Regression tree methods for subgroup identification I

Xu He

Academy of Mathematics and Systems Science,
Chinese Academy of Sciences

March 25, 2014

Outline

- The problem and notation
- The Virtual Twins method
- The SIDES method
- The Gc (GUIDE Classification) method
- Other methods and results be covered in Part II

Introduction

- Personalized medicine.
- Genome-wide association studies (GWAS) on the whole genomes for disease association and susceptibility.
- Complex diseases what are likely due to gene-gene interactions.
- Identification of the markers that define subgroups with large differential treatment effects in placebo-controlled studies.

A quick set-up of the data set

- A binary response Y
- A binary treatment variable Z
- Even split between the treatment and placebo groups
- 100 Genetic markers X, which are three-level categorical variables
- Six possible single-marker subgroups for each marker X_j, namely, $X_j = 0$, $X_j = 1$, $X_j = 2$, $X_j \neq 0$, $X_j \neq 1$, and $X_j \neq 2$
- Several markers are true; others are pure noise

Our goal

- Prognostic variables are those have marginal effects on Y but do not interact with Z
- Predictive variables are those interact with Z
- Find predictive variables
- Find subgroups with large differential treatment effects, defined by predictive variables
- When there is no predictive variables, say no
- Generalize to other types of covariates
- Generalize to other types of response, continuous or survival time with censoring

List of methods

Name	Authors
Model-based	T. Hothorn, K. Hornik, and A. Zeileis
Interaction tree	X. Su, T. Zhou, X. Yan, J. Fan and S. Yang
Quint	E. Dusseldorp and I. Van Mechelen
Virtual Twins	J. Foster, J. Taylor and S. Ruberg
SIDES	I. Lipkovich, A. Dmitrienko, J. Denne and G. Enas
GUIDE methods	W.Y. Loh, X. He and M. Man

The Virtual Twins Method

- Build a random forest on Y
- Obtain twin estimates
- Estimate treatment effect for subjects
- Build a single regression tree on the treatment effect
- Y must be binary

The forest step

- Response Y
- Candidate splitting variables X_j, T, $X_j I(T = 0)$, $X_j I(T = 1)$ if X are continuous variables
- Build a random forest
- Obtain the estimate $P(Y = 1|X, T)$.

Random Forest

- R function randomForest
- Build default of 1000 independent trees
- Each tree is built from a bootstrap sample of the data
- Each tree is built with a randomly selected subset of covariates
- Trees are not pruned
- Each tree yields a classifier $f_k(X, T)$ for Y.
- For classification problems, the final classifier is obtained by majority voting,

$$f(X, T) = \begin{cases} 1, & \text{if } \sum_k f_k(X, T) > 500 \\ 0, & \text{if } \sum_k f_k(X, T) < 500 \end{cases}$$

- For regression problems, the final classifier is obtained by averaging,

$$f(X, T) = \sum_k f_k(X, T)/1000$$

Predicting $P(Y = 1)$ from a forest

- Predicting Y for a new input

$$P(Y = 1|X, T) = \sum_k f_k(X, T)/1000$$

- Out-of-bag sample for a tree f_k: observations not included in the bootstrap data
- $o(i, k) = 1$ if observation i is an out-of-bag sample for tree k; $o(i, k) = 0$ otherwise.
- Predicting Y_i for a trial in the training data set

$$P(Y_i = 1|X = X_i, T = T_i) = \left[\sum_k \{f_k(X_i, T_i)o(i, k)\}\right] / \left\{\sum_k o(i, k)\right\}$$

Twin estimates

- Twin data sets: (X, T) and $(X, 1-T)$.
- Obtain $P(Y_i = 1|X = X_i, T = T_i)$ using out-of-bag estimate
- Obtain $P(Y_i = 1|X = X_i, T = 1 - T_i)$ using average estimate
- Twin estimates: $P_{i1} = P(Y = 1|X = X_i, T = 1)$ and $P_{i0} = P(Y = 1|X = X_i, T = 0)$
- Estimate of treatment effect for subject i: $D_i = P_{i1} - P_{i0}$

The single tree step

- Regression tree from R function rpart
- Use D as the response
- Use X as covariates
- Obtain the final tree on treatment effect
- The tree can be pruned
- Alternatively, in the classification method, use D^* as the response, where

