A Survey on Educational Data Mining in Field of Education

Dr. P. Nithya, B. Umamaheswari, A. Umadevi
Department of CS, PSG CAS, Coimbatore.
Tamil Nadu, India.

Abstract: Educational Data Mining refers to techniques, tools, and research designed for automatically extracting meaning from large repositories of data generated by or related to people' learning activities in educational settings.[1] It is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in.[2]

Keywords- Data Mining, Educational Data Mining, Prediction, Clustering, Relationship Mining.

I. INTRODUCTION

While the analysis of educational data is not itself a new practice, recent advances in educational technology, including the increase in computing power and the ability to log fine-grained data about students' use of a computer-based learning environment, have led to an increased interest in developing techniques for analyzing the large amounts of data generated in educational settings. This interest translated into a series of EDM workshops held from 2000-2007 as part of several international research conferences.[3]In 2008, a group of researchers established what has become an annual international research conference on EDM, the first of which took place in Montreal, Canada.[4]

As interest in EDM continued to increase, EDM researchers established an academic journal in 2009, the Journal of Educational Data Mining, for sharing and disseminating research results. In 2011, EDM researchers established the International Educational Data Mining Society to connect EDM researchers and continue to grow the field.

With the introduction of public educational data repositories in 2008, such as the Pittsburgh Science of Learning Centre's (PSLC) Data Shop and the National Center for Education Statistics (NCES), public data sets have made educational data mining more accessible and feasible, contributing to its growth.[5]

II. GOALS FOR EDUCATIONAL DATA MINING IN EDUCATIONAL FIELD

Baker and Yacef[6] describes the following four goals of EDM:

1) Predicting student's future learning behavior

2) Discovering or improving domain models

3) Studying the effects of educational support

4) Advancing scientific knowledge about learning and learners

Predicting student's future learning behavior - With the use of student modeling, this goal can be achieved by creating student models that incorporate the learner's characteristics, including detailed information such as their knowledge, behaviors and motivation to learn.

Discovering or improving domain models - Through the various methods and applications of EDM, discovery of new and improvements to existing models is possible.

Studying the effects of educational support - It can be achieved through learning systems.

Advancing scientific knowledge about learning and learners - By building and incorporating student models, the field of EDM research and the technology and software used.

III. METHODS OF EDUCATIONAL DATA MINING

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

1. Prediction: Ryan S. J. d. Baker has given a detail explanation of prediction in his paper. He mentioned that " In prediction, the goal is to develop a model which can infer a single aspect of data from some combination of other aspects of data. If we study prediction extensively then we get three types of prediction: classification, regression and density estimation. In any category of prediction the input variables will be either categorical or continuous. In case of classification, the categorical or binary variables are used, but in regression continuous input variables are used. Density estimation can be done with the help of various kernel functions.

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

3. Relationship Mining: Relationship mining generally refers to contrive new relationships between variables. It can be done on a large data set, having a no of variables. Relationship mining is an attempt to discover the variable which is most closely associated with the specified variable. There are four types of relationship mining: association rule mining, correlation mining, sequential pattern mining and causal data mining. Association data mining is based on if- then rule that is if some particular set of variable value appears then it generally have a specified value. In correlation mining, the linear correlations are discovered between variables. The aim of sequential pattern mining is to extract temporal relationships between variables.

4. Discovery with Models: it includes the designing of model based on some concepts like prediction, clustering and knowledge engineering etc. This newly created model's predictions are used to discover a new predicted variable.

5. Distillation of Data for Human Judgment: There are two objectives for human judgment for which distillation of data can be done: Identification and Classification. As per phenomenon of identification, data is represented in a way that human can easily recognize the well specified patterns.

IV. TRENDS IN EDUCATIONAL DATA MINING METHODS

Romero and Ventura's survey of Educational Data Mining research from 1995 to 2005, 60 papers was stated that developed EDM methods to answer research questions of applied interest. Relationship mining methods of various types were the most prominent type of EDM research between 1995 and 2005. 43% of papers in those years involved relationship mining methods. Prediction was the second most prominent research area, with 28% of papers in those years involving prediction methods of various types. Human judgment/exploratory data analysis and clustering followed with 17% and 15% of papers [7]. The full distribution of methods across papers is shown in Figure 1.

