ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 5, Issue 11, November 2018

Students' Academic Performance Prediction Using MadaBoost Algorithm

V.A.Malathi , C.V.Banupriya

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: Educational Data Mining (EDM) is the field of study concerned with mining educational data to find out interesting patterns and knowledge in educational organizations. A new boosting algorithm MadaBoostis proposed by modifying the weighting system of AdaBoost. MadaBoostis used to classify and predict values. Educational Data Mining is no exception of this fact, hence, it was used in this research paper to analyze collected students' information from machine learning dataset repository. It affords classification based on the collected data and predict students' performance in their upcoming semester. The objective of this study is to identify relations between students 'personal and social factors with their academic performance. This newly discovered knowledge can help students as well as instructors in carrying out better enhanced educational quality by identifying possible underperformers at the beginning of the year and apply more attention to them in order to help them in their education process and get better marks. In fact, not only underperformers can benefit from this research but also possible well performers can benefit from this study by employing more efforts to conduct better projects and research through having more help and attention from their instructors.

KEYWORDS: Educational data mining, AdaBoost, MadaBoost, Dataset, Classification, Prediction.

I.INTRODUCTION

Data Mining (DM) is a process used by companies to turn raw data into useful information. Data mining is looking for hidden, valid and potentially useful patterns in huge data sets. Data Mining is all about discovering unsuspected or previously unknown relationships in the midst of the data. Data mining is also called as knowledge discovery, knowledge extraction, data or pattern analysis and information harvesting etc. 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. Data Mining is a process that analyzes a large amount of data to find new and hidden information that improves business efficiency. The ultimate goal of data mining is prediction. Predictive data mining is the most common type of data mining and that has the most direct business applications.

A. Knowledge Discovery Database in Data Mining [KDD]

Knowledge Discovery in Databases is an emerging field combining techniques from databases, statistics and artificial intelligence, which is concerned with the theoretical and practical issues of extracting from volumes of low level data. The process of finding and interpreting patterns from data involves the repeated application of the following steps:

Data Cleaning: Data cleaning is defined as removal of noisy and irrelevant data from collection.

- Cleaning in case of missing values.
- Cleaning noisy data, where noise is a random or variance error.
- Cleaning with data discrepancy detection and data transformation tools.

Data Integration: Data integration is defined as heterogeneous data from multiple sources combined in a common source or called Data Warehouse.

- Data integration using data migration tools.
- Data integration using data synchronization tools.

ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 5, Issue 11, November 2018

- Data integration using ETL (Extract-Load-Transformation) process.

Data Selection: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.

- Data selection using neural network.
- Data selection using decision trees.
- Data selection using naive bays.
- Data selection using clustering, regression, etc.

Data Transformation: Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure. Data Transformation is a two step process:

- Data mapping: Assigning elements from source base to destination to capture transformations.
- Code generation: Creation of the actual transformation program.

Data Mining: Data mining is defined as clever techniques that are applied to extract patterns potentially useful.

- Transforms task relevant data into patterns.
- Decides purpose of model using classification or characterization.

Pattern Evaluation: Pattern Evaluation is defined as identifying strictly increasing patterns representing knowledge based on given measures.

- Find interestingness score of each pattern.
- Uses summarization and visualization to make data understandable by user.

Knowledge representation: Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.

- Generate reports and tables
- Generate discriminate rules, classification rules, characterization rules etc.
-

B. Educational Data Mining (EDM)

Applying data mining in education is an emerging interdisciplinary research field also known as Educational Data Mining (EDM).Data Mining is 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.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. Educational Data mining describes the following four goals:1) Predicting student's future learning behavior 2) Discovering or improving domain models 3) Studying the effects of educational support 4) Advancing scientific knowledge about learning and learners.

Educational Data Mining Methods

Prediction: Prediction in data mining is to identify data points purely on the description of another related data value. It is not necessarily related to future events but the used variables are unknown.Prediction derives the relationship between a thing to know and need to predict for future reference.

Clustering: The cluster is actually a collection of data objects; those objects are similar within the same cluster. That means the objects are similar to one another within the same group and they are rather different or they are dissimilar or unrelated to the objects in other groups or in other clusters.

