S05-2 (02/16 09:25-09:50)

FACT: High-Dimensional Random Forests Inference

Chien Ming Chi
Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

Abstract

Random forests is one of the most widely used machine learning methods over the past decade thanks to its outstanding empirical performance. Yet, because of its black-box nature, the results by random forests can be hard to interpret in many big data applications. Quantifying the usefulness of individual features in random forests learning can greatly enhance its interpretability. Existing studies have shown that some popularly used feature importance measures for random forests suffer from the bias issue. In addition, there lack comprehensive size and power analyses for most of these existing methods. In this paper, we approach the problem via hypothesis testing, and suggest a framework of the self-normalized feature-residual correlation test (FACT) for evaluating the significance of a given feature in the random forests model with bias-resistance property, where our null hypothesis concerns whether the feature is conditionally independent of the response given all other features. Such an endeavor on random forests inference is empowered by some recent developments on high-dimensional random forests consistency. The vanilla version of our FACT test can suffer from the bias issue in the presence of feature dependency. We exploit the techniques of imbalancing and conditioning for bias correction. We further incorporate the ensemble idea into the FACT statistic through feature transformations for the enhanced power. Under a fairly general high-dimensional nonparametric model setting with dependent features, we formally establish that FACT can provide theoretically justified random forests feature p-values and enjoy appealing power through nonasymptotic analyses. The theoretical results and finite-sample advantages of the newly suggested method are illustrated with several simulation examples and an economic forecasting application in relation to COVID-19. This is a joint work with Dr. Yingying Fan (University of Southern California) and Dr. Jinchi Lv (University of Southern California).

Keywords:
Random Forests; Nonparametric Inference; High Dimensionality; Nonasymptotic Theory; Size and Power; Bias-Resistance; Sparsity and Conditional Independence.