International Journal of Internet, Broadcasting and Communication Vol.11 No.1 47-53 (2019)
http://dx.doi.org/10.7236/IJIBC.2019.11.1.47

IJIBC 19-1-6

Ensemble Methods Applied to Classification Problem

ByungJoo Kim

Department of Computer Engineering, Youngsan University, Korea
bjkim@ysu.ac.kr

Abstract

The idea of ensemble learning is to train multiple models, each with the objective to predict or classify a set of results. Most of the errors from a model's learning are from three main factors: variance, noise, and bias. By using ensemble methods, we're able to increase the stability of the final model and reduce the errors mentioned previously. By combining many models, we're able to reduce the variance, even when they are individually not great. In this paper we propose an ensemble model and applied it to classification problem. In iris, Pima indian diabeit and semiconductor fault detection problem, proposed model classifies well compared to traditional single classifier that is logistic regression, SVM and random forest .

Keywords: *Ensemble model, Decision trees, Bagging, Overfitting*

1. INTRODUCTION

The idea of ensemble learning is to train multiple models, each with the objective to predict or classify a set of results. Most of the errors from a model's learning are from three main factors: variance, noise, and bias. By using ensemble methods, we're able to increase the stability of the final model and reduce the errors mentioned previously. By combining many models, we're able to reduce the variance, even when they are individually not great, as we won't suffer from random errors from a single source. The main principle behind ensemble modelling is to group weak learners together to form one strong learner. bagging, is shorthand for the combination of bootstrapping and aggregating. Bootstrapping is a method to help decrease the variance of the classifier and reduce overfitting, by resampling data from the training set with the same cardinality as the original set. The model created should be less overfitted than a single individual model.

A high variance for a model is not good, suggesting its performance is sensitive to the training data provided. So, even if more the training data is provided, the model may still perform poorly. And, may not even reduce the variance of model. In this paper we propose an ensemble method that combines a single classifier. Paper is organized as follows. In Section 2 we will briefly explain the bagging method and decision tree algorithm. Experimental results to evaluate the performance of proposed method is shown in Section 3. Discussion of proposed method and future work is described in Section 4.

Manuscript Received: Jan. 3, 2019 / Revised: Jan. 10, 2019 / Accepted: Jan. 17, 2019
Corresponding Author: bjkim@ysu.ac.kr
Tel: +82-55-380-9447, Fax: +82-55-380-9447
Department of Computer Engineering, Youngsan University, Korea

2. BAGGING

Bagging is an effective method when you have limited data, and by using samples you're able to get an estimate by aggregating the scores over many samples. The simplest approach with bagging is to use a couple of small subsamples and bag them, if the ensemble accuracy is much higher than the base models, it's working; if not, use larger subsamples.

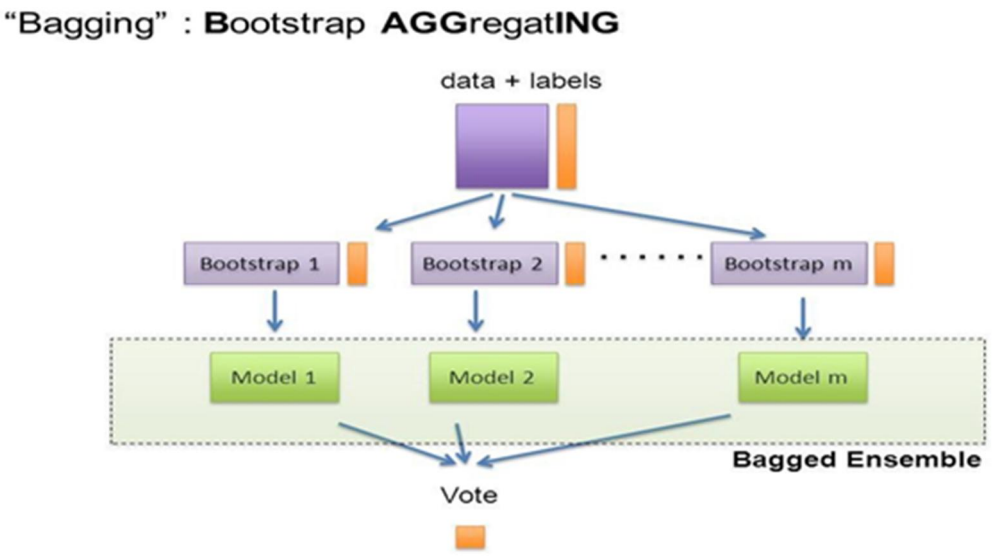

Figure 1. Structure of Bagging

Once the bagging is done, and all the models have been created on different data, a weighted average is then used to determine the final score.

