A Survey on Educational Data Mining-Prediction and Classification

Ashna Sethi, Charanjit Singh

Abstract: ***Educational Data Mining (EDM) is an upcoming field examining and exploring data in educational context by implementing different Data Mining (DM) techniques/tools. It provides knowledge of teaching and learning as a process for effective education planning. In this survey work focuses on highlighting Techniques and educational Outcomes. In this paper, Various DM techniques are discussed and comparison of classifiers is made. A general Methodology for classification and Prediction is mentioned.***

Keywords: ***Educational Data Mining (EDM), EDM Components, DM Methods, Education Planning***

I. INTRODUCTION

Educational Data Mining (EDM) is an upcoming field examining and exploring data in educational context by implementing different Data Mining (DM) techniques/tools. EDM inherits properties from areas like Statics, Learning Analytics, Psychometrics, Artificial Intelligence, Information Technology, Machine Learning, Database Management System, Computing and Data Mining. The huge growth of educational data from heterogeneous sources results an urgent need for research in EDM. This can help to meet the objectives and to determine specific goals of education. The method of extracting information from large set of databases and using it to make important business decisions is termed as **data mining**. It involves the process of examining data from different aspects and grouping it into useful information.

- **Educational Data Mining (EDM)** is an emerging discipline.
- It is concerned with **developing methods** for exploring the unique and increasingly large-scale data.
- The data in EDM comes from **educational firms**.
- Those methods are used to better **understand** students.

The data mining in **field of Education** is called EDM

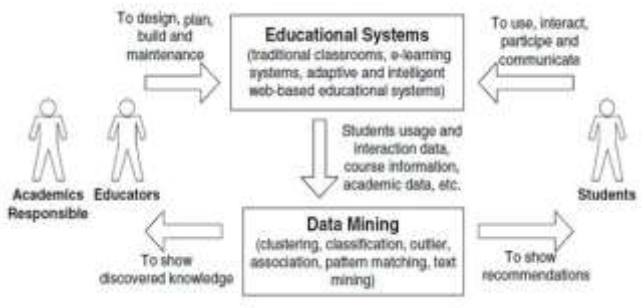

Fig: 1 Educational Data Mining

Revised Version Manuscript Received on May 06, 2017.

Ashna Sethi, Department of Computer Science and Engineering, Regional Institute of Management and Technology, Mandi Gobindgarh (Punjab)-147301, India.

Charanjit Singh, Assistant Professor, Department of Computer Science and Engineering, Regional Institute of Management and Technology, Mandi Gobindgarh (Punjab)-147301, India.

II. EDM OBJECTIVE CAN BE CLASSIFIED IN THE FOLLOWING WAY:

(1) Academic Objectives

- Person oriented (related to direct participation in teaching and learning process) E.g.: Student learning, cognitive learning, modeling, behavior, risk, performance analysis, predicting right enrollment decision etc. both in traditional and digital environment and Faculty modeling- job performance and satisfaction analysis.
- Educational Firms oriented (related to particular department/institutions with respect to sequence, time and demand). E.g.: Redesign new courses according to requirements; identify problems to effective research and learning process.
- Domain Oriented (related to particular branch/institutions) E.g.: Designing Methods-Knowledge Discovery based Decision Support System (KDDS) for specific application, Tools, Techniques.

(2) Administrative Objectives

- Administrator Oriented (related to direct involvement of higher authorities/administrator) E.g.: Resource (Infrastructure as well as Human) utilization, Industry academia relationship, marketing for student enrollment in case of private institutions and establishment of network for innovative research and practices.
- To explore heterogeneous educational data by analyzing the authors' views from traditional to intelligent educational systems in the decision making process.
- To explore intelligent tools and techniques used in EDM.

A. Applications of EDM:-

- Financial Data Analysis
- Telecommunication Industry
- Biological Data Analysis
- Scientific Applications
- Mining of Clusters

III. EDM COMPONENTS

The key components of EDM are Stakeholders of Education, DM Methods-Tools and Techniques, Educational data, Educational task and Outcomes.

