Journal of Analysis and Computation (JAC)
(An International Peer Reviewed Journal), www.ijaconline.com, ISSN 0973-2861
National Conference on: Research Challenges in Information Systems of Computer Science (NCRCISCS)".

AN OVERVIEW OF EDUCATIONAL ENVIRONMENT MISSING VALUES IN DATA MINING

M. Dharunya

Mphil Research Scholar, Department of Computer Science, NGM College, Pollachi, India

Dr. Antony Selvadoss Thanamani

Professor and Head Department of Computer Science, NGM College, Pollachi, India

ABSTRACT: *Education DataMining is a research field related with the utilization of data mining, statistics and machine learning to data produced from educational settings from colleges furthermore, universities. This paper is an investigation of the different ideas utilized in Education DataMining whose results are exceedingly invaluable to the understudies for their execution to be redesigned in a systematized way. Educational data mining (EDM) is an emerging discipline that focuses on applying data mining tools and techniques to educationally related data. Numerous different methodologies, for example, Clustering, decision tree algorithm and association rule mining have been discussed about which very spotlights on the furtherance of the student's academic betterment .Educational Data Mining (EDM) is a rising field investigating information in instructive setting by applying distinctive Data Mining (DM) systems. It gives natural information of teaching and learning process for effective education arranging. In this work focus around component, inquire about examples from 1998 of EDM highlighting its related Tools, Techniques and educational Outcomes. It likewise features the Challenges EDM. The aim of this paper is to discuss about the educational data mining and its component, goals and method.*

Keyword Terms: Data mining, Clustering, Educational Data Mining.

I. INTRODUCTION

Educational data mining is a surfacing field which investigates statistical data, machine learning and other data mining algorithms to find interesting examples with regards to educational database. The process of extracting important and useful information from large sets of data is called Data Mining. So far all the educational institutes have restricted themselves to predict only the academic performance of the students by considering internal, external marks and grades etc. However, it tends to be demonstrated that different expectation models can be utilized for anticipating the future subtleties like profession alternatives and plausibility for a youngster to get vicious in future. There are different mining procedures like Association rule mining, clustering and classification that serve this purpose. Through the proposed framework, it is recommended that different classifiers like ID3 (Iterative Dichotomizer) and C4.5 algorithm can be utilized to give career suggestion and to locate the rough conduct in a student. Finally a comparative study is made with the different decision trees to find a suitable algorithm. EDM has emerged as a research area in recent years aimed at analyzing the unique kinds of data that arise in educational settings to resolve educational research issues. Data mining is a term which refers to the processing or handling of huge amount of database called as big data which can't be prepared effectively. These days, since the amount of data accessible in any institutions is so high, it is very hard to process them. Because of this, the challenges faced are analyzing the data, catching, visualizing the data, making secure condition for the prepared data. Analysis of data collections is tedious to the point that it needs a protected and effective tool like hadoop and cloudera. Educational huge data are the information gathered from colleges which are taken for mining with the goal that the expectation to come learning behavior of the students can be done very easily. Data collected from learning systems can be aggregated over large numbers of students and can contain many variables that data mining algorithms and techniques can explore for model building.

Educational data mining researchers view the following as the goals for their research:

1. Predicting students' future learning behavior by creating student models that incorporate such detailed information as students' knowledge, meta-cognition, motivation, and attitudes.

2. Discovering or improving domain models that characterize the content to be learned and optimal instructional sequences.

3. Studying the effects of different kinds of pedagogical support that can be provided by learning software; and

4. Advancing scientific knowledge about learning and learners through building computational models that

incorporate models of the student, the software's pedagogy and the domain.

EDM analyze data created by any type of information system supporting learning or education in schools, colleges, universities and other academic or professional learning institutions providing conventional model of teaching, as well as easy learning. Predicting students' results and student modeling have been the primary goals of educational data mining.

II. DATA MINING TECHNIQUES

Data mining, also popularly known as Knowledge Discovery in Database, refers to extracting or "mining" knowledge from large amounts of data. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. Data mining is also one step in an overall knowledge discovery process, where organizations want to discover new information from the data in order to aid in decision-making processes. Knowledge discovery and data mining can be thought of as tools for decision-making and organizational effectiveness.

Various algorithms and techniques like Classification, Clustering, Regression, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases. These techniques and methods in data mining need brief mention to have better understanding.

