International Journal of Advanced Technology in Engineering and Science
Vol. No.3, Issue 11, November 2015
www.ijates.com
ijates
ISSN 2348 - 7550

REVIEW ON EDUCATIONAL DATA MINING TECHNIQUES

Geeta Kashyap[1], Ekta Chauhan[2]

[1]Student of Masters of Technology, [2]Assistant Professor,

Department of Computer Science and Engineering, AP Goyal Shimla University, (India)

ABSTRACT

The educational data mining is emerging field that focuses on analyzing educational data to develop models for improving learning experiences and improving institutional effectiveness. Increasing interest in data mining and educational systems, make educational data mining as a new growing research community. Educational Data Mining means to extract the hidden knowledge from large educational databases with the use of techniques and tools. Educational Data Mining develops new methods to discover knowledge from educational database and it is used for decision making in educational system. In this paper we focus on comparative analysis of various educational data mining techniques with their algorithms. We compare the accuracy of these techniques with their algorithms on weka tool. The compared techniques and algorithms are presented together with some experimental data that give rise to the final conclusion.

Keyword: Association Rule, Classification, Clustering, Educational Data Mining (EDM), KDED (Knowledge Discovery In Educational Databases), Web Mining, Weka.

I. INTRODUCTION

Data Mining (DM)(knowledge Discovery in databases) is the process of extraction of interesting(non-trivial, implicit, previously unknown and potentially useful) pattern or information from large databases using various data mining techniques such as classification, clustering, association rule etc which helps in various decision making[12].

Education Data Mining (EDM) is the application of data mining related to educational data and Educational Data Mining is a learning analytics and quantitative observation method in order to understand how student respond to educational system and their responses impact their learning. Its objective is to analyze educational data in order to resolve educational research issues. In recent years there is rapid growth in education sector which leads to growing of education data so mining of education data become important to understand student behavior during learning process or to understand student problems[2].

Traditionally, educational researchers have been using methods such as surveys, interviews, focus groups, and classroom activities 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. The scale of such studies is also usually limited [1]. The Education Data Mining techniques overcome problems which are faced seldom. The emerging fields of learning analytics and Educational Data Mining (EDM) have focused on analyzing structured

data obtained from course management systems (CMS), classroom technology usage, or Controlled online learning environments to inform educational decision-making [1].

Our research study analyzing both the structured data i.e. the data that comes from course management system (CMS) or the data that comes from controlled online learning environment and the social web data (twitter, you tube) that comes from uncontrolled environment to understand student learning experiences (i.e. to better understand their problems, performance, behavior during learning process) with the use of educational data mining techniques and tools. After that the comparative analysis of different techniques with their algorithms is on weka tool.

1.1 Educational Data

Decision-making in the field of academic planning involves extensive analysis of huge volumes of educational data. Data are generated from heterogeneous sources like diverse and distributed, structured and unstructured data. These data are mostly generated from the offline or online sources:

Offline Data. Offline Data are generated from traditional and modern classroom, Interactive teaching/learning environments, learner/educators information, students attendance, Emotional data, Course information, data collected from the academic section of an institution etc..

Online Data. Online Data are generated from the geographically separated stake holder of the education, distance educations, web based education and computer supported collaborative learning used in social networking sites and online group forum.

E.g.: Web logs, E-mail, Spreadsheets, and Tran scripted Telephonic Conversations, Medical records, Legal Information, Corporate contracts, Text data, publication databases etc [3].

