The RandomForest algorithm

((From RNews (3) 2002, p.18))
The random forests algorithm (for both classification and regression) is as follows:

1. Draw $n_{\texttt{tree}}$ bootstrap samples from the original data.

2. For each of the bootstrap samples, grow an *unpruned* classification or regression tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample $m_{\texttt{try}}$ of the predictors and choose the best split from among those variables. (Bagging can be thought of as the special case of random forests obtained when $m_{\texttt{try}} = p$, the number of predictors.)

3. Predict new data by aggregating the predictions of the $n_{\texttt{tree}}$ trees (i.e., majority votes for classification, average for regression).

An estimate of the error rate can be obtained, based on the training data, by the following:

1. At each bootstrap iteration, predict the data not in the bootstrap sample (what Breiman calls "out-of-bag", or OOB, data) using the tree grown with the bootstrap sample.

2. Aggregate the OOB predictions. (On the average, each data point would be out-of-bag around 36% of the times, so aggregate these predictions.) Calculate the error rate, and call it the OOB estimate of error rate.

Our experience has been that the OOB estimate of error rate is quite accurate, given that enough trees have been grown (otherwise the OOB estimate can bias upward; see Bylander(2002).

Extra information from Random Forests The `randomForest` package optionally produces two additional pieces of information: a measure of the importance of the predictor variables, and a measure of the internal structure of the data (the proximity of different data points to one another).