A bias correction algorithm for the Gini variable importance measure in classification trees

Marco Sandri and Paola Zuccolotto*

University of Brescia - Department of Quantitative Methods

C.da Santa Chiara 50 - 25122 Brescia - Italy.

February 16, 2008

Abstract

This paper considers a measure of variable importance frequently used in variable selection methods based on decision trees and tree-based ensemble models, like CART, Random Forests and Gradient Boosting Machine. It is defined as the total heterogeneity reduction produced by a given covariate on the response variable when the sample space is recursively partitioned. Some authors showed that this measure is affected by a bias that, under certain conditions, may have potentially dangerous effects on variable selection. The aim of our work is to present a simple and effective method for bias correction, focusing on the easily generalizable case of the Gini index as a measure of heterogeneity.

Keywords. Variable importance, variable selection, learning ensemble, bias.

*Corresponding author: Paola Zuccolotto, Dipartimento Metodi Quantitativi, Università di Brescia, C.da Santa Chiara 50, 25122 Brescia, Italy. Email: zuk@eco.unibs.it

1 Introduction

Statistical and machine learning techniques for regression and classification based on recursive partitioning are becoming popular tools for variable selection. In recent years, a growing number of papers have appeared in the scientific literature, where applications of CART [Breiman *et al.*(1984)], Random Forests [Breiman(2001a)], Gradient Boosting Machine [Friedman(2001)] and similar methods were proposed for solving problems of variable selection and feature extraction in different research areas ([Bureau *et al.*(2003), Guha and Jurs(2004), Díaz-Uriarte and Alvarez de Andrés(2006), Lunetta *et al.*(2004)], to name a few).

These variable selection methods are usually based on the computation of one or more measures of variable importance (VI henceforth) for each variable in the set $\mathbf{X} = \{X_1, \cdots, X_p\}$ of potential predictors of the response variable Y. For example, in the context of Random Forests, [Breiman(2002)] proposed some measures of VI of the covariate X_i based on two different approaches: (a) the evaluation of the reduction of predictive accuracy after a random permutation of the values assumed by X_i; and (b) the total heterogeneity reduction produced by X_i on the response variable, obtained by adding up all the decreases of the heterogeneity index in the tree nodes where X_i is selected for splitting.

The present paper is focused on the class of VI measures described in (b) above, originally introduced by [Breiman *et al.*(1984)] in the context of CART. Investigations and applications of these measures can be found, sometimes with little modifications, in influential theoretical works [Breiman(2001a), Friedman(2001)] and in many empirical works [Friedman and Meulman(2003), Svetnik *et al.*(2005), Menze *et al.*(2007), De'ath(2007)]. In addition, these measures are often set as the default in many software for data mining, like the `randomForest` package in R ([Breiman *et al.*(2006)]), the `gbm` package in R ([Ridgeway(2007)]), the `boost` Stata command [Schonlau(2005)], the `MART` package in R ([Friedman(2002)]).

Some authors showed that these VI measures are biased in a way that may have, under certain conditions, potentially dangerous effects on variable selection. [Breiman *et al.*(1984)] first noted that they are biased in favor of those variables having more values (i.e., less missing values, more categories or distinct numerical values) and thus offering more splits. This means that variable selection may be affected by covariate characteristics other than information content. Subsequently, [White and Liu(1994), Kononenko(1995), Dobra and Gehrke(2001), Strobl(2005)] investigated in greater detail the nature of the bias in information-based VI measures and elucidated the relation between bias and the number of values of the covariate.

When the Gini gain is used as the splitting criterion for the tree nodes, the resulting total heterogeneity reduction is called the 'Gini VI measure'. [Strobl *et al.*(2007b)] reinterpreted and systematized previous results about this measure and identified three fundamental sources of bias: (a) the bias and (b) the variance of the Gini estimator; and (c) the effects of multiple comparisons.

In recent years, some authors proposed methods for eliminating bias from the Gini VI measure. [Loh and Shih(1997), Kim and Loh(2001)] propose to avoid selection bias by the modification of the algorithm for the construction of a CART. While the common approach simultaneously finds that the covariate and the split point to minimize some node impurity criterion, the authors show that the separation at each node of variable selection from split point selection eliminates bias. In the work of [Strobl *et al.*(2007a)], the alternative implementation of Random Forests developed by [Hothorn *et al.*(2006a)] is proposed as a means for unbiased estimation of the Gini VI measure. When this method is applied using subsampling without replacement, extensive simulations show that resulting VI measures can be reliably used for variable selection even in situations where the potential predictor variables vary in their scale level or their number of categories. Another interesting approach is presented in [Strobl *et al.*(2007b)], where the exact distribution of the maximally selected Gini gain is derived by means of a combinatorial approach and the resulting p-value is suggested as an unbiased split selection criterion in recursive partitioning algorithms.

The aim of the present work is to develop a simple and effective heuristic procedure for the correction of the bias of th the Gini VI measure in tree-based ensemble models. Our method is, to a certain degree, connected to the strategy recently proposed by [Wu *et al.*(2007)].

The paper is organized as follows. In section 2 some preliminary definitions are given. Section 3 discusses the notion of informative and uninformative splits. Section 4 analyzes the central idea behind our bias correction method, and in section 5 an algorithm is derived. Empirical analysis is carried out on simulated and real data (sections 6 and 7). Concluding remarks follow in section 8.

2 Basic notions

Variable selection (or feature subset selection, in the jargon of AI and machine learning) has a traditional close link with the notion of importance (or relevance) of variables. The majority of the techniques developed in this field directly or indirectly make use of VI measures to evaluate the 'goodness' of feature subsets and to select the optimal one.

The concept of importance has been extensively investigated in the philosophical, AI, machine learning and statistical literature. Many authors proposed possible ways to formalize and quantify this notion (see [Bell and Wang(2000)] for a brief overview of the current lines of research). In the present work, we follow [Pearl(1988)] and identify unimportance with conditional independence of random variables. Importance is identified by the negation of unimportance. In other words, let $\mathbf{X}_i = \mathbf{X} - X_i = \{X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_p\}$ be the set of all covariates except X_i, if X_i is stochastically independent of the response variable Y conditionally on $\mathbf{X}_i$, then we say that X_i is unimportant or *uninformative* for the prediction of Y. We write

$X_i \perp_P Y|\mathbf{X}_i$, where P is the joint probability law of Y and X_i given $\mathbf{X}_i$. We call $\mathbf{U}$ the set of all uninformative variables, $\mathbf{U} \equiv \{X_i \in \mathbf{X} \mid X_i \perp_P Y|\mathbf{X}_i,\ i = 1, 2, \ldots, p\}$. The set of *informative* variables $\mathbf{I}$ is thus defined as $\mathbf{I} \equiv \mathbf{X} - \mathbf{U}$. Starting from an alternative definition of importance, [John *et al.*(1994)] proposed to distinguish between strong relevance and weak relevance and [Yu and Liu(2004)] suggested to further differentiate between weakly relevant but non-redundant features and weakly relevant and redundant features. In this study, we group together weakly (redundant and non-redundant) and strongly relevant features in the set $\mathbf{I}$ and we call them informative variables, without distinction.

