Literature Survey on Educational Data Mining

Prasanna M. Kothawale
Assistant Professor
Department of Computer Science,
Dnyanprassarak Mandal's College and Research Centre, Assagao, Bardez, Goa 403 507, India

Abstract: All the researches referred in the present review have revealed some significant areas in the field of education where prediction with data mining has provided systematic valuable information on various aspects of students and education related issues including management and policy designing. It is found that there are specific issues of concern; sometimes they are institution specific that needs to be addressed for better results. Attributes like emotional intelligence, self-management, work experiences and life experiences are also important attributes in gauging and lifting student's performance in the academics. One author has recommended open source Waikato Environment for Knowledge Analysis (WEKA) software for convenient data preprocessing, cleaning and handling missing values, and this is very common application software that can be used in each school / college for initial collection of data.

1.INTRODUCTION

Data are generated daily on many aspects of student's lives and most of this generated data goes unprocessed. If it is used and analyzed properly we will find interesting and valuable information. Data mining is a tremendously vast area that includes employing different techniques and algorithms for pattern finding. In case of education system, which is an area selected in this paper, large amounts of data from a variety of sources such as classes, students, administration, faculty members etc. are collected daily. Hence mining of this data is essential. Educational Data Mining (EDM) is one of the emerging trends that concerns with developing methods for exploring the unique types of data that come from educational settings, and using it for better understanding about students, and institutional set ups which they learn in. EDM is an efficient tool for data recovery, analytics, recommending system, understanding student psychological aspects etc. Real value of information from large amount of data generated can be experienced by systematic analysis using techniques of data mining. Analysis of data of any educational institute using EDM according to their need base and their own perspectives will provide patterns that will certainly help to place any given educational institute in a strategic position to resolve issues for better output.

2. ANALYSIS OF RESEARCH LITERATURE

The possible achievable goals such as predicting student's future learning behavior, studying the effects of educational support, advancing scientific knowledge about learning and learners, and discovering or improving domain models through EDM is presented by N. Bhagoriya et al. [1]. According to author all the promoted methods of EDM lie in one of the following specified categories such as predictions, clustering, relationship mining, discovery with models and refinement of data for human judgment. Samira ElAtia et al. in her paper, [2] considered application of data mining in educational systems as an iterative cycle of hypothesis formation, testing, and refinement and concluded that research using data mining techniques can be useful in three contexts: providing information that enhances learning from the students' perspective, providing insight into learning and teaching for university teachers; and providing the institution and its administrators with valuable knowledge that would be an asset for decision-making. The focus of EDM was on learning, teaching, managing the institution from an academic experience and managing the institution from an infrastructure perspective.

Carla Silva et al. [3] investigated the different data mining approaches and techniques which can be applied on educational data to build up a new environment giving new predictions on the data. This study also looks into the recent applications of big data technologies in education and presented a literature review on EDM and learning analytics. Some of the data mining algorithms on clustering, classification, predication approaches applied on education related areas were discussed. The reference of new term "connectivism" which is characterized as the "amplification of learning, knowledge and understanding through the extension of personal network" is highlighted by author.

The study of Amjad Abu Saa [4] explores multiple factors that theoretically assumed to affect students' performance in higher education, and finds a qualitative model which best classifies and predicts the students' performance based on related personal and social factors. In this work multiple decision tree techniques and algorithms were reviewed, and their performances and accuracies were tested and validated. Tree techniques such as *ID3* (Iterative Dichotomiser 3) *decision tree, (CART)* Classification and Regression Tree, *and* CHi-squared Automatic Interaction Detection (CHAID) *decision tree* were analyzed. It was noted CART had the best accuracy of 40%, which was significantly more than the expected (default model) accuracy, CHAID with 34.07% and the least accurate was ID3 with 33.33%. In this research paper, multiple data mining tasks were used to create qualitative predictive models which were efficiently and effectively able to predict the students' grades from a collected training dataset using implementation of the Naïve Bayes classification technique. First, a survey was constructed that has

