Fitting Classification Trees to Complex Survey Data

Jean Opsomer
Westat

Joint work with Minsun Riddles

May 31, 2023

Outline

1. Background and motivation
2. Recursive partitioning and CHAID
3. Survey CHAID (sCHAID)
4. Theoretical results
5. Implementation

1) Background

- Surveys continue to be an important data collection source for government agencies and other organizations
- Survey weights need to be used in estimation to account for the sampling design
- These weights also adjust for:
  - unknown eligibility (non-contact, locating errors)
  - nonresponse
- Modeling (implicit or explicit) is required for these adjustments, so it is crucial to do this in a theoretically valid, transparent and reproducible manner

Weighting @ Westat

1. Identify base weights (inverse of inclusion probabilities, known)
2. Create unknown eligibility adjustment weights, using cell-based model
3. Create nonresponse adjustment weights, using cell-based model
4. Calibrate weights to control totals
5. Create replicate weights for variance estimation based on (1) and repeating (2)-(4) for each replicate

Response homogeneity group (RHG) model

- RHG model:
  - Population is partitioned into groups in which the propensity to respond is constant
  - The groups are defined in terms of variables known for all sampled units, treated as fixed
  - Response propensity within each group is not known
- Alternative model:
  - Assume known parametric model for propensity function (with same covariates as above)
  - Model provides response propensity for each sampled unit

  $\rightarrow$ In practice, estimated propensities are binned by quantiles, effectively resulting in cell-based model

