An Adventure in randomForest – A Set of Machine Learning Tools

Andy Liaw
Biometrics Research, Merck Research Laboratories
Rahway, New Jersey, USA

Abstract:

Random Forest is a relatively new method in machine learning. It is built from an ensemble of classification or regression trees that are grown with some element of randomness. While algorithmically it is a relatively simple modification of bagging, conceptually it is quite a radical step forward. The algorithm has been shown to have desirable properties, such as convergence of generalization errors (Breiman, 2001). Empirically, it has also been demonstrated to have performance competitive with the current leading algorithms such as support vector machines and boosting. We will briefly introduce the algorithm and intuitions on why it works, as well as the extra features implemented in the R package randomForest, such as variable importance and proximity measures. Some examples of application in drug discovery will also be discussed.