targeted university students and collected multiple personal, social, and academic data related to them. Second, the collected dataset was preprocessed and explored to become appropriate for the data mining tasks. Third, the implementation of data mining tasks was presented on the dataset in hand to generate classification models and testing them. Finally, interesting results were drawn from the classification models, as well as, interesting patterns in the Naïve Bayes model was found. In the current study, it was slightly found that the student's performance is not totally dependent on their academic efforts, in spite; there are many other factors that have equal to greater influences as well. Author also systematically presented research work on a group of 50 students enrolled in a specific course program across a period of 4 years using various performance parameters such , semester marks of previous year, grades of internal class test, seminar, assignments, lab work, attendance, semester end marks etc. The ID3 decision tree algorithm was used to finally construct a decision tree to help the teacher as well as the students to better understand and predict students' performance at the end of the semester. Furthermore, they extended their work to identify those students which needed special attention to reduce fail ratio and taking appropriate action for the next semester examination.

The survey of the work carried out by Rajni Jindal et al. [5] focuses on components and research trends during the period of 1998 to 2012 of EDM highlighting its related tools, techniques and educational outcomes. It also highlights the challenges of EDM to meet the objectives and to determine specific goals of education. The objectives of the EDM were classified into two parameters academic objectives and administrative objectives. The work also highlighted on EDM components such as stakeholders, environments, data, methods, tools, research trends and future scope of EDM. Various algorithms are discussed by P. Nithya et al. in their paper [6] such as Support Vector Machine (SVM), Apriori, Expectation Maximization Algorithm and Page Ranker to achieve several of educational support through EDM. Brijesh Kumar Baradwaj et al. [7] in this task extracted knowledge that describes students' performance in end semester examination and identifying the dropouts and students who need special attention that will allow the teacher to provide appropriate advising/counseling through EDM. In present investigation various domain values of the course were defined including previous semester marks (PSM), class test grade (CTG), seminar performance (SEM), assignment performance (ASS), general proficiency performance (GP), attendance of student (ATT), Lab Work (LW) and end semester marks (ESM). The data set of 50 students used in this study was obtained from VBS, Purvanchal University, Jaunpur (Uttar Pradesh) Computer Applications department of course MCA (Master of Computer Applications) from session 2007 to 2010. Information such as attendance, class test, seminar and assignment marks were collected from the students previous database, to predict the performance at the end of the semester.

Author Ryan S.J.d. Baker [8] explained four areas of applications firstly, improving student models, models that provide detailed information about a student's characteristics or states, such as knowledge, motivation and attitudes. A second key area of application is in discovering or improving models of the knowledge structure of the domain. A third key area of application is in studying the pedagogical support provided by learning software. A fourth key area of application of educational data mining is for scientific discovery about learning and learners. In this work, a brief case study is discussed, as a concrete "best practices" example of how the educational data mining method of learning decomposition (a type of relationship mining) was used to determine the relative efficacy of different types of learning material presented to students. A case study [9] was done by Agathe Merceron et al. to show how using data mining algorithms can help discovering pedagogically relevant knowledge contained in databases obtained from Web-based educational systems. These findings can be used both to help teachers with managing their class, understand their students' learning and reflect on their teaching and to support learner reflection and provide proactive feedback to learners. Study is focused on to identify students those who have not trained enough and are at risk of failure. Clustering and cluster visualization were used to identify a particular behavior among failing students, when students try out the logic rules of the pop-up menu of the tool. A timely and appropriate warning to students at risk could help preventing failing in the final exam.

