International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 12, Number 2 (2016), pp. 227-237

http://www.ripublication.com

Education Data Mining: An Overview

Balwinder Kaur
DCSA, PU, Chandigarh-160014, India
Email-neeru.saini1@gmail.com

Abstract

Abstract- Educational Data Mining (EDM) is a integrative research area dealing with building new techniques to examine and analyse the data coming from educational settings to acquire valuable knowledge. The information acquired can help not only to enhance teaching process, learning process as well as management related process of academic institutes at various levels. The paper poses an attempt to study EDM, different entities involved, and EDM components are discussed briefly. The paper also lists some of the EDM techniques. The paper presents an understanding into the procedure of knowledge discovery in EDM and an experimental study on classification using decision tree by implementing frequently used DT algorithms like J48, Random forest, Reptree and NB tree. A comparison is performed on the basis of accuracy of algorithms. It is found that J48 and Reptree performed quiet well on the given dataset. The study also presents challenges/ issues related to EDM which suggest new scope for research work to be taken up in future for better outcomes in EDM.

Keywords: Educational Data Mining (EDM), EDM process, Classification, Decision Tree.

1. INTRODUCTION

Data mining (DM) in education is an area which brings into usage machine learning, statistical techniques and data-mining algorithms on various types of educational data. The objective behind is to understand and analyse the data to discover valuable knowledge and resolve academic research issues [1]. It is mainly concerned with building methods or techniques to explore the data coming from educational environment, which helps to not only understand the students but also the setting in which they study [2]. There has been tremendous increase in the different types of data available related to student's as there is an increase in web-based education as well as academic institutes are also storing lot on information related to students, courses etc [3]. EDM uses these repositories of data to understand learners and their learning patterns in a better way, and to design and develop different approaches for computation that can merge the data and theory for the benefit of the student's as well as institutes. These days, there exist diverse educational environments namely; classroom, "Learning

Management Systems" (LMS), "E-learning", and web-based, etc, the data coming from these environments are different hence different problems arise that need to be resolved using DM methods [3].

A few EDM objectives on the basis of perspectives of different researchers are [4,5]-

1. To create student models based on their behaviour, performance, surroundings/ environment, and learning style.
2. To develop the system for studying the effects of pedagogical support.
3. To study the effects of resources related to institutional infrastructure, human resource, and Industry-academic relationship in the organization.

The rest of the paper is organized as follows: section2: related work, section3: components of EDM, section4: DM Process, Section5: Experimental Study, Section6: Discussion and Section 7 Conclusion.

2. RELATED WORK

Han and Kamber [6] describes DM as a process that allow users to analyse data from distinct or separate perspectives. It also allows to create summaries, categorize as well as find relationships with the data. Romero, C., Ventura, S. [7] in their review paper present that EDM is being used as a prediction tool and helps in various decision-making situations in the education sector. R. Jindal et al.,[8] presents an elaborated study on EDM, they present a detailed comparison of various tools available for EDM. B.K. Baradwaj and S. Pal [9], describes that to provide quality education is the fundamental objective of academic institutions. They applied classification task to evaluate student's performance and out of different algorithms, the decision tree algorithm is used. Surjeet Kumar Yadav et al.,[10] studied and used decision tree classifiers. they conducted experiments to discover the desirable classifier that predict the student's future drop-out possibility with optimum accuracy. B. K. Baradwaj et al., [11] applied classification on student dataset to predict the student's grades based on previous results. Amjad Abu Saa [12], explores multiple elements that may influence the students' performance at higher education level, and tries to discover a model that is able to classify and predict the performance of students on the basis of social as well as personal factors. Out of CART, CHAID and ID3, CART had the best accuracy. Kumar, S. Anupama, and M. N. Vijayalakshmi in their study [13] used a dataset of 60 students and applied Rule based classification techniques and it is found that Rule based algorithm can be efficiently used to predict students' performance as compared to other classification techniques. Dorina Kabakchieva [14], applied rule learners: "OneR" and "JRip", decision tree classifier: C4.5, two popular Bayes classifiers: NaiveBayes and BayesNet and a Nearest Neighbour classifier on a student dataset for predicting the performance of students' based on their personal as well as pre-university features, at university level. The results showed that the classifiers perform individually for 5 different classes. Ramaswamy N, [15] performed classification using Neural Net, Decision Tree (DT), and Bayesian Net, Naïve Bayes (NB). The paper classifies students' as a slow learner or fast Lerner and compared the accuracy of Data mining techniques. Srecˇko Natek , and Moti Zwilling [16], studied the applicability of DM on small student data set to predict students' final grade . Different DT algorithms- J48,

