JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

PERFORMANCE EVALUATION OF ENSEMBLE LEARNING ALGORITHMS FOR VARIOUS CLASSIFIERS

Ajay Kumar[1], Preeti Sondhi[2]
[1]M.tech Student,[2]Assistant Professor
Universal Group of Institutions Lallru Punjab,
Punjab Technical University

Abstract: ***In the present scenario with the advancement of digital technology extensive amount of data and information are generated. With such extensive amount of data, powerful approaches are required for data interpretation that help the human being for decision making process. Handling, analyzing, processing and finding relevant information from this extensive amount of data is very interesting and fast growing research area. Data mining is the process of extracting or discovering useful information from the large amount of data and then transforms them in to an understandable form for future use. Data mining offers various methods or techniques that are used to predict the accuracy of various classes of object. This research focus on various ensemble learningh techniques like bagging and boosting also enhance the accuracy of various base classifiers like NaiveBayes, DecisionStump, DecisionTable and J48. All the techniques are compared on the basis of four evaluation parameters like accuracy, precision, recall and root mean squared error. The finding are also supported with justification by conducting an experimental survey at the end of research. RapidMiner tool will be used to perform the simulation using 10 fold cross validation.***

KEYWORD: Data Mining, Ensemble Learning Techniques, Bagging, Boosting, Weka Tool, Rapid Miner Tool

1.INTRODUCTION

With the advancement of computer technology, huge amount of data and information is being accumulated day by day. Data and information is available either in online and offline mode. Extraction of relevant information from data is a challenging task. It is not possible to make effective decision by using manual analysis of such huge amount of data. So we require data mining, Data mining has made it possible to analyze data efficiently from different aspect (1). It is an interdisciplinary sub field of computer science. Data mining is the process of mining relevant or meaningful information from the extensive amount of data or we can also say that it is the process of discovering hidden patterns and information from the existing data using various powerful techniques and then transforms them in to an understandable structure for future uses. The main role of data mining technique is in decision support system. These techniques are used to predict unknown class label using known class label (2). Data mining is the analysis step in the process of KDD (Knowledge Discovery in Databases). The Knowledge Discovery in Databases(KDD) field is concerned with the development of methods and techniques for making sense of data or we can say that this process consist of various steps leading from raw data collection to some form of new meaningful knowledge(3). Various free and open source data mining tools are available in the market such as Weka, RapidMiner, ImageJ, ITK, KNIME, Orange, Pentaho, Keplers(4)

1.1 Classification algorithms

Classification is the most popular data mining technique. It is the act of looking for a model in such way so that the model can be used to predict unknown class label. In general classification process consist of two steps**.** In the first step the model is prepared by using classification algorithm on training data set. This step is also called training phase or training step where classification algorithms are used to prepare the model from a training data set which is made up of database tuples and their associated class labels. In the second step or phase the previous prepared model is tested against a predefined test dataset to measure the accuracy and performance of trained model. The rule can be applied to the new data tuples if the accuracy is acceptable. So we can called classification is a process to assign the class label from dataset whose class label is unknown [5]. Some of the common classification algorithms are:

1.1.1 NaiveBayes Algorithm

NaiveBayes algorithm is based on Bayesian theorem. It is a type of supervised learning method. This algorithm is practically used when dimensionality of the input is high. It is a simple probabilistic classifier that is capable of calculating the most possible output based on input [5].

1.1.2 RandomTree Algorithm

This algorithm consists of number of simple trees, which are used to determine the final outcome. It is a type of supervised learning method.

1.1.3 Decision Stump Algorithm

It is a type of decision tree algorithm, in which root node connected to leaf nodes. This algorithm usually used in Conjunction with boosting algorithm. In DecisionStump, a single input feature is sufficient to make prediction for new data. This algorithm can perform regression based on mean squared error and classification based on entropy [6].

1.2 Ensemble Learning Technique

Ensemble learning technique is one of the best technique of data mining. Ensemble learning technique uses several learning algorithm together for the same task with the aim to have better prediction than the individual learning algorithm. It is also called committee based learning or learning multiple classifier system. This method try to construct a set of learners and then combined them or train multiple learner to solve the same problem.

Common Ensemble Learning Methods are:

- Boosting
- Bagging

1.2.1 Boosting

Boosting is one of the robust ensemble algorithm which is capable for reducing both bias and variance. This technique facilitate the conversion of weak learner to strong learner. Boosting creates strong classification tree because it forces new classifier to focus on the error produced by the previous ones.

1.2.2 Bagging

Bagging Technique is also known as bootstrap aggregation. It is an ensemble meta-algorithm that are designed to improve the stability and accuracy of learning algorithm used in statistical classification and regression. This algorithm also reduce variance and helps to avoid overfitting [7].

1.3 Dataset

The datasets used for analysis purpose are downloaded from UCI repository and data.world website. The data set is PIMA Indian Diabetes.

1.3.1 Pima Indian Diabetes

In this datasets the prediction task is to determine whether a person is diabetes or not. This data set contains 9 attributes that are number of time pregnant, plasma glucose concentration a 2 hours in an oral glucose tolerance test, diastolic blood pressure (mm Hg),triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml), body mass index (weight in kg/(height in m)^2), diabetes pedigree function, age (years) and Result. The 9th attribute Result is a class label used to divide the data into group (diabetes or not diabetes).

