UGC Approved Journal

IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 5, Issue 9, September 2017

A Comparative Study on Student Academic Performance Prediction using Classification Algorithm

R. Sasikala[1], G. Seenuvasan[2]

Research Scholar, P.G and Research Department of Computer Science, Swami Vivekananda Arts and Science College, Orathur, Villupuram, Tamilnadu, India[1]

Assistant Professor, P.G and Research Department of Computer Science, Swami Vivekananda Arts and Science College, Orathur, Villupuram, Tamilnadu, India[2]

Abstract: Data mining is the process of mining the exact data from a large amount of database In other words it can be said that data mining is mining the knowledge from data. A database is collection of inter related data. Normally, the performance of the student is based on the training provided by the teacher and the facilities provided by the institution. J48 classifies the simple c4.5 decision tree for classification. The binary tree is a decision tree approach is most useful in classification problem. The current system uses Naive Bayes, C4.5 and ID3 algorithm for analyzing the performance of the students is used for classifying the students into high and low category. This paper provides comparative study of predicting student level performance using classification algorithm such as Naive Bayes, C4.5 and ID3.

Keywords: Data Mining, Classification algorithm, Naive Bayes, C4.5, ID3 algorithm and Random Forest algorithm.

I.INTRODUCTION

Data Mining is a process to explore required data from large amounts of data - typically business or market related - also known as "big data". The user search for reliable patterns or efficient relationships between variables confirm the findings by applying the detected patterns to new subsets of data. There are many classification algorithms used for classifying the data. The classification algorithms can be classified into four categories. Basic learning/mining tasks deals with the mining of data from the database based on query of search and Inferring rudimentary rules provides for mechanism that generates rules by concentrating on specific class at time.

Decision tree learning model maps the observation about the item and provide conclusion about the target value and the Covering algorithm Remove positive examples covered by this rule. This paper used for classifying the student dataset into high and low performance students. This paper uses random forest algorithm for classifying the student based on their performance.

II.RELATED WORK

Many of the earlier researchers have analyzed performance of the students based on their tenth, twelth marks scored by the students, and on living location, medium of teaching, mother's qualification, students other habit, family annual income and student's family status, etc.

Bharti Thakur et. al [1], in Data Mining with Big Data Using C4.5 and Bayesian Classifier compares C4.5 and Bayesian Classifier algorithm using the performance of the students. Comparison of these two algorithms and classified the data set into different classes and in different phases. Navie Bayes algorithm provides best accuracy level 81% and compared C4.5 algorithm

Brijesh Kumar Baradwaj et. al [2], in Data Mining Educational Data to Analyze Students Performancel. They use ID3 algorithm for classifying the dataset. This paper used dataset obtained from VBS Purvanchal University, Jaunpur. These analyze the performance of the students based on their Previous Semester Marks, Class Test Grade, Seminar Performance, and Assignment and end semester marks. From this paper, the students with low performance can be easily identified and high concentration can be provided in order to improve the performance level of such students

Ajay Kumar Pal et. al [3] , Classification Model of Prediction for Placement of Studentl, apply various classification algorithm with Navie Bayes, MLB, and J48 for analysis the student's academic performance for Training and

DOI 10.17148/IJIREEICE.2017.5905

placement. This model determines the relations between academic achievement of students and their placement in campus selection.J48 algorithm gives the best accuracy level of 86.15% then other classification algorithms.

Mrinal Pandey et. al [4], used decision tree algorithm for prediction of students' academic performance in higher education. Data Mining: A Decision Tree Algorithm Pertaining to the Student Performance Analysis and Prediction‖. These learn will helpful for identify the students. They are used different decision tree algorithm J48, NB tree and simple card algorithm.J48 method identify two factors weak and ―success‖ students. It gives the best accuracy level is 80.15%

Brijesh Kumar Bhardwaj in Data Mining:A prediction for performance improvement using classification‖ [5] raw data was preprocessed in terms of filling up lost ethics, transform ethics in one shape into one added and relevant attribute variable selection. Bayesian classification method is used on student database to predict the students division on the basis of previous year database. This study helped the students and the teachers to improve the division of the student.

Surjeet Kumar Yadav et. al [6], used C4.5, ID3 and CART decision tree algorithms are applied on engineering student's data to predict their performance in the final exam in ―Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification‖. This paper generate the result of the decision tree predicted the number of students and the C4.5 algorithm has given highest level of accuracy 67.77% compared to other classification.

