© 2016 IJSRSET | Volume 2 | Issue 5 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

# Empirical Studies and Analysis of Ensemble Learning Techniques in Data Mining

**Bhavesh Patankar*[1], Dr. Vijay Chavda[2]**

*[1]Research Scholar, Department of Computer Science, Hemchandracharya North Gujarat University, Patan, Gujarat, India.
[2]NPCCSM, Kadi Sarva VishwaVidyalaya, Gandhinagar, Gujarat, India.

## ABSTRACT

This Classification using ensemble generally combines multiple classifiers that results in the improvement in the accuracy of the classification. Experimenting with the same dataset using the single classifier provides lesser accuracy than ensemble techniques. Many researches have been carried out using the technique of combining the predictions of multiple classifiers to generate a single classifier. The produced classifiers provide more accurate results than any individual classifier. This paper focuses on the ability of ensemble techniques to improve the accuracy of basic J48 algorithm. Ensemble techniques like Bagging and Boosting improved the efficiency of the J48 classifier. Experiments have been carried out on many datasets taken from UCI repository to investigate the effects of ensemble techniques on J48 and Naïve Bayes algorithm. WEKA tool is used to measure the effectiveness of a classifier model.

**Keywords:** Ensemble, Boosting, Bagging, J48, Naïve Bayes

## I. INTRODUCTION

Ensemble learning techniques have been called the most prominent development in Data Mining especially in Machine Learning during the past decades. They combine multiple learning models into one which usually more accurate than the best of its constituents. Ensembles can generate a serious boost to various real world challenges -- from rain forecasting to discovery in medical science, and scam detection to endorsement systems -- where predictive accuracy is more essential than model understanding. Ensembles are useful with all modeling algorithms, but this paper- focuses on decision trees to explain them most clearly. The paper first provide information on ensemble learning, J48 algorithm then an empirical studies is carried out using various UCI repository datasets. Finally results are shown with conclusion at the end of the paper. It also shows the future work can be carried out in this field.

The paper is organized as follows: Section 2 provides literature review an effective method of analysing the findings in the research. Section 3 shows the experiment and analysis of the results. Section 4 presents conclusion and future work.

## II. METHODS AND MATERIAL

### Literature Review

Muhlbaier et. Al. have presented an incremental learning algorithm Learn++, which learns new information from successive data sets by creating an ensemble of classifiers with each data set, and uniting them by weighted majority voting. They have introduced dynamically weighted consult and vote (DW-CAV), a new voting tool for combining classifiers: individual classifiers check with each other to determine which ones are most qualified to do classification of a given instance, and assign how much weight, if need to have any, each classifier's assessment should carry[1]. Nishida et.al. focused on becoming accustomed to various types of concept drift which is important for dealing with real-world online learning problems. To achieve this, they previously reported an online learning mechanism that uses an ensemble of classifiers, the adaptive classifiers-ensemble (ACE) system. The

adaptive classifiers ensemble comprises of one classifier, also many batch classifiers, and a drift detection technique. In order to increase the performance of adaptive classifiers ensemble, they have improved the weighting method, which combines the results of classifiers, and also they have included a novel classifier pruning method [2]. Li et.al. have carried out their work on computer-aided diagnosis (CAD). In computer-aided diagnosis (CAD), machine learning methods have been extensively applied to learn a hypothesis from diagnosed samples to help the medical professionals in generating a diagnosis. In order to learn a well-performed assumption, a huge amount of diagnosed samples are needed. While the samples can be easily collected from regular medical investigations, it is generally difficult for medical professionals to make a conclusion for each of the gathered samples. If an assumption could be learned in the existence of a large amount of undiagnosed samples, the substantial burden on the medical professionals could be free. In this paper, a novel semi-supervised learning method named Co-Forest is projected. It covers the co-training model by consuming a wide spread ensemble learning scheme entitled Random Forest, which permits Co-Forest to calculate the classification buoyancy of undiagnosed samples and effortlessly generate the final assumption [3]. Zhao et.al. have carried out a survey on neural network ensembles. A neural network ensemble combines a fixed number of neural networks or other types of forecasters, which are given training concurrently for a common classification job. In comparison with a sole neural network, the ensemble is much more efficient to improve the generalization ability of the classifier. The main aim of their work in the paper is to present existing research effort on the neural network ensembles, comprising real analysis, common implementation stages of ensembles, and conventional tools for training part neural networks [4].