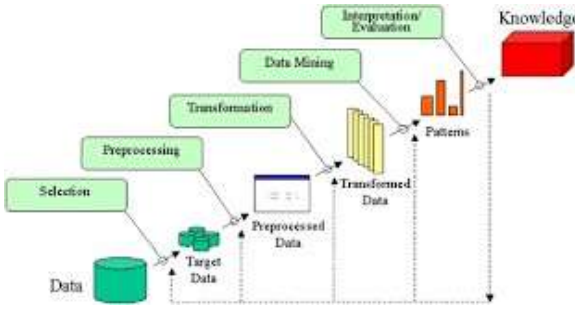

Figure 1: The steps for extracting knowledge from data *A. Classification:*

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks. Classification is the most commonly applied data mining technique, which employs a set of preclassified examples to develop a model that can classify the population of records at large. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm.

B. Clustering

Clustering can be said as identification of similar classes of objects. Clustering, in the context of databases, refers to the ability of several servers or instances to connect to a single database. An instance is the collection of memory and processes that interacts with a database, which is the set of physical files that actually store data. Clustering is one of the main tasks in exploratory data mining and is also a technique used in statistical data analysis. While clustering is not one specific algorithm, it is a general task that can be solved by means of several algorithms. Some of the popular clustering methods that are used include hierarchical, partitioning, density-based and model-based. Clustering is also known as clustering analysis. Clustering is to find point's that naturally group together, splitting, and full data set into a set of clusters. Some of the example application are grouping students based on their learning difficulties and communication patterns, such as how and how much they use tools in a learning management system, and grouping users for purposes of recommending actions and resources to similar patterns.

C. Regression

Regression is a data mining technique used to predict a range of numeric values (also called continuous values), given a particular dataset. ... Regression is used across multiple industries for business and marketing planning, financial forecasting, environmental modeling and analysis of trends. Advanced techniques, such as multiple regressions, predict a relationship between multiple variables. It involves predictor variable (the values which are known) and response variable (values to be predicted).

D. Neural networks

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.

E. Association Rules

Association rules are if-then statements that help to show the probability of relationships between data items within large data sets in various types of databases. Association rule mining has a number of applications and is widely used to help discover sales correlations in transactional data or in medical data sets. Association rule mining is a procedure which is meant to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories. *F. Decision Trees*

A decision tree is a graphical representation of specific decision situations that are used when complex branching occurs in a structured decision process. A decision tree is a predictive model based on a branching series of Boolean tests that use specific facts to make more generalized conclusions. The main components of a decision tree involve decision points represented by nodes, actions and specific choices from a decision point. Each rule within a decision tree is represented by tracing a series of paths from root to node to the next node and so on until an action is reached. A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.

G. Nearest Neighbor Method

Neural network is an adaptive system that changes its structure during a learning phase. Neural networks are used to model complex relationships between inputs and outputs or to find patterns in data. The term "neural network" usually refers to models employed in statistics, cognitive psychology and artificial intelligence.

H. Genetic Algorithm

A genetic algorithm is a heuristic search method used in artificial intelligence and computing. It is used for finding optimized solutions to search problems based on the theory of natural selection and evolutionary biology. Genetic algorithms are excellent for searching through large and complex data sets.

III. EDUCATION IN DATAMINING

Data mining in higher education is a recent research field and this area of research is gaining popularity because of its potentials to educational institutes. Data Mining can be used in educational field to enhance our understanding of learning process to focus on identifying, extracting and evaluating variables related to the learning process of students Mining in educational environment is called Educational Data Mining. EDM can be defined as the application of data mining (DM) techniques to this specific type of dataset that come from educational environments to address important educational questions. Educational data mining (EDM) describes a research field concerned with the application of data mining, machine learning and statistics to information generated from educational settings in colleges.

Challenges of EDM:

- Educational data is incremental in nature
- Lack of Data Interoperability
- Possibility of Uncertainty
- Research Expertise Relation between StudentTeacher

Han and Kamber [1] describes data mining software that allow the users to analyze data from different dimensions, categorize it and summarize the relationships which are identified during the mining process.

