Search for the Smallest Random Forest

by
Heping Zhang
Yale University
300 George Street
New Haven, CT 06510
heping.zhang@yale.edu

Abstract

Random forests have emerged as one of the most commonly used nonparametric statistical methods in many scientific areas, particularly in analysis of high throughput genomic data. A general practice in using random forests is to generate a sufficiently large number of trees, although it is subjective as to how large is sufficient. Furthermore, random forests are viewed as b

In this work, we address a fundamental issue in the use of random forests: how large does a random forest have to be? We conclude that a random forest does not have to be large. We propose a specific method to find small forests (e.g., in a single digit number of trees) that can achieve the prediction accuracy of a large random forest (in the order of thousands of trees). By reducing the size of a random forest to a manageable size, the random forest is no longer a black-box. Our conclusion is supported by both exten-sive simulation studies and a real study on prognosis of breast cancer.

This is joint work with Minghui Wang.