The notion of importance considered here is closely related to the definition of relevance given by [Bell and Wang(2000)] using the information theoretic formalism: the importance of X_i to Y given $\mathbf{X}_i$ is measured by the (relative) reduction of uncertainty of Y when X_i and $\mathbf{X}_i$ are known, that is

$$VI_P(X_i; Y|\mathbf{X}_i) = \frac{H(Y|\mathbf{X}_i) - H(Y|X_i, \mathbf{X}_i)}{H(Y|\mathbf{X}_i)},$$

where $H(Y|Z)$ is the entropy of Y given Z and P is the joint probability distribution of X_i, $\mathbf{X}_i$ and Y.

Similarly, we define variable importance as the (absolute) reduction of impurity/uncertainty of the response variable Y given by the knowledge of X_i and $\mathbf{X}_i$ and by binary recursive partitions of the sample space. The VI measure originated by this notion consists in the summation, over the set J of nonterminal nodes of the t tree, of the heterogeneity reductions due to the splits made by that variable along the whole tree [Breiman *et al.*(1984)]. It represents the default VI measure in most implementations of classification trees and, with minor modifications, of classification tree ensembles.

Let d_{ij} be the decrease in the heterogeneity index allowed by the X_i variable at node $j \in J$. The X_i variable is used to split at node j if $d_{ij} > d_{kj}$ for all variables in the dataset, $k = 1, 2, \ldots, p$, $k \neq i$. The VI of X_i for the t-th tree is measured by:

$$\widehat{\mathrm{VI}}_{X_i}(t) = \sum_{j \in J} d_{ij}\, I_{ij} \tag{1}$$

where I_{ij} is the indicator function which equals 1 if the i-th variable is used to split at node j and 0 otherwise.

In the context of tree ensemble predictors the VI measure is given by the average of $\widehat{\mathrm{VI}}_{X_i}(t)$ over the set of T trees:

$$\widehat{\mathrm{VI}}_{X_i} = \frac{1}{T}\sum\nolimits_{t=1}^{T} \widehat{\mathrm{VI}}_{X_i}(t) \tag{2}$$

This is the VI measure called 'M4' and proposed by [Breiman(2002)] in Random Forests. In the gradient TreeBoost algorithm of ([Friedman(2001)]), importance of variables is evaluated using a slightly different version of (1) called 'influence of input variables', with d_{ij}^2 in place of d_{ij} and $\widehat{\mathrm{VI}}_{X_i}$ rescaled by assigning a value of 100 to the most influential variable.

Different measures of heterogeneity are available for selecting the best splitting variable. When Y is categorical, the most frequently used is the Gini index. Given a sample from the joint distribution of $(Y, \mathbf{X})$, in the case of a binary response Y and for a given split s of the variable X_i at a given node j, the following contingency table can be specified:

	L	R	
	$X_i \leq s$	$X_i > s$	Σ
$Y = 0$	n_0	$N_0 - n_0$	N_0
$Y = 1$	n_1	$N_1 - n_1$	N_1
Σ	N_L	$N_R = N - N_L$	N

where N is the number of sample units at node j, N_L and N_R the number of units in the left and right nodes after splitting, N_0 and N_1 the number of units with response $Y = 0$ and $Y = 1$, respectively. The empirical Gini heterogeneity index is defined as $\widehat{G} = 2\hat{p}(1 - \hat{p})$, $\hat{p} = N_1/N$ and the impurity reduction (Gini gain) at node j produced by splitting at cutpoint s is given by:

$$d_{ij} = \widehat{\Delta G} = \widehat{G} - \left(\frac{N_L}{N} \widehat{G}_L + \frac{N_R}{N} \widehat{G}_R \right), \tag{3}$$

where $\widehat{G}_L$ and $\widehat{G}_R$ are the Gini indexes calculated in the left and right nodes, respectively.

Following [Dobra and Gehrke(2001)], we state that a split criterion in a tree-based model is unbiased if the selection of a split variable X_i is based only on the importance of X_i, regardless of other characteristics of X_i. Otherwise the split selection criterion is biased. [Strobl *et al.*(2007b)] outline three important sources of bias when the measure given in (3) is used: (a) an estimation bias of the Gini index: $\text{Bias}(\widehat{G}) = -G/N$, where $G = 2p\,(1 - p)$ is the 'true' Gini index; (b) a variance effect of the empirical Gini index: $\text{Var}(\widehat{G}) = 4\,G/N\,(1/2 - G) + O(1/N^2)$; and (c) the effect produced by multiple statistical tests when looking for the best split. The estimation bias (a) leads to a preference of variables with small N, i.e., variables with many missing values. In combination with (a), the variance effect (b) again tends to favor variables with many missing values because $\widehat{\Delta G}$ can take more extreme values. The multiple comparisons effect (c) gives an advantage to covariates with many possible partitions: with many categories (for categorical or ordinal variables), few missing or few ties (for continuous variables).

3 Informative and uninformative splits

As a consequence of recursive partitioning and of the definition of unimportance given above, at each node of a tree-based model, uninformative variables always remain uninformative. Differently, informative variables

can continue to be informative or can become uninformative.

For example, let X be a continuous variable and Y be a binary 0/1 variable, with $P(Y = 1|X > a) = 7/10$, $P(Y = 1|X \leq a) = 1/5$ and $P(X > a) = 1/2$, where a is a given threshold value. In the root node of the tree X is, of course, an informative variable. After the first splitting, the sample space is partitioned in two parts: $X \leq a$ and $X > a$. The Gini gain is given by $\Delta G(X) = G - (P(X \leq a) \cdot G_L + P(X > a) \cdot G_R) = 1/8$. Within the two daughter nodes, X is conditionally independent of Y and thus uninformative.

Within a single node of a tree, each covariate X_i belongs to one of these three classes: (a) informative variables, (b) informative variables which became uninformative by the effect of partitioning and (c) uninformative variables. The finer the partitioning of the sample data, the higher the number of informative variables which became uninformative.

When there is at least one informative variable within a node, the split will be made by using the best variable, in terms of heterogeneity reduction d_{ij}. In other words, only informative variables participate to the 'competition' for the best splitting variable and the heterogeneity reduction d_{ij} of the winner, say X_i, is a direct result of the importance of X_i. We define this circumstance as an *informative split*.

When there are no informative variables within a node, only uninformative variables and/or informative variables which became uninformative participate to the competition for the best split. This is the case of an *uninformative split*. As stated before, because of the bias affecting the Gini gain, in this competition some variables may have an artificial advantage with respect to other variables (e.g. by the action of the estimation effects and/or the multiply comparisons effect). Supposing that the winner is X_i, the heterogeneity reduction d_{ij} added to the computation of $\widehat{\mathrm{VI}}_{X_i}(t)$ in (1) is therefore not attributable to the information content (the 'true' importance) of the variable but depends on the variable's characteristics. In this sense we can say that $\widehat{\mathrm{VI}}_{X_i}$ is biased.