2.1 DECISION TREES

A decision tree is a flowchart-like tree structure, where each internal node represents a test on an attribute, each branch represents an outcome of the test, class label is represented by each leaf node. Given a tuple X, the attribute values of the tuple are tested against the decision tree. A path is traced from the root to a leaf node which holds the class prediction for the tuple. It is easy to convert decision trees into classification rules. Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value[1]. It is one of the predictive modelling approaches used in statistics, data mining and machine learning[2]. Tree models where the target variable can take a finite set of values are called classification trees, in this tree structure, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision tree can be constructed relatively fast compared to other methods of classification. SQL statements can be constructed from tree that can be used to access databases efficiently. Decision tree classifiers obtain similar or better accuracy when compared with other classification methods[3]. A number of data mining techniques have already been done on educational data mining to improve the performance of students like regression, genetic algorithm, k-means clustering, associate rules, prediction etc. Data mining techniques can be used in educational field to enhance our understanding of learning process to focus on identifying, extracting and evaluating variables related to the learning process of students. Classification is one of the most frequently. We will briefly explain the famous decision tree algorithms in section 2.2.

2.2 DECISION TREES ALGORITHM

2.2.1 ID3 ALGORITHM

Iterative Dichotomiser 3 is a simple decision tree learning algorithm introduced in 1986 by Quinlan Ross[4]. It is serially implemented and based on Hunt‟s algorithm[5]. The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. In the decision tree method, information gain approach is generally used to determine suitable property for each node of a generated decision tree. Therefore, we can select the attribute with the highest information gain (entropy reduction in the level of maximum) as the test attribute of current node. In this way, the information needed to classify the training sample subset obtained from later on partitioning will be the smallest. So, the use of this property for partitioning the sample set contained in current node will make the mixture degree of different types for all generated sample subsets reduced to a minimum. Hence, the use of an information theory approach will effectively reduce the required dividing number of object classification.

2.2.2 C4.5 ALGORITHM

C4.5[6] is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification and for this reason C4.5 is often referred to as a statistical classifier. As splitting criteria, C4.5 algorithm uses information gain. It can accept data with categorical or numerical values. Threshold is generated to handle continuous values and then divide attributes with values above the threshold and values equal to or below the threshold. C4.5 algorithm can easily handle missing values, as missing attribute values are not utilized in gain calculations by C4.5. The algorithm C4.5 has following advantages. Handling each attribute with different cost. Handling training data C4.5 allows attribute missing. Missing attribute values are not used in gain and entropy calculations. Handling both continuous and discrete attributes to handle continuous attributes, C4.5 creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it. Pruning trees after creation, C4.5 goes back through the tree once it has been created, and attempts to remove branches that are not needed, by replacing them with leaf nodes.

2.2.3 CART ALGORITHM

CART stands for classification and regression Trees. It was introduced by Breiman in 1984[7]. It builds both classifications and regression trees. The classification tree construction by CART is based on binary splitting of the attributes. CART also based on Hunt's algorithm and can be implemented serially. Gini index is used as splitting measure in selecting the splitting attribute. CART is different from other Hunt's based algorithm because it is also use for regression analysis with the help of the regression trees. The regression analysis feature is used in forecasting a dependent variable given a set of predictor variables over a given period of time. CARTS supports continuous and nominal attribute data and have average speed of processing.

2.2.3.1 CART ADVANTAGE

The advantage of CART algorithm is as follows. First it is non parametric, second it performs automatically variable selection, third it uses any combination of continuous or discrete variables and it is very nice feature i.e ability to automatically bin massively categorical variables into a few categories. Finally it establishes interactions among variables.

Table 1: Comparisons between different decision tree algorithms

Features	ID3	C4.5	CART
Type of data	Categorical	Continuous and Categorical	continuous and nominal attributes data
Speed	Low	Faster than ID3	Average
Boosting	Not supported	Not supported	Supported
Pruning	No	Pre-pruning	Post pruning
Missing Values	Can't deal with	Can't deal with	Can deal with
Formula	Use information entropy and information Gain	Use split info and gain ratio	Use Gini diversity index

3. EXPERIMENT

3.1 IRIS DATA

To evaluate the performance of accuracy on ensemble model we take Iris data[8]. This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are not linearly separable from each other. Attribute information is consists of 5, that is 1. sepal length 2. sepal width 3. petal length 4. petal width 5. Class. Class is divided into 3 categories such as Iris Setosa, Iris Versicolour and Iris Virginica. In this experiment number of trees are 100. Mean and overall accuracy is 0.9409 and 95.5555 respectively.

