Educational Data Mining: A state-of-the-art survey on tools and techniques used in EDM

Ginika Mahajan, Bhavna Saini
Manipal University, Jaipur.

Abstract

Over the last decade, limitless data is generating in the field of education. To process this huge raw data, enormous potential is required beyond the manual and time consuming tasks.With the span of time, analytics and data mining is used to extract useful information from large data repositories. Educational Data Mining (EDM) exploits statistical, machine learning and data mining in the domain of education to analyze and predict the educational data using various approaches. EDM seeks to use online modes of learning to better understand learners and their learning, and develop computational approaches to analyze the facts and figures so as to benefit learners. Various machine learning algorithms and research tools are used in Educational Data Mining for analysis and prediction on different types of educational data. This paper presents a survey of applications and tools used in Educational Data Mining. Also, it presents detailed review of current trends in EDM where techniques and results of recent work done in this field are compared.

1 Introduction

Educational Data Mining (EDM) is an evolving interdisciplinary research domain that deals with development of computational and statistical models to analyze and explore educational data [3]. The core objective of EDM is to analyze education-based different types of data to resolve research issues in the domain of education [12]. When applied in the domain of education, these approaches are referred to as learning analytics (LA) and educational data mining (EDM). These two research communities, educational data mining and learning analytics, have embraced similar perspectives on analysis of educational data. The only difference is that EDM researchers are more concerned in applying automated models to get information from educational data while researchers in learning analytics are more towards human-led methods for exploring educational data [4]. Practically EDM allows to determine new knowledge based on students learning data in order to endorse and evaluate educational system, and hence to improve some aspects of quality of education [10].

2 Background

2.1 Data Mining and its Applications

Over the last decade, limitless data is generating in every field. To process this huge raw data, enormous potential is required beyond the manual and time consuming tasks. Data mining or "Knowledge discovery in databases" are the processes that extract the useful or most relevant insights out of the large datasets. There are a number of applications where data mining has been applied to discover the hidden facts and to predict the future trends. Healthcare is one of the emerging field where huge data was collected and analyzed to understand the illness, symptoms and complications [23][24]. Various previous studies had helped in prediction and in taking right decision at early stage [25][26][27]. Agriculture is another major application area of data mining. Using previous agriculture data, an early detection can be done for the food quality, plant diseases, color of affected area, nutrient deficiency and many more factors [28][29]. Along with this, data mining is also proving helpful in understanding the agricultural supply chain and operations such as production, storage, retails, distribution, etc. [30][31] Other than these banking sector is utilizing the data mining methods in evaluating the customer satisfaction and their needs [33]. It also helps in generating credit rating and identifying credit card frauds [34].

Data Mining methods are also major role in e-commerce, to understand the diversity of market, to meet the increasing demand of customers, etc. [36]. It also helped a lot in improving customer relationship [37].

2.2 Educational Data Mining

In this era, teaching and learning platforms have completely change. Nowadays everything is online whether its related to teaching content, teacher, evaluation, tests and quizzes. There is an abundant diversityin educational systems and environments like E-learning, Learning Management System (LMS), Adaptive Hypermedia (AH) educational systems, tests or quizzes, contents. Various other platforms are available that aid in learning such as various social networks, forums, educational game-based learning environment, Learning repositories, virtual environments, ubiquitous computing environments, etc. [3].

Traditional classroom method which is considered as offline mode, various Psychometrics and statistical techniques have been applied to understand the student/learners behavior and performance in classroom. In online mode like learning management system (LMS) and E-learning, the EDM techniques are applied on student's data which is stored digitally in database. Extracting various features and factors from this data, EDM techniques can be applied to gain relevant information which can then be analyzed for further improvements.

EDM applies computational techniques for analysis and visualization of educational data. This analysis can be used to predict student's performance or student's strong and weak skills and knowledge. It can be used to detect undesirable student's behaviors and providing recommendations for students. These models can assist instructors in grouping students, getting feedback, developing course with proper planning and scheduling. This paper presents a review of most relevant and current studies done in EDM.