Kalpesh Adhatrao et. al [7], Predicting students' performance with Id3 and c4.5 using classification algorithms‖ analyze the data set containing information concerning student, such as sexual category, marks score in the panel examinations of lessons X and XII, symbols and position in entrance examination and results in first year of the previous batch of students. Compared to other classification algorithms ID3 (Iterative Dichotomies 3) produces better accuracy level is 67% and C4.5 categorization algorithms Decision tree algorithm and Naive Bayesian Classifier algorithm are applied on pre-processed student data to reveal classification accuracy between 93.33 % and 71.67 % in Dr. A. Padmapriya, Prediction of Higher Education Admissibility using Classification Algorithms [8]. The highest accuracy is achieved for the Decision Tree model (93.33%). The Decision Tree model predicts with higher accuracy the brawny class, while the other three models execute enhanced for the scrawny class. The data attribute connected to the students' personal data and under-graduation data are among the factors influencing most the classification process. A Study on Student Data Analysis Using Data Mining Techniques‖ by Umamaheswari. K, S. Niraimathi [9], explores the socio-demographic variables times, masculinity, name, lesser class rating, upper class blot, degree expertise and extra knowledge or skill, etc. This paper used classification algorithm to categorize the level of students. Clustering analysis separate tool to discover data sources distribution of information, as well as other data mining algorithm as a preprocessing rung, the collect psychoanalysis has been into the meadow of data mining is an important research topic. This groups the students according to their grade and proficiency.

T.Miranda Lakshmi, A. Martin, R.Mumtaj Begum and Dr.V.Prasanna Venkatesh [10] made An Analysis on Performance of Decision Tree Algorithms using Student's qualified data. This paper compares the ID3, C4.5 and CART algorithm. The performance based on Parent qualification, Living Location and Economic Status, Friend and Relative Support, Attendance Result. CART shows the best classification accuracy when compared other classification. It produce the highest accuracy level is 55.83%.

III.PROPOSED METHODOLOGY

Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. Tree sculpt where the intention variable can take a finite set of values are called classification trees.

In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target changeable can take continuous values are called regression trees. The proposed system uses random forest algorithm which is among decision tree to classify the student dataset into high and low category. The uploaded dataset of the student is classified based on the performance of the student.
An issue with Naive-Bayes is that it has no occurrences of a class label and a certain attribute value together then the frequency-based probability estimate will be nothing. Agreed Naive-Bayes' provisional independence supposition, when all the probability are multiply it will get zero and this will affect the posterior probability estimate.
The major nuisances of the C4.5 algorithm are as follows.

DOI 10.17148/IJIREEICE.2017.5905

Values generated using this algorithm neither contribute to generate rules nor help to construct any class for classification task. It crafts the tree bigger and more complex. Branches reduce the usability of decision trees.
Some of the issues ID3 algorithm is time complexity is high when compared to other algorithms. It cannot be provide with exact accuracy of classified students. This problem happens when samples are being drawn from a population and the drawn vectors are not fully representative of the inhabitants. For prediction a new sample is pushed down the tree. It is dole out the label of the preparation sample in the terminal node it ends up in. This process is iterated in excess of all foliage in the band, and the customary vote of all trees is report as random forest prediction.

Random forest Algorithm

Random Forest Prediction s= $\frac{1}{k}\sum_{k-1}^{k} k^{th}$

The prediction of the random forest is taken to be the average of the predictions of the tree. Where the index K run over the individual trees in the forest.
This algorithm run time is fast efficiently on large data base they are able to deal with unbalanced and missing data. It can handle thousands of input variable without variable deletion .This method effective for estimating missing data and maintains accuracy.

Data Selection

Random forest algorithms developed by Leo Breiman and Adele Cutler in 2001. Random forests are an ensemble learning method for classification, regression and other tasks, that function by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the person trees. Random forest correct for choice trees habit of over fitting to their training set.
It is one of the most accurate learning algorithms available for many data set it produce a highly accurate classifier. Random forest classifier used for number of decision trees in order to improve the classification rate this method combines the Breiman's in ―bagging‖ idea and the random selection features.
It is a group of classifier that consists of many decision trees and outputs the class that is mode of the class output by individual trees. Decision trees are individual learners that are combined for one of the popular learning methods commonly used for data exploration. Ensembles are a divide and conquer approach used to improve performance.
Each classifier independently is a weak learner‖ whiles all the classifiers taken together are a strong learner. Random forest techniques examines a large ensemble data set .It is first generating a random sample of the original data with replacement and a user defined number of variable selected at random from all the of the variables to determine node splitting. When a new input is entered into the system it is run down all the trees. The result may either be an average or weighted average of all the terminal nodes that are reached. Each of the trees is grown to the largest extent possible there is no pruning.
The variables, description and possible values of the variables are listed below in table1. The data sets collected from the College are used for comparison of Naïve Bayes algorithm and Random forest algorithm.

