Survival Trees and Forest for Thyroid Cancer Prognostication

Mousumi Banerjee*
University of Michigan, Ann Arbor, USA - mousumib@umich.edu

Daniel Muenz
University of Michigan, Ann Arbor, USA

Megan Haymart
University of Michigan, Ann Arbor, USA

Thyroid cancer is becoming an increasingly common cancer and yet little is known about the prognostic factors associated with survival. Controversy also exists over appropriate treatment for thyroid cancer. Prognostic models are needed to determine correlates of overall survival and identify subgroups of patients with poor prognosis who may benefit from earlier therapeutic intervention. In this talk we present a tree-based model for thyroid cancer prognostication using data from the US National Cancer Database. Tree-based methods partition the covariate space into a set of rectangles, leading to a fitted model that is piecewise constant over regions of the covariate space. Trees are conceptually simple yet powerful, and are being increasingly used in biomedical studies for analyzing censored survival data where the primary goal is prognostication of patients. This talk will firstly compare and contrast several split criteria in growing survival trees. We will focus on empirical comparisons based on the thyroid cancer dataset, and further supplemented by simulations. To gain accuracy in prediction and address instability in a single tree, we next present an analysis based on growing an ensemble of trees and aggregating. Lastly, we present a methodology for identifying the most representative tree from the ensemble based on several tree distance metrics. Out of bag error based on the cumulative hazard estimate is computed for the representative tree. Based on the thyroid cancer data as well as simulations, we found strong concordance between the various split criteria used in the survival data setting. The representative tree from the ensemble was able to identify five distinct prognostic groups, defined by age, gender, lymph node involvement, tumor size, and metastasis status. Five year survival rates in these groups ranged between 65.3% and 99%. The worst prognosis was for patients above 65 years of age with distant metastasis. The prognostic groups derived can provide guidance for patient management, clinical trial design, and future treatment policy. The representative tree itself can be used as a decision making tool in the clinical setting.

Keywords: tree-based methods, censored data, splitting rules, ensemble, representative tree thyroid cancer.