IJBCS
ISSN: 2147-1428

Original Article

International Journal of Basic and Clinical Studies (IJBCS)
2018; 7(1): 24-35, Yay M.

Ensemble Methods: Boosting, Bagging and Their Applications on Two Different Data Sets

Meral YAY*

*Dr. , Mimar Sinan Fine Arts University, Department of Statistics
meral.yay@msgsu.edu.tr

Abstract

Bagging and boosting are ensemble learning methods that make groups more powerful when they come together and they are frequently used in data analysis. These methods combine a diversity of classifiers using the same learning algorithm for the base-classifiers. Ensemble methods also include logistic regression and linear discriminant analysis, effective on small and medium-sized data sets, and decision/regression trees, support vector machines, and artificial neural networks especially effective on large-sized data sets. However, the existence of a large number of classification methods brings with it the problem of selection. At this point, as an alternative, ensemble algorithms can be used to improve the performance of the selected classification method. The mostly referred ensemble methods are called “Bagging” and "Boosting”. This study aimed to show that how ensemble techniques work on different data sets and the results about bagging and boosting algorithms were evaluated on two different data sets.

Keywords: Boosting, Bagging, Classification, Ensemble.

Introduction

Classification is a set of algorithms that combine different methods on a single model to reduce variance and bias. It is also known as a meta-algorithm because it contains more than one algorithm. It can be grouped under two headings by structurally: sequential and parallel methods. Base learners are created by sequentially and the process continues with boosting the sample with the highest weight each time in the sequential methods. There is dependence between learners in the sequential one. On the other hand

base learners are created by parallel and it is advisable to take the average to reduce the error because there is independence between the learners in the parallel ensemble methods.

The classification can be defined statistically as the determination of the group memberships of the observations in the data set. The fact that classification is one of the main questions of scientific research has led to the development of many techniques for classifying data of different disciplines. In particular, artificial neural networks, decision tree algorithms, Bayesian classification algorithms and support vector machines can be indicated as examples. The classification methods can be classify hypothetically according to their parametric or non-parametric nature. Commonly used parametric techniques are linear regression analysis, generalized linear regression, logistic regression and separation analysis; it is necessary to provide assumptions about the distribution of data. On the non-parametric classification methods do not require this assumption. Because of this feature they can be applied in a much wider area statistically.

One of the main problems of scientific research is that the observations in the dataset can be classified correctly. This has led to the development of many techniques for classifying data of different disciplines. Especially when the dataset is multidimensional and the training set is smaller, it is difficult to create a good classification rule. Furthermore, a classification based on small-scale training clusters will have biased and large variance. As a result, the classifier is weak and has poor performance. In addition, if minor changes in the training set are caused by major changes in the classification, it can be said that the performance of the classifier is not good and that this has been caused by different factors. For example; the classifier's performance is adversely affected if the assumption is made incorrectly about the model while creating the classifier. As a result, it is possible to call such a classifier "weak classifier". However, researchers can use various means to improve the performance of the classifier and produce different

IJBCS
ISSN: 2147-1428

Original Article

International Journal of Basic and Clinical Studies (IJBCS)
2018; 7(1): 24-35, Yay M.

solutions. One of these solutions is to create multiple weak classifiers instead of a single one, and to combine these weak classifiers with a strong decision rule. Among these methods known as merging algorithms, the most common ones are bagging and boosting algorithms based on repetitive sampling methods.

In bagging and boosting algorithms, models based on the algorithm called base classifier are used (6). The common feature of these algorithms is that they both make the majority vote and bring similar models together. These two algorithms show similarities as well as fundamental differences. For example, in the bagging algorithm, while each model is constructed to be independent of the other, each model in the boosting algorithm is affected by the performance of the model first established. Also, while the bagging algorithm gives equal weight to all models, In the boosting algorithm, weighting is related to the performance of the model. It is possible to find in the literature examples of how the boosting algorithms provide better results in problem-free data sets (7).