3 Surveys on EDM

Various researchers are working in the field of EDM to analyse and do prediction in the field of education. This paper presents recent survey of various tools and techniques used in EDM.

Hui Chun et. al. [1] worked to explore learning behaviors of students in blended learning courses. Dataset was collected from a university in northern Taiwan where two classes of Python programming related courses of first-year students were considered for experimentation. Experimentation values of f1-score of random forest model was evaluated as 0.83. This score was better as compared with decision tree and logistic regression. Also authors implemented machine learning and symmetry-based learning algorithms to explore student's learning behavior. Huei-TseHou et.al. [7] analyzed the videotaped learning process behaviors from 86 college students in simulation game-based learning activities. This study used an integrated technique of sequential analysis with cluster analysis to simulate learner behavioral patterns in games. Authors also categorized three clusters of learners with diverse pattern of learning processes. The results indicate that by using this integrated model of unsupervised learning one can explore the learners' reflective behavior pattern in simulation games. Castro et.al. [2] used Machine Learning algorithms on Learning Management System. The proposed work has applied Learning Management System in teaching and learning process of Bulacan State University (BulSU) Graduate School (GS) Program. Authors applied Support Vector Machine which is a supervised machine learning algorithm for classification and to identify best video lecture topic-wise.

Acharya et.al. [8] predicted students' performance using ML techniques - Naive bayes , C4.5, MLP (multi-layer perceptron), sequential minimal optimization (SMO), and KNN (1-Nearest Neighborhood).They considered four features and applied correlation-based feature selection (CBFS).

The results show that SMO attains effective average testing accuracy of 66% which is higher than other ML techniques used.

Paper	Author(s)	Year	Techniques	Results
[1]	Hui-Chun Hung, Fan Liu, Che-Tien Liang and Yu-Sheng Su	2020	Random forest, decision tree and logistic regression	Experimentation values of f1-score of random forest model was evaluated as 0.83. This score was better as compared with decision tree and logistic regression.
[7]	Huei-TseHou	2015	Integrated cluster and sequential analysis	Applying this integrated model explored three clusters of learners with different patterns of learning processes and that learners with advanced flow levels reach more complete reflective process.
[2]	Castro Mayleen Dorcas Bondoc and Tumibay Gilbert Malawit	2020	Support Vector Machine algorithm	Used SVM to classify and identify the best video topic-wise on LMS
[8]	Acharya and Sinha	2014	Naive bayes , C4.5, MLP (multi-layer perceptron), SMO (sequential minimal optimization) and KNN (1-Nearest Neighborhood)	SMO attains effective average testing accuracy of 66% which is higher than other ML techniques used.
[13]	Shuangyan Liu , Mathieu Aquin	2017	K-Prototypes clustering Algorithm	Using this algorithm, one can identify groups of students (successful and weak students group) based on demographic characteristics and interactions learning in online environment, and can examine the learning achievement of each group.
[5]	EkanshMaheshwari, Chandrima Roy, Manjusha Pandey, and SiddharthSwarupRautray	2018	K-means, Naïve Bayes and Random Forest	On applying naïve Bayes, a maximum accuracy of 82.35% and minimum accuracy of 57.35% is achieved. Using Random Forest, the maximum accuracy of 66.17% and minimum accuracy of 47.05% is achieved.
[9]	Elaf Abu Amrieh , ThairHamtini and Ibrahim Aljarah	2016	Artificial Neural Network, Naïve Bayesian and Decision tree ensemble methods - Bagging, Boosting and Random Forest (RF).	The ANN model outperformed other techniques, while Boosting was the best ensemble method.
[11]	Tsiakmaki,Kostopoulos, Koutsonikos, Pierrakeas, Kotsiantis, and Ragos	2018	LR, RF, 5NN, M5 Rules, M5, SMOreg, GP, Bagging	The obtained results show that RF, Bagging and SMOreg took precedence over other methods with MAE value ranging from 1.217 to 1.943.

