International Journal of Engineering & Technology, 7 (2.6) (2018) 93-97

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Review Paper

EDM – survey of performance factors and algorithms applied

Deepali R Vora[1]*, Kamatchi Iyer[2]

[1] *Research Scholar, Amity University, Mumbai*
[2] *Department of Computer Science and Engineering, Amity University, Mumbai*
**Corresponding author E-mail: deepali_as@yahoo.com*

Abstract

Educational Data Mining (EDM) is a new field of research in the data mining and Knowledge Discovery in Databases (KDD) field. It mainly focuses in mining useful patterns and discovering useful knowledge from the educational information systems from schools, to colleges and universities. Analysing students' data and information to perform various tasks like classification of students, or to create decision trees or association rules, so as to make better decisions or to enhance student's performance is an interesting field of research. The paper presents a survey of various tasks performed in EDM and algorithms (methods) used for the same. The paper identifies the lacuna and challenges in Algorithms applied, Performance Factors considered and data used in EDM.

Keywords: *EDM, Algorithms, Performance Factors, Deep Learning*

1. Introduction

Data mining is a process to extract information from a data set and transform it into an understandable structure for further use. Educational Data Mining (EDM) relates to the inter-disciplinary research that deals with the development of various methods and techniques to explore the data generated from different educational sources. Analysing educational data could provide information of student's behaviours, based on which education policies would be made properly.[1] Educational Data Mining refers to techniques, tools, and research designed for automatically extracting meaning from large repositories of data generated by or related to people's learning activities in educational settings. Recent advances in educational technology, including the increase in computing power and the ability to log fine-grained data about student's use of a computer-based learning environment, have led to an increased interest in developing techniques for analysing the large amounts of data generated in educational settings.

This paper presents a study on current state of EDM and identifies the algorithms applied. Section 2 talks about the Goals and Methods in EDM, section 3 proposes purpose of study and presents the study in summarized fashion. Section 4 talks about role and study about performance factors. Section 5 presents the findings from the survey followed by conclusion.

2. Goals and Methods of Educational Data Mining

Major goals [2][3] of EDM are:

- Providing Feedback for Supporting Instructors

EDM can provide very input about the students' to the supporting instructors. This can help instructors modelling their way of teaching to suit every student.

- Detecting Student Behaviour

By mining the characteristics of students' it's possible to get important insights through mining characteristics of students'. The entire education can be modelled according to the students' behaviour.

- Predicting Student's Performance

This is the utmost important goal of the EDM. Identification of characteristics which plays a vital role in improving performance of student is important. EDM can be effective in identifying crucial factors which has effect on students' performance.

- Recommendations for Students

EDM can be used effectively to provide various education related recommendations to the students.

- Constructing Courseware

EDM can help teachers to better understand the students' need and prepare courseware accordingly. Online courses are mend according to students' ability and need.

- Planning and Scheduling

A course can be planned as per the characteristics of the students. EDM can help better understand the students and help to achieve the proper scheduling and planning of courses.

Major methods in EDM are as shown below:

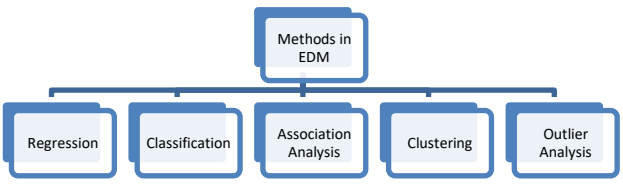

Fig 1: Methods in EDM

Regression can be used for prediction like predicting students' performance. Classification and Clustering is used to categorize the students and then tasks like courseware planning can be applied. Association Analysis is used to find association among students' and is effective in recommending educational courses, books, sites etc. to students. Outlier analysis is used effectively to find out the odd student out of the majority of students. This analysis is helpful in terms of behaviour analysis or finding drop out students etc.

3. Purpose of study

•Objective 1: Recent State of EDM and application area

The paper performs survey analysis of the work done in EDM, various tasks performed and listing various goals to be achieved.

•Objective 2: Algorithmic advances in EDM

The paper presents a summary of algorithms used in performing various EDM tasks; as well highlights the challenges faced.

