© 2015 IJSRSET | Volume 1 | Issue 6 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Improving Classification Accuracy through Ensemble Technique in Data Mining

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

[1]Hemchandracharya North Gujarat University, Gujarat, India

[2]NPCCSM, Kadi SarvaVishwaVidyalaya, Gandhinagar, Gujarat, India

ABSTRACT

Data Mining is the study to get the knowledge from the huge data sources. It is a technology with huge potential to help the corporate ventures focus on the most important information in their data warehouses or database, so that it will help in making business decisions. Decision making with data mining is very much complex task. Ensemble technique is one of the common strategies to improve the accuracy of classifier. In general ensemble learning is an effective technology that combines the predictions from multiple base classifiers. Most commonly used ensemble techniques are bagging and Boosting. Stacking is also one of the techniques, but it is less widely used. In this paper, we are focusing on bagging technique. An experiment is carried out using bagging with different datasets from UCI repository to study the classification accuracy improvement

Keywords: Data Mining; Classification; Ensemble Learning; Bagging;

I. INTRODUCTION

Data mining refers to digging out knowledge from large amounts of data available from different data sources which are accumulated in data warehouse. It is an interdisciplinary field, which covers different areas like data warehousing, statistical methods, database management systems, artificial intelligence, information retrieval, data visualization etc.. Other contributing areas include pattern recognition, spatial data study, signal processing, image databases and many more other application fields, like business, economics, and bioinformatics.[1]

In data mining, classification is one of the tasks which are performed on the given datasets. Accuracy of classification is one of the very much important factors. To improve the classification accuracy various strategies have been identified. Ensemble learning is one of the ways to improve the classification accuracy. Ensemble methods are learning techniques that builds a set of classifiers and then classify new data sets on the basis of their weighted vote of predictions. The original ensemble method is Bayesian averaging, but more recent algorithms include Bagging, boosting etc.[2]. In this paper reviews for these methods have been made and explained why ensembles can often perform better than any base classifier. Combining outputs from multiple classifiers, known as ensemble learning, is one of the standard and most important techniques for improving classification accuracy in machine learning. Out of these, bagging and boosting are the most renowned methods of ensemble learning. In bagging, from the training data, a training set is randomly sampled k times with replacement which produces k training sets with exactly the same size as what we have in original training set. As the original data set is sampled with replacement, it may happen that some training instances are repeated in the new training sets, and it is quite possible that some are not present at all. The obtained sample sets are used to train base classifiers like CART etc. which in turn will give k different predictors. These k different predictors are used to classify the new dataset.

The classification for each data instance is obtained by equal weight voting on all k predictors. Voting gives a significant improvement in classification accuracy and stability. Boosting, on the other hand, induces the ensemble of classifiers by adaptively changing the distribution of the training set based on the accuracy of the previously created classifiers and uses a measure of classifier performance to weight the selection of training examples and the voting.

Various empirical studies, suggest that combining classifiers gives optimal improvements in accuracy if the classifiers are not correlated. It is stated in Ref. [3], the most effective method of achieving such autonomy is by training the members of an ensemble on qualitatively different feature (sub)sets. In other words, attribute partitioning methods are capable of performance superior to data partitioning methods (e.g. bagging and boosting) in ensemble learning. There are a growing number of publications that investigate performance of classifier ensembles trained using attribute

The paper is divided into five parts. In introduction, data mining is briefly explained and ensemble learning technique is discussed. In section 2, literature review is done in which related works done by various authors are elaborated. Here, analysis of classification accuracy on different datasets using ensemble learning is done. In section 3 Bagging process is described and bagging algorithm is discussed. In section 4, Experimental setup and strategy evaluation for accuracy estimation on different datasets using ensemble learning are described. Also result of the experiment is discussed. In Final section, a conclusion and some future directions are highlighted in section.