Consider the case where $\mathbf{X} = \{X_1, X_2, X_3\}$ are three continuous and independent covariates and Y is a binary 0/1 dependent variable generated by the following data generating process: $P(X_1 > a) = P(X_2 > b) = 1/2$, $P(Y = 1|X_1 \leq a \cap X_2 \leq b) = P(Y = 1|X_1 \leq a \cap X_2 > b) = 1/5$, $P(Y = 1|X_1 > a \cap X_2 \leq b) = 3/5$ and $P(Y = 1|X_1 > a \cap X_2 > b) = 4/5$, where a and b are threshold values. At the root node, X_1 has 'more power' than X_2 for reducing the heterogeneity of Y by means of a binary split because $\Delta G(X_1) = 1/8 > \Delta G(X_2) = 1/200$. At the root node, X_3 is uninformative, X_1 and X_2 are informative and X_1 will be chosen as splitting variable. This is an informative split. In the daughter node $X_1 > a$, variable X_1 becomes uninformative, X_3 is uninformative and X_2 is informative. Data in this node are therefore partitioned by X_2 and an informative split follows, with $\Delta G(X_2) = 1/50$. In contrast, in the daughter node $X_1 \leq a$, variables X_1, X_2 and X_3 are all uninformative because $\Delta G(X_2) = 0$. The subsequent split of sample data is therefore an uninformative split and is a source of bias for the Gini VI measure.

It follows that $\widehat{\mathrm{VI}}_{X_i}(t)$ can be expressed as the sum of two components:

$$\widehat{\mathrm{VI}}_{X_i}(t) = \sum_{j \in J_{(I)}} d_{ij} I_{ij} + \sum_{j \in J_{(U)}} d_{ij} I_{ij} \equiv \mathrm{VI}_{X_i}(t) + \varepsilon_{X_i}(t) \tag{4}$$

where $J_{(I)}$ and $J_{(U)}$ are the nodes characterized respectively by informative and uninformative splits ($J_{(I)} \cup J_{(U)} = J$, $J_{(I)} \cap J_{(U)} = \emptyset$). $\mathrm{VI}_{X_i}(t)$ is the part of the VI measure attributable to informative splits and directly related to the 'true' importance of X_i. On the contrary, the term $\varepsilon_{X_i}(t) \in \Re^+$ is a noisy component associated with the selection of X_i within uninformative splits and is the source of the bias of $\widehat{\mathrm{VI}}_{X_i}$. The analytical results and the numerical simulations of [Strobl *et al.*(2007b)] indicate that $E[\varepsilon_{X_i}(t)]$ is an increasing function of the number of possible cutpoints of X_i .

4 Bias elimination

The idea behind the algorithm for bias correction proposed in this paper is related to the notion of *phony variables* of [Wu *et al.*(2007)].

Consider the sample data $(\mathbf{Y}, \mathbf{X})$, where $\mathbf{Y}$ is $N \times 1$ and $\mathbf{X}$ is $N \times p$. Suppose that $\mathbf{Z}_r$ is a $N \times p$ matrix of realizations of the p uninformative random pseudocovariates $\mathbf{Z} = \{Z_1, \ldots, Z_p\}$. We add this matrix to the set of p covariates $\mathbf{X}$. Hence, for each covariate X_i, there is now a corresponding pseudovariable Z_i. Let $\widehat{\mathrm{VI}}_{X_i}(\mathbf{Y}, \mathbf{X}, \mathbf{Z}_r)$ be the measure of the importance of X_i, with $i = 1, 2, \ldots, p$, according to (2) and obtained applying the ensemble tree predictor on the augmented dataset $\tilde{\mathbf{X}}_r = (\mathbf{X}, \mathbf{Z}_r)$.

The addition of the set of variables $\mathbf{Z}_r$ produces no effect on informative splits because they are all uninformative. They participate in the competition for the best split in uninformative splits only. Therefore, $\mathrm{VI}_{X_i}(t)$ in formula (4) is not affected by the insertion of $\mathbf{Z}_r$. Modifications occur on $\varepsilon_{X_i}(t)$, the noisy component.

For each covariate X_i and the corresponding pseudovariables Z_i, the following two assumptions are made:

$$\begin{aligned}
&\text{(A1)} \quad E[\widehat{\mathrm{VI}}_{X_i}(\mathbf{Y}, \mathbf{X}, \mathbf{Z})] &&= E[\widehat{\mathrm{VI}}_{Z_i}(\mathbf{Y}, \mathbf{X}, \mathbf{Z})] \qquad \forall i \in \mathbf{U} \\
&\text{(A2)} \quad E[\widehat{\mathrm{VI}}_{X_i}(\mathbf{Y}, \mathbf{X}, \mathbf{Z})] &&= E[\mathrm{VI}_{X_i}(\mathbf{Y}, \mathbf{X}, \mathbf{Z})] + E[\widehat{\mathrm{VI}}_{Z_i}(\mathbf{Y}, \mathbf{X}, \mathbf{Z})] \quad \forall i \in \mathbf{I}
\end{aligned}$$

Assumption (A1) states that each unimportant variable and the corresponding pseudovariable have the same expected VI measure; (A2) states that the expected VI measure of each important variable is given by the sum of a component originated by its 'true' importance and the expected VI measure of its corresponding pseudovariable. From equation (4), these assumptions are equivalent to the condition $E[\widehat{\mathrm{VI}}_{Z_i}(\mathbf{Y}, \mathbf{X}, \mathbf{Z})] =$

$E[\varepsilon_{X_i}]$. In other words, for each (informative or uninformative) covariate X_i, (A1) and (A2) require the existence of a corresponding random pseudovariable Z_i that has the same probability of X_i to win the competition within uninformative splits.

Thus, if (A1) and (A2) are verified, after an adequate number of replications R, the quantity:

$$\widehat{\mathrm{VI}}^{*}_{X_i} = \frac{1}{R} \cdot \sum_{r=1}^{R} \left[\widehat{\mathrm{VI}}_{X_i}(\mathbf{Y}, \mathbf{X}, \mathbf{Z}_r) - \widehat{\mathrm{VI}}_{Z_i}(\mathbf{Y}, \mathbf{X}, \mathbf{Z}_r) \right] \qquad i = 1, 2, \ldots, p \tag{5}$$

can be used as an unbiased VI measure for X_i.

5 The algorithm

Assumptions (A1) and (A2) considered in the previous section provide guidance for generating pseudovariables. The objective is to generate pseudovariables so that their average importance is equal to the bias of the corresponding covariates. We are aware that these assumptions are almost certain to be violated and in any case are virtually unverifiable. Thus, we recognize that our method is only approximate and regard (A1) and (A2) more as guiding principles rather than as crucial mathematical conditions justifying the method.