Reptree, and M5P were implemented to predict the grades of the students. Out of all the implemented algorithms J48 showed highest accuracy of 98%.

3. COMPONENTS AND ENTITIES INVOLVED IN EDM

The entire EDM presented in figure 1, can be perceived as a working system made up of separate yet linkable entities and components namely; Users/Stakeholders, environment, Educational data, Educational tasks, and DM methods [17].

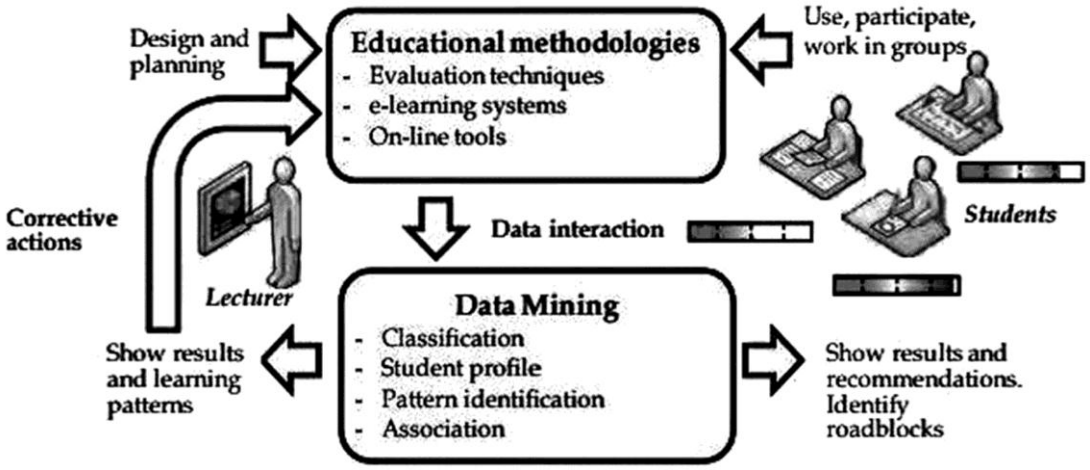

Figure 1: working of an Educational data mining system [18]

Stake holders/User

Starting from primary to secondary, and higher education different users and stakeholders of education they look at the data with different perspective. The stakeholders and their objectives are described in table1:

Table 1: User and their objectives [19]

User/ Stake Holder	Objective
Student	To support a learner's by providing feedbacks or recommendations, to enhance academic performance, etc.
Teachers	To understand learning processes of students' and enhance teaching methods, and other related objectives.
Researchers	To develop and differentiate DM techniques. To recommend the beneficial one for specific academic issue, to measure the effectiveness of learning while using separate methods and settings and methods, etc.
Administrators	To access the most beneficial way to arrange and distribute the institutional resources (human and material) and so on

Environment

It is categorized into two categories-Informal and Formal. Informal-in this indirect interaction take place for e.g. e-learning. Formal- the conventional class room setting, where one on one interaction takes place [17, 20].

Educational Data

Academic systems produce tremendous amount of data related to students, results, and courses etc. which can be used for analysing the students, improving teaching as well as learning process and decision-making process of administrators. The data is generated from numerous sources can be grouped into following groups [3,8]:

Offline Data: it is generated from different interactive environment, classroom interaction (traditional and modern), learners' information, educator's information, student's attendance, various course information's so on [8].