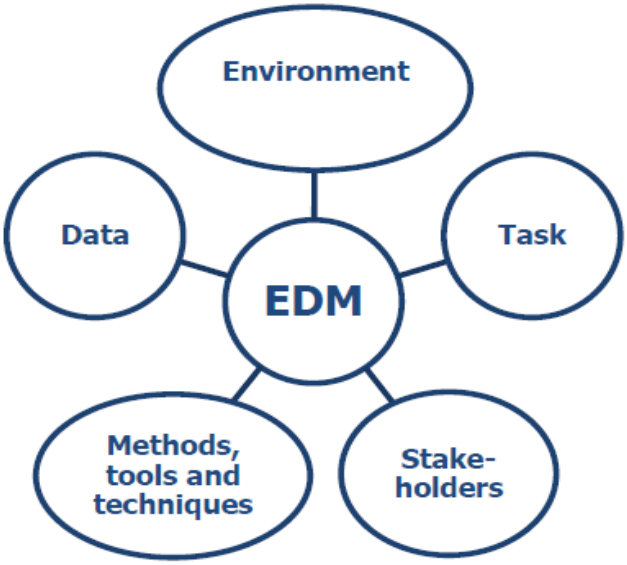

Fig 2: Components of EDM

IV. DM METHODS

Data mining methods are one of the vital components in EDM. Following DM methods are popular with the EDM research community.

A. Classification

It is a two way technique (training and testing) which maps data into a predefined class. Classification analysis is the organization of data in given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The algorithm learns from the training set and builds a test model. The model is implemented to classify new objects. This Technique is useful for success analysis with low, medium, high risk students used in, student monitoring systems, predicting student performance, misuse detection used in etc.

Clustering

It is similar to classification that organizes the data in classes. In clustering, class labels are unknown and it depends upon the clustering algorithm to discover acceptable classes. It is a technique to compare similar data which is formed into clusters in a way that groups are not predefined. This technique is useful to distinguish learner with their preference in using interactive multimedia system used in, Students comprehensive character analysis used in and suitable for collaborative learning used in.

Applications of Clustering:-

- Pattern Recognition
- Market Research
- Spatial Data Analysis
- Image Processing

B. Prediction

It has attracted considerable attention given the potential implications of successful forecasting in a business context. There are two major types of predictions: one can either try to predict some unavailable data values or pending trends, or predict a class label for some data. It is a technique which predicts an upcoming state rather than a state. This technique is useful to predict success rate, drop out used in Dekker et al, and retention management used in of students.

Most of the prediction techniques are strongly based upon mathematical models:

– Simple statistical models: Regression
– Non-linear statistics: Power Series
– Neural networks, RBFs, etc

All are based on finding a relationship from the predictors to the predicted

C. Neural Network

It is a technique to improve the interpretability of the learned network by using extracted rules for learning networks.[6] This technique is useful to determine residency, ethnicity used in, to predict academic performance used in, accuracy prediction in the branch selection used in and explores learning performance in a TESL based e-learning system.

D. Association Rule Mining

Association is one of the data mining techniques. In this method, a pattern is discovered based on a relationship between items in the same transaction. That's the reason why association technique is also known as *relation technique*. The association technique is mainly used in *market analysis* to identify a set of products that customers frequently purchase together. Retailers are using this technique to research customer's purchasing habits. Based on sale data, retailers might find out that customers always buy butter when they buy bread, and, therefore, they can put bread and butter next to each other to save time for customer and increase sales. It is a technique to identify specific relationships among data. This technique is useful to identify students' failure patterns ,parameters related to the admission process, migration, contribution of alumni, student assessment, co-relation between different group of students, to guide a search for a better fitting transfer model of student learning etc.

E. Web Mining

It is a technique for mining web data. This technique is useful for building virtual community in

Computational Intelligence used in, to determine misconception of learners used in and to explore cognitive sense. Apart from the above methods, mentioned two new methods i.e. distillation of data for human judgment and discovery with models to analyze the behavioral impact of students in learning environments.

F. Sequential Patterns

Sequential patterns analysis is a DM technique that is widely used to discover regular events, similar patterns or trends in transaction of data over a business period. In sales, with historical transaction data, businesses can identify a set of items that customers purchase together in a year at different times. [15] Then businesses can use this information to recommend customers to buy it with better deals based on their purchasing frequency in the past.