Dietterich et.al. have carried out study on ensemble techniques. In the paper they have shown that ensemble methods are learning schemes that build a set of classifiers and after that classify new data samples by taking a (weighted) vote of their estimates. Bayesian averaging is the original ensemble learning technique, but more modern learning schemes include bagging, and boosting. In the paper they reviewed these techniques and explained why ensembles can often out perform any single classifier [5].

## III. RESULTS AND DISCUSSION

### Experiment

J48 is a dominant decision tree method that performs well on the many of the dataset. In this experiment we are going to explore whether we can improve upon the result of the J48 algorithm using ensemble learning schemes. Boosting and bagging, the popular ensemble learning schemes have been used to perform the experiment. The experiment is carried out using Weka [6]. Also datasets used in this experiment has been take from the UCI machine learning repository [7].

In this experiment Ionosphere, glass and soybean datasets have been used. These datasets are taken from the UCI machine repository. A base classifier is used to perform classification on the chosen datasets. Here no filter is applied on the chosen datasets. Testing scheme 10-fold cross validation is used. Two learning algorithms J48 and Naïve Bayes are used to perform the classification.

Three datasets namely glass, Ionosphere and soybean are taken from UCI machine repository. Details of the datasets is given in below table.

**Table 1.** Datasets used in the experiment

| Sr. No. | Dataset | No. of Instances | No. of Attributes |
|---|---|---|---|
| 1. | Glass | 214 | 10 |
| 2. | Ionosphere | 351 | 35 |
| 3. | Soybean | 683 | 36 |

Below figure show the experiment setup.

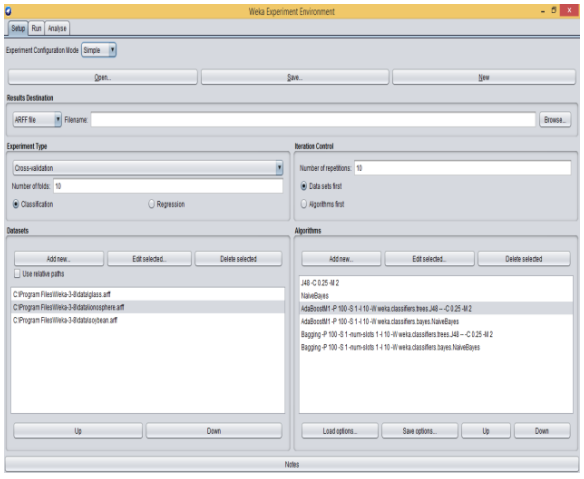

**Figure 1.** Experiment setup in Weka

Three datasets namely glass, Ionosphere and soybean have been added on the left side. Two base algorithms J48 and Naïve Bayes have been added on the righ hand side. Along with this algorithm ensemble learning algorithms namely boosting and bagging have also been added using above two base classifiers. So in all total six algorithms are added here and along with them three datasets are also added to perform the test. After setting up the experiment, it is run using the run tab on Weka experimenter.

Experiment results are shown in tabular form in the given below figure.

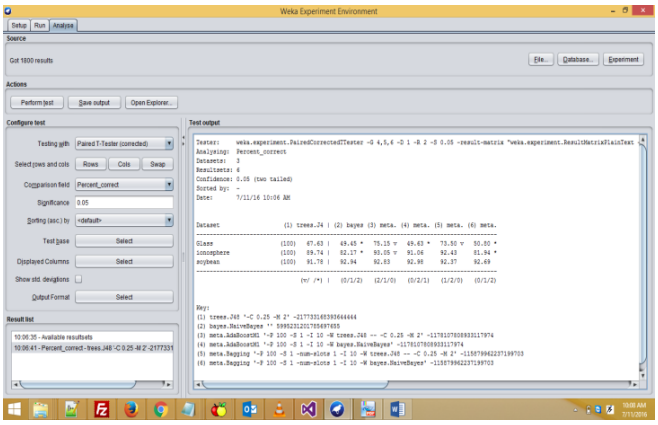

**Figure 2.** Experiment results in Weka.

From the results, it can be quite clear that ensemble learning scheme out perform the base classifier in terms of classification accuracy.

