International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 02 Issue: 09 | Dec-2015 www.irjet.net p-ISSN: 2395-0072

MINING EDUCATIONAL DATA USING DATA MINING TECHNIQUES AND ALGORITHMS –A REVIEW

[1]S. Padmapriya, Dr. L. Jayasimman[2], Dr. Nisha Jebaseeli[3], B. Senthil Kumar[4]

[1] S. Padmapriya Assistant Professor, Department of CS, Srimad Andavan Arts & Science College, India
[2] Jayasimman Assistant Professor, Department of CS, Srimad Andavan Arts & Science College, India
[3]Dr. Nisha Jebaseeli, Assistant Professor, Department of CS, Bharathidasan University constituent College, India
[4]B. Senthil Kumar Former Principal, Department of Animation, ES Pan Asian College, India

---

Abstract – ***This research paper discusses various algorithms of data mining in educational data mining which has greater impact to improve the students' academic performance and experience. Students' satisfaction and their academic performance plays a vital role in the success of any educational communication and educational institution. . Improving the performance and quality of education in an educational environment is one of the significant issues in educational communication. Prediction of students' exam performance depends upon so many factors which are complex to evaluate but factors such as students' personal, social and psychological differences are very important to predict the accuracy in evaluation. So, developing the accurate performance monitoring, alarming and evaluating system is the need of the hour. Students' performance monitoring system helps to identify the students who are all facing the risk of failures. Also, it helps the academicians to take necessary actions and steps to educate the students to improve their performance. Data mining plays a vital role in the educational monitoring, evaluation, management to mine the raw educational data. This research paper is also discusses various data types in order to study its application in educational institutions***

***Key Words:** Data mining, Educational data mining, Students' performance, Educational communication, Classification, Clustering*

1. INTRODUCTION

Educational data mining is one of the fruitful ways to analyse the students' satisfaction level in a learning environment [1]. Particularly in the web learning environment, it is one of the best way to solve the problems related to teaching and learning. Data mining finds useful information hidden in large volumes of educational data such as students' class attendants' data, personal bio-data, geographical area details, learning styles, learning system usage patterns, mobile device usage patterns, educators details, health information, students' attitude details, heredity, etc. which need to be interpreted into useful information. So it also called KDD-Knowledge Discovery in Databases. KDD deals with huge amounts of data which are kept in the database. Data mining is the analysis of data and the use of software techniques for finding hidden patterns and regularities. Knowledge discovery from the large data set has many processes such as processing the data, cleaning the data, integrating the data, selecting the data and interpreting the data [2]. It includes the processes such as pattern evaluation and knowledge representation. Data cleaning is the initial phase in which irrelevant data are removed from the large data. The increase in necessity of finding pattern from huge data is improved by means of data mining algorithms and techniques.

Data mining can be applied in many fields like genetics, software engineering, educational technology, business, sales, forensic science, biotechnology, etc. This paper study examines the educational mining using a case study from dataset from students' behaviour. It explains how and what data could be collected how it should be processed. And the result of the analysis will make the educational domain people to be benefited. Clustering and classification based on the students' behaviour patterns is one of the effective ways to predict the accuracy. It gives us meaningful information based on various evaluation factors.

Types of data:

Data is classified into

- Stream data
- Spatial, temporal, spatiotemporal and multimedia data
- Text, web and unstructured data
- Visual data

Stream data:

Stream data refers to vast volume and dynamically changing data like scientific and engineering. Satellite data flow, web click streams and network flow are some example of stream data. It contains multi-dimensional features. Stream data analysis counts the approximate the frequency from the infinite data.

Spatial, temporal, and spatial-temporal and multimedia data:

Spatial data mining is the method of processing hidden useful information from large spatial data. Generally scientific and engineering data has space, time and multimedia data such as color image, audio and video. Weather reports, google map reports, YouTube information, satellite image reports, spatial and digital data are the example of this kind of data.

Web, text, and unstructured data:

Digital libraries, research literatures, office automation systems, biological information, computer aided design and information, semi structured and unstructured information have large volume of data. Text data mining includes information extraction, topic tracing, categorization, clustering, summarizing, linkage of concepts, etc.

Visual data:

Charts, graphs, histograms, box plots, scatter plots, pictures, X-ray filmed visuals contains numerous data to analysis. Visualization tools like DataScope are very useful to extract useful information. Generally this type of visualization tools classifies the data using hierarchical and geographical techniques.