G. Decision Trees

The A decision tree is one of the common used data mining techniques because its model is easy to understand for users. In this technique, the root of the decision tree is a condition or simple question that has multiple answers. Each answer then leads to a set of questions or conditions that help us determine the data so that we can make the final decision based on it.

H. Naïve Bayes

Naive Bayes classifiers is a simplest probabilistic classifiers which is based on a Bayesian Theorem with strong independent assumptions between the different features. It is highly scalable, requiring a number of parameters with number of variables (features/predictors) in a learning problem. In this Method, maximum training can be done by evaluating a closed-form expression, which takes some time, instead of using iterative approximation ,Hence less expensive as compared to other classifiers.. In the statistics and computer science literature, Naive Bayes models are known under a variety of names, including simple Bayes and independence Bayes.

Table 1: Classifiers Comparison

S. No	Tools	Advantage	Disadvantage
1	Neural Network	Nonlinear application Works with incomplete data Capability of updating and reasoning	Needs a lot of data to train Difficult to deal with missing data Does not support mixed variable black-box behavior
2	Decision Tree	Nonlinear application Mixed variable Simple interpretation Capability of reasoning	Needs a lot of data Difficult to deal with missing data
3	K-Nearest Neighbor	Nonlinear application Works with incomplete data Mixed variable Capability of updating	Needs a lot of data Difficult to deal with missing data
4	Bayesian Networks	Nonlinear application Works with incomplete data Mixed variable Simple inter-predation Capability of updating and reasoning	Needs large training data
5	Neuro-Fuzzy	Nonlinear application Works with incomplete data Capability of learning and reasoning	Does not support mixed variable
6	Support Vector Machine	Nonlinear application Accuracy in small data set	Does not support mixed variable Difficult to deal with missing data

V. LITERATURE SURVEY

Amornsinlaphachai. P, [1] has studied that to select a data mining model to predict learners' academic performance in computer programming subject to group learners for cooperative learning by comparing the efficiency of the models created from data mining with classification technique. To develop a model for cooperative learning via web using the selected data mining model to group learners. The efficiency of seven models created from data mining with classification technique by using seven algorithms that are Artificial Neural Network, K-Nearest Neighbor, Naive Bayes, Bayesian Belief Network, JRIP, ID3 and C4.5 is compared and it was found that the models created from C4.5 has the best efficiency.

Buniyamin et al. [2] highlights the importance of using student data to drive improvement in education planning. He describes the development of a tool that will enable faculty members to identify, predict and classify students based on academic performance measured using Cumulative Grade point average (CGPA) grades. His work elaborates a brief overview of the most commonly used classifiers techniques in educational data mining and an outline of the use of Neuro-Fuzzy classification in a case study research to predict and classify students' academic achievement in an Electrical Engineering faculty of a Malaysian public university.

Abaidullah et al. [3] presented the analysis of student's feedback data for decision making by educational community responsible for monitoring and reviewing the effectiveness of educational programs and for improving the quality of teaching and learning experience for their students. For this purpose k-means clustering algorithm is used.

La Red et al. [16] described the various mining models and discussed the main results. Mining models of clustering, classification and association were considered especially. It seeks to analyze patterns of success and failure for students in academics, therefore predicting the likelihood of dropping out or having poor academic performance of students, with the advantage of being able to do it early, allowing addressing action to reverse this situation.

Shaukat.K et al. [8] has focused on recognizing, extracting and calculating data associated to the learning method and improving student's performance. The purpose of our study is to evaluate the performance of students by taking different attributes like academic achievements (CGPA), gender, class test grade, environment of class, Fund/Scholarships/Private etc. In our research we will use classification and clustering techniques to analyze student performance. The techniques used in our work are decision tree, Bayesian classification-mean algorithms, neural networks, Naive's Bayes, Web based system and nearest neighbor methods.

Shahiria. A.M et al. [7] has highlighted on how the prediction algorithm can be used to identify the most important attributes in a student's data. We can really improve student's achievement and success more effectively in an efficient way using EDM techniques. It could bring the benefits and impacts to students, educators and academic institutions.

