Performance Evaluation of Ensemble Learning Algorithm for Various Base Classifiers

Ajay Prashar[1], Dr. K. L. Bansal[2]
[1, 2] Dept of Computer Science
[1, 2] Himachal Pradesh University, Shimla

Abstract- *In present scenario huge amount of data and information is available to everyone either offline or online. Apart from being available on web or offline, different kind of databases and information repository offer data storage. With such huge amount of data, powerful techniques are required for data interpretation that exceeds the human ability for comprehension and decision making in a better way. Data mining offers various method or techniques that may be used to predict the accuracy of different classes of object. This research focuses on various ensemble learning techniques like Boosting and Bagging and also compares Boosting, Bagging ensemble learning data mining techniques with each other by using various base classifiers like J48, NaiveBayes, OneR, DecisionStump and DecisionTable. All the techniques are compared on five parameters like precision, recall, f-measure, means absolute error and accuracy. The findings are also supported with justification by conducting an experimental survey at the end of research.*

Keywords- WEKA, Ensembling Algorithm, Boosting, Bagging, Data Mining

I. INTRODUCTION

Currently databases and data repositories contains large amount of data and information that it becomes impossible to analyze them manually for accurate decision making. So we require data mining [1]. Data mining is the process of extracting or mining useful information from the large amount of data or we can also say that it is the process of discovering hidden patterns and information from the existing data and transform them in to an understandable structure for further uses. It is an interdisciplinary subfield of computer science [2]. In data mining; the main need to primarily concentrate on cleansing the data so as to make it feasible for furthers processing [3]. Data mining is the analysis step in the process of knowledge discovery in databases or KDD. The KDD process consist of various steps leading from raw data collection to some form of new meaningful knowledge. This process consist of following steps:

- **Data cleaning** also known as data cleansings, in this phase noise data and irrelevant data removed from the databases.
- **Data integration** at this step, multiple data source may be combined in to a common source.
- **Data selection** at this step, the data relevant to analysis task is decided and retrieved from the database.
- **Data transformation** in this step selected data are transformed into forms appropriate for mining by performing summary or aggregation operations, for instance. This is also called data consolidation,
- **Data mining** is an essential step where intelligent methods are applied in order to extract data patterns.
- **Pattern evaluation** in this step strictly interesting patterns identify representing knowledge based on some interestingness measures.
- **Knowledge presentation** is the last step where visualization and knowledge representation techniques are used to represent the mined knowledge to the user [4].

1.1 Ensemble Learning Technique

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

The main advantages of Ensemble learning method are:

- **Reduced variance:** This method solve overfitting problem. Low variance means that model independent of training data so the results are less dependent on features of single model and training set.
- **Reduced bias:** This method solve under fitting problem. Low bias means linear regression applied

on linear data, 2^{nd} degree polynomial applied to quadratic data. Combination of multiple classifiers produce reliable classification than the single classifier.

- **Improve prediction:** This method improve prediction. [2]

Common Ensemble learning methods are:

- Bagging
- Boosting
- Sampling

This paper works on two of the methods i.e. Boosting and Bagging.

1.1.1 Bagging

Bagging is also called bootstrap aggregation. It is an ensemble meta-algorithm 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.

1.1.2 Boosting

Boosting is a robust ensemble algorithm that is capable of reducing both bias and variance and also 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.

These ensemble learning method are compared with each other by using various base classifiers like:

- **J48 decision tree classifier** is an enhanced version of C4.5 decision tree classifier and also become one of the popular decision tree classifier This classifier build model using tree structure
- **NaiveByes classifier** is based on Bayesian theorem and 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.
- **OneR classifier** is a type of rule based classifier. This classifier uses the minimum error attribute for prediction, discretizing numeric attribute.
- **DecisionStump classifier** is a type of decision tree algorithm. This algorithm usually used in conjunction with boosting algorithm. This classifier can perform regression based on mean squared error and classification based on entropy.
- **DecisionTableclassifer**isa type of rule based classifier. Decision table consist of a hierarchical table in which each entry in a higher level table gets broken down by the values of a pair of additional attribute to form another table.

