Relevance of Data Mining Techniques in Edification Sector

S. Anupama Kumar and M. N. Vijayalakshmi

***Abstract*—The various data mining techniques like classification, clustering and relationship mining can be applied on educational data to study the behavior and performance of the students. This study will enable the learner and teaching community to grow. These techniques can also be combined with specific discovery models and distillation of data for human judgment to enhance the learning community. This paper explores the various approaches and techniques of data mining which can be applied on Educational data to build up a new environment so as to give the new predictions on the data.**

***Index Terms*—Classification, clustering, discovery with models, EDM, prediction, relationship mining.**

I. Introduction

Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data. The process of converting raw data into useful information consists of a series of steps from data preprocessing to post processing of data mining results. Data mining has been applied in a great number of fields, including retail sales, bioinformatics, and counter-terrorism. In recent years, interest in the use of data mining has been increased in the field of education and termed as Educational Data Mining.

Educational data mining can be defined as “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”.

EDM is the process of transforming raw data compiled by education systems in useful information that could be used to take informed decisions and answer research questions [1].

Three objectives could be identified to use EDM as a technology in the field of education. They are listed as below:

1) *Pedagogic objectives* – To help the students to improve in academics, designing the content of the course in a better way etc
2) *Management objectives* - To optimize the organization and maintenance of educational infrastructures, areas of interest, more requested courses
3) *Commercial objectives* - It allows to make market segmentation and facilitates students recruitment in schools, high schools, colleges and universities

EDM is an emergent discipline on the intersection of data mining and pedagogy. On one hand, pedagogy contributes with the intrinsic knowledge of learning process. On the other hand, data mining adds the analysis and information modeling techniques. There are a number of data mining methods and patterns which help us achieve the above set objectives. The paper is mainly organized into two sections: Section II describing the various approaches of data mining in educational data mining and conclusion in Section III.

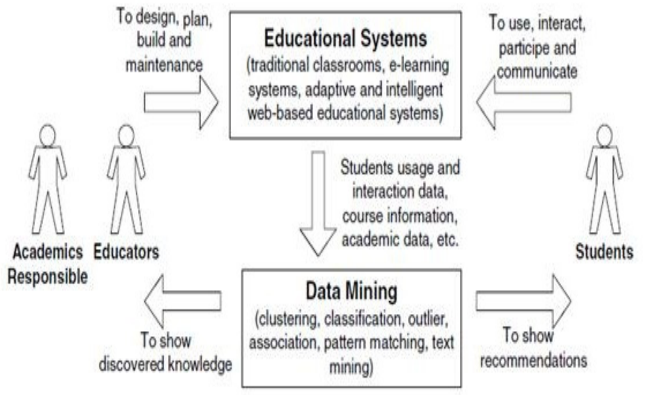
To design, plan, build and maintenance
Educational Systems
(traditional classrooms, e-learning systems, adaptive and intelligent web-based educational systems)
To use, interact, participe and communicate
Students usage and interaction data, course information, academic data, etc.
Academics Responsible
Educators
Students
Data Mining
(clustering, classification, outlier, association, pattern matching, text mining)
To show discovered knowledge
To show recommendations

Fig. 1. Data mining applications in the education sector [2]

II. Approaches of Data Mining in Educational Data

Data mining tasks are divided into two major categories. (i) Predictive task (ii) Descriptive task.

Data mining methods like prediction, clustering and relationship mining are hugely used in the field of marketing and finance. They are mainly used for classifying and forecasting the market. These methods can be efficiently applied on educational data to classify the students, institutions etc and also to predict the outcome of the student. These methods are often integrated with methods from the machine learning and data mining literatures to achieve meaningful information about the learners. Educational data can be either collected directly from the students or from the data which they submit in their institutions. Data can also be created depending upon their performance in the examination either internally or through board exams. There are a wide variety of methods popular within educational data mining. These methods include Prediction, clustering, relationship mining, discovery with models, and distillation of data for human judgment. The first three categories are largely common methods used in all the areas of data mining and the last two are special methods which are used within educational data mining.

A. Classification

Classification is used to develop a model which can infer a single aspect of the data (predicted variable) from some

Manuscript received September 07, 2012; revised November 18, 2012.

The authors are with the Department of M.C.A., R. V. College of Engineering, Bangalore, India (e-mail: kumaranu.0506@gmail.com, mnviju74@gmail.com).

DOI: 10.7763/IJMLC.2013.V3.260

combination of other aspects of the data (predictor variables). Prediction can be classified into: classification, regression, and density estimation. In classification, the predicted variable is a binary or categorical variable. Some popular classification methods include decision trees, logistic regression. And support vector machines. In regression, the predicted variable is a continuous variable. Some popular regression methods within educational data mining include linear regression, neural networks, and support vector machine regression. In density estimation, the predicted variable is a probability density function. Density estimators can be based on a variety of kernel functions, including Gaussian functions

Prediction methods can be used to study what features of a model are important for prediction, giving information about the underlying construct**.** This is a common approach in programs of research that attempt to predict student educational outcomes without predicting intermediate or mediating factors.. For example we can predict a student's outcome depending upon his previous scores in the test. Prediction methods can also be used to predict what the output value would be if the intermediate or mediating factors are considered. This may not be a right method because some of the mediating factor would definitely affect the outcome of the result. For example if the mediating factor is the students' interest towards the subject or his absence in the internal examination, the prediction of his outcome would not be always correct.