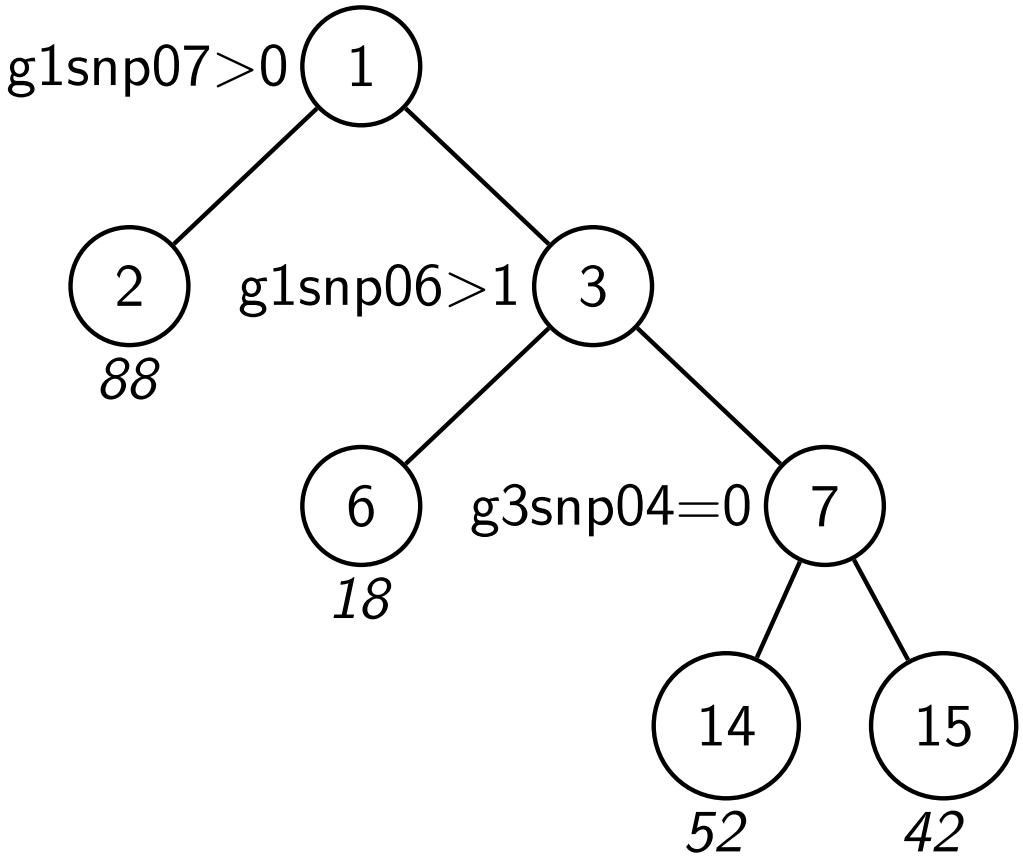
$$D_i^* = \begin{cases} 1, & \text{if } D_i > c, \\ 0, & \text{if } D_i \leq c, \end{cases}$$