Table 1.3 Data Set

Dataset Name	Abbreviation Used	Data Types	Default Task	Instances	Attributes	Size(KB)
Pima Indian Diabetes	DS	Multivariate	Classification	768	9	30.7 KB

Table 1.3 Data set Attribute Information

Sr. No.	Attributes	Attribute Information
1	number of time pregnant	How many time a lady pregnant
2	plasma glucose concentration a 2 hours in an oral glucose tolerance test	Glucose tolerance test

3	diastolic blood pressure (mm Hg)	Measure the pressure of in blood vessel. Its normal value is 120/80 mmHg
4	triceps skin fold thickness (mm)	Used to estimates the body fat
5	2-Hour serum insulin (mu U/ml)	Insulin test
6	body mass index (weight in kg/(height in m)^2)	It is a measure of body fat based on your weight in relation to your height.
7	diabetes pedigree function	It provides diabetes pedigree function value
8	age (years)	Age of the person
9	Result	Result is a class label used to divide the data into group (diabetes or not diabetes)

1.4 Parameters Choosen for Evaluation

1.4.1 Accuracy: is defined as a relative number of correctly classified instances or in other words percentage of correctly classified instances.

1.4.2 Precision: It is defined as relative number of correctly as positive classified example among all examples classified as positive. It is also called positive predictive value.

1.4.3 Recall: Recall specifies the relative number of correctly as positive classified example among all positive examples. It is also called true positive rate.

1.4.4 Mean absolute error: It measures the average magnitude of error in the set of forecasts, without considering their direction. It is linear score which means that all the individual difference are weighted equally in the average.

II Literature Review

E. Suriyapriya, M. Praveena (2017) discussed a novel approach for developing a cluster and booster on the basis of data mining. Clustering with boosting improve the quality of mining process. Boosting is the iterative process whose main goal is to improve the predictive accuracy of the learning algorithms. In this research paper various boosting problem and their proposed solutions are discussed and also stated some clustering technique. In order to performance enhancement, integrate the boosting methodology with fuzzy c means (FCM) [8]. **Prajakta S. Kasbe et al. (2017)** did a survey on road accident analysis methods in data mining. In this research Self Organization Map (SOM) was used to find a number of pattern to analysis the road accident data which help to find the prediction of accident reason and improve the accuracy of analysis [9]. **Anand Kishor Pandey, Dharmveer Singh Rajpoot (2016)** did a comparative study of various classification algorithms with the help of data mining tool named WEKA on dataset of alcohol consumption by school students. The experimental results showed that among all the algorithms experimented, Decision Stump (95.44%) algorithm performs the better classification [10**]. Sumouli Choudhury, Anirban Bhowal (2015)** did a survey and comparative analysis of machine learning algorithm along with classifiers for network intrusion detection. In this paper multiple classification technique and machine learning algorithm have considered to categorize the network traffic. Performance of various classifiers compared by using WEKA tool and concluded that RandomForest and BayesNet are suitable for this purpose. The machine learning algorithm have also compared and it can be deducted that Boosting is the best algorithm [11**]. Pooja Shrivastava, Manoj Shukla (2013)** did a comparative analysis of bagging, stacking and random subspace algorithms using Weka tool. On the basis of experiment, it is found that using 20 fold cross validation the performance of the stacking with decision stump and decision table improve the prediction accuracy of the classifier. So stacking is better and straight forward to interpret other. Stacking algorithm built accurate classifier model and consume less time [12].

III. RESULT AND DISCUSSION

Based on the parameters precision,accuracy,recall, and root mean squared error. Used RapidMiner and Weka tools for data analysis . Algorithm used for the data evaluation are NaiveBayes, DecisionStump,RandomTree with Ensemble learning algorithm such as Bagging and Boosting

3.1 Accuracy

RapidMiner and Weka tool are used for the analysis purpose. The algorithms used for the evaluation are NaiveBayes, DecisionStump,RandomTree and Ensemble learning algorithm such as Bagging andBoosting. The results are shown in the graph o observe the performance of algorithms in datasets.

3.1.1 Using NaiveBayes as Base Classifier on Weka

We have took data set value and perfomed with NaïveBayes as Base Classifier with ensemble learning algorithm such as Bagging and Boosting Weka tool gives more accuracy when we used Bagging Ensemble learning algorithm as compare to the RapidMiner tool andEnsemble learning algorithms does not enhance the performance of NaiveBayes algorithm in RapidMiner tool . weka tool give more precision as compared to rapid miner tool.Weka tool perform with all parameter better then rapid miner tool.

