Course view	692	234
		Observe	85	27

The decision graph shown is based on (1) estimation through heat diffusion and (2) calculating the minimum distance of δi points closest to higher density. At this stage, the cluster center identification process is obtained by implementing the decision graph as shown in Fig. 3a, where the outlier acts as a potential cluster center and is shown by a different color. The clusters that have been found are displayed with various color schemes according to Fig. 3b, where 2D non-classical multidimensional scaling applies to visualizing datasets. In this process, the recommended CFSFDP-HD approach is adaptive, so the available parameters do not need to be explicitly set.

Based on the categories of students obtained through classification based on Figs. 3a and b, the approach of teaching methods can begin to be adapted to each student's learning needs. So there will be improvements related to student performance because the methods they face to improve performance quality are following their characteristics. The student achievement index can also be further analyzed to improve the quality of an educational institution's system.

The grouping method implemented in this study consists of K-means, K-medoid, AHCT, and DBSCAN, which are then simulated through a dataset of 600 students and outlined in Table 2 below.

Based on the comparison of approaches recommended by CFSFDP-HD through K-Means, K-Medoids, AHCT, and DBSCAN, the resulting decision graph is proven to help provide information in-depth insights to select clusters that have the potential to meet all needs related to educational aspects. The implementation of K-means

Table 2 Clustering method based on the indicator value

Method	Indicator
K-means	The number of clusters (k = 7) The number of iterations (n = 40)
K-medoids	The number of clusters (k = 5) Predefined the number of iterations (n = 40)
DBSCAN	Epsilon (eps = 0.5) Minimun points (minPts = 10)
AHCT	The number of clusters (k = 6)
CFSFDP-HD	Does not require the number of clusters and iterations

and K-medoids has been done repeatedly with various input settings (such as cluster count and iteration) to get the most optimal cluster and provide the right solution of multiple iterations that have been done. Meanwhile, the AHCT approach explicitly defines the number of clusters, making it easier to check, visualize, and compare, as shown in Fig. 2b. This is in contrast to the data visualizations presented in Fig. 3a–d.

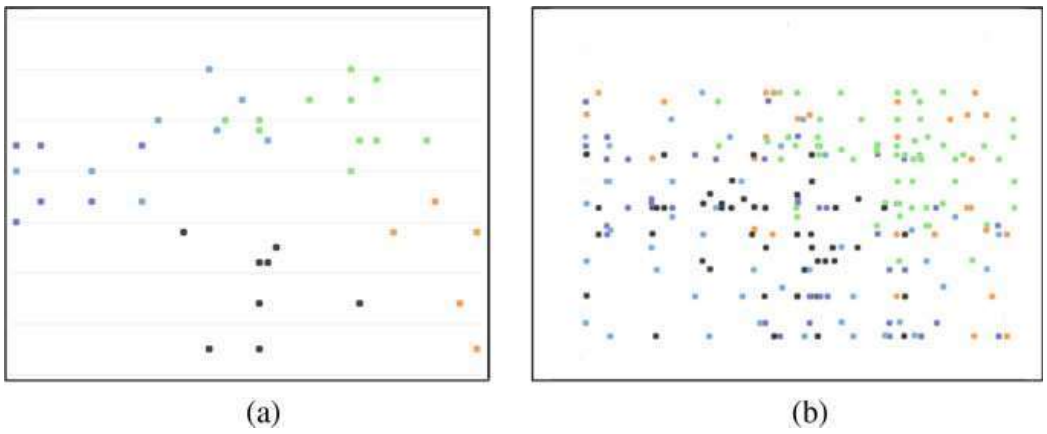

Fig. 2 **a** Decision graph with adjusted parameters, **b** Identify clusters with different color schemes