Table 2. Mean and overall accuracy of ensemble model

Model	0	1	2	3	4	5	6	7	8	9
Accuracy	1.0	1.0	0.90	1.0	1.0	0.8	1.0	1.0	0.7	0.94

We compared the proposed model to other traditional single classifier i.e logistic regression, random forest and support vector machine(SVM). In SVM RBF kernel is taken.

Table 3. Performance comparison of ensemble method and other classifier

Classifier	Accuracy
Logistic Regression	94.66
Random Forest	94.02
SVM(RBF kernel)	92.0
Proposed Ensemble model	95.55

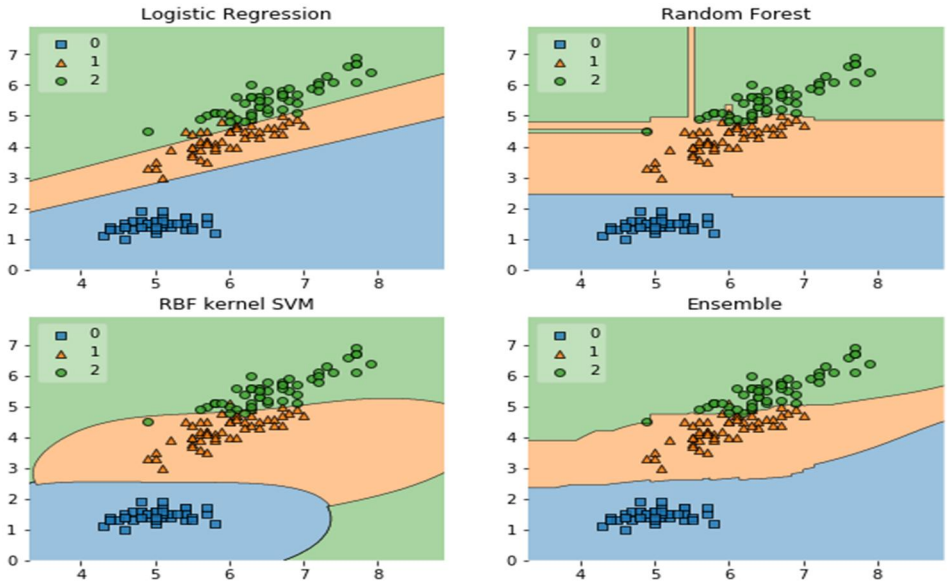

Figure 2. Comparison of classification performance of single classifier and ensemble model

3.2 WINE DATA

We will be experiment the wine dataset that is deposited on the UCI machine learning repository[9].These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. The attributes are (1) alcohol, (2) malic acid, (3) ash, (4) alcalinity of ash, (5) magnesium, (6) total phenols, (7) flavanoids, (8) nonflavanoid phenols, (9) proanthocyanins, (10)color intensity, (11)hue, (12)OD280/OD315 of diluted wines and finally 13)proline. Number of instances of each wine class is class1 is 59, class2 is71 and class3 is 48.