TABLE I. STUDENT RELATED VARIABLES

Variable	Description	Possible Values
PSM	Previous Semester Marks	{First > 60% Second >45 & <60% Third >36 & <45% Fail < 36%}
CTG	Class Test Grade	{Poor , Average, Good}
SEM	Seminar Performance	{Poor , Average, Good}
ASS	Assignment	{Yes, No}
GP	General Proficiency	{Yes, No}
ATT	Attendance	{Poor , Average, Good}
LW	Lab Work	{Yes, No}
ESM	End Semester Marks	{First > 60% Second >45 & <60% Third >36 & <45% Fail < 36%}

The domain values for some of the variables were defined for the present investigation as follows:

- PSM – Previous Semester Marks/Grade obtained in BCA course. It is split into five class values: *First – >60%, Second – >45% and <60%, Third – >36% and < 45%, Fail < 40%.*
- CTG – Class test grade obtained. Here in each semester two class tests are conducted and average of two class test are used to calculate sessional marks. CTG is split into three classes: *Poor – < 40%, Average – > 40% and < 60%, Good –>60%.*
- SEM – Seminar Performance obtained. In each semester seminar are organized to check the performance of students. Seminar performance is evaluated into three classes: *Poor – Presentation and communication skill is low, Average* – Either presentation is fine or Communication skill is fine, *Good* – Both presentation and Communication skill is fine.
- ASS – Assignment performance. In each semester two assignments are given to students by each teacher. Assignment performance is divided into two classes:

  Yes – student submitted assignment, No – Student not submitted assignment.
- GP - General Proficiency performance. Like seminar, in each semester general proficiency tests are organized. General Proficiency test is divided into two classes: *Yes* –

student participated in general proficiency, No – Student not participated in general proficiency.

- ATT – Attendance of Student. Minimum 70% attendance is compulsory to participate in End Semester Examination. But even through in special cases low attendance students also participate in End Semester Examination on genuine reason. Attendance is divided into three classes*: Poor - <60%, Average - > 60% and <80%, Good - >80%*.
- LW – Lab Work. Lab work is divided into two classes: *Yes – student completed lab work, No – student not completed lab work.*
- ESM - End semester Marks obtained in BCA semester and it is declared as response variable. It is split into five class values: *First – >60%, Second – >45% and <60%, Third – >36% and < 45%, Fail < 40%*.

Educational data systems now store large amounts of data and its origin can come from different sources, different formats and different granularity levels. The problems of educational data mining, must be analyzed particularly due to their specific objective determines a singularity when it is solved by data mining techniques. Data mining in education can analyze the data generated by any system of learning and focus on diverse aspects, both individual and group and take into account underlying, administrative, demographic and motivational data which in turn contain multiple levels of hierarchy, contexts, levels of granularity and historical data. It is called interdisciplinary educational data mining because it can involve the analysis of social networks, educational psychology, cognitive psychology, psychometrics among others.

IV.CONCLUSION

In this paper, Educational data mining is the best method for improving the performance of the students. The data mining technique applied for educational dataset proves to be of immense use for the benefit of the students. Various techniques applied in the data mining technique such as decision tree, clustering are highly efficient in analyzing the educational big data. There are increasing research interests in using data mining in education. Educational data mining (EDM) is an area full of exciting opportunities f or researchers and practitioners. It analyses new patterns and relationships among a large amount of data. Classification, clustering and regression methods will support to develop validated models of a variety of complex constructs that have been embedded into increasingly sophisticated student models. This paper is just a simple review to this emerging field EDM and aims to highlight the importance of its study. EDM applies techniques coming from statistics, machine learning, and data mining to analyze data collected during teaching and learning, tests learning theories, and informs decisionmaking in educational practice.

REFERENCES

[1] J. Han and M. Kamber, "Data Mining: Concepts and Techniques,"Morgan Kaufmann, 2000.

[2] Agarwal, S., Pandey, G. N., & Tiwari, M. D. (2012). "Data Mining in Education: Data Classification and Decision Tree Approach".

[3] A. Villanueva, L.G. Moreno & M.J. Salinas "Data mining techniques applied in educational environments: Literature review".

[4] Dr. M. Thangamani, T.Selvakumar "Exploring Educational Dataset using Data Mining Technique", (IJARBEST) Vol.2, Issue.2, February 2016.

[5] Brijesh Kumar Baradwaj, Saurabh Pal, "Mining Educational Data to Analyze Students Performance", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No .6, 2011.

[6] Brahamjit Pannu1, Mr. Puneett Sharma, A Comparative data mining technique on education data, International Journal For Technological Research In Engineering, Vol. 2, Pp. 2124 -2127, Issue 9, May-2015.

[7] Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, M. Inayat Khan, "Data mining model for higher education system", Europen Journal of Scientific Research, Vol.43, No.1, pp.24-29, 2010.

[8] P. Gulati and A. Sharma, "Educational data mining for improving educational quality". Int. J. Comput. Sci. Inf. Technol. Secur., vol. 2, no. 3, pp. 648–650, 2012.