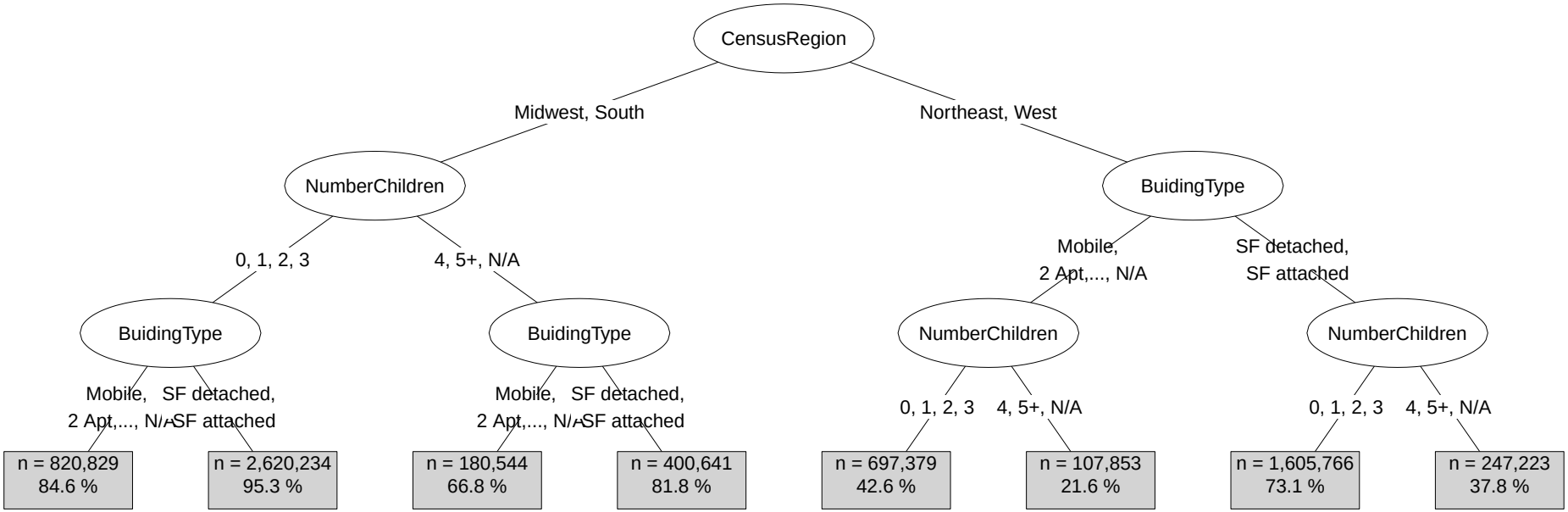
Determining the groups

- In many situations, there are (too) many possible covariates, so it is necessary to determine groups based on the data
  - $\Rightarrow$ Use recursive partitioning method to fit a *classification tree* (Breiman et al, 1984) predicting response status
- In survey context, exchangeability (*iid*) assumptions of classification tree methods do not apply
- Goal: develop a modification of an existing classification tree method that accounts for the design-based setting

2) Recursive partitioning

- Recursive partitioning: define successive splits of the data into subsets, based on a selection criterion to chose the splits, until the stopping criteria are met
  - Selection criterion: split node which results in largest decrease in estimated mean squared error, entropy, ...
  - Stopping criteria: no eligible splits, number of nodes, depth of tree, minimum size of node, pure nodes, etc

CHAID example



CHAID

- Chi-square Automatic Interaction Detection (Kass, 1980): method for recursive partitioning that decides on splitting based a χ^2 tests (or F, for continuous variables)
  - Selection criterion in CHAID: p-value
- Popular in marketing applications, especially market segmentation and prediction
- Typical implementation of CHAID selection step:
  1. Choose best possible split for each categorical variable
  2. Select which variable to split on among the splits identified in (1)

  $\rightarrow$ Only consider splits with p-value smaller than predetermined cutoff, otherwise do not split

CHAID advantages/disadvantages

+ Flexible and easy to implement: can be applied to categorical, ordinal, continuous data
+ Local decision criterion: decision on splitting does not depend on other splits
+ Can use p-values as tuning parameters, which are easy to interpret
– Tends to select too many splits on variables with many categories
– Does not actually estimate an overall model, making it difficult to evaluate statistically
– Does not accommodate the survey design
- We developed *sCHAID* (survey CHAID) to address the last issue

3) Survey design and estimation

- Finite population $U = \{1, \ldots, N\}$, sample s of size n selected according to sampling design $p(s)$ with inclusion probabilities $\pi_i = \Pr(i \in s), \pi_{ij} = \Pr(i, j \in s)$
- If full sample observed, Horvitz-Thompson estimator

$$\widehat{t}_\pi = \sum_s \frac{y_i}{\pi_i}$$

  is unbiased for population total $t_y = \sum_U y_i$
- Define the response indicator $R_i, i \in U$

$$R_i = \begin{cases} 1 & \text{if } i \text{ would respond to the survey if selected} \\ 0 & \text{otherwise} \end{cases}$$

  The R_i are independent with $\mathrm{E}(R_i) = p_i$
- If the p_i known, estimator with nonresponse

$$\widehat{t}_{\pi,p} = \sum_s \frac{y_k}{\pi_i} \frac{R_i}{p_i}$$

  remains unbiased for population total t_y

Survey estimation under RHG model

- Under the RHG model, the population is partitioned into groups $U_g^*, g = 1, \ldots, G$ of size N_g^*, with P_g^* the (constant, unknown) propensity to respond in group g

$$p_i = P_g^* \quad \text{for} \quad i \in U_g^*$$

- Sample s partitioned into $s_g^*, g = 1, \ldots, G$
- Naive RHG estimator

$$\widehat{T}_{\text{RHG}} = \sum_{g=1}^{G} \frac{\sum_{s_g^*} \frac{1}{\pi_i}}{\sum_{s_g^*} \frac{1}{\pi_i} \frac{R_i}{P_g^*}} \sum_{s_g^*} \frac{y_i}{\pi_i} \frac{R_i}{P_g^*} = \sum_{g=1}^{G} \sum_{s_g^*} \frac{y_i}{\pi_i} \frac{R_i}{\widehat{P}_g^*}$$

with

$$\widehat{P}_{g^*} = \frac{\sum_{s_g^*} \frac{R_i}{\pi_i}}{\sum_{s_g^*} \frac{1}{\pi_i}}$$

