A STUDY AND ANALYSIS OF EDUCATION DATA USING DECISION TREE ALGORITHMS

[1]E. Venkatesan, [2]S. Anu H Nair, [3]K. P. Sanal Kumar

Guest Lecturer[1], Assistant Professor[2,3], [1,3]PG Department of Computer Science, RV Government Arts college, Chengalpattu, India, [2] Department of CSE, Annamalai University, Chidambaram, India[Deputed to WPT Chennai

venkatelumalai@yahoo.co.in[1], anu_jul@yahoo.co.in[2], sanalprabha@yahoo.co.in[3]

ABSTRACT

In today scenario education data mining is a best tool for improving education method and managing academic institutions. Many researchers studied about education data mining in technology and predict new methods, it is useful for academic institutions and higher education if it will implement in education system people get a quality education. **Statistical analysis**: A classification technique used to the huge amount of data analysis, and predicts future decision and applying in many fields. This proposed study has three stage process, the first stage is input the data to pre processing, second stage applying classification algorithms, namely J48, random forest (RF) tree and reduced-error pruning (REP) tree and final stage is to identify performance. **Findings:** To find out the classification algorithm's performance, this method used education data as input. Particularly, this work carries out to compare the three algorithms in decision tree algorithms to predict the performance accuracy in the usage of education data.

Keywords: Education data, Classification algorithms, RF Tree, REP Tree, J 48.

INTRODUCTION

Data processing is the most powerful way to compile and analyze useful information from a data warehouse. The task of data mining is predicting hidden information, make decision making and extraction processes. Classification is an over-the-top machine learning technique that assigns labels or classes but not the same objects or groups. Classification is a two-step process, the first step is model construction which is defined as analysis in the training records of a database. The second step is used to classify the model application of the structured model, then calculate classified accuracy is estimated by the percentage of correctly classified test samples or records. In this Classification fouse area in scientific experiments, clinical diagnosis, weather forecast, credits approval, client division, text mining, web mining, target marketing has been successfully applied to a wide variety of application areas such misdiagnosis and data mining techniqes, and expanded its application in educational activities based on student, staff and performance in administrative decisions [3,1].The main function in education dataset predicted purpose to update UGC regulations and issues, such as the university grant commission provides some guidelines, which quality education and teaching faculty qualification and student enrollment ratios should be increased the main objective, next steps motivate innovative research this method implements one of the best source for the education data mining, then education data only considered not possible higher education rule processing, other data also useful to for such data is below the poverty line and economic weaker section data also helpful, recently the data mining techniques very support to enhance and evaluation in process higher education task. Some researcher education related research contribution in new methods and proposed research have useful for higher educations and academic institution for effective manage. The main point of this research work is to identify the best classification algorithms to explore the performance of undergraduate student in academic data set. To findout the best classification algorithms can be found by comparing the performance in different traits in predictiong student's performance in

the final semester exam using different types of classification algorithms such as random forest Tree (RF Tree) reduced-error pruning (REP) tree and j48 [2].This research work is structured as follows, section 2 discusses Materials and Methods, trial evaluation and comparative analysis is given in section 3 and conclusion of this proposed work is given on section 4, Finally, key references are mentioned in section 5.

MATERIALS AND METHODS

The research work main focus by five steps first one is education data input to pre processing then second steps also use to decision tree algorithms namely such as j48, random fores tree (RF) Tree and reduced-error pruning (REP)Tree thus algorithms input education dataset, next step third also manipulated was produced results could be identifying future decision for students performance and academic institution to management, *fourth* step is error measures for both algorithms *result, such* as recall, f-measure, precision, *sensitivity, specificity* and accuracy, thus error measure analysis fifth sept is comparative study base error measure is which one highest accuracy produced *an algorithm* for education dataset suitable to *the decision tree method*.

RESEARCH DATA SET

This study used education data gathered from in Tamilnadu colleges. In this data below describe about attributes like age, registration number, sex, semester marks, grade, subjects, attendance percentage, assignment marks, class test marks and internal marks. This study, carried out totally more than 100 samples used both male and female undergraduate students. This data created in CVs excel format, thus formatted input for classification algorithms, then education data used to find out the best performance of algorithms.

Table1. Describe of the Education Data Set.

Sr. no	Variables	Details
1	Id number	Identifications
2	Age	16 above
3	Sex	Both male and female
4	Semester marks and percentage	Range 40 to100 and fail marks start 01to 39
5	Grade or percentage	Gold medal, distinction, A++,A,B or first, second and third class
8	Subject percentage	Language, major and allied
9	Attendance percentage	The student attens the class
10	Assignment marks	Maximum 10 marks
11	Class test marks	Maximum 75 marks
12	Internal marks	Maximum 25 marks

Table 1 shows the education data attributes, the main purpose of this research work to find out the students' performance, and also classification algorithm efficiency.