Materials and Methods

Boosting Ensemble Method

In the boosting algorithm, the weak classifiers are trained hierarchically to learn harder and harder parts of a classification problem (10) . The Boosting algorithm developed by Robert Schapire and Yoav Freund (1999) is used to combine weak classifiers to obtain a classification rule. Freund (1995) proposed a "boost by majority" variation which combined many weak learners simultaneously and improved the performance of the simple boosting algorithm of Schapire (5) . Boosting is a general method for improving the accuracy of any given learning algorithm (4). At each step of the boosting algorithm, the training set is re-weighted in such a way that the incorrectly classified observations in the previous step have greater weight. Thus, the researcher has the opportunity to maximize the gaps between observations in the training set.

In the boosting algorithm, the training set is denoted by $X = \left(X_1, X_2, ..., X_n \right)$ and its

IJBCS
ISSN: 2147-1428

Original Article

International Journal of Basic and Clinical Studies (IJBCS)
2018; 7(1): 24-35, Yay M.

weighted form is denoted by $X^* = \left(w_1^b X_1, w_2^b X_2, ..., w_n^b X_n \right)$. Using these two forms of X, $C^b(x)$ is obtained. It should not be overlooked that the algorithm is $b = 1, 2, ..., B$ at all stages and the operations are repeated B times. The estimated values of the error probabilities that are processed in the misclassification are obtained as follows:

$$err_b = \frac{1}{n} \sum_{i=1}^{n} w_i^b \xi_i^b \qquad (1)$$

If X_i is correctly classified ξ_i^b is equal to "1"; if it is incorrectly classified ξ_i^b is equal to "0" (9). Error probabilities are used to obtain c_b weights:

$$c_b = \frac{1}{2} \log \left(\frac{1 - err_b}{err_b} \right) \qquad (2)$$

If the error probabilities are between $0 < err_b < 0.5$; the weights can be obtained as in Eq 3.

$$w_i^{b+1} = w_i^b \exp\left(c_b \xi_i^b \right), \; i = 1, .., n \qquad (3)$$

The total weight is adjusted to be equal to "n" as follows:

$$\sum_{i=1}^{n} w_i^{b+1} = n \qquad (4)$$

If the total weight is not equal to "n", the algorithm is restarted by adjusting the weight so that all weights are $w_i^b = 1, \; i = 1, 2, ..., n$. In the final stage a majority order is made according to c_b weights to determine the rule of the decision and to combine $C^b(x)$ the classifiers. The decision rule is defined as follows (4):

$$\beta(x) = \arg\max_{y \in (-1,1)} \sum_b c_b \delta_{\text{sgn}\left(C^b(x)\right), y} \qquad (5)$$

The Boosting algorithm is often referred to as "AdaBoost (Adaptive Boosting)" in practice. Adaboost is an algorithm that is often used to combine weak classifiers in order to obtain the classifier with the best performance in combination in recent years (2). In the case of binary classification, the best-known boosting algorithm is the Adaboost algorithm. In the first step of the algorithm, the weights $w_i^{[0]} = 1/n \; ; \; i = 1, 2, ..., n$ are determined for each sampling unit and the number of iterations at the beginning is "0". The process iterates

IJBCS
ISSN: 2147-1428

Original Article

International Journal of Basic and Clinical Studies (IJBCS)
2018; 7(1): 24-35, Yay M.

by increasing the number of iterations (m) in every step. In the second step, the classifier $\hat{g}^{[m]}$ is expanded to the entire data set with the weights given below. In the third step, the misclassification rate is obtained as follows:

$$err^{[m]} = \sum_{i=1}^{n} w_i^{(m-1)} I\left(Y_i \neq \hat{g}^{[m]}(X_i)\right) / \sum_{i=1}^{n} w_i^{[m-1]} \quad (6)$$

For the misclassification rate $\alpha^{[m]} = \log\left(\frac{1-err^{[m]}}{err^{[m]}}\right)$ and $\tilde{w}_i = w_i^{[m-1]} \exp\left(\alpha^{[m]} I\left(Y_i \neq \hat{g}^{[m]}(X_i)\right)\right)$ are assumed as in the equations and the weightining is obtained as in Eq 7. :

$$w_i^{[m]} = \tilde{w}_i / \sum_{j=1}^{n} \tilde{w}_j \quad (7)$$

The second and third steps continue until $m = m_{stop}$. In the last step, an ensembled classifier is obtained with the help of weighted majority vote and expressed (3) as in Eq 8. :

$$\hat{f}_{Adaboost}(x) = \arg\max_{y \in \{0,1\}} \sum_{m=1}^{m_{stop}} \alpha^{[m]} I\left(\hat{g}^{[m]}(x) = y\right) \quad (8)$$

Bagging Ensemble Method

The bagging (bootstrap aggregating) introduced by Leo Breiman (1996) is a method that can be used to show the benefits of bootstrap and ensemble approaches. In the algorithm the multiple bootstrap training clusters are created for a given training cluster. The basic model is called the learning algorithm. It is possible to develop accurate classification models by combining randomly generated training clusters. It is aimed to reduce the variance related to the classification by making equal weight assignments to each of the models developed with the bagging algorithm (1).

