International Journal of Data Mining Techniques and Applications
Volume: 05 Issue: 02 December 2016, Page No.167-171
ISSN: 2278-2419

A Survey on Educational Data Mining Techniques

A.S. Arunachalam[1], T.Velmurugan[2]
[1]Research scholar, Department of Computer Science, Vels University, Chennai, India.
[2]Associate Professor, PG and Research Department of Computer Science, D.G.Vaishanav College, Chennai, India.
Mail:arunachalam1976@gmail.com, velmurugan_dgvc@yahoo.co.in

Abstract - Educational data mining (EDM) creates high impact in the field of academic domain. The methods used in this topic are playing a major advanced key role in increasing knowledge among students. EDM explores and gives ideas in understanding behavioral patterns of students to choose a correct path for choosing their carrier. This survey focuses on such category and it discusses on various techniques involved in making educational data mining for their knowledge improvement. Also, it discusses about different types of EDM tools and techniques in this article. Among the different tools and techniques, best categories are suggested for real world usage.

Key words: Educational Data Mining, Web Mining, E-Learning, Data Mining Techniques

I. INTRODUCTION

Collecting relevant student record and analyzing the same from huge record set always remain difficult task for researchers. Data mining process of extracting hidden information from large database provides a meaningful solution for educational data mining. The researcher also faces many problems in implementing the developed system for educational data mining in different platform. Huge number of developments in educational courses always remains difficult task for students in choosing best course. Current web based course applications doesn't provide static learning materials by understanding students mentality. User friendly environment for web based educational system always remain a good solution to richer learning environment. In traditional education system, students share their learning experiences one to one interaction and continual evaluation process [1]. Classroom Evaluations processes by observing student's attitude, analyzing record, and student appraisal in teaching strategies. The supervision is not possible when the students working in IT field; pedagogue chooses for other techniques to get class room data. Institutions, which run websites for distance learning, collect huge data, collecting server access log and web server by automatically. Web based learning analysing tools available online increses the interaction data between acdemicion and students [2]. Most effective learninig environment can be carried by following data mining techniques. The data mining techniques stages starts from pre processing to post processing techniques by following KDD process of identifying necessary educational data. Web based domain area E-commerce uses data mining techniques in advancing educational mining. E-Learning process gives optimal solution for improving the educational data mining process. Some differents in E-learning and E-Commerce systems are disscussed below [3].

E-commerce technology is used for communicating client with server for commercial purpose. Web based applications are used for carrying out this technology. Web access log files stores E Commerce transaction data for money transfer. The commercial web sites transaction data passes the control to bank website application and purchases happen. E Learning technology is used for learning from web sites. The users can gather necessary knowledge from web based learning process. The web applications used in this type of websites are very useful in providing knowledge for learners. Web based applications stores the information through web access log files, which eventually stores information about users working on the websites.

In educational system the knowledge assessment techniques applied to improve students' learning process. The formative assessment process evaluating continues improvement of students learning capacity. The formative system helps the educator to improve instructional materials. The data mining techniques helped the educator to make academic decision when designing or editing the teaching methodology. The educational data mining follow the common data mining methods. Extracted information should enter the circle of the system and guide, fine tuning and refinement of learning [4]. This data not only becoming the knowledge, it improves the mined knowledge for decision making. The rest of the survey paper is planned as follows. Section 2 discusses about the basic concepts of educational data mining. In section 3, it is discussed about tools used for educational data mining. The applications or techniques of educational data mining are illustrated in section 4. Finally, section 5 concludes the survey work.

II. EDUCATIONAL DATA MINING

Educational data mining play a major role in society and educational area. The data mining sequence applying in educational system can be clearly represented with a diagram shown in Figure1.

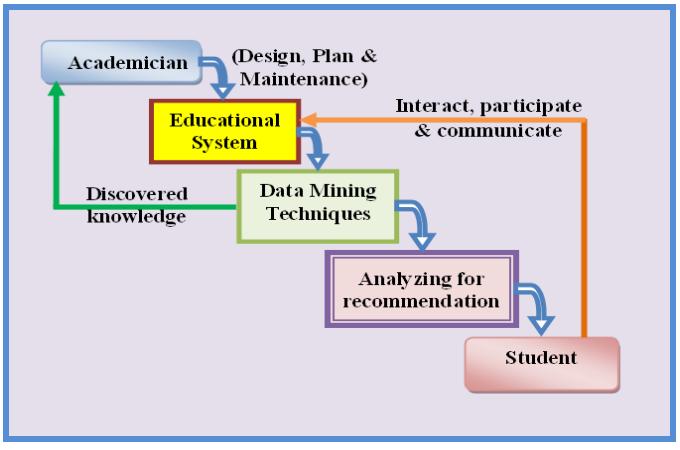

Figure 1: Data mining sequence applying in Educational System

a. Academician

Academician plays a major role in designing educational system. The educational system should be constructed for better benefits to students. This effort may cause dramatic changes in educational environment and society. The academicians and teachers should analyze students' records and construct the better educational system. Data mining techniques provides an additional benefit to academicians for analyzing student's behavior based on historical data.