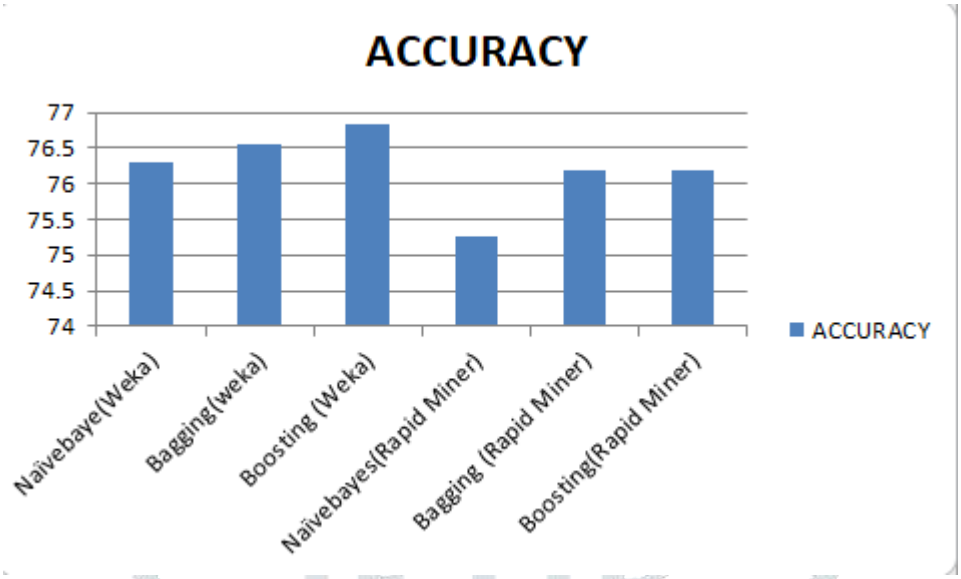

Fig. 1 Accuracy using NaïveBayes as Base Classifier

3.1.2 Using DecisionStump as Base Classifier

Using DecisionStump we have took dataset, Weka tool gives more accuracy as compare to the RapidMiner tool and both Ensemble learning algorithm enhance the accuracy of DecisionStump in both the tools. InRapidMiner tool both the Ensemble learning algorithm gives same accuracy but in Weka tool Boosting algorithm gives more accuracy as compare to the Bagging algorithm. Weka tool performs well as compare to the Rapid miner tool and Ensemble learning algorithm enhance the accuracy of DecisionStump in Weka but in case of Rapid miner tool Ensemble learning algorithm does not enhance the much accuracy of DecisionStump as compared to weka.

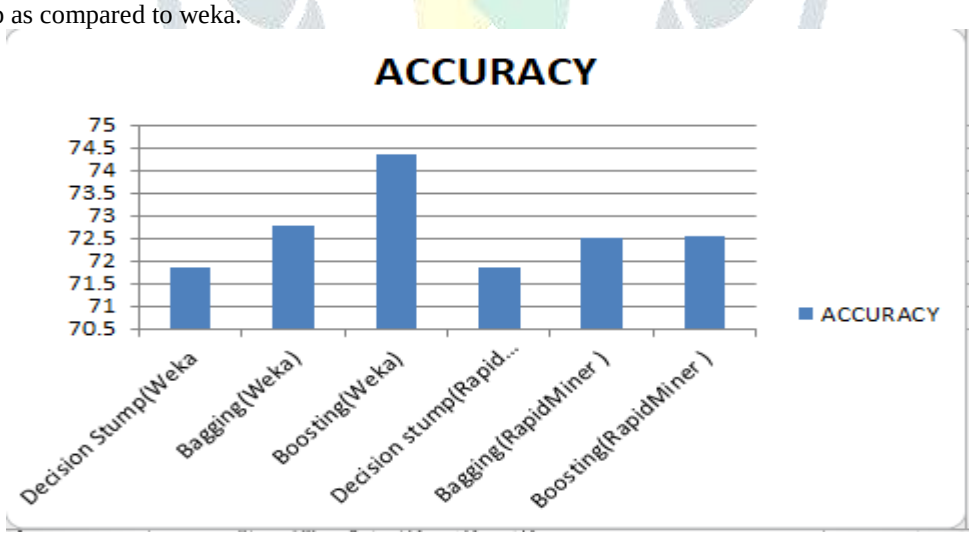

Fig. 2 Accuracy using DecisionStump as Base Classifier

3.1.3 Using RandomTree as Base Classifier:

Weka tool perform well as compare to the RapidMiner tool and Ensemble learning algorithm enhance the performance of RandomForest algorithm in both the tools. Boosting algorithm gives more accuracy in Weka tool while Bagging gives more accuracy in Rapid miner tool. Weka tool both the Ensemble learning algorithm. enhance the performance of RandomForest algorithm and Boosting algorithm perform well as compare to Bagging algorithm. In Rapid miner tool both the algorithm enhance the performance of RandomForest algorithm and Begging perform well as compared to Boosting

