Survey of Ensemble Classifiers

AJAY VIJAYAKUMARAN NAIR
University of North Carolina at Charlotte

An interesting class of algorithms in the Machine Learning domain are the ensemble classification algorithms that try to improve the predictive power of weak classifiers through averaging or aggregation. The goal of this survey is to evaluate the building blocks of a typical ensemble classifier. An ensemble classifier, in most cases, performs far superior than its underlying individual base classifier. The survey begins by exploring some of the typical base classifiers that are used in an ensemble. Through the evaluation of ensemble techniques such as Bagging, Boosting and Bayes Optimal Classifier, the intuition and rationale of why an ensemble classifier outperforms a weak individual classifier is presented. An analysis of Adaboost learning classifier is also presented to review the mechanics of how an ensemble classification scheme achieves the boosted predictive power.

Additional Key Words and Phrases: Classifier, ensemble, bagging, boosting, bayes optimal classifier, adaboost

1. INTRODUCTION

A typical individual base classifier, in a supervised classification scheme, tries to reduce the in sample error there by expecting the in sample error to closely reflect the out of sample error. Trying to model the hypothesis purely based on the training data set to fit often results in a complex model that reduces the in sample error but not necessarily the out of sample error. The effectiveness of a classifier is usually measured in terms of the error which is a combination of primarily two factors - Bias and Variance. For real world problems, it is not practical to reduce both of the terms at the same time. A decrease in bias often leads to an increase in variance and vice-versa.

Ensemble classifiers try to solve this problem by training multiple classifiers that learn different patterns in the data and then combine the results to yield a conclusion that an otherwise individual classifier would not have been able to learn. The bagging technique does this by training multiple weak classifiers non-sequentially and later combining the predictions based on majority vote to make a decision in a binary classification problem. Boosting follows a sequential approach where a single weak classifier is trained sequentially in multiple boosting rounds to progressivley improve the model. A weak classifier is usually termed as a classifier that can predict an outcome better than chance.

2. BASE CLASSIFIER

A typical base classifer can be any of the well known classifiers such an k-NN, decision trees, perceptron or even SVMs. For this survey, we would restrict classifiers to be any of k-NN, decision trees or perceptron to focus on the ensemble techniques.

A decision tree is a tree with nodes of the tree representing decisions based on a selected attribute of the data set. The branches are labelled with the outcome of the decision and represent the consequence of the decision made at the parent. A traversal till the leaf of the tree constitute a decision for a given data point and signifies the outcome. For decision tree ensembles, a decision stump, a one level decision tree, is often chosen as the weak learner. The decision stump consists of a single root node which constitutes a decision.

For k-NN classification algorithms, k neighbors are chosen to decide the label of a test point. Given a test data point, the classification is based on how close the given point is to its neighbors. The number of neighbors is the value of k. For higher dimensional data vectors, a distance measure such an Euclidean or Manhattan distance is used.

3. BIAS AND VARIANCE

Error in a classification algorithm scheme is typically expressed as the sum of bias, variance and irreducible error.

$$E = bias^2 + variance + irreducible\ error \tag{1}$$

Error due the bias factor can be thought of as how far on average is our model differing from the expected result. Error due to variance can be thought of as how different are the predictions from each other for a given data vector. Bias and variance are important aspects that needs consideration because, trying to reduce one over the other is infact, a correlation on undefitting or overfitting of the model in question. Reducing one would be at the expense of the other and, in a practical scenario, a good choice would be to balance both of the error factors.

4. BAGGING

Bootstrap Aggregation or Bagging is an ensemble technique which averages the prediction over a collection of bootstrap samples. A bootstrap sample is a sampling of the training data set to build weak models from. The sampling can be either with replacement so that the generated data samples for each of the models do not conform to a predefined bias, or it can be based on a model with a gaussian noise added so that the effect of bias in the sample selection is reduced.

With n bootstrap samples, n classifiers are trained to predict the outcome of a test data point x. The final outcome is decided based on averaging of all the outcomes of individual classifiers.

$$\bar{f}(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \tag{2}$$

Here, $\bar{f}(x)$ represents the bagged prediction, combining the predictions of the individual base classifiers.

This report is the draft to a full masters study report in the making required as part of the Masters in Computer Science program at UNC Charlotte.