1.2 Evaluation Parameter

Boosting and Bagging ensemble learning algorithm are compared by taking a base classifiers on the basis of following evaluation parameter such as:

- **Precision** is also called positive predictive value. It is defined as relative number of correctly as positive classified example among all examples classified as positive
- **Recall** is the true positive rate also referred as sensitivity. This parameter specifies the relative number of correctly as positive classified example among all positive examples.
- **F-measure**is the combination of the precision and recall i.e. $f=2pr / (p + r)$

Where f, r and p are f-measure, recall and precision.

- **Mean absolute error**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.
- **Accuracy**is defined as a relative number of correctly classified instances or in other words percentage of correctly classified instances.

The rest of the paper is divided into sections as follows:Section 2 contains the literature review.Section 3 contains the result and discussion.Section 4 contains conclusion.

II. LITERATURE REVIEW

Swati Singhal, Monika Jena (2013) did a study on weka tool and introduce a brief introduction of the key principle of data pre-processing, classification, clustering and introduction of WEKA tool and also described various steps how to use WEKA tool for these particular technologies. [5].Yanmin Sun, Yang Wang et al. (2006)did a study on cost-sensitive boosting for classification of imbalanced data. Meta techniques were explored which were applicable to most classifier learning algorithm, with the aim to advance the classification of imbalance data.AdaBoostwas applied to an associative classifier for both learning time reduction and accuracy improvement [6]. A comprehensive survey on Data Mining was done by Sunita B Aher, Lobo L.M.R.J. (2011) in Educational System using WEKA. This paper surveys an application of data mining in education system and also represents their result analysis using WEKA tool [7].Poonam Pandey, Radhika Prabhakar (2016)didan Analysis of Machine Learning Techniques (J48 and AdaBoost) for classification. Based on number of experiment which performed, their results shows that AdaBoost provides better accuracy than Decision Tree (J48 algorithm) when the number of class label in the dataset are exactly two whereas the decision tree generate rules faster than the AdaBoost and when number of class label in the dataset are more than two then J48 algorithm performs better than AdaBoost [8]. RushiLongadge, Snehlata S. Dongre et al. (2013)discuss class imbalance problem in data mining. In this research paper awell-defined study of each approach is defined which gives the right direction for research in class imbalance problem [9].Taghi M. Khoshgoftaar, Van Hulse et al. (2011) compared the performance of various boosting and bagging techniques in the context of learning from imbalanced and noisy binary-class data.Based on number of experiment which performed, their results shows that the bagging techniques generally outperform boosting, so in case of noisy data environment, bagging is the most efficient method for handling class imbalance [10].Arvind Sharma, P.C. Gupta (2012)did a survey and predictedthenumber of blood donors based on their age and blood group. For this research work J48 algorithm and WEKA tool was used. Through training and evaluations, the experimental analysis shows that the generated classification rules performed well in the classification of blood donors, whose accuracy rate reached 89.9% maximum [11].Pooja Shrivastava, Manoj Shukla (2013)showed the comparing results using bagging, stacking and random subspace algorithm on forest fire dataset using WEKA tool. Stacking algorithm among all built accurate classifier model and consume less time [12].

III. RESULTS AND DISCUSSION

The two ensembling learning algorithm are compared with each other by taking different data mining algorithm as the base classifiers like: J48, NaiveBayes, OneR, DecisionStump and DecisionTable

The dataset used is downloaded from the UCI Repository website.The data mining tool used is Weka Tool.

