Volume 6, Issue 12, December 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Review on Predicting Students' Academic Performance Earlier, Using Data Mining Techniques

P. Kavipriya

Assistant Professor, Department of Computer Science, CMS College of Science and Commerce,
Coimbatore, Tamilnadu, India

Abstract: The methods of Data mining are used in evaluating the given data and to mine the unknown facts and knowledge which greatly supports the researchers to take effective decisions. Due to the tremendous growth in recent technology like social media, it may divert the students from their actual track, and this is one of the reasons for the students to perform poor in academic activities and it even leads to course drop outs. This paper reviews the previous research works done on students' performance prediction, analysis, early alert and evaluation by using different methods of data mining.

Keywords: Performance Prediction, analysis, Academic activities, early alert, Data Mining Methods

I. INTRODUCTION

In the recent decays, the exploration in the educational field is increasing rapidly due to the colossal growth of data that related to the performance of students' academics. It increases the accuracy and quality of students' performance by predicting it in-advance. Early Prediction of students' performance is to enhance the quality of education in various traits. It helps in predicting at risk students during the course time itself and not at the end of the course. It acts as an effective tool that provides information to change educators' practices and make an alert to help students get back on track. It will be a successful method for improving academic success and retention. This early prediction benefits the students to take necessary steps in advance to avoid poor performance and to improve their academic scores. It benefits both the course instructors as well as the students whose performance is lagging in class.

The feature of this prediction system is that it can be used early and as needed in prior of semester for faculty to communicate their concerns to students when signs of risk occur. So that students graduate on time without reappearing in the semester and are prepared to flourish in actual course and get back to careers on time. Student's academic performance in educational environment is based upon the mental and environmental factor can predicted by using various data mining techniques.

The data collected for this educational practice i.e. data from traditional learning system will greatly avoid proxy method where as in E-Learning system it is not possible. This approach will support to predict the students' performance by analyzing the student's academic record such as internal assessment marks, assignment submission, and attendance percentage. The techniques reviewed in this paper are classification, clustering, ensemble and many others.

In this paper the study is structured as follows. In Section 2, we reviewed previous work related to students' performance. Section 3, we explained and evaluated the Existing System with the techniques they used. Section 4, describes the various data mining techniques with their merits. Section 5, presents the discussion for proposed work. Section 6, concludes the Review and in Section 7, scope for future study.

II. RELATED PREVIOUS WORK

(Xinga, Chen, Steinc, & Marcinkowsk, 2016) **[1]** , present a paper to predict a dropouts in online courses: for this prediction they applied, two algorithms called General Bayesian Network (GBN) and decision tree (C4.5) are implemented to reduce the higher dropout rates in online education .

(Han Hu a, Lo a & Shih, 2014) **[2]**, Develops the paper to alert students by some warning to improve their performance earlier. In their study they used three well-known single classification techniques, C4.5, CART, and LGR and made a comparative study to develop a system to predict student's E- learning performance by giving early alert.

(Wanli, Eva & Sean, 2015) **[3]**, presents a paper to predict final year student performance as a model through Genetic Programming, and to Integrate the learning analytics, educational data mining and theory they demonstrate the structure for connecting trace data to a hypothetical framework, the processing data using the algorithm of Genetic Programming approach which outperforms the traditional models in prediction rate and also in interpretability.

(Campagni, Merlini, Sprugnoli, & Verri 2015) **[4]**, Develops a model for student careers, in this paper they present different approaches based on clustering and sequential patterns techniques in order to identify strategies for improving the performance of students and the scheduling of exams.

(Sembiring, et. al., 2011) **[5]**, Made the presentation on Predicting students' performance academically by applying some of the techniques of data mining. The results of this study stated a model of student academic performance predictors by employing (psychometric factors) as variables predictors.

(Kalpana, et. al., 2014) **[6]**, in this paper they made analysis on students Intellectual Performance by Using Data Mining Techniques. This presentation intends to analysis the student's performance in different categories of measurements.

