Educational Data Mining:Review

Paramjit Kaur[#1], Kanwalpreet Singh Attwal[#2]
[#1]*M.Tech Scholar,Department of Computer Engineering,Punjabi University Patiala Punjab ,India*
[1]paramjitkaur335@gmail.com
[#2]*Astt.Professor,Department of Computer Engineering,Punjabi University Patiala Punjab ,India*
[2]kanwalp78@yahoo.com

Abstract— *As the cost of processing power and storage is coming down, data storage became easier and cheaper so the amount of data stored in educational databases is increasing rapidly and in order to find precious hidden information from such large data we can use different data mining techniques. Data mining has potential in analyzing and uncovering previously unknown information from huge size educational data which is hard and very time consuming if to be done manually. The purpose of this review is to look into how the data mining is used in educational research and find out the capabilities of data mining in context of educational data.*

Keywords—Data mining, EDM, Classification, Clustering ,Prediction

I. INTRODUCTION

Data Mining is used to extract meaningful information and to developed significant relationship among variables stored in large data warehouse [1]. "Educational data mining is an emerging discipline concern with developing methods for exploring the unique type of data that come from educational setting and using those methods to better understand students and the setting which they learn in" as defined by the educational data mining community [2] . Education is an essential element for the progress of country. Mining in educational environment is called educational data mining. It is concerned with developing new methods to discover knowledge from educational database [3]. Educational data mining provides a set of techniques, which can help the educational system to overcome these issues [4].

Educational systems have special characteristics that require a different treatment of the mining problem. As a consequence, some specific data mining techniques are needed to address in particular the process of learning (Li & Zai¨ane, 2004; Pahl & Donnellan, 2003). Some traditional techniques can be adapted, some cannot [5].

Educational Data Mining (EDM) is the application of Data Mining (DM) techniques to educational data, and so ,its objective is to analyse these type of data in order to resolve educational research issues [6].EDM involves different groups of users or participants .Different groups look at educational information from different angles according to their own mission, vision and objectives for using data mining [7].

For example ,knowledge discovered by EDM algorithms can be used not only to help teachers to manage their classes, understand their students' learning processes and reflect on their own teaching methods, but also to support a learner's reflections on the situation and provide feedback to learners [8].

II. DIFFERENT USERS OR ACTORS OF EDM

A. Student

Student uses EDM to personalize e-learning.EDM recommend activities, resources, learning tasks, courses, relevant discussions, books etc that could further improve students' learning;

B. Educators

Educators use EDM To get objective feedback about instruction; to analyze students' learning and behaviour; to detect which students require support; to predict student performance; to classify learners into groups; to find a learner's regular as well as irregular patterns; to find the most frequently made mistakes; to determine more effective activities; to improve the adaptation and customization of courses, etc.

C. Course Developers

Course Developers use EDM to evaluate and maintain courseware; to improve student learning; to evaluate the structure of course content and its effectiveness in the learning process; to automatically construct student models and tutor models; to compare data mining techniques in order to be able to recommend the most useful one for each task; to develop specific data mining tools for educational purposes; etc.

D. Universities

Universities use EDM to enhance the decision processes in higher learning institutions; to streamline efficiency in the decision making process; to achieve specific objectives; to suggest certain courses that might be valuable for each class of learners; to find the most cost-effective way of improving retention and grades; to select the most qualified applicants for graduation; to help to admit students who will do well in university, etc.

E. School Administrators

To develop the best way to organize institutional resources (human and material) and their educational offer; to utilize available resources more effectively; to enhance educational program offers and determine the effectiveness of the distance learning approach; to evaluate teacher and curricula; to set parameters for improving web-site efficiency and adapting it to users(optimal server size, network traffic distribution, etc.).

III. EDM Tasks

Education data mining general task is to Guide student learning efforts, develop or refine student models, improved teaching support, predict student performance and behaviour etc.

Baker [10],[11] suggests four key areas of application for EDM:

- Improving student models.
- Improving domain models.
- Studying the pedagogical support provided by learning software.
- Scientific research into learning and learners.

and five approaches/methods: prediction, clustering, relationship mining, distillation of data for human judgment and discovery with models.

