IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

A Comparison Between Classification Algorithms on Different Datasets Methodologies using Rapidminer

Ashmeet Singh[1], R Sathyaraj [2]

Student, School of Computing Science and Engineering, VIT, Vellore, India [1]

Assistant professor (Senior),School of Computing Science and Engineering, VIT, Vellore, India [2]

Abstract: Data Mining techniques are helpful in finding out patterns between data attributes and results in probalistic prediction of the label attribute.The paper discusses different classification techniques on small and large datasets.The two datasets are example datasets used from repository sites depending upon the number of instances.These datasets were applied in different classifier like Random Forest, Naive Bayes and Decision Tree to identify the best classifier for small dataset and large dataset. This paper gives the study and analysis of various methodologies used for prediction Based on the study, Naïve Bayes is most suitable for small datasets and Decision Tree is suitable for large datasets based on the evaluation done in this paper using various methodologies driven by RapidMiner tool while equating precision, recall and accuracy.

Keywords: Naive Bayes, Random Forest, Decision Tree,RapidMiner tool

I. INTRODUCTION

The main objective of paper is to study the impact of different classification algorithms in the prediction of unknown label attributes. The parameters for judging the algorithms are accuracy, recall and precision. These are helpful when training data is used instead of testing data, i.e. finding out the value of known values and comparing them to know the accuracy, recall and precision of the particular algorithm. This paper is catalogued as follows. Section II lists a related work. Section III presents the methodology and discusses the aspects of classification algorithm and respective datasets. Section IV elaborates Experiment and finalizes the results produced by the algorithms. Section V provides the conclusion.

II. RELATED WORK

Mrs. M.S. Mythili and Dr. A. R. Mohamed Shanavas used data mining methodologies, such as decision tables, IB1, J48, Multilayer Perceptron and Random Forest, to study and analyze the performance of the school students. The conclusion came out was that the Random Forest is the best classifier for analyzing the school students' performance result. It consumes less time and has good accuracy in [3]. The classification results of Jehad Ali show that the Random Forest gives better results for large datasets keeping the same number of attributes while J48 is a best and easy approach for small datasets i.e. less number of instances in [4]. Amit Gupte and his team too found Random Forest at top of all the other algorithms on their dataset of sentiment analysis. Sentiment analysis is a task which involves the extraction of information from customers' feedbacks and other authentic sources such as survey agencies. Considering sentiment analysis the Random Forest classifier again has high accuracy and performance, simplicity in understanding, and improvement in results over a period of time. This results the classifier to best fit for situations like sentiment analysis in [5].

III. METHODOLOGY

The following steps are included in the classification process in this paper,

- Two types of datasets are chosen based on the number of instances.
- Three different classifiers are chosen- Naive Bayes, Random Forest and Decision Tree.
- RapidMiner tool is used to analyse the predicted values by each of the classifier.
- The precision, recall and accuracy of each classifier is calculated.
- Finally the result is analyzed and the best suited algorithm for a particular type of dataset is found.

A. Dataset Used

Datasets discussed in the paper are purely judged on the number of instances. There are basically two datasets used to judge the potential of different algorithms. The number of instances of two datasets are 498 and 30161.

1) Dataset of 30161 Instances (Large):

This dataset aimed at the case of customers â€™ default payments in Taiwan. The dataset enrolls a binary variable, default payment ('<=50' or ' >50' i.e. boolean), as the response variable. This dataset used the 14 variables as regular attributes and one as a label attribute. The dataset has 32561 instances in total among which 2400 instances have missing values in [9].

2) Dataset of 498 Instances (Small):
The CM1/Sotware defect prediction creator was a NASA spacecraft instrument. It was a NASA metrics data program which was written in C language. These metrics were segment based or it may call as a function or method. CM1 has 498 numbers of instances which is the priority focus of using this dataset. Unlike the large dataset, none of the instances in this dataset had a missing value in [8].

B. Classifiers
1) Naive Bayes:
Naïve Bayes classifier is the simplest instance of a probabilistic classifier. It gives the probability PR (C|d) as the output of a probabilistic classifier where a document d belongs to a class C¬. 1 It is used as a supervised learning method as well as a statistical method for classification. The Bayes Theorem: P (h/D) = P (D/h) P (h) P (D); P (h): Prior probability of hypothesis h; P (D): Prior probability of training data D; P (h/D): Probability of h given D; P (D/h): Probability of D given h in [6].

2) Random Forest:
Random forest is a collection of decision trees. It is presented independently with some controlled modification. Trees and the results included in Random Forest are based on majority of accurate output. Random forest is the best classifier for large datasets. 1) If 'n' is the number of cases in the training set, then 'n' cases are to be sampled randomly but with replacement, from the original data. This sample will act as a training set for growing the tree. 2) If input variables are 'M' in number, a number mM is specified such that at each node, m variables randomly selected out of the 'M' input variables and among all these 'm',the best split is used to split the node. The value of m is kept constant during the forest growing. 3) Each tree is made to grow to the largest extent possible. Pruning is restricted just to get more accuracy compromising increased execution time in [7].

