International Journal of Electrical, Electronics and Computer Engineering **6**(1): 24-28(2017)

ISSN No. (Online): 2277-2626

A Study of various Techniques for Predicting student Performance under Educational Data Mining

Ankita Katare* and Shubha Dubey**

**Department of Computer Science and Engineering RITS, Bhopal, (Madhya Pradesh), INDIA*
***Department of Computer Science and Engineering RITS, Bhopal, (Madhya Pradesh), INDIA*

(Corresponding author: Ankita Katare, ankita.katare27@gmail.com)

(Received 27 December, 2016 Accepted 28 January, 2017)

(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Data mining techniques applied on educational field increases day by day. Data mining techniques when applied to the educational environment named as Educational Data Mining. A research community of such emerging field involves student learning experiences by predicting student's performance. As compared to all data mining techniques, classification is most important one. The techniques used for improving students performance and finding the finest formation of set of courses for existing environment is foremost plan behind this research. This survey mostly focuses on classification algorithms and their utilization for evaluating student performance.

Keywords: Data Mining, EDM, Classification Algorithms, Performance Prediction

I. INTRODUCTION

To make better decisions the new generation of data mining tools and techniques are used. The field of data mining grew out of the limitations of current data analysis techniques in handling the challenges posed by these new types of datasets. Data mining does not replace other area of data analysis, but rather takes them as the foundation for much of its work. By simply applying data analysis techniques the challenges of analyzing new types of data cannot be met in separation from those who recognize the data and domain in which it resides. Repeatedly, proficiency in building multidisciplinary teams has been as responsible for the achievement of data mining projects as the formation of new and modern algorithms.

Knowledge Discovery in Databases Process

The KDD is an automatic, exploratory data analysis and modeling of large data sources. The KDD is the organized process of identifying valid, novel, useful, and human eye understandable patterns from huge and difficult data sets. Data Mining is the foundation of the KDD process, relating the linking of algorithms that search the data, build up the model and determine previously unknown patterns. The KDD knowledge discovery process is cyclic, interactive, and consists of nine steps.

The unifying purpose of the KDD process is to dig out valuable information from data in the framework of large databases. Data mining refers to the set of computational methods that extract important patterns from original data. In addition, KDD process is concerned about manipulation with immense data, scaling algorithms for better presentation, appropriate analysis of retrieved information, and human interaction with the overall process.

"Educational Data Mining is an rising authority, 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 be taught in." On one hand, the raise in both instrumental educational software as well as state database of students information have shaped great repositories of data dazzling how students be trained. On the other hand, a new context has created by using Internet in education known as e-learning or web-based education in which information in large amount in relation to teaching–learning interaction are continuously generated and universally available. These data repositories used by EDM to better recognize learners and learning, and to develop computational approaches that merge data and theory to convert practice to profit learners.

Number of questions can be answered by educational data mining from student data the patterns obtained such as

1) Who are the students at risk?
2) What are the chances of placement of student?
3) Who are the students likely to fall the course?
4) What is the quality of student contribution?
5) Which courses the organization should present to attract more students?

In order to perform EDM, researchers use a variety of sources of data such as intelligent computer tutors, classic computer-based educational systems, online class discussion forums, electronic teacher grade books, and school-level data on student enrollment, and standardized tests. Many of these sources have existed for decades. What has recently changed is the fast development in storage and communication provided by computers, which greatly simplifies the job of collecting and collating large data sets. This bang of data has revolutionized the way we study the learning process.

Student learning assumptions have been divided into five hierarchical categories: learning as the quantitative raise in knowledge, learning as memorization, learning as gaining of facts and procedures that can be retained and/or utilized in practice, learning as the concept of meaning and learning as an interpretative process meant at the concerned of reality.