Relationship Mining: There are different types of relationship in mining techniques such as association rule mining or any relationship between variables, sequential pattern mining, correlation mining or linear correlations between variables and causal data mining or causal relationship between variables. In EDM, relationship mining has been used to identify relationships in learners' behavior patterns and diagnosing students' learning difficulties or mistakes that frequently occur together.

Discovery with Models: The goal of discovering with models is to use a previously validated model of a phenomenon as a component in another analysis such as prediction or relationship mining.

Distillation of Data for Human Judgment: The goal is to represent data in intelligible ways using summarization, visualization and interactive interfaces to highlight useful information and support decision making. In

ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 5, Issue 11, November 2018

EDM, it also known as distillation for human judgment and it has been used for helping educators to visualize and analyze the students' course activities and usage information.

Regression: In statistical terms, a regression analysis is the process of identifying and analyzing the relationship among variables. It can help you understand the characteristic value of the dependent variable changes, if any one of the independent variables is varied. This means one variable is dependent on another, but it is not vice versa. It is generally used for prediction and forecasting.

Outlier Detection: 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 students with learning difficulties, deviations in the learners or educator's actions or behaviors and for detecting irregular learning processes.

II.LITERATURE REVIEW

Prediction of student's academic performance by **JyotiBansode2016** suggested that in data mining prediction can be done by using students' academic background and family background. In this study, using decision tree students' performance can be predicted. The students, whose performances are poor, can be warned. The management can take necessary action to improve their performance by giving more attention, taking extra lectures etc. Due to such measures student performance can be improved. The number of failures can be reduced. Ultimately college results also get improved.

Mining Educational Data to analyse studentsperformanc eproposed by **Brijesh Kumar Baradwaj***et.al.***2011**is designed to justify the capabilities of data mining techniques in context of higher education by offering a datamining model for higher education system in the university.Inthis study, the classification task is used to evaluate student's performance and as there are many approaches that are used fordata classification, the decision tree method is used here. Classification taskis used to evaluate student's performance and as there are many approaches that are used for data classification, the decision tree method is used here. Information's like attendance, classtest, seminar and assignment marks were collected from the student's management system to predict the performance at the end of the semester.

To predict students at risk of poor performance by **ZahyahAlharbi***et.al.***2016** described data mining can be used to highlight performance problems early on and propose remedial actions. Data collected through the admission process and through the students' degrees. In this study, predict good honors outcomes based on data at admission and on the first year module results. To validate the results, evaluate data relating to students with different characteristics from different schools.

Predicting dropout student by ErmanYukselturk*et.al.***2014**concluded that to classify the dropout students, four data mining approaches were applied based on k-Nearest Neighbour (k-NN),Decision Tree (DT), Naive Bayes (NB) and Neural Network (NN). These methods were trained and tested using 10-fold cross validation. The detection sensitivities of 3-NN, DT, NN and NB classifiers were 87%, 79.7%, 76.8% and 73.9% respectively. Also, using Genetic Algorithm (GA)based feature selection method, online technologies self-efficacy, online learning readiness, and previous online experience were found as the most important factors in predicting the dropouts.

III.RESEARCH METHODOLOGY

A. EXISTING SYSTEM-ID3 and k-NN

Iterative Dichotomiser 3 – [ID3]

In decision tree learning, one of the algorithms is the **ID3** algorithm or the Iterative Dichotomiser 3 algorithm. It is used to generate a decision tree from a dataset and also is considered as a precursor to the c4.5 algorithm.ID3 decision tree classification algorithm adopt a greedy or non backtracking approach in which decision trees are constructed in a top-down recursive divide and conquer manner. Most algorithms for decision tree induction also follow such a top-down approach, which starts with a training set of tuples and their associated class labels. The training set is recursively partitioned into smaller subset as the tree is being built.

K-Nearest Neighbor[k-NN]

The K-Nearest Neighbor (k-NN) classifieris one of the supervised learning simple algorithms. It has high accuracy and versatile. It is useful for classification or regression. The k-NN is a simple technique, that it is easily implemented. It well suited for multi-modal classes and records with multiple class labels.K-nearest neighbours uses the local neighbourhood to obtain a prediction. Both ID3 and k-NN algorithms are used in the existing system.

ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 5, Issue 11, November 2018

B. PROPOSED SYSTEM

The proposed work consists of the following essential steps. In first step, researcher collected data from machine learning repository. Pre processing is the main step after collected data. Applying proposed algorithm MadaBoost on dataset for predicting students' academic performance.

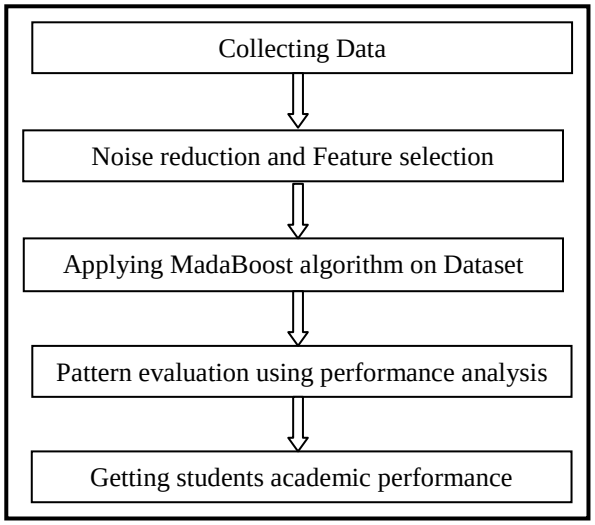

Fig1: Steps for Proposed Work

C. MadaBoost - MODIFIED ADAPTIVE BOOSTING ALGORITHM

AdaBoostis defined for the sub sampling framework, where a sample of sufficient size, which is randomly selected before the boosting is fixed throughout all the boosting process and distributions are defined only with respect to the sample.Themodification for MadaBoostis very simple. It's just bound the weight assigned to every example by its preliminary probability. In this way, the weights of the examples cannot become randomly large as it happens in AdaBoost.MadaBoostworks under the filtering framework and it belongs to the statistical query model.

This newly discovered knowledge can help students as well as instructors in carrying out better enhanced educational quality by identifying possible underperformers at the beginning of the semester or year, and apply more attention to them in order to help them in their education process and get better marks. In fact, not only underperformers can benefit from this research but also possible well performers can benefit from this study by employing more efforts to conduct better projects andresearch through having more help and attention from their instructors.

Researcher proposed a new boosting algorithm that mends some of the problems that have been detected in the so far most successful boosting algorithm, these problems are:

(1) AdaBoost cannot be used in the boosting by filtering framework

(2) AdaBoost does not seem to be noise resistant.

In order to solve them, investigator proposes a new boosting algorithm MadaBoost by modifying the weighting system of AdaBoost. This algorithm proved that the new boosting algorithm can be casted in the statistical query learning model and thus, it is robust to random classification noise.

Advantages of MadaBoost Algorithm

- This new boosting algorithm can be casted in the statistical query learning model.
- MadaBoost Algorithmis robust to random classification noise.
- It provides a much improved analysis of its correctness and performance.
- By adding more variables into the data set, the MadaBoost Algorithm can produce more accurate result in an easy mode.
- It is easily predict future academic performance of students' activities.
- Students can improve their performance and teachers also changing their classroom teaching interaction techniques.

ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 5, Issue 11, November 2018

IV.DESIGN AND IMPLEMENTATION

A. STUDENTS' ACADEMIC PERFORMANCE PREDICTION SYSTEM

Students' academic performance prediction system used modification of adaptive boosting classification algorithm to predict the future performance of the students. In this proposed work makes use of 21 students' academic, social and economic related attributes for the predication of student performance from 500instances or data.

B. MODULE DESIGN

View Data Set

View Data Setis the first module of student academic performance prediction system. This module contains 500 instances, each instance described with 21 attributes of students' personal, academic and social related attributes.

Dataset Pre-processing

Once the right data is selected, pre-processing includes selection of the right data from the complete dataset and building a training set. Three steps involved in data pre-processing such as, Organise and format, Data cleaning, Feature extraction. Classroom data sets of various attributes scattered in data base which needs to be organized together to form a dataset. Cleaning refers to mainly dealing with the missing values and removal of unwanted characters from the data. One has to find out which features are important for prediction and select them for faster computations and low memory consumption.