Online Data: This data is produced from the spatially distinct users of the education like web-based education, online group forum and social networking site etc [8].

Educational tasks

These are continuous processes which help students to address different problems. It helps to achieve administrative as well as academic objectives. Divide into 2 categories [8]:

Decision oriented tasks: are associated with administrative and academic decision making.

Learner oriented tasks: It involves active participation of learner and educator to enhance learning as well as to fulfil academic related objectives.

EDM Methods

There are various frequently used EDM methods [3]. These methods come from various sources; data mining, analytics, statics and so on. Romero and Ventra in 2007 and later Baker defined topology and divided EDM methods into following categories [4,7,20,21] **:**

- **Prediction**

"Prediction" builds a model can find data value of a feature known as predicted variable on the basis of some other features called predictor variables [6]. Prediction are categorised into "Classification", "Regression", and "Density estimation" [20]. In EDM, prediction is used to predict student academic performance etc [19].

- **Clustering**

It is "unsupervised learning". It is to use to divide the instances into group based on similarity. The groups thus created are called clusters. Once the clusters are created, new instances are inserted into the clusters having similar instances [6].

- **Outlier Detection**

The objective of "outlier detection" is to search for the instances that don't match with rest of the instance data [6, 17].

- **Relationship Mining**

"Relationship mining" is to search linkage among features of a data set and to convert them into rules for future use [6]. "Association rule", "Correlation mining", "Sequential pattern mining", and "Causal data mining" [2, 21] are Relationship mining techniques which are used frequently.

- **Discovery with Model**

Under this, a previously created model using techniques like classification, clustering, and so forth is used to develop another model or is used in analysis with techniques like relationship mining [4].

- **Distillation of Data for Human Judgment**

"Distillation of Data for Human Judgment" presents results to the user using techniques like visualization [21].

4. DM PROCESS

DM refers to "extracting" or "mining" new and valuable knowledge from big data repositories [6]. DM methods and techniques are utilized to work upon these repositories for discovering underlying patterns as well as associations and linkages useful for decision makers [9]. The phases or stages of DM process for discovering underlying valuable knowledge are presented in Figure 2.

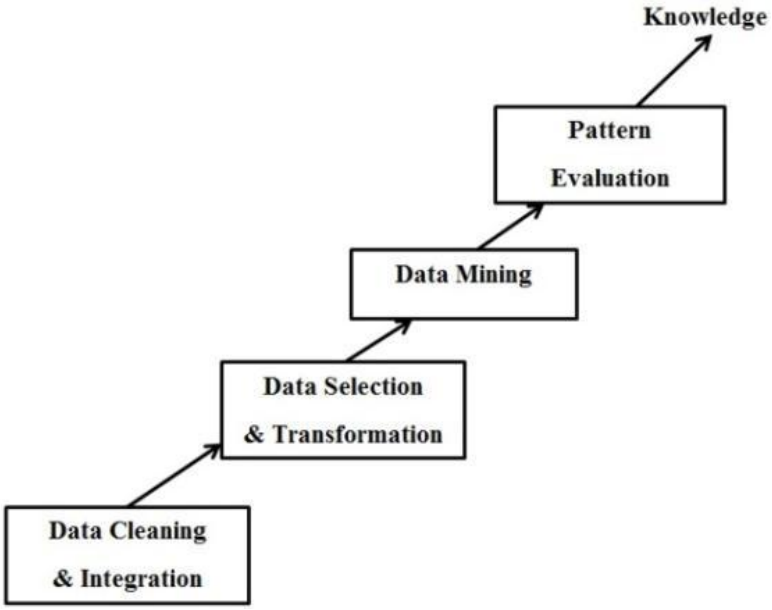

Figure 2: Stages of DM process [6]

Figure 3 show the process followed in EDM for knowledge discovery which is almost similar to DM process implemented in various other applications such as medicine, genetics etc. As in standard DM process, similar steps are executed in EDM process which are divided into 3 categories [16]:

• **Pre-processing**: in this phase the data is transformed into format as required for DM algorithms. It involves following steps like [6, 16]:

a. Data cleaning is the first task after collecting data, it deals with deleting irrelevant data from item sets that are not required for mining [6].
b. Data transformation consist of deriving new attributes from already existing attributes, converting data types like changing numerical values to nominal etc.
c. Data integration is concerned with integrating and synchronizing data from various sources which can be heterogeneous in nature [6].
d. Data reduction is concerned with reducing the dimensions of data [6].
e. Attribute selection allows to select the attributes which are relevant and required.