2. Literature Review:

Oyelade considered 79 students record set and applies k-means clustering algorithm to cluster the students' academic performance Graded Point Average (GPA) they performed different cluster based on different cluster size [2]. This analysis clearly shows the grade level of the students. And it helps the mentors to pay attention to the poor scorer. This analysis helps any educational institution to identify the poor scorer and makes them to perform better. Durairaj and Vijitha used k-means clustering and Naviebayes clustering techniques to cluster the students based on learning behaviour and their academic performance they used students semester marks in different subjects they evaluate the performance based on. Certain parameters like FP rate, TP rate, Recall F-measure, precision and ROC area. They also did the comparison between the decision stump tree technique and Naviebayes algorithm and the results shows that the Naviebayes produces accurate result then the stump tree technique [3].

Abdul Hamid M. Ragap underwent a study on educational Data mining. They used various parameters such as GPA, Gender, Race, Family income, Hometown, University entry mode, etc. for analysing the academic performance. They used NBC to the selected parameters to identify the useful pattern in SAP. They applied NRC and decision tree in wekatool to predict. They purposed a framework for predicting the students' academic performance. They suggested that their system will be very much helpful to act like a warning system to predict the students' performance and necessary action plan can be taken to prevent failures in advance [4]. Rajesh Kumar Arora has undergone a study on the students' academic performance semester wise. They considered the data from III Semester to VII Semester because from this semester onwards the subjects are specialized. They used Tanagra tool for the analysis purpose. They applied k-means cluster & deterministic model for exploratory data analysis. Their work aims to help the academicians to improve the students' performance in semester wise [4].

Mining Educational Data to analyse Students Performance - Brijesh Kumar Beradwaj done a research work on the evaluation of students' performance using classification techniques and used decision tree method from VBS the dataset they used for this study was obtained Purvanchal University, Janupur. They used information like Seminar, Assignment, Attendance, Class test as the performance parameters. They proved that ID3 decision tree proves to be better to identify the students' performance [5].

Educational data mining:

Educational data mining includes machine learning and data mining techniques. Data related to students' machine usage level like keystroke level, eye movements, timing level, answering level, etc. are very important to find out the students' interest in computer assisted learning and its environment. If the data analysis construct the students' individual differences in a similar way, then it will be very easy to cluster them in a group [6]. This cluster data and interpretation of data will lead to find out the useful information in a better way. Even though, discovery and maintenance of data set in large educational data is complex, the availability of new technologies such as cloud computing can be utilized. Dynamic data set computation is more complex than the ordinary educational data. If the data set is updated, then it will form another type of large data set. It is more complex for computation and interpretation. There are various types of algorithms and methods available to find out the useful data from the educational data. They are classified into Association rule, Multi-level association rule, Genetic Algorithm, Fuzzy FP Tree Algorithm, etc.

Data mining: A prediction for performance improvements using classification:

Brijesh Kumar Bhardwaj conducted a study an improvement of students' performance based on Bayes classification techniques. They considered social, Psychological and Personal or Performance evaluation parameters [5]. The size of their data set was 300. For 5 colleges out of which two was an urban based, aided and women's college and the others two was rural based, aided, co-education & the information like academic, demographic of socio-economic were used for analysis purposes. They found that the students' senior secondary examination, medium of teaching, mother's qualification, and income are high potential variables which influences on the students' performance.

Clustering:

Clustering is the discipline focused at revealing groups of similar entities in data set. Cluster is the process of making a group of physical or abstract objects. Clustering is the most common unsupervised data mining method [7]. It can also be defined as a process of partitioning a set of data into a set of meaningful sub-classes, called clusters. It helps the users to understand the natural grouping or structure in a data set. Cluster analysis is used in data mining and is a common technique for statistical data analysis used in many fields of study, such as the medical & life sciences, behavioural & social sciences, engineering, and computer sciences [2].

A large number of clustering algorithms have been developed in a variety of domain for different kind of applications. There are some of the characters which are strongly affects the cluster analysis. They are High Dimensionality, Size, Capability, Attributes, Interpretability, Shape, etc., There are various clustering methods have been proposed and they are classified as partitioning methods [6], density based, grid based, model based method and hierarchical methods and so on. The main objective of clustering is to find data points that naturally group together and splitting the data into clusters. Clustering is useful when the common points within the data not known well in advance. Clustering algorithms can be started with a hypothesis or can be started without a known hypothesis. Clustering is a phase in which relevant data are grouped and irrelevant noisy data are removed [6].

Computation of Clustering:

Grid based-

It includes input into hyper rectangular cells, eliminating low density cell, and combines high density cells, etc.

Locality based- Local conditions (neighbouring data) are used to cluster the data.

Partitioning based-Distance is used to partitioning the objects in order to cluster the data [8].

Hierarchical based- Hierarchy can be followed to cluster bottom-up or top-down of data [1].