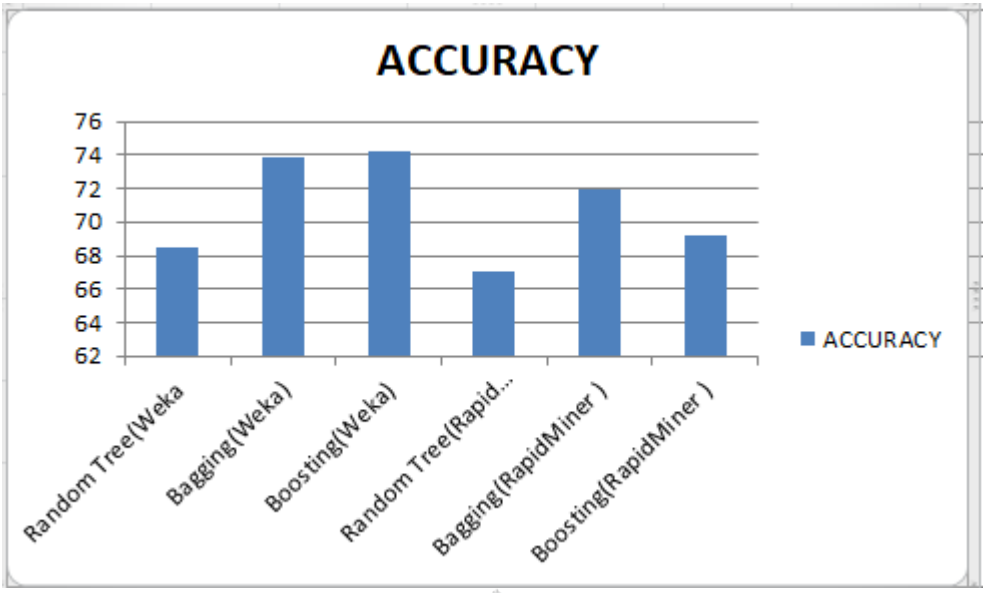

Fig. 3 Accuracy using RandomTree as Base Classifier

3.2 Precision

RapidMiner and Weka tool are used for the analysis purpose on the basis of evaluation parameter precision. The algorithms used for the evaluation are NaiveBayes, DecisionStump, RandomTree and Ensemble learning algorithm i.e.Bagging and Boosting.The results are shown in the graph to observe the performance of algorithms with dataset.

3.2.1 Using NaiveBayes as Base Classifier:

we evaluate the precision for dataset DS on RapidMiner and Weka. Weka tool gives more value of precision when we used NaiveBayes and Boosting Ensemble learning algorithm as compare to the RapidMiner tool. Ensemble learning algorithm does not enhance value of precision in RapidMiner tool. Bagging Ensemble learning algorithm enhance the base classifier in RapidMiner tool.

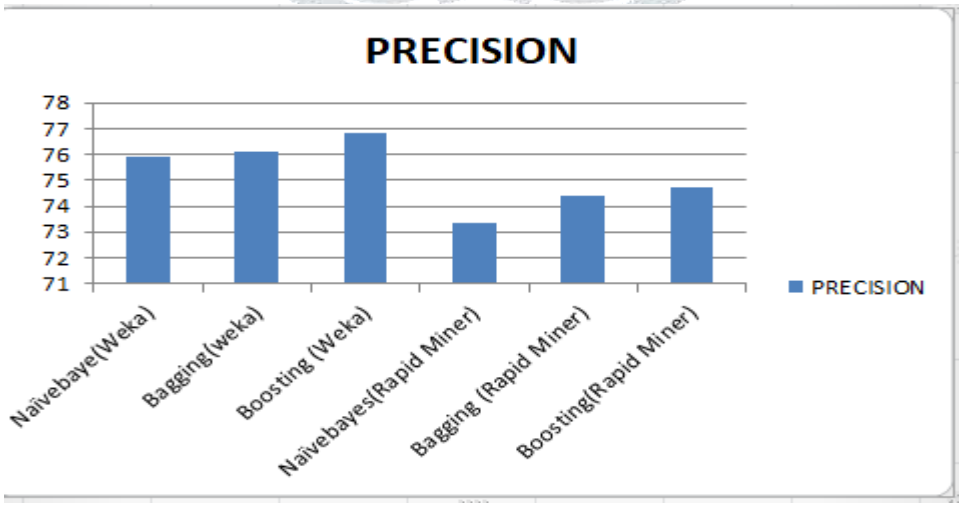

Fig. 4 Precision using NaiveBayes as Base Classifier

3.2.2 Using DecisionStump as Base Classifier

In the given dataset RapidMiner tool gives more value of precision as compare to Weka tool and Ensemble learningalgorithms does not enhance the value of precision in RapidMiner tool but in Weka tool Ensemble learning algorithms enhance the value of precision. RapidMiner tool performs well as compare to the Weka tool and both Ensemble learning algorithm enhance the precision of base classifier DecisionStump in Weka tool but in Rapid miner tool both Ensemble learning algorithms does not enhance the precision..