We have studied two methods to generate pseudovariables according to the above assumptions. In the first method, each Z_i is obtained by randomly permuting the elements of the single X_i. In the second, the N rows of $\mathbf{Z}_r$ are obtained by randomly permuting the rows of $\mathbf{X}$. In both methods, the pseudovariables are stochastically independent of Y and of covariates $\mathbf{X}$; each Z_i has the same distribution, the same number of missing values and the same number of possible cutpoints of the corresponding X_i. In addition, in the second method the sample multiple relationships existing among the p variables in $\mathbf{X}$ are preserved when creating the corresponding pseudovariables in $\mathbf{Z}_r$. Our simulation studies (not reported here) show a significant advantage when adopting the second method. We also compared sampling with and without replacement in the construction of $\mathbf{Z}_r$. Simulations shows that sampling without replacement moderately outperforms the other method.

The proposed algorithm for bias correction can be summarized as follows:

(1) Generate $\mathbf{Z}_r$ according to one of the methods described above.

(2) Apply the ensemble tree prediction method using $\mathbf{Y}$ as dependent variable and $\tilde{\mathbf{X}}_r = (\mathbf{X}, \mathbf{Z}_r)$ as the set of explanatory variables.

(3) Applying equation (2), compute $\widehat{\mathrm{VI}}_{X_i}$ and $\widehat{\mathrm{VI}}_{Z_i}$ for each independent variable X_i and each pseudovariable Z_i $(i = 1, 2, \ldots, p)$.

(4) Repeat steps (1), (2) and (3) R times.

(5) Calculate the value of $\widehat{\mathrm{VI}}^*_{X_i}$, $i = 1, 2, \ldots, p$, given in (5).

6 Simulation studies

In this section, the effectiveness of the proposed algorithm is investigated by a set of numerical simulations. We consider a binary 0/1 response variable Y and a set $\mathbf{X} = \{B, O6, O11, N6, N11, C\}$ of mutually independent covariates: a binary variable, an ordinal variable with 6 categories, an ordinal variable with 11 categories, a nominal variable with 6 categories, a nominal variable with 11 categories, and a numerical variable with a standard normal distribution $N(0, 1)$, respectively. The sample size is $N = 250$. For each generated sample, categorical variables have an equal number of units in their categories. For example, in each sample of 250 units, B has absolute frequencies $n_i = 250/2 = 125$, $i = 1, 2$.

We consider 4 cases:

Null case: all covariates are equally uninformative;

Power case I: covariate B is informative ($B \in \mathbf{I}$); the data generating process is a logistic regression model with $P(Y = 1|B = x) = e^{\beta x}/(1 + e^{\beta x})$, where $\beta = 0.8$ and B assumes values -1 and 1;

Power case II: $\{B, O6, C\} \in \mathbf{I}$; the data generating process is a logistic regression model $P(Y = 1 \mid [B, 06, C] = \mathbf{x}) = e^{\mathbf{x}\beta}/(1 + e^{\mathbf{x}\beta})$, where $\beta = [0.8, 0.8, 0.8]$; the three variables have been opportunely standardized;

Power case III: the interaction of B and C is informative; the data generating process is defined as follows: $P(Y = 1|B = 1 \cap C > 0) = 0.75$ and $P(Y = 1|B = -1 \cup C \leq 0) = 0.25$.

We apply our bias correction method on two tree-based ensemble models: Random Forests and Gradient Boosting Machine. We use the R packages `randomForest` and `gbm`. The number of trees for the two models is $T = 1000$ and the minimum number of observations in the trees terminal nodes is 10. In `randomForest` the number of variables randomly sampled as candidates at each split is `mtry` $= 3$. In `gbm` the maximum depth of variable interactions is `interaction.depth` $= 2$ and the learning rate is `shrinkage` $= 0.005$. We set the number of replications (without replacement) R defined in (5) to $R = 100$. For each simulation study the number of samples analyzed is $S = 100$.

For comparison purpose, we compute the Gini importance of the 6 variables by the `cforest` command of the `party` package of R ([Hothorn *et al.*(2006a)]). This command implements Random Forests and bagging ensemble algorithms by utilizing conditional inference trees as base learners. The test statistic used is `quad`, a univariate test statistics based on a quadratic form. The distribution of the test statistic has been computed by the Bonferroni-adjusted method. The number of trees is $T = 1000$. Because the `varimp` command of this package calculates only the measure of importance based on the mean decrease accuracy, we developed an R function for the calculation of the Gini index in this class of models.

The results of our simulation studies are shown in figures (1), (2), (3) and (4), where the distribution of the S values of $\widehat{\mathrm{VI}}^*_{X_i}$ are visualized by boxplots with whiskers ranging from 2.5-th to 97.5-th quantile. In the (a) part of each figure we show the distributions of the 6 Gini importances for Conditional Inference Random Forests. The gray boxes in part (b) refer to the raw Gini importances in Random Forests and white boxes to the corresponding bias-corrected importances. Similarly, gray and white boxes in part (c) refer to raw and bias-corrected importances calculated by the Gradient Boosting Machine.

[Figures (1), (2), (3) and (4) approximately here]

The simulation and benchmarking experiments in this section support two conclusions: (1) the proposed method is effective in removing bias from the Gini VI measure, and (2) the capability of our algorithm of identifying informative and uninformative variables is comparable to that of Conditional Inference Random Forests. It is apparent that the distributions of the bias-corrected VI measures show different patterns for informative and uninformative covariates. In the case of an uninformative variable, the distribution is centered around 0, showing that, on average, the variable has no power in reducing the heterogeneity of Y. On the contrary, the most part (95%) of the values of the bias-corrected measures of the informative variables are positive and the distribution is centered away from zero.

7 Application to real-life datasets

In this section we further investigate the performances of the proposed method by means of 4 real-life datasets.[1]

The first dataset (ulcer data) contains rebleeding (13.3%) and no rebleeding of 738 patients with bleeding ulcers. There are 32 covariates related to patient history, magnitude of bleeding and endoscopic findings: 19 binary, 3 ordinal, 3 nominal and 7 numerical variables. The dataset is described in ([Guglielmi *et al.*(2002)]). The aim of this study was to identify risk factors for recurrence of hemorrhage. The authors estimated a logistic regression model and selected a set of informative variables by means of statistical evidences (AIC stepwise) and medical experience: ulcer size, systolic blood pressure (`sbp`), Forrest index, ulcer location, hematemesis, liver cirrhosis (`livcir`) and recent surgery (`recsurg`). Conditional Inference Random Forests and the proposed bias-correction method substantially confirm these findings (see Fig.5) and suggest two additional informative covariates: `shock` and `symptoms`. This fact can be explained considering that these covariates are strongly correlated to other informative variables: a low level of the systolic blood pressure

[1] We warn the reader that the figures of this section are substantially different from those drawn using simulated data. the boxplots here display the distribution of the R values of the difference $\widehat{\mathrm{VI}}_{X_i}(\mathbf{Y},\mathbf{X},\mathbf{Z}_r) - \widehat{\mathrm{VI}}_{Z_i}(\mathbf{Y},\mathbf{X},\mathbf{Z}_r)$ given in (5) for bias-corrected VIs, while in section 6 the boxplots refer to the distributions of the mean values of the VIs for the S simulated samples.

is one of the clinical sign of shock and symptoms is a nominal variables with the following 4 categories: melena, hematemesis, hematemesis+melena, other. Furthermore, in 4 numerical covariates (age, heart rate `hr`, haematocrit `ht` and haemoglobin level `hb`) the effects of the bias of the Gini measure appeared more pronounced and dangerous. These covariates show high values of the raw VI and, after bias removal, they became uninformative or slightly informative.

