Analysis of Educational Data Mining

Ravinder Ahuja, Animesh Jha, Rahul Maurya and Rishabh Srivastava

Abstract This research paper aims to compare the performance of various clustering and classification algorithms which are applied on the same educational dataset. Educational Data Mining (EDM) uses these algorithms to explore educational statistics to discover patterns and predictions in data that illustrate learner's performance. Various design challenges such as accuracy, objective and functionality, and overheads when the data set is extremely large, etc., have been highlighted. The algorithms discussed here can be classified as centroid-based clustering, graph-based clustering, and various supervised classification algorithms. Further, after comparison of these algorithms, this paper aims at using the cited literature survey to determine the most suited algorithm according to the need of EDM clustering or classification. The regular search is on to understand students better and to know the patterns in which they learn to make it more efficient for them. Nowadays, many educational institutions have educational databases that can be utilized in various ways to make it effective for students but are unutilized. Powerful tools are required to get benefits from these educational databases. EDM is one of those emerging tools that analyzes the data collected from learning and teaching and then applies the techniques from machine learning and data mining for predicting student's future behavior by learning detailed information such as student's grades, knowledge, achievements, motivation, and attitude.

Keywords Clustering methods · Data mining · Educational mining
Analytic study · Educational data

R. Ahuja (✉) · A. Jha · R. Maurya · R. Srivastava
Jaypee Institute of Information Technology, Noida, India
e-mail: ahujaravinder022@gmail.com

A. Jha
e-mail: jhaanimesh1996@gmail.com

R. Maurya
e-mail: rahulmaurya97@rediffmail.com

R. Srivastava
e-mail: rishabh22dec@gmail.com

N. Yadav et al. (eds.), *Harmony Search and Nature Inspired Optimization Algorithms*,
Advances in Intelligent Systems and Computing 741,
https://doi.org/10.1007/978-981-13-0761-4_85

1 Introduction

Last decade, the increase in the number of higher educational institutions has been enormously high, which resulted in more increase in the number of undergraduates and postgraduates every year. Many universities have made changes in their teaching methods or conducting examinations but are still facing issues of unemployment and dropout students. Understanding factors for low performance or increase in dropout rate is a difficult task which includes past and present performances of students and their discipline records. EDM is one of those powerful tools to analyze and predict their future behavior and performances of the students [1].

Although the institutions/universities have maintained good educational databases, they are not being utilized in any decision-making and to increase the students performance or decrease the dropout rates or unemployment. The aim of EDM is to discover methods to progress education and students efficiency by lowering their dropout rate. EDM uses preprocessed clustering and classification for mining the data. Clustering is an unsupervised method for examining information. It consists of grouping together of similar data items to form a cluster or a group. The best clustering algorithms are considered to be those that in which the data items within the same cluster have the maximum similarity and the data items belong to different clusters have minimum or no similarity at all. Classification is a supervised approach of learning, which is based on utilizing the known results and predicting the future results based on given constraints. EDM uses classification algorithms is various prediction fields for institutional effectiveness. The success of classification algorithms depends on how varying is the data that is provided and many other factors. Universities need to have meeting among them and identify the common reasons behind the low performances of students and their dropping out so that they can predict up to the mark of students' performance and behavior, which is ultimately going to help students and institution as students will be able to know their weakness and could take actions in advance or prepare themselves from getting indulged into those factors. It will result in winning of all participants of universities, i.e., management, teachers, students, and parents. Teachers could plan about their lectures and can advise such students to perform better. Students can work on their problems and take advice from teachers to improve. Parents can be assured of better performances of their child in such guidance. Management will be able to provide better facilities for both teachers and students.

1.1 Clustering Algorithms

Clustering is a way of grouping together similar data items. According to statistical notation, clustering is the most significant unsupervised clustering algorithm [2]. Clustering is often used for preprocessing and hence it removes not required data and groups them into clusters, this makes datasets easy for further analysis. Since clustering reduces the size of the data, it leads to loss of information and one must understand and work according to that.

Clustering algorithms can be classified into two ways:

a. **Hierarchical Clustering**

In this type of clustering algorithms, clusters have a hierarchy among them.

b. **Partitioned Clustering**

In this type of clustering algorithms, the data items are partitioned into separate clusters. For example, *k*-means clustering algorithms, PSO clustering algorithm. etc.

Classification algorithms can be classified as follows:

a. **Decision Tree Based Classification**

These types of algorithms use the construction of decision tree on the (training data) to predict the output for the test data. For examples, Naïve Bayes algorithm, KNN algorithm, and C4.5 algorithm.

b. **Ensemble Learning**

These types of algorithms use the collection of statistical classifiers to classify a given test data item based on the trained data.