•Objective3: Deep Learning is effective in EDM applications

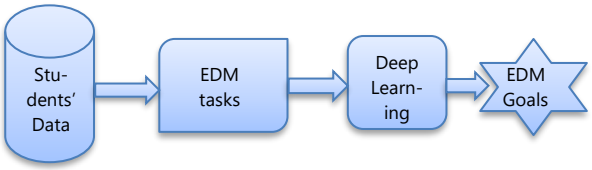

Fig. 2: Objective of the study

Deep Learning is the new field of machine learning applied in various areas like natural language processing, image and video processing etc. The paper reviews whether Deep Learning can be applied and will it be effective in EDM tasks to achieve the goals accurately.
Figure 2 depicts the objectives clearly.

4. Algorithm advances in performing EDM tasks

Three major tasks are frequently used in EDM as Classification, Clustering and Association rule mining. Following figure depicts the algorithms that can be used in performing these tasks.

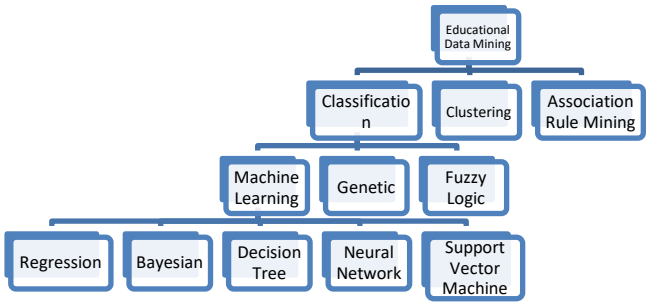

Fig. 3: EDM tasks and Algorithms Applied

Table 1 summarizes the Algorithms used in EDM and challenges faced.

5. Factors contributing to model the performance of students

For achieving the goals of EDM various factors related to students' are monitored. These factors play a vital role while predicting their behaviour or performance in Educational System. In a broad sense the factors can be classified as cognitive or non-cognitive factors. Cognitive factors refer to characteristics of the person that affect performance and learning. But non cognitive factors also play an important role in various EDM goals [22].

Xin Chen, et al [23] has studied social data to identify the factors which affects the behaviour or performance of students as study-life balance, lack of sleep, lack of social engagement, and lack of diversity. Michail N. Giannakos, et al[24] has identified various cognitive factors like academic performance, attendance etc. and its effect on students' performance. Xing, W., et al. [25] monitored closed classroom learning of students and identified the factors which affects the performance. The participation of student in various activities was majorly considered.

Syed Tahir Hijazi1 and S.M.M. Raza Naqvi[26] and Ali, Shoukat, et al.[27] has studied the impact of various factors like ; Attendance in class ,Family income, Study hours per day after college ,Mother's age, Mother's Education and gender, age, faculty of study, schooling, father/guardian social economic status, and residential area, medium of schooling; tuition trend, daily study hours ;on students' performance.

Irfan Mushtaq & Shabana Nawaz Khan[28] has proved that communication, learning facilities, proper guidance and family stress has direct impact on students' performance. As well Omar Augusto Echegaray-Calderon; Dennis Barrios-Aranibar[29] used many factors for study and identified which factors played a vital role in students' performance.

Angellia Debora Suryawan, Eryco Putra[30] has done a detail survey to identify the factors which affects students GPA. Also regression tests and correlation analysis are done on various factors. It proved that entrance exam and attendance in class are important factors. Lecturer quality is also important and has effect on GPA. Jae-Young Park, Heng Luo, Won Ho Kim[31] has studied factors like gender, academic performance of previous semesters, derailed enrolment, major related, credit, stop out years, age are considered. It is found that gender and age are not significant, previous academic performance is important.

In an interesting article Pooja Mondal [32] has identified various factors like intellect, learning, physical, mental, social and economic as factors which affect students' behaviour and performance.

From the survey of above papers it can be seen that cognitive and non-cognitive both factors are playing a role in identifying behaviour as well as predicting students' performance. So identification of factors to achieve goals of EDM is an important aspect.