The Hepatitis dataset is available at the UCI Machine Learning Repository and it contains survival (79.3%) and nonsurvival of 155 chronic hepatitis patients with 19 covariates (13 binary and 6 numerical). The covariates protime, alk phosphate and albumin are characterized by high percentage of missing values: 43%, 19% and 11%, respectively. Using a three-steps procedure based on forward logistic regression, the Gregory's rule described in [Gong (1986)] identified 4 predictive covariates: albumin, spiders, bilirubin and sex. [Kim and Loh(2001)] estimated classification trees with binary and multiway splits (CART, QUEST, CRUISE 1D and CRUISE 2D) finding that the top three predictors are protime, bilirubin, and albumin. [Breiman(2001b)] calculated variable importance in Random Forests by the reduction of predictive accuracy after random permutation of covariates and concluded that virtually all the predictive capability is provided by a single variable (either ascites or albumin). Our analysis evidences that the most important covariates are albumin, protime, ascites, histology and bilirubin (Fig.6). These results partially confirm the findings of the preceding analyses and shed light on the role of a neglected covariate: histology. Conditional Inference Forests and the unbiased Gini VI in Random Forests show that the importance of this variable is comparable to that of ascites and bilirubin. The estimation of the out-of-bag prediction error of a Random Forest with and without histology (the other covariates considered are albumin, ascites and bilirubin) indicate a moderate predictive power of this covariate: 18.7% vs. 20.0%. Another interesting result evidenced by our analysis is the marked level of bias in the Gini VI measure for the numerical covariates age, agot and alk phosphate. Evaluating the Gini VI measure without bias correction would lead to a misleading attribution of importance to these uninformative covariates.

[Figures (5) and (6) approximately here]

8 Concluding remarks

The Gini VI measure is frequently used in classification trees and in tree-based learning ensembles. This measure has long been recognized by many authors to be affected by bias. The main consequence of this systematic deviation is that variable selection may be influenced by covariate characteristics other than information content. In spite of this potentially dangerous effect, the Gini index is often set as the default VI measure in many software for data mining, without any correction.

In the present paper we proposed a bias correction heuristic strategy and investigated its performances both on simulated and real data. The idea behind the algorithm is to artificially add a set of uninformative pseudovariables to the original data whose VIs, under certain conditions, can approximate the unknown bias. The results show that the method is capable of efficiently removing bias in many practical circumstances. In addition, there is substantial agreement between our algorithm and the unbiased variable selection method of [Hothorn *et al.*(2006a)].

Although the paper is focused on classification trees with the Gini gain as splitting criterion, preliminary investigations give indications that the proposed strategy is also effective in regression trees and can be extended to other heterogeneity measures (e.g. entropy-based measures). Another attractive advantage of our method is that it can be easily integrated with minor efforts into any traditional algorithm for recursive partitioning and might thus prove manageable and useful to applied scientists.

9 Acknowledgements

We are grateful to the Editor and two anonymous referees for their valuable comments and suggestions for improving the quality of the paper. A special thank to Roberto Perli and Samantha Sartori for their help. The usual disclaimer applies.

References

[Bell and Wang(2000)] Bell, D. and Wang, H. (2000): A formalism for relevance and its application in feature subset selection. *Machine Learning*, 4(2), 175–195

[Breiman(2001a)] Breiman, L. (2001): Random Forests. *Machine Learning*, 45, 5–32.

[Breiman(2001b)] Breiman, L. (2001): Statistical Modeling: The Two Cultures *Statistical Science*, 16 (3), 199–231.

[Breiman(2002)] Breiman, L. (2002): Manual on setting up, using, and understanding Random Forests v3.1. *Technical Report*, ftp://ftp.stat.berkeley.edu/pub/users/breiman/Using_random_forests_v3.1.pdf.

[Breiman *et al.*(1984)] Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984): *Classification and Regression Trees*. Chapman & Hall, London.

[Breiman *et al.*(2006)] Breiman L, Cutler A, Liaw A, Wiener M (2006): Breiman and Cutlers Random Forests for Classification and Regression. R package version 4.5-18 http://cran.r-project.org/doc/packages/randomForest.pdf

[Bureau *et al.*(2003)] Bureau, A., Dupuis, J., Hayward, B., Falls, K. and Van Eerdewegh, P. (2003): Mapping complex traits using Random Forests. *BMC Genetics*, 4(Suppl.1):S64, http://www.biomedcentral.com/1471-2156/4/s1/S64

[Cummings and Myers(2004)] Cummings M.P. and Myers D.S. (2004): Simple statistical models predict C-to-U edited sites in plant mitochondrial RNA. *BMC Bioinformatics*, 5, 132, http://www.biomedcentral.com/1471-2105/5/132

[De'ath(2007)] De'ath, G. (2007): Boosted trees for ecological modeling and prediction. *Ecology*, 88(1), 243–251.

[Díaz-Uriarte and Alvarez de Andrés(2006)] Díaz-Uriarte, R. and Alvarez de Andrés, S. (2006): Gene selection and classification of microarray data using random forest. *BMC Genetics*, 7:3, http://www.biomedcentral.com/1471-2105/7/3

[Dobra and Gehrke(2001)] Dobra A, Gehrke J (2001): Bias Correction in Classification Tree Construction. In *Proceedings of the Seventeenth International Conference on Machine Learning, Williams College, Williamstown, MA, USA*. Edited by Brodley CE, Danyluk AP, pp. 90-97.

[Friedman(2001)] Friedman, J.H. (2001): Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29, 1189–1232.

[Friedman(2002)] Friedman, J.H. (2002): Tutorial: getting started with MART in R. *Technical Report*, Standford University, http://www-stat.stanford.edu/ jhf/r-mart/tutorial/tutorial.pdf.

[Friedman *et al.*(2001)] Friedman, J.H., Hastie, T. and Tibshirani, R. (2001): *The Elements of Statistical Learning; Data Mining, Inference and Prediction*. Springer, New York.

[Friedman and Meulman(2003)] Friedman, J.H. and Meulman, J.J. (2003): Multiple additive regression trees with application in epidemiology. *Statistics in Medicine*, 22, 1365–1381.

[Gong (1986)] Gong, G. (1986): Cross-Validation, the Jackknife, and the Bootstrap: Excess Error Estimation in Forward Logistic Regression. *Journal of the American Statistical Association*, 81(393), 108–113.

[Guglielmi *et al.*(2002)] Guglielmi, A., Ruzzenente, A., Sandri, M., Kind, R., Lombardo, F., Rodella, L., Catalano, F., De Manzoni, G. and Cordiano, C. (2002): Risk assessment and prediction of rebleeding in bleeding gastroduodenal ulcer. *Endoscopy*, 34, 771–779.