The skill to observe the growth of student's academic presentation is a serious subject to the academic area of advanced learning. A scheme for analyzing student's outcome based on classification study and uses standard statistical algorithms to position their scores data according to the height of their performance is described. There are two main factors in predicting student's performances, which are attributes and prediction methods. Main objective of this survey is to use classification algorithms for predicting and analyzing the student performance and compare with other algorithms and methods based on some parameters. Large database in educational environment contain secret information which we test on a standard dataset to predict the student performance and to make improvement.

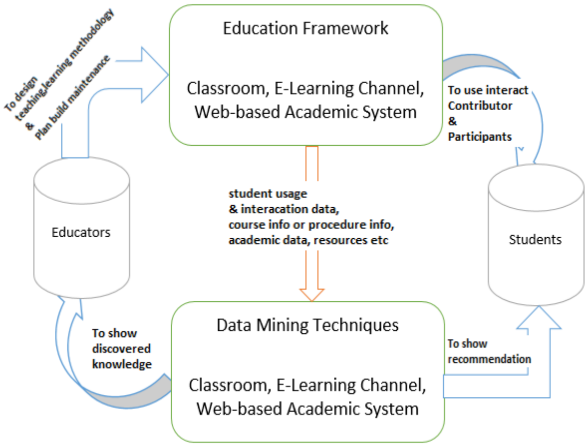

Fig. 1. Data Flow for Educational Data Mining (EDM).

II. LITERATURE REVIEW

Nowadays educational data mining is emerged as a very active research area because there are lots of things in this field are not exposed. Work connected to student performance, student behavior analysis, faculty performance and impact of this factor on student final performance need much attention.

C. Anuradha *et.al.* [1] In recent years it has been observed in the survey on student performance prediction to be done by using ID3 and C4.5 classification algorithms and compared them on the basis of utilization. Dropout rates for students in correspondence and open courses are on increase. There is a need of analysis of factors causing increase in dropout rate. Data mining technique are used for analysis and prediction by them to find the dropout reasons Hina Gulati *et.al.* [2], Krian Parmar *et.al.* [3]. In this paper classification techniques are used for prediction of student performance in distributed environment. Aim of the Distributed Data Mining (DDM) is getting useful models and patterns from distributed databases for decision making for accurate results and this will increase performance on system also. Data Clustering and predictive modeling applied on dataset to discuss an analytical approach to deal with e-learning data in virtual learning environment (VLE) Alana m.de Morais *et.al.* [4]. Students data to be evaluated using decision tree which is helpful in predicting student results and identifying weak students. Here they calculated the entropy of attributes (Information gain) taken in datasets Pratiyush guleria *et.al.* [5]. This paper show how social media conversation of students help to predict the performance based on their activities on social media by using Naïve Bayes multi label classifier Xin Chen *et.al.* [6].

Hashima Hamsa *et.al.* [7] uses two decision tree and fuzzy genetic algorithms to predict students academic performance of bachelor's and master's degree based on their academic marks. Review on analysis of data of student using different data mining techniques and analyzed that classification methods : Decision tree and Neural network are two methods highly used by researchers for predicting student performance Amirah Mohamed Shahiri *et.al.* [8]. This papers discuss the comparison of genetic algorithm to J48 with respect to the accuracy and size of tree to evaluate the performance of engineering students and comparison of GA with SVM classifier where Support vector machine algorithm is most appropriate for predicting student performance Ruhi R. Kabra *et.al.* [9] P. Usha *et.al.* [10]. Here the classification algorithms including rule learner, a decision tree classifier, a neural network and nearest neighbor classifier applied on dataset.

The performance of these algorithms is analyzed and compared Dorina Kabakchieva *et.al*. [11].