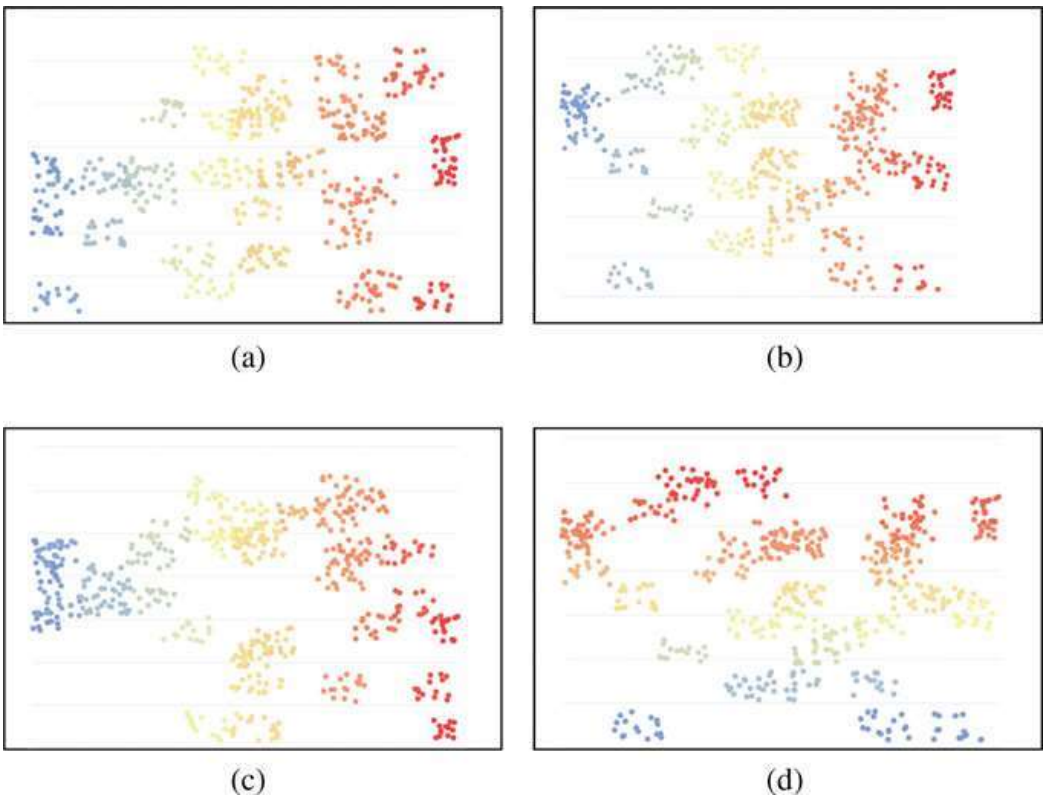

Fig. 3 Visualization of **a** K-means clustering, **b** K-medoids clustering, **c** AHCT clustering, **d** DBSCAN clustering

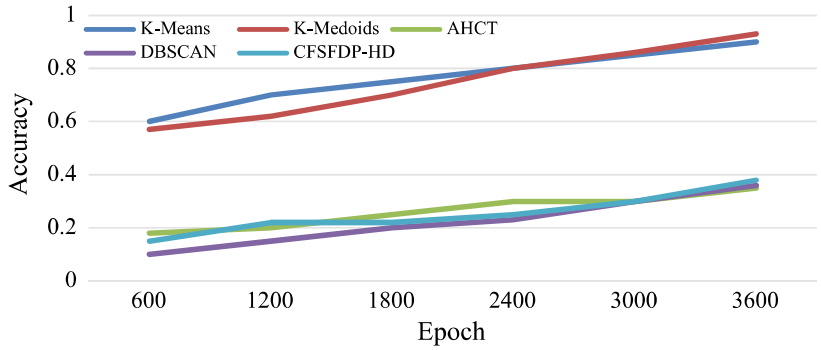

Fig. 4 Time execution comparison based on the different number of students

The representated data point in Fig. 2b with Fig. 3 is also different given the differences in the location of the data displayed. Based on ρ and δ grades, students with exemplary achievements are on the upper side in Fig. 2b. While in Fig. 3, students who have exemplary achievements are on the right side. To get the cluster form with the implemented approach in this study, it's important to know the number of existing clusters and the possibility of difficulties in detecting low or below average student performance. Of course, this possibility can lead to limitations in finding hidden patterns that are present in the process of classifying data. On that basis, the CFSFDP-HD method is recommended to find hidden patterns without even knowing the technicalities related to the data referenced by the study.

In this study, it was also found that the CFSFDP-HD method is known to be more adaptive and has significant results when compared to some previous approaches. The study also calculated the estimated time in executing the CFSFDP-HD approach and determining whether optimization efforts are required to predict the number of big data sets clusters accurately. Figure 4 below shows the average compute time of K-means, K-medoids, AHCT, DBSCAN, and CFSFDP-HD approaches with different student count benchmarks: 600, 1200, 1800, 2400, 3000, 3600. Each available approach has been run several times to take the required average time. In Fig. 5 below, the number of students is displayed on the x-axis, while the y-axis displays the execution time (in seconds).

Based on the results represented in Fig. 4, it is known that the recommended approach, namely CFSFDP-HD, runs quite effectively and efficiently compared to K-means and K-medoids. The recommended approach also takes less time than AHCT on smaller datasets. Although the DBSCAN approach takes less time on smaller data sets, AHCT and DBSCAN methods are not recommended because they are less efficient when applied to larger data sets. Through this consideration, the CFSFDP-HD approach is recommended to accurately measure the patterns and behaviors of students and present execution results related to significant and real datasets.

4 Conclusion

The available data mining approach provides convenience in e-learning systems to run effectively and efficiently. In this study comes the adaptive data grouping approach "CFSFDP" integrated into the e-learning system architecture where most queries are responded to without the need for intelligence and heuristics. This study has also been conducted an assessment of the application of grouping in big data. Experiments to find a suitable approach to grouping educational data need to be done by knowing the number of clusters and the shortcomings concerning the size of the identified data cluster. In this case, the approach recommended in this study, "CFSFDP" is considered capable of analyzing big data to strengthen the education system. This study also shows that the approach has formed a real cluster with less execution time compared to other approach methods. Besides, this approach also serves in solving various obstacles and needs related to elements in the field of education.