Block diagram:

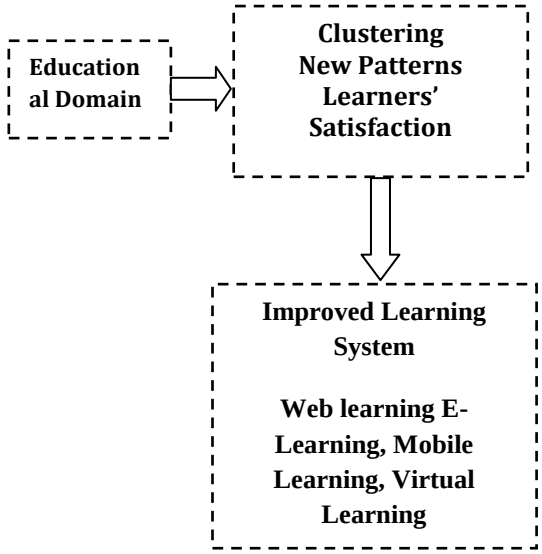

Figure 1: Flow diagram of educational data mining

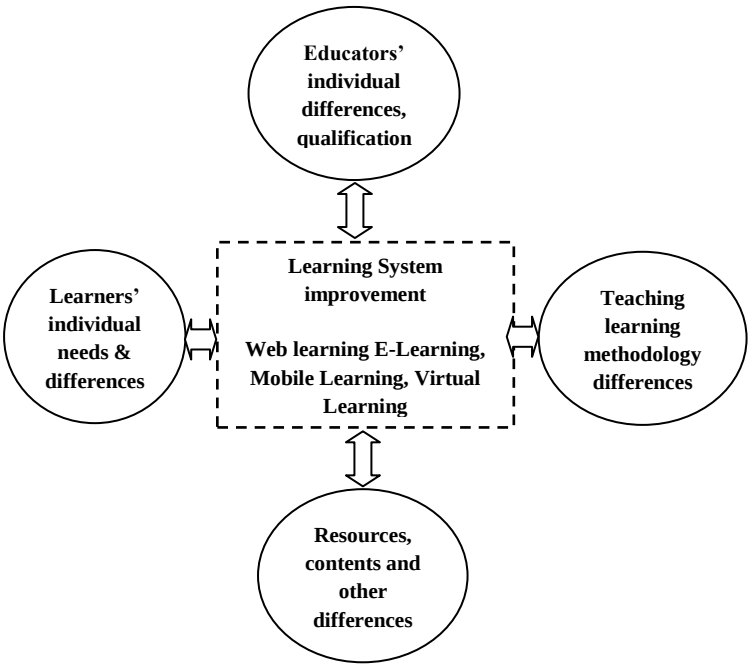

Figure 2: Flow diagram for system improvement

Conclusions:

The result of the review of the paper work indicates that the data mining technique is an efficient method for identifying the students' performance. It acts like an alarming system to predict the students' failure well in advance and make them to get succeed in their examinations. This kind of analysis will help the mentors to identify the students those who need special care to avoid failures. In this way the academicians and the academic planners achieves the high success rate of percentage. There are many technologies are available to predict the students' performance but this review work show that clustering is the efficient method among the existing methods. This paper concludes that the data mining application produces powerful and accurate results in prediction and clustering.

REFERENCES

[1] S. Padmapriya et al. "Enhancing web learning system using cluster algorithm based on cognition", IRJET, ISSN: 2395 -0056. Sep' 2015.

[2] Oyelade et al. "Application of k Means Clustering algorithm for prediction of Students Academic Performance", Computers and Society (cs.CY) International Journal of Computer Science and Information Security, IJCSIS, Vol. 7, No. 1, pp. 292-295, January 2010, USA.

[3] Durai raj et al. "Educational Data mining for Prediction of Student Performance Using Clustering Algorithm", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, ISSN.5987-5991

[4] Abdul Hamid, et al, **"**A Comparative Analysis of Classification Algorithms for Students College Enrollment Approval Using Data mining", IDEE, Proceedings of the 2014 Workshop on Interaction Design in Educational Environments ISBN: 978-1-4503-3034-3 Pages 106, ACM, 2014.

[5] Brijesh et al. "Data Mining: A prediction for performance improvement using classification", Computers and Society (cs.CY) International Journal of Computer Science and Information Security, IJCSIS, Vol. 9, No. 4, April 2011, pp 136-140, April 2011, USA.

[6] Madani,k., Lohi, M., A Comparative Study of Selected Classifiers with Classification Accuracy in User Profiling, Proceedings of the IEEE conference on Computer Science and Information Engineering, Vol.3, PP. 708-712, 2009.

[7] Boris et al, "clustering a Data Recovery approach", Computer Science and data analysis", Taylor and Francis group, CRC Press, USA, Edition 2, ISBN: 978-4398. 2013.

[8] E. Kolatch et al. "Clustering Algorithms for Spatial Databases: A Survey", downloaded from http://citeseerx.ist.psu.edu",