DOI: 10.1145/1559755.1559763

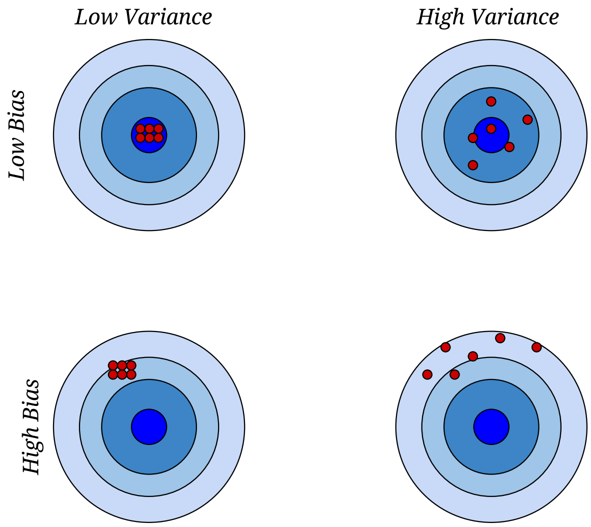

Fig. 1. Figure that explains bias and variance

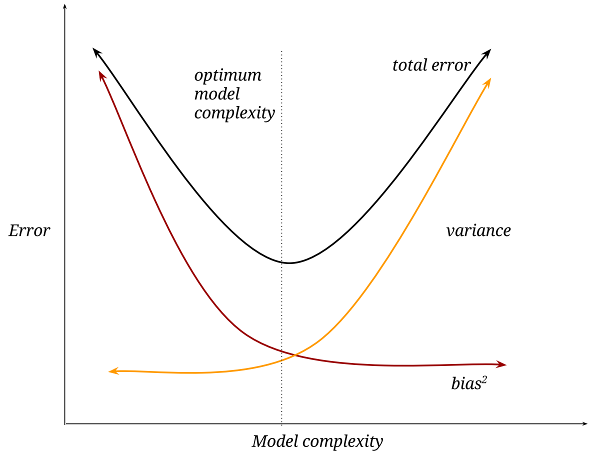

Fig. 2. Figure depicting the relationship between total error, bias and variance

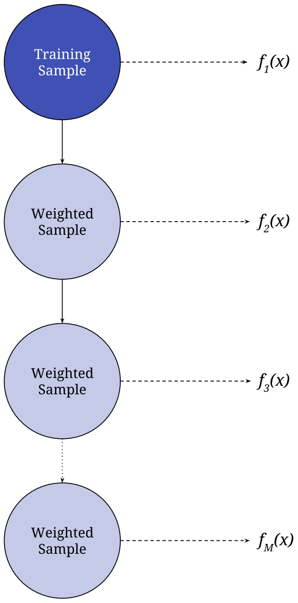

Fig. 3. Figure depicting the boosting rounds

5. BOOSTING

Boosting is also a technique which combines the output of many weak classifiers. It is similar only superficially to bagging in that boosting sequentially improves the predictive power of a weak classifier, whereas bagging involves combining outputs of many classifiers. Boosting involves weighting the data points sequentially based on the outcome of the previous output and there by producing a series of weak classifiers. The below equation shows how the boosted classifier $f(x)$ arrives at its final prediction.

$$f(x) = sign(\sum_{i=1}^{M} \alpha_i f_i(x)) \tag{3}$$

Here, the α_i's are the weights associated with each classifier. For adaboost,

$$\alpha_i = \frac{1}{2} \ln \frac{1 - \epsilon_i}{\epsilon_i} \tag{4}$$

5.1 Adaboost

Given: $(x_1, y_1), (x_2, y_2), ...(x_m, y_m)$ where $x_i \epsilon X, y_i \epsilon Y = -1, +1$

Algorithm 1 Adaboost algorithm

```
(1) Initialize $D_1(i) = \frac{1}{m}$
(2) For $t = 1, ....T$
    —Train base learner using distribution $D_t$
    —Get base classifier $f_t : X \longrightarrow R$
    —Choose $\alpha_t \epsilon R$
    —Update:
        $$D_{t+1}(i) = \frac{D_t(i)exp(-\alpha_t y_i h_t(x_i))}{Z_t} \quad (5)$$
     $Z_t$ is a normalization factor
(3) Output the final classifier:
        $$f(x) = sign(\sum_{i=1}^{M} \alpha_i f_i(x)) \quad (6)$$
```

6. BAYES OPTIMAL CLASSIFIER

The Bayes Optimal Classifier is also an ensemble technique which looks at all possible hypotheses in the hypothesis space. As such, in practical scenarios, it becomes computationally expensive to do an exhaustive search of the entire hypothesis space. In machine learning contexts, what is of interest, is the best hypothesis h out of the hypothesis space H, having observed the training data distribution D. The best hypothesis is the most probable hypothesis.

According to Bayes Theorem,

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \tag{7}$$

—$P(h)$ is the prior probability of the hypotheses h
—$P(D)$ is the probability of the data
—$P(D|h)$ is the probability of the data D given the hypothesis h
—$P(h|D)$ is the probability of the hypothesis h having observed the data

The most probable hypothesis also called the maximum a posteriori hypothesis (MAP) can be expressed as

$$\begin{aligned} h_{MAP} &= argmax_{h \epsilon H} P(h|D) \\ &= argmax_{h \epsilon H} \frac{P(D|h)P(h)}{P(D)} \\ &= argmax_{h \epsilon H} P(D|h)P(h) \end{aligned} \tag{8}$$

since $P(D)$ is a constant, dropping it from the RHS.

In cases where all the hypothesis are equally likely, the above equation can further be simplified to obtain a maximum likelihood hypothesis (ML). The equation for h_{ML} is given below:

$$h_{ML} = argmax_{h \epsilon H} P(D|h) \tag{9}$$

Now, the Bayes Optimal Classifier predicts the output by combiming the predictions of all the hypothesis which are weighed by their respective posterior probabilities.Expressed as:

$$P(v_j|D) = \sum_{h_i \epsilon H} P(v_j|h_i)P(h_i|D) \tag{10}$$

where $P(v_j|D)$ is the probability that v_j is the right classification.

With this, the Bayes Optimal Classifier can be expressed as:

$$argmax_{v_j \epsilon V} \sum_{h_i \epsilon H} P(v_j|h_i)P(h_i|D) \tag{11}$$

7. CONCLUSION

Ensemble techniques such as random forests, which are fast algorithms, are used in Xbox for real time predictions.