For example, AdaBoost algorithm.

c. **Rule Mining/Formulae Based**

These algorithms use a specified set of rules and mathematical formula to train data and classify the given data set by fitting them onto these rules.

For example, Apriori Algorithm, SVM algorithm.

Clustering algorithms can be applied to Big Data with ease. Big Data means voluminous data that keeps on increasing with very fast velocity for example bookkeeping of student's examination and results records. Now we have seen that several studies have been conducted on how clustering can help grouping together the student's data and analyzing them to predict what subject the student is interested in.

So, educational data system can be divided in two ways one is E-Learning (web-based study, online forums, etc.) and the other is classroom learning.

2 Educational Data Mining by Means of Clustering

We have seen that clustering techniques can be classified into various algorithms. Various studies have been conducted using hierarchical and nonhierarchical clustering. This has been explained in Table 1.

1. Wook et al. [3] applied a combination of ANN and Farthest-first method as *k*-Means and a decision tree classification as classifier. They used a data set from (NUDM). As a result of their study, they suggested to use PSO algorithm for clustering instead of *k*-means.

Table 1 Research papers published in clustering in educational data

Educational data and clustering		Published paper
Hierarchical clustering algorithm	Agglomerative clustering	[7, 8]
Nonhierarchical algorithm	K-Means	[9–11]
	C-Means	[12]
	Co-Operative Particle Swarm Optimizer (PSO)	[13]
	Farthest first	[14]
	Expectation Maximization(EM)	[15]

2. Parack et al. [4] used various DM algorithms to analyze the behavior and computation cost using c-means. This paper provided no specific directions as to from where the data set was derived. They basically worked on identifying key variables that affect a student's performance. As a result, they found that K-means and C-means cost more than the computed cost of an optimized PSO algorithm.
3. Chi et al. [5] conducted a research with a objective of identifying student's profile based on the research of their online browsing. For this, they used K-means as step 1 clustering and collaborative classifier as the classification tool.
4. Chen et al. [6] in their study applied clustering algorithm (Farthest-first) and EM algorithm to recognize the association between student's marks and their education online or offline. It was found out as a result of their study that a positive relationship exists between a student with high quality marks and the student pursuing online or E-Learning.

3 Integrated Summary

Data mining is the process of extraction or mining of useful (nontrivial, absolute, previously unknown, and imaginably useful) enlightenment from large datasets using variety of data-driven decision making such as classification, clustering, association rule mining, etc., which helps in decision-making and answering of questions. The type of data mining that deals with how a student performs, learns, reads, and writes is studied using educational data and this study or mining of information is the branch of data mining that is called Educational Data Mining (EDM). The objective of EDM was and has always been to analyze and pursue research in the field of Education. In recent years, there have been massive strides in educational sector which has led to the overall growth of educational data and as a result, mining of educational data has gained significant importance to understand behavior of student's during learning process and if occurs then understand student's performance. Conventionally, educational investigators have been using approaches such as studies, conferences, effort sets, and classroom actions to collect data related to student's learning experiences. These methods are usually very time consuming, thus cannot be duplicated or repeated with high frequency. To solve such problems and for overall enhancement, many surveys, interviews, and classroom discussions were conducted. But since this process is very time consuming and tiring, they cannot be achieved with high frequency. To overcome such difficulties, EDM was introduced. The field of the EDM uses newer learning analytics and techniques of mining that focuses on analyzing data from various transcripts, course management systems, and researches from schools and institutions and hence facilitates a quality decision-making. Our research study analyzes both the structured data and the unstructured data that come from unorganized environment to understand student learning acquaintance with the use of educational data mining tools and techniques (Table 2).

4 Conclusion

Analysis of over three decade's research on educational data mining has been presented by this paper. Several existing literatures have been reviewed and further future avenues based on the insights of EDM have been identified by this paper. The clustering methods help us to know how key variables such as learning in groups, learner's behavior in class, time spent on learning a particular topic, how motivated the student is, and the environment of the classroom. Clustering on EDM provides various useful insights and it can be multilevel nonhierarchical and hence the researchers must carefully choose the algorithm and the variables that result in better and accurate clusters and hence provide useful information.