6. Findings from the survey

Following observations are noted based on the survey:

a) Algorithm

Identification of right algorithm and methodology to perform EDM tasks to achieve the goals is important. From the figure 3 and table 1 it is clear that Deep Learning is still not given a thought in EDM. Usage of Deep Learning in various areas is surveyed in the paper [33] which strengthens our findings. Looking at the growing details getting captured in Educational Systems; use of Deep Learning will be beneficial to attain the goals of EDM

b) Performance factors

Factors; cognitive and non-cognitive; plays a role in determining behaviour and performance of students. Based on this, course planning can be achieved accurately. So analysis of educational data considering right combination of performance factors is very important. Though lot of work is done based on various factors, still non-cognitive factors are not given due weightage. So behaviour and performance analysis with consideration for non-cognitive factors is required.

c) Data under consideration

EDM work is done majorly for the students in high schools; for predicting there performance or behaviour and determining their career options. There is a need to analyse the data of Professional courses like Architecture, Engineering, and Medical etc. Now a day dropout rate of students in these courses is more. As well performance and behaviour analysis of these students may help Educationist to design the professional courses in a better way.

7. Conclusion

Educational Data Mining (EDM) is an interesting field of research for Educationist. Though a lot of research is going on in this field the main focus is always the School children. With the digitization trend lots of data about students is available to researchers for study. For the growth of any nation; producing good professionals is the key to success. So it is important to study data of students studying professional courses. Identifying the correct factors which affect their behaviour or performance is an important task. Thus analysing data of students' studying in professional courses and producing results related to their behaviour or performance is an important and interesting task.

References

[1] Karan Sukhija, Dr. Manish Jindal, Dr. Naveen Aggarwal, "The Recent State of Educational Data Mining: A Survey and Future Visions", IEEE 3rd International Conference on MOOCs,Innovation and Technology in Education, 2015.

[2] "What is educational data mining" , http://edtechreview.in/dictionary/394whatiseducationaldataminining, Accessed on August 2016.

[3] Ryan S.J.d. Baker,George Siemens ,"Educational Data Mining and Learning Analytics", Cambridge Handbook of the Learning Sciences, 2013

[4] Wattana Punlumjeak, Nachirat Rachburee, "A Comparative Study of Feature Selection Techniques for Classify Student Performance", 7th International Conference on Information Technology and Electrical Engineering (ICITEE), Chiang Mai, Thailand, 2015

[5] John Jacob, Kavya Jha, Paarth Kotak, Shubha Puthran , "Educational Data Mining Techniques and their Applications," IEEE International Conference on Green Computing and Internet of Things (ICGCloT), 2015

[6] Norlida Buniyamin, Usamah bin Mat, Pauziah Mohd Arshad, "Educational Data Mining for Prediction and Classification of Engineering Students Achievement," IEEE 7th International Conference on Engineering Education (ICEED),2015

[7] Camilo Ernesto López Guarín, Elizabeth León Guzmán, and Fabio A. González, "A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining," IEEE Journal of Latin-American Learning Technologies (IEEE-RITA), VOL. 10, NO. 3, AUGUST 2015

[8] Alana M. de Morais and Joseana M. F. R. Araújo, Evandro B. Costa, "Monitoring Student Performance Using Data Clustering and Predictive Modelling," IEEE, 2014

[9] Lu Thi Kim Phung, Vo Thi Ngoc Chau, Nguyen Hua Phung, "Extracting Rule RF in Educational Data Classification from a Random Forest to Interpretable Refined Rules," IEEE International Conference on Advanced Computing and Applications, 2015

[10] Asmaa Elbadrawy, Agoritsa Polyzou, Zhiyun Ren, Mackenzie Sweeney, George Karypis, Huzefa Rangwala, "Predicting Student Performance Using Personalized Analytics," IEEE , April 2016

[11] Anjana Pradeep, Smija Das, Jubilant J Kizhekkethottam, "Students Dropout Factor Prediction Using EDM Techniques," International Conference on Soft-Computing and Network Security (ICSNS -2015), Coimbatore, INDIA, Feb. 25 – 27, 2015

[12] Wanli Xing, Rui Guo, Eva Petakovic and Sean Goggins, "Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory", Computers in Human Behavior, vol. 47, pp. 168-181, June 2015.

[13] Harwati, Ardita Permata Alfiani and Febriana Ayu Wulandari, "Mapping Student's Performance Based on Data Mining Approach (A Case Study)", Agriculture and Agricultural Science Procedia, vol.3, pp. 173-177, 2015.

[14] Wen-Lung Shiau and Patrick Y.K. Chau, "Understanding behavioral intention to use a cloud computing classroom: A multiple model comparison approach", Information & Management, vol. 53, no.3, pp. 355-365, April 2016.