The bagging algorithm is configured with the help of a large number of iteratively created bootstrap samples (11). Different bootstrap samples are classified by voting method. In the algorithm, firstly a training set $X = (X_1, X_2, ..., X_n)$ is considered. A bootstrap repeat sample is drawn from this data set by random sampling with replacement method

IJBCS
ISSN: 2147-1428

Original Article

International Journal of Basic and Clinical Studies (IJBCS)
2018; 7(1): 24-35, Yay M.

and it can be expressed by $X^b = \left(X_1^b, X_2^b, ..., X_n^b \right)$. If the misleading observations in the bootstrap training set are as low as possible, it means that the performance of the classifier to be created increases. In the next step in the algorithm, a classifier $C^b(x)$ expressed as a bootstrap sample X^b is created and all these operations are repeated "B" times. Generated classifiers are combined with a simple majority rule to determine the final decision rule. When the number of bootstrap samples is expressed as $b = 1, 2, ..., B$, it is obvious that the number of classifiers will be B as well. The final decision rule is expressed as follows:

$$\beta(x) = \arg\max_{y \in (-1,1)} \sum_b \delta_{\text{sgn}\left(C^b(x)\right), y} \quad (9)$$

There are bagging algorithms that work with smaller training sets and use different strategies. In these strategies, different algorithms are used which are composed of bagging and cross-validation algorithms (8).

Results

In this study two different data sets are used to application of the ensemble methods:
(1) birth
(2) obesity.

The first set of data, called "birth", consists of eight independent variables and one binary dependent variable. Dependent variable Y expresses the form of birth, with "1" normal and "2" cesarean birth. Independent variables affecting the form of the birth; mother's age, mother's height (cm), mother's weight (gr), placental thickness (mm), baby's weight (gr), baby's gender (1 : female; 2 : male), pregnancy hypertension (1: present; 2: not present) and mother's smoking habits (1: present; 2: not present). The data set consisting of 334 observations was divided into 80% learning and 20% validity set, and the decision tree was obtained as follows.

IJBCS
ISSN: 2147-1428

Original Article

International Journal of Basic and Clinical Studies (IJBCS)
2018; 7(1): 24-35, Yay M.

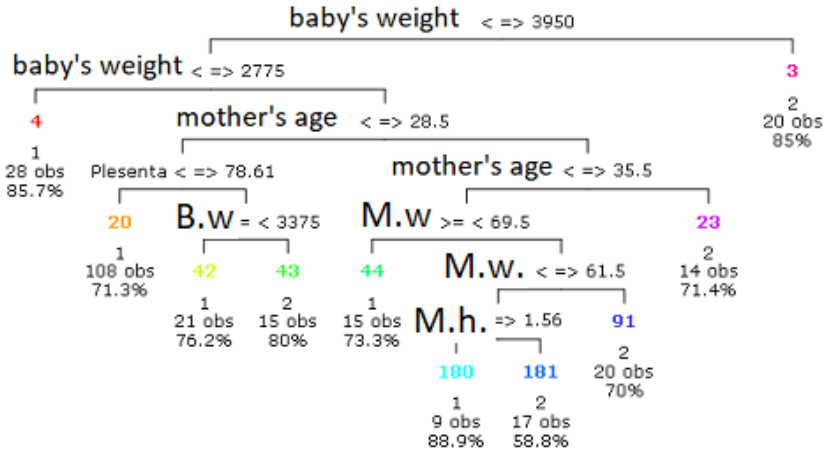

Figure 1. Decision Tree for the "Birth" Data Set

It is expected that the classification error of the technique used in classification problems is not very high. It is seen that the decision error generated on the "birth" dataset corresponds to the expected state, and the classification error is relatively low. However, in practice, depending on the data set, decision trees may not reach the expected classification performance. Especially, it is seen that the decision tree does not capture the desired classification success if the number of variables in the data set is too large, if there are dependent variables, or if the number of observations is small. Ensemble algorithms introduced at this point are used to increase the success of weak decision trees. The most commonly used Bagging and Boosting algorithms have been applied to the "Birth" data set to improve the classification success achieved by the decision tree. The applications were implemented using the R project software.