III. DATA MINING TECHNIQUES

Data mining techniques are implemented in several organizations as a standard process for analyzing the great volume of on hand data, extracting valuable information and knowledge to sustain the major decision-making processes. EDM concerns with innovative methods and techniques by searching into unusual kind of data from educational settings to recognize students learning skill. In educational area, data mining techniques are incredibly valuable for enhancing the present educational principles and managements. These techniques offer a path to a several stage of ranking, a result which gives a new observation of how people can become proficient in these educational sectors. As a result of this, EDM has given rise to hypothesis concerned with the scientific study of human sciences. We known that wide range of data is stored in educational databases, so in order to get required data and to find the hidden relationship, for that different data mining techniques are developed and used.

Classification is a simple process of discovering a prototype (or function) that recognize the salient features of data classes or concepts, for the purpose of being capable to use the model to examine the class of items whose class label is not known. It forecast distinct and unordered labels in huge data sets. As with classification, the test set is used to build a predictor but an independent test set should be used to assess its accuracy. The data classification process involves learning and classification. In learning classification algorithm analyzed the training data. In classification the accuracy of the classification rules estimated by using test data. The classifier training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The EDM Classification is used to categorize the students to shape their learning styles and inclination.

Educational Data mining can be implemented in many techniques such as decision trees, neural networks, k-nearest Neighbor, Naive Bayes, support vector machines and many others. Using these methods many kind of knowledge can be discovered such as association rules, classification, clustering, and pruning the data. Some of the Classification algorithms mentioned here for the proposed work have provided a better understand in educational resources.

A. Decision Tree Classifier

Decision Tree classifiers are used in data mining to produce trees after studying the training set and will be used to create predictions. Decision tree classifiers are one of the admired and influential tools for classification. Normally, decision tree classifiers have a tree-like structure which starts from root attributes, and ends with leaf nodes. It also has several branches consisting of dissimilar attributes, the leaf node on each branch representing a class or a kind of class distribution. Decision tree algorithms explain the relationship with attributes, and the comparative significance of attributes. The benefit of decision trees are that they characterize rules which could simply be understood and interpreted by users, do not need complex data preparation, and do well for numerical and categorical variables. The core algorithm for constructing decision trees called ID3 [12]. ID3 uses Entropy and Information Gain to build a decision tree. It is based on the theory of entropy which is a common way to calculate contamination. When Entropy is higher that means Information Content is additional.

B. Bayesian Classifiers

A Bayesian classifier is based on the scheme that the position of a (natural) class is to analyze the values of features for members of that class. Examples are grouped in classes because they have common values for the features. Such classes are frequently called natural kinds. In this section, the target feature corresponds to a separate class, which is not fundamentally binary.[13]

The thinking at the back of Bayesian classifier is that, if an agent knows the class, it can guess the principles of the other features. If it does not recognize the class, Bayes rule can be used to examine the class given (some of) the feature values. In a Bayesian classifier, the learning agent builds a probabilistic model of the features and uses that model to predict the classification of a fresh example.

Naive Bayes classifiers are extremely scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Naive Bayes is a easy technique for building classifiers: models that allocate class labels to problem instances, represented as vectors of feature values, where the class labels are strained from a few finite set. It is not a particular algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers suppose that the value of a particular feature is independent of the value of any other feature, given the class variable. Despite their adolescent design and apparently oversimplified assumptions, naive Bayes classifiers have worked relatively healthy in many difficult real-world situations.

A benefit of naive Bayes is that it only requires a small number of training data to calculate approximately the parameters compulsory for classification.

C. K-Nearest Neighbor Classifier (K-NN)

The k-Nearest Neighbor algorithms (k-NN) organize objects based on the neighboring training examples in the feature space. K-NN is a kind of instance-based learning, or lazy learning, where the function is only approximated nearby and the entire calculation is delayed in anticipation of classification. The main problem of k-NN algorithm is that its accuracy can be strictly ruined by the existence of loud or inappropriate features. Likewise, its accuracy becomes unfortunate if the feature balance are not reliable with their importance. In this paper, five classifiers that can run incrementally: the Naïve Bayes, Decision tree, Neural Network, SVM and Nearest neighbor (KNN) have been compared and observed that nearest neighbor algorithm gives enhanced accuracy compared to others if applied on Student Evaluation dataset [14].

