Review On Educational Data Mining And Predicting Students Performance: Machine Learning Theory

Shubham Agrawal* & Sumit Nigam**  
*(CSE) SIMS Indore, R.G.P.V. University (MP).  
**Head, Department of Computer Science, SIMS Indore, R.G.P.V. University (MP).

Received: Feb. 25, 2018  
Accepted: April 01, 2018

ABSTRACT  
Data mining is very emerging technology that is used in each and every system. Education data mining is very useful for disciplines because the amount of data in education system is increasing day by day. In higher education it is relatively new but its importance increases because of increasing database. There are many approaches for measuring students’ performance. Data mining is one of them. With the help of data mining the hidden information in the database is get out which help for improvement of students performance. Education data mining is used to study the data available in education field to bring the hidden data i.e. important and useful information from it. There are many methods of data mining is used to analysis of students performance. Classification method like decision tree is most used to measure the students’ performance. With the help of these it is easy to improve the result and future of students. More methods like clustering, regression, time series, and neural network can also be applied.

Key Words: Data mining, EDM, Decision tree, Clustering, Neural network, Students data.

I. INTRODUCTION  
Higher education is very important for students’ life. Higher education institute are focus on analysis of every objects because of private participation. Data mining techniques is applied in many fields like marketing, medicine, fraud detection, web, engineering etc. The main aim of data mining is to know hidden knowledge. DM provides various methods analysis; these include classification, association, k-means, decision tree, clustering, regression, time series, neural network, etc. Application of data mining in the educational system is directly help to analysis of participants in the education system. The students also recommend many activities and task [1]. Data mining is also used to show how students use material of particular course. In teaching environment trainer are able to obtain feedback on students [5].

II. RELATED WORK  
Edin Osmanbegovic [1]-In these paper supervised data mining algorithm were applied. Different method of data mining was compared. The data were collected from the survey conducted during the summer semester at the University of Tuzla. Many variable like Gender, GPA, Scholarships, High school, Entrance Exam, Grade, etc. are taken for the performance. Naïve Bayes algorithm, multilayer Perceptron, J48 issued. The result indicates that the naïve Bayes classifier outperforms in predication decision tree and neural network method. These will help the student for future.

Qasem A. Al-Radelideh [2]-The title of the paper is “Mining student data using decision tree”. They use data mining process for student performance in university courses to help the higher education management. Many factors affect the performance. They use classification technique for building the reliable classification model, the CRISP-DM (cross-industry standard process for data mining) is adopted. These method consist of five steps i.e. collecting the relevant features of the problem, Preparing the data, Building the classification model, Evaluating the model and finally future prediction. The data were collected in table in proper format, the classification model were building using the decision tree method. Many rules were applied. The WEKA toolkit is used. Different classification methods were used like ID3, C4.5 and naïve Bayes and accuracy were in the table as result.

J.K. JothiKalpana [3]-“Intellectual performance analysis of students by using data mining techniques” This paper focus on the prediction of school in different level such as primary, secondary, higher level. Clustering method such as centroid based distribution based and density based clustering are used. The data were collected from Villupuram College. There method used for improving the performance as the students.

Cristobal Romero [4]-“Educational data mining: A Review of the state of the art”. EDM i.e. educational data mining is emerging discipline. EDM process converts raw data coming
from educational system into useful information. DM techniques are used i.e. association rule mining for selecting weak students. Several classification algorithms were applied in order to group students. EDM tools were designed for educators.

Romero [5] - "Educational data mining survey from 1995 to 2005" There is also web-based education in the computer aided instruction in the specific location. Web based education is so popular now a days that predication its level is also become useful. Data processing is done for transform the original data into suitable shape. Web mining is there for extract knowledge from the web. Clustering, classification is used. In these it says that the predication of performance in e-learning is also so important.

S. Kotsiantis [6] - "Predicating students' performance in distance learning using machine learning techniques" Many university are giving distance learning education so predicating performance of students in that become so important. Machine learning algorithm is so effective for many types of learning tasks. This paper Use ML techniques to predict students' performance in distance learning system. Set of rules are planned. Decision tree are used, ANN is also inductive learning based on computational models. Set of attribute are taken and divided into groups. There is ANOVA test result. It showed that best algorithm is naïve Bayes with 66.49% accuracy in the data it taken.

Pooja Thakar [7] - "Performance analysis and prediction in education data mining: A Research Travelogues" Lots of data is collected in educational databases. In order to get benefits from such big data tools are required. University produces lots of students and its performance predication is important. Set of weak students are taken and predication with data mining techniques is used. This paper says that many models are required for an instruction.

V. Shanmugarajeshwari [8] - "Analysis of students' performance evaluation using classification techniques" The author used the classification techniques for predication of student's performance in education system. The data is collected and preparation is done the preprocessing for checking. It calculates the entropy, Info Gain, Ratio then the information gain for evaluating these. Classification technique is used. Decision tree is build and finally gain ratio is evaluated.

Mashael A [11] - These researches has applied decision tree for predicting students final GPA. It used WEKA toolkit. It collect the data from C. s. College at king save university in the year 2012 were collected from the institute. Each student record with different attributes. Student name, student id, final GPA, semester of graduation etc. It is important to improve the final GPA of the student.

Ben Daniel [13] - It applied big data analysis in higher education. KDD is an interdisciplinary area focusing on method for identifying and extracting pattern from large data sets. Big data help provide insight to support students learning needs.

Tisimy Devasia [18] - It used classification technique to predict the student performance. Naïve theorem is used various information like group action, class text, semester and assignment marks were collected from the students previous information to predict performance of the student.