IJBCS
ISSN: 2147-1428

Original Article

International Journal of Basic and Clinical Studies (IJBCS)
2018; 7(1): 24-35, Yay M.

Table 1. Misclassification Ratios Using Decision Tree for Learning and Validity Sets

		Learning Set		Validity Set	
		Birth		Birth	
		Normal	Cesarean	Normal	Cesarean
Normal		136	23	29	8
Cesarean		45	63	15	14
Misclassification Error Rate		0,2546		0,3484	

As a result of classification with decision tree, misclassification ratios were obtained as 0.2546 for learning set and 0.3484 for validity set. Although these ratios may seem low, it may be possible to improve by using different combining algorithms, that is to reduce the ratios. For this purpose the misclassification ratios of the boosting and bagging algorithms are summarized in Table 2 and Table 3.

Table 2. Boosting Misclassification Rates for "Birth" Data Set

		Learning Set		Validity Set	
		Birth		Birth	
		Normal	Cesarean	Normal	Cesarean
Normal		148	11	33	4
Cesarean		34	74	10	19
Misclassification Error Rate		0,1685		0,212	

IJBCS
ISSN: 2147-1428

Original Article

International Journal of Basic and Clinical Studies (IJBCS)
2018; 7(1): 24-35, Yay M.

Table 3. Bagging Misclassification Rates for "Birth" Data Set

	Learning Set		Validity Set	
	Birth		Birth	
	Normal	Cesarean	Normal	Cesarean
Normal	156	3	36	1
Cesarean	90	18	12	17
Misclassification Error Rate	0,3483		0,1969	

When Table 2. and Table 3. are examined, it is seen that the boosting algorithm is more successful in the learning set than the classification obtained by the decision tree. However, this result is not available for the bagging algorithm.

The other set of data used by the application is "obesity" and dependent variable Y expresses the obesity, with "0" non-obese and "1" obese. The independent variables in the second data set are age groups (1 : "20-29"; 2 : "30-39"; 3 : "40-49"; 4 : "50-59; 5 : "60-69", place of residence (0: country side; 1: urban area) body mass index (kg/m2), systolic blood pressure (mm/Hg), diastolic blood pressure (mm/Hg), total cholesterol (mg / dL); Hdl (mg / dL), triglyceride (mg), waist circumference (cm) and Ldl (mg/dL). "Obesity" dataset consists of 4249 observations. The data are divided into the learning and the validity set, which are classified in the decision tree.

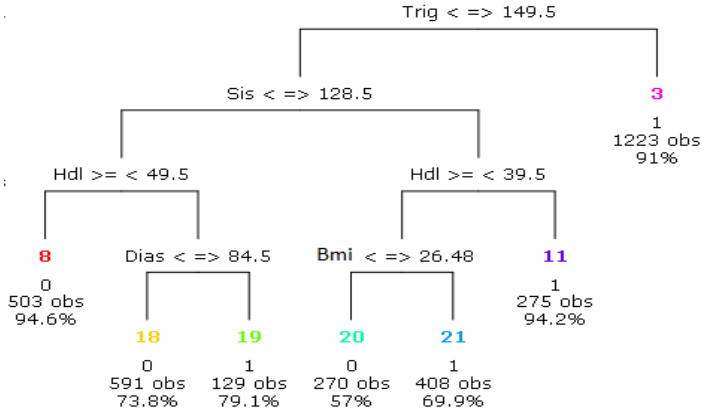

Figure 2. Decision Tree for the "Obesity" Data Set

The classification made by the decision tree for the "obesity" dataset. It is quite successful with a false classification rate of 0.1688 as in Table 4. However, it is possible to further reduce the misclassification rate.

The results obtained with the boosting and bagging methods used to reduce the false classification rate are summarized as in the Table 4 and Table 5.