Jaya Srivastava in her paper [10] reviews the application of various data mining tools and techniques that can be effectively used in answering the issues of predictions of student's performance and their profiling. Author highlighted on broad challenges of Indian education system such as a gap between supply and demand, poor quality of teaching and learning and limited research capacity. Various functional areas included for applications of data mining by author are predicting students' admission in higher education, predicting students' profiling, predicting students' performance, teachers' teaching performance, curriculum development, students' targeting, and predicting students' survival in a course, predicting students' course selection and predicting students' placement opportunities. In this paper a student data from a community college database has been taken and various classification approaches have been performed and a comparative analysis has been done. The research [11] work on Support Vector Machines (SVM) is established as a best classifier with maximum accuracy and minimum root mean square error (RMSE) by Sonali Agarwal et al. The study also includes a comparative analysis of all SVM kernel types and radial basis kernel is identified as a best choice for SVM. A decision tree approach is proposed which may be taken as an important basis of selection of student during any course program. The paper is aimed to develop a faith on data mining techniques so that present education and business system may adopt this as a strategic management tool. The dataset has 4 attributes and 2000 records of student performance details. The attributes are MAT score, verbal ability score, quantitative ability score and likelihood of placement. Author recommends WEKA software to use as it is suitable and is an open source. WEKA can efficiently work with limited data. WEKA also provides convenient data preprocessing, cleaning and handling missing values. It takes data from excel file in Comma Separated Values (CSV) format, which is a very common application software to be used in each school / college for initial collection of data. This software contains tools for a whole range of data mining tasks like data pre-processing, classification, clustering, association and visualization.

The work under the project by S. Lakshmi Prabha [12] presents broad areas of applications of educational data mining for e-learning using user modeling, user grouping or profiling, domain modeling and trend analysis. This modeling was implemented on MathsTutor for school students of 6th, 7th and 8th grade designed in 3 schools in the state of Tamilnadu- India. This paper provides only limited number of screen shots applied on student data. It is suggested that by identifying the knowledge level of a student and grouping them will make easier for the teacher to concentrate the areas for week students. So application of EDM methods on student data opens up the possibility for students to develop skills in monitoring their own learning and to see directly how their effort improves their success. Teachers get views into students' performance that assist them adapt their teaching or commence tutoring, tailored assignments, and the like. Using the data, administrators can make policies, execute programs, and adapt the policies and programs to progress teaching, learning, and completion/retention rates.

The result of the study by Mahendra Tiwari et al. [13] is aimed to develop a faith on data mining techniques so that present education system may adopt this as a strategic management tool. In this work data of B.Tech second year (CS & IT branch) student from database management system course held at the United College of Engineering and Research Naini Allahabad (Affiliatedto GBTU) in fourth semester of 2011/2012 was collected. In this study questionnaire were collected to have the real data describing relationships between learning behavior of students and their academic performance. The variable for judging learning and academic behavior of students used in questionnaire were assignment, attendance, internal sessions marks, GPA (grade point average for general performance in lab or extracurricular), and final grade. In this case data mining techniques with association and classification rules were used to predict the students' performance and were clustered into groups using k-means clustering algorithm.

The focus by Nitya Upadhyay et al. [14] is on the educational data mining and classification techniques in his paper. The attributes for the prediction of student's behavior and academic performance by using WEKA open source data mining tool and various classification methods like decision trees, C4.5 algorithm, ID3 algorithm etc were reported. A comparative analysis on different existing approaches and methods of classification of data sets like Naïve Bayesian classification, Multilayer Perceptron, J48 and ID3 etc were made. Advantages and shortcomings of each algorithm applied to data set were also analyzed to provide a beneficial glance of existing solution for classification with their advantages and shortcomings. Total of 40 tools frequently used for data mining/analytics in the area of education were reviewed by Slater, S et al. [15]. Different approaches to different problems were represented each with their own particular strengths and weaknesses. Important discoveries were suggested using combination of tools for complex analyses.

The significance of use of algorithms such as C4.5, SVM, EM, Page Ranker, Naïve Bayes, Apriori and CART in the field of Education data mining is highlighted in his work by Shafiq Aslam et al.[16]. These algorithms has produced remarkable improvement in strategies like course outline formation, teacher student understanding and high output and turn out ratio. The paper by Falguni Ranadev et al. proposes the methodology to predict students' performance based on their psychological characteristics, internal and final examination results to address number of issues such as high dropout rates, identifying students in need, personalization of training and predicting the quality of student interactions [17]. The Big Five personality traits or Five Factor Model (FFM) including openness, conscientiousness, extraversion, agreeableness and neuroticism were used as scientific measure of personality and have been extensively researched. The proposed methodology can be extended to accommodate the evaluation of the method after the group/groups of students who are weak in the expertise area are allocated to teacher having respective expertise.