Ryan S. J. D. Baker [19] - "The state of educational data mining in 2009: A review and future vision" In these paper author review the trend in 2009 in field of educational data mining. The year 2009 finds research communizing of EDM and these moment in EDM bring unique opportunity. EDM categories in web mining, Statistics and Visualization, Clustering, Relationship mining i.e. Association rule mining and causal data mining. There are many application of edm. These papers discuss about the EDM.

Pooja M. Dhekanekar [22] - "Analysis of student performance by using data mining concept" Data mining technique is used in many area and in the educational field it become so important for future of the students. Students classification is done on the basic of students mark. Association rule mining and causal data mining. There are many application of edm. These paper discuss about the EDM.
III. MACHINE LEARNING

It is the branch of science that works with the system in such that they automatically learn. It means that recognizing and understanding the input data and moving decision on the support data. The name machine learning was come in 1959 by Arthur Samuel. They evolved from the study of pattern, AI, computational theory. Machine learning constructs the algorithm that can learn and make predictions. Machine learning closely related to statistics which help in prediction. It is very difficult to take the division for their problem and algorithm is developed. There algorithm are based like statics logic etc.

Application of machine learning:-
- Vision processing.
- Language processing.
- Forecasting.
- Pattern recognition.
- Games.
- Data mining.
- Robotics.
- Expert system.

Types of Machine Learning:-

Supervised Learning:-
In there is desired input with desired output. In addition to take feedback about the accuracy of predication. It can be apply what are learned in the past to the new set of data using the suitable example to feedback future events. There are known training data set and starting from the analysis; the learning algorithm produces as function to make prediction about the output. The system is able to provide targets from any new input. The learning algorithm can also compare its output and find error in order to modify the model.

Unsupervised Learning:-
In these we do not have any target to predict. It is used for clustering in different groups. It is used when the information used to train is neither classified nor solved. It studies how systems can information function to describe hidden structure from unlabeled data. It examples positive data and can draw information from data set to describe hidden structure.

Semi-Supervised M.L.:-
It is between supervised and unsupervised learning. It means it used both labeled and unlabeled data for training. It can be said that it used small amount of labeled data and the large amount of unlabeled data. The system in there is for learning accuracy.

Types of Supervised Learning Algorithms
- Logistic Regression.
- Decision Tree.
- Support vector machine.
- K-Nearest neighbors.
- Naive Bayes.
- Linear Regression.

Types of Unsupervised Learning:-
- K - Mean clustering.
- Hierarchical clustering.
- Hidden markov model.

Data Mining - Data mining is process of discovering patterns in large data set that involving the methods at intersection of machine learning, and the system. With the help of data mining tools the prediction can be made easily, There is large data set and pattern is identify and establish relationship to solve problem through data analysis. It means analyzing hidden patterns of data according to different perspectives for categorization into useful information.

KNN - K-nearest neighbors is the classification algorithm. On pattern recognition it is the non-parametric method used for both classification and regression. In both cases the input consists by the K-closest training example in the feature space. The output depends on whether k-nearest. It is to simple even with its simplicity, It give highly competitive results.

KNN has different names is K-nearest neighbor.

Memory – Based Neighbors

Example – Based Reasoning
- Instance – Based Learning
- Care – Based Reasoning
- Lazy Learning

KNN = No. any neighbors
If K = 1
Select the nearest neighbor
If K > 1
For classification select the method frequent neighbor
For regression calculate the over one as K-neighbors
Distance – categorical Variable

How did it do that?
Step1 – charge to no. k as neighbors
Step2 – Take the K nearest neighbors as the new data point
Step3 – Among there K neighbors loud the number as data point in each category
Step4 – Assign the new point to the category where you’re counted the mast neighbors.
So step is simple but it can solve complete problem.
Decision tree algorithm—It is flow chart like structure. The tree consists of the root node branches and the leafnodes. The internal node in the tree denotes test on an attribute. Each branch denotes the outcome of the test and class label. Tree pruning is performed in order to remove anomalies in the training data due to noise.

Classification technique’s—It is a method in which data is classified into different classes. The classification analysis is used to receive useful information from data or Meta data. In classification techniques we will apply algorithm to classify the data.

Association Rule—It refer to method that can help to identify pattern in database these techniques is used to identify hidden pattern in the data that can be used to identify variable within the data that appear frequently in the dataset these rule is mostly used in analysis customer behavior.

Clustering Analysis—cluster is collection of similar data within a group. It means that the data of same type are kept in within a group and other are in other group. Clustering analysis is analysis of discovering groups and cluster in the data in such a way that the degree of association is highest.

K-Means Clustering—It is positive method that is mostly used for cluster analysis. K – Mean clustering aim to partition n observation into k-clusters in which each observation belongs to the cluster with the nearest mean.

SVM—Support vector machine is supervised learning algorithm, it is a classification concept. When we have dataset with features and class; labels both then we can use support vector machine, but if an dataset does not have the output it is consider as unsupervised algorithm.

There are two types of SVM—Linear Nonlinear

In linear model the training example are plotted in space. In nonlinear model our dataset is generally dispersed up to some expert.

IV CONCLUSION

Machine learning is very emerging technology that every placed it used. Now days in bank, labs, telecom, industrial each and every place machine learning is used. Data mining is part of it which helps in prediction, future prediction is very important in many place which help so much. Many algorithm is build and more and more research is going on every technology used the concept of it. We survey many papers for prediction of students’ performance. Students performance is so important for their future it not only help student but also help teachers institute parents. Many big institutes used the concept of AI for prediction.

V. REFERENCES


2. Qasem A. Al-Radaideh, Emad Al-Shawakfa— Mining Student Data Using Decision Trees, Research Gate Article 2006.


14. Fadhilah Ahmad, NurHafieza Ismail and Azwa Abdul Aziz “The Prediction of Students' Academic
Leadership is the capacity to translate vision into reality.

~ Warren G. Bennis