ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

Techniques Used for Mining Data in Educational System

Kapila Kundu

Research Scholar, Department of Computer Science and Engineering, GJUS&T University, Hisar, India

ABSTRACT: Educational Data Mining concerned with developing methods for exploring the unique types of data that come from educational settings. The obtained result is used to better understand students and the settings which they learn in. Various methods are used under the umbrella of educational data mining. Goal of this paper is to describe the all method used in educational data mining specially included in the taxonomy of Baker and Romero &Ventura. Prediction, clustering and association rule mining are the most common methods applied in the literature.

KEYWORDS: Data mining, Educational data mining, classification, prediction, clustering.

I. INTRODUCTION

Educational data mining has emerged as an independent research area in recent years, culminating in 2008 with the establishment of the annual International Conference on Educational Data Mining, and the Journal of Educational Data Mining [1,2]. Educational Data Mining 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 [3]. The obtained knowledge can then be used to offer suggestions to the academic planners in higher education institutes to enhance their decision making process, to improve students' academic performance, to decrease failure rates, to understand students' behaviour in a better way, to assist instructors, to improve teaching, and to construct regression models and decision trees to predict student performance in terms of their grades or percentage [4].

In last decade, Educational data mining can be applied to data coming from two types of educational systems: traditional classroom and distance education. It is necessary to deal separately with the application of data mining techniques in each type due to the fact that they have different data sources and objectives [romero 95-05]. Nowadays, there is a great variety of scholastic systems/ environments such as: the traditional classroom, e-learning, LMS, adaptive hypermedia (AH) educational systems, ITS, tests/quizzes, texts/contents, and others such as: learning object (LO) repositories, concept maps, social networks, forums, scholastic game environments, virtual environments, ubiquitous computing environments, etc [5].

Apart from the environment from where data is coming out, methods and techniques are also important. Educational data mining methods are drawn from a variety of literatures, including data mining and machine learning, psychometrics and other areas of statistics, information visualization, and computational modeling [3]. Romero and Ventura [2007] categorize work in educational data mining into the following categories [6]:

- Statistics and visualization
- Web mining
- Clustering, classification, and outlier detection
- Association rule mining and sequential pattern mining
- Text mining

In other viewpoint on educational data mining is given by Baker [2009], which classifies work in educational data mining as follows:

- Prediction
  - Classification
  - Regression

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

- Density estimation

- Clustering
- Relationship mining
  - Association rule mining
  - Correlation mining
  - Sequential pattern mining
  - Causal data mining
- Distillation of data for human judgment
- Discovery with models

Some techniques like clustering, prediction and association rule mining are common in both the taxonomy. These are the categories which are frequently seen in the literature but some are specific. Statistics and visualization from the Romero and Ventura taxonomy has had a prominent place both in published EDM research and in theoretical discussions of educational data mining [3]. The fifth category of Baker's Educational Data Mining taxonomy is perhaps the most unusual category, from a classical data mining perspective. In discovery with models, a model of a phenomenon is developed through any process that can be validated in some fashion (most commonly, prediction or knowledge engineering), and this model is then used as a component in another analysis, such as prediction or relationship mining [3].

II. TECHNIQUES USED IN THE DOMAIN

There are numerous techniques used in the domain of educational data mining. Some techniques are from the data mining domain like classification and clustering etc. and others like regression are used specifically in educational data mining. All the techniques used in the domain of educational data mining are described in this paper.

- Prediction

In prediction, the goal is to develop a model which can predict a single variable from some combination of predictor variables. Prediction requires having labels for the output variable for a one data set, where a label represents information about the output variable's value [2]. Prediction can be applied on two types of data. The prediction method which is applied on categorical data is called classification and which is applied on continuous data is called regression. Prediction is applied in two aspects. One is, to predict the important feature of the model. In other aspect, prediction is used to find the target value (previously known) for the model [2].

- Classification

In classification, the predicted variable can either be a binary or categorical variable. Classification is a supervised learning technique which classifies the data on the basis of known class variable. Some popular classification methods include decision trees, random forests, and decision rules. Classifiers are typically validated using cross-validation, where part of the data set is repeatedly and systematically held out and used to test the goodness of the model. Cross-validation should be conducted at multiple levels. It is typically standard to cross-validate at the student level in order to ensure that the model will work for new students. Some common metrics are used for cross-validation in classification i.e. AUC, kappa, precision, recall, and accuracy [1, 8]. Some popular classification methods include decision trees, logistic regression, k-nearest neighbour, Multilayer Perceptron and support vector machines [2].

- Regression

Regression deals with continuous data. It is used to find out the relationship between the predicted (dependent) variable and other independent variables. The most popular regression method is linear regression, and logistic regression. Another method, regression tree is also fairly popular. Regressors can be validated using Kappa, linear correlation or root mean squared error (RMSE) [1, 8]. Some popular regression methods are linear regression, neural networks, and support vector machine regression [2].

- Clustering

Clustering is unsupervised data mining technique used to classify data into different groups based on their inherent characteristics. Many researchers have applied clustering in student performance modelling, e.g., to divide the students of interest into different groups to look at the common characteristics of the target groups under consideration like the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

students who have failed, the student who have achieved average score and the students who have done very well [1]. Clustering algorithms can either start with no prior hypotheses about clusters in the data (such as the k-means algorithm with randomized restart), or start from a specific hypothesis, possibly generated in prior research with a different data set (using the Expectation Maximization algorithm to iterate towards a cluster hypothesis for the new data set). A clustering algorithm can postulate that each data point must belong to exactly one cluster (such as in the k-means algorithm), or can postulate that some points may belong to more than one cluster or to no clusters (such as in Gaussian Mixture Models) [2].