CLASSIFICATION ALGORITHMS

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 one can get new ideas can be discovered like association rules, classification, clustering, and pruning the data. Some of the Classification algorithms mentioned for the proposed work have provided excellent results in educational resources [6, 7]. Data mining is a technique to identify, explore and model large amounts of data detection of unidentified patterns or relationships that give a correct result. [4].

J48 ALGORITHMS

J48 examines the resulting standardized data development based on data capture by selecting a elements. Each phase of the information is intensified into smaller subdivisions that appear at one end and to craft the conclusion, the elements extreme regular data growth is utilized. Intense techniques bring to a halt if there is a subgroup related to the same type in all the cases. J48 creates a result node using projected values of the class. J48 can select specific attributes and lost attribute values as well as element values.

Step 1: If the events apply to an equal set then first the leaf node is considered with the same set

Step 2: All attributes and possible sequence will be considered and information development will be selected from the verified attribute.

Step 3: The best element identifier obtained from the current identification control will be found [5].

RANDOM FOREST TREE (RFT)

The reduced-error pruning (REP) tree as a decision tree-learning algorithm can be considered a fast classifier based on the principles of computing information gain with entropy and minimizing the error arising from variance.REP Tree produce multiple trees and uses regression logic to convert iterations. Then the algorithm selects the best one from all the trees. The algorithm creates a regression decision tree based on variance and result information. Also, this method enhances the pruning tree using the back matching method and reducing error pruning. Creating multiple decision trees to improve Random Forest Classification ratio and can overcome compatibility issues. Random Forest uses the decision tree to view a data mining technique classification. K Random Forrest to create countless tress each time chose different part of the data set. Random Forest uses the decision tree to view a data mining technique classification. Test data used for all trees built in Random Forest and frequently output will be assigned as a label for tested data. Random Forest has the idea of the real forest and it has more trees will be much stronger. Also, Random forest will give better accuracy if it has a large number of trees.

Step1: First, start with the selection of random samples from a given dataset.

Step2: Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.

Step3: In this step, voting will be performed for every predicted result.

Step 4: At last, select the most voted prediction result as the final prediction result [8].

REDUCED-ERROR PRUNING (REP TREE)

The reduced-error pruning (REP) tree as a decision tree-learning algorithm can be considered a fast classifier based on the principles of computing information gain with entropy and minimizing the error arising from variance. reduced-error pruning (REP)Tree produce multiple trees and uses regression logic to convert

iterations. Then the algorithm selects the best one from all the trees. The algorithm creates a regression decision tree based on variance and result information. Also, this method enhances the pruning tree using the back matching method and reducing error pruning [9].

EXPERIMENTAL RESULTS AND DISCUSSION

Classification algorithms are used in this research. The main object of this study is to identify slow learners and expose students' knowledge. And then the classification algorithms analyzes various statistics to analyze the performance and predicts its effectiveness. Details of statistics are , error measure for recall, precision, f-measure, sensitivity, specificity and accuracy and running time, which of these classification algorithm gives the highest performance, it has potential most suitable for classifying students data, these results discuss below.

RESULT OF J48 ALGORITHM

The table2 show result of j48 algorithm precision value is 0.931, recall 0.931 and f-measure value 0.930, next sensitivity value is 93.0693 and specificity value 6.9307. In the statistics values compare to precision and recall both same, but f-measure values also reduce 1 value only. The result of j48 algorithm sensitivity value is 93.0693 and specificity 6.9307, both values is comparatively very highest values is sensitivity.

Table2. Describe of the J48 Algorithm Result.

Precision	Recall	F-Measure	Sensitivity	Specificity
0.931	0.931	0.930	93.0693 %	6.9307 %

The result of Reduced-error pruning, tree algorithm

The table3 show theresult of reduced –error pruning, tree algorithm statistics values compared is precision value 0.871, recall 0.861 and f-measure 0.862, thus three statistics values compare to precision value is very highest other than values are low. Next statistics values derived is sensitivity 86.1386 and specificity 13.8614, both compare to sensitivity is a very high value present.

Table 3.Description of the reduced-error pruning tree algorithm result.

Precision	Recall	F-Measure	Sensitivity	Specificity
0.871	0.861	0.862	86.1386 %	13.8614 %

The result of random forest tree algorithm

The algorithm described in table 4 shows the result of random forest tree algorithm statistics measure precision value is 0.944, recall 0. 941and f-measure 0.941, then the sensitivity value is 94.0594 and specificity 5.9406. In this result compares the algorithm carry out both statistics measures very highest values specifying to sensitivity.