(Kaura, Singh & Josanc, 2015) **[7]** ,In this paper they made a comparison study to predict slow learners in educational sector using Classification and prediction built data mining algorithms, in this paper they targeted the slow learners and the output dataset is tested and analyzed with five classification algorithms which are Multilayer Perception, Naïve Bayes, SMO, J48 and REP Tree.

(Sullare, Thakur, Mishra, 2016) **[8]**, presents a paper on students' Performance, based on Grouping up of Neighbors Students in Progressive Education Datasets. In this paper they used Naive Bayes clustering method to assess student's performance in the end semester examination from education databases.

(Nagar et. al., 2015) **[9]**, presents a paper on (Data Mining Clustering Methods), in this paper they made a detailed comparison study on different clustering techniques, an unsupervised learning method which makes the cluster of bits and pieces or forms according to their similarity and dissimilarity bases. This paper gives review about various clustering methods.

(Baradwaj, Pal, 2011) **[10],** presents a paper to mine the educational data by analyzing the students' academic performance, for this analysis they used decision tree method for identifying the dropouts and predict students who need special attention and makes the work of educators easier in providing some appropriate warning or advising.

III. EXISTING APPROACHES FOR PREDICTING TECHNIQUES ON STUDENTS DATA

[1]Presents General Bayesian Network (GBN) and decision tree (C4.5) are implemented to reduce the higher dropout rates in MOOCs online education. This study only works for online dataset and not suited for traditional learning system. [2] Developing early warning systems to predict students' learning performance using C4.5, CART, and LGR techniques. Only applicable for online courses, and accuracy can be greatly improved by using various other techniques.[3] Presents student final performance prediction model using Genetic Programming, in this paper the prediction rate described can be fine-tuned.[4]Developed a model for student careers presents different approaches based on clustering and sequential patterns can greatly improve its performance and data set can be fine-tuned.[5] Presentation on Prediction of student academic performance by an application of data mining techniques results student academic performance predictors by employing psychometric factors as variables predictors seems to be reduced performance. [6] Made the presentation on Intellectual Performance Analysis of Students by Using Data Mining Techniques intends to analysis the student's performance in different categories of measurements. The parameters of this study can be fine-tuned. [7] Presents to predict slow learners in education sector, the performance can be improved by using advanced data mining techniques. [8] Develops a Performance Based Grouping of Neighbors Students in Progressive Education Datasets by Naive Bayes clustering method to assess student's performance in the end semester examination from education databases. The speed can be greatly improved by using some other classification techniques. [9] Presents Data Mining Clustering Methods gives review about various clustering methods. Performance can be fine-tuned in this study.[10] Developed a Mining Educational Data to Analyze Students' Performance used decision tree method that helps in identifying the dropouts and students who need special attention and allow the teacher to provide appropriate advising/counseling. Speed can be improved and it is applicable only for post graduate students.

IV. DATA MINING TECHNIQUES

Data mining is concerned with the analysis of data and they use different software techniques to find the hidden and unexpected patterns and their relationships in data set. The work of mining the data is to extract the information that is unknown and unpredicted. Generally data mining contains several algorithms and techniques for finding out interesting patterns from large size of data sets. The techniques of data mining are classified into two groups: they are supervised learning and unsupervised learning. In supervised learning, a model is built earlier for the analysis and then it applies the algorithm to the data in order to estimate the parameters for the model. Classification, Association Rule Mining ,Decision Tree, Bayesian Classification, Neural Networks, etc. are some examples of supervised learning. In unsupervised learning, we do not create a model or assumption earlier for the analysis; it just applied the algorithm directly to the dataset to detect its result. Later a model can be created on the basis of the obtained results. Clustering is one of the examples that falls under the category of unsupervised learning. Various data mining techniques such as Classification, Decision Tree, Bayesian Classification, Neural Networks, Clustering, Association Rule Mining, Prediction, Time Series Analysis, Sequential Pattern and Genetic Algorithm and Nearest Neighbor have been used for knowledge discovery from large data sets. Some of the common and useful data mining techniques have been discussed.