Table 2 Review of various papers published in Educational Data Mining

Category	Year, Author(s)	Methodology	Key findings
Survey of papers published in Educational Data Mining	2014, Peña-Ayala, Alejandro [16]	Statistical and clustering processes	Identified kinds of educational systems, disciplines, tasks, methods, and algorithms
	2010, Romero, Cristóbal, and Sebastián Ventura [17]		Listed tasks in educational area resolved through data mining and future lines. Suggested to develop more unified and collaborative studies
	2009, Baker, Ryan SJD, and Kalina Yacef [18]		Identified key features of researches in EDM as discovery with models, emergence of public data, and tools
	2007, Romero, Cristóbal, and Sebastian Ventura [19]		Presented survey on application of data mining on traditional educational systems. Emphasized on the need of much more specialized work
Predicting academic performance with pre/post enrollment factors	2014, Saranya, S., R. Ayyappan, and N. Kumar [20]	Naive Bayes algorithm	Graphically represented Institutional Growth Prognosis and Students' Progress Analysis
	2014, Archer, Elizabeth, Yuraisha Bianca Chetty, and Paul Prinsloo [21]	Experimental usage of employee profiling software	Experimented the usage of a commercial product generally used for employee profiling in corporate, for higher education environment
	2014, Hicheur Cairns, Awatef, et al. [22]	Clustering technique	Professionals' data was analyzed during training of a consulting company

(continued)

Table 2 (continued)

Category	Year, Author(s)	Methodology	Key findings
	2014, Arora, Rakesh and Dharmendra Badal [23]	Association analysis algorithm	Found set of weak students based on graduation and postgraduation marks
	2012, Osmanbegović, Edin, and Mirza Suljić [24]	Chi-Square Test, One R-Test, Info Gain and Ratio Test, Naive Bayes, DTree	Found predicting model for academic performance that is user friendly for professors or nonexpert users
	2012, Sukanya, M., S. Biruntha, Dr. S. Karthik, and T. Kalaikumaran [25]	Bayesian classification method	Analyzed and assisted the low academic achievers in higher education
	2011, Torenbeek, M., E. P. W. A. Jansen, and W. H. A. Hofman [26]	Structural equations modeling, correlation matrix	Examined two variables, Pedagogical approach and skill development in the first 10 weeks of enrollment
	2011, Yongqiang, He, and Zhang Shunli [27]	Association rules analysis	Guidance provided for scientific management and comprehensive evaluation of students
	2011, Sakurai, Yoshitaka, Tsuruta, and Rainer Knauf [28]	Decision tree	Estimated success chances of curricula by implementing student profiling with storyboard system
	2011, Aher, Sunita B., and L. M. R. J. Lobo [29]	Classification and clustering	Analyzed the performance of final year students
	2010, Ayesha, Shaeela, Tasleem, Ahsan, Inayat [30]	K-Means clustering	Analyzed students' learning behavior to check the performance of students and predicted weak students
	2010, Kovacic, Zlatko [31]	Classification tree models	Investigated enrolment attributes to pre-identify success of students

(continued)

Table 2 (continued)

Category	Year, Author(s)	Methodology	Key findings
	2010, Al-shargabi, Asma A., and Ali N. Nusari [32]	Clustering, association rules and decision trees	Analyzed students' academic achievement, students' drop out, and students' financial behavior
	2010, Yan, Zhi-min, Qing Shen, and Bin Shao [33]	Rough set theory	Students' grades were analyzed
	2010, Ningning, Gao [34]	Neural network, rough set theory	Predicted dropouts from course
	2010, Knauf, Rainer, Yoshitaka Sakurai, Setsuo Tsuruta, and Kouhei Takada [35]	Decision tree	Analyzed successful Storyboard (e-learning system) success paths for students.
	2010, Youping, Bian Xiangjuan Gong [36]	Decision tree	Evaluated the high school students and studying effectiveness
	2010, Liu, Zhiwu, and Xiuzhi Zhang [37]	Decision tree	Built forecasting model for students' marks to identify negative learning habits or behaviors of students
	2009, Zhu, Li, Yanli Li, and Xiang Li [38]	Association rule	Predicted student's achievement systematically and improved teaching management
	2009, Nayak, Amar, Jitendra, Vinod, Shadab [39]	Proposed use of ontology, RDF, XML	Proposed enterprise framework to identify suitable semantic data related to students, faculties, and courses
	2009, Wang, Pei-ji, Lin Shi, Jin-niu Bai, Yu-lin Zhao [40]	Apriori algorithm	Improved algorithm used to mine the students' data table
	2009, Ramasubramanian, Iyakutti, and Thangavelu [41]	Rough set theory	Predicted weak students
	2008, Selmoune, Nazih et al. [42]	Association rules	Found the success and failure factors of students

