

IDENTIFYING REPRESENTATIVE TREES IN RANDOM FOREST FOR SURVIVAL DATA

Mousumi Banerjee¹, Ying Ding¹, & Anne-Michelle Noone²

*Departments of Biostatistics¹ and Epidemiology², School of Public Health,
University of Michigan, USA*

Tree-based methods have become popular for analyzing right censored survival data where the primary goal is prognostic grouping of patients. Ensemble techniques such as random forest improve the accuracy in prediction and address the instability in a single tree by growing an ensemble of trees and aggregating. However, individual trees are lost in the forest. In this paper, we propose a methodology for selecting the most representative trees in a forest for survival data, based on three tree similarity metrics. For any two trees, the metrics are chosen to (1) measure similarity of the covariates used to split the trees; (2) reflect similar clustering of patients in the terminal nodes of the trees; and (3) measure similarity in predictions from the two trees. While the latter focuses on prediction, the first two metrics focus on the architectural similarity between two trees. The most representative trees in the forest are chosen based on the average similarity score assigned to each tree corresponding to each of the three metrics. Out of bag estimates of error are computed for the most representative trees using a neighbourhood of similar trees. Finally the methods are illustrated using data from a cohort study of breast cancer patients to model recurrence free survival.