II. METHODS AND MATERIAL

A. Literature Review

The attraction that this topic exerts on machine learning researchers is based on the premise that ensembles are often much more accurate than the individual classifiers that make them up. Most of the research on classifier ensembles is concerned with generating ensembles by using a single learning base classifier, such as CART. Various classifiers are created by changing the training set (as done in boosting or bagging), changing the input features, changing the output targets or injecting randomness in the learning algorithm. In present stacking method we are not manipulating the training set or manipulating the feature set. Many researches have been done on finding good Meta learner at Meta level. By applying the boosting method, training set will be manipulated and diversity will be increased among the base classifiers so accuracy will be improved. For the large data set training time is very large, so with use of features.

Numerous methods have been suggested for the creation of ensemble of classifiers. As many methods of ensemble creation have been proposed, there is as yet no guarantee of which method is best out of all the methods. So, an active area of research in supervised learning is the study of methods for the construction of good ensembles of classifiers. Mechanisms that are used to build ensemble of classifiers include: (i) using different subsets of training data with a single learning method, (ii) Using different training parameters with a single training method (e.g., using different initial weights for each neural network in an ensemble) and (iii) using different learning methods. [4]

Breiman (1996) made the important observation that in order to make bagging to be more effective, instability (i.e. responsiveness towards the changes in the training data) is a requirement. A committee of classifiers that all agree in all circumstances will give identical performance to any of its members in separation. Reduction in variance process will have no effect if there is no variance. If there is too little data, the gains achieved via a bagged ensemble cannot compensate for the decrease in accuracy of individual models, each of which now considers an even smaller training set. Besides that, if the dataset is enormously large and computation time is not a problem of concern, even a single classifier can be fairly sufficient. Another method that uses different subsets of training data with a single learning method is the boosting approach (Freund and Schapire 1997). Boosting is similar in overall structure to bagging, except that it keeps track of the performance of the learning algorithm and concentrates on instances that have not been correctly learned. Instead of choosing the t training instances randomly using a uniform distribution, it chooses the training instances in such a manner as to favor the instances that have not been precisely learned. After quite a few iterations, the prediction is performed by taking a weighted vote of the predictions of each and every classifier, with the weights are being relative to each and every classifier’s accuracy on its training set.

AdaBoost is a practical version of the boosting approach. Adaboost requires less instability than bagging, because Adaboost can make much larger changes in the training set.[5] A number of studies that compare AdaBoost and bagging suggest that AdaBoost and bagging have quite different operational profiles (Bauer and Kohavi 1999; Quinlan 1996). In general, it appears that bagging is more consistent, increasing the error of the base learner less frequently than does AdaBoost. However, AdaBoost appears to have greater average effect, leading to substantially larger error reductions than bagging on average.

Generally, bagging tends to decrease variance without unduly affecting bias (Breiman 1996; Schapire et al. 1998; Bauer and Kohavi 1999). On the contrary, in empirical studies AdaBoost appears to reduce both bias and variance (Breiman 1996; Schapire et al. 1998; Bauer and Kohavi 1999). Thus, AdaBoost is more effective at reducing bias than bagging, but bagging is more effective than AdaBoost at reducing variance.

The decision on limiting the number of sub-classifiers is important for practical applications. To be competitive, it is important that the algorithms run in reasonable time. Quinlan (1996) used only 10 replications, while Bauer and Kohavi (1999) used 25 replications, Breiman (1997) used 50 and Freund and Schapire (1997) used 100. For both bagging and boosting, much of the reduction in error appears to have occurred after ten to fifteen classifiers. However, Adaboost continues to measurably improve test-set error until around 25 classifiers for decision trees (Opitz and Maclin 1999).

As mentioned in Bauer and Kohavi (1999), the main problem with boosting seems to be robustness to noise. This is expected because noisy instances tend to be misclassified, and the weight will increase for these instances. They presented several cases where the performance of boosted algorithms degraded compared to the original algorithms. On the contrary, they pointed out that bagging improves the accuracy in *all* datasets used in the experimental evaluation.

Thomas [6] carried out experiments which show that in situations where there is little or no classification noise, randomization is competitive with (and perhaps slightly superior to) bagging but not as accurate as boosting. In situations with considerable classification noise, it is found that bagging is much better than boosting.

B. Bagging

Bagging, which is also known as bootstrap aggregating, is a technique that repeatedly samples (with replacement) from a dataset according to a uniform probability distribution. Each bootstrap sample has the same size as the original data. Because the sampling is done with replacement, some instance may appear several times in the same training set, while others may be omitted from the training set.[7] On average, a bootstrap sample S_i contains approximately 63% of the original training data. The bagging procedure is summarized as below.