Table 1. Survey of various techniques used by different authors and results obtained

Shuangyan et.al.[13] focused on performance of students in distance learning. forecasted how demographic variables and online learning activities affect student's performance using unsupervised learning algorithm. They used clustering algorithm, k- prototype, to know and gather information about two group of students- Successful students group and weak students group. Categorizing these groups may help faculty to focus on students with poor learning outcomes and who need special attention. Using this algorithm, one can identify groups of students (successful and weak students group) based on demographic characteristics and interactions learning in online environment, and can examine the learning achievement of each group. Ekansh et.al [5] gave a prediction model to know direct and indirect factors that are affecting the dropout rate of Primary to High School Students in India. They used K-means algorithm to find classes and factors, for male is the percentage schools having toilet and the number of male teachers, while for female it is the percentage of schools having girl's toilet and number of female teachers. For prediction naive Bayes classifier and Random forest

was applied on the dataset. On applying naïve Bayes, a maximum accuracy of 82.35% and minimum accuracy of 57.35% is achieved. Using Random Forest, the maximum accuracy of 66.17% and minimum accuracy of 47.05% is achieved. Elaf et.al. [9] proposed student's performance prediction model based on student's behavioral features. The obtained results forecasts a strong relationship between learner's behavior and their academic achievement. Artificial Neural Network, Naïve Bayesian and Decision tree classifiers are used to evaluate the student's performance in this model. Also to improve the performance of these classifiers, three ensemble methods are used- Bagging, Boosting and Random Forest (RF). The proposed model achieves an accuracy of 25.8% using ensemble methods. The ANN model outperformed other techniques, while Boosting was the best ensemble method. Tsiakmaki et.al. [11] compared eight supervised learning algorithms on Weka environment – Linear Regression, Random Forests (RF), 5-NN, M5 Rules, M5 algorithm, Sequential Minimal Optimization algorithm for regression problems using SVM (SMOreg), Gaussian processes (GP), Bootstrap Aggregating (Bagging) − for predicting students' marks. The evaluation metric used in this study to determine efficiency of regression methods is Mean Absolute Error (MAE). The obtained results show that RF, Bagging and SMOreg took precedence over other methods with MAE value ranging from 1.217 to 1.943.Table 1 provides a detailed survey of various techniques used by different authors and results obtained.

4 Tools used in EDM

Various machine learning algorithms and research tools are used in Educational Data Mining for analysis and prediction on different types of educational data. Although EDM is a rapidly emerging field and new tools and techniques are developing continuously, Table 2 provides a survey of various tools used in EDM for analysis in educational domain.

Tool	Description
RapidMiner[14]	A package for conducting data mining analyses and creating models.
KNIME [15]	Data cleaning and analysis package
Orange [16]	Orange is a data visualization and analysis package
SPSS [17]	A statistical package used for statistical tests, regression frameworks, correlations, and factor analyses.
KEEL [18]	KEEL used for analysis has classification and regression algorithms
The EDM Workbench [19]	tool for automated cleaning, organizing, and creating data with feature distillation and data labeling.
Spark MLLib [20]	A framework for large-scale processing of data acrossmultiple computer processors, in a distributed fashion.
D3js [21]	Data visualization tool used for complex data visualizations that require data handling
PSLC DataShop [22]	Integrates data collection, construction, analysis, and visualization.

Table 2. Survey of various tools used in EDM

Conclusion

Educational data mining is an emerging research area which is needed in educational domain. It exploits statistics, machine learning and data mining techniques to get various students and instructor parameters for analysis and predictions. From a practical standpoint, this survey can be proved valuable for researchers working in the domain of Educational Data Mining. This paper presents a survey of tools used in EDM. Also it presents review of current trends in EDM where techniques of related work done in this field are compared. It is hoped that this paper will play a role in raising the profile of the educational data mining field and research community.

