University of California, Irvine
Statistics Seminar

Constructing Balanced Treatment and Control Groups for Observational Study Data Using Random Forest

Juanjuan Fan
Professor of Statistics
San Diego State University

Thursday, January 11, 2018
4 p.m., 6011 Bren Hall
(Bldg. #314 on campus map)

In order to derive unbiased inference from observational data, matching methods are often applied to produce balanced treatment and control groups in terms of all background variables. Propensity score has been a key component in this research area. However, propensity score based matching methods in the literature have several limitations, such as model mis-specifications, categorical variables with more than two levels, difficulties in handling missing data, and nonlinear relationships. Random forest, averaging outcomes from many decision trees, is nonparametric in nature, straightforward to use, and capable of solving these issues. More importantly, the precision afforded by random forest may provide us with a more accurate and less model dependent estimate of the propensity score. In addition, the proximity matrix, a by-product of the random forest, may naturally serve as a distance measure between observations that can be used in matching. The proposed random forest based matching methods are applied to data from the National Health and Nutrition Examination Survey (NHANES). Our results show that the proposed methods can produce well balanced treatment and control groups. An illustration is also provided that the methods can effectively deal with missing data in covariates.

For directions/parking information, please visit https://uci.edu/visit/maps.php and http://www.ics.uci.edu/about/visit/index.php