Training Data Set Selection

Separating data into training and testing sets is an important part of evaluating data mining models. In a data set for implementing process, most of the data is used for training and a smaller portion of the data is used for testing. Training data is also known as a training set, training dataset or learning set.In this proposed study, the dataset contains 500 instances were obtained from Kaggle Machine Learning 'Students' academic performance data set repository'.

MadaBoost is a popular boosting technique which helps to combine multiple weak classifiers into a single strong classifier. A weak classifier is simply a classifier that performs poorly but performs better than random guessing. Each weak classifier should be trained on a random subset of the total training set. AdaBoost assigns a "weight" to each training example, which determines the probability that each example should appear in the training set. Examples with higher weights are more likely to be included in the training set. After training a classifier, MadaBoost increases the weight on the misclassified examples so that these examples will make up a larger part of the next classifiers training set and hopefully the next classifier trained will perform better on them.

Predictions with MadaBoost

MadaBoost isthe first practical boosting algorithm invented by Freund and Schapire (1995). It is based on Vapnik and Chervonekis idea that for a trained classifier to be effective and accurate in its predictions, it should meet these three conditions:1) Classifier should be trained on enough training examples2) It should provide a good fit to these examples by producing low training error3) It should be simple

C. DATASET DESCRIPTION

A Dataset can often viewed as a collection of data objects. These data objects are described by an attributes that capture the basic characteristics of an object. This proposed dataset contains a total of 500 instances, after preprocessing 486 instances were taken for the prediction process. A total of 500 instances were obtained from the Kaggle Machine Learning Data set repository. It contains 21 main attributes like Gender, Previous semester marks, Parent Qualifications etc. from the database are considered for current proposed work. The testing algorithm is classified students performance prediction into two categories 1. Student will get improve 2. Student will not get improve by analyzing the available student data with selected data mining methods from classification. Java NetBeans IDE is free and open source software used for this proposed work.

V.EXPERIMENTAL RESULTS

A. RESULTS AND DISCUSSIONS

The experimental results proved that after classifying the data and predict the values using MadaBoost algorithm, it plots performance of the current proposed system in an easily understandable way. The performance chart displayed comparison of precision values between existing algorithms (ID3, k-NN) and proposedalgorithm (MadaBoost). In the same way, Recall and Accuracy values are displayed clearly. MadaBoost has highest accuracy 92% than the ID3 (78%) and k-NN (70.1%) algorithms.

ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 5, Issue 11, November 2018

PRECISION

Precision takes all retrieved data into account, but it can also be evaluated at a given cut-off rank, considering only the topmost results returned by the academic performance prediction system. Precision is the number of correct results divided by the number of all returned results. Precision is also used with recall, the percent of all relevant data that is returned by the search.

RECALL

Recall in information retrieval is the fraction of the datasets that are relevant to the query that are successfully retrieved.The text search on a set of dataset recall is the number of correct results divided by the number of results that should have been returned. In classification, recall is called sensitivity.

ACCURACY

The True Positive (TP) is the number of data is correctly identified as relevant data. False Positive (FP) is the number of non-relevant data that are incorrectly identified, True Negative (TN) refers to the number of non-relevant data that are correctly identified as non-relevant data and False Negative (FN) is the number of data that are incorrectly identified as non-relevant data. This method is fast, robust and efficient.

Table1:Precision, Recall and Accuracy Values

Number of Data	Precision Values			Recall Values			Accuracy Values		
	ID3	k-NN	Mada Boost	ID3	k-NN	Mada Boost	ID3	k-NN	Mada Boost
100	70.3	65.9	81.9	76.5	67	88	76.2	68	88.9
200	72.1	66.4	84.9	76.9	68.7	88.3	76.9	68.4	89
300	74.5	67.9	87.3	77.4	69.1	89.1	77	69.1	89.3
400	75.1	68.4	90.2	77.9	69.4	89.3	77.5	69.5	90
500	79.6	70.5	92	78	70	91	78	70.1	92

B. PERFORMANCE CHARTS

Precision, Recall and Accuracy values calculated and displayed both existing and proposed algorithms through performance chart.

