CANSSI-CRT Workshop
Modern Methods in Survey Sampling
July 8-10, 2024

Fitting Classification Trees to Complex Survey Data

Jean Opsomer

WESTAT

Classification tree algorithms are a convenient method to perform variable selection and obtain interpretable structures relating covariates and an outcome of interest. When fitting classification trees to survey data, it is common to ignore sampling weights as well other design characteristics such as stratification and clustering. However, unless the survey design is uninformative, there is a risk that the inference for the classification tree is incorrect. We propose an extension of the popular Chi-square Automatic Interaction Detector (CHAID) approach that accounts for the design in its classification criterion. We discuss the statistical properties of the resulting algorithm under a design-based framework. We compare its performance to existing weighted and unweighted algorithms, and we illustrate the use of the method using data from the U.S. American Community Survey.