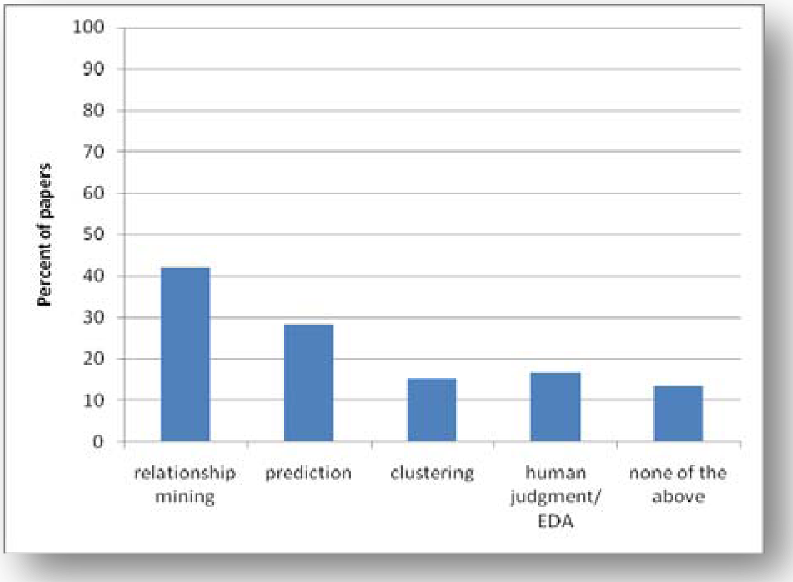


Figure-1.

Whereas relationship mining was leading between 1995 and 2005, in 2008- 2009 it slipped to fifth place, with only 9% of papers involving relationship mining. Prediction, which was in second place between 1995 and 2005, moved to the leading position in 2008-2009, representing 42% of EDM2008 papers. Human judgment/exploratory data analysis and clustering remain in approximately the same position in 2008-2009 as 1995-2005, with (respectively) 12% and 15% of papers [2].

V. APPLICATIONS OF DATA MINING IN EDUCATION FIELD

Some applications of Data Mining in education sector are given below:

A. Analysis and Visualization of Data

It is used to highlight meaningful information and support decision making. In the educational sector, for example, it can be helpful for course administrators and educators for analyzing the usage information and students' activities during course to get a brief idea of a student's learning. Visualization information and statics are the two main methods that have been used for this task. Statistical analysis of educational data can give us information like where students enter and exit, the most important pages students browse, how many number of downloads of e-learning resources, how many number of different type of pages browsed and total amount of time for browsing of these different pages. It also provides information about reports on monthly and weekly user trends, usage summaries, how much material students will study and the series in which they study topics, patterns of studying activity, timing and sequencing of activities. Visualization uses graphical methods to help people in understanding and analyzing data. There are number of studies related to visualization of different educational data such as patterns of hourly, daily, seasonal and annual user behavior on online forums.[8]

B. Predicting Student Performance

In student performance prediction, we predict the unknown value of a variable that defines the student. In educational sector, the mostly predicted values are student's performance, their marks, knowledge or score.

Classification technique is used to combine individual items based upon quantitative traits or based upon training set of previously labeled items. Student's performance prediction is very popular application of DM in education sector. Different techniques and models are applied for prediction of student's performance like decision trees, neural networks, rule based systems, Bayesian networks etc. This analysis is helpful for someone in predicting student's performance i.e. prediction about student's success in a course and prediction about student's final grade on the basis of features taken from logged data. Several regression techniques are used for prediction of student's marks such as linear regression to predict student's academic performance, stepwise linear regression to predict time spent by a student on a learning page and multiple linear regression for identification of variables that are helpful for predicting success in colleges courses and for prediction of exam results in distance education courses. [9][10][11][13][14]

C. Enrolment Management

Enrolment management is frequently used in higher education to explain well-planned strategies and ways to shape the enrolment of college to meet planned goals. It is an organizational concept and also a systematic set of activities designed to allow educational institutions to exert more influence over student's enrolments. Such practices often include retention programs, marketing, financial aid awarding and admission policies.[14]

D. Grouping Students

In this case groups of students are created according to their customized features, personal characteristics, etc. These clusters/groups of students can be used by the instructor/developer to build a personalized learning system which can promote effective group learning. The DM techniques used in this task are classification and clustering. Different clustering algorithms that are used to group students are hierarchical agglomerative clustering, K-means and model-based clustering. A clustering algorithm is based on large generalized sequences which help to find groups of students with similar learning characteristics like hierarchical clustering algorithm which are used in intelligent e-learning systems to group students according to their individual learning style preferences discriminating features and external profiling features. [8]