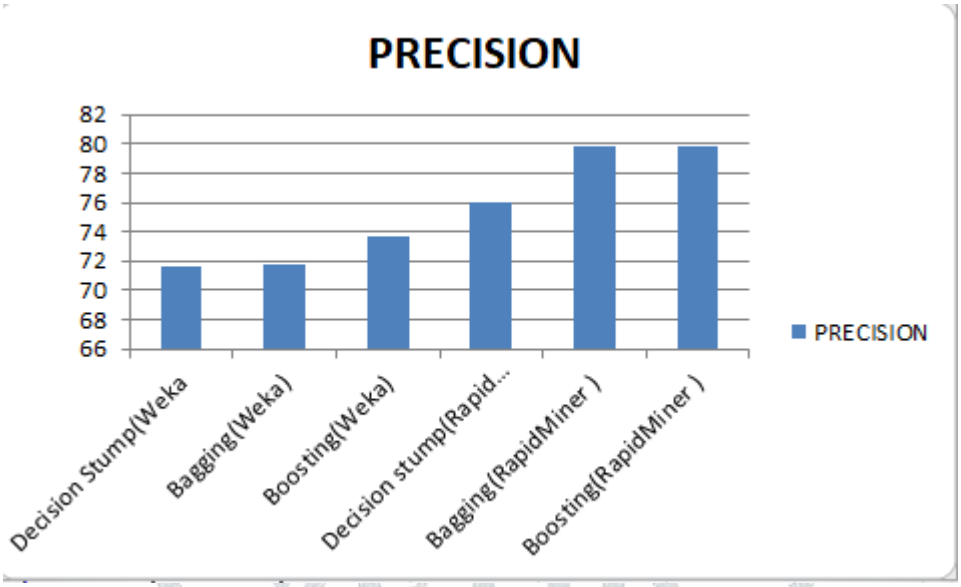

Fig. 5 Precision using DecisionStump as Base Classifier

3.2.3 Using RandomTree as Base Classifier

We evaluate the precision of dataset on RapidMiner and Weka tool. Weka tool using Boosting Ensemble learning algorithm gives more value of precision as compare to the RapidMiner tool. In the RapidMiner tool ensemble learning algorithm begging enhance the precision of base classifier and gave better accuracy then boosting.

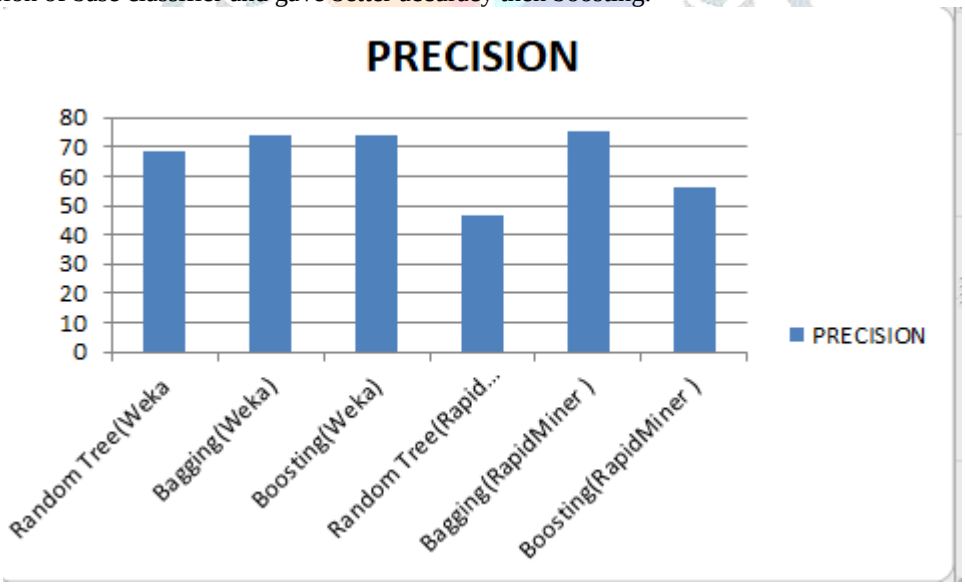

Fig. 6 Precision using RandomForest as Base Classifier

3.3 Recall

RapidMiner and Weka tool are used for the analysis purpose on the basis of evaluation parameter recall. The algorithms used For the evaluation are NaiveBayes, DecisionStump, RandomForest and Ensemble learning algorithm i.e. Bagging and Boosting.The results are shown in the graph to observe the performance of algorithms in different datasets with different size.

3.3.1 Using NaiveBayes as Base Classifier

In the given dataset, In case of DS2, Weka tool gives more recall as compare to the RapidMiner tool and Bagging Ensemble learning algorithm enhance the value of recall of NaïveBayes in RapidMiner tool. Boosting Ensemble learning algorithm enhance the value of recall of base classifier NaïveBayes in RapidMiner tool. Weka tool performs well as compare to the RapidMiner tool and Bagging Ensemble learning algorithm enhance the value of recall of base classifier NaïveBayes in RapidMiner tool.

