ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2016.0177

ANALYSE THE PERFORMANCE OF ENSEMBLE CLASSIFIERS USING SAMPLING TECHNIQUES

M. Balamurugan[1] and S. Kannan[2]
Department of Computer Applications, Madurai Kamaraj University, India
E-mail: [1]apkbala107@gmail.com, [2]skannanmku@gmail.com

Abstract

In Ensemble classifiers, the Combination of multiple prediction models of classifiers is important for making progress in a variety of difficult prediction problems. Ensemble of classifiers proved potential in getting higher accuracy compared to single classifier. Even though by the usage ensemble classifiers, still there is in-need to improve its performance. There are many possible ways available to increase the performance of ensemble classifiers. One of the ways is sampling, which plays a major role for improving the quality of ensemble classifier. Since, it helps in reducing the bias in input data set of ensemble. Sampling is the process of extracting the subset of samples from the original dataset. In this research work, analysis is done on sampling techniques for ensemble classifiers. In ensemble classifier, specifically one of the probability based sampling techniques is being always used. Samples are gathered in a process which gives all the individuals in the population of equal chances, such that, sampling bias is removed. In this paper, analyse the performance of ensemble classifiers by using various sampling techniques and list out their drawbacks.

Keywords:

Ensemble of Classifiers, Sampling, Random Forest, Boosting

1. INTRODUCTION

In the classification, dataset is divided into training dataset and testing dataset. Model is constructed using training dataset and then performs the prediction on testing dataset using the model. Here labels of training dataset are known [1].

Ensemble classifier is a combination of more than one prediction model of classifier. Its' prediction is better than the traditional classification algorithm. Generally there are two types of ensemble framework available. They are dependent framework and independent framework. In dependent framework, each classifier is sequentially trained. In independent framework, each classifier is trained in parallel manner. The main difference between the independent and dependent framework is the execution of time and combined fashion. For this reason, we took the algorithms one from dependent framework (Boosting) and another from independent framework (Random Forest). Sample is the input for the ensemble classifier to build the prediction model. Sampling is the process of taking the subset of sample from original dataset [2]. Sample should be a quality one that having the prosperities of original dataset. This sample is generated by applying sampling technique. Sampling plays the major role for improving the accuracy of ensemble classifier. Simple random sampling, systematic sampling and stratified sampling are different sampling techniques used commonly. This paper examines the performance of sampling techniques with two ensemble algorithms boosting and random forest.

The paper is organised into five sections. Section 2 contains description about Ensemble Classifier, its building blocks and Ensemble algorithms. Section 3 describes various Sampling techniques. Section 4 discusses in detail about experimental results. Section 5 gives conclusion.

2. ENSEMBLE CLASSIFIER

Ensemble of classifier means combine the prediction models of more than one classification algorithms into single prediction model to improves the performance of classifiers [1]. Training dataset is the input for the ensemble classifier to perform classification. After classification, combine the outputs of classification algorithms into single prediction model based on combine strategy such as voting, weighting, etc. Finally combined prediction model is applied on testing dataset to perform prediction.

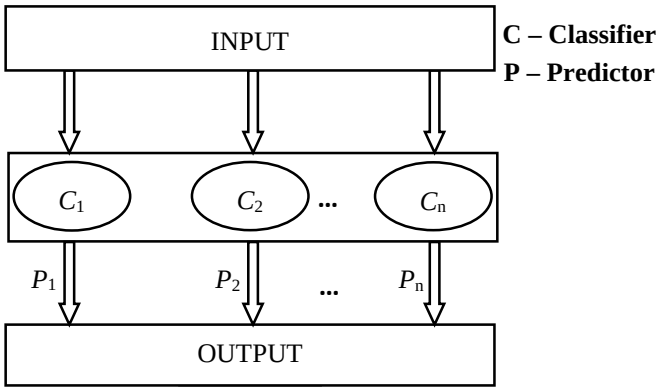

Fig.1. Process of Ensemble classifiers

2.1 ENSEMBLE ALGORITHMS

Building blocks of ensemble algorithms can be classified into dependent ensemble methods and independent ensemble methods. In a dependent method, the output of one classifier is used in the construction of the next classifier. It is for improving the performance of a weak learner (such as classification rules or decision trees). In an independent method, each classifier is built independently and their outputs are combined in some fashion. It is for improving the predictive power of classifiers or decreasing the total execution time.

2.1.1 Boosting:

Boosting is one of the algorithms in dependent ensemble framework [3] [4] [5]. Training dataset is the input for each classifier in boosting to perform classification. Initially each classifier assigned zero weight. After that classifier is trained based on using the training dataset and model trained in the previous pass. That is assigned the larger weight if the classifier is incorrectly perform classification otherwise assigned the smaller weight.