References

1. Bresfelean, V.P., Bresfelean, M., Ghisoiu, N., Comes, C.-A.: Determining students' academic failure profile founded on data mining methods. In: Presented at the ITI 30th International Conference Information Technology Interfaces, pp. 317–322 (2008)
2. Madhulatha, T.S.: An overview on clustering methods. [Online]. Available: https://arxiv.org/abs/1205.1117 (2012)
3. Wook, M., Yahaya, Y.H., Wahab, N., Isa, M.R.M., Awang, N.F., Seong, H.Y.: Predicting NDUM student's academic performance using data mining techniques. In: Proceedings of 2nd International Conference Computer Electrical Engineering, vol. 2, pp. 357–361 (2009)
4. Parack, S., Zahid, Z., Merchant, F.: Application of data mining in educational databases for predicting academic trends and patterns. In: Proceeding of IEEE International Conference on Technology Enhanced Education (ICTEE), pp. 1–4 (2012)
5. Chi, C.-C., Kuo, C.-H., Lu, M.-Y., Tsao, N.-L.: Concept-based pages recommendation by using cluster algorithm. In: Presented at the 8th IEEE International Conference Advanced Learning Technology, Santander, Spain (2008)
6. Chen, C.-M., Li, C.-Y., Chan, T.-Y., Jong, B.-S., Lin, T.-W.: Diagnosis of students' online learning portfolios. In: Proceeding of 37th Annual Frontiers Education Conference-Global Engineering, Knowledge Borders, Opportunities Passports (FIE), pp. T3D-17–T3D-22 (2007)
7. Van To, T., Win, S.S.: Clustering approach to examination scheduling. In: Presented at the 3rd International Conference Advanced Computer Theory Engineering (ICACTE), pp. V5-228–V5-232 (2010)
8. Bouchet, F., Harley, J.M., Trevors, G.J., Azevedo, R.: Clustering and profiling students according to their interactions with an intelligent tutoring system fostering self-regulated learning. J. Educ. Data Mining **5**(1), 104–146 (2013)
9. Govindarajan, K., Somasundaram, T.S., Kumar, V. S., Kinshuk: Particle swarm optimization (PSO)-based clustering for improving the quality of learning using cloud computing. In: Proceeding of IEEE 13th International Conference Advanced Learning Technology (ICALT), pp. 495–497 (2013)
10. Hani, H., Hooshmand, H., Mirafzal, S.: Identifying the factors affecting the success and failure of e-learning students using cluster analysis. In: Presented at the 7th International Conference e-Commerce Developing Countries, Focus e-Security (ECDC), pp. 1–12 (2013)
11. Chen, J., Huang, K., Wang, F., Wang, H.: E-learning behavior analysis based on fuzzy clustering. In: Proceeding of 3rd International Conference on Genetic Evolutionary Computing (WGEC), Guilin, China, pp. 863–866 (2009)
12. Sagiroglu, S., Sinanc, D.: Big data: a review. In: Proceeding of International Conference Collaboration Technology Systems (CTS), pp. 42–47 (2013)
13. Govindarajan, K., Somasundaram, T.S., Kumar, V.S., Kinshuk: Particle swarm optimization (PSO)-based clustering for improving the quality of learning using cloud computing. In: Proceeding of IEEE 13th International Conference Advanced Learning Technology (ICALT), pp. 495–497 (2013)
14. Chen, H.-M., Cooper, M.D.: Using clustering techniques to detect usage patterns in a web-based information system. J. Amer. Soc. Inf. Sci. Technol. **52**(11), 888–904 (2001)
15. Shatnawi, S., Al-Rababah, K., Bani-Ismail, B.: Applying a novel clustering technique based on FP-tree to university timetabling problem: a case study. In: Presented at the International Conference on Comput. Eng. Syst. (ICCES), pp. 314–319 (2010)
16. Peña-Ayala, A.: Educational Data Mining Application and Trends. ISBN 978-3-319-02738-8, p. XVIII, 468, 139 illus
17. Romero, C., Ventura, S.: Educational data mining: a review of the state of the art. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **40**(6), 601–618 (2010)
18. Baker, R.S.J.D., Yacef, K.: The state of educational data mining in 2009: a review and future visions. J. Educ. Data Min. **1**(1), 3–16 (2009)
19. Romero, C., Ventura, S.: Educational data mining: a survey from 1995 to 2005. Expert Syst. Appl. **33**(1), 135–146 (2007)