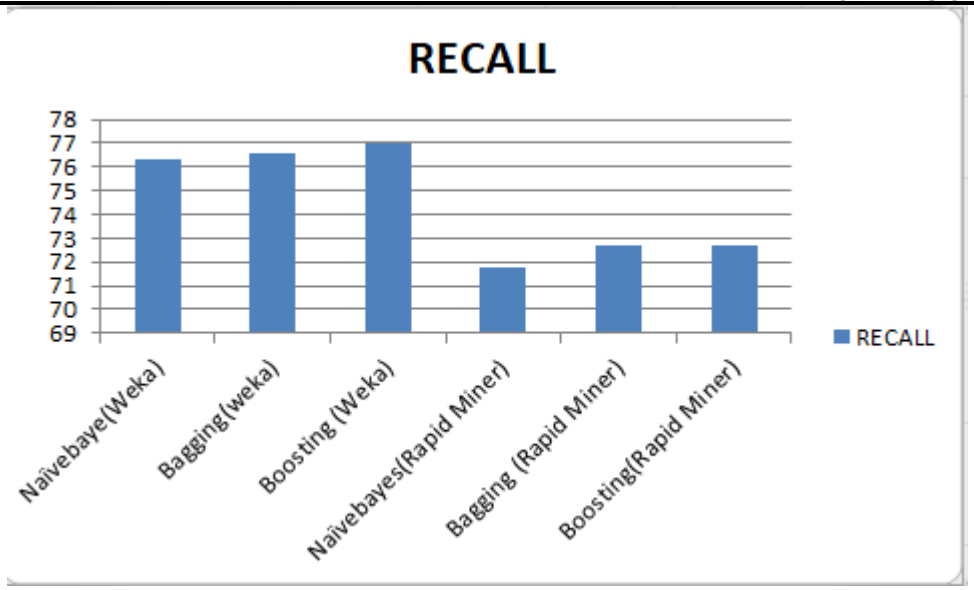

Fig. 7 Recall using NaiveBayes *as Base* Classifier

3.3.2 Using DecisionStump as Base Classifier

Weka tool gives more value of recall as compare to the RapidMiner tool . Weka tool using Boositng Ensemble learning algorithm gives more value of recall as compare to the RapidMiner tool. Weka tool using Boositng Ensemble learning algorithm gives more value of recall as compared to the RapidMiner tool.

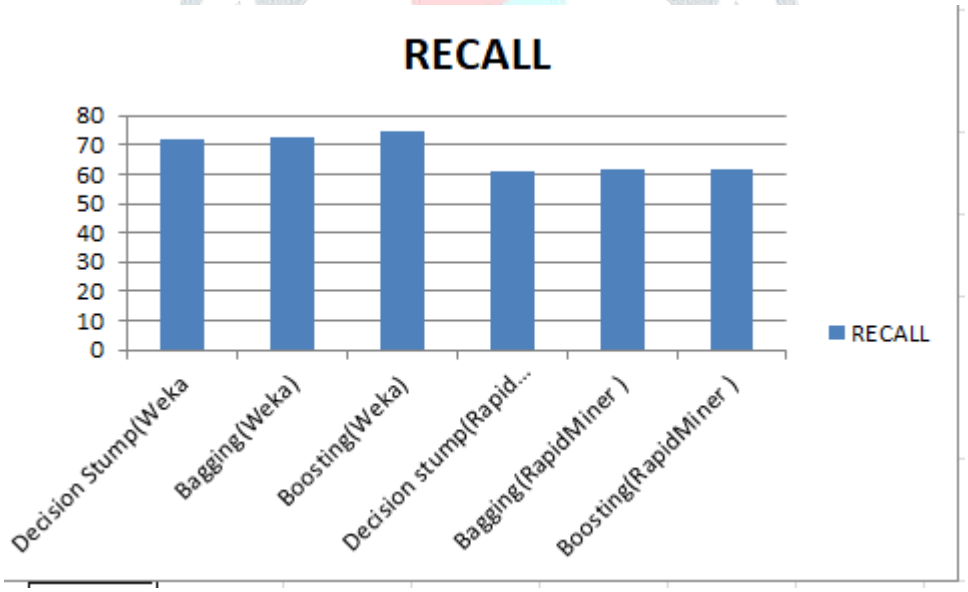

Fig. 8 Recall using DecisionStump as Base Classifier

3.3.3 Using RandomTree as Base Classifier

Weka tool using Boosting Ensemble learning algorithm gives more value of recall as compare to the RapidMiner tool.in RapidMiner tool Bagging ensemble learning algorithm enhance rmore recall vaue than boosting learning algorithm. Weka Tool Perform well than Rapid miner tool.

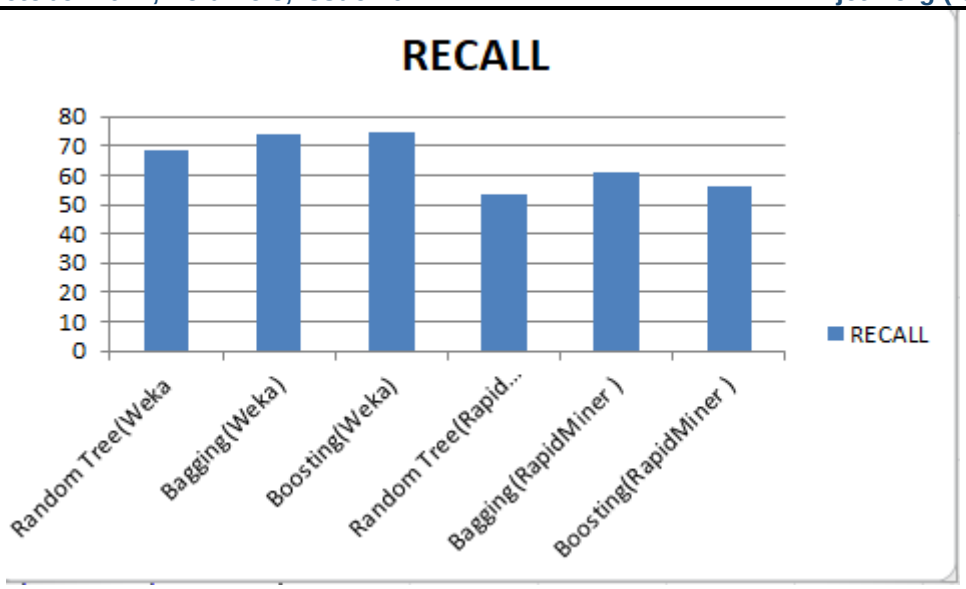

Fig. 9 Recall using RandomForest as Base Classifier

3.4 Root Mean Squared error

RapidMiner and Weka tool are used for the analysis purpose on the basis of evaluation parameter root mean squared error denoted as RMSE. The algorithms used for the evaluation are NaiveBayes, DecisionStump, RandomForest and Ensemble learning algorithm i.e. Bagging and Boosting. The results are shown in the graph to observe the performance of algorithms in dataset

