Assessing the Performance and Merit of the Random Survival Forest and Cox Models on a Pancreatic Cancer Data Set

Haiming Zhou, Ph.D. (Chair)
Alan Polansky, Ph.D.
Duchwan Ryu, Ph.D.

July 1, 2019
1:00 p.m.
DuSable Hall 248

Random Survival Forest (RSF) is one of the most powerful and easily applied machine learning models for survival data. RSF sacrifices some of the interpretability of the decision trees used to grow the forest in order to significantly reduce the bias and variance of the basic classification and regression tree (CART) paradigm. The lessened interpretability and higher computational intensity of RSF means that it may not always be the preferred method, even in settings where black-box methods are readily used. By contrast, the Cox Proportional Hazards (PH) model is incredibly flexible, resistant to overfitting, and transparently estimable. The tradeoff for the Cox PH model is the difficulty in construction when rigorous best practices are followed and model assumptions are violated, requiring complex extensions.

This thesis finds the best performing RSF, as measured by the Concordance Index (CI), for a cohort of pancreatic cancer patients from the Surveillance, Epidemiology and End Results (SEER) index and compares the predictive power of that model to a carefully constructed Cox model, while providing commentary on the relative strengths and weaknesses of each approach. Our conclusion marries the pros and cons of the classical and contemporary to cap a complete and cohesive narrative that insists on analyst customization for the machine learning technique and careful construction in the context of the Cox approach.