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 10, October 2018

A Survey on Educational Data Mining and its Techniques

V.A.Malathi[1], C.V.Banupriya[2]

Research Scholar, Department of Computer Science, SJSMV College of Arts & Science, Coimbatore, India.[1]

Assistant Professor, Department of Computer Science, SJSMV College of Arts & Science, Coimbatore, India.[2]

ABSTRACT: Data mining is a very useful in the field of education especially when examining students' learning behavior. Data mining is a powerful new technology with great potential to help schools and universities focus on the most important information in the data they have collected about the behavior of their students and potential learners. Educational data mining is an emerging discipline, concerned with developing methods for exploring the unique and increasingly large scale data that come from educational settings and using those methods to better understand students, and the settings which they learn in. This proposed work compared various educational data mining research works with their attributes, tools, various data mining algorithms and their accuracy while predicting students' future performance. In educational data mining used different techniques like regression, classification, clustering etc. also discussed in this proposed paper.

KEYWORDS: Data mining; Educational data mining; student learning behavior; prediction; tools; techniques

I. INTRODUCTION TODATA MINING

Data miningis the process of discovering interesting patterns andknowledge from large amounts of data. The data sources can include databases, datawarehouses, theweb, other information repositories, or data that are streamed into thesystem dynamically.It is a component of a wider process called Knowledge Discovery from databases. It involves scientists from a wide range of disciplines, including mathematicians, computer scientists and statisticians. They are working in the fields such as machine learning, artificial intelligence, information retrieval and pattern recognition.Data Mining is an interdisciplinary of Computer Science. The goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.Some of the important areas where data mining is widely used,

- Sales / Marketing
- Medicine
- Education
- Research analysis
- Transportation
- Health care and Insurance
- Banking / Finance

Data mining, also popularly referred to as knowledge discovery from data (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories or data streams.Data mining is a process that analyzes a large amount of data to find new and hidden information that improves business efficiency. Data mining can be used by an institution to take accurate decisions and also to predict the results of the student. With the results the institution can focus on what to teach and how to teach. Learning pattern of the students can be captured and used to develop techniques to teach them.

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 10, October 2018

II. STEPS INVOLVED IN KNOWLEDGE DISCOVERY PROCESS

1. Data cleaning (to remove noise and inconsistent data)
2. Data integration (where multiple data sources may be combined)
3. Data selection (where data relevant to the analysis task are retrieved from thedatabase)
4. Data transformation (where data are transformed and consolidated into formsappropriate for mining by performing summary or aggregation operations)
5. Data mining (an essential process where intelligent methods are applied to extractdata patterns)

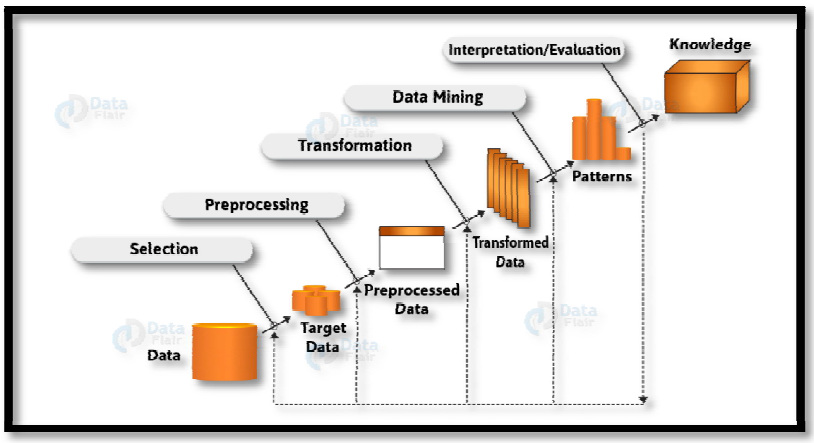

Fig.2.1 Steps in Knowledge Discovery Process

6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on interestingness measures)
7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users)

Data mining tasks can be classified in two categories – descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in database. Predictive mining tasks perform inference on the current data in order to make predictions.

III. EDUCATIONAL DATA MINING

The new emerging field called Educational Data Mining [EDM], concerns with developing methods that discover knowledge from data originating from educational environments. Educational data mining is an upcoming field related to several well-established areas of research including e-learning, adaptive hypermedia, intelligent tutoring systems, web mining, data mining, etc.The main objective of any educational system is to improve the quality of education. Educational data mining is concerned with developing, researching, and applying computerized methods to detect patterns in large collections of educational data that would otherwise be hard or impossible to analyze due to the enormous volume of data within which they exist. Its objective is to analyze educational data in order to resolve educational research issues.