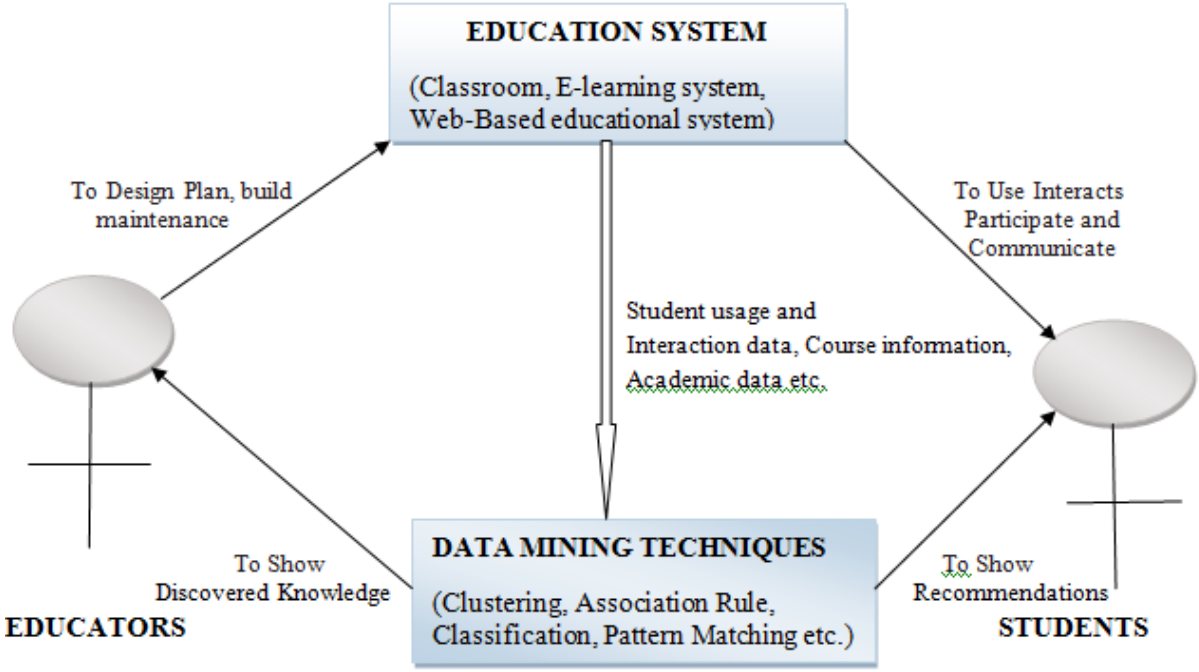


Figure 1(EDM)

In fig.1(EDM) we represent the need of Educational Data Mining. The Academicians and educationists worked upon the educational system to enhance the performance of students. In this diagram it is shown that

- Educators want to design the educational system then plan to build that system and most importantly to maintain that educational system. Educational systems include traditional classrooms and some innovative learning methods like e-learning system, intelligent and adaptive web based educational system etc.
- The data set can be extracted from students as students are directly connected with educational system.
- Now the data is given as input to data mining process and in result it gives recommendations to students and to extract new knowledge to the educators by using various data mining techniques like clustering, classification, pattern matching etc[2].

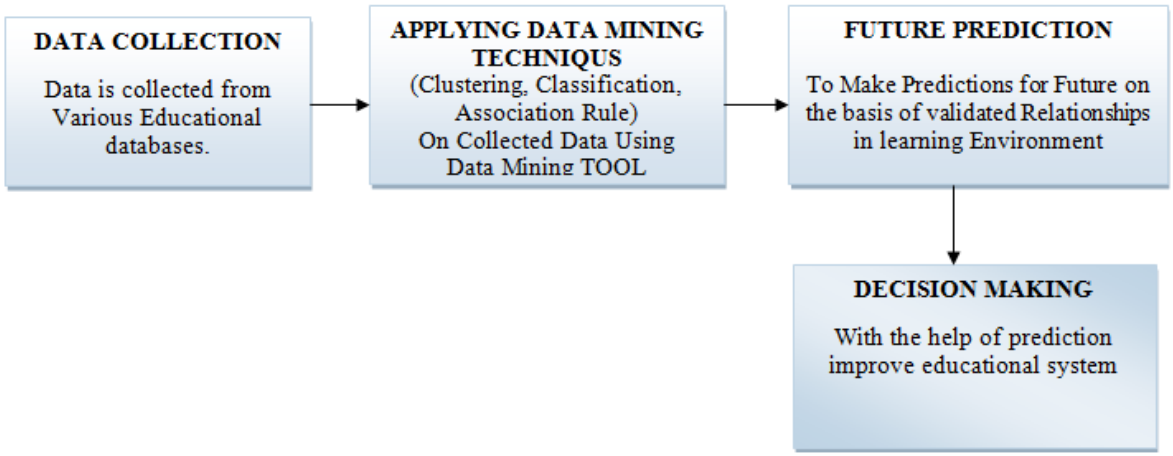
II. OBJECTIVES OF EDUCATIONAL DATA MINING (EDM)