is asymptotically unbiased and does not require knowledge of the P_g^*

RHG tree structure

- The U_g^* are defined as the intersections of categorical and/or ordinal variables $X_k, k = 1, \ldots, K$, which can take L_k different values
  - For simplicity, $L_k = L$ and categories are labeled $(1, \ldots, L)$ for all k
  - $G = K \times L$
- When G is large relative to n, the naive RHG estimator is unstable
  - Number of respondents in some s_g^* can become small or even zero
  - Too many cells leads to increased weight variation and estimator variance
- We want to reduce the number of cells to a more manageable number by collapsing cells with the same (or similar) response propensities

Defining splits

- For each variable X_k, we can divide the population into L non-overlapping "population slices" containing all units i with $X_{ik} = l$ for $l = 1, \ldots, L$
- To define a binary split for X_k, we consider combinations of these slices into

$$U_{kl} = \{i \in U : X_{ik} \leq l\} \text{ and } U_{kl}^c = \{i \in U : X_{ik} > l\} \qquad \text{(ordinal)}$$
$$U_{kl} = \{i \in U : X_{ik} = l\} \text{ and } U_{kl}^c = \{i \in U : X_{ik} \neq l\} \qquad \text{(categorical)}$$

- We write $\mathcal{A}_{kl}, \mathcal{A}_{kl^c}$ for the set of indices g such that

$$\bigcup_{g \in \mathcal{A}_{kl}} U_g^* = U_{kl} \text{ and } \bigcup_{g \in \mathcal{A}_{kl^c}} U_g^* = U_{kl}^c$$

with sizes $N_{kl} = \sum_{g \in \mathcal{A}_{kl}} N_g^*$, $N_{kl^c} = \sum_{g \in \mathcal{A}_{kl^c}} N_g^*$

Response propensities on splits

- Group propensities P_{kl} (and P_{kl^c})

$$P_{kl} = \frac{\sum_{i \in U_{kl}} p_i}{N_{kl}} = \frac{\sum_{g \in \mathcal{A}_{kl}} N_g^* P_g^*}{\sum_{g \in \mathcal{A}_{kl}} N_g^*}$$

- Estimated by

$$\widehat{P}_{kl} = \frac{\sum_{i \in s_{kl}} R_i/\pi_i}{N_{kl}} = \frac{\sum_{g \in \mathcal{A}_{kl}} N_g^* \widehat{P}_g^*}{\sum_{g \in \mathcal{A}_{kl}} N_g^*}$$

with

$$\widehat{P}_{g^*} = \frac{\sum_{s_g^*} \frac{R_i}{\pi_i}}{\sum_{s_g^*} \frac{1}{\pi_i}}$$

sCHAID splitting criterion

- To decide whether to split the dataset at $X_k = l$, we perform a statistical test of the form

$$\begin{aligned} H_0 &: P_{kl} = P_{kl^c} \\ H_a &: P_{kl} \neq P_{kl^c} \end{aligned}$$

- Survey-weighted χ^2 test statistic

$$\widehat{W}_{kl} = n\frac{(\widehat{P}_{kl} - \widehat{P}_{kl^c})^2}{\widehat{V}_{kl}}$$

with p-value

$$\widehat{q}_{kl} = \Pr(\chi^2_1 > \widehat{W}_{kl})$$

- The same test statistic is computed for other covariates, and the one with the smallest p-value is selected for splitting the dataset
- What is $\widehat{V}_{kl}$, and what is distribution of $\widehat{W}_{kl}$?