References

[1] Hung, H. C., Liu, I. F., Liang, C. T., & Su, Y. S. (2020). Applying Educational Data Mining to Explore Students' Learning Patterns in the Flipped Learning Approach for Coding Education. Symmetry, 12(2), 213.

[2] Bondoc, C. M. D., &Malawit, T. G. (2020). Classifying relevant video tutorials for the school's learning management system using support vector machine algorithm. Global Journal of Engineering and Technology Advances, 2(3), 001-009.

[3] Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601-618.

[4] Baker, R. S., &Inventado, P. S. (2014). Educational data mining and learning analytics. In Learning analytics (pp. 61-75). Springer, New York, NY.

[5] Maheshwari, E., Roy, C., Pandey, M., &Rautray, S. S. (2020). Prediction of Factors Associated with the Dropout Rates of Primary to High School Students in India Using Data Mining Tools. In Frontiers in Intelligent Computing: Theory and Applications (pp. 242-251). Springer, Singapore.

[6] Tsiakmaki, M., Kostopoulos, G., Kotsiantis, S., &Ragos, O. (2020). Implementing AutoML in Educational Data Mining for Prediction Tasks. Applied Sciences, 10(1), 90.

[7] Hou, H. T. (2015). Integrating cluster and sequential analysis to explore learners' flow and behavioral patterns in a simulation game with situated-learning context for science courses: A video-based process exploration. Computers in human behavior, 48, 424-435.

[8] Acharya, A., & Sinha, D. (2014). Application of feature selection methods in educational data mining. International Journal of Computer Applications, 103(2).

[9]Amrieh, E. A., Hamtini, T., &Aljarah, I. (2016). Mining educational data to predict student's academic performance using ensemble methods. International Journal of Database Theory and Application, 9(8), 119-136.

[10] Romero, C., Ventura, S., & De Bra, P. (2004). Knowledge discovery with genetic programming for providing feedback to courseware authors. User Modeling and User-Adapted Interaction, 14(5), 425-464.

[11]Tsiakmaki, M., Kostopoulos, G., Koutsonikos, G., Pierrakeas, C., Kotsiantis, S., &Ragos, O. (2018, July). Predicting University Students' Grades Based on Previous Academic Achievements. In 2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA) (pp. 1-6). IEEE.

[12] Romero, C., Romero, J. R., & Ventura, S. (2014). A survey on pre-processing educational data. In Educational data mining (pp. 29-64). Springer, Cham.

[13] Liu, S., &d'Aquin, M. (2017, April). Unsupervised learning for understanding student achievement in a distance learning setting. In 2017 IEEE Global Engineering Education Conference (EDUCON) (pp. 1373-1377). IEEE.

[14] Hofmann, M., &Klinkenberg, R. (Eds.). (2016). RapidMiner: Data mining use cases and business analytics applications. CRC Press.

[15] Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., ... &Wiswedel, B. (2009). KNIME-the Konstanz information miner: version 2.0 and beyond. AcM SIGKDD explorations Newsletter, 11(1), 26-31.

[16] Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., ... &Štajdohar, M. (2013). Orange: data mining toolbox in Python. The Journal of Machine Learning Research, 14(1), 2349-2353.

[17] SPSS Inc. (2005). SPSS Base 14.0 user's guide. Prentice Hall.

[18] Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., & Herrera, F. (2011). Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. Journal of Multiple-Valued Logic & Soft Computing, 17.

[19] Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., ... & Xin, D. (2016). Mllib: Machine learning in apache spark. The Journal of Machine Learning Research, 17(1), 1235-1241.

[20] Rodrigo, M., Mercedes, T., d Baker, R. S., McLaren, B. M., Jayme, A., &Dy, T. T. (2012). Development of a Workbench to Address the Educational Data Mining Bottleneck. International Educational Data Mining Society.