3.4.1 Using NaiveBayes as Base Classifier

Weka tool using Boosting Ensemble learning algorithm gives least root mean squared error as compared to the RapidMiner toolRapidMiner tool using Begging Ensemble learning algorithm gives least root mean squared error as compared to then RapidMiner tool

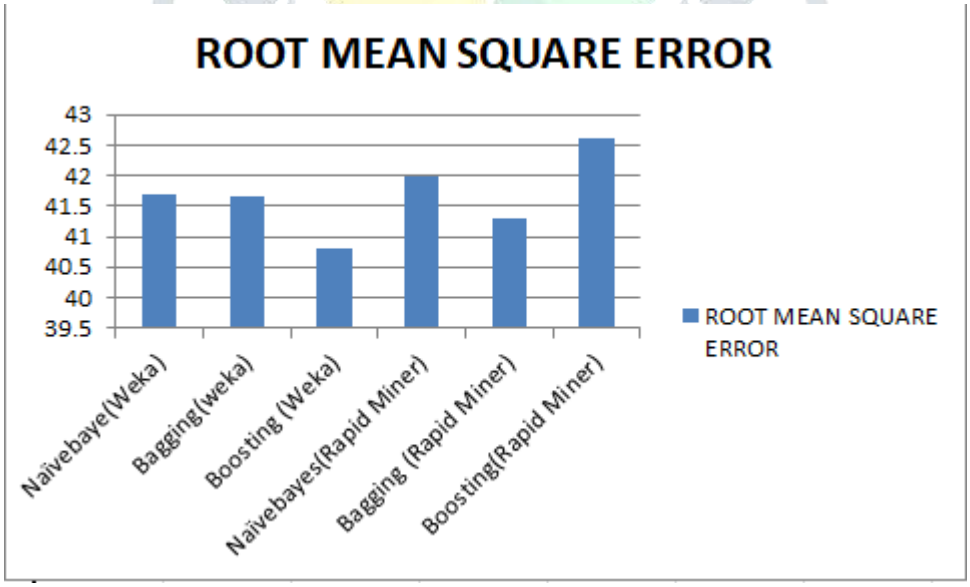

Fig. 10 RMSE using NaiveBayes as Base Classifier

3.4.2 Using DecisionStump as Base Classifier

Weka tool using Boosting Ensemble learning algorithm gives least root mean squared error as compared to the RapidMiner tool. In RapidMiner tool Bagging Ensemble learning algorithm give least root mean squared error as compare to boosting.

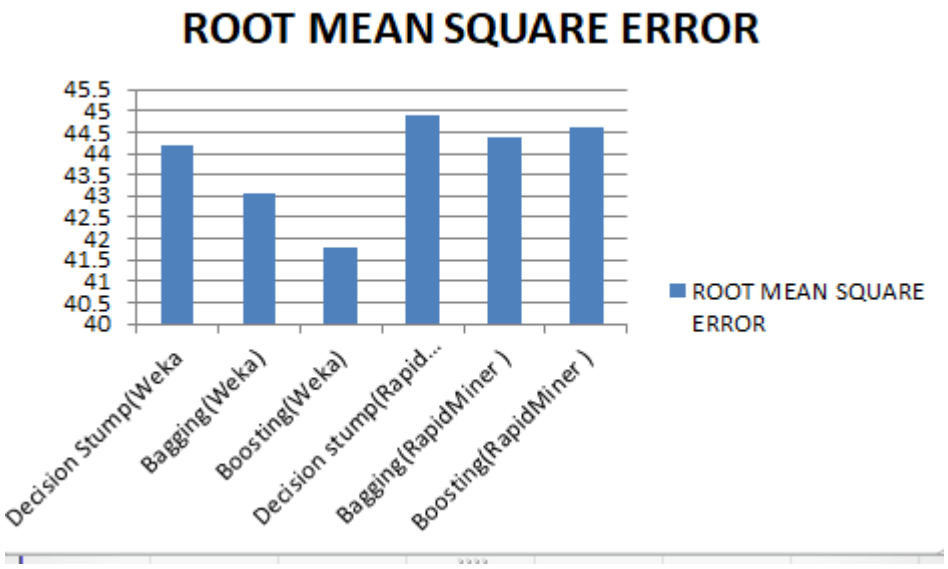

Fig. 11 RMSE using DecisionStump as base Classifier

3.4.3 Using RandomTree as Base Classifier

Weka tool gives least root mean squared error as compared to the RapidMiner tool. Boosting Ensemble learning algorithm gives least root mean squared error in Weka tool. In RapidMiner tool Begging Ensemble learning algorithm gives least root mean squared error. In RapidMiner tool, Begging Ensemble learning algorithm gives least root mean quared error but in Weka tool RandomTree Bagging algorithm gives least root mean squared error