[Guha and Jurs(2004)] Guha, R. and Jurs, P.C. (2004): Development of Linear, Ensemble, and Nonlinear Models for the Prediction and Interpretation of the Biological Activity of a Set of PDGFR Inhibitors. *J. Chem. Inf. Comput. Sci.*, 44, 2179–2189.

[Hothorn *et al.*(2006a)] Hothorn, T., Hornik, K. and Zeileis, A. (2006): Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, 15 (3), 651–674.

[Hothorn *et al.*(2006b)] Hothorn, T., Hornik K., Zeileis, A. (2006): party: A Laboratory for Recursive Part(y)itioning. R package version 0.9-11. http://cran.r-project.org/doc/vignettes/party/party.pdf

[John *et al.*(1994)] John, G.H., Kohavi, R., and Pfleger, K. (1994): Irrelevant features and the subset selection problem. In: Cohen, W.W. and Hirsch, H. (eds), *Proceedings of the 11th international conference on machine learning*, Morgan Kaufmann, New Brunswick, NJ, 121-129.

[Kim and Loh(2001)] Kim H., Loh W. (2001): Classification Trees with Unbiased Multiway Splits. *Journal of the American Statistical Association*, 96, 589-604.

[Kononenko(1995)] Kononenko I. (1995): On Biases in Estimating Multi-Valued Attributes. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Montréal, Canada.* Edited by Mellish C., 1034-1040.

[Loh and Shih(1997)] Loh, W.-Y., Shih, Y.-S. (1997): Split selection methods for classification trees. *Statististica Sinica*, 7, 815–840.

[Lunetta *et al.*(2004)] Lunetta, K.L., Hayward, B.L., Segal, J. and Van Eerdewegh, P. (2004): Screening large-scale association study data: exploiting interactions using random forests. *BMC Genetics*, 5:32, http://www.biomedcentral.com/1471-2156/5/32

[Menze *et al.*(2007)] Menze, B.H., Petrich, W. and Hamprecht F.A. (2007): Multivariate feature selection and hierarchical classification for infrared spectroscopy: serum-based detection of bovine spongiform encephalopathy. *Anal. Bioanal. Chem.*, doi:10.1007/s00216-006-1070-5.

[Pearl(1988)] Pearl J. (1988): Probabilistic reasoning in intelligent systems: networks of plausible inference. *Morgan Kaufmann Publishers, Inc.*, San Francisco, California.

[Ridgeway(2007)] Ridgeway, G. (2007): Generalized Boosted Models: A guide to the gbm package. http://i-pensieri.com/gregr/papers/gbm-vignette.pdf

[Ripley(1996)] Ripley, B. (1996): Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge

[Schonlau(2005)] Schonlau, M. (2005): Boosted Regression (boosting): A Tutorial and a Stata plugin. *The Stata Journal*, 5(3), 330–354.

[Strobl(2005)] Strobl, C. (2005): Statistical Sources of Variable Selection Bias in Classification Trees Based on the Gini Index. *Technical Report*, SFB 386, http://epub.ub.uni-muenchen.de/archive/00001789/01/paper_420.pdf

[Strobl *et al.*(2007a)] Strobl, C., Boulesteix, A.-L., Zeileis, A. and Hothorn, T. (2007): Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics*, 8:25, doi:10.1186/1471-2105-8-25

[Strobl *et al.*(2007b)] Strobl, C., Boulesteix, A.-L. and Augustin, T.(2007): Unbiased split selection for classification trees based on the Gini Index. *Computational Statistics & Data Analysis*, doi:10.1016/j.csda.2006.12.030

[Svetnik *et al.*(2005)] Svetnik, V., Wang, T., Tong, C., Liaw, A., Sheridan, R.P. and Song Q. (2005): Boosting: An Ensemble Learning Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Model.*, 45, 786–799

[White and Liu(1994)] White, A.P. and Liu, W.Z. (1994): Bias in Information-Based Measures in Decision Tree Induction. *Machine Learning*, 15, 321–329.

[Wu *et al.*(2007)] Wu, Y., Boos, D.D. and Stefanski, L.A.(2007): Controlling Variable Selection by the Addition of Pseudovariables. *Journal of the American Statistical Association*, 102 (477), 235–243.

[Yu and Liu(2004)] Yu, L. and Liu, H. (2004): Efficient Feature Selection via Analysis of Relevance and Redundancy. *Journal of Machine Learning research*, 5, 1205-1224.

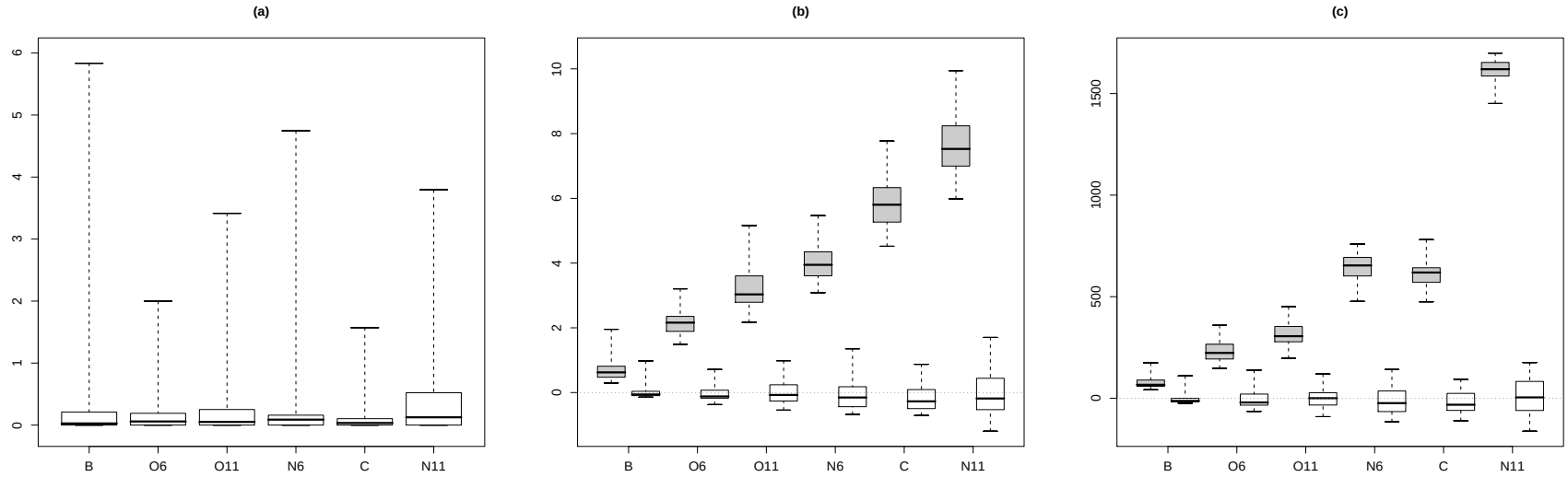

Figure 1: Null case: all uninformative variables. (a) Gini importance in Conditional Inference Forests, (b) in Random Forests and (c) in Generalized Boosted Regression Models. Gray boxes indicate raw (biased) importance and white boxes bias-corrected importance.