[21] Zhu, N. Q. (2013). Data visualization with D3. js cookbook. Packt Publishing Ltd.

[22] Koedinger, K. R., Baker, R. S., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. Handbook of educational data mining, 43, 43-56.

[23] Sharma, Manik, Samriti Sharma, and Gurvinder Singh. "Performance Analysis of Statistical and Supervised Learning Techniques in Stock Data Mining." Data 3.4 (2018): 54.

[24] Kaur, P., Sharma, M., Mittal, M.: Big data and machine learning based secure healthcare framework. In: International Conference on Computational Intelligence and Data Science (ICCIDS 2018), Elsevier Procedia Computer Science, vol. 132, pp. 1049–1059 (2018).

[25] Ambigavathi, M., &Sridharan, D. (2020). A Survey on Big Data in Healthcare Applications. In Intelligent Communication, Control and Devices (pp. 755-763). Springer, Singapore.

[26] Sharma, Manik, and Prableen Kaur. "A Comprehensive Analysis of Nature-Inspired Meta-Heuristic Techniques for Feature Selection Problem." Archives of Computational Methods in Engineering (2020): 1-25.

[27] Pramanik, M.I., Lau, R.Y.K., Demirkan, H., Azad, M.A.K.: Smart health: big data enabled health paradigm within smart cities. J. Expert Syst. Appl. 87, 370–383 (2017).

[28] Sharma, R., Kamble, S. S., Gunasekaran, A., Kumar, V., & Kumar, A. (2020). A systematic literature review on machine learning applications for sustainable agriculture supply chain performance. Computers & Operations Research, 104926.

[29] Balducci, F., Impedovo, D. and Pirlo, G., 2018. Machine Learning Applications on Agricultural Datasets for Smart Farm Enhancement. Machines, 6(3), p.38.

[30] Kaur, Prableen, and Manik Sharma. "Analysis of data mining and soft computing techniques in prospecting diabetes disorder in human beings: a review." Int. J. Pharm. Sci. Res 9 (2018): 2700-2719.

[31] Borodin, V., Bourtembourg, J., Hnaien, F. and Labadie, N., 2016. Handling uncertainty in agricultural supply chain management: A state of the art. European Journal of Operational Research, 254(2), pp.348-359.

[32] Sharma, R., Kamble, S. S., Gunasekaran, A., Kumar, V., & Kumar, A. (2020). A systematic literature review on machine learning applications for sustainable agriculture supply chain performance. Computers & Operations Research, 104926.

[33] Dinçer, H., Yüksel, S., Canbolat, Z. N., &Pınarbaşı, F. (2020). Data Mining-Based Evaluating the Customer Satisfaction for the Mobile Applications: An Analysis on Turkish Banking Sector by Using IT2 Fuzzy DEMATEL. In Tools and Techniques for Implementing International E-Trading Tactics for Competitive Advantage (pp. 320-339). IGI Global.

[34] Sharma, M., G. Singh, and R. Singh. "Stark assessment of lifestyle based human disorders using data mining based learning techniques." IRBM 38.6 (2017): 305-324.

[35] Shahbazi, F. (2020). Using Decision Tree Classification Algorithm to Design and Construct the Credit Rating Model for Banking Customers. IOSR Journal of Business and Management, 21(3, Series 2), 24-28.

[36] Dixit, V. S., & Gupta, S. (2020). Personalized Recommender Agent for E-Commerce Products Based on Data Mining Techniques. In Intelligent Systems, Technologies and Applications (pp. 77-90). Springer, Singapore.

[37] Xu, D., Ruan, C., Korpeoglu, E., Kumar, S., &Achan, K. (2020, January). Product Knowledge Graph Embedding for E-commerce. In Proceedings of the 13th International Conference on Web Search and Data Mining (pp. 672-680).