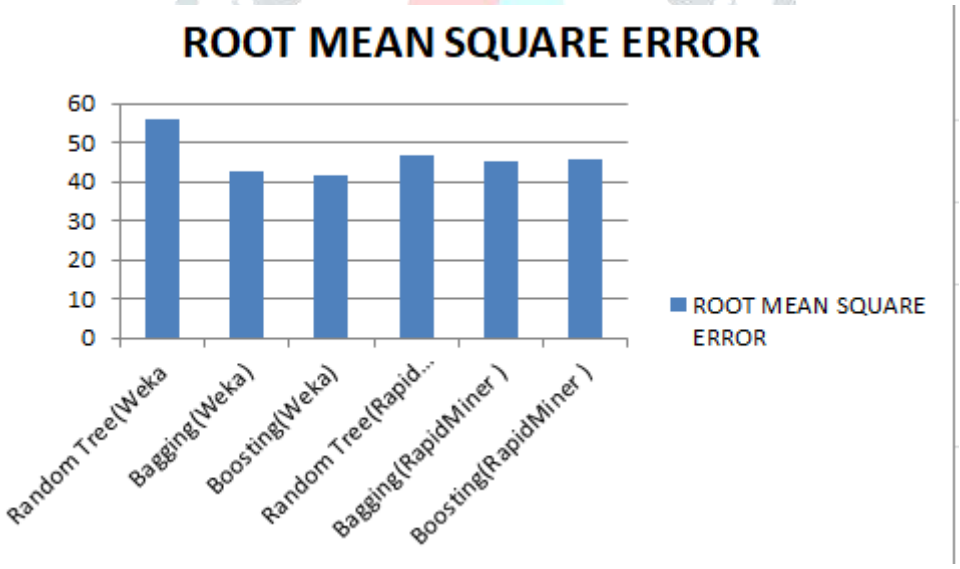

Fig. 12 RMSE using RandomForest as Base Classifier

IV. Conclusion and Future Scope

From the above result it has been observed that Ensemble learning algorithms have highest accuracy, precision, recall and least root mean squared error in both the tools as compared to the base classification algorithm. In some cases Ensemble learningalgorithms have same value as the base classification algorithm but it has never decreased the performance of any base classification algorithm in both the tools. These technique indicate that Ensemble learning algorithms are most efficient algorithms as compare to the other one. Although considering the performance of two tools, it is hard to conclude which one is performingbetter. The behaviour of tool varies with the type of dataset and algorithm used. Also every base classifier is behaving differently in two tools. Although both ensemble learning techniques enhances the performance of base algorithm but Boostingoutperforms Bagging in terms of Precision, Recall, and Accuracy but in case of DecisionStump, NaïveBayes, Bagging outperform Boosting. So it can be concluded that there is a "no-freelunch" policy for the ensembling algorithm. The performance of ensemble learning algorithm depends on the base algorithm.

For future scope the same algorithms can be run on different datasets and another data mining tools can be used instead of RapidMiner and Weka to analyze the performance of algorithms on different datasets. The impact of change in number of iteration in algorithms can also be observe.

REFERENCES

[1] Ajay Prashar, K.L. Bansal, "Analysis of various learning ensemling algorithm using increasing data set "jetir" ",vol. 5(10) pp 89-98,2018
[2] Jiawei Han and Micheline, "Data Mining: Concepts and Techniques", 2nd ed., Morgan Kaufmann, 2000.
[3]Usama Fayyad, Grgory Piatetsky-Shapiro and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Database", AAAI, vol. 17(3), pp 37-54, 1996
[4] Mikut, R. and Reischl, M., "Data mining tools. Wiley interdisciplinary reviews: data mining and knowledge discovery", vol. 1(5), pp.431-443, 2011.
[5]Sagar S. Nikam, "A Comparative Study of Classification Techniques in Data Mining Algorithms," Oriental Journal of Computer Science and Technology, vol. 8(1), pp 13-19, 2015.
[6]V. Krishnaiah et al., "Survey of Classification Techniques in Data Mining," International Journal of Computer Sciences and Engineering, vol. 2(9), pp 65-74, 2014.
[7] Margaret H. Dunham, "Data Mining Introductory and Advanced Topics",Dorling KindersleyvPvt.Ltd.India,Sixth Edition,2013
[8] E. Suriyapriya, M. Praveena, Clustering and Boosting in Data Mining, International Journal of Engineering Science and Computing, Vol 7(8), 2017.
[9] Prajakta S. Kasbe, Apeksha V. Sakhare, A Review On Road Accident Data Analysis Using Data Mining Techniques, IEEE, 2017.
[10] Anand Kishor Pandey, Dharmveer Singh Rajpoot, A Comparative Study of Classification Techniques By Utilizing WEKA, IEEE, 2016.
[11] Sumouli Choudhury, Anirban Bhowan et al., Comparative Analysis of Machine Learning Algorithms along with Classifiers for Network Intrusion Detection, International Conference on Smart Technologies and Management (ICSTM), pp 89-95, 2015.
[12] Pooja Shrivastava, Manoj Shukla, Comparative Analysis of Bagging, Stacking and Random Subspace algorithms, IEEE, 2015.