Author Varun Kumar et al. in his paper [18] reported empirical study on the applications of data mining in educational institution to extract useful information from the huge data sets and providing analytical tool to view and use this information for decision making processes by taking real life examples. The objective is to identify the potential areas in which data mining techniques can be applied in the field of higher education and to identify which data mining technique is suited for what kind of application. Application of data mining for example, to efficiently assign resources with an accurate estimate of how many male or female will register in a particular program by using the prediction techniques was reported. Detecting cheating in online examination, predicting student performance, predicting registration of students in an educational program and organization of syllabus to maintain a high quality educational program was discussed. Through the research work, Ancelmo Castro et al. [19] concluded that the decision tree is the main method applied in EDM field. 45% of the algorithms apply decision tree method which has graphic knowledge representation, which could help experts to better interpret the elicited evidences, and to postulate causes and effects. The algorithms C5.0, C4.5, and K-means, were highlighted methods in this study. These three mining algorithms are most commonly used to analyze data in large scale, especially in education.

In the paper by Shruthi P. et al. [20], has used classification rule on student database to predict the student's performance in the upcoming semester on the basis of previous student's database using Naïve Bayes algorithm. Information's like attendance, seminar and assignment marks were collected from the student's previous database, to predict the performance at the end of the semester. The other attributes are collected by students and their respective faculties who know the behavior of students. This study was aim to help students and the teachers to improve the result of the students who are at the risk of failure by identifying those students who needed special attention to reduce failure ratio and taking appropriate action for the next semester

examination. In the research project of Umesh Kumar et al. [21] authors are trying to analyze the success ratio, performance & productivity of the student by applying data mining tools & techniques such as: classification, association rule, clustering and decision tree. This paper tried to put emphasize on the different learning techniques such as offline educational system/traditional educational system, web mining/e-learning and intelligent tutorial system. By adopting all these learning techniques student and institutions could attain better enhancement and enrichment to obtain the knowledge in the field of academic curriculum. It is also beneficial to educator, recommender, policy maker, instructor and stakeholder to design course curriculum according the need of students.

A study on developing a research methodology for educational data mining was presented by comparing the Indian education system with another university data in USA in by Nidhi Chopra et al. [22]. The study predicts that model of our current education system where most people like to study or focus on their career value addition in their twenties or early thirties is very true. Students find it easy to do humanities courses and urban areas students show poor life style. The article by Z. Lustigova et al. describes the results of a data mining to observe students' behavior during their work in virtual environment [23]. Simple data mining and text mining techniques were used to reveal individual user's behavioral patterns, to detect disengagement, and to compare learning outcomes and student preferences. It was found that even though students are able to setup the apparatus, to start the measurement, to finish it correctly and to save the measured data, finally they mostly preferred data download in virtual lab. The credibility given by students to pre-measured data is very high. Students do not trust to their own results. This behavior in the present research was associated with the learning and teaching paradigm change (single to team work), lack of supervision and not used to increased uncertainty in the virtual environment.

Authors P.Veeramuthu et al. in their research work [24] have developed software that facilitates the use of the generated rules to built higher education systems to predict the student's loyalty. The aim is to improve the current trends in the higher education systems to understand the factors that might create loyal students by using valid management and processing of the students database using various data mining techniques. The areas like optimization of resources, prediction of retaining faculties in the university, to find the gap between the numbers of candidates applied for the post, number of applicants responded, number of applicants appeared, selected and finally joined was considered in this work. The study by Hung, J.-L et al. [25] has investigated an innovative approach of program evaluation through analyses of student learning logs, demographic data, and end-of-course evaluation surveys in an online K–12 supplemental program. Researcher developed a program evaluation model for decision making on teaching and learning at the K–12 level using case study on 7,539 student's sample. Clustering analysis was applied to reveal students' shared characteristics, and decision tree analysis was applied to predict student performance and satisfaction levels toward course and instructor. Few resulted outputs of the clustering analysis includes such as students with higher engagement levels usually had higher performance, female students were more active than male students in online discussions and younger students who lived in larger cities were more successful than those in smaller cities and older students.