Response propensity tree

- A CHAID tree fitted to a sample s consists of an ordered sequence of R_s splits, denoted $\widehat{\mathcal{T}}_s$
- Each of the splits is defined by two sets of indices $(\mathcal{A}_{kl}, \mathcal{A}_{kl^c})$ among the eligible groups
- We would like $\widehat{\mathcal{T}}_s$ to converge to a non-random split sequence $\mathcal{T}_U$ that depends on the population and the sample design, but not on the realized sample

4) Theoretical results

- $\widehat{P}_{kl}$ and $\widehat{P}_{kl^c}$ are of the form

$$\widehat{P}_{kl} = \frac{\sum_{g \in \mathcal{A}_{kl}} N_g^* \widehat{P}_g^*}{\sum_{g \in \mathcal{A}_{kl}} N_g^*}$$

- We can obtain variance estimator $\widehat{V}_{kl}$ and properties of $\widehat{W}_{kl}$ based on those of the vector of Horvitz-Thompson estimators $\widehat{P}_g^*$, $g = 1, \ldots, G$

Asymptotic framework

- Design-based asymptotic framework:
  - Sequence of finite populations U_N with $N \to \infty$
  - Associated sequence of sampling designs $p_N(s)$ with $n \to \infty$
  - Asymptotic design normality of Horvitz-Thompson estimators
  - Regularity conditions on π_i, π_{ij}, such that $\mathrm{Var}(\widehat{P}_g^*) = O(1/n)$
- RHG model generating response indicators R_i in population
  - In each group U_g^*, the R_i are *iid* Bernoulli(P_g^*)
  - The number of groups G is fixed
- We will consider properties of the estimators under the combined design-model distribution

Asymptotic distribution of the group propensity estimators

- Under stated assumptions, the vector of estimators $\widehat{\boldsymbol{P}}^* = (\widehat{P}_1^*, \ldots, \widehat{P}_G^*)^T$ has the following asymptotic distribution

$$\sqrt{n}\left(\widehat{\boldsymbol{P}}^* - \boldsymbol{P}^*\right) \Rightarrow \mathcal{N}(\boldsymbol{0}, \boldsymbol{V}^*)$$

with $\boldsymbol{V}^* = \boldsymbol{V}_1^* + \boldsymbol{V}_2^*$, where $\boldsymbol{V}_1^*$ is a matrix with elements

$$[\boldsymbol{V}_1^*]_{gg'} = \frac{n}{N_g^* N_{g'}^*} \sum_{U_g^*} \sum_{U_{g'}^*} (\pi_{ij} - \pi_i \pi_j) \frac{p_i}{\pi_i} \frac{p_j}{\pi_j}$$

and

$$\boldsymbol{V}_2^* = \operatorname{diag}\left\{ \frac{n}{N_g^{*2}} \sum_{U_g^*} \frac{1}{\pi_i} P_g^*(1 - P_g^*) \right\}$$

- Variance terms can be consistently estimated based on sample

Asymptotic distribution of the test statistic

- $V_{kl} = n\mathrm{Var}(\widehat{P}_{kl} - \widehat{P}_{kl^c})$ is complicated linear combination of terms in $\boldsymbol{V}^*$, estimated by $\widehat{V}_{kl}$ based on $\widehat{\boldsymbol{V}}^*$
- Under the stated assumptions, the test statistic for the split on variable X_k at category l,

$$\widehat{W}_{kl} = n\frac{(\widehat{P}_{kl} - \widehat{P}_{kl^c})^2}{\widehat{V}_{kl}}$$

has an asymptotic non-central χ^2_1 distribution, with non-centrality parameter equal to

$$\lambda_{kl} \quad = \quad n\frac{(P_{kl} - P_{kl^c})^2}{V_{kl}}$$

Convergence of the sample-based tree

- For simplicity, consider a simple sCHAID procedure that performs R splits and stops
- Suppose that we are choosing splits by selecting the smallest p-values among K possible splits at each step
- With probability going to 1, the smallest p-value corresponds to the split with the largest value for

$$\lambda_{kl} \quad = \quad n\frac{(P_{kl} - P_{kl^c})^2}{V_{kl}}$$

  among those considered at each step
- With probability going to 1, the sample-based tree $\widehat{\mathcal{T}}_s$ converges to a population tree $\mathcal{T}_U$, with sequence of splits defined at each step by

$$(\mathcal{A}_{kl}, \mathcal{A}_{kl^c}) \quad = \quad \arg_{k,l} \max \left((P_{kl} - P_{kl^c})^2 / V_{kl} \right)$$

- Interpretation?