and c is prespecified

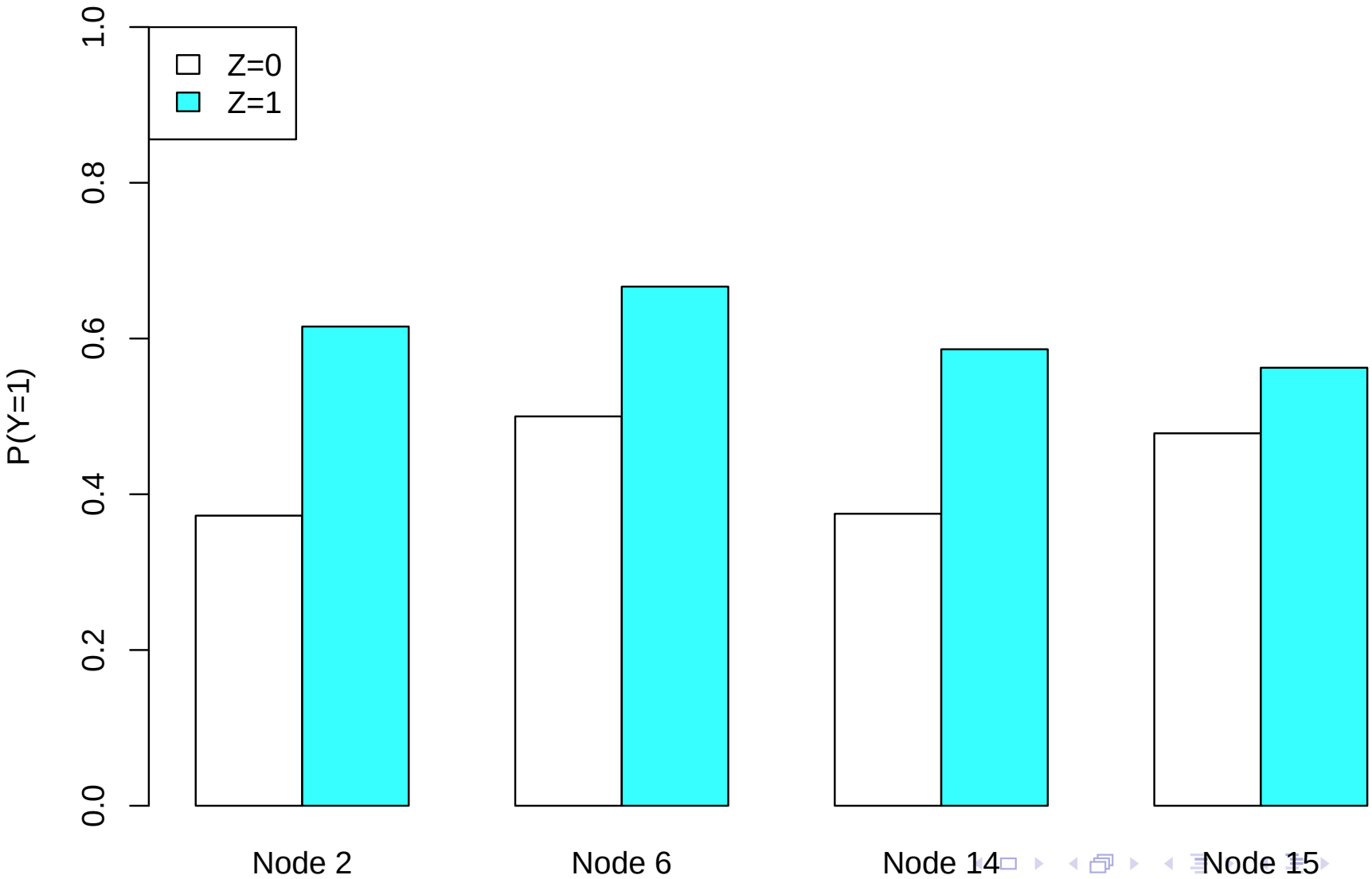
When X are categorical

- How to create X_j, T, $X_j I(T=0)$ and $X_j I(T=1)$?
- Consider the case with categorical X_j with three levels
- Solution one: Use covariates $I(X_j=1)$, $I(X_j=2)$, T, $I(X_j=1)I(T=1)$, $I(X_j=2)I(T=1)$, $I(X_j=1)I(T=0)$ and $I(X_j=2)I(T=0)$
- Solution two: Use covariates $I(X_j=1)$, $I(X_j=2)$, $I(X_j=0)$, T, $I(X_j=1)I(T=1)$, $I(X_j=2)I(T=1)$, $I(X_j=0)I(T=1)$, $I(X_j=1)I(T=0)$, $I(X_j=2)I(T=0)$ and $I(X_j=0)I(T=0)$
- We use Solution two in our simulation because it performs much better than Solution one

An example VT tree



An example VT tree



Comments on the Virtual Twin method

- Take advantage of available tree methods
- Prone to overfitting
- Slow in computation
- Although using the "treatment effect" as the response in the last step, splits with both predictive variables and prognostic variables

SIDES

- Balanced allocation procedure to create training data set and testing data set
- Balance with respect to the treatment variable and all prespecified covariates
- Recursively partition to get many candidate subgroups
- Use the testing data set or adjusted p-value to select subgroups
- Select tuning parameters using cross-validation

Subgroup identification procedure

- Start with the entire training data set
- Try all covariates, one by one
- For each covariate, try all possible splits
- For each split, retain the subgroup with larger positive treatment effect
- Select M best subgroups
- Recursively split on the newly obtained subgroups
- A subgroup can not be split by the same covariate twice
- Result in a lot of terminal subgroups

Splitting criterion

- Test the efficacy of the treatment for each of the two child subgroups
- Let E_1 and E_2 denote the test statistics of efficacy.
- Criterion 1:

$$p_1 = 2\left[1 - \Phi\left(\frac{|E_1 - E_2|}{\sqrt{2}}\right)\right],$$

  where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution

- Criterion 2:

$$p_2 = 2\min(1 - \Phi(E_1), 1 - \Phi(E_2))$$

- Criterion 3:

$$p_3 = \max(p_1, p_2)$$

- We use Criterion 1 in simulation

Splitting Criterion 4

- Use the testing data set to test the efficacy of the treatment for each of the two child subgroups
- Let T_1 and T_2 denote the test statistics of efficacy from the testing sample.
- Criterion 4:

$$p_4 = 2\left[1 - \Phi\left(\frac{|w(E_1 - E_2) + (1-w)(T_1 - T_2)}{\sqrt{2(w^2 + (1-w)^2)}}\right)\right],$$

where w is a prespecified weight

Stopping criteria

- The maximal number of covariates, L
- The minimal sample size, S
- The maximum number of best candidate covariates considered, M
- If $L = 3$ and $M = 5$ as in default, there will be at most $5 + 25 + 125$ terminal subgroups
- Subgroups must meet a continuation criteria