EDM aims to improve several aspects of educational system. EDM Objectives depend on the view-point of the final users (learner, educator, administrator and researcher) and it helps to clear their problems:

1) **Student Modeling:** User modeling in the educational domain incorporates such detailed information as student's characteristics or states such as knowledge, skills, motivation, satisfaction, meta-cognition, attitudes, experiences and learning progress, or certain types of problems that negatively impact their learning outcomes. The common objective here is to create or improve a student model from usage information.
2) **Predictive Modeling:** Predicting students' performance and learning outcomes. The objective is to predict a student's final grades or other types of learning outcomes (such as retention in a degree program or future ability to learn) based on data from course activities.
3) **Generating Recommendations:** The objective is to recommend students that content (or tasks or links) which is the most appropriate for them at the current time.
4) **Analyzing learner's behavior**: This takes on several forms: Applying educational data mining techniques to analyze learner behavior.
5) **Maintaining and improving courses**: The objective here is to determine how to improve courses (contents, activities, links, etc.), using information (in particular) about student usage and learning. Discovering or improving models that characterize the subject matter to be learned (e.g. math, science, etc.), identify fruitful pedagogical sequences, and suggest how these sequences might be adapted to student's needs. Studying the effects of varied pedagogical enhancements on student learning.
6) **Learners:** To support a learner's reflections on the situation, to provide adaptive feedback or recommendations to learners, to respond to student's needs, to improve learning performance, etc.
7) **Educators:** To understand their student's learning processes and reflect on their own teaching methods, to improve teaching performance, to understand social, cognitive and behavioral aspects, etc.
8) **Administrators:** To evaluate the best way to organize institutional resources (human and material) and their educational system [3].

2.1 Phases of Educational Data Mining

Educational Data Mining is concerned with translation of new hidden information from the raw data collected from educational systems. EDM generally consist of following phases:



- The data is collected which is to be mined from different educational system resources i.e. from course management system (different institutes), E-learning environment, web based data (i.e. YouTube, twitter)which is relevant to students activities during learning process(i.e. their academic grades, students posts on social networking sites etc)
- **Educational Data Mining Process Phases**

1) The first phase of educational data mining is to find the relationships between the data of educational environment using data mining techniques i.e. classification, clustering, regression etc.

2) The second phase of educational data mining is validation of discovered relationships between data so that uncertainty can be avoided.

3) The third phase is to make predictions for future on the basis of validated relationships in learning environment.

4) The fourth phase is supporting decision making process with the help of predictions [2].

2.2 Educational Data Mining Methods

EDM not apply only data mining techniques Classification, clustering, and association analysis, but also apply methods and techniques drawn from the variety of areas related to EDM (statistics, machine learning, text mining, web log analysis, etc.).

There are so many methods of educational data mining but all kind of methods lie in one of following categories:

1) **Prediction:** The goal is to develop a model which can infer a single aspect of the data (predicted variable) from some combination of other aspects of the data (predictor variables). Types of predictions methods are classification, regression (when the predicted variable is a continuous value), or density estimation (when the predicted value is a probability density function).

- **Regression:** Regression is an inherently statistical technique used regularly in data mining. Regression analysis establishes a relationship between a dependent or outcome variable and a set of predictors.

Regression is supervised learning data mining technique. Supervised learning partitions the database into training and validation data. There are two type of regression technique.