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 10, October 2018

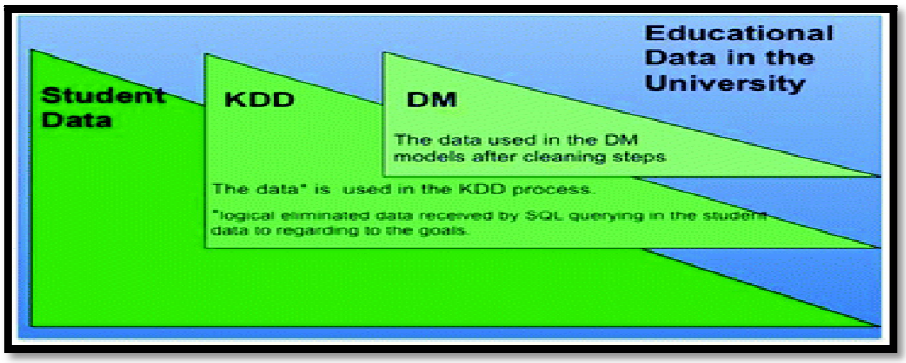

Fig.3.1 Educational Data mining process

EDM is the process of transforming raw data compiled by educational systems in useful information. In the field of education techniques "Data Mining" has also been used to analyze the curriculum and subject of the current research topics, as well as to analyze the students' performancethat can be used to take informed decisions and answer research questions.

Goals of Educational Data Mining:

- Predicting students' future learning behavior by creating student models that incorporate such detailed information as students' knowledge, motivation, metacognition, and attitudes.
- Discovering or improving domain models that characterize the content to be learned and optimal instructional sequences.
- Studying the effects of different kinds of pedagogical support that can be provided by learning software.
- Advancing scientific knowledge about learning and learners through building computational models that incorporate models of the student, the domain, and the software's pedagogy.
- EDM methods may also be used to categorize the students who require support, to analyze students' learning and cluster them according to their strengths and weaknesses for placement related activities.

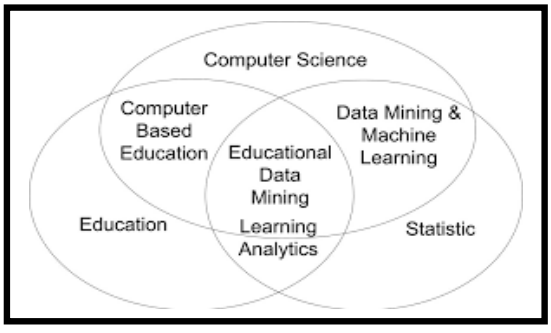

Fig.3.2Application of data mining in Education

IV. LITERATURE SURVEY ON EDUCATIONAL DATA MINING

Literature survey enables a researcher to become an expert in the specific area. To understand the purpose and expectations of the prompt for research so as to place appropriate emphasis in the analysis and summary.In this table analyzed various educational data mining papers to predict the student academic performances in their future semester.Various algorithms, tools and different data sets were used to find out the future academic performance of the students.

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 10, October 2018

Table I: Comparison of various Educational Data Mining studies

S.No	Title of the Paper	Author Name	No of Attributes	Dataset	Tools Used	Algorithms Used	Accuracy Percentage
1	Educational Data Mining & Students' performance Prediction	Amjad Abu Saa	21	270 Records	Rapid Miner Weka	1.C4.5 2. ID3 3. CART 4.CHAID 5. Naive Bayes	1.35.19% 2.33.33% 3.40% 4.34.07% 5.36.40%
2	Data Mining approach for predicting Student Performance	EdinOsmanbegovic MirzaSuljic	12	257 Records	Weka	1.Naive Bayes 2.Multilayer Perception 3.J48	1.76.65% 2.71.2% 3.73.93%
3	Mining Educational Data to Improve Students' Performance: A Case Study	Mohammed M.AbuTair Alaa Mustafa El-Halees	18	3360 Records	Rapid Miner	1.Rule Induction 2.Naive Bayesian 3.K means algorithm 1.Cluster-0 2. Cluster-1 3. Cluster-2 4. Cluster-3	1.71.25% 2.67.50% 1.23% 2.31% 3.31% 4.15%
4	A CHAID Based Performance Prediction Model in Educational Data Mining	M.Ramaswami R.Bhaskaran	35	1000 Records	STATISTICA 7	CHAID classification tree	44.69%
5	Predictive Modeling of Student Dropout Indicators in Educational Data Mining Using Improved Decision Tree	SubithaSivakumar SivakumarVenkataraman RajalakshmiSelvaraj	32	240 Records	Renyi Entropy	1.ID3 2.Improved Decision Tree	1.92.50% 2.97.50%
6	Predicting Students Performance by Using Data mining methods for Classification	DorinaKabakchieva	14	10330 Records	Weka	1.Decision Tree classifier (C4.5) 2.Bayesian classifier 3.K-Nearest Neighbour 4.Rule learner (JRIP)	1.Most reliable 2.Less accurate 3.Unable to predict 4.Most reliable
7	Efficiency of Decision Trees in Predicting Students Academic Performance	Anupama Kumar Vijayalakshmi	5	117 (Internal Marks) 116 (External Marks	Weka	1.ID3 2.C4.5 3.J48	1.Less accurate 2.Less accurate 3.More Accurate