Continuation criteria

- Test the efficacy of the treatment for both the parent subgroup and the child subgroup
- Let p_P and p_C denote the p-value of efficacy for the parent subgroup and the child subgroup, respectively
- Retain child subgroups with $p_C \leq \gamma p_P$
- $\gamma = (\gamma_1, \ldots, \gamma_L)$
- γ is either prespecified or selected by cross-validation

Selection criterion

- Sort all terminal subgroups by p-value on testing the efficacy of the treatment variable
- If there is no testing data set, use a resampling-based method with 500 permutations to control the overall Type I error
- The resampling-based method yields a justified bound of p-value
- Using of a continuation criteria usually gives a weaker bound
- Control the Type I error of all resulted subgroups simultaneously
- If there is testing data set, use the testing sample p-value to select resulted subgroups

An example SIDES result

Subgroup	size	p-value
Full data	200	0.0023
g4snp03 $>$ 0	96	0
g4snp03 $>$ 0 ; g1snp03 $<=$ 1	86	0
g4snp03 $>$ 0 ; g4snp01 $<=$ 1	84	0
g2snp06 $<=$ 1	186	0.002
g2snp06 $<=$ 1 ; g1snp03 $<=$ 1	168	0.0013
g2snp06 $<=$ 1 ; g4snp03 $>$ 0	89	0
g2snp06 $<=$ 1 ; g1snp08 $<=$ 1	173	7e-04
g2snp06 $<=$ 1 ; g1snp07 $<=$ 1	170	8e-04
...		

Comments on the SIDES method

- Gives many result subgroups
- Higher chance to select both true covariates and false covariates
- Testing data set or permutation method for selection
- Slow in computation if using cross-validation for γ
- Compare two p-values
- One covariate be used only once in a subgroup

GUIDE classification tree

- Recursively partition terminal nodes
- Select splitting variable without specifying split set
- Choose the best split of the selected variable
- Avoid selection bias
- Prune the tree to the best size by 10 fold cross-validation

Selecting splitting variables

- Convert continuous variables into categorical variables of four or three levels by quantiles
- For the response of J levels and a variable of L levels, create an $J \times L$ contingency table
- Obtain the test statistic of the Chi-square test on independence
- Convert the test statistic corresponding to that with one degree of freedom
- Choose the covariate with the highest converted test statistic
- Options to incorporate interaction splits or linear splits

Node impurity measure

- Let $p(j|t)$ be the proportion of class j in node t.
- Node impurity measure

$$i(t) = \phi(p(\cdot|t)) \geq 0,$$

where ϕ is a symmetric function with maximum value $\phi(J^{-1}, \ldots, J^{-1})$ and $\phi(1, 0, \ldots, 0) = \phi(0, 1, 0, \ldots, 0) = \cdots = \phi(0, \ldots, 0, 1) = 0$.

- Entropy:

$$i(t) = -\sum_{j=1}^{J} p(j|t) \log(p(j|t))$$

- Gini index:

$$i(t) = 1 - \sum_{j=1}^{J} p(j|t)^2$$

Choosing split set

- Goodness of a split s as

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R),$$

  where t_L and t_R are the left and right subnodes of t and p_L and p_R are the probabilities of being in those subnodes

- Choose split set to maximize the goodness of a split
- Shortcut algorithm for categorical split
- Break ties

Pruning

- Obtain the resubstitution estimate of expected misclassification cost of a tree T, $M(T)$.
- Cost-complexity function

$$M_\alpha(T) = M(T) + \alpha|\tilde{T}|$$

- For each α, there is a tree T that minimizes the cost-complexity
- Obtain a nested sequence of trees with decreasing α
- Create 10 fake data sets, each with 9 folds of data
- Build a sequence of trees for each fake data set
- Choose α that minimizes the misclassification rate from testing sample
- k-SE rule: Choose a little shorter tree within k standard deviations