b. Educational System

Proper educational system provides a healthy environment and society. The students are one of the major factors in educational system that eventually makes the society healthy. The rules and regulations are made possible by the use of experienced academicians.

c. Data mining Techniques

Historical Students records are collected from various localities and data mining techniques are applied. The unwanted data are removed by applying preprocessing technique. There are multiple data mining techniques have been used in EDM process including classification, association, prediction, clustering, sequential pattern and decision tree.

d. Analyzing for Recommendation

The goal to have constraint about how to develop educational system efficiency and get used to it to the performance of their students, have measures about how to enhanced categorize institutional resources (human and material) and their educational offer, enhance educational programs offer and determine effectiveness of the new computer mediated distance learning approach.

e. Students

The objective is absorb students learning experience,resources, their activities and interest of learning based on the responsibilities already done by the student and their successes and on errands made by other similar learners, etc.

III. TOOLS FOR EDUCATIONAL DATA MINING

Zaiane and Luo discusses about Web Utilization Miner (WUM) for students' feedback system. This added value by specific event recording on the E- learning side will give click steams and the patterns discovered a better meaning and interpretation [5]. Silva and Vieira uses web usage ranking technique to identify student information web pages. MultiStar textually presents the patterns it finds. The patterns resulting from the classification task are expressed through certain rules [6]. Shen implements student academic performances visualization through statistical graphs [7]. Romero discovers interesting prediction rules from student usage information to improve adaptive web courses like AHA!(Adaptive Hypermedia for All). He also uses a visual tool (EPRules) discover prediction rules and it is oriented to be used by the teacher [8]. Tane propose an ontology-based tool to build the majority of the resources available on the web. He uses text mining and text clustering techniques in order to group papers according to their topics and similarities [9]. Merceron and Yacef , used traditional SQL queries to mining student data captured from a web-based tutoring tool and association rule and symbolic data analysis. Their objective is to find mistakes that often occurs together [10]. Becker introduced Sequential patterns can expose which content has provoked the access to other contents, or how tools and contents are tangled in the learning process [11]. Avouris, N., Komis, V., Fiotakis, G., Margaritis, M. and Voyiatzaki, E. develop automatically generated log files by introducing contextual information as additional events and by associating comments and static files [12]. Mazza and Milani introduced a tool GISMO/CourseVis which features Information visualization techniques can be used to graphically render multidimensional, complex student tracking data collected by web-based educational systems these techniques facilitate to analyze large amounts of information by representing the data in some visual display [13]. Mostow used Listen tool for understanding of their learners and become aware of what is happening in distance classes [14]. Damez, M., Dang, T.H., Marsala, C. and Bouchon-Meunier, B., used a fuzzy decision tree for user modeling and discriminating a learner from an experimented consumer automatically. They use an agent to learn the cognitive characteristics of a user's relations and classify users as experimented or not [15]. Bari and Benzater recover data from pdf hypermedia productions for serving the assessment of multimedia presentations, for statistics purpose and for extracting relevant data. They recognize the major blocks of multimedia presentations and recover their internal properties [16]. Qasem A. Al.Radaideh has introduced CRISP Framework for mining student related academic data. He have used decision tree as a classification technique for rule mining in academic data [17]. Cristobal Romero also introduced a tool KEEL which is a software tool to access evolutionary algorithms to solve various data mining problems in regression, classification and unsupervised learning [18]. C.Marquez-Vera has introduced an SMOTE algorithm for classifying rebalanced students success rate using 10 Fold Cross Validation, which the rules was implemented and tested in WEKA [19]. Dragan Gašević identifies the critical topics that require immediate research attention for learning analytics to make a sustainable impact on the research and practice of learning and teaching by using online learning tool and video annotation tool [20].