Castro [12] suggests the following EDM subjects/tasks:

- Applications dealing with the assessment of the student's learning performance
- Applications that provide course adaptation and learning recommendations based on the student's learning behaviour
- Approaches dealing with the evaluation of learning material and educational web based courses,
- Applications that involve feedback to both teacher and students in e-learning courses,
- Developments for detection of atypical students' learning behaviours.

In Romero & Ventura's survey of EDM research from 1995 to 2009, 300 papers were reported that utilized EDM methods to answer research questions of applied interest. Romero &Ventura [13] suggests the following EDM tasks:

- Analysis and visualization of data, Providing feedback for supporting instructors, to make recommendations directly to the students,
- Predicting student performance,
- To develop cognitive models of human users/students, including a modelling of their skills and declarative knowledge,
- To discover/detect those students who have some type of problem or unusual behaviour,
- To create groups of students according to their customized features, personal characteristics, Social network analysis.

IV. Importance of EDM

Education is an essential element for the betterment and progress of a country. It enables the people of a country civilized and well mannered. Mining in educational environment is called Educational Data mining, concern with developing new methods to discover knowledge from educational database in order to analyze student's trends and behaviours towards education. Lack of deep and enough knowledge in higher educational system may prevents system management to achieve quality objectives, data mining methodology can help bridging this knowledge gaps in higher education system.

The educational system in India is currently facing several issues such as identifying students need, personalization of training and predicting quality of student interactions. Educational data mining (EDM) provides a set of techniques which can help educational system to overcome this issue in order to improve learning experience of students as well as increase their profits [6]. Manual data analysis has been around for sometimes now, but it creates bottleneck for large data analysis. The transition won't occur automatically; in this case, we need the data mining. Data mining software allow user to analyzed data from different dimensions categorized it and summarized the relationship, identified during mining process [9].

V. DM techniques in EDM

DM methods are one of the main components in EDM. As per the different purpose it can be broadly divided into two groups [14]:

- Verification Oriented (Traditional Statistics-Hypothesis test, Goodness of fit, Analysis of Variance etc.)
- Discovery Oriented (Prediction and Description-Classification, Clustering, Prediction, Relationship Mining, Neural Network, Web mining etc.)

Following DM methods are popular with the EDM research community:

A. Classification

Classification is a procedure in which individual items are placed into groups based on quantitative information regarding one or more characteristics inherent in the items and based on a training set of previously labeled items [15].Other comparisons of different data mining algorithms are made

- To classify students(predict final marks) based on Moodle usage data [16];
- To predict student performance (final grade) based on features extracted from logged data [17]
- To predict University students' academic performance [18].

B. Prediction

Prediction of a student's performance is one of the oldest and most popular applications of DM in education, and different techniques and models have been applied (neural networks,

Bayesian networks, rule-based systems, regression and correlation analysis). A comparison of machine learning methods has been carried out to predict success in a course (either passed or failed) in Intelligent Tutoring Systems [19]. It is a technique which predicts a future state rather than a current state. This technique is useful to predict success rate, drop out used in Dekker et al. [20,21], and retention management used in[22] of students.

C. Statistics

It is a technique to identify outlier fields, record using mean, mode etc. and hypothetical testing. This technique is useful to improve the course management system & student response system[23].

D. Clustering

It is a technique to group similar data into clusters in a way that groups are not predefined. Cluster analysis or clustering is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster have some points in common [37]. This technique is useful

- To distinguish learner with their preference in using interactive multimedia system used in [25],
- Students comprehensive character analysis used in [21]
- For collaborative learning used in [26,27].

E. Neural Network

It is a technique to improve the interpretability of the learned network by using extracted rules for learning networks. This technique is useful

- To determine residency, ethnicity used in [28].
- To predict academic performance used in [29],
- Accuracy prediction in the branch selection used in[30]
- To explores learning performance in a TESL based e-learning system [31].

F. Association Rule Mining

It is a technique to identify specific relationships among data. Association rule mining has been used to provide new, important and therefore demand-oriented impulses for the development of new bachelor and master courses [38]. This technique is useful to identify

- Students' failure patterns [32],
- 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. used in [33].