E. Predicting Students Profiling

Data mining can help management to identify the demographic, geographic and psychographic characteristics of students based on information provided by the students at the time of admission. Neural networking technique can be used to identify different types of students. [11]

F. Planning and scheduling

Planning and scheduling is used to enhance the traditional educational process by planning future courses, course scheduling, planning resource allocation which helps in the admission and counseling processes, developing curriculum, etc. Different DM techniques used for this task are classification, categorization, estimation, and visualization. Decision trees, link analysis and decision forests have been used in course planning to analyze enrollee's course preferences and course completion rates in extension education courses. Educational training courses have been planned through the use of cluster analysis, decision trees, and back-propagation neural networks in order to find the correlation between the course classifications of educational training. Decision trees and Bayesian models have been proposed to help management institutes to explore the probable effects of changes in recruitments, admissions and courses. [8]

G. User Modeling

User modeling encompasses what a learner knows, what the user experience is like, what a learner’s behavior and motivation are, and how satisfied users are with online learning. EDM can be applied in modeling user knowledge, user behavior and user experience. [12]

H. Organization of Syllabus

Presently, organization of syllabi is influenced by many factors such as affiliated, competing or collaborating programs of universities, availability of lecturers, expert judgments and experience. This method of organization may not necessarily facilitate students' learning capacity optimally. Exploration of subjects and their relationships can directly assist in better organization of syllabi and provide insights to existing curricula of educational programs. One of the applications of data mining is to identify related subjects in syllabi of educational programs in a large educational institute. [14]

I. Detecting Cheating in Online Examination

Now a day's exams are conducted online remotely through the Internet and if a fraud occurs then one of the basic problems to solve is to know: who is there? Cheating is not only done by students but the recent scandals in business and journalism show that it has become a common practice. Data mining techniques can propose models which can help organizations to detect and to prevent cheats in online assessments. The models generated use data comprising of different student’s personalities, stress situations generated by online assessments, and common practices used by students to cheat to obtain a better grade on these exams. [14]

VI. APPLICATIONS OF ALGORITHMS IN EDUCATION MINING

Number of universities and students is increasing day by day; we think that data mining technology can help improving the education standard and consequently causing high ratio of successful candidate, low ratio of students' drop-out and maximizing education system efficiency. Following is a detail of the algorithms used in education mining.

A. C 4.5

A classifier system takes input from the cases described by values and attributes and output a classifier that can accurately predict classes of new cases. C 4.5 is a descendant of CLS and IDE, creates classifier and generated decision tree. It can also make classifier in most comprehensive rule-set forms.

B. Support Vector Machine (SVM)

Support Vector Machine (S.V.M) is considered an efficient tool to train data. If offers accurate methods among algorithms. SVM is the most worked upon algorithm for training purposes and a lot of research is still going on. SVM can find classification function in two- class learning tasks. The metric for "best" can be realized geometrically. It is good because of its generalization ability.

C. Apriori

A popular way to find frequent data item sets [14] is by comparing explosion. After data items are obtained then we can easily generate association rules.
There are different steps for that.

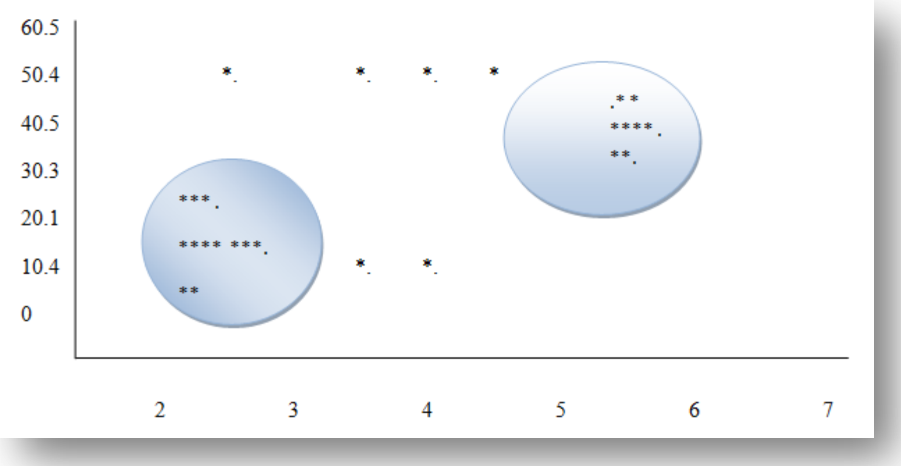
1- Generate Ck+1 for item sets of k+1

2- Calculate support

3- Put items for minimum support for fk+1

D. The Expectation Maximization Algorithm

It provides a flexible mathematical approach to modeling and clustering of data on randomly observed basis. This can be used to cluster continuous data. Expectation Maximization algorithm is used to model distribution of random phenomenal data.