5) sCHAID implementation

- Modified version of existing R-package *CHAID*[1]
- Requires *survey* R-package and creates an object of class *constparty*
- Uses *p*-values of second-order (Satterthwaite) Rao-Scott adjusted χ^2 tests (Rao and Scott, 1987)
- At a given step,
  - For each variable, choose best binary split by merging most similar categories
    - Continuous variables: create ordinal variable using deciles
    - Ordinal variables: collapse adjacent categories with largest p-value
    - Nominal variables: collapse any two categories with largest p-value
  - Split using best binary split (with the smallest p-value) among binary splits with p-value $< \alpha$
  - Stop if the node is pure, no splits have p-value $< \alpha$, or any other user-specified stopping criterion is satisfied

[1] https://rdrr.io/rforge/CHAID/man/chaid.html

Illustration

- Simulation on real data: 2017-2021 American Community Survey Public Use Microdata Sample (ACS PUMS)[2]
  - 6,680,469 household-level records
- Treat ACS PUMS as sampling frame, draw 1,000 stratified unequal-probability two-stage samples of households (HHs)
  - Strata: 9 census divisions
  - PSU: Public Use Microdata Areas (PUMAs), simple random sample of 10
  - Elements: HHs in selected PUMA, simple random sample of 100

[2] https://www.census.gov/programs-surveys/acs/microdata/access.html

Sampling design

Census Division	# PUMAs	# sampled PUMAs	Average # HHs per PUMA	# sampled HHs per PUMA
New England	109	10	2,946	100
Middle Atlantic	310	10	2,790	100
East North Central	339	10	3,026	100
West North Centra	159	10	3,008	100
South Atlantic	455	10	2,950	100
East South Central	138	10	2,984	100
West South Central	294	10	2,599	100
Mountain	180	10	2,791	100
Pacific	367	10	2,642	100

Response mechanism

- The response status $R_i \sim \text{Bernoulli}(p_i)$ was generated for each household i with

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{k=1}^{6}\sum_{l=1}^{L_k-1}\beta_k X_{ik,l}$$

  with covariates
  - census region indicators (Northeast, Midwest, South, West)
  - building type (one-family house detached home or not)
  - tenure status (owned or not)
  - presence of children in the household (yes/no)
  - health insurance coverage status (yes/no)
  - property value (quintiles)
- The parameters were obtained by fitting a logistic model for early response, with 1 if the ACS response mode was web or mail and 0 if the response mode was telephone or in-person
- Average response rate was 74.5%

Response propensity modeling

- As potential covariates for constructing trees, the same variables were provided, plus:
  - telephone services status (yes/no)
  - family type (married, not married, non-family)
  - number of persons in family (1, 2, 3, 4, 5+)
  - access to internet (yes/no)
- Methods considered
  - CHAID (unweighted, no design information)
  - sCHAID
  - rpms (R implementation of Toth and Eltinge (2011))

Results

Algorithm	Proportion of trees containing all correct covariates	Proportion of trees containing incorrect covariates
CHAID	0.99	0.11
sCHAID	0.99	0.06
rpms	0.95	0.07

Results (2)

Estimate the mean household income, using cells determined by selected trees and RHG estimator

Algorithm	Relative bias (%)	Relative root mean squared error (%)
CHAID	0.20	0.68
sCHAID	0.21	0.52
rpms	0.33	0.59

6) Conclusions

- Classification trees are useful approach to create response propensity adjustment cells, a crucial part of survey weighting
- We propose sCHAID as a design-based recursive partitioning method
- Method can easily be extended to other survey applications outside the nonresponse adjustment context

Contact: JeanOpsomer@westat.com