- Linear regressions.
- Non linear regressions.

2) **Classification:** classifies a data item into some of several predefined categorical classes .The algorithm used for classification are:

- Decision tree
- Naive biased classification
- Generalized Linear Models (GLM)
- Super vector machine etc.

3) **Clustering**: In clustering technique, the data set is divided in various groups, known as clusters. As per clustering phenomenon, the data point of one cluster and should be more similar to other data points of same cluster and more dissimilar to data points of another cluster. There are two ways of initiation of clustering algorithm: Firstly, start the clustering algorithm with no prior assumption and second is to start clustering algorithm with a prior postulate.

4) **Relationship mining**: It is used for discovering relationships between variables in a dataset and encoding them as rules for later use. There are different types of relationship in mining techniques such as association rule mining (any relationships between variables), sequential pattern mining (temporal associations between variables), correlation mining (linear correlations between variables), and causal data mining (causal relationships between variables). In EDM, relationship mining is used to identify relationships between the student's on-line activities and the final marks and to model learner's problem solving activity sequences..

5) **Discovery with Models**: Its goal is to use a validated model of a phenomenon (using prediction, clustering, or knowledge engineering) as a component in further analysis such as prediction or relationship mining. It is used for example to identify the relationships between the student's behavior and characteristics.

6) **Outlier Detection:** The goal of outlier detection is to discover data points that are significantly different than the rest of data. An outlier is a different observation (or measurement) that is usually larger or smaller than the other values in data. In EDM, outlier detection can be used to detect deviations in the learner's or educator's actions or behaviors, irregular learning processes, and for detecting students with learning difficulties [2, 8].

2.3 Hallenges of Educational Data Mining

The research trends on EDM since the year 1998 to 2012 and found that maximum research focuses were on academic objectives. The other issues are:

1) **Educational data is incremental in nature**: Due to the exponential growth of data, the maintaining the data in data warehouse is difficult. To monitor the operational data sources, infer the student interest, intentions and its impact in a particular institution is the main issue. Another issue is the alignment and translation of the incremental educational data. It should focus on appropriating time, context and its sequence. Optimal utilization of computing and human resources is another issue of incremental.

2) **Lack of Data Interoperability:** Scalable Data management has become critical considering wide range of storage locations, data platform heterogeneity and a plethora of social networking sites.E.g: Metadata Schema Registry is a tool to enhance Meta data interoperability. So there is a need to design a model to classify/ cluster the data or find relationships. Examples of clustering applications are grouped students based on their learning and interaction patterns used in and grouping users for purposes of recommending actions and resources to similar users. It is possible to introduce Neuro-Fuzzy mining technique to remove the gap of data interoperability.
3) **Possibility of Uncertainty:** Due to the presence of uncertain errors, no model can predict hundred percent accurate results in terms of student modeling or overall academic planning.
4) Research Expertise Relation between Student-Teacher. In most of the higher Educational institutions (e.g. Engineering Institutions) final year students have a compulsory project work which are a research work based on their area of interest. Generally Supervisors are assigned as per availability and area of expertise in the respective department. But still it is not possible to assign all the students –supervisor with similar area of interest hence the result of the project is not applicable to real scenarios. There is need to find the relation between areas of interest, students' interest, applicability of the project/research and mining cross faculty interest. It will be beneficial to introduce using Association Mining to optimize this issue [3].

III. COMPARISION OF EDUCATIONAL DATA MINING(EDM) TECHNIQUES

3.1 Predictive Analytics in Higher Education

In this paper Jindal Rajni and Dutta Borah Malaya [2015] implemented a prediction analysis method that can help to improve the education quality in higher education for ensuring organization success at all level. They used the C5.0, C4.5-A2, C4.5-A1 algorithms for prediction analysis, after that they compare their results. The result of C5.0 is best in performance. Then they applied NN (Neural Network) and CRT algorithms on same data set for prediction analysis. After that they compared the result of C5.0 with Neural Network and CRT algorithms result. This paper analyzes the accuracy of algorithm in two ways; the first is by comparing the result of C5.0 with C4.5-A2, and C4.5-A1. After that the C5.0 algorithm is comes out to be best algorithm in accuracy. Then its result is compared with NN (Neural Network) and CRT [13].