3) Decision Tree:
A decision tree is a classifier which classifies an input sample into one of its possible classes.It is a tree structured classifier which makes decision rules from the large amount data to extract knowledge. A decision tree classifier uses a simple form which is concisely stored and that efficiently classifies new data.

The advantages of decision tree in data mining 1) Its ability to handle different input data types such as, numerical, textual and nominal. 2) It can even take care of datasets whose instances have missing values and errors. 3) It is available in various packages of data mining and a number of platforms as in [2].

C. Factors Considered for Calculating Performance of Classifiers
The accuracy of the classifiers are given by true positive rate, false positive rate, precision, recall and F-measures using RapidMiner tool. RapidMiner is a powerful software platform that gives an integrated environment for machine learning, data mining, text mining and other business and prediction analysis. The average of measures from all the classes has been taken to give the overall measure for classifiers.For example, to give the overall precision for a classifier for a given dataset, average of precisions of both (true/false) classes is calculated.

1) Accuracy:
Accuracy is calculated as number of instances predicted positively divided by Total number of instances.This means accuracy is the percentage of the accurately predicted classes among the total classes. In the experiment the values of the accuracy posted into table in the basis of 0 to 100, not from 0 to 1.
Accuracy = ((True Positive + True Negative) / (P + N))*100

2) Precision:
Precision is the preciseness or exactness of truly classified class, therefore known as positive predictive value.It is the proportion of instances which truly have class x / Total classified as class x. So basically high precision stated the accurate results and it takes all relevant data but returns only topmost results. In short, it is the number of chosen items which were related.

Precision =(True Positive / (True Positive + False Positive))*100

3) Recall:
Recall gives sensitivity of problem and it process values or product quantity or completeness. It returned most relevant and part of the documents that are relevant as result from the query. In other words, modules that are really recognize as difficult to maintain from the total number of modules. In short, it is the number of related objects that were chosen.
Recall = (True Positive / (True Positive + False Negative))*100

4) True Positive (TP):
True positive are the positive tuples which were correctly labelled by the classifier.It is the proportion categorized as class x / Actual total in class x. True positive projected by the modules that are predicted positively as the results specified at the end.
True Positive rate = (True Positive / (True Positive + False Negative))*100

5) False Positive (FP):
False positive, proportion incorrectly categorized as class x / Actual total of all classes, except x. It is incorrectly predicted compared to original results.
False Positive rate = (False Positive / (False Positive + True Negative))*100

6) F-Measure:
F-Measure categorized as (2*Precision*Recall / (Precision+ Recall))*100. It is a combined measure for precision and recall.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

IV. EXPERIMENTS AND RESULTS

A. Experiment

In the analysis of datasets, one attribute was taken as label attribute which was used for classification of instances. Using Rapidminer tool, CM1(of 498 instances) and E-commerce(of 30161 instances) datasets were applied to classifiers Naive Bayes, Random Forest and Decision Tree algorithms. The rapidminer has been used to classify the testing data which was done using 10-fold validation. In the case of random forest and decision tree, pruning and pre-pruning is applied for better results by compromising the execution time

B. Results

The Results of following analysis on two datasets are clearly given by the tables I,II,III and IV. Tables I and III have given the instances correctly classified and in-accurately classified with total number instances in dataset using different classifiers. Tables II and IV listed the accuracy, true positive rate, false positive rate, precision, recall and F-measures to analyse the classifiers. Also it provides best classifier by highlighted based on precision value.

TABLE I CLASSIFIED INSTANCES OF SMALLER DATASET (498 INSTANCES)

Method	Appropriately Classified Instances	Appropriately Not Classified Instances	Total Instances
Naive Bayes	420	78	498
Random Forest	445	53	498
Decision Tree	418	80	498

TABLE II ANALYSIS ON SMALLER DATASET (498 INSTANCES)

	Accu-racy	Precision	Recall	TP Rate	FP Rate	F-Measu-re
Naive Bayes	82.74	78.26	72.19	72.19	27.81	75.10
Random Forest	77.11	87.64	58.22	58.22	41.79	70.0
Decision Tree	81.42	81.21	75.55	75.55	24.46	78.27

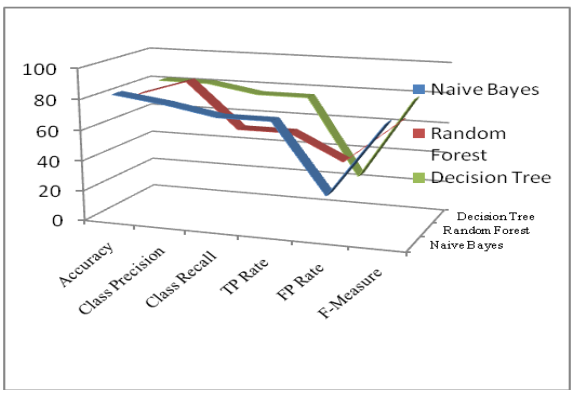

Fig. 1 Describes the ratio of each classifier for each dataset based on Table 1.