G. Web mining

It is a technique for mining web data. This technique is useful for building virtual community in computational Intelligence used in [34], to determine misconception of learners used in [35] and to explore cognitive sense.

Apart from the above methods,[27] mentioned two new methods i.e. distillation of data for human judgment and discovery with models to analyze the behavioural impact of students in learning environments [36].

VI. EDUCATIONAL DATA MINING TOOLS

Tool name	Source(Open/ free/commercial)	Mining Task
Intelligent Miner (IBM) Windows, Solaris, Linux	Commercial	Association Mining, Classification, Regression, Predictive Modelling, Deviation detection, Clustering, Sequential Pattern Analysis
MSSQL Server 2005 (Microsoft) .	Commercial	Integrates the algorithms developed by third party vendors and application users
Oracle Data Mining (Oracle Corporation)	Commercial	Association Mining, Classification, Prediction, Regression, Clustering, Sequence similarity search and analysis.
iData Analyzer (Microsoft)	Open /Free	Heuristic Agent, Neural Network, Rule Maker and Report generator
WEKA (University of Waikato, New Zealand)	Open/free	Data pre-processing, classification, regression, clustering, association rules, and visualization
Carrot Provides	Open/free	Clustering

VII. CHALLENGES OF EDM

A. Mining tools complexity

Data mining tools are normally designed more for power and flexibility than for simplicity. Most of the current data mining tools are too complex to use for educators and their features go well beyond the scope of what a educator may want to do .So, these tools must have a more intuitive and easy to use interface, with parameter-free data mining algorithms to simplify the configuration and execution, and with good visualization facilities to make their results meaningful to educators and e-learning designers.

B. Integration with the e-learning system

The data mining tool has to be integrated into the e-learning environment as another author tool. All data mining tasks (pre processing, data mining and post processing) have to be carried out into a single application. Feedback and results obtained with data mining can be directly applied to the e-learning environment.

C. Specific data mining techniques

More effective mining tools that integrate educational domain knowledge into data mining techniques. Education-specific mining techniques can help much better to improve the instructional design and pedagogical decisions. Traditional mining algorithms need to be tuned to take into account the educational context.

VIII. CONCLUSIONS

Educational data mining is an upcoming field related to several well-established areas of research including e-learning, adaptive hypermedia, intelligent tutoring systems, web mining, data mining, etc.
Data Mining can be used in educational field to enhance our understanding of learning process to focus on identifying, extracting and evaluating variables related to the learning process of students as described by Alaa el-Halees [24].

REFERENCES

- Shaeela Ayesha,(2010),data mining model for higher education system, J. of scientific research, vol -43, pp24-29.
- Sunita B. Aher,(2011), data mining in education system using WEKA, International J. of Computer Applications, pp20-25.
- Alaa el- Halees,(2009), Mining Students Data to Analyze e-learning behavior. A case study.