D. Neural Network

Neural network is one more admired technique used in educational data mining. The benefit of neural network is that it has the capability to identify all possible connections between predictors variables. Neural network might also do a full detection without having any uncertainty even in difficult nonlinear relationship between dependent and independent variables. Therefore, neural network technique is chosen as one of the finest prediction method. Through the meta-analysis study, eight (8) papers have been published using Neural Network method. The papers shows an Artificial Neural Network model to guess student's performance [15]. The attributes analyzed by Neural Network are admission data, students thoughts towards self-regulated learning and scholarly performance.

Neural networks are being applied to an rising huge number of real world troubles. Their major benefit is that they can resolve problems that are too difficult for conventional technologies; troubles that do not have an algorithmic way out or for which an algorithmic solution is too complex to be defined. In general, neural networks are compatible to problems that community are good at solving, but for which computers usually are not. These problems contain pattern recognition and forecasting, which requires the recognition of trends in records. The proper supremacy and advantage of neural networks lies in their capability to signify both linear and non-linear relationships and in their ability to study these relationships openly from the data being modeled. Traditional linear models are simply not enough when it comes to modeling data that contains non-linear characteristics. The most familiar neural network model is the Multi-Layer Perceptron (MLP). This type of neural network is identified as a supervised network since it requires a preferred output in order to be trained. The objective of this kind of network is to make a model that properly maps the input to the output by means of historical data so that the model can then be used to generate the output when the desired output is unknown.

E. Support Vector Machine

In machine learning, support vector machines (SVMs, also maintain vector networks [16]) are supervised learning models with related learning algorithms that examine information used for classification and regression analysis. Given a set of training examples, all marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns latest examples to one class or the other, building it a non-probabilistic binary linear classifier. An SVM model is a symbol of the examples as points in space, mapped so that the examples of the split categories are separated by a understandable space that is as wide as feasible. Novel examples are then mapped into that identical space and predicted to fit in to a category based on which area of the gap they drop.

In addition to performing linear classification, SVMs can powerfully execute a non-linear classification by means of what is called the kernel trick, completely mapping their inputs into high-dimensional feature spaces.

When records are not labeled, supervised learning is not probable, and an unsupervised learning approach is necessary, which attempts to locate natural clustering of the data to groups, and then plot fresh data to these shaped groups. The clustering algorithm which provides an progress to the support vector machines is called support vector clustering and is repeatedly used in manufacturing applications either when data are not labeled or when only some data are labeled as a preprocessing for a classification pass.

SVMs are supportive in text and hypertext categorization as their application can extensively decrease the need for labeled preparation instances in both the standard inductive and transductive settings.

IV. PREDICTIVE ANALYTICS USING EDM

Using Educational Data mining techniques actually improves student's achievement and success more successfully in an efficient way. Students, Educators and Academic institutions get the profit and impacts. To identify most important attribute in the student dataset Prediction algorithms are used.

To obtain the prediction model based on attributes identified is the main aim of prediction algorithm. There are various classification algorithms that can be applied on the dataset using data mining tools. The induction rules and decision tree got by different classification algorithms help in understanding prediction model and compare their accuracy with other algorithms. We can identify the attributes that influence the classification the most and those that may not have much effect on prediction by analyzing the decision tree and induction rules. To find the most efficient way for our dataset classification we evaluate and compare accuracy of different cases [17].

V. CONCLUSION

To improve teaching and learning process of educators and learners performance prediction is very useful. This paper discusses the related work on evaluating the performance of students using different analytical methods. It talks about the techniques which applied on datasets to find out hidden information and pattern from database of educational environment. In EDM, the frequently used prediction technique is classification. Under classification different methods like Decision tree, Neural network, k-nearest neighbor, SVM, Bayes algorithm are used to predict student performance. The performance of all these methods is evaluated based on the parameters like accuracy, precision and recall. Here First, we will work on Entropy based feature selection and after that perform classification process using KNN-GA on standard student dataset to observe accuracy and compare with some previous results.