20. Saranya, S., Ayyappan, R., Kumar, N.: Student progress analysis and educational institutional growth prognosis using data mining. Int. J. Eng. Sci. Res Technol. (2014)
21. Archer, E., Chetty, Y.B., Prinsloo, P.: Benchmarking the habits and behaviours of successful students: a case study of academic-business collaboration. Int. Rev. Res. Open Distrib. Learn. **15**(1), 62–83 (2014)
22. Ariouat, H., Cairns, A.H., Barkaoui, K., Akoka, J., Khelifa, N.: A two-step clustering approach for improving educational process model discovery. In: 2016 IEEE 25th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), Paris, pp. 38–43 (2016)
23. Rakesh Kumar, Arora, Dharmendra, Badal: Mining association rules to improve academic performance. Int. J. Comput. Sci. Mob. Comput. **3**(1), 428–433 (2014)
24. Osmanbegović, E., Suljić, M.: Data mining approach for predicting student performance. Econ. Rev. J. Econ. Bus. **X**(1) (2012)
25. Sukanya, M., Biruntha, S., Karthik, S., Kalaikumaran, T.: Data mining: performance improvement in education sector using classification and clustering algorithm. In: International Conference on Computing and Control Engineering (ICCCE 2012), 12 and 13 Apr 2012
26. Torenbeek, M., Jansen, E.P.W.A., Hofman, W.H.A.: Predicting first-year achievement by pedagogy and skill development in the first weeks at university. Teach. High. Educ. **16**(6), 655–668 (2011)
27. He, Y., Zhang, S.: Application of data mining on students' quality evaluation. In: 2011 3rd International Workshop on Intelligent Systems and Applications, Wuhan, pp. 1–4 (2011)
28. Sakurai, Y., Tsuruta, S., Knauf, R.: Success chances estimation of university curricula based on educational history, self-estimated intellectual traits and vocational ambitions. In: 2011 IEEE 11th International Conference on Advanced Learning Technologies, Athens, GA, pp. 476–478 (2011)
29. Sunita Aher, B., Lobo, L.M.R.J.: Article: combination of clustering, classification & association rule based approach for course recommender system in e-learning. Int. J. Comput. Appl. **39**(7), 8–15 (2012)
30. Ayesha, S., Mustafa, T., Sattar, A.R., Khan, M.I.: Data mining model for higher education system. Eur. J. Sci. Res. **43**(1), 24–29 (2010)
31. Kovačić, Z.: Early prediction of student success: mining students enrolment data. In: Proceedings of Informing Science and IT Education Conference (InSITE), pp. 647–665 (2010)
32. Al-shargabi, A.A., Nusari, A.N.: Discovering vital patterns from UST students data by applying data mining techniques. In: 2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE), Singapore, pp. 547–551 (2010)
33. Yan, Z., Shen, Q., Shao, B.: The analysis of student's grade based on Rough Sets. In: 2010 3rd IEEE International Conference on Ubi-Media Computing, Jinhua, pp. 345–349 (2010)
34. Ningning, G.: Proposing data warehouse and data mining in teaching management research. In: 2010 International Forum on Information Technology and Applications, Kunming, pp. 436–439 (2010)
35. Knauf, R., Sakurai, Y., Takada, K., Tsuruta, S.: Personalizing learning processes by data mining. In: 2010 10th IEEE International Conference on Advanced Learning Technologies, Sousse, pp. 488–492 (2010)
36. Xiangjuan, B., Youping, G.: The application of data mining technology in analysis of college student's performance. In: The 2nd International Conference on Information Science and Engineering, Hangzhou, China, pp. 5477–5480 (2010)
37. Liu, Z., Zhang, X.: Prediction and analysis for students' marks based on decision tree algorithm. In: 2010 Third International Conference on Intelligent Networks and Intelligent Systems, Shenyang, pp. 338–341 (2010)
38. Zhu, L., Li, Y., Li, X.: Research on early-warning model of students' academic records based on association rules. In: 2009 WRI World Congress on Computer Science and Information Engineering, Los Angeles, CA, pp. 121–125 (2009)
39. Nayak, A., Agarwal, J., Yadav, V.K., Pasha, S.: Enterprise architecture for semantic web mining in education. In: 2009 Second International Conference on Computer and Electrical Engineering, Dubai, pp. 23–26 (2009)

40. Wang, P., Shi, L., Bai, J., Zhao, Y.: Mining association rules based on apriori algorithm and application. In: IFCSTA'09 Proceedings of the 2009 International Forum on Computer Science-Technology and Applications, vol. 1, pp. 141–143, 25–27 Dec 2009
41. Ramasubramanian, P., Iyakutti, K., Thangavelu, P., Jeya, G.J., Shameera Begam, S.: Data mining techniques for teaching result analysis using rough set theory. In: 2008 International Conference on Computing, Communication and Networking, St. Thomas, VI, pp. 1–8 (2008)
42. Selmoune, N., Alimazighi, Z.: A decisional tool for quality improvement in higher education. In: 2008 3rd International Conference on Information and Communication Technologies: From Theory to Applications, Damascus, pp. 1–6 (2008)