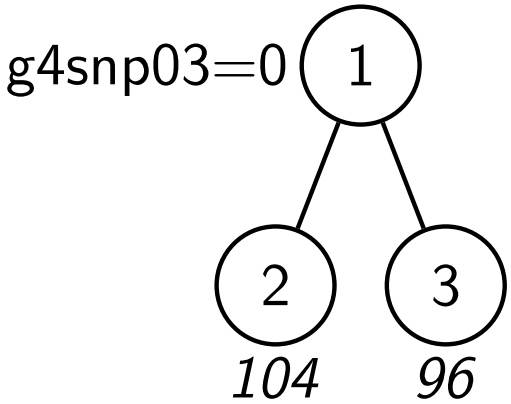
The Gc method

- Response: the interaction between Y and Z:

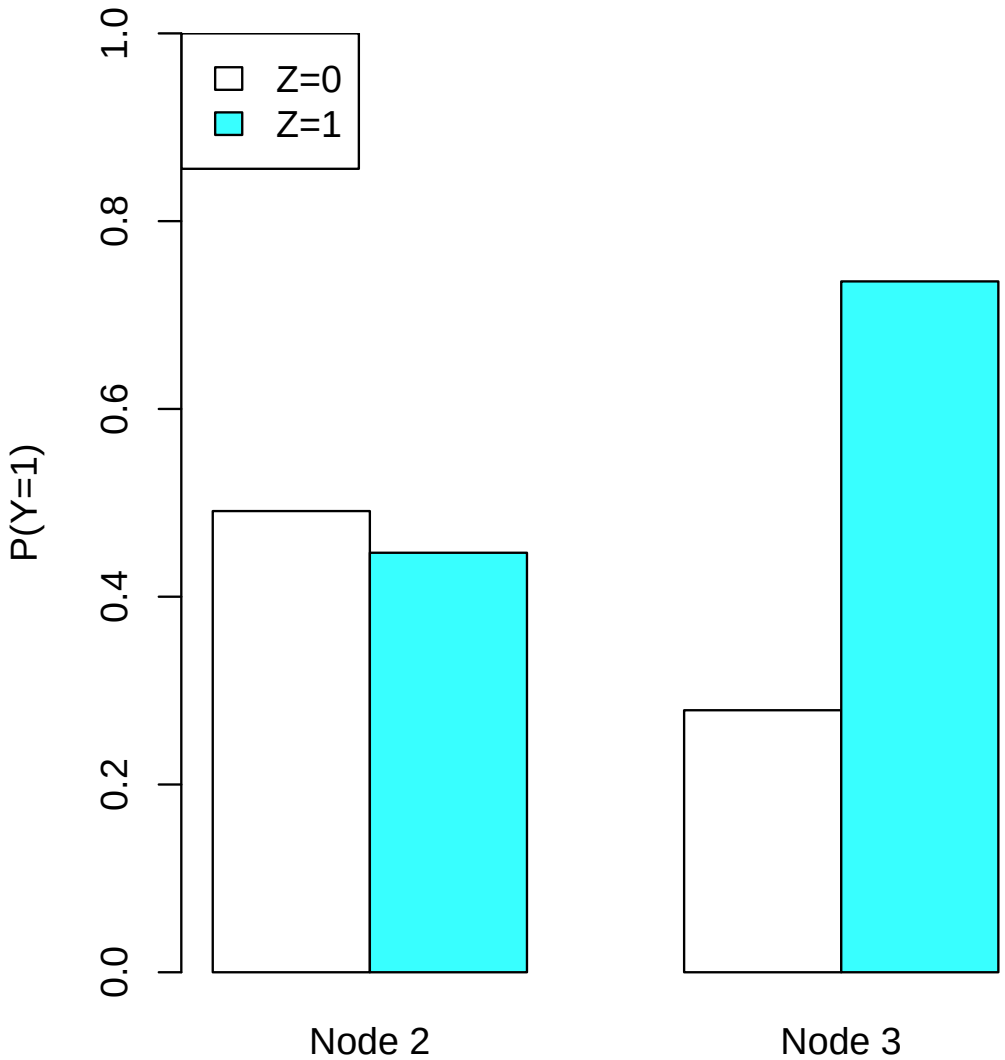
$$V = \begin{cases} 1, & \text{if } \{Y = 1 \text{ and } Z = 1\} \text{ or } \{Y = 0 \text{ and } Z = 0\}, \\ 0, & \text{if } \{Y = 0 \text{ and } Z = 1\} \text{ or } \{Y = 1 \text{ and } Z = 0\} \end{cases}$$

- Covariates as candidate splitting variables
- Adopt GUIDE classification technique
- Aim at predictive variables

An example Gc tree



An example Gc tree



To be continued in Part II

Thank You!