Improving Predictions for Tree Ensembles using Distributions of Estimated Probabilities with Applications in Record Linkage

Samuel Ventura[1], **Rebecca Nugent**[1]

1. Department of Statistics, Carnegie Mellon University
*Contact author: sventura@stat.cmu.edu

Keywords: Random Forests, Classification Trees, Prediction, Record Linkage, Entity Resolution

A random forest is an ensemble of classification trees built using randomized subsets of the covariates and bootstrap samples of the observations at each split in each of the T trees in the ensemble. In a two-class problem, random forests aggregate the predictions of each underlying tree using a "majority vote" scheme, assigning the class with the majority vote amongst the underlying trees as the predicted class of the random forest. Predicted/estimated probabilities for a given class are obtained by finding the proportion of trees that voted for that class. However, much information is discarded in this process, including the individual predicted probabilities from each of the T underlying trees, which we show to often have multimodal or heavily skewed distributions. We introduce new prediction approaches that extract and incorporate information from these distributions of tree probabilities for the two-class problem. In these approaches, we calculate distributional summary statistics and use them to build an additional classifier that makes more accurate predictions. We assess these approaches on a large, labeled record linkage dataset of death records from the Syrian Civil War conflict.