3.1 Comparison of ensembling algorithm using J48 decision tree algorithm

Boosting and Bagging ensemble learning algorithm are compared by taking J48 decision tree algorithm as a base classifiers on the basis of following evaluation parameter such as: Precision, Recall, F-measure, Mean absoluteerror and accuracy and is shown in both tabular as well as graphical form

Table3.1Comparison of ensembling algorithm using J48 decision tree algorithm

Algorithm	Precision	Recall	F-Measure	Mean absolute error	Accuracy
J48	0.990	0.990	0.990	0.0225	0.990
AdaBoostM1	0.995	0.995	0.995	0.0188	0.995
Bagging	1.00	1.00	1.00	0.0381	1.00

In Table 3.1 J48 algorithm is compared with AdaBoostM1 and Bagging algorithm. From this comparison it is observed that the value of precision, recall, f-measure and accuracy of Bagging algorithm is high as compare to the other. So in this case Bagging is more efficient algorithm. This can also be represented in the form of graph.

The figure shows the graphical representation of the result interpreted from the table. This column chart have three different coloured column, blue column represent J48 algorithm, orange column represent AdaBoostM1 and grey column represent Bagging algorithm. This graph clearly shows that Bagging algorithm have highest precision, recall, f-measure and accuracy value. So in this case Bagging is more efficient algorithm.

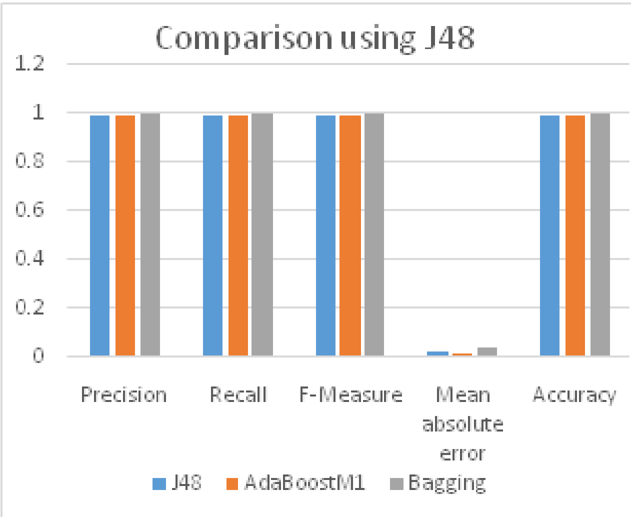


Figure 3.1 Comparison of ensembling algorithm using J48 decision tree algorithm

3.2 Comparison of ensembling algorithm usingNaïveBayes algorithm

Boosting and Bagging ensemble learning algorithm are compared by takingNaïveBayes algorithm as a base classifier. The number of iterations taken are 10.

Table 3.2 Comparison of ensembling algorithm usingNaïveBayes algorithm

Algorithm	Precision	Recall	F-Measure	Mean absolute error	Accuracy
NaiveBayes	0.956	0.950	0.951	0.0479	0.950
AdaBoostM1	0.981	0.980	0.980	0.0194	0.980
Bagging	0.960	0.955	0.955	0.0456	0.955

In Table 3.2 NaiveBayes algorithm is compared with AdaBoostM1 and Bagging algorithm. From this scomparisonit is observed that the value of precision, recall, f-measure and accuracy of AdaBoostM1 algorithm is high as compare to the other and also having least mean absolute error value. So in this caseAdaboostM1 is more efficient algorithm. This can also be represented in the form of graph.