Devasai.T. et.al. [4] Has used compared different techniques of data mining on EDM like: Naive Bayesian, Regression, Decision Tree, Neural networks the proposed system is a web based application which makes use of the Naive Bayesian mining technique for the extraction of useful information. The experiment is conducted on 700 students' with 19 attributes in Amrita Vishwa Vidyapeetham, Mysuru. Result proves that Naïve Bayesian algorithm provides more accuracy over other methods like Regression, Decision Tree, Neural networks etc., for comparison and prediction.

Alshareef et al. [15] justified the capabilities of data mining techniques in the context of higher education by offering a data mining model for the higher education system at Sebha University. In this research, association rules were used to evaluate students' performance by applying the apriori algorithm on survey data. In this task author extract knowledge that describes students' performance, which helps in identifying earlier trends in the choices of major and in helping new students to select their major.

VI. CONCLUSION

EDM is a data mining process applied on data from Educational Field. There is a necessity in today's time to work on EDM as number of educational institutes is increasing. So there is a need to identify the trend of student's performance. And for that a system is needed to build where the performance, Dropout Rate, Result can be predicted. By predicting the above mentioned things, we can take suitable measures in advance to overcome the problems that come in Education system. Number of techniques has been used for this like Artificial Neural Networks, Naïve Bayesian, Decision Tree, C4.5 etc. But still there is wide scope of improvement. EDM in coming future is going to be the field of research for more.

