Classification Tree Method Implementation in C and R for Precision Medicine
CPSC 490 Project Proposal

Catherine Xue

Advised by Prof. Heping Zhang and Prof. John Lafferty

Fall 2017

Background and Motivation

For many medical conditions several different treatments are available, whose relative efficacy are typically assessed in clinical trials. Clinical trials are usually run as double-blind, randomized controlled trials and report average relative performance for the subjects. This model obscures meaningful interactions between these treatments and subject subgroups. These subgroups are determined by pre-treatment or baseline characteristics, which, for example, might include genetic differences or symptoms of the medical condition. These characteristics can be indicative of particular reactions to a treatment that differs from the average over all clinical trial subjects. Detecting and understanding such treatment-subgroup interactions are important for selection of the optimal treatment strategy for patients and overall advancing the field of precision medicine, which the National Institutes of Health defines as “an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person.”

While attempts have been made to run clinical trials specific to subgroups of interest, these efforts depend on a priori hypotheses of treatment behavior and what subgroups will interact with the treatment. In general, such a priori hypotheses cannot easily be made for randomized controlled trials, especially with novel treatments. However, clinical trial data include a wide range of subjects’ pre-treatment
characteristics, and it is thus more efficient to detect significant treatment-subgroup interactions from existing clinical trial data.

In recent years, several clustering methods have been developed to identify subgroups of interest in clinical trial data, including some tree-based methods. Tree-based methods recursively partitions patients into subgroups, and evaluate the candidate trees. In this project, I seek to implement one such method for identifying subgroups of interest and the associated treatment-subgroup interactions.

Proposed Work

For this project, I propose to develop an R package that implements the classification tree method developed by Zhang, et al. (2010) and extended by Tsai, et al. (2016), built on an R-callable C program.

The method operates on binary outcomes (i.e. which of two treatment options is better) and binary classifiers from baseline conditions. Candidate trees are evaluated using a $U$-measure:

$$0 \leq U = \frac{\sum_{k=1}^{m} n_k |p_{k,A} - p_{k,B}|}{N} \leq 1,$$

where $m$ is the number of leaf nodes in the candidate trees and $n_k, p_{k,T}$ are the number of patients in the node and proportion of them responding better to treatment $T$, respectively. A larger value of $U$ indicates better separation between the two states at each branch point, and therefore maximizing $U$ produces the most informative tree. The R-callable C program format was selected to take advantage of C’s relative speed at highly iterative operations (such as the classification tree method) while acknowledging that R is the most useful platform for use by the deliverable’s intended audience in the biomedical field. I will be testing the tool on a simulated data set and against results from an existing, unpublished implementation written entirely in R.

This work will be completed under the direction of Prof. Heping Zhang at the Center for Collaboration for Statistics in Science (C2S2) at the Yale School of Public Health and advised by Prof. John Lafferty.
Deliverables

The deliverables for this project will be:

1. R package built on R-callable C program for executing the modified classification tree method developed by Zhang, et al.,
2. Source code for the above, and
3. Final report detailing the findings and obstacles encountered during this project.

References