A. Decision tree

Decision tree in data mining is one of the simplest and easiest methods which are most frequently used by the researchers on their work. The root node of the decision tree is a top node resembles simple question also called as a posture that bears multiple branches called sub nodes with answers for the root node question. In turn each answer related to a set of questions or conditions that help us to predict the data, on which the final decision is made.

ID3 and C4.5 are called as induction algorithm of decision tree developed by the researcher called Ross Quinlan. Both algorithm supports greedy method, top-down recursive in divide-and-conquer manner and they does not support backtracking.C4.5 is also known as superset of ID3.The advantages of this technique are, it doesn't require detailed knowledge, it deals with complex data, these are easy to understand, and data Classification becomes simpler, makes learning easier, it produces very accurate end result.

B. Naive Bayes Algorithm

A simple probabilistic classifier that based on applying Bayes theorem (from Bayesian statistics) with strong (naive) independence assumptions is the Naive Bayes classifier. "Independent feature model" is a more descriptive term for the underlying probability model. In simple terms, a Naive Bayes classifier adopts a particular feature in the presence (or absence) of a class that is unrelated to any other feature. For example, a vegetable may be considered to be a carrot if it is orange in color, it shaped as in cone, and about 10 to 15 centimeter long in length. If these features match on each other or match upon the existence of the other features, a Naive Bayes classifier then considers all these properties independent contribution to the probability that this vegetable is a carrot. Based on the exact nature of the probability model, the data set is then trained very effectively by a Naive Bayes classifier in a supervised learning setting. In most of the practical applications, the parameter estimation for Naive Bayes models uses the method of maximum likelihood; in other words, one can also work with the Naïve Bayes model without considering the Bayesian probability or by using any Bayesian methods. In spite of their over-simplified assumptions of a Naive design, the Bayes classifiers will work effectively in solving many complex real-world circumstances. There are several advantages of Naive Bayes classifier such as; it requires a small amount of training data set to estimate the parameters (the means and variances of the variables are used as parameters) essential for classification. This is Because, the independent variables are assumed only to the variances of the variables of each class that is needed to be determined and not the entire covariance matrix, irrelevant features are removed to improves the classification performance, results High Performance and take very less computational time.

C. Support Vector Machine

A powerful Support Vector Machine (SVM) which was first proposed by Vapnik and it has a great potency of interest in the machine learning research community. Several past studies have reported that the SVM generally has a proficient of delivering the high accuracy in classification when compared to other data classification algorithms. Though, for certain datasets, the achievement of SVM is very subtle in determining the cost parameter and kernel parameters. As in the case of closure, to figure out the most encouraging condition environment the user normally needs to conduct extensive cross validation. Basically this technique is baited to as a model selection. A superior asset of this SVM technique is that, concurrent miniaturize the projected classification error and make best use of the geometric margin, So SVM is also named as paramount Margin Classifiers. It is found on the Structural Risk Minimization (SRM), SVM can be used for both classification and prediction. There are several advantages of SVM such as it uses maximum marginal hyper plane for classifying linearly separable data, Data can be separated clearly into rations, extends by itself in order to classify the linearly inseparable data.

D. DT-J48

The J48 Decision tree classifier is an implementation of ID3 classifier used to create a decision tree based on the attribute values of the available training data set. So, whenever it encounters a set of items i.e. training set it identifies the attribute that discriminates the various instances most visibly. This feature tells us about the data instances so that we can classify them clearly and the best is said to have the highest information gain. Now, among the possible values of this feature, if there is any value for which there is no ambiguity, that is, for which the data instances falling within its category have the same value for the target variable, then we terminate that branch and assign to it the target value that we have obtained.

E. Neural Networks

The area of neural networks probably belongs to the border line between the artificial intelligence and approximation algorithm. A neural network is a pool of neurons like processing units with weighted connection between the units. It composes of many elements, called nodes which are connected in between. The connection between two nodes is weighted and by the adjustment of this weight, the training of the network is performed. There are many advantages of neural networks such as adaptive learning ability, self-organization, real time operation and insensitivity to noise. Neural networks are used to identifying patterns or trends in data and well suited for prediction or forecasting needs. There are several neural network algorithms such as Back Propagation, NN Supervised Learning, and Radial Base Function (RBF) Network etc.