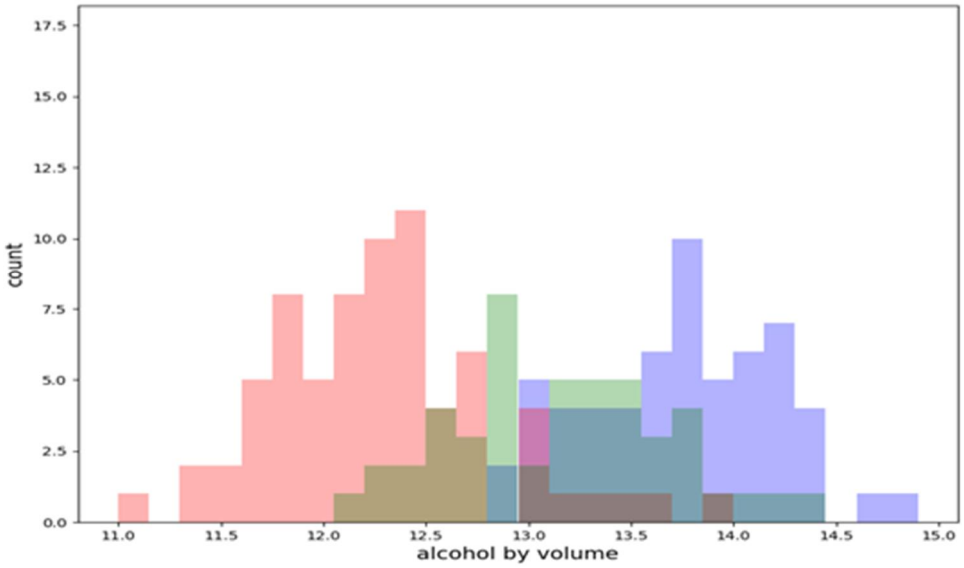

Figure 3. Wine data set - Distribution of alcohol contents

Table 4. Performance comparison of ensemble method and other classifier

Classifier	Accuracy
Logistic Regression	95.26
Random Forest	96.05
SVM(RBF kernel)	96.43
Proposed Ensemble model	98.93

3.3 SEMICONDUCTOR DATA

We extend our experiment on real world classification problem. Fault detection in semiconductor manufacturing process is a hot research topic as more and more sensor data are being collected throughout the industrial process. we take 1300 data and each data has 18 variables. Detailed variables attributes are shown in Table 1. Among data, the number of normal is 800and rest of them is abnormal. Data were collected and recorded at one second intervals during the etch for each of these sensors. Since our primary concern in this work was to detect faults occurring from one wafer to the next, we took the average value of each variable during the etch process for each wafer, resulting in a 1x18array of values for each wafer.

Table 5. Tool-state variables used for process monitoring

1	TCP Top Power	**10**	RF Impedance
2	TCP Tune	**11**	RF Power
3	TCP Load	**12**	TCP Reflected Power
4	TCP Phase Error	**13**	RF Bottom Reflected Power
5	TCP Impedance	**14**	Pressure
6	RF Bottom Power	**15**	BCl3 Flow
7	RF Tune	**16**	Cl2 Flow
8	RF Load	**17**	He Pressure
9	RF Phase Error	**18**	Vat Value

4. COMPARISON WITH SVM

Recently SVM has been a powerful methodology for solving problems in nonlinear classification. To evaluate the classification accuracy of the proposed system it is desirable to compare with SVM. Generally a disadvantage of the incremental method is its accuracy compared to the batch method even though it has the advantage of memory efficiency. According to Table 2 and Table 3 we can see that the proposed method has better classification performance compared to batch SVM. Through this result we can show that the proposed classifier has remarkable classification accuracy, although it is worked in an incremental way.

Table 6. Performance comparison of ensemble method and other classifier

	Training	Generalization
SVM(RBF kernel)	100%	95.34%
Random Forest	100%	95.01
Logistic Regression	100%	93.74
Ensemble method	100%	96.74%

5. CONCLUSION

The idea of ensemble learning is to train multiple models, each with the objective to predict or classify a set of results. Most of the errors from a model's learning are from three main factors: variance, noise, and bias. By using ensemble methods, we're able to increase the stability of the final model and reduce the errors mentioned previously. By combining many models, we're able to reduce the variance, even when they are individually not great. In this paper we propose an ensemble model and applied it to classification problem. In iris, Pima indian diabeit and semiconductor fault detection problem, proposed model classifies well compared to traditional single classifier that is logistic regression, SVM and random forest .

ACKNOWLEDGMENT

This work was supported by a 2018 research grant from Youngsan University, Republic of Korea.

REFERENCES

[1] L. Hyafil and L. Rivest R., "Constructing optimal binary decision trees is NP-complete," Information Processing Letters, Vol. 5, No.1 pp. 15-17, 1976.
[2] H. Zantema, and H. L. Bodlaender, "Finding Small Equivalent Decision Trees is Hard," International Journal of Foundations of Computer Science, Vol. 11, No. 2, pp. 343-354, 2000.
[3] G.E.Naumov, "NP-completeness of problems of construction of optimal decision trees," Soviet Physics Vol. 36, No. 4, pp.270-271, 1991.
[4] J.R. Quinlan, "Induction of decision trees, Machine Learning," Vol. 1, pp.81-106, 1986.
[5] Available : https://www.navodayaengg.in
[6] J. Quinlan, C4.5 1st Edition Programs for Machine Learning, Morgan Kaufmann, 2014.
[7] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees, Wadsworth Int. Group, 1984.
[8] Available : https://archive.ics.uci.edu/ml/datasets/iris
[9] Available : http://archive.ics.uci.edu/ml/datasets/Wine
[10] B.J. Kim, "Model Selection in Artificial Neural Networks," International Journal of Advanced smart Convergence, Vol. 7 No.4, pp.66-74,2018.
DOI: https://doi.org/10.7236/IJASC.2018.7.4