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 10, October 2018

V. TECHNIQUES OF EDUCATIONAL DATA MINING

One of the most important tasks in data mining is to select the correct data mining technique. A generalized approach has to be used to improve the accuracy and cost effectiveness of using data mining techniques.This techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful for decision-making. Various algorithms and techniques such as Prediction, Classification, Clustering, Regression, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbour method etc., are used for knowledge discovery from databases.

A. Statisticsin EDM

Data mining techniques statistics is a branch of mathematics which relates to the collection and description of data. Statistical technique helps to discover the patterns and build predictive models. For this reason data analyst should possess some knowledge about the different statistical techniques. Through statistical reports people can take smart decisions.

B. Prediction in EDM

The goal of prediction is to infer a target attributes or single aspect of the data (predicted variable) from some combination of other aspects of the data (predictor variables). Types of predictions methods are

- classification (when the predicted variable is a categorical value)
- regression (when the predicted variable is a continuous value)
- density estimation (when the predicted value is a probability density function)

C. Classification inEDM

Classification models describe data relationships and predict values for future observations. In classification test data is used to estimate the accuracy of the classification rules. If the accuracy is acceptable, the rules can be applied to the new data tuples. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination.

Classification classifies a data item into some of several predefined categorical classes. The algorithm used for classification are:

- Decision tree
- Naive Bayes classification
- Generalized linear models
- Support vector machine etc.

D. Regression inEDM

In statistical terms, a regression analysis is the process of identifying and analyzing the relationship among variables. Regression is an inherently statistical technique used regularly in data mining. Regression analysis establishes a relationship between a dependent or outcome variable and a set of predictors. Regression is supervised learning data mining technique. It can help to understand the characteristic value of the dependent variable changes, if any one of the independent variables is varied. It is generally used for prediction and forecasting.

E. Decision Tree inEDM

Decision tree can be considered as a segmentation of the original dataset where segmentation is done for a particular reason. Each data that comes under a segment has some similarities in their information being predicted. A decision tree provides results that can be easily understood by the user. This technique can be used for prediction and data pre-processing.

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 10, October 2018

F. Clustering inEDM

Clustering can be defined as the identification and classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters). 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. In EDM, clustering can be used for grouping similar course materials or grouping students based on their learning and interaction patterns.

G. Relationship Mining inEDM

It is used for discovering relationships between variables in a dataset and encoding them as rules for later use. There are different types of relationship in mining techniques such as association rule mining (any relationships between variables), sequential pattern mining (temporal associations between variables), correlation mining (linear correlations between variables), and causal data mining (causal relationships between variables). In EDM, relationship mining is used to identify relationships between the student's on-line activities and the final marks and to model learner's problem solving activity sequences.

H. Outlier Detection inEDM

The goal of outlier detection is to discover data points that are significantly different than the rest of data. An outlier is a different observation (or measurement) that is usually larger or smaller than the other values in data. In EDM, outlier detection can be used to detect deviations in the learner's or educator's actions or behaviors, irregular learning processes and for detecting students with learning difficulties.

I. Support Vector Machine inEDM

Support Vector Machine (SVM) is one of the data mining techniques. It is a supervised learning method and can be used for both regression and classification. SVM classifiers are used for the prediction of placement of students as in many cases, students focus only on their regular curriculum of studies besides on other educational trends which are necessary for overall development of students and their placements.

J. Neural Network inEDM

Neural Network is another important technique used by people these days. This technique is most often used in the starting stages of the data mining technology. Artificial neural network was formed out of the community of artificial intelligence. Neural networks are very strong predictive modelling technique. Every neural network model has different architectures and these architectures use different learning procedures.

K. Association Rule Technique inEDM

This technique helps to find the association between two or more items. It helps to know the relations between the different variables in databases. It discovers the hidden patterns in the data sets which is used to identify the variables and the frequent occurrence of different variables that appear with the highest frequencies.

L. K Nearest Neighbourin EDM

KNN, or k-Nearest Neighbours, is a classification algorithm. K-Nearest Neighbors is one of the most basic yet essential classification algorithms in machine learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection. KNN can be used for both classification and regression predictive problems.