Table 4. Described of the random forest tree algorithm result.

Precision	Recall	F-Measure	Sensitivity	Specificity
0.944	0.941	0.941	94.0594 %	5.9406 %

COMPARATIVES RESULTS

This study manipulate three algorithms namely j48, RFTree and REP Tree, compared the results of precision, recall, f-measure, sensitivity, specificity and accuracy, analysis and predicted algorithms efficiency and find out

students performance are defined in figure 1 and table 1 to 5 shows the classification algorithms results, random forest tree sensitivity 94.0594 %, specificity 5.9406 % and j48 sensitivity 93.0693 %, specificity 6.9307 %, reduced-error pruning tree method sensitivity 86.1386 % and specificity 5.9406 %, and finally analysed random forest algorithm sensitivity and specificity value is highest compare than other two algorithm, so it is perfectly suitable for classifying education data.

Table 5.Described of the comparative results

Algorithms	Accuracy
RFTree	94%
REPTree	83%
J48	93%

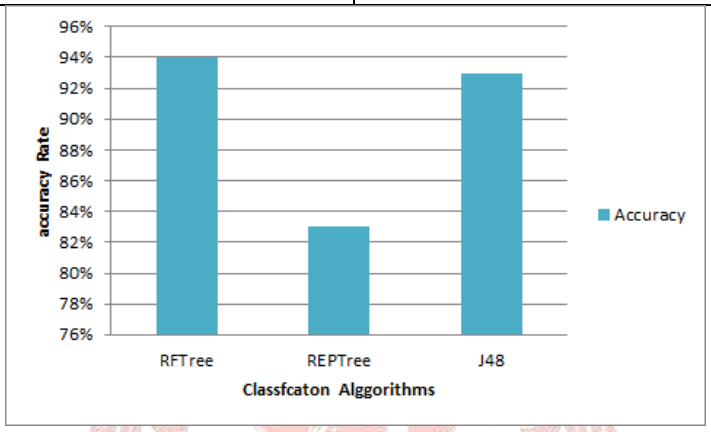

Figure 1:Described of the comparative results

CONCLUSION

An important challenge in data mining has precisely identified efficient classifiers of education mining applications. In this study, the accuracy of classification techniques is evaluated based on the selected characteristics of education data. The classifiers, J48 accuracy 93%, REP Tree 83% and highest point value 94% found in RF Tree. So from the results obtained, it can be seen that the RF Tree algorithm is the best way to classify the education data. Similarly, in the future, other classification algorithms can be used to predict the performance of academic data.

REFERENCES

[1] Arunachalam A.S and Velmurugan.T.,"Analyzing student performance using evolutionary artificial neural network algorithm", International Journal of Engineering & Technology, Vol. 7(2.26), pp. 67-73, 2018.

[2] Velmurugan.Tand Anuradha.C.," Performance Evaluation of Feature Selection Algorithms in Educational Data Mining", International Journal of Data Mining Techniques and Applications,Vol 05(02). pp.131-139,2016.

[3] Latha.U and Velmurugan.T.,"Analyzing Agricultural Text Data using Classification Algorithms",The 3rd International Conference on Small & Medium Business, January 19 - 21, 2016, Nikko Saigon Hotel, Hochiminh, Vietnam,2016.

[4] Venkatesan.E and Velmurugan.T., "Performance Analysis of Decision Tree Algorithms for Breast Cancer Classiication",Indian Journal of Science and Technology, Vol 8(29),pp.1-8,2015.

[5] Saravananathan.k and Velmurugan.T.,"Analyzing Diabetic Data using Classification Algorithms in Data Mining", Indian Journal of Science and Technology, Vol 9(43),pp.1-6,2016.

[6] Padmapriya.B and Velmurugan.T.,"Classification Algorithm Based Analysis of Breast Cancer Data",International Journal of Data Mining Techniques and Applications, Vol 5(01),pp.43-49,2016.

[7] Anuradha.C and Velmurugan.T.," A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance",Indian Journal of Science and Technology, Vol 8(15),pp.1-12,2015.

[8] Mohammed Hikmat Sadiq and Nawzat Sadiq Ahmed.," Classifying and Predicting Students' Performance using Improved Decision Tree C4.5 in Higher Education Institutes ", Journal of Computer Science,Vol 15(9),pp.1291-1306, 2019.

[9] Alaa Khalaf Hamoud, Ali Salah Hashim and Wid Aqeel Awadh.," Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis", International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 5(2),pp.26-31,218.