REFERENCES

1. Amornsinlaphachai.P. (2016). Efficiency of data mining models to predict academic performance and a cooperative learning model.2016 IEEE 8th International Conference on Knowledge and Smart Technology (KST)
2. Buniyamin, N., Mat, U. bin, & Arshad, P. M. (2015). Educational data mining for prediction and classification of engineering student's achievement. 2015 IEEE 7th International Conference on Engineering Education (ICEED), 49–53.
3. Abaidullah, A. M., Ahmed, N., & Ali, E. (2014). Identifying Hidden Patterns in Students' Feedback through Cluster Analysis. International Journal of Computer Theory and Engineering, 7(1), 16–20
4. Devasia. T, T P.V, Hegde. V .(2016). Prediction of Students Performance using Educational Data Mining.2016 IEEE International Conference on Data Mining and Advanced Computing (SAPIENCE)
5. Borah.M.D et.al(2012). Application of knowledge based decision technique to predict student enrollment decision. IEEE 2011 International Conference on Recent Trends in Information Systems (ReTIS).
6. Banumathi and A. Pethalakshmi, "A novel approach for upgrading Indian education by using data mining techniques," in Technology Enhanced Education (ICTEE), 2012 IEEE International Conference on, 2012, pp. 1-5.
7. Shahiria.A.M, Husaina.W, Rashida.N.A.(2015) A Review on Predicting Student's Performance using Data Mining Techniques. 3rd International Conference on Information Systems
8. Shaukat.K, Nawaz.I , Aslam.S, Zaheer.S, Shaukat.U. (2017). Student's performance in the context of data mining. 19th International Conference on Multi-Topic Conference (INMIC)
9. Gobert, J. D., Kim, Y. J., Sao Pedro, M. A., Kennedy, M., & Betts, C. G. (2015). Using educational data mining to assess students' skills at designing and conducting experiments within a complex systems microworld. Thinking Skills and Creativity, 18(September 2016), 81–90.
10. F. Yi and Z. Chunyuan, "Improving the Quality of Graduate Education by Association Rules Analysis," in Fuzzy Systems and Knowledge Discovery, 2008. FSKD '08. Fifth International Conference on, 2008, pp. 570-573.
11. Mayilvaganan, M., &Kalpanadevi, D. (2015). Cognitive Skill Analysis for Students through Problem Solving Based on Data Mining Techniques. Procedia Computer Science, 47, 62–75
12. Banumathi and A. Pethalakshmi, "A novel approach for upgrading Indian education by using data mining techniques," in Technology Enhanced Education (ICTEE), 2012 IEEE International Conference on, 2012, pp. 1-5.
13. Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. Computers and Education, 61(1), 133–145.
14. Kaur, P., Singh, M., &Josan, G. S. (2015). Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector. Procedia Computer Science, 57, 500–508.
15. Alshareef, A., Ahmida, S., Bakar, A. A., Hamdan, A. R., &Alweshah, M. (2015). Mining survey data on university students to determine trends in the selection of majors. Proceedings of the 2015 Science and Information Conference, SAI 2015, 586–590.
16. La Red Martínez, D. L., &Podestá Gómez, C. E. (2014). Contributions from Data Mining to Study Academic Performance of Students of a Tertiary Institute. American Journal of Educational Research, 2(9), 713–726.
17. Lee, G., and Chen, Y.C. (2012), "Protecting sensitive knowledge in association pattern mining", John Wiley & Sons, Inc .2, pp.60-68.,DOI:10.1002/widm.50.
18. Calders, T., and Pechenizkiy, M. (2012), "Introduction to the special section on Educational Data Mining", SIGKDD Explorations.Vol.13, No.2, pp.3-6.
19. Huebner, R.(2012), "A.Educational data-mining research", Research in Higher Edu. Journal, pp.1-13.
20. Vandamme, J.P. et al. (2007), "Predicting academic performance by Data Mining methods", Taylor and Francis group Journal Education Economics.Vol.15, No.4, pp.405-419.
21. Mansmann, S.and Scholl, H. (2007), "Decesion Support System for managing Educational Capacity Utilization",IEEE Transaction on Education, Vol.50,No.2,pp.143-150,DOI: 10.1109/TE.2007.893175
22. Romero,C. et al.,(2008), "Data Mining in course management systems: Moodle case study and tutorial", Computer and Education, Elsevier publication. Vol. 51, No. 1,pp.368-384.
23. Delavari,N. et al.(2008), "Data Mining Application in Higher Learning Institutions",Journal on Informatics in Education.Vol.7,No.1,pp.31-54.
24. Perera,D. et al.(2009), "Clustering and sequential pattern mining of online collaborative learning data", IEEE Transactions on Knowledge and Data Engineering,Vol.21, No.6,pp.759-772.
25. Zurada, J.M.et al.(2009), "Building Virtual Community in Computational Intelligence and Machine Learning", IEEE Computational Intelligence Mazazine.pp.43-54,.DOI: 10.1109 / MCI. 2008. 9309 86.
26. Baker, R.S.J.D.,and Yacef, K.(2009), "The state of Educational Data Mining in 2009:A review and future vision" Journal of Educational Data Mining, Vol.1,No. 1,pp.3-17.
27. Chrysostomu K. el al.(2009), "Investigation of users' preference in interactive multimedia learning systems: a data mining approach",Taylor and Francis online journal Interactive learning environments. Vol. 17,No. 2.
28. Romero, C., and Ventura, S. (2010), " Educational Data Mining: A review of the state of the Art", IEEE Trans.on on Sys. Man and Cyber.-Part C: Appl. and rev., Vol.40, No.6, pp. 601-618.
29. Al-shargabi, A.A. and Nusari, A. N (2010), "Discovering Vital Patterns From UST Students Data by Applying Data Mining Techniques", in Proc. Int. Conf. On Computer and Automation Engineering, China: IEEE, 2010, 2,547-551.DOI:10.1109/ICCAE.2010.5451653.
30. Jafar, M.J., (2010), "A Tool based approach to teaching Data Mining Methods", Journal of Information Technology Education: Innovations in Practice..Vol.9, pp. 1-24.
31. Zhang Y.et al. (2010), "Using data mining to improve student retention in HE: a case study", in Proc.12th Int. Conf. on Enterprise Information Systems, Volume 1: Databases and Information Systems Integration. Portugal, pp.190-197.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

32. Dalip, D.H., Gonclaves, M. A. (2011), "Automatic Assessment of Document Quality in Web collaborative Digital Libraries", ACM Journal of Data and Information Quality.Vol.2,No.3, pp.14.DOI 10.1145/2063504.2063507
33. Baradwaj,B.K., and Pal,S.(2011), "Mining Student Data to Analyze Students' Performance" International Journal of advanced Computer Science and applications.2,6.
34. Wang and Liao.(2011), "Data Mining for adaptive learning in a TESL based e-learning system", in Elsevier journal Expert systems with applications,Vol.38,No.6,pp.6480-6485.
35. Alberg, D., Last, M., and Kandel, A (2012), "Knowledge discovery in data streams with regression tree methods" John Wiley & Sons, Inc,2,pp.69-78,DOI:10.1002/widm.51