Finally the classifier which is having the smaller weight can be considered as output of boosting. It takes less number of times to build classification model than other ensemble algorithms. Since, the sample is not necessary to divide. The advantage of this ensemble algorithm is to decrease the bias error and helps in building strong predictive models. The disadvantage is because of sometimes over fit on the training data which takes more execution time.

Algorithm:

1. Apply the sampling technique on given data set to extract sample.
2. Set the sample as training data and remaining as testing data.
3. Initially assign the zero weight to three weak classifiers in the boosting.
4. First classifier is trained by using random subset of training data.
5. Next classifier is trained on a subset only half of which is correctly classified by previous classifiers, and the other half is misclassified.
6. Compare the model of previous classifiers and next classifier; update the small weight to classifier which is highly classified in correct manner.
7. Process the step 5 and step 6 until all classifiers are built in the model.
8. Choose the model of which classifier has large weight.
9. Apply the final model on testing data to perform classification.

2.1.2 Random Forest:

Random forest is the independent ensemble framework algorithm. First we have to divide the dataset into subsets by sampling the records of original dataset at random but with replacement. Subset is called as bootstrap sample. Each bootstrap sample is the training data for growing the trees in random forest [6] [7] [8]. Select the subset of feature from the original features that are used to split the node in growing each tree in the forest. The value of subset of feature is held constant during forest growing. Finally combine the votes of all trees and the class which is having the maximum votes is considered as the output of random forest to classify the new instances. There is no need to prune the tree in the forest. It takes more time to train its classifier than other ensemble algorithms. The advantages of random forest are less execution time, no input preparation and perform implicit feature selection. The main drawback of random forest is the model size.

Algorithm:

1. Apply the sampling technique on given data set to extract sample.
2. Set the sample as training data and remaining as testing data.
3. Give the training data as input to random forest to train the model
4. Training data is split into number of samples based on the number of classifiers in random forest.
5. Each classifier is independently built its own classification model by using its sample.
6. Choose the model which one has majority vote by classifiers.
7. Perform classification by applying final model on testing data.

3. SAMPLING TECHNIQUES

In statistics, sampling is concerned with the selection of a subset of individuals from within a statistical population to estimate characteristics of the whole population. The two main advantages of sampling are; (1) the cost is lower and the accuracy and (2) quality of the data can be easily improved [9].

3.1 SIMPLE RANDOM SAMPLING

In this sampling technique, sample is randomly taken from the original dataset. That sample consists of random instances that are extracting from the dataset. All instances of dataset are having equal chance of selection [9] [10]. So it can avoid bias during sample selection. Based on the size of sample, the random numbers of instances are selected. One of the drawbacks is when the dataset contains large number of class labels; random sample may be not perfectly representing the original dataset [11]. For such situations, bias can be generated i.e. there is a possibility of sample belongs to one particular class label only.

3.2 SYSTEMATIC SAMPLING

Systematic sampling is also called interval sampling, which means each element is selected from original dataset based on the fixed interval [9] [11] [12]. First we have randomly selected the beginning element. Then select the fixed range of elements from the beginning element based on the size of sample. Fixed interval can be calculated from the sample size and dataset size.

$$\text{Fixed interval} = \frac{\text{Sample size}}{\text{Dataset size}} \quad (1)$$

It takes less time to generate the sample than Simple random sampling technique. When the elements of the dataset are homogeneous it will not give representative sample.

3.3 STRATIFIED SAMPLING

It overcomes the problem of simple random sampling technique, which fails to extract sample with multiclass from the dataset. In stratified sampling, split the dataset into groups based on the class labels [9]. Then extract the sample from each group by applying the simple random sampling technique on groups. Each group is called as strata. It increases the homogeneity within strata and increases the heterogeneity between strata. Compare to simple random sampling technique with stratified sampling, stratified can generate less number of biases in sample [10] [11] [12] because sample contains the equal proportion of class in the original dataset.

3.3.1 Need of Sample in Ensemble Classifier:

Sampling plays the major role to generate the quality sample for ensemble classifier to improve its accuracy. In Ensemble, sample is named as training dataset. Sampling techniques try to provide sample to ensemble with several features. They are:

1. Sample with greater accuracy.
2. Sample with much reliability.
3. Sample with proper class distribution like that in original dataset.

So quality sample is helpful to build accurate prediction ensemble model instead of using whole dataset.

4. RESULTS AND DISCUSSION

In this research, accuracy is considered as prime. Confusion Matrix is used to finding the accuracy for ensemble classifiers. Prediction results of ensemble method are input for confusion matrix to find the accuracy.

