Use of blending classifier to predict a biological response

Cheng-Bin Ren

Zhejiang University, Hangzhou 310027, China

**Abstract.** In the age of Big Data, the usage of data is more and more often in all kinds of industries.

1. Introduction
2. Related Work
   1. SVM
   2. Random Forrest

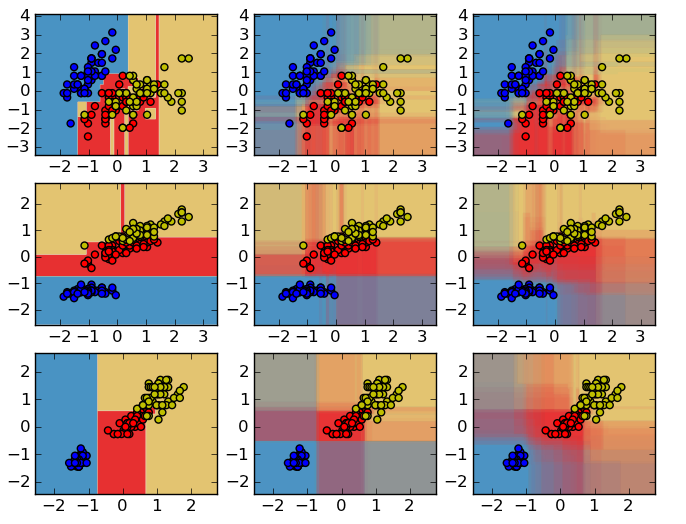
Random Forrest has been proposed by Breiman (2001) as an enhancement of Tree Bagging. To build a tree it uses a bootstrap replica of the learning sample, and the CART algorithm (without pruning) together with the modification used in the Random Subspace method. At each test node the optimal split is derived by searching a random subset of size K of candidate attributes (selected without replacement from the candidate attributes). Empirical studies have shown that Random Forests significantly outperform Tree Bagging and other random tree ensemble methods in terms of accuracy. In terms of degree of randomization, this algorithm is stronger than Tree Bagging, specially, if K is small compared to the number of attributes, n. It is also stronger than Random Subspace since it combines this method with bootstrap sampling. Randomization is both implicit (attribute and cut-point) and explicit (attribute). We use the notation RFK, with K = d for the default setting, and K =∗ for the best result over the range K = 1, … , n.

In Random Forrest, each tree in the ensemble is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set. In addition, when splitting a node during the construction of the tree, the split that is chosen is no longer the best split among all features. Instead, the split that is picked is the best split among a random subset of the features. As a result of this randomness, the bias of the forest usually slightly increases (with respect to the bias of a single non-random tree) but, due to averaging, its variance also decreases, usually more than compensating for the increase in bias, hence yielding an overall better model.

* 1. Extremely Randomized Trees

Extremely Randomized Tress is a tree-based ensemble method for supervised classification and regression problems. It essentially consists of randomizing strongly both attribute and cut-point choice while splitting a tree node. In the extreme case, it builds totally randomized trees whose structures are independent of the output values of the learning sample. The strength of the randomization can be tuned to problem specifics by the appropriate choice of a parameter.

In extremely randomized trees, randomness goes one step further in the way splits are computed. As in random forests, a random subset of candidate features is used, but instead of looking for the most discriminative thresholds, thresholds are drawn at random for each candidate feature and the best of these randomly-generated thresholds is picked as the splitting rule. This usually allows to reduce the variance of the model a bit more, at the expense of a slightly greater increase in bias.



**Fig. 1.** Decision surfaces of a decision tree, of a random forest and of an extra-trees classifier

* 1. Blending Classifier

1. Methodology
   1. Datasets

The data is provided by Boehringer Ingelheim for research. The data is in the comma separated values (CSV) format. Each row in this data set represents a molecule. The first column contains experimental data describing an actual biological response; the molecule was seen to elicit this response (1), or not (0). The remaining columns represent molecular descriptors (d1 through d1776), these are calculated properties that can capture some of the characteristics of the molecule - for example size, shape, or elemental constitution. The descriptor matrix has been normalized.

* 1. Blending Method

### **.** Random Forrest

In contrast to the original publication, in this research, the implementation combines classifiers by averaging their probabilistic prediction, instead of letting each classifier vote for a single class.

* 1. Evaluation Method

Predicted probabilities that a molecule elicits a response are evaluated using the log loss metric.

Log loss is defined as:

,

where N is the number of samples, log is the natural logarithm, is the posterior probability that the sample elicited a response, and is the ground truth ( means the molecule elicited a response, means that it did not).

* 1. Experiments

1. Results

**References**