• **Data Mining**: it is the main step in the entire process. Appropriate DM techniques-clustering, classification, association rule mining etc. is applied after pre-processing data. The aforementioned techniques are implemented using DM tools [16].

• **Post – Processing**: the final phase where testing of model is performed and interpretation and analysis of the results is performed.

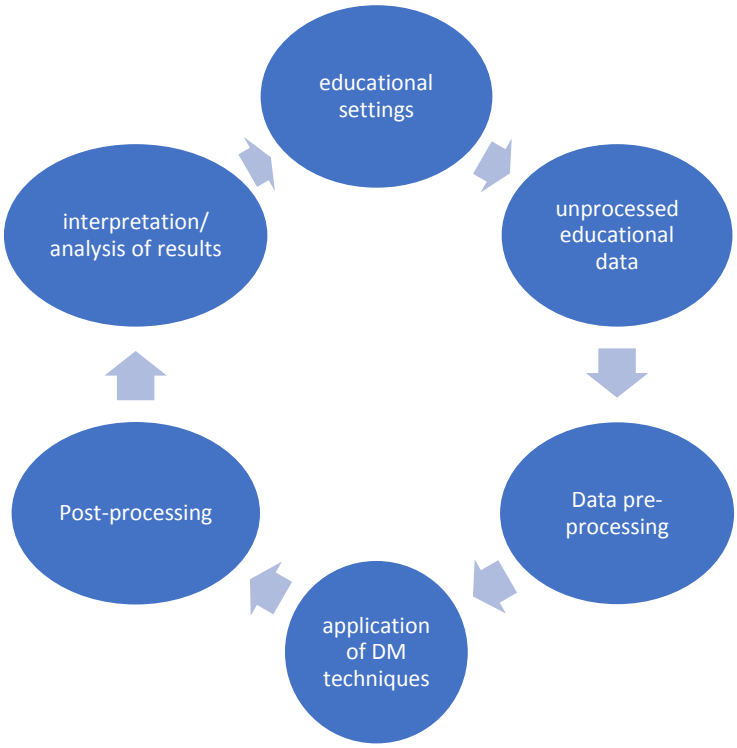

Figure 3: Stages EDM process [16]

5. EXPERIMENTAL STUDY

A small experimental study is performed to get a deeper understanding of use of classification technique for student academic performance prediction on the basis of the aforementioned study.

Data Set used: A student data set of 105 instances containing academic and demographic attributes is collected from a department of an institute.

Data Pre-processing: in this stage data has been cleaned and only relevant fields are selected. The data description is shown in table 2. The predictor variables as well as the response variables have been derived from the dataset.

Table 2: Data Description of Data set

S. No	Attribute	Description
1.	Study year	Year of Study
2.	Stud #	Student Number
3.	Gender	Male, Female, other
4.	DOB	Date of Birth
5.	Employment	Whether student is employed or not
6.	Reg Stat	Registration Status-old or new registration
7.	Study type	Study type can be full time or part time
8.	Exam cond.	Exam condition (yes / No)
9.	Act. Part	Activity participation (marks in points out of 50)
10.	Exam pts	Exam Points – marks scores in exam out of 50
11.	Final pts.	Final Points 100 (activity + exam)
12.	Final grade	Final grade out of 10
13.	grade	Grade is the response variable divided into 4 classes poor, low, medium and high.