F. Clustering

The method of grouping object that have similar characteristics is referred to us as Data Clustering. The criterion for checking the similarity is implementation dependent. Clustering is often confused with classification, but there are vast differences between them. In classification the objects are assigned to a predefined class, where as in clustering the classes are also to be defined. Clustering methods may be divided into two categories based on the structure of cluster, they are named as hierarchical cluster and partitioning cluster.

G. Ensemble Clustering

Cluster ensembles is the assignment to pool multiple clustering's of a group of objects into a single consolidated clustering, often referred to as the consensus clustering. It is also called as aggregation of clustering. Compare to an individual clustering method it can be used to generate more stout and stable clustering results. Ensemble is to perform distributed computing under privacy or in sharing constraints, or to reuse existing knowledge.

Cluster ensemble can produce either by

- Using different clustering algorithms
  E.g. K-means, Hierarchical Clustering, Fuzzy C-means, Spectral Clustering, Gaussian Mixture Model…
- Running the same algorithm for many times with different parameters or initializations.

E.g., run K-means algorithm N times using randomly initialized clusters centers, or by using different dissimilarity measures, or by various number of clusters.

V. DISCUSSION RELATED TO PROPOSED WORK

This paper, make a review on different students' performance based on data mining techniques under different circumstances. From this review, we can improve the speed of predicting the result. in the previous study lot of time taken in training the dataset i.e., more than 60% time in training and the rest in testing .if we reduce the time in training ,it greatly improve speed .the step taken to reduce the training dataset may be by using clustering techniques. And also to improve the quality of students' performance to make an early prediction with some classification techniques and ensemble clustering can also be used for the same.as an enhanced work of proposed system the real time data from any reputed colleges or from universities can be used for the betterment of effective result.

The below mentioned table showed some of the actual accuracy rate predicted in previous study by using some classification techniques.

Table 1: Comparison of Classification Methods .

Classification Methods	Accuracy Rate
Decision Tree	94%
Naive Bayes	96%
Neural Network	91%
Support Vector Machine	97%
DT-J48	95%

VI. CONCLUSION

In this paper, we reviewed different classification method used on student database to predict the student's performance in the upcoming semester on the basis of previous student's database and the work done on this till now. As we have seen, predicting students' performance earlier is a difficult task because it is a multifaceted problem and because the available data are normally imbalanced. To resolve this problem and to improve the accuracy and quality, the Support Vector Machine algorithm can be used which is showing the greatest accuracy among other techniques. Clustering technique and ensemble cluster can also be used to fine tune the quality of resulting dataset. Information's like Attendance, Seminar and Assignment marks were collected from the student's database, to predict the performance at the mean time of the course. The other attributes are collected by students and their respective faculties who know the behavior of students. This study will help to the students and the teachers to improve the performance of the students who are at the risk of failure. This study will also work to identify those students who needed special attention to reduce fail ration and taking appropriate action for the current semester examination.

VII. FUTURE WORK

In place of future work, supposed to do the research by using various classifications and clustering applications to enhance the prediction speed and accuracy in the field of education.

REFERENCES

[1] Nkitaben Shelke, Shriniwas Gadage,"A Survey of Data Mining Approaches in Performance Analysis and Evaluation", (2015), International Journal of Advanced Research in Computer Science and Software Engineering

[2] Harwatia, Ardita Permata Alfiania, Febriana AyuWulandaria, "Mapping Student's Performance Based on Data Mining Approach (A Case Study)" ScienceDirect,Agriculture and Agricultural Science Procedia 3 (2015) 173 – 177.

[3] Renza Campagni, Donatella Merlini, Renzo Sprugnoli, Maria Cecilia Verri "Data mining models for student careers", at Science Direct Expert Systems withApplications,pp55085521,2015,www.elsevier.com.