4.1 CONFUSION MATRIX

In predictive analytics, confusion matrix is a two dimension table (actual and predicted) with two rows and two columns that reports the number of false positives, false negatives, true positives, and true negatives. For example, in true positives, true represents actual result is true and positive represents decision is true.

From the table of confusion, accuracy can be calculated as,

$$\text{Accuracy} = \frac{(\text{True Positives} + \text{True Negatives})}{\left(\begin{array}{l}\text{True Positives} + \text{True Negatives} + \\ \text{False Positives} + \text{False Negatives}\end{array}\right)} \quad (2)$$

4.2 DATASETS

There are two datasets taken for performing classification and finding the accuracy. They are given below.

4.2.1 Car Seats:

It is a simulated data set containing the sales of child car seats at 400 different stores. It is an unbalanced dataset, because it has 164 objects out of 400 for one class (Sales is high) and 236 objects out of 400 for another class (Sales is low). This dataset can be used by importing ISLR package in R.

Objective	- predict the sales of car seats
No. of instances	- 400
No. of attributes	- 11

4.2.2 Iris:

Iris dataset is balanced that consists of 50 samples from each of three species of Iris (Setosa, Virginica and Versicolor). It is available in UCI Repository.

Objective	- predict the class of iris plant
No. of instances	- 150
No. of attributes	- 5

4.3 ENSEMBLE CLASSIFIER USING SAMPLING TECHNIQUES

Ensemble classifier need sample to build classification model. Sample should be appropriate to original dataset that is input for the ensemble classifier. Specifically training dataset is the main thing for the ensemble to build model. If the sample is accurate then the good model is built by ensemble algorithm. So we need good sampling techniques to extract sample from the original dataset. In this work, there are three sampling techniques simple random, systematic and stratified which are used to generate samples for ensemble classifier. All sampling techniques are using randomization concept to extract samples from original dataset. Since, randomization will avoid the bias in the sample. Simple random sampling uses the randomization concept to its whole process. Systematic sampling uses the randomization in choosing a starting point. Stratified uses the randomization to extract the samples from each stratum (group).

Generally, a dataset may have either balanced data or unbalanced data. If data set has balanced data then its classes are balanced in the target attribute. So, we may avoid bias in choosing sample. Otherwise classes are unequal in the dataset. There may be possible to bias in choosing sample from the dataset. Setting the sample can be (Training: Testing) 50:50 or 60:40 or 70:30 or 80:20. In this work, two datasets Car seats (Unbalanced dataset) and Iris (Balanced dataset) are used for Boosting and Random forest algorithms. Ensemble classification models are built using 50% of original dataset as training dataset which are selected using the above three sampling techniques.

Table.1. Accuracy of Unbalanced dataset (Car seats) using Ensemble classifier

Sampling	Random Forest	Boosting
Simple random	0.79794	0.82167
Systematic	0.79135	0.8266
Stratified	0.80144	0.82923

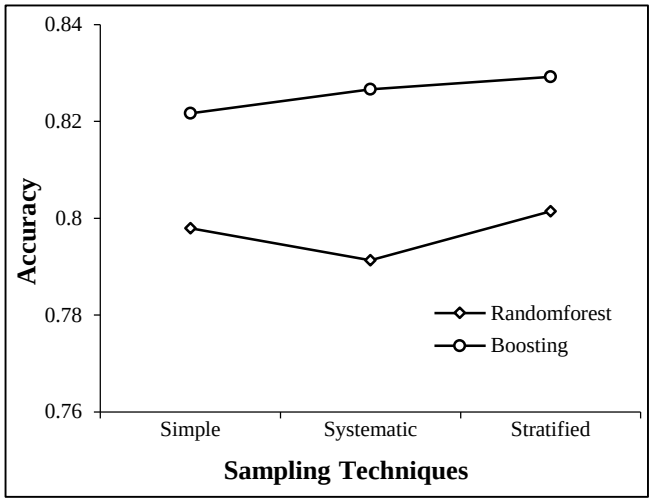

Fig.2. Comparison of Accuracy values for Unbalanced dataset

Then these models are tested using remaining 50% of dataset as test dataset. Each algorithm with each sampling technique is experimented ten times. The Accuracies of these ten trials are averaged. This average is considered as a final result. Accuracies of applying random forest and boosting algorithm using sampling techniques for unbalanced dataset are given in Table.1 and that for balanced dataset are listed out in Table.2. Each sampling technique gives the little variation in sampling, except systematic

sampling. Because of systematic sampling, it can perform better only in ordered dataset.