IV. TECHNIQUES FOR EDUCATIONAL DATA MINING

Cristobal Romero compares different type of data mining classification techniques with the use of Moodle usage data of Cordoba University. He modeled a rebalanced preprocessing technique for classifying original numerical data [21]. Ramasami has developed a Predictive data mining model for identifying slow learners and to study the influence of the dominating factor of their academic performances. He also introduced CHAID model for predicting slow learners in an accurate manner [22]. Edin Osmanbegovie successfully implemented a datamining technique in higher education socio-demographic variables and analysis high school entrance exam and attribute related to it. He also uses some of supervised algorithms for analyzing the student data [23]. Sujeet Kumar Yadav has used three decision trees and three machine learning algorithms (ID3, C4.5, and CART) for obtaining students predictive model. He also classifies by giving accuracy value in time and identifies student's success and failure ratio [24].

International Journal of Data Mining Techniques and Applications
Volume: 05 Issue: 02 December 2016, Page No.167-171
ISSN: 2278-2419

Table 1: Educational Data Mining Tools

Authors	Tool	Mining task	Findings
Zaiane and Luo (2001)	WUM	Association and Patterns	E- learning side will give click steams and the patterns discovered a better meaning and interpretation.
Silva and Vieira (2002)	MultiStar	Association and classification	MultiStar textually presents the patterns it finds. The patterns resulting from the classification task are expressed through certain rules.
Shen et al. (2002)	Data Analysis Center	Association and classification	E- Learning system for solving problem like students and teacher's interaction problem in assignments and other problems.
Romero et al. (2003)	EPRules	Association	Grammar based genetic programming with multi-objective optimization techniques for providing a feedback to course ware authors was developed and identifies increasing relationship in student's usage data.
Tane et al. (2004)	KAON	Text mining and Clustering	The Course ware Watch dog addresses the different needs of teachers and students. It integrates the Semantic Web vision by using ontologies and a peer-to-peer network of semantically annotated learning material.
Merceron and Yacef (2005)	TADA-ED	Classification and Association	Implementing traditional SQL queries to mining student data from a web based tutoring tool. The objectives of fining mistakes that occur often are done with accurate rating.
Becker et al. (2005)	O3R	Sequential patterns	The proposed filtering functionality 1) They have the support of the ontology to understand the domain, and establish interesting filters, and 2) Direct manipulation of domain concepts and structural operators minimize the skills required for defining filters.
Mostow et al. (2005)	Listen tool	Visualization	Log each distinct type of tutorial event in its own table. Include student ID, computer, start time, and end time as fields of each such table so as to identify its records as events.
Merceron and Yacef (2005)	TADA-ED	Classification and Association	Implementing traditional SQL queries to mining student data from a web based tutoring tool. The objectives of fining mistakes that occur often are done with accurate rating.
Avouris et al. (2005)	Synergo/ ColAT	Statistics and visualization	Main features of two tools that facilitate analysis of complex field data of technology mediated learning activities, the Synergo Analysis Tool and ColAT.
Mazza and Milani (2005)	GISMO/ CourseVis	Visualization	GISMO has been implemented based on authors previous experience with the CourseVis research, and proposes some graphical representations that can be useful to gain some insights on the students of the course.
Damez et al. (2005)	TAFPA	Classification	Four new steps were expected as four questions were asked, but that made a notice that one user missed a question by double-clicking accidentally on the button "Show next question". It can lead to some mistakes to use the LCS.
Qasem A. Al.Radaideh (2006)	CRISP Classifier	Classification	The classification algorithms ID3, C4.5 and Naive Bayes correctly classification parentage accuracy rating is not so high.
Cristobal Romero (2009)	KEEL	Régression, classification and unsupervised Learning	Association rule mining has been used to provide new, important and therefore demand-oriented impulses for the development of new bachelor and master courses
C.Marquez-Vera (2010)	10 Fold Cross Validation using WEKA Tool	Classification	Rule induction algorithms such as JRip, NNge, OneR, Prism and Ridor; and decision tree algorithms such as J48, SimpleCart, ADTree, RandomTree and REPTree are used for experiment because these algorithms can be used directly for decision making and provides detailed classification results.
Dragan Gasevic (2015)	Online learning tools and video annotation tool	Classification and segmentation	Transition graphs are constructed from a contingency matrix in which rows and columns were all events logged by the video annotation tool.