TABLE III
CLASSIFIED INSTANCES OF SMALLER DATASET (30161INSTANCES)

Method	Appropriately Classified Instances	Appropriately Not Classified Instances	Total Instances
Naive Bayes	24951	5210	30161
Random Forest	24069	6092	30161
Decision Tree	25558	4603	30161

TABLE IV
ANALYSIS ON SMALLER DATASET (30000 INSTANCES)

	Accu-racy	Preci-sion	Recall	TP Rate	FP Rate	F-measu-re
Naive Bayes	84.34	58.84	60.41	60.41	39.60	59.61
Random Forest	89.96	61.82	50.80	50.80	47.50	55.77
Decision Tree	89.97	65.24	51.71	51.71	43.45	57.69

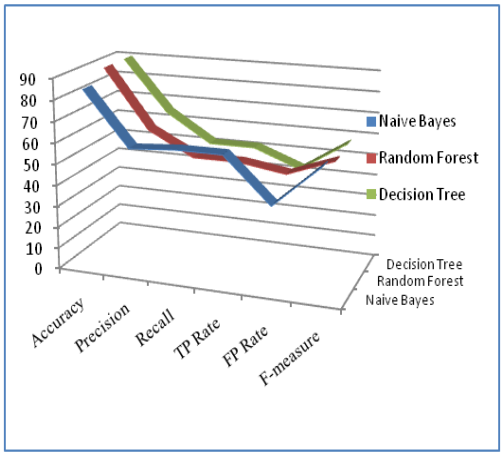

Fig. 2 Describes the ratio of each classifier for each dataset based on Table 2.

V. CONCLUSION

In this study, Naive Bayes results better results than other two in smaller dataset whereas Decision Tree is best for larger dataset.Therefore , Random forest acts as an average in both the cases. This happened because random forest takes large set of data to learn but the fails in these datasets as they have one thing in common. i.e. much lesser amount of data for true instances.

Therefore,as the number of instances having 'True' value were less in number, it was easier for Naive Bayes classifier to learn and respond better than others

ACKNOWLEDGEMENT

Every endeavour requires the effort many people at many levels. This Paper is no different. I express my Gratitude

to My Professor Sathyaraj R. His guidance was instrumental in the realisation of this Paper. I would also like to thank my peers Ishan Arora for helping my oraganise and structure the paper. I would also like to thank people who anonymously have contributed to my knowledge and motivated me towards the completion of this project.

REFERENCES

[1] S.L. Ting, W.H. IP, Albert H.C. Tsang "Is Naïve Bayes a Good Classifier for Document Classification?", International Journal of Software Engineering and Its Applications, Vol. 5, No. 3, July, 2011.
[2] Shahrukh Teli, Prashasti Kanikar " A Survey on Decision Tree Based Approaches in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, 2015.
[3] Mrs. M.S. Mythili, Dr. A. R. Mohamed Shanavas, "An Analysis of students' performance using classification algorithms", IOSR Journal of Computer Engineering, 2278-8727Volume 16, Issue 1, Ver. III (Jan. 2014).
[4] Jehad Ali, Rehanullah Khan, Nasir Ahmad, Imran Maqsood," Random Forests and Decision Trees", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012.
[5] Amit Gupte, Sourabh Joshi, Pratik Gadgul, Akshay Kadam," Comparative Study of Classification Algorithms used in Sentiment Analysis", International Journal of Computer Science and Information Technologies, Vol. 5 (5), 2014.
[6] Khosrow-Pour, Mehdi,"Encyclopedia of Information Science and Technology", First Edition, January 31, 2005 IGI Global.
[7] Sub Kumar, Dr. Manish mann,"E-Mail Filtering For The Removal Of Misclassification Error", International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 2, Issue 12, December 2015.
[8] CM1 dataset , Promise Software Engineering Repository, "http://promise.site.uottawa.ca/SERepository/datasets/cm1.arff", December 2, 2004.
[9] Credit Screening, UCI Machine Learning Repository, " http://archive.ics.uci.edu/ml/machine-learning-databases/credit-screening/ ".

BIOGRAPHIES

Mr. Ashmeet Singh is pursuing BTech, in Computer Science.He is studying in the School of Computer Science and Engineering, Vellore Institute of Technology University,Vellore. His fields of research interest are data mining, networks, text mining and embedded systems.

Mr. Sathyaraj R. has completed M.E Computer Science and Engineering. He is working as Assistant professor (Senior) in the School of Computer Science and Engineering, VIT University,Vellore. His fields of research interest software testing, data mining and soft computing. He has published papers in the international journals and presented research papers in international and national conferences.