Table. I. Variables and Description

Variable	Description	Possible Values
AM	Assignment Mark	<5
SM	Seminar Mark	<10
IM	Internal Mark	<50
SEM	Semester Mark	<100
CFAC	College Facility Internet facility Library facility	Brilliant, Superior, Reasonable, Meager

If the low performance students are high the reason for low performance is taken from the Test and CFAC. This helps to improve the performance of the students in future.

IV. RESULT AND DISCUSSION

The accuracy level of accessing algorithms such as ID3, Naïve Bayes (NB) and C4.5 are compared. The ID3 algorithm provides 67% of accuracy in [7] when comparing the student's concert. The Naïve Bayes algorithm provides 81% of accuracy level in [1] and C4.5 provides 75.145%.

DOI 10.17148/IJIREEICE.2017.5905

The accuracy level for the accessing algorithm is provided in the.

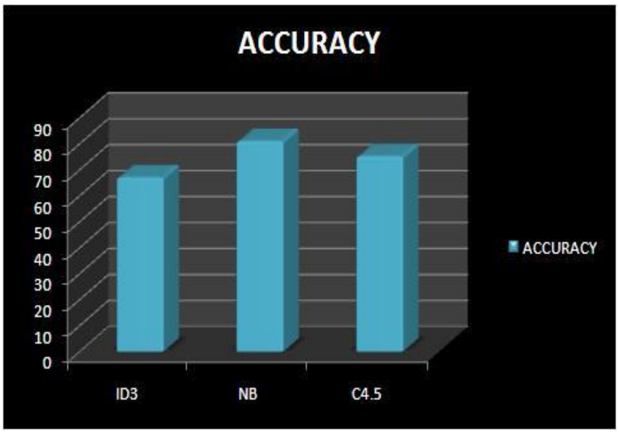

Figure.1. Performance accuracy

V. CONCLUSION

Predicting student academic performance has long been an important research topic. Among the issues of higher education institutions, questions concerning admissions remain important. The main objective of the admission system is to determine the candidates who would likely perform well after being accepted into the university. The quality of admitted students has a great influence on the level of research and training within the institution. The failure to perform an accurate admission decision may result in an unsuitable student being admitted to the program. Hence, admission officers want to know more about the academic potential of each student. Accurate predictions help admission officers to distinguish between suitable and unsuitable candidates for an academic program, and identify candidates who would likely do well in the university. The results obtained from the prediction of academic performance may be used for classifying students, which enables educational managers to offer them additional support, such as customized assistance and tutoring resources. The results of this prediction can also be used by instructors to specify the most suitable teaching actions for each group of students, and provide them with further assistance tailored to their needs. In addition, the prediction results may help students develop a good understanding of how well or how poorly they would perform, and then develop a suitable learning strategy. Accurate prediction of student achievement is one way to enhance the quality of education and provide better educational services. The accessing algorithms such as ID3 (Iterative Dichotomies 3), Naïve Bayes, C4.5 algorithm provided classification accuracy based on the provided dataset. This paper compares the Naïve Bayes, ID3, C4.5 algorithm and aims at displaying that random forest algorithm performs better than the other classification algorithm with more than 81% of accuracy.

REFERENCES

[1] Bharti Thakur, "Data Mining With Big Data Using C4.5 and Bayesian Classifier", International Journal of Advanced Research in Computer Science and Software Engineering , Volume 4, Issue 8 , August 2014.

[2] Brijesh Kumar Baradwaj and Saurabh Pal, "Mining Educational Data to Analyze Students Performance", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.

[3]]Ajay Kumar Pal, Saurabh Pal, I.J.Modern , "Classification Model of Prediction for Placement of Students Education and Computer Science", 2013, 11, 49-56, Published Online November 2013 in MECS.

[4] Mrinal Pandey and Vivek Kumar Sharma, "Data Mining: A Decision Tree Algorithm Pertaining to the Student Performance Analysis and Prediction‖", International Journal of Computer Applications (0975 – 8887) Volume 61– No.13, January 2013.

[5] Brijesh Kumar Bhardwaj, "Data Mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security, Vol. 9, No. 4, April 2011.

[6] Surjeet Kumar Yadav, "Data Mining: A Prediction for Performance improvement of Engineering Students using Classification‖", World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221- 0741 Vol. 2, No. 2, 51-56, 2012.

[7] Kalpesh Adhatrao, "Predicting students' performance using Id3 and c4.5 classification algorithms", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.5, September 2013.

[8] Dr. A. Padmapriya, "Prediction of Higher Education Admissibility using Classification Algorithms", International Journal of Advanced Research in Computer Science and Software Engineering.

[9] Umamaheswari and K, S. Niraimath , "A Study on Student Data Analysis Using Data Mining Techniques, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8, August 2013.

[10] Umamaheswari and K, S. Niraimath , "A Study on Student Data Analysis Using Data Mining Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8, August 2013.

DOI 10.17148/IJIREEICE.2017.5905