Table 2: Techniques for Educational Data Mining

Authors	Techniques	Mining task	Findings
Cristobal Romero (2007)	Moodles	Classification	C4.5 and CART algorithms are simple for instructors to understand and interpret. GGP algorithms have a higher expressive power allowing the user to determine the specific format of the rules.
M. Ramasami (2010)	CHIAD model	Association	Features whose chi-square values were greater than 100 were given due considerations and the highly influencing variables with high chi-square values. These features were used for the CHAID prediction model construction.
Edin Osman begovie (2012)	L86 Classifier	Classification and Association	The Results shows that the naïve Bayes algorithms performance and accuracy level for decision tree is much more than that of the neural network method for decision tree classification.
Sujeet Kumar Yadav (2012)	C4.5,ID3,CART Algorithm based tool	Classification	ID3, C4.5 and CART machine learning algorithms that produce predictive models with the best class wise accuracy. Classifiers accuracy True positive rate for FAIL is 0.786 for ID3 and C4.5. The created model successfully classified.
Dorina Kabakchieva (2013)	CRISP-DM	Classification	The J48 classifier correctly classifies about 66% of the instance with 10-fold cross-validation testing and 66.59 % for the percentage split testing and produces. The achieved results are slightly better for the percentage split testing option.
Abeer Badr El Din Ahmed (2014)	Decission Tree (ID3 Algorithm)	Classification	For measurements of best attribute for a particular node in the tree Information Gain are used with attribute A, relative to a collection of sample S.
Xing Wanli (2015)	GP-ICRM for rule generation and work flow	Classification	Implemented EDM, theory and application to solve the problem of predicting student's performance in a CSCL learning environment with small datasets. Model for performance prediction is evaluated using a GP algorithm.
Ashwin Satyanarayana (2016)	Bootstrap Averaging	Classification and Clustering	Single Model: Decision trees (J48) was used for single filtering base model. Online Bagging: Implemented online bagging using Naive Bayes as the base model. Ensemble Filtering: The proposed algorithm uses the following classifiers: J48, RandomForest and Naive Bayes.

Dorina Kabakchieva implemented CRISP (Cross-Industry Standard Process) approach for Data mining model for non-property, freely available and application neutral standard for data mining projects. Author also discusses about decision tree classifier of NaiveBayes and BayesNet with J48 10 fold cross validation and J48 percentage split and identifies weighted average. Same J48 10 fold cross validation and J48 percentage split comparison testing was carried for K-NN Classifier (with k=100 and k=250) and OneR and JRip classifiers [25]. Abeer Badr El Din Ahmed uses decision tree method for predicting students' performance with the help of ID3 Algorithm [26]. Xing Wanli introduces student prediction messures by using different rules and uses gnetic operator for classification and evaluate offspring for analysing student participations [27]. Ashwin Satyanarayana uses multiple classifiers such as J48, NaiveBayes, and Random Forest for classifying students' prediction. He also uses K-means clustering algorithm for calculating similar cluster cancroids average in student cluster [28].

V. CONCLUSION

Educational data mining is the most valuable research area which makes society a better one by giving nice prediction techniques for academician, teachers and students. The papers discussed in this survey will give the detailed thought of educational data mining and core paths of EDM. The techniques and tools discussed in this survey will provide a clear cut idea to the young educational data mining researchers to carry out their work in this field. Also, this research work carried out on the areas which make data mining process with educational data mining in a batter way. Finally, it is confirmed that most of the classification algorithms perform in a better way of understating the current trends of EDM by the students as well as academicians.

References

[1] Sheard. J, Ceddia. J, Hurst. J & Tuovinen. J, "Inferring student learning behaviour from website interactions: A usage analysis", Journal of Education and Information Technologies, 2003, Vol: 8(3), pp. 245-266.
[2] Sheard. J, Ceddia. J, Hurst. J & Tuovinen. J, "Determining website usage time from interactions: Data preparation and analysis", Journal of Educational Technology Systems, 2003, Vol: 32(1), pp.101-121.
[3] Muehlenbrock, Martin.,"Automatic action analysis in an interactive learning environment", Proceedings of the 12th International Conference on Artificial Intelligence in Education. 2005, pp.452-455.
[4] Anuradha. C and T. Velmurugan, "A Data Mining based Survey on Student Performance Evaluation System", IEEE

Int. Conference on Computational Intelligence and Computing Research, 2014, pp. 452-455.

[5] Zaiane. O, & Luo. J, "Web usage mining for a better web-based learning environment", In Proceedings of conference on advanced technology for education, Banff, Alberta, 2001, pp. 60–64.

[6] Silva, D., & Vieira, M. "Using data warehouse and data mining resources for ongoing assessment in distance learning". In IEEE international conference on advanced learning technologies, Kazan, Russia, 2002, pp. 40–45.

[7] Shen, Ruimin, Fan Yang, and Peng Han. "Data analysis center based on e-learning platform", The Internet Challenge: Technology and Applications. Springer Netherlands, 2002, pp.19-28.

[8] Cristobal Romero, Sebastian Ventura, Paul de Bra & Carlos de Castro, "Discovering prediction rules in AHA! Courses", International Conference on User Modeling, Springer Berlin Heidelberg, 2003, pp.25-34.