B. Clustering

Clustering is used to find data points that naturally group together, splitting the full data set into a set of clusters. Clustering is particularly useful in cases where the most common categories within the data set are not known in advance. If a set of clusters is optimal, within a category, each data point will in general be more similar to the other data points in that cluster than data points in other clusters. Clusters can be used to group schools to investigate the similarities and differences between schools, students can be clustered together to study the differences in their behavior etc. [3]

C. Association Rule Mining

Relationship mining is used discover relationships between variables, in a data set with a large number of variables. This may take the form of attempting to find out which variables are most strongly associated with a single variable of particular interest, or may take the form of attempting to discover the strongest relationships between any two variables. Broadly, there are four types of relationship mining: Association rule mining, Correlation mining, Sequential pattern mining, and Causal data mining. In association rule mining, the goal is to find if-then rules of the form that if some set of variable values is found, another variable will generally have a specific value. A rule might be found of the form to find when student is frustrated or when the student has stronger goal of learning than goal of performance or when a student frequently asks for help. Correlation mining is used to find the positive or negative linear correlations between variables. Sequential pattern mining is used to find temporal associations between events. It can be used to determine what path of student behaviors leads to an eventual learning event of interest. Causal data mining is used to find whether one event was the cause of another event either by analyzing the covariance of the two events or by using information about how one of the events was triggered.

D. Discovery with Models

Discovery with a model is a phenomenon which is developed via prediction, clustering, or in some cases knowledge engineering. In the prediction case, the created model's predictions are used as predictor variables in predicting a new variable. Analysis of complex constructs such as gaming the system within online learning have generally depended on assessments of the probability that the student knows the current knowledge component being learned [4] These assessments of student knowledge have in turn depended on models of the knowledge components in a domain, generally expressed as a mapping between exercises within the learning software and knowledge components. In the relationship mining case, the relationships between the created model's predictions and additional variables are studied. This can enable a researcher to study the relationship between a complex latent construct and a wide variety of observable constructs. Often, discovery with models leverages the validated generalization of a prediction model across contexts. For example, the use of predictions of gaming the system across a full year of educational software data to study whether state or trait factors were better predictors of how much a student would game the system [4]. Generalization in this pattern relies upon appropriate validation that the model accurately generalizes across contexts.

E. Distillation of Data for Human Judgment

In some cases, human beings can make inferences about data, when it is presented appropriately, that are beyond the immediate scope of fully automated data mining methods. The methods in this area of educational data mining are information visualization methods. These methods are different than those most often used for other information visualization problems in their specific structure, and the meaning embedded within that structure, often present in educational data. Data is distilled for human judgment in educational data mining for two key purposes: identification and classification. When data is distilled for identification, data is displayed in such a way it enable a human being to easily identify well-known patterns that are nonetheless difficult to formally express. For example[3], one classic educational data mining visualization is the learning curve, which displays the number of opportunities to practice a skill on the X axis, and displays performance (such as percent correct or time taken to respond) on the Y axis. A curve with a smooth downward progression that is steep at first and gentler later indicates a well specified knowledge component model. A sudden spike upwards, by contrast, indicates that more than one knowledge component is included in the model .Alternatively; data may be distilled for human labeling, to support the later development of a prediction model. In this case, sub-sections of a data set are displayed in visual or text format, and labeled by human coders. These labels are then generally used as the basis for the development of a predictor. This approach has been shown to speed the development of prediction models of complex phenomena such as gaming the system by around 40 times, relative to prior approaches for collecting the necessary data

TABLE I: SUMMARY OF METHODS AND APPLICATIONS.

Name of the Method	Applications in EDM
Classification	Detecting student behaviors, Developing domain models; Predicting and understanding student educational outcomes
Clustering	Discovery of new student behavior patterns; Investigating similarities and differences between schools
Relationship Mining	Discovery of curricular associations in course, sequences
Discovery with Models	Discovery of relationships between student behaviors, and student characteristics or contextual variables; Analysis of research question across wide variety of contexts
Distillation of Data for Human Judgment	Human identification of patterns in student learning, behavior, or collaboration; Labeling data for use in later development of prediction model

III. CONCLUSION

The various data mining algorithms which are applied in business can also be effectively used in the field of education for the betterment of student performance and the institution. The Table I summarize the different methods and the key areas where it can be applied.

REFERENCES

[1] C. Heiner, R. Baker, and K. Yacef, *Proceedings of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems*, Taiwan, 2006.

[2] K. H. R. Anushka and P. Msit, "Data mining applications in the education sector," Carnegie Mellon University, 2010.

[3] R. S. J. D. Baker, "Data mining in education," *Journal of Educational Data Mining*, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. 2009.

[4] R. S. J. D. Baker and P. B. Peterson, "International encyclopedia of education (3rd edition)," *Journal of Educational Data Mining*, *Elsevier- Data Mining for Education,* Oxford, UK, 2009.

[5] S. Ayesha, T. Mustafa, and A. R. Sattar, M. I. Khan, "Data mining model for higher education system," *European Journal of Scientific Research*, vol. 43, no. 1, pp. 24-29, 2010.

[6] C. Romero, S. Ventura, P. G. Espejo, and C. Hervás, "Data mining algorithms to classify students," *Journal of Educational Data Mining*, 2010

Mrs. S. Anupama Kumar has 13 years of teaching experience. She has completed her Master of Philosophy from Alagappa University and her Masters from Bharathidasan University. She has published research papers in national and international conferences / journals. Her research interests are in the area of data mining and artificial intelligence. She is a member of IAENG and IACSIT.

Dr. M. Vijayalakshmi had completed her PhD from Mother Teresa Women's university, Kodaikanal in 2010. She has 13 years of teaching experience and 6 years of Research experience. She is a recognised research guide in VTU and Prist University. She has published many research papers in the national and international conferences / journals. She has got many research projects to her credit funded by different agencies. Her research interests are Pattern recognition, data mining, neural networks, Image Processing. She is a life member of ISTE, CSI, IACSIT.