Bagging Algorithm

```
1. Let k be the number of bootstrap samples
2. for i=1 to k do
3.     Create a bootstrap sample of size X, $S_j$,
4.     Train a base classifier $C_j$ on the bootstrap sample $S_j$
5. end for
6. The class with the maximum number of votes is chosen as the label for test data x.
```

A detail bagging procedure is shown below.

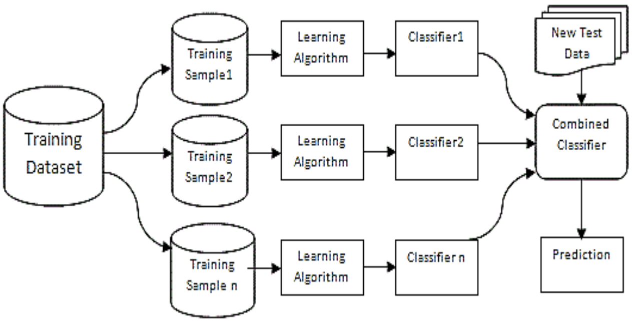

Figure 1: Bagging Illustration

As seen from the above figure, training dataset is divided into multiple test data. The training sets are generated by doing random sampling with replacement. After that each training sample is used with the learning algorithm to generate the classifier. This process is carried out until the last training set it used to train and generate the classifier. So if there are n training subsets then n different classifiers are generated from that n different training set with the learning algorithm. Finally all the classifiers are combined and the new unseen test data is used with the model to generate the prediction.

Research studies reveal that combined classifiers outperform the single classifier. Hence the predicted accuracy is better than the accuracy of the single classifier.

III. RESULTS AND DISCUSSION

Experiment of the algorithm

To carry out the experiment, Weka tool is used. Weka (Waikato Environment for Knowledge Analysis) is a popular machine learning tool written in JAVA. Weka is free open source software available under the GNU General Public License. Firstly, the experiment is carried out on base classifier and then accuracy is measured. After that experiment is carried out on the classifier with bagging. The experiment is carried out using dataset collected from UCI machine repository. Finally results are compared and conclusion is made.
In our experiment, we've taken following datasets from the UCI Machine Learning Repository.

Sr.No	**Dataset Information**		
	Dataset	***Instances***	***Attributes***
1	Iris	150	5
2	Zoo	101	18
3	Vehicle	846	19

The experiment is carried out on RepTree, Decision Sump and J48 classifier. The datasets are chosen and no filter is applied while carrying out the experiment. Firstly experiment is carried out using single base classifier then experiment is carried out using single base classifier with bagging. The experiment is carried out using weak 3.6.12.

Accuracy of the base single classifier and base classifier with bagging is measured which is displayed in below table.

Classifier	**Datasets**		
	Iris	***Zoo***	***Vehicle***
RepTree	94.0	40.59	72.34
RepTree with Bagging	94.67	42.57	75.17
Decision Sump	66.67	60.39	39.95
DecisionSump with Bagging	72.00	61.38	40.07
J48	96.00	92.07	72.45
J48 with Bagging	94.47	93.06	73.64

We can see the result of the classifiers when used alone and when used with bagging. The columnar chart clearly shows the effect of base classifier with bagging.

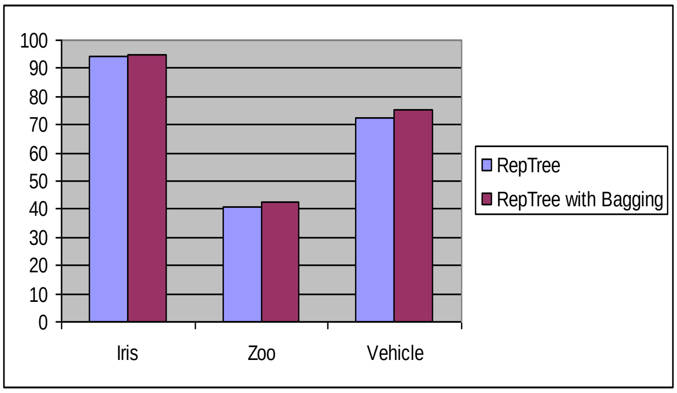

Figure 2 : RepTree and ensemble RepTree comparison

It is clearly seen that when RepTree is used alone with iris, zoo and vehicle dataset, the accuracy of classifier is lesser than when it is used with bagging.