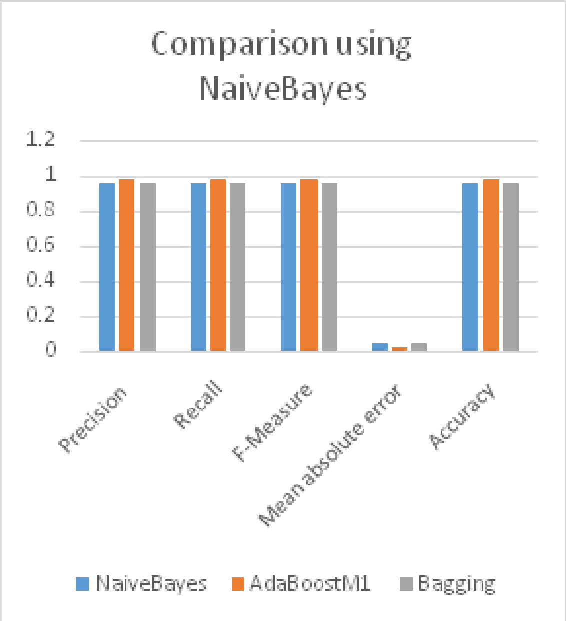


Figure 3.2Comparison of ensembling algorithm usingNaiveBayes algorithm

The above figure shows the graphical representation of the result interpreted from the table. This column chart have three different coloured column, blue column represent NaiveBayes, orange column represent AdaBoostM1 and grey column represent Bagging algorithm. This graph clearly shows that AdaBoostM1 algorithm have highest precision, recall, f-measure, accuracy and least mean absolute error value. So in this case AdaBoostM1 is more efficient algorithm.

3.3 Comparison of ensembling algorithm usingOneR algorithm

Boosting and Bagging ensemble learning algorithm are compared by takingOneR algorithm as a base classifier on the basis of following evaluation parameter such as: Precision, Recall, F-measure, Mean absolute error and Accuracy and is shown in both tabular as well as graphical form.

Table 3.3Comparison of ensembling algorithm usingOneR algorithm

Algorithm	Precision	Recall	F-Measure	Mean absolute error	Accuracy
OneR	0.922	0.920	0.920	0.0800	0.920
AdaBoostM1	0.975	0.975	0.975	0.0277	0.975
Bagging	0.927	0.925	0.925	0.0903	0.925

In Table 3.3 OneR algorithm is compared with AdaBoostM1 and Bagging algorithm. From this comparison it is observed that the value of precision, recall, f-measure and accuracy of AdaBoostM1 algorithm is high as compare to the other and also having least mean absolute error value. So in this case AdaBoostM1 is more efficient algorithm. This can also be represented in the form of graph.

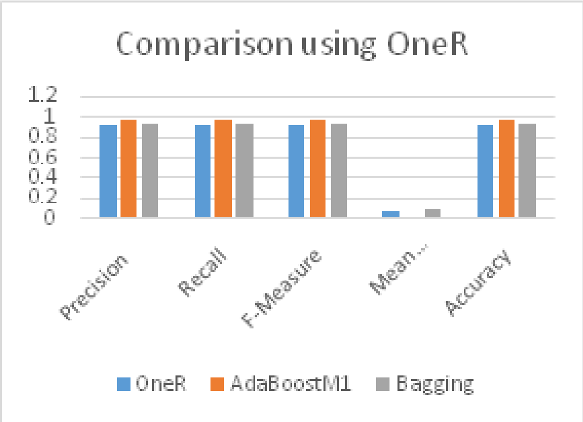


Figure 3.3 Comparison of ensembling algorithm usingOneR algorithm

The above figure shows the graphical representation of the result interpreted from the table.This column chart have three different coloured column, blue column represent OneR, orange column represent AdaBoostM1 and grey column represent Bagging algorithm. This graph clearly shows that AdaBoostM1 algorithm have highest precision, recall, f-measure, accuracy and least mean absolute error value. So in this case AdaBoostM1 is more efficient algorithm.

3.4 Comparison of ensembling algorithm usingDecisionStump algorithm

Boosting and Bagging ensemble learning algorithm are compared by taking DecisionStump algorithm as a base classifier on the basis of following evaluation parameter such as: Precision, Recall, F-measure, Mean absolute error and Accuracy and is shown in both tabular as well as chart forms.

Table 3.4 Comparison of ensembling algorithm usingDecisionStump algorithm

Algorithm	Precision	Recall	F-Measure	Mean absolute error	Accuracy
DecisionStump	0.926	0.920	0.921	0.1362	0.920
AdaBoostM1	0.990	0.990	0.990	0.0182	0.990
Bagging	0.932	0.928	0.928	0.1398	0.9275

In Table 3.4 DecisionStump algorithm is compared with AdaBoostM1 and Bagging algorithm. From this comparison it is observed that the value of precision, recall, f-measure and accuracy of AdaBoostM1 algorithm is high as compare to the other and also having least mean absolute error value. So in this case AdaBoostM1 is more efficient algorithm. This can also be represented in the form of graph.