M. Text Mining in EDM

Text mining methods can be viewed as an extension of data mining to text data and it is very much related to web content mining. It is an interdisciplinary area involving machine learning and data mining, statistics, information

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 10, October 2018

retrieval and natural language processing. Text mining can work with unstructured or semi-structured datasets such as full-text documents, HTML files, emails, etc.

VI. DATA MINING TOOLS

A. ***WEKA*** -WEKA stands for Waikato Environment for Knowledge Learning. It was developed by the University of Waikato, New Zealand. It is a collection of machine learning algorithms for data mining tasks. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

B. ***RAPIDMINER***- It is the world-leading open-source system for data mining. It is an environment for providingdata mining and machine learning procedures including: data loading and transformation (ETL), data pre-processingand visualization, modelling, evaluation, and deployment. Types of graphs and visualization techniquesavailable in rapid miner,scatter matrices, line, bubble, histograms, area, bar charts, Quartile etc.

C. ***R – TOOL*** - R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories by John Chambers and colleagues. It is software which provides an environment in which we can perform statistical analysis and produce graphics.

D. ***ORANGE***- Orange is an open source data visualization and analysis for novice and experts. It contains components for machine learning. Much of Orange is devoted to machine learning methods for classification, or supervised data mining. These methods rely on the data with class-labelled instances.

E. ***MATLAB***- MATLAB is a high level language and interactive environment for numerical computation, visualization and programming. Using MATLAB we can analyze data, develop algorithms and create models and applications. The language, tools and built-in math functions enable us to explore multiple approaches and reach a solution faster than with spreadsheets or traditional programming languages.

VII. CONCLUSION

Educational data mining offers many opportunities to improve effectiveness of teaching. It is an emerging multidisciplinary research area. Educational Data Mining (EDM) is the process of discovering useful information from raw data generated and collected from educational systems which can be used by the different stakeholders. This survey paper, analyzed different educational data mining papers to find out student academic performance in future using different algorithms and tools.The application of data mining methods in the educational sector is an interesting phenomenon. Various educational data mining techniques and tools are also discussed in this paper. Data mining techniques in educational organizations help us to learn student performance, student behavior, designing course curriculum and to motivate students on various parameters.

REFERENCES

1. Amjad Abu Saa, “Educational Data Mining & Students’ performance Prediction”, International Journal of Advanced Computer Science and Applications, Vol.7- No.5 2016.
2. Arun K Purari, “Data Mining Techniques”, University Press, 2004.
3. Anupama Kumar, Vijayalakshmi,“Efficiency of Decision Trees in Predicting Students Academic Performance”,D.C. Wyld, et al.(Eds): CCSEA 2011, CS & IT 02, pp.335-343, 2011.
4. DorinaKabakchieva, “Predicting Students Performance by Using Data mining methods for Classification”, Cybernetics and Information Technologies, Volume 13, No.1, 2013.
5. Dr. T. Karthikeyan, V.A.Kanimozhi, “A Review on Heart Disease Prediction System Using Data Mining Tools”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 9, September 2016.
6. EdinOsmanbegovic, MirzaSuljic , “Data Mining approach for predicting Student Performance”, Economic Review – Journal of Economics and Business, Vol.X- Issue 1, May 2012.
7. Jiawei Han, MichelineKamber, Jian Pei, “Data Mining: concepts and techniques”, 2011.
8. Mohammed M.AbuTair, Alaa Mustafa El-Halees, “Mining Educational Data to Improve Students’ Performance: A Case Study”, International Journal of Information and Communication Technology Research, Volume.2- No. 2, Feb 2012.

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 10, October 2018

9. M.Ramaswami, R.Bhaskaran, "A CHAID Based Performance Prediction Model in Educational Data Mining", International Journal of Computer Science Issues, Vol.7- Issue 1, No. 1, Jan 2010.
10. Romero C, Ventura S, Pechenizky M, Baker R. Handbook of Educational Data Mining. Data Mining and Knowledge Discovery Series. Boca Raton, FL: Chapman and Hall/CRC Press; 2010.
11. SubithaSivakumar, SivakumarVenkataraman, RajalakshmiSelvaraj, "Predictive Modeling of Student Dropout Indicators in Educational Data Mining Using Improved Decision Tree", Indian Journal of Science and Technology, Vol 9(4), Jan 2016.
12. V.A.Kanimozhi ,Dr. T. Karthikeyan, , " A Survey on Machine Learning Algorithms in Data Mining for Prediction of Heart Disease", International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 4, April 2016.

BIOGRAPHY

Ms. Malathi V.A pursuing M.Phil. Degree in computer science at SJSMV College of Arts &Science, Coimbatore, Tamil Nadu. Her research interest is data mining.

Ms.Banupriya C.V working as Assistant Professor in SJSMV College of Arts & Science, Coimbatore, Tamil Nadu. She has published many national and international research papers. Her specialization is data mining.