Random Forests

- Author
  LEO BREIMAN
  Statistics Department, University of California, Berkeley, CA 94720
- Presenter
  Ryoichi Omae
  Applied Mathmatics department, Tokyo University of Science, Japan

1.2. Outline of paper

Section 2 gives some theoretical background for random forests. Use of the Strong Law of Large Numbers shows that they always converge so that overfitting is not a problem. We give a simplified and extended version of the Amit and Geman (1997) analysis to show that the accuracy of a random forest dependes on the strength of the individual tree classifiers and a measure of the dependence between them (see Section 2 for definitions).

Section 3 introduces forests using the random selection of features at each node to determine the split. An immportant question is how many features to select at each node. For guidance, internal estimates of the generalization error, classifier strength and dependence are computed. These are called out-fo-bag estimates and are reviewed in Section 4. Section 5 and 6 give empirical results for two different forms of random features. The first uses random selection from the original inputs; the second uses random linear combinations of inputs. The results compare favorably to Adaboost.

The results turn out to be insensitive to the number of features selected to split each node. Usually, selecting one or two features gives near optimum results. To explore this and relate it to strength and correlation, an empirical study is carried out in Section 7.

Adaboost has no random elements and grows an ensemble of trees by successive reweightings of the training set where the current weights depend on the past history of the ensemble formation. But just as a deterministic random number generator can give a good imitation of randomness, my belief is that in its later stages Adaboost is emulating a random forest. Evidence for this conjecture is given in Section 8.

Important recent problems, i.e., medical diagnosis and document retrieval, often have the property that there are many input variables, often in the hundreds or thousands, with each one containing only a small amount of information. A single tree classifier will then have accuracy only slightly better than a random choice of class. But combining trees grown using random features can produce improved accuracy. In Section 9 we experiment on a simulated data set with 1,000 input variables, 1,000 examples in the training set and a 4,000 example test set. Accuracy comparable to the Bayes rate is achieved.

In many applications, understanding of the mechanism of the random forest "black box" is needed. Section 10 makes a start on this by computing internal estimates of variable importance and binding these together by resuse runs.

Section 11 looks at random forests for regression. A bound for the mean squared generalization error is derived that shows that the decrease in error from the individual trees in the forest depends on the correlation between residuals and the mean squared error of the individual trees. Empirical results for regression are in Section 12. Concluding remarks are given in Section 13.

2. Characterizing the accuracy of random forests

2.1. Random forests converge

Given an ensemble of classifiers $h_1(\mathbf{x}), h_2(\mathbf{x}), ..., h_{\boldsymbol{K}}(\mathbf{x})$, and with the training set drawn at random from the distribution of the random vector $\mathbf{Y}, \mathbf{X}$, define the margin functions as

$$mg(\mathbf{Y}, \mathbf{X}) = av_k I(h_k(\mathbf{X}) = Y) - \max_{j \neq Y} av_k I(h_k \mathbf{X}) = j).$$

where $I(\cdot)$ is the indicator function. The margin measures the extent to which the average number of votes at $\mathbf{x}, Y$ for the right class exceeds the average vote for any other class. The larger the margin, the more confidence in the classification. The generalization error is given by

$$PE^* = P_{\mathbf{X},Y}(mg(\mathbf{X}, Y) < 0)$$

where the subscripts $\mathbf{X}, Y$ indicate that the probability is over the $\mathbf{X}, Y$ space.

In random forests, $h_k(\mathbf{X}) = h(\mathbf{X}, \Theta_k)$. For a large number of trees, it follows from the Strong Law of Large Numbers and the tree structure that:

Theorem 1.2. As the number of trees increases, for almost surely all sequences $\Theta_{1,...} PE^*$ converges to

$$P_{\mathbf{X},Y}(P_\Theta(h(\mathbf{X}, \Theta) = Y) - \max_{j \neq Y} P_\Theta(h(\mathbf{X}, \Theta) = j) < 0).$$