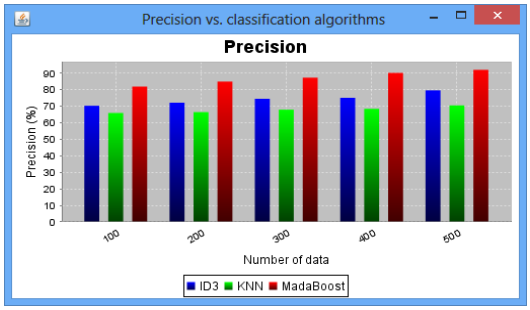

Fig2: Precision Comparison

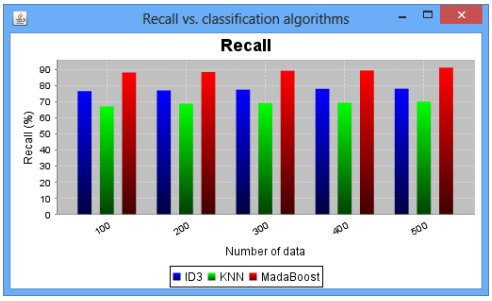

Fig3: Recall Comparison

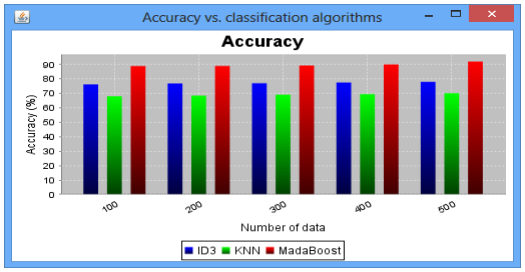

Fig4: Accuracy Comparison

ISSN: 2350-0328

International Journal of AdvancedResearch in Science, Engineering and Technology

Vol. 5, Issue 11, November 2018

VI.CONCLUSION AND FUTURE ENHANCEMENTS

In this proposed work, multiple data mining tasks were used to create qualitative predictive models which were efficiently and effectively able to predict the students' grades from a collected training dataset. The collected dataset was pre-processed and explored to become appropriate for the data mining tasks. Third, the implementation of data mining tasks was presented on the dataset in hand to generate classification models and testing them. Finally, interesting results were drawn from the classification models as well as interesting patterns in the MadaBoost was found. ID3 and k-NN algorithms have been implemented as well as with the MadaBoostalgorithm.

In this study, it was slightly found that the student's performance is not totally dependent on their academic efforts in spite, there are many other factors that have equal to greater influences as well. In conclusion, this study can motivate students as well as staff members to perform data mining tasks on their students' data regularly to find out interesting results. This study is also helpful for those students' who need special attention and will also reducing failure ratio by taking proper action for the higher education. The experiment can be extended for students who studied distance mode also to predict their academic future performances. The parameters like students' extracurricular activities and other vocational courses completed by the students also will be considered as performance predictable variables.

REFERENCES

[1] Brijesh Kumar Baradwaj, Saurabh Pal, "Mining Educational Data to analyze Students' Performance", *International Journal of Advanced Computer Science and Applications*, Vol. 2, No. 6, 2011.
[2] Cristobal Romero, Sebastian Ventura, Mykola Pechenizkiy, and Ryan S.J.d. Baker "Handbook of Educational Data Mining", CRC Press, 2011.
[3] Erman Yukselturk, Serhat Ozekes, Yalın Kılıç Türel, "Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program", *European Journal of Open, Distance and e-Learning*, ISSN: 1027-5207, 2014.
[4] Jyoti Bansode, "Mining Educational Data to Predict Student's Academic Performance", *International Journal on Recent and Innovation Trends in Computing and Communication*, ISSN: 2321-8169, Volume: 4 Issue: 1, Jan 2016.
[5] Suchita Borkar, K. Rajeswari, "Predicting Students Academic Performance Using Education Data Mining", *International Journal of Computer Science and Mobile Computing*, ISSN: 2320–088X, Vol. 2, Issue. 7, July 2013.
[6] Zahyah Alharbi, James Cornford, Liam Dolder, Beatriz De La Iglesia "Using Data Mining Techniques to Predict Students at Risk of Poor Performance", *SAI Computing Conference*, July 2016, London, UK.

AUTHOR'S 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.