- Connolly T., C. Begg et al,(1999) Database System: A practical approach to design, Implementation and management(3 rd edition),
Harlow; Addison-Wesley,687.
- Li, J., & Zaı¨ane, O. (2004). Combining usage, content, and structure data to improve web site recommendation. In International conference on ecommerce and web technologies (pp. 305–315).
- Barnes, T., Desmarais, M., Romero, C., Ventura, S. (2009). Educational Data Mining 2009: 2nd International Conference on Educational Data Mining, _Proceedings. Cordoba, Spain.
- Hanna, M. (2004). Data mining in the e-learning domain. In Campus-Wide Information Systems, Volume 21, Number 1, 29-34.
- Merceron, A., Yacef, K. (2005). Educational Data Mining: a Case Study. In International Conference on Artificial Intelligence in Education, Amsterdam, The Netherlands, 1-8.
- ZhaoHui. Maclennan.J, (2005). Data Mining with SQL Server 2005 Wihely Publishing, Inc.
- Baker, R., Yacef, K. (2009) The State of Educational Data Mining in 2009: A Review and Future Visions. Journal of Educational Data Mining, 1, 1, 3-17.
- Baker, R (2010). Data Mining for Education. To appear in McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education(3rd edition). Oxford, UK: Elsevier
- Castro, F., Vellido, A., Nebot, A. Mugica, F. (2007). Applying Data Mining Techniques to e-Learning Problems. In: Jain, L.C., Tedman, R. and Tedman, D. (eds.) Evolution of Teaching and Learning Paradigms in Intelligent Environment. Studies in Computational Intelligence, 62,Springer-Verlag, 183-221.
- 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.
- Mainman,O., and Rokach,L. (Ed.)(2010) Data Mining and Knowledge Discovery Handbook, 2nd ed..,DOI:10.1007/987-0-387-09823-4_1,Springer Science Business Media,LLC.
- Espejo, P., Ventura, S., Herrera, F. (2010) A Survey on the Application of Genetic Programming to Classification. IEEE Transactions on Systems, Man, and Cybernetics-Part C. 40, 2, 121-144.
- Romero, c., ventura, s., hervás, c., gonzales, p. (2008). Data mining algorithms to classify students. In International Conference on Educational Data Mining, Montreal, Canada, 8-17.conference Intelligent Systems, Varna, Bulgaria, 581-586.
- Minaei-bidgoli, B., Kashy, D.A., Kortmeyer, G., Punch, W.F. (2003).Predicting student performance: an application of data mining methods with an educational Web-based system. In International Conference on Frontiers in Education, 13-18.
- Ibrahim, Z., Rusli, D. (2007). Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression. In Annual SAS Malaysia Forum, Kuala Lumpur, 1-6.
- Hämäläinen, W., Vinni, M. (2006). Comparison of machine learning
methods for intelligent tutoring systems. In international conference in
intelligent tutoring systems, Taiwan, 525-534.
- Dekker, G., Pechenizkiy, M., and Vleeshouwers J.(2009), "Predicting students drop out: A case study", In Proceedings of the 2nd International Conference on Educational Data Mining, pp.41-50.
- 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.
- Lin, S.H.(2012), "Data Mining for student retention management" *ACM journal of Computing Sciences in CollegesI*. Vol.27,No.4,pp. 92-99.
- Campbell, J.P., and Oblinger, D.G. (2007,Oct.). Academic Analytics. EDUCAUSE, Washington,D.C. [Online]. Available: http://net.educause.edu/ir/library/pdf/pub6101.pdf,2007.
- Alaa el-Halees, "Mining students data to analyze e-Learning behavior: A Case Study", 2009..
- 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.
- 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.
- 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.
- Yu et al. (2010), "A data mining approach for identifying predictors of student retention from sophomore to junior year", *Journal of Data Science.*Vol.8, pp.307-325.
- 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.
- Dutta Borah, M., Jindal, R., Gupta,D.,Deka,G.C (2011), "Application of knowledge based decision technique to Predict student's enrolment decision. *in Proc. Int. Conf. on Recent Trends in Information Systems*,

India:IEEE,pp.180-184,DOI:10.1109/ReTIS.2011.6146864.

- 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.
- Oladipupo,O.O.,Oyelade,O.J.(2009), "Knowledge Discovery from Students' Result Repository:Association Rule Mining Approach", *International Journal of Computer Science &Security*,Vol.4,No.2,pp.199-207.
- Freyberger,J., Heffernan, N., Ruiz, C.(2004), "Using association rules to guide a search for best fitting transfer models of student learning", *Workshop on Analyzing Student-Tutor Interactions Logs to Improve Educational Outcomes at ITS Conference*.
- 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. 2008.930986.
- 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.
- Cocea,M., and Weibelzahl,S (2009), "Log file analysis for disengagement detection in e-learning environments", Springer Journal User Modeling and User Adapted Interaction. Vol.19, No.4, pp.341-385. DOI: 10.1007/s 11257-009-9065-5.

[37] Romesburg, H.C. (2004). Cluster Analysis for Researchers. Krieger Pub.

[38] Schönbrunn, K., Hilbert, A. (2007). Data Mining in Higher Education.In Annual Conference of the Gesellschaft für Klassifikation e.V., FreieUniversität,Berlin,489-496.