[4] S. Archna, Dr. K. Elangovan ―Survey of Classification Techniques in Data Mining‖, International Journal of Computer Science and Mobile Applications vol 2, Issue 2,, February 2014, p.g. 65-71

[5] Rajni Jindal and Malaya Dutta Borah, "A SURVEY ON EDUCATIONAL DATA MINING AND RESEARCH TRENDS, International Journal of Database Management Systems (IJDMS) Vol.5, No.3, June 2013.

[6] Random Forest Algorithm : Data Science Control

[7] An Overview of Data Mining Techniques Excerpted from the book Building Data Mining Applications for CRM by Alex Berson, Stephen Smith, and Kurt Thearling [12] Kriegel, Hans-Peter; Kröger, Peer; Sander, Jörg; Zimek, Arthur (2011). "Density-based Clustering". WIREs Data Mining and Knowledge Discovery 1 (3): 231–240.doi:10.1002/widm.30.

[8] Advantages of Bayesian Networks in Data Mining and Knowledge Discovery By Petri Myllymäki , Ph.D., Academy Research Fellow, Complex Systems Computation Group, Helsinki Institute for Information Technology.

[9] Sajadin Sembiring, M. Zarlis, Dedy Hartama, Ramliana S, Elvi Wani "Prediction of Student Academic Performance by an Application of Data Mining Techniques" 2011 International Conference on Management and Artificial Intelligence IPEDR vol.6,pp 110-114,2011.

[10] J. K. Jothi Kalpana, K. Venkatalakshmi " Intellectual Performance Analysis of Students by Using Data Mining Techniques" ,International Journal of Innovative Research in Science, Engineering and Technology, Volume 3, Special Issue 3, March 2014, & 2014 IEEE International Conference on Innovations in Engineering and Technology (ICIET'14) On 21st&22ndMarch, Organized by K.L.N. College of Engineering, Madurai, Tamil Nadu, India ISSN (Online) : 2319 - 8753 ISSN (Print) : 2347 – 6710.pp 1922-1929, 2014.

[11] Manoj Bala et al., "Study of Application of Data Mining Technique in Education", International Journal of Research in Science and Technology, Vol. No. 1, Issue No. IV, Jan-March, 2012.

[12] Kotsiantis, S. B., & Pintelas, P. E. (2005). Predicting students marks in hellenic open university. Paper presented at the 5th IEEE international conference on advanced learning technologies, Kaosiung, Taiwan.

[13] Kotsiantis, S. B., Pierrakeas, C. J., Zaharakis, I. D., & Pintelas, P. E. (2003)." Efficiency of machine learning techniques in predicting studentsa¯ performance in distance learning systems", Paper presented at the Symposium on recent advances in mechanics, Athens, Greece.

[14] P. Golding, S. McNamarah, ―Predicting Academic Performance in the School of Computing & Information Technology (SCIT),‖ Proceedings of 35th ASEE /IEEE Frontiers in Education Conference, 2005.

[15] P. Golding, O. Donaldson, ―Predicting Academic Performance‖, Proceedings of 36th ASEE /IEEE Frontiers in Education Conference, 2006.

[16] J. Zimmermann, K. H. Brodersen, J. P. Pellet, E. August, J. M. Buhmann, ―Predicting graduate-level performance from undergraduate achievements,‖ Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, the Netherlands. July 6-8, 2011.

[17] T. N. Nghe, P. Janecek, P. Haddawy, ―A Comparative Analysis of Techniques for Predicting Academic Performance,‖ Proceedings of 37th ASEE /IEEE Frontiers in Education Conference, 2007.

[18] D. Kabakchieva , K. Stefanova, V. Kisimov, Analyzing University Data for Determining Student Profiles and Predicting Performance, Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, the Netherlands. July 6-8, 2011.

[19] Amirah Mohamed Shahiria,∗, Wahidah Husaina, Nur'aini Abdul Rashida," A Review on Predicting Student's Performance using Data Mining Techniques",Science Direct, Procedia Computer Science 72 (2015) 414 – 422.