Acknowledgements This research is the result of the basic research scheme of the Indonesian Dikti grant B/112/E3/RA.00/2021.

References

1. Hamoud A, Hashim AS, Awadh WA (2018) Predicting student performance in higher education institutions using decision tree analysis. Int J Interact Multimedia Artif Intell 5:26–31
2. Triayudi A, Sumiati S, Dwiyatno S, Karyaningsih D, Susilawati (2021) Measure the effectiveness of information systems with the naïve bayes classifier method. IAES Int J Artif Intell 10(2)
3. Lu OH, Huang AY, Huang JC, Lin AJ, Ogata H, Yang SJ (2018) Applying learning analytics for the early prediction of students' academic performance in blended learning. Educ Technol Soc 220–232
4. Zaffar M, et al (2017) Performance analysis of feature selection algorithm for educational data mining. 2017 IEEE conference on big data and analytics
5. Zhang W, Li J (2015) Extended fast search clustering algorithm: widely density clusters, no density peaks. arXiv preprint arXiv:1505.05610
6. Yang F, Cao J, Zhou K, Zhang P, Wang Y (2018) An adaptive clustering algorithm based on CFSFDP. In: 2018 33rd youth academic annual conference of Chinese association of automation (YAC). pp 404–408
7. Xu H, Yao S, Li Q, Ye Z (2020) An improved k-means clustering algorithm. In: 2020 IEEE 5th international symposium on smart and wireless systems within the conferences on intelligent data acquisition and advanced computing systems. pp 1–5
8. Wang Y, Chen Q, Kang C, Xia Q (2016) Clustering of electricity consumption behavior dynamics toward big data applications. IEEE Trans Smart Grid 7(5):2437–2447
9. Sheng K, Liu Z, Zhou D (2017) An adaptive resampling algorithm based on CFSFDP. In: 2017 2nd IEEE international conference on computational intelligence and applications (ICCIA). pp 41–45
10. Silva C, José F (2017) Educational data mining: a literature review. In: Europe and MENA cooperation advances in information and communication technologies. Springer, Cham, pp 87–94

11. Nakayama M, Mutsuura K, Yamamoto H (2018) Using note taking instructions to reform student's note taking activities and improve learning performance in a blended learning course. Int Conf Inf Visualisation 326–331
12. Hernández-Blanco A, Herrera-Flores B, Tomás D, Navarro-Colorado B (2019) A systematic review of deep learning approaches to educational data mining. Complexity 2019
13. Salloum S, Muhammad A, Ashraf E, Khaled S (2020) Mining in educational data: review and future directions. In: Joint European-US workshop on applications of invariance in computer vision. Springer, Cham pp 92–102
14. Agung T, Widyarto OW, Vidila R (2020) CLG clustering for mapping pattern analysis of student academic achievement. ICIC Express Lett 14(12):1225–1234
15. Ahuja R, et al (2019) Analysis of educational data mining. In: Harmony search and nature inspired optimization algorithms. Springer pp 897–907
16. Baker RS (2019) Challenges for the future of educational data mining: the baker learning analytics prizes. JEDM, J Educ Data Min 11(1):1–17
17. Kausar S, Huahu X, Hussain I, Wenhao Z, Zahid M (2018) Integration of data mining clustering approach in the personalized e-learning system. IEEE Access 6:72724–72734
18. Fynn A et al (2018) A comparison of the utility of data mining algorithms in an open distance learning context. S Afr J High Educ 32(4):81–95
19. Rosalina V, et al (2020) Measuring citizen readiness to adopt electronic citizen relationship management (E-CIRM) using technology readiness index (TRI). J Theor Appl Inf Technol 98(21): 3416–3425
20. Ferraz R, Marcus V, Renan A, Luc Q (2019) Implementation of a distance learning program focused on continuing medical education with the support of patent-based data mining. Revista de Gestão
21. Wang R (2021) Exploration of data mining algorithms of an online learning behaviour log based on cloud computing. Int J Continuing Eng Educ Life Long Learn 31(3):371–380
22. Islam O, Siddiqui M, Aljohani NR (2019) Identifying online profiles of distance learning students using data mining techniques. In: Proceedings of the 2019 The 3rd international conference on digital technology in education. pp 115–120
23. Qi Z (2018) Personalized distance education system based on data mining. Int J Emerg Technol Learn 13(7)
24. Gonçalves AFD, Maciel AMA, Rodrigues RL (2017) Development of a data mining education framework for data visualization in distance learning environments. In: International conference on software engineering and knowledge engineering
25. Xu Y, Miao Z, Zhigang G (2019) The construction of distance education personalized learning platform based on educational data mining. In: International conference on applications and techniques in cyber security and intelligence. Springer, Cham pp 1076–1085