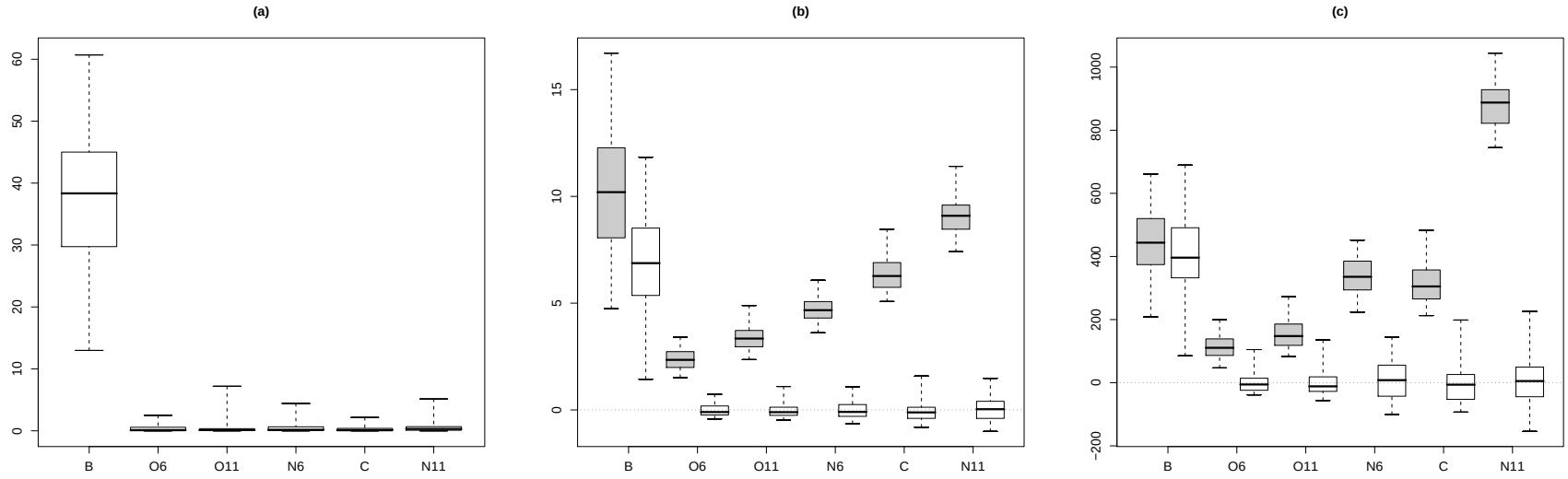

Figure 2: Power case I: the only important variable is B. (a) Gini importance in Conditional Inference Forests, (b) in Random Forests and (c) in Generalized Boosted Regression Models. Gray boxes indicate raw (biased) importance and white boxes bias-corrected importance.

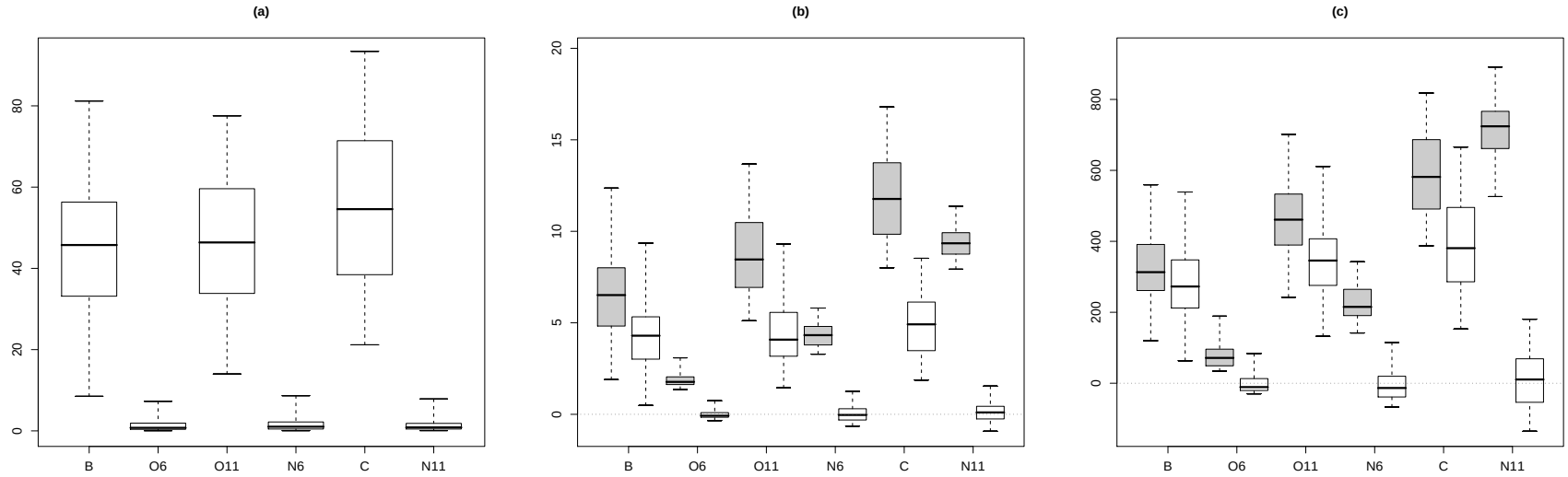

Figure 3: Power case II: three important variables (B, O11, C). (a) Gini importance in Conditional Inference Forests, (b) in Random Forests and (c) in Generalized Boosted Regression Models. Gray boxes indicate raw (biased) importance and white boxes bias-corrected importance.