Table.2. Accuracy of Balanced dataset (Iris) using Ensemble classifier

Sampling	Random Forest	Boosting
Simple random	0.9406	0.94701
Systematic	0.9375	0.94666
Stratified	0.94	0.94665

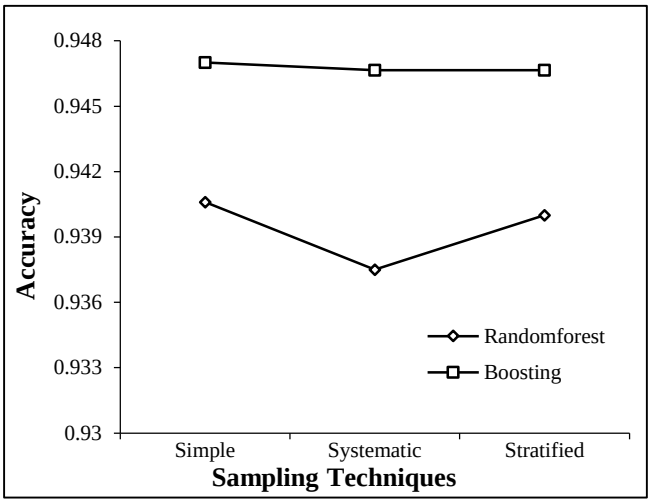

Fig.3. Comparison of Accuracy values for balanced dataset

From the Fig.2, *Stratified Sampling* gives the better accuracy than other sampling techniques with ensemble algorithms. Because in unbalanced dataset, classes are unequal, Stratified samplings split the dataset into groups based on class labels and then extract equal samples from each group by applying simple random sampling. In the result, sample has equal number of classes. So it is possible to avoid bias error in the samples generated. But, there is a possibility to occur more bias error with other sampling techniques.

As shown in the Fig.3, among the Sampling techniques used with ensemble algorithm the *Simple random sampling* performs better. But for the balanced dataset, there is more possibility to reduce the bias error using the simple random sampling techniques compare to other sampling techniques. Because recursively applying the simple random sampling technique on dataset, it choose sample with less bias error.

5. CONCLUSION

An ensemble classifier consists of a set of individually trained classifiers whose predictions are combined in some fashion. Sample is the input for ensemble classification algorithm to build classification model. Sampling techniques are used to extract the samples from original dataset. Quality of sample could improve the accuracy and performance of ensemble methods. In this work, confusion matrix is used to find the accuracy of random forest and boosting ensemble methods for both balanced and unbalanced dataset. From the analysis, stratified sample is suitable for unbalanced dataset and simple random sample is suitable for balanced dataset. Still with sampling techniques there is a possible to introduce bias error in sample. Existing sampling technique fails to extract proper samples for both balanced dataset and unbalanced dataset. So it is necessary to improve a new sampling technique to extract the quality samples for all datasets.

REFERENCES

[1] Lior Rokach, "Ensemble based Classifiers", *Artificial Intelligence Review*, Vol. 33, No. 1, pp. 1-39, 2009.

[2] Lior Rokach, "Decision Forest: Twenty Years of Research", *Information Fusion*, Vol. 27, pp. 111-125, 2016.

[3] Robert E. Schapire, "The Strength of Weak Learn Ability", *Machine Learning*, Vol. 5, pp.197-227, 1990.

[4] Jerome H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine", pp. 1-39, 1999.

[5] Cheng Li. "A Gentle Introduction to Gradient Boosting", Available at:http://www.ccs.neu.edu/home/vip/teach/MLcourse/4_boosting/slides/gradient_boosting.pdf

[6] Tin Kam Ho, "Random Decision Forests", *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Vol. 1, pp. 278-282, 1995.

[7] Vrushali Y Kulkarni and Pradeep K Sinha, "Random Forest Classifiers: A Survey and Future Research Directions", *International Journal of Advanced Computing*, Vol. 36, No. 1, pp. 1144-1153, 2013.

[8] Leo Breiman, "Random Forests", *Machine Learning*, Vol. 45, No. 1, pp. 5-32, 2001.

[9] William G. Cochran, "*Sampling Techniques*", 3rd Edition, John Wiley and Sons, 1977.

[10] Iain A MacDonald, "Comparison of Sampling Techniques on the Performance of Monte Carlo based Sensitivity Analysis", *Proceedings of 11th International Conference on Building Simulation*, pp. 992-999, 2009.

[11] James D. Nelson and Robert C. Ward, "Statistical Considerations and Sampling Techniques for Ground Water Quality Monitoring", *Ground Water*, Vol. 19, No. 6, pp. 617-626, 1981.

[12] Jr. Ronald D. Fricker, "Sampling Methods for Web and E-Mail Surveys", *Fielding: Online Research Methods*, pp. 195-217, 2008.