REFERENCES

[1]. C. Anuradha, T. Velmurugan, "A Data Mining Based Survey on Student Performance Evaluation System", *IEEE International Conference on Computational Intelligence and Computing Research,* DOI: 10.1109/ICCIC.2014.7238389, September 2015.

[2]. Hina Gulati, "Predictive Analytics using Data mining Technique", 2nd International Conference on Computing for Sustainable Global Development, May 2015.

[3]. Kiran Parmar, Dinesh Kumar Vaghela, Priyanka Sharma, "Performance Prediction of Students Using Distributed Data Mining" *IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems,* DOI: 10.1109/ICIIECS.2015.7192860, August 2015.

[4]. Alana M.de Morais, Joseana M.F.R. Araujo, Evandro B.Costa, "Monitoring Students Performance Using Data Clustering and Predictive Modelling", Frontiers in Education Conference, DOI: 10.1109/FIE.2014.7044401, February 2015.

[5]. Pratiyush Guleria, Niveditta Thakur, Manu Sood, "Predicting Student performance Using Decision Tree Classifiers and information Gain", *International Conference on Parallel, Distributed and Grid Computing*, DOI: 10.1109/PDGC.2014.7030728, February 2015.

[6]. Xin chen, Mihaela Vorvoreanu, Krishna Madhavan, "Mining Social Media Data for Understanding Students Learning Experiences", *IEEE Transactions on Learning technologies,* Vol. **7**, No.3,September 2014.

[7]. Hashmia Hamsa, Simi Indiradevi, Jubilant J. Kizhakkethottam, "Student Academic Performance Prediction Model Using Decision Ttree and Fuzzy Genetic Algorithm", *Procedia Technology*, **25**, pp.326-332,2016.

[8]. Amirah Mohamed Shahiria, Wahidah Husaina, Nuraini Abdul Rashida," A Review on Predicting Student's Performance using Data Mining Techniques", *Procedia Computer Science*, **72** pp.414 – 422,2015.

[9]. Ruhi R. Kabra, R. S. Bichkar, "Students Performance Prediction Using Genetic Algorithm", *International Journal of Computer Engineering and Applications*, Vol. **6**, Issue 3, June 14.

[10]. P. Usha, "Predicting student Performance Using Genetic and SVM Classifier", *International Journal of Computer Engineering*, Vol. **3**, No. 2, pp. 97–102, December 2011,

[11]. Dorina Kabakchieva, "Student Performance Prediction by Using Data Mining Classification Algorithms', *International Journal of Computer Science and Management Research,* Vol. **1**, Issue 4, November 2012.

[12]. Himani Sharma, Sunil Kumar "A Survey on Decision Tree Algorithms of Classification in Data Mining", *International Journal of Science and Research (IJSR)* ISSN (Online): 2319-7064.

[13]. Henry Xiao "Bayesian Approaches Data Mining Selected Technique", CISC 873 Data Mining – p. 1/17.

[14]. Devroye, L., Gyorfi, L. & Lugosi, G. "A probabilistic theory of pattern recognition", In Springer 1996. ISBN 0-3879-4618-7.

[15]. Chady El Moucary, Marie Khair, Walid Zakhem "Improving Student's Performance Using Data Clustering and Neural Networks in Foreign-Language Based Higher Education", *The Research Bulletin of Jordan ACM - ISWSA*; ISSN: 2078-7952 (print); 2078-7960 (online).

[16]. Vapnik, V.N. "Statistical learning theory" John Wiley & Sons. 1998.

[17]. Pooja Thakar, Anil Mehta, Manisha "Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue", *International Journal of Computer Applications* (0975 – 8887) Volume **110** – No. 15, January 2015.