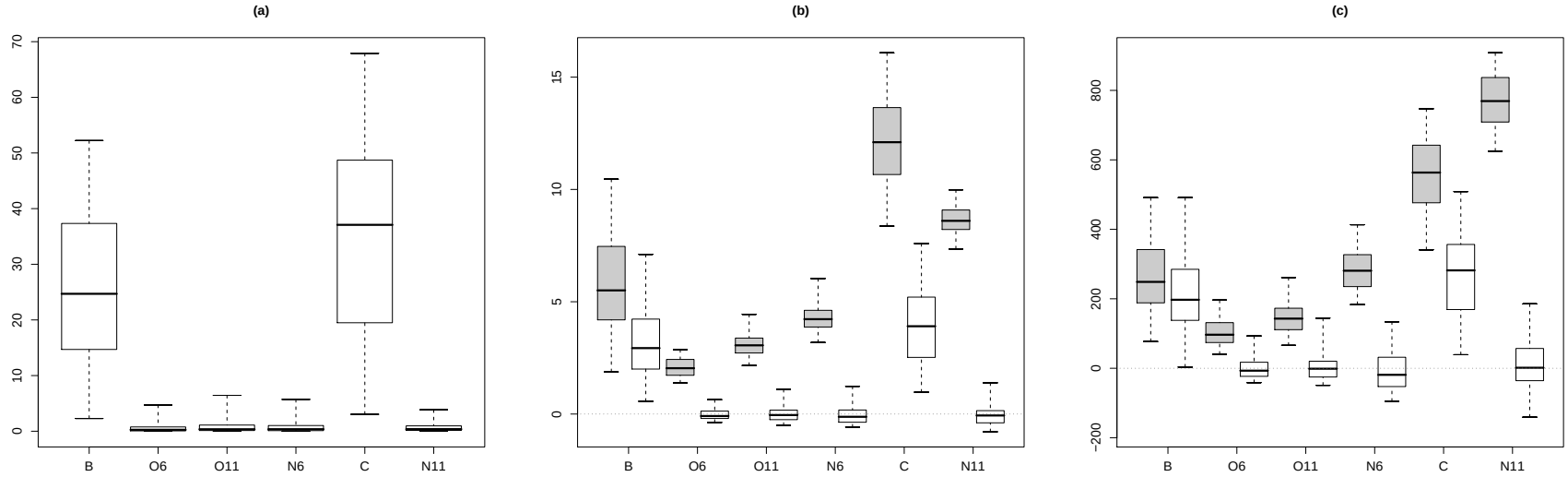

Figure 4: Power case III: the important variable is the interaction between covariates B and C. (a) Gini importance in Conditional Inference Forests, (b) in Random Forests and (c) in Generalized Boosted Regression Models. Gray boxes indicate raw (biased) importance and white boxes bias-corrected importance.

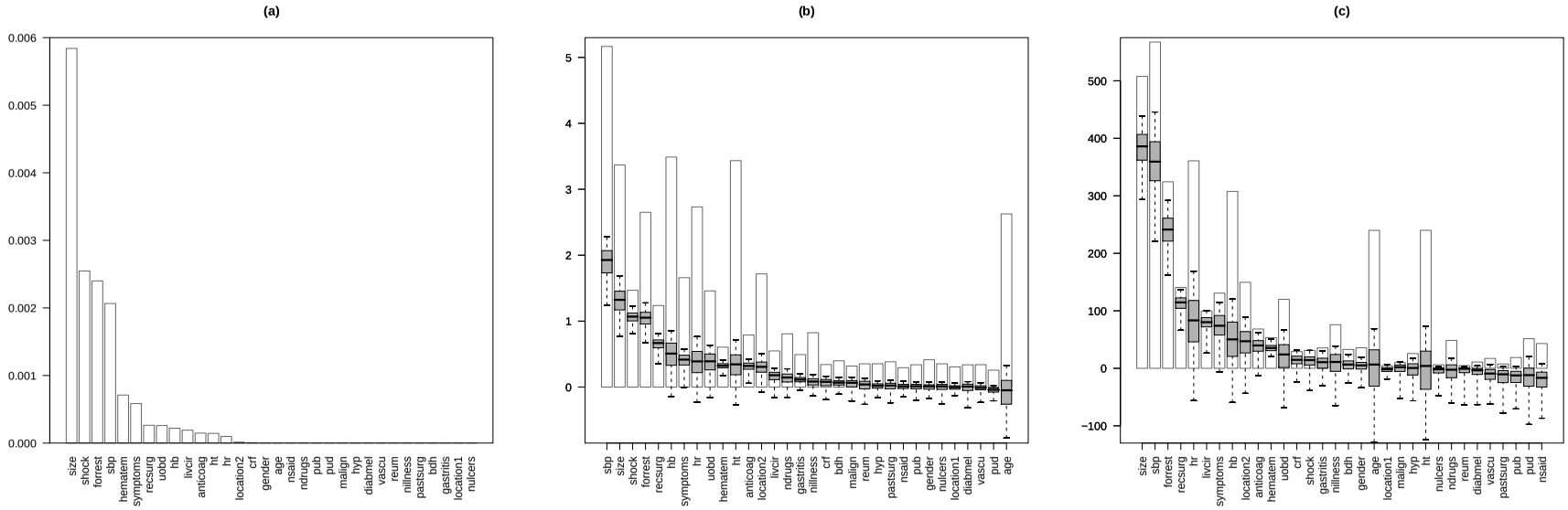

Figure 5: Ulcer data: (a) Gini importance in Conditional Inference Forests, (b) in Random Forests and (c) in Generalized Boosted Regression Models. In (b) and (c) figures, bars indicate raw (biased) importance and gray boxes bias-corrected importance.

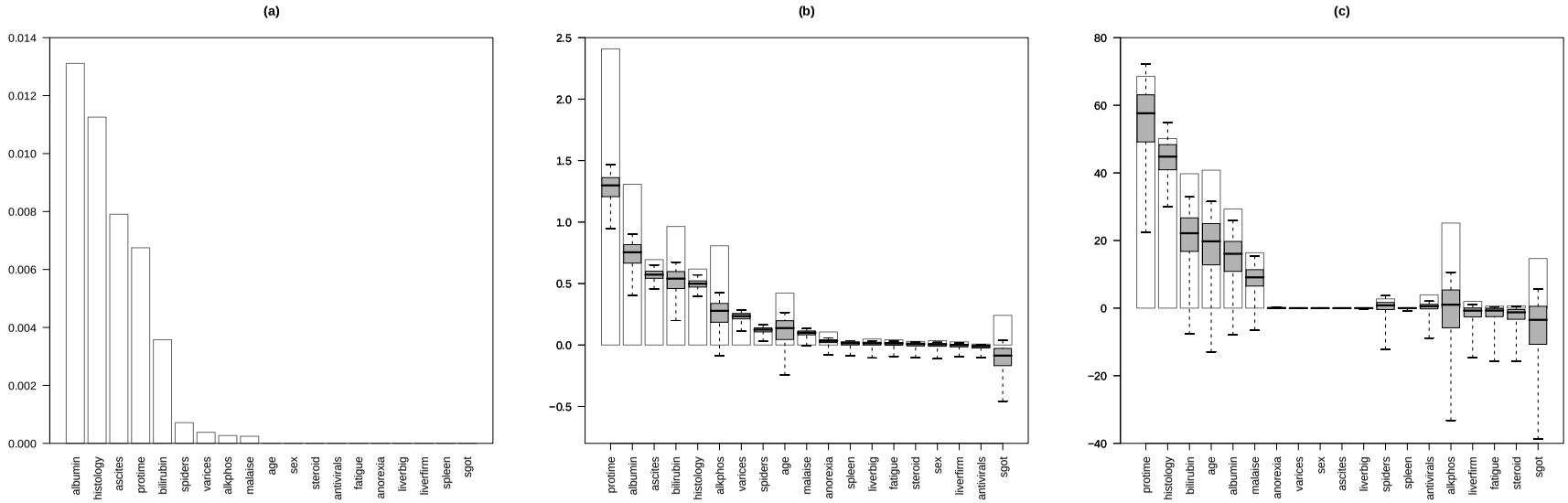

Figure 6: Hepatitis data: (a) Gini importance in Conditional Inference Forests, (b) in Random Forests and (c) in Generalized Boosted Regression Models. In (b) and (c) figures, bars indicate raw (biased) importance and gray boxes bias-corrected importance.