[15] Sadaf Ashtari and Ali Eydgahi, "Student perceptions of cloud applications effectiveness in higher education", Journal of Computational Science, January 2017.

[16] Fernando Koch, Marcos D. Assunção, Carlos Cardonha and Marco A.S. Netto, "Optimising resource costs of cloud computing for education", Future Generation Computer Systems, vol.55, pp. 473-479, February 2016.

[17] Humphrey M. Sabi, Faith-Michael E. Uzoka, Kehbuma Langmia and Felix N. Njeh, "Conceptualizing a model for adoption of cloud computing in education", International Journal of Information Management, vol.36, no. 2, pp.183-191, April 2016.

[18] Janice D. Gobert, Yoon Jeon Kim, Michael A. Sao Pedro, Michael Kennedy and Cameron G. Betts, "Using educational data mining to assess students' skills at designing and conducting experiments within a complex systems microworld", Thinking Skills and Creativity, vol. 18, pp. 81–90, December 2015.

[19] Wanli Xing, Rui Guo, Eva Petakovic and Sean Goggins, "Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory", Computers in Human Behavior, vol. 47, pp. 168-181, June 2015.

[20] Harwati, Ardita Permata Alfiani and Febriana Ayu Wulandari, "Mapping Student's Performance Based on Data Mining Approach (A Case Study)", Agriculture and Agricultural Science Procedia, vol.3, pp. 173-177, 2015.

[21] Pedro J. Munoz Merino, José A. Ruiperez-Valiente, Carlos Alario-Hoyos, Mar Pérez-Sanagustín and Carlos Delgado Kloos, "Precise Effectiveness Strategy for analyzing the effectiveness of students with educational resources and activities in MOOCs," Computers in Human Behavior, vol. 47, pp. 108-118, June 2015.

[22] "The important role of Non Cognitive Factors in School Performance," http://singteach.nie.edu.sg/issue25-hottopic/", Accessed on 20/08/2017

[23] Xin Chen, Mihaela Vorvoreanu, and Krishna Madhavan , "Mining Social Media Data for Understanding Students' Learning Experiences," IEEE transactions on Learning Technologies, 2013

[24] Michail N. Giannakos, Trond Aalberg, Monica Divitini, Letizia Jaccheri, Patrick Mikalef, Ilias O. Pappas, Guttorm Sindre , "Identifying Dropout Factors in Information Technology Education: A Case Study," IEEE Global Engineering Education Conference (EDUCON), 2017

[25] Xing Wanli, Guo Rui, Petakovic Eva , Goggins Sean , "Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analyt-

ics, educational data mining and theory," Elsevier- Computers in Human Behavior, 2014
[26] Syed Tahir Hijazi1 and S.M.M. Raza Naqvi , "Factors Affecting Students' Performance," Bangladesh e-Journal of Sociology. Volume 3. Number 1,2006
[27] Ali, Shoukat, et al , "Factors Contributing to the Students Academic Performance: A Case Study of Islamia University Sub-Campus," American Journal of Educational Research 1.8 (2013): 283-289, 2013
[28] Irfan Mushtaq & Shabana Nawaz Khan , "Factors Affecting Students' Academic Performance", Global Journal of Management and Business Research Volume 12 Issue 9 Version 1.0 June 2012
[29] Omar Augusto Echegaray-Calderon; Dennis Barrios-Aranibar , "Optimal selection of factors using Genetic Algorithms and Neural Networks for the prediction of students' academic performance ", IEEE- Latin America Congress on Computational Intelligence (LA-CCI), 2015
[30] Angellia Debora Suryawan, Eryco Putra , "Analysis of Determining Factors for Successful Student's GPA Achievement", 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS), Yogyakarta, Indonesia, 2016
[31] Jae-Young Park, Heng Luo, Won Ho Kim "Factors affecting students' completion: A study of an online Master's program", International Conference of Educational Innovation through Technology, 2015
[32] "7 Important Factors that May Affect the Learning Process", http://www.yourarticlelibrary.com/learning/7-important-factors-that-may-affect-the-learning-process/6064/, Accessed on 16/06/2017
[33] Deepali Vora, Dr. Kamatchi Iyer , "A Survey of Inferences from Deep Learning Algorithms ", 7th International Conference on Computing, Engineering and Information Technology (ICCEIT 2017).