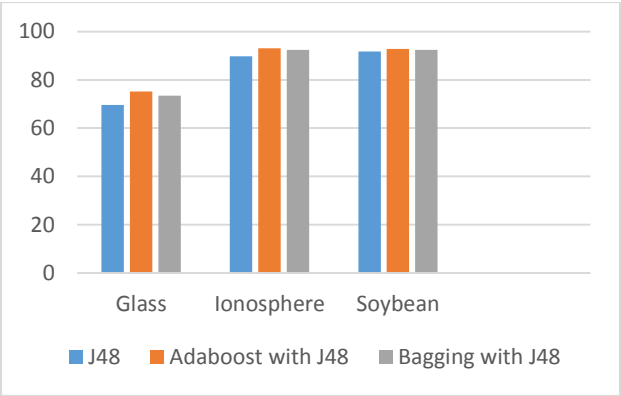


**Figure 3.** Comparison of Boosting and Bagging vs J48 as base classifier

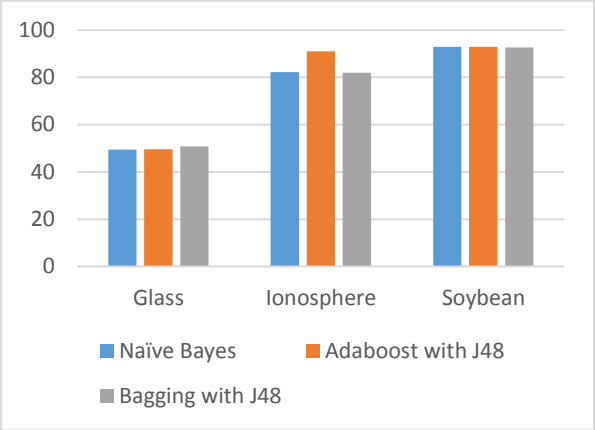


**Figure 4.** Comparision of Boosting and Bagging vs single classifier using Naïve Bayes as base classifier

## IV.CONCLUSION AND FUTURE WORK

The paper compares the effect of boosting and bagging on classification accuracy by using J48 and Naïve Bayes as the base classifiers. The experiment shows the effect of boosting and bagging on J48 and Naïve Bayes base classifiers. It was observed that for all three different datasets, the classification accuracy increases when ensemble learning technique is used instead of a single classifier. The results show that bagging and boosting improves the performance of the base classifier. It is concluded that ensemble learning scheme of boosting and bagging assists in improving the accuracy of classification. Future work can be carried out to check the effects of changing the base classifier learner from J48 and Naïve Bayes to some other classifiers.

## V. REFERENCES


[1] Muhlbaier, Michael D., Apostolos Topalis, and Robi Polikar. "Learn. NC: combining ensemble of classifiers with dynamically weighted consult-and-vote for efficient incremental learning of new classes." IEEE transactions on neural networks 20.1 (2009): 152-168.

[2] Nishida, K. Y. O. S. U. K. E., and Koichiro Yamauchi. "Adaptive classifiers-ensemble system for tracking concept drift." 2007 International Conference on Machine Learning and Cybernetics. Vol. 6. IEEE, 2007.

[3] Li, Ming, and Zhi-Hua Zhou. "Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples." IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 37.6 (2007): 1088-1098.

[4] Zhao, Ying, Jun Gao, and Xuezhi Yang. "A survey of neural network ensembles." 2005 International Conference on Neural Networks and Brain. Vol. 1. IEEE, 2005.

[5] Dietterich, Thomas G. "Ensemble methods in machine learning."*International workshop on multiple*

[6] http://www.cs.waikato.ac.nz/ml/weka

[7] http://archive.ics.uci.edu/ml