EM clustering of Old Faithful Data

E and M parts EM algorithms give (ML) estimation of normal components. On (k+1) turn of EM, log of the data is taken.

E. Page Ranker

Page Ranker [15] was given forth by Brin Karruy Page in 1998. On this algorithm's basis they built Google, which has an excellent success ratio. It produces a static ranking of different web pages in sense that pager value is determined offline and does not depend on the online queries.

Pager Ranker formula:

1- A hyperlink point the value of the page is an implicit conveyance for authority. Thus more links a page receives more prestige it has.

2- Pages that go to I is also considered good

Page Rank algorithm given by Lawrence Page and Sergey Brin is in a lot of publications. It is as under…

PR(A) = (1-d) + d (pr (ti) / C (ti) + …..... + PR (Tn) /C (tn))

Where,

PR(A) is a Page Rank of Page A
PR(Tn) is Page Rank.

VII. CONCLUSION

Data mining is a tremendously vast area that includes employing different techniques and algorithms for pattern finding. The algorithms discussed in this paper are the ones used in education mining. These algorithms have shown a remarkable improvement in strategies like course outline formation, teacher student understanding and high output and turn out ratio. ICDM conference encourages employment and development of algorithms helpful in data mining. An appreciable research is still being done on various algorithms. I hope this review paper appreciates the current algorithm researchers and inspires the new ones to explore further.

References

[1] "EducationalDataMining.org". 2013. Retrieved 2013-07-15.

[2] BakerRSJd, Yacef K. The state of educational datamining in 2009: A review and future visions. J EduData Min 2009.

[3] C. Romero, S. Ventura. Educational Data Mining: A Review of the State-of-the-Art. IEEE Transaction on Systems, Man, and Cybernetics, Part C: Applications and Reviews. 40(**6**), 601-618, 2010.

[4] "http://educationaldatamining.org/EDM2008/" Retrieved 2013-09-04.

[5] Baker, Ryan. "Data Mining for Education" (PDF). oxford, UK: Elsevier. Retrieved 9 February 2014.

[6] Baker, R.S.; Yacef, K (2009). "The state of educational data mining in 2009: A review and future visions". JEDM-Journal of Educational Data Mining **1** (1): 2017.

[7] Romero, C. and Ventura, S. (2007) 'Educational data mining: A Survey from 1995 to 2005', Expert Systems with Applications (33).

[8] Monika Goyal and Rajan Vohra, "Applications of Data Mining in Higher Education", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012.

[9] Naeimeh Delavari, Somnuk Phon-Amnuaisuk, "Data Mining Application in Higher Learning institutions", Informatics in Education, Vol. 7, No. 1, 2008.

[10] Dina Abdulaziz Alhammadi, "Data Mining in Education- An Experimental Study", International Journal of Computer Applications Volume 62, No.15, January 2013.

[11] Dr. Mohd Maqsood Ali, “Role of data mining in education sector”, International Journal of Computer Science and Mobile Computing Vol. 2, Issue. 4, April 2013.

[12] S. Lakshmi Prabha, Dr.A.R.Mohamed Shanavas, “Educational data mining applications”, Operations Research and Applications: An International Journal (ORAJ), Vol. 1, No. 1, August 2014.

[13] Mrs. M.S. Mythili, Dr. A.R.Mohamed Shanavas, “An Analysis of students’ performance using classification algorithms”, IOSR Journal of Computer Engineering, Volume 16, Issue 1, January 2014.

[14] U. K. Pandey, and S. Pal, “A Data mining view on class room teaching language”, (IJCSI) International Journal of Computer Science Issue, Vol. 8, Issue 2, pp. 277-282, ISSN:1694-0814, 2011.

[15] Romero, C., Ventura, S. and De-Bra, P. “Knowledge Discovery with Genetic Programming for Providing Feedback to Courseware Authors, Kluwer Academic Publishers, Printed in the Netherlands, 30/08/2004.”