Table 4. Misclassification Ratios Using Decision Tree for Obesity Data Set

		Learning Set		Validity Set	
		Obese	Non-obese	Obese	Non-obese
	Obese	1066	276	259	68
	Non-obese	298	1759	65	457
Misclassification Error Rate		0,1688		0,1566	

Table 5. Boosting Misclassification Rates for "Obesity" Data Set

		Learning Set		Validity Set	
		Obese	Non-obese	Obese	Non-obese
	Obese	1112	230	293	34
	Non-obese	220	1837	26	496
Misclassification Error Rate		0,1324		0,071	

IJBCS
ISSN: 2147-1428

Original Article
International Journal of Basic and Clinical Studies (IJBCS)
2018; 7(1): 24-35, Yay M.

Table 6. Bagging Misclassification Rates for "Obesity" Data Set

		Learning Set		Validity Set	
		Obese	Non-obese	Obese	Non-obese
	Obese	933	409	230	97
	Non-obese	281	1776	75	447
Misclassification Error Rate		0,203		0,202	

The misclassification rate obtained after the boosting algorithm is more successful than the decision tree.
As in the previous data set, the classification performance achieved by the boosting algorithm is lower than the bagging algorithm.

Discussion

The data sizes used in scientific research are increasing day by day. Accordingly, the analyzes made are in a continuous development process depending on the data size and structures. One of the biggest problems encountered during analysis is to correctly classify the data. It is observed that classical statistical methods are insufficient to solve classification problems in very large data sets. Decision trees are one of the most frequently used techniques for classifying large data sets. It is possible to list the reasons: does not require assumptions, can be apply on different scale variables and to perform much faster than other techniques. However, it cannot be argued that the decision trees are very successful in the data sets with a low number of observations. Various ensemble methods have been developed to remove the problem of classification. The most common ones of these methods are boosting and bagging.

In this study, it was researched whether it is possible to improve the classification ratios obtained after applying the decision tree on two different data sets at different sizes. For this, boosting and bagging algorithms for both sets of data were performed using the R programming language. The results obtained show that the boosting algorithm can

improve the decision tree for both data sets. This success has not been achieved for the bagging algorithm. The boosting algorithm in decision trees gives better results in practice. (1). In this study, boosting algorithm has been shown to perform a more successful classification than the bagging algorithm.

References

1. Aljamaan H I, Elish M O, An Emprical Study of Bagging and Boosting Ensembles for Identifying Faulty Classes in Object-Orianted Software, IEEE Symposium on Computational Intelligence and Data Mining, 2009; 187-194.

2. Bauer E, Kohavi R, An Emprical Comparison of Voting Classification Algorithms: Bagging, Boosting and Variants, Machine Learning, 1999; 36:105-139.

3. Bühlmann P, Hothorn T, Boosting Algotihms:Regularization, Prediction and Model Fitting, Statistical Science, 2007; 22: 477-505.

4. Freund Y, Schapire R, , A Short Introduction to Boosting, Journal of Japanese Society for Artificial Intelligence, 1999; 14(5):771-780.

5. Friedman J, Hastie T, Tibshirani R, Additive Logistic Regression: A Statistical View of Boosting , The Annals of Statistics, 2000: 28(2): 337-407.

6. Hamza M, Larocque D,An Emprical Comparison Of Ensembles Methods Based On Classification Trees, Journal of Statistical Computation and Simulation, 2005; 75: 629-643.

7. Kotsiantis, S B, Pintelas P E, Combining Bagging and Boosting, International Journal of Computational Intelligence, 2004; 1:324-333.

8. Machova K, Barcak F, Bednar P, A Bagging Method Using Decision Trees in Role of Base Classifier, Acta Polytechnica Hungarica, 2006; 3:121-132.

9. Skurichina M, Duin P W R, Bagging, Boosting and The Random Subspace Method for Linear Classifier, Pattren Analysis and Applications, 2002; 5:121-135.

10. Tzenk Y C, Chen KS, Using Modified Bagging and Boosting Algorithms in Multiple Classifiers System for Remote Sensing Image Classification, Journal of Photogrammetry and Remote Sensing, 2007;12:241-256.

11. Zang G, Fang B, LogitBoost classifier for discriminating thermophilic and mesophilic proteins, Journal of Biotechnology, (2007); 127: 417–424