3.2 Mining Social Media Data for Understanding Student Learning Experiences

In this paper X.chen, M.Vorvoreanu, K.Madhavan [2014] suggested how social media sites data is helpful in Understanding student learning experience. They collected data about student's problems from twitter. They also developed a workflow to integrate both qualitative analysis and large-scale data mining techniques. They focused on engineering student's twitter posts to understand issues and problems in their educational experiences. They used Naive Bayes Multi-Label Classifiers for tweets classification then after that they compare the result of Naïve Bayes Multi-Label Classifiers with the most used and accurate classifier used in many machine learning tasks i.e. Super Vector Machine (SVM) and Max margin Multi Label Classifier [1].

3.3 Comparison and Analysis of Various Clustering Methods in Data mining On Education data set using weka tool.

In this paper Suman, P Mittal [2014] implemented different clustering algorithms (K-Means algorithm, K-Medoids algorithm, Hierarchical clustering algorithm, Grid based algorithm, Density based clustering algorithm, Optics clustering algorithm) on educational data set. They collected educational data set related to student's record from different engineering branches. Then different clustering algorithms are applied on that educational data set. After that the comparison of different clustering algorithms is done on weka tool [4].

3.4 A Comparative Study of Association Rule Algorithms for Course Recommender System in E-learning

Sunita B. Aher, Lobo L.M.R.J[2012] implemented different association rule algorithm(Apriori Algorithm, Predictive Apriori Algorithm, Tertius Algorithm & Filtered Associator algorithm) for course recommendation system in E-learning Environment then apply these algorithm on Weka tool. After that comparative analysis of different association rule algorithm is done [5].

3.5 Comparision Table

PAPER	AUTHOR	TOOL USED	TECHNIQUE USED	ALGORITHM USED	ACCURACY	ADVANTAGE
Predictive Analytics in Higher Education	Jindal Rajni and Dutta Borah Malaya	Weka	Classification	C5.0	99.91%	C5.0 is better in speed, memory and efficiency than C4.5-A1, C4.5-A2 and the C5.0 algorithm provides maximum Information Gain.
				Neural Network(NN)	98.09%	
				CRT	98.72%	
Mining Social Media Data for Understanding Student's Learning Experiences	X.chen, M.Vorvoreanu and K.Madhavan	Weka	Classification	Naive Bayes Multi- Label Classifier	The Accuracy of Naïve Bayes Multi-Label Classifier is better and more accurate than Super Vector Machine(SVM) and Max Margin Multi-Label Classifier	Shortcoming of Naïve Bayes Multi-Label classifier is removed by comparing their result with most accurate classifiers.
				Super Vector Machine (SVM) Multi-Label Classifier		
				Max Margin Multi- Label Classifier		
Comparison and Analysis of Various Clustering Methods in Data Mining On Education Data Set Using the Weak tool	Suman, P. Mittal	Weka	Clustering	K-Means Clustering	The Accuracy Of K-Mean algorithm is better and more accurate than K-Medoids, Hierarchical, Grid Based and Density based, Optics Clustering Algorithms.	K-Mean algorithm is less time consuming than any other clustering algorithms.
				K-Medoids Clustering		
				Hierarchical Clustering		
				Grid based Clustering		
				Density based Clustering		
				Optics Clustering		
A Comparative Study of Association Rule Algorithms for Course Recommender System in E-learning	Sunita B.Aher, Lobo L.M.R.J	Weka	Association Rule	Apriori Algorithm	Accuracy of Apriori algorithm is better and more accurate than that of Predictive Apriori and Tertius & FilteredAssociator algorithm	The Apriori Algorithm is easy to implement and its implementation consume less time and its results is more accurate as compared to other algorithms
				Predictive Apriori Algorithm		
				Tertius Algorithm and Filtered Associator algorithm		