4. CONCLUSION:

The various discussed algorithms in this review have revealed significant areas in the field of education where predicting with EDM can give benefits with regards to weak students, faculty evaluation, students' dropout, course planning and formulate appropriate strategies for students. EDM will guide us to take a step of systematic approach for collecting, storing, and analyzing important data and valuable knowledge could be harvested from the large scattered data. In conclusion, this review can motivate and help universities to perform data mining tasks on their students' data regularly to find out interesting results and patterns which can help both the university as well as the students in multiple ways. It is possible to analyze big data and get good answers within a reasonable time using various data mining algorithms. It is hoped that this review will be helpful to researchers working in the area of educational data mining.

REFERENCES:

1. Nupur Bhagoriya and Priyanka Pande, International Journal of Engineering Sciences and Research Technology, Vol. 6(4), (April 2017)
2. Samira ElAtia, Donald Ipperciel and Ahmed Hammad, Canadian Journal of Education 35, 2 101-119 (2012)
3. Carla Silva and José Fonseca, Advances in Intelligent systems and Computing, (September 2017)
4. Amjad Abu Saa, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 5, (2016)
5. Rajni Jindal and Malaya Dutta Borah, International Journal of Database Management Systems (IJDMS) Vol.5, No.3, (June 2013)
6. P. Nithya, B. Umamaheswari and A. Umadevi, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Vol. 5-1, (January 2016)
7. Brijesh Kumar Baradwaj and Saurabh Pal, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2-6, (2011)
8. Ryan S.J.d. Baker, Data Mining for Education. McGaw, and International Encyclopedia of Education (3rd edition). Oxford, UK: Elsevier

9. Agathe Merceron and Kalina Yacef, Educational data mining a case study.
10. Jaya Srivastava, Special Conference Issue: National Conference on Cloud Computing & Big Data.
11. Sonali Agarwal, G. N. Pandey, and M. D. Tiwari, International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 2 -2, (April 2012)
12. S. Lakshmi Prabha and A.R.Mohamed Shanavas, Operations Research and Applications: An International Journal (ORAJ), Vol. 1-1, (August 2014)
13. Mahendra Tiwari, Randhir Singh and Neeraj Vimal, International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 2-2, pg.53 – 57 (February 2013)
14. Nitya Upadhyay and Vinodini Katiyar, International Journal of Computer Applications Technology and Research,Vol. 3-11, 725 - 728, (2014)
15. Slater, S., Joksimovic, S., Kovanovic, V., Baker, R.S. and Gasevic, (Article) Tools for educational data mining: a review
16. Shafiq Aslam and Imran Ashraf, International Journal of Advance Research in Computer Science and Management Studies, Vol.2-7, (July 2014)
17. Falguni Ranadev and Dhaval Mehta, (Article) Improving Students' Performance using Educational Data Mining
18. Varun Kumar and Anupama Chadha, (IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 2-3, (March 2011)
19. Ancelmo Castro, Leandro Garcia, David Prata, Marcelo Lisboa, and Monica Prata, International Journal of Information and Education Technology, Vol. 7-5, (May 2017)
20. Shruthi P. and Chaitra B.P., International Journal of Advanced Research in Computer Science and Software Engineering, Vol.6-3, (March 2016)
21. Sen and Umesh Kumar, International Journal of Advanced Research in Computer Science & Technology (IJARCST 2015), Vol. 3-1, (January 2015)
22. Nidhi Chopra and Manohar Lal, Research Methodology for Educational Data Mining in India, (2012)
23. Z. Lustigova and , P. Brom, iJAC – Vol.7-1, (2014)
24. P.Veeramuthu and R.Periasamy, International Journal of Innovative Research in Advanced Engineering (IJIRAE), Vol.1-5 (June 2014)
25. Hung, J.-L., Hsu, Y.-C., & Rice, K. Educational Technology & Society, Vo. 15-3, pg: 27–41(2012)