[9] Tane, Julien, Christoph Schmitz, and Gerd Stumme, "Semantic resource management for the web: an e-learning application", Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters, ACM, 2004, pp. 1-10.

[10] Merceron, Agathe, and Kalina Yacef, "Tada-ed for educational data mining", Interactive multimedia electronic journal of computer-enhanced learning, 2005, Vol:7(1), pp: 267-287.

[11] Vanzin, Mariangela, Karin Becker, and Duncan Dubugras Alcoba Ruiz , "Ontology-based filtering mechanisms for web usage patterns retrieval", International Conference on Electronic Commerce and Web Technologies, Springer Berlin Heidelberg, 2005, pp. 267-277.

[12] Avouris, N., Komis, V., Fiotakis, G., Margaritis, M. and Voyiatzaki, E., "Logging of fingertip actions is not enough for analysis of learning activities", In 12th International Conference on Artificial Intelligence in Education, AIED 05 Workshop1: Usage analysis in learning systems, 2005, pp.1-8.

[13] Mazza. R , and Milani. C , "Exploring usage analysis in learning systems: Gaining insights from visualizations", Workshop on usage analysis in learning systems at 12th international conference on artificial intelligence in education, 2005, pp. 65-72.

[14] Mostow. J, Beck. J, Cen. H, Cuneo. A, Gouvea. E, & Heiner. C, "An educational data mining tool to browse tutor-student interactions: Time will tell", Proceedings of the Workshop on Educational Data Mining, National Conference on Artificial Intelligence. AAAI Press, 2005, pp. 15-22.

[15] Damez. M, Marsala. C, Dang. T, & Bouchon-Meunier. B, "Fuzzy decision tree for user modeling from human–computer interactions", In International conference on human system learning: Who is in control?,2005, pp.287–302.

[16] Bari. M, & Benzater. B, "Retrieving data from pdf interactive multimedia productions". In International conference on human system learning: Who is in control? ,2005, pp.321–330.

[17] Qasem A. Al-Radaideh, Emad M. Al-Shawakfa, Mustafa I. Al-Najjar, "Mining Student Data Using Decision Trees", The 2006 International Arab Conference on Information Technology, Jordan,2006, pp.1-5.

[18] Romero. C, Alcala-Fdez. J, Sanchez. L, Garcia. S, del Jesus. M. J, Ventura. S, Garrell. J. M,& Fernandez. J, "KEEL: a software tool to assess evolutionary algorithms for data mining problems." Soft Computing, 2009, Vol:13(3), pp. 307-318.

[19] Marquez-Vera. C, ROMERO. C, "Predicting School Failure Using Data Mining", Educational Data Mining, 2010,2011.

[20] Dragan Gasevic, "Let's not forget: Learning analytics areabout learning", Association for Educational Communications and Technology, 2015, Vol: 59(1), pp. 64-71.

[21] Romero. C, Ventura. S, Espejo. P. G, & Hervas. C, "Data mining algorithms to classify students", Educational Data Mining 2007, 2007, pp:1-10.

[22] Ramaswami. M, and R. Bhaskaran, "A CHAID based performance prediction model in educational data mining", 2010, arXiv preprint arXiv:1002.1144.

[23] Edin Osmanbegovic & Mirza Suljic , "DATA MINING APPROACH FOR PREDICTING STUDENT PERFORMANCE", Journal of Economics and Business, 2012, Vol: 10(1), pp. 3-12.

[24] Surjeet Kumar Yadav, Saurabh Pal, "Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification", World of Computer Science and Information Technology Journal (WCSIT), 2012, Vol: 2(2), pp. 51-56.

[25] Dorina Kabakchieva, "Predicting Student Performance by Using Data Mining Methods for Classification", Cybernetics and Information Technologies, 2013, Vol: 13(1), pp.61-72.

[26] Abeer Badr El Din Ahmed and Ibrahim Sayed Elaraby, "Data Mining: A prediction for Student's Performance Using Classification Method", World Journal of Computer Application and Technology, 2014, Vol: 2(2), pp. 43-47.

[27] Xing Wanli , Guo Rui, Petakovic Eva & Goggins Sean, "Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, Educational data mining and theory", Computers in Human Behavior, 2015, Vol: 47, pp. 168–181.

[28] Ashwin Satyanarayana, Gayathri Ravichandran, "Mining Student data by Ensemble Classification and Clustering for Profiling and Prediction of Student Academic Performance", 2016 ASEE Mid-Atlantic Section Conference, 2016.