IV. CONCLUSION

This paper describes goals of Educational Data Mining, phases of Educational Data Mining, Educational data Mining Techniques as well as the challenges of Educational Data Mining. In this paper, we did the comparative study of different Education Data Mining Techniques with their algorithms on educational data sets using Weka tool. We also did the comparative analysis on the basis of accuracy percentage. We also analyzed the advantages of algorithms that applied to educational data set. It is difficult to say that which technique of education data mining is best because each technique has its own advantage and limitations and it also depend upon the purpose for which educational data is to be mined. But according to our comparative study on Educational Data Mining techniques on weka tool we can say that in Classification technique C5.0, Naïve Bayes Classification is the best algorithm in performance and in Clustering Technique K- Mean clustering algorithm is best algorithm or in Association Rule Technique Apriori algorithm is best and more accurate as compared to other algorithms. Our research study will be beneficial to the researchers in educational data mining for the selection of educational data mining techniques with their algorithm according to their research area. So we can say that this paper will provide a beneficial glimpse of existing solution for education data mining techniques with their accuracy and advantages.

REFERENCES

[1] X.chen, M.Vorvoreanu, K.Madhavan, Mining Social Media Data for Understanding Student's Learning Experiences, *Ieeexplore.Ieee.Org*, *7(3)*, 2014, 246–259.

[2] N. Upadhyay, V. Katiyar, A Survey on the Classification Techniques in Educational Data Mining, *International Journal of Database Management System (IJDMS)*, *3(11)*, 2014, 725–728.

[3] R.Jindal, M.D Borah, A Survey on Educational Data Mining and Research trends, *International Journal of Database Management System (IJDMS)*, *5(3)*, 2013, 53–73.

[4] Suman, P.Mittal, Comparison and Analysis of Various Clustering Methods in Data mining On Education Data set using the Weka tool, *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, *3(2)*, 2014, 240–244.

[5] SunitaB.Aher, Lobo.L.M.R.J, A Comparative Study of Association Rule Algorithms for Course Recommender System in E-learning, *International Journal of Computer Applications*, *39(1)*, 2012, 48–52.

[6] M.A Khan, W.Gharibi and S.K.Pradhan (2014), Data Mining Techniques for Business Intelligence in Educational System: a Case Mining, 978-1-4799-3351-8/14/$31.00 ©2014 IEEE.

[7] V.T.N.Chau and N.H. Phung, Imbalanced Educational Data Classification: An Effective Approach with Resampling and Random Forest. Proceedings - 2013 RIVF International Conference on Computing and Communication Technologies: Research, Innovation, and Vision for Future, RIVF 2013, 135–140.

[8] Baradwaj, B. Kumar, Mining Educational Data to Analyze Students Performance. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *2(6)*, 2011, 63-69.

[9] Z.K .Dehkordi (2013).New Approach to the Design of Decision Support System to Improve E-Leaming Environments, 26–29,The 4th International Conference on E-learning and E-teaching (ICELET) 2013.

[10] Q.Liu, Y.Peng (2013), Research on the Method of Unstructured Information Process in Computer Teaching Evaluation System Based on Data Mining Technology, *Energy Procedia*, *11*, 5095–5103, International Conference on Communication System and Network Technologies (2013).

[11] A. M Morais, J. M. F. R. Araújo and E.B. Costa (2014), Monitoring Student Performance Using Data Clustering and Predictive Modelling,978-1-4799-3922-0/14/$31.00 ©2014 IEEE.

[12] H.Sahu, S.Shrma, S.Gondhalakar, A Brief Overview on Data Mining Survey, *International Journal of Computer Technology and Electronics Engineering (Ijctee), 1(3),* 2008, 114–121.

[13] R.Jindal, M.D Borah (2015), Predictive Analytics in Higher Education, 1520-9202/15/$31.00 © 2015 IEEE