Table 1: Algorithms used in EDM, features and challenges

Authors	**Methodology/ Algorithm**	**Features**	**Challenges**
Wattana Punlumjeak,Nachirat Rachburee [4]	Feature Selection • Genetic Algorithm • SVM • Information Gain • Minimum redundancy and maximum relevance Classification • KNN • Naïve Bays • Decision Tree • Neural Network	• Minimum redundancy and maximum relevance with KNN gave more accuracy • Limited features and mostly Students' academic performance is only considered	• Parameters used are very few • More parameters like personal details of students can be considered.
John Jacob, Kavya Jha, Paarth Kotak,Shubha Puthran [5]	• K Means • Regression Analysis	• Different parameters are used for K means and Regression analysis. • GPA grades of the students is predicted • Simple algorithms are used	• Selection of Parameters
Norlida Buniyamin, Usamah bin Mat, Pauziah Mohd Arshad[6]	• Neuro fuzzy algorithm	• Six input and one output is defined • The computation of parameters will be facilitated via gradient vector	• Fuzzy rules requires linguistic variables and labels
Camilo Ernesto López Guarín, Elizabeth León Guzmán, and Fabio A. González[7]	• naïve Bayes • decision tree classifier	• Decision tree classifier was consistent • Bayesian classifier was Computationally faster • Data collected was limited	• Accuracy may decrease with increase in data
Alana M. de Morais and Joseana M. F. R. Araújo, Evandro B. Costa[8]	• Ward algorithm	• E learning data is considered • Teaching Learning methods were the focus • Clustering was used	• Pre-processing was specific to algorithm • Identification of number of clusters is difficult
Lu Thi Kim Phung, Vo Thi Ngoc Chau, Nguyen Hua Phung[9]	• Rule extraction algorithm	• follows a greedy approach with two phases: rule refinement and rule extraction • Handles both discrete and continuous attributes in the educational data sets	• Algorithm is complex • Compacting the rules is main challenge
Asmaa Elbadrawy et al.[10]	• Regression	• Students performance was estimated with multiple linear regression method • Varying regression models were used	• Only linear analysis is indicated
Anjana Pradeep, Smija Das, Jubilant J Kizhekkethottam [11]	• Induction rules • Decision tree	• High dimensionality was reduced by using WEKA feature selection algorithm	• Imbalanced data affecting accuracy • Students' social, economic data was not used

Authors	Methodology/ Algorithm	Features	Challenges
Wanli et al. [12]	• Genetic programming	• Interpretable algorithm • Produced an optimized prediction rate	• Lesser consideration of the qualitative aspects • Difficult time for replicating the results in a different context
Harwati et al. [13]	• K-mean Cluster algorithm	• Better classification • Computationally faster than hierarchical clustering	• Difficult to predict K-value • Does not work well with global cluster
Wen and Patrick [14]	• Statistical modelling	• Exhibited adequate explanatory power • Comprehensive understanding of the factors	• Combine direct and indirect evidence, lead to uncertainty • Difficulty in handling lead time bias
Sadaf et al. [15]	• Statistical modelling	• Higher education settings • Maximum self-efficacy	• Needs thorough validation of scale • Cannot support the changes in size and population
Fernando et al. [16]	• Maximum likelihood estimation	• Provides cost reduction • Attain maximum error cancellation	• The mathematics is often non-trivial, particularly if confidence intervals for the parameters are desired • It is sensitive to the choice of starting values.
Humphrey et al. [17]	• Partial least square	• Provide proper decision making • Empirical data back valuable information	• Difficulty in interpreting loadings of independent latent variables • Lack of model test statistics
Janice et al. [18]	• Decision tree	• High potential power • Broad scalability in multiple domains	• Tree structure prone to sampling errors • Tree splitting is locally greedy
Wanli et al. [19]	• Genetic programming	• Interpretable algorithm • Produced an optimized prediction rate	• Lesser consideration of the qualitative aspects • Difficult time for replicating the results in a different context
Harwati et al. [20]	• K-mean Cluster algorithm	• Better classification • Computationally faster than hierarchical clustering	• Difficult to predict K-value • Does not work well with global cluster
Pedro et al. [21]	• Precise effectiveness strategy	• Extracted relevant high-level information	• Failed to evaluate the usefulness of these visualizations with teachers and other stakeholders • Complexity in using more than two or three variables