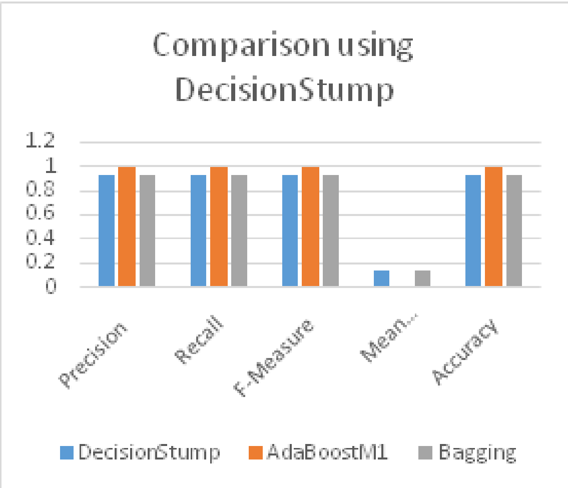


Figure 3.4 Comparison of ensembling algorithm usingDecisionStump algorithm

The above figure shows the graphical representation of the result interpreted from the table.This column chart we have three different coloured column, blue column represent DecisionStump, orange column represent AdaBoostM1 and grey column represent Bagging algorithm. This graph clearly shows that AdaBoostM1 algorithm have highest precision, recall, f-measure, accuracy and least mean absolute error value. So in this case AdaBoostM1 is more efficient algorithm.

3.5 Comparison of ensembling algorithm usingDecisionTable algorithm

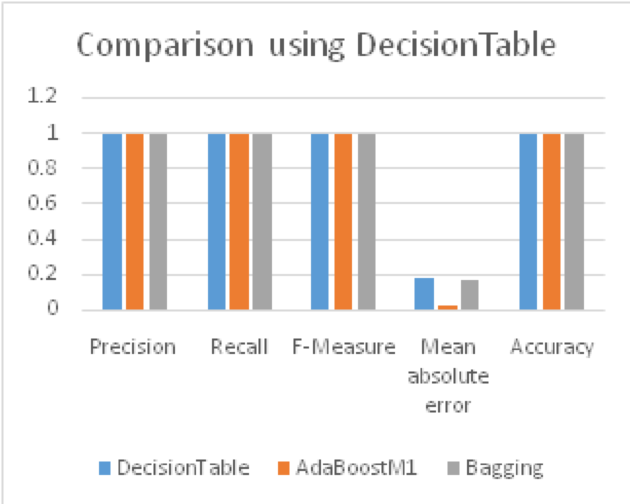
Boosting and Bagging ensemble learning algorithm are compared with DecisionTable algorithm on the basis of following evaluation parameter such as: Precision, Recall, F-measure, Mean absolute error and Accuracy and is shown in both tabular as well as graphical forms.

Table 3.5 Comparison of ensembling algorithm usingDecisionStump algorithm

Algorithm	Precision	Recall	F-Measure	Mean absolute error	Accuracy
DecisionTable	0.990	0.990	0.990	0.1815	0.990
AdaBoostM1	0.995	0.995	0.995	0.0265	0.995
Bagging	0.988	0.988	0.988	0.1637	0.9875

In Table 3.5 DecisionTable algorithm is compared with AdaBoostM1 and Bagging algorithm. From this comparison we know that the value of precision, recall, f-measure and accuracy of AdaBoostM1 is high as compare to the other and also having least mean absolute error value. So in this case AdaBoostM1 is more efficient algorithm. This can also be represented in the form of graph.