- Density estimation

Density estimation, the predicted variable is a probability density function. Density estimators can be based on a variety of kernel functions, including Gaussian functions. For each type of prediction, the input variables can be either categorical or continuous; different prediction methods are more effective, depending on the type of input variables used [1,2].

- Relationship mining

In relationship mining, the goal is to 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 which relationships between any two variables are strongest. Broadly, there are four types of relationship mining: association rule mining, correlation mining, sequential pattern mining, and causal data mining. Relationships found through relationship mining must satisfy two criteria: statistical significance, and interestingness [1, 2].

- Association rule mining

Association means relationship or togetherness or connection of objects. Association rule mining indicates associated relationship between the set of objects/items. An example of an association could be that, 90% of the people who buy cookies also buy milk (60% of all grocery shoppers buy both) [1].

- Correlation mining

In correlation mining, the goal is to find linear correlations between variables. It can be positive correlation or negative correlation [2].

- Sequential pattern mining

In sequential pattern mining, the goal is to find temporal associations between events – for example, to determine what path of student behaviours leads to an eventual learning event of interest [2].

- Causal data mining

In causal data mining, the goal is to find whether one event (or observed construct) was the cause of another event (or observed construct), either by analyzing the covariance of the two events [2].

- Distillation of data for human judgment

Distillation of data for human judgement means to infer the meaning and structure of the data with the help of visualization. This visualization is different from the one which is used in educational data mining as a method. 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 ways that enable a human being to easily identify well-known patterns that are nonetheless difficult to formally express. For example, 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 [2].

- Discovery with models

In discovery with a model, a model of a phenomenon is developed via prediction, clustering, or in some cases knowledge engineering (within knowledge engineering, the model is developed using human reasoning rather than automated methods). This model is then used as a component in another analysis, such as prediction or relationship mining. In the prediction case, the created model's predictions are used as predictor variables in predicting a new variable [2].

- Visualization and Statistics

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

Student's usage statistics are used in student learning system, although they are usually not considered as data mining techniques. Formal statistical inference is driven from the data and tested against the model. Data mining, in contrast, is discovery driven in the sense that the hypothesis is automatically extracted from the data [6, 9].

- Wed mining

Web mining is data mining techniques used to extract knowledge from web data. There are three main web mining categories from the used data viewpoint: web content mining is the process of extracting useful information from the contents of web documents; web structure mining is the process of discovering structure information from the web; and web usage mining (WUM) that is the discovering of meaningful patterns from data generated by client–server transactions on one or more web localities [6, 9].

III. CONCLUSION

Educational data mining has emerged as an independent research area in recent years, culminating in 2008 with the establishment of the annual International Conference on Educational Data Mining, and the Journal of Educational Data Mining. There are numerous techniques used in the domain of educational data mining. Some techniques are from the data mining domain like classification and clustering etc. and others like regression are used specifically in educational data mining. All the methods used under the umbrella of educational data mining are described in this paper. Some methods like classification, regression, clustering and association rule mining are mostly used by the researcher.

REFERENCES

[1] J. Han, M. Kamber, "*Data Mining: Concepts and Techniques"*, Morgan Kaufmann Publishers, 2006.
[2] R. Baker, "Data mining for education," In *International Encyclopaedia of Education*, B.McGaw, P. Peterson, and E. Baker, Eds., 3rd ed. Oxford, U.K.: Elsevier, 2010.
[3] R. S. J. D. Baker, & K. Yacef, "The state of educational data mining in 2009: A review and future vision", *Journal of Educational Data Mining*, vol. 1, issue 1, pp. 1–15, 2009.
[4] E. Osmanbegovic, and M. Suljic "Data Mining Approach for Predicting Student Performance", In Economic Review - Journal of Economics and Business, Vol. X, Issue 1, 2011.
[5] Anoopkumar M, A. M. J. Md. Zubair Rahman, "A Review on Data Mining Techniques and Factors Used in Educational Data Mining to Predict Student Amelioration", In International conference on Data mining and Advanced Computing (SAPIENCE), 2016 , pp. 1-12.
[6] C. Romero, & S. Ventura, "Educational data mining: a survey from 1995 to 2005", *Expert Systems with Applications*, vol. 33, issue 1, pp. 135–146, 2007.
[7] J. Jacob, K. Jha, P. Kotak, S. Puthran, "Educational Data Mining Techniques and their Applications", *In International Conference on Green Computing and Internet of Things (ICGCloT*, 2015, pp. 1344-1348.
[8] R. Pelanek, "Metric for evaluation of student model", *Journal of Educational Data Mining*, vol. 7, issue. 2, 2015.
[9] C. Romero, & S. Ventura, "Educational data mining: a review of the state of the art", *IEEE Transactions on Systems, Man, and Cybernetics, part C: Applications and Reviews*, vol. 40, issue 6, pp. 601–618, 2010.
[10] C. Romero, S. Ventura, M. Pechenizkiy, & S. J. d. R. Baker (Eds.), "*Handbook of educational data mining, data mining and knowledge discovery series"*, Florida: Chapman & Hall/CRC, 2010.
[11] R. S. J. D. Baker & P. S. Inventado, "Educational Data Mining and Learning Analytics", *Learning Analytics: from research to Practice*, pp. 61-75, 2014.