Classification – Data Mining Technique

Classification: a "Supervised learning technique" [6]. In this technique the class labels are pre-defined/ known on the basis of the data-set instances. Classification process is executed in 2 stages- the training stage and testing stage. During training stage the model built is trained using the training set to perform classification, as class label are known for the training set. Once the training is performed the model is tested and validated using the test data -set which has similar type of instances without class label [23]. The validation can be performed using k-cross fold validation. In this the entire data-set is partitioned into k parts, and using k-1 sub set for training and 1 for testing in

round robin fashion iteratively. There are number of algorithms like- Naïve Bayes (NB), Rules, Functions, and Decision Tree (DT) etc. used for classification. Out of all these Decision Tree is used frequently by the researchers.

Decision Tree

DT - "flow-chart" like tree structure, all internal nodes are the test nodes that can split into 2 or more child nodes. Test nodes, test the value of an attribute related function or expression [6]. The paths between the internal nodes and child nodes are known as arcs which are labelled with test outcome and each leaf node is labelled as class label [10].

Experiment Results

The data set used for performing the experimental study has been obtained from one of the departments of an educational institutions. Initially size of the data is 108 records.

Experimentation has been performed using WEKA – an open and free DM software [23]. Only DT algorithms; J48, REPTree, NBTree, and Random Tree were implemented as they are easy to study and interpret for the initial understanding. All the algorithms were implemented using the default setting and validated using 10 – cross fold validation [23]. The response variable or class variable is grade which is divided into 4 classes- High, Medium, Low and Poor. The comparison of the results is presented in the table 3 below:

Table 3: Comparative Results of Decision Tree algorithms

Tree	Correctly classified	Incorrectly classified	Accuracy
J48	108	0	100%
REPTree	108	0	100%
NBTree	106	2	98.14%
Random Forest	101	7	93.51%

The comparison is performed on the basis of accuracy, correctly classified and incorrectly classified instances. Comparative results show the highest prediction accuracy of J48 and Reptree followed by Random forest and NB tree which are also performing quite well on the given data set.

6. DISCUSSION AND FUTURE SCOPE

EDM is associated with various fields -e- learning, ITS, data mining, statics and so on. In an academic environment's, there are different stake-holders and user who interpret the data and corresponding results based on their respective objectives. There are many factors that affect the educational environment which has led to different challenges associated the EDM field. These challenges are needed to be resolved; hence more work

is required to done in EDM. A few of these challenges are discussed below:

- There is need user friendly and convenient tools for both non-expert users and educators.
- The tools must be flexible, powerful as well as intuitive with a user-friendly interface and with visualization [3, 7].
- There is a need of standardization of data as well as pre-processing and post-processing stages [3, 7].
- The database availability and size of data set available.
- Apart from size, another issue that needs to be addressed is the versatility of dataset.
- There is a need to upgrade academic specific techniques taking into consideration instructional design as well as pedagogical decisions [3, 7].
- There is a strong requirement for securing the privacy of learner.
- Incremental nature of educational data is a big challenge which leads to issues related to maintenance of data [8].

EDM needs to be applied practically by the educators and decision makers not only to enhance the performance of individual learner but of educational institute as whole. EDM tools are to be merged and integrated into educational systems at all levels. DM tools must be able to facilitate all the stage of EDM process. Aforementioned information obtained from the new-fangled EDM comprehensively recognizes the necessity for further research.

7. CONCLUSION

The study presented is a step towards understanding various aspects and components of EDM. The paper first acquaint with the EDM concept, explains the components along with the process. It also presents an outline of the present state of EDM. The paper presents a small experimental study, where decision tree algorithms-J48, NBTree, REPTree and Random Tree of classification technique are implemented on a dataset of almost 100 instances. Out of all the implemented algorithms J48 and REPTree showed 100% accuracy. EDM is a phenomenon that has capability to change the scenario of current education completely. It has advanced over the time but for maturing more exploration is still required. The discussion on challenges and issues tries to bring forward the future scope where improvements can be done.

REFERENCES

[1] T. Barnes, M. Desmarais, C. Romero, and S. Ventura, In 2nd International Conference on Educational Data Mining, Cordoba, Spain, 2009.