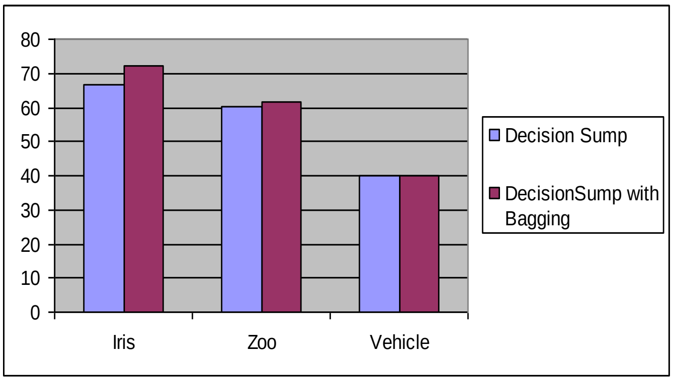

Figure 3 : DecisionSump comparison with ensemble

It is clearly seen that when DecisionSump is used alone with iris, zoo and vehicle dataset, the accuracy of classifier is lesser than when it is used with bagging.

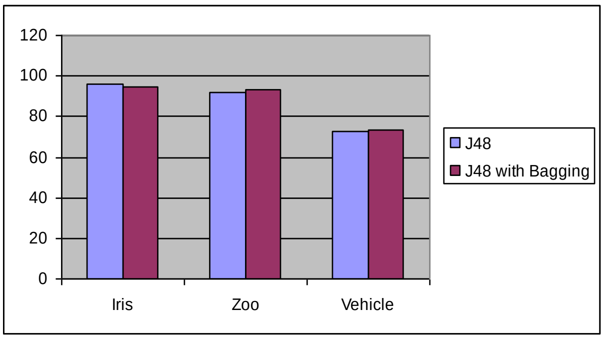

Figure 4 : J48 comparison with ensemble

It is clearly seen that J48 is used alone with zoo and vehicle dataset, the accuracy of classifier is lesser than when it is used with bagging. Here one exception is there that is when same thing is performed with iris dataset the ensemble accuracy goes down. So from above experiment, we can say that bagging improves the classification accuracy.

IV. CONCLUSION

The paper shows the effect of bagging on classification accuracy by using different classifiers. The experiment was carried out using weak 3.6.12 and showed the effect of bagging on various base classifiers. Adding to it, it was observed that for all the three datasets, the classification accuracy increases when we use ensemble learning instead of a single classifier, exception was the iris dataset with J48 classifier. In a nutshell ensemble learning technique of bagging assists in improving the accuracy of classification. Future directions can include the effects of changing the base classifier learner like naive bayes, neural network etc. Further study can be made on combining the heterogeneous classifiers to improve the accuracy.

V. REFERENCES

[1] Han, Jiawei, and Micheline Kamber. "Data mining: concepts and techniques (the Morgan Kaufmann Series in data management systems)." (2000).

[2] Dietterich, Thomas G. "Ensemble methods in machine learning." Multiple classifier systems. Springer Berlin Heidelberg, 2000. 1-15.

[3] D.K. Tumer, J. Ghosh, Classifier combining: analytical results and implications, Working notes from the Workshop ‘Integrating Multiple Learned Models’, 13th National Conference on Artificial Intelligence, August 1996, Portland, Oregon.

[4] Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." (2007): 3-24.

[5] Schapire, Robert E. "The boosting approach to machine learning: An overview." Nonlinear estimation and classification. Springer New York, 2003. 149-171.

[6] Dietterich, Thomas G. "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization." Machine learning 40.2 (2000): 139-157.

[7] Pang-Ning, Tan, Michael Steinbach, and Vipin Kumar. "Introduction to data mining." Library of Congress. 2006.