Table 3.5 Comparison of ensembling algorithm usingDecisionTable algorithm



The above figure shows the graphical representation of the result interpreted from the table. This column chart have three different coloured column, blue column represent DecisionTable, orange column represent AdaBoostM1 and grey column represent Bagging algorithm. This graph clearly shows that AdaBoostM1 algorithm have highest precision, recall, f-measure, accuracy and least mean absolute error value. So in this case AdaBoostM1 is more efficient algorithm.

From these comparison we conclude that both ensemble learning techniques enhances the performance of base algorithm but Boosting outperforms Bagging in terms of Precision, Recall, F-measure and Accuracy but in case of DecisionStump, NaïveBayes, Bagging outperform Boosting.

So it can be concluded that there is a “no-free lunch policy” for the ensembling algorithm. The performance of ensemble learning algorithm clearly depends on the base classifier used.

IV. CONCLUSION AND FUTURE SCOPE

Many advance tool for data mining techniques are available currently. In this research work Weka has been used as a data mining tool. It is free software licensed under GNU General Public License. In this paper various ensemble learning techniques like Boosting and Bagging have been discussed. This paper compares Boosting and Bagging ensemble learning data mining techniques by using many classifier or learning algorithmlike J48, NaiveBayes, OneR, DecisionStump and DecisionTable. All the techniques are compared on five parameters like precision, recall, f-measure, means absolute error and accuracy. A dataset of chronic kidney diseases has been used to train and test the various classification model that are opted to compare. It has been observed that the ensemble learning algorithm have highest precision, recall, f-measure, and accuracy value which indicate that ensemble learning techniques are most efficient algorithm as compare to the other one. Although both ensemble learning techniques enhances the performance of base algorithm but Boostingoutperforms Bagging in terms of Precision, Recall, F-measure 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 ofensemble learning algorithm depends on the base algorithm.

For future scope the same algorithm can be run on different application domain or another data mining tool can be used instead of weka to analyze the performance of ensembling algorithm. The impact of change in number of iteration can also be observed.

REFERENCES

[1] Goebel et al. , A Survey of Data Mining and Knowledge Discovery Software Tools, ACM SIGKDD Explorations Newsletter, Vol 1(1), pp 20-33,1999

[2] Jiawei Han and Micheline, Data Mining: Concepts and Techniques, Morgan Kaufmann Second Edition, 2000.

[3] M.S.B. PhridviRaj and C.V. GuruRao, Data Mining- Past, Present and Future – A Typical Survey on Data Streams, Procedia Technology, Vol 12, pp 255-63,2014

[4] Zaïane OR, Principles of knowledge discovery in databases, Department of Computing Science University of Alberta, 20,1999.

[5] Swati Singhal and Monika Jena, A Study on WEKA Tool for Data Preprocessing, Classification and Clustering, International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol 2(6), 2013.

[6] Yanmin Sun, Yang Wang, et al., Cost-Sensitive Boosting for the Classification of Multi-Class Imbalanced Data, IEEE, 2006

[7] Sunita B Aher, Lobo L.M.R.J. , Data Mining in Educational System using WEKA, International Conference on Emerging Technology Trends (ICETT), Vol 3, pp 20-25, 2011.

[8] Poonam Pandey, Radhika Prabhakar, An Analysis of Machine Learning Techniques (J48 &AdaBoost) for Classification, IEEE, pp 1-6, 2016

[9] RushiLongadge, Snehlata S. Dongre et al., Class Imbalance Problem in Data Mining: Review, International Journal of Computer Science and Network (IJCSN), Vol 2(1), 2013

[10] Taghi M. Khoshgoftaar, Van Hulse et al., Comparing Boosting and Bagging Techniques with Noisy and Imbalanced Data, IEEE, Vol 41(3), 2011.

[11] Arvind Sharma, P.C. Gupta, Predicting the Number of Blood Donors through their Age and Blood Group Using Data Mining Tool, International Journal of Communication and Computer Technologies (IJCCTS), Vol 01(6), 2012.

[12] Pooja Shrivastava, Manoj Shukla, Comparative Analysis of Bagging, Stacking and Random Subspace algorithms, IEEE, 2015.