[2] R. Baker, "Data mining for education," Int. Encyclopaedia of Education, B.

McGaw, P. Peterson, and E. Baker, Eds., 3rd edition. Oxford, U.K. 2010 (Elsevier).

[3] C. Romero, "Educational Data Mining: A Review of the State of the Art", In IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, Vol. 40, I. No. 6, Nov 2010

[4] Baker, R. S. J. D.," Data Mining for Education", Int. Encyclopaedia of Education, Elsevier, UK, 3rd ed., ed. by B. McGaw, P. Peterson, and E. Baker. Oxford, 2011.

[5] Baker, R. S. J. D., and Yacef, K., "The State of Educational Data Mining in 2009: A Review and Future Visions", Journal of Educational Data Mining, Volume. 1, no.1, pages. 3–17, 2009.

[6] Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques, 2nd edition, 2006.

[7] C. Romero, S. Ventura, "Educational data mining: A survey from 1995 to 2005", Expert Systems with Applications, Volume- 33, pp. 135–146, 2007.

[8] Jindal R., and Dutta B.M., "A Survey on Educational Data Mining and Research Trends", Int. Journal of Database Management Systems, Volume. 5, no.-3, pages. 53-73, 2013.

[9] Brijesh Kumar Baradwaj, Saurabh Pal, Data mining: machine learning, statistics, and databases, 1996.

[10] S.K. Yadav, B. Bharadwaj, and S. Pal, "Mining Education Data to Predict Student's Retention: A comparative Study", 2012.

[11] Brijesh Kumar Baradwaj, Saurabh Pal, Mining Educational Data to Analyze Students‟ Performance, 2011.

[12] Saa, Amjad Abu., "Educational data mining & students' performance prediction.", International Journal of Advanced Computer Science and Applications, Vol. 7, Issue no. 5, pages: 212-220

[13] Kumar S. Anupama, and M. N. Vijayalakshmi, "Mining of Student Academic Evaluation Records in Higher Education.", International Conference on Recent Advances in Computing and Software Systems, IEEE, pages: 67-70, 2012.

[14] Dorina Kabakchieva, "Predicting Student Performance by Using Data Mining Methods for Classification.", Cybernetics and Information Technologies, Volume. 13, Issue no. 1, pages: 61-72, 2013.

[15] Ramaswami, M., and R. Rathinasabapathy, "Student performance prediction.", International Journal of Computational Intelligence and Informatics, Vol. 1, Issue no. 4, 2012.

[16] Srecˇko Natek, Moti Zwilling, "Student Data Mining Solution– Knowledge Management System Related to Higher Education Institutions", Expert Systems with Applications, Volume 41, Issue 14, 15 October 2014, Pages 6400-6407, (SCI- Extended), Elsevier

[17] C. Romero, and S. Ventura," Data mining in education", Data Mining and Knowledge Discovery, Wiley Interdisciplinary Reviews, Volume. -3, Issue-1, pages. 12-27, 2013

[18] R. S. J. D. Baker and K. Yacef., "The State of Educational Data Mining in 2009: A Review and Future Visions", Journal of Educational Data Mining, Vol. 1, Issue.1, pages. 3–17.

[19] Cristobal Romero and Sebastian Ventura, "Data mining in education", WIREs Data Mining and Knowledge Discovery Vol. 3, pages: 12–27, 2013.

[20] D Baker, and S.J. Ryan, "Mining data for student models.", In Advances in Intelligent Tutoring Systems, Springer, Berlin, Heidelberg, pages. 323-337, 2010.

[21] S.J Ryan, and D. Baker, "Learning analytics and educational data mining", In: Proceedings 2nd International Conference on Learning Analytics and Knowledge - LAK '12, 2012.

[22] R. Llorente, and M. Morant. "Data mining in higher education", In New fundamental technologies in data mining, InTech, 2011.

[23] I. Witten, E. Frank, L. Trigg, M. Hall, and G. Holmes, "Weka: Practical machine learning tools and techniques with Java implementations," 1999.