Stat Comput (2010) 20: 393–407
DOI 10.1007/s11222-009-9132-0

Analysis and correction of bias in Total Decrease in Node Impurity measures for tree-based algorithms

Marco Sandri · Paola Zuccolotto

Received: 29 September 2008 / Accepted: 11 May 2009 / Published online: 27 May 2009

Abstract Variable selection is one of the main problems faced by data mining and machine learning techniques. These techniques are often, more or less explicitly, based on some measure of variable importance. This paper considers Total Decrease in Node Impurity (TDNI) measures, a popular class of variable importance measures defined in the field of decision trees and tree-based ensemble methods, like Random Forests and Gradient Boosting Machines. In spite of their wide use, some measures of this class are known to be biased and some correction strategies have been proposed. The aim of this paper is twofold. Firstly, to investigate the source and the characteristics of bias in TDNI measures using the notions of informative and uninformative splits. Secondly, a bias-correction algorithm, recently proposed for the Gini measure in the context of classification, is extended to the entire class of TDNI measures and its performance is investigated in the regression framework using simulated and real data.

Keywords Impurity measures · Ensemble learning · Variable importance

1 Introduction

In the last decades, with the proliferation of large datasets, the problem of variable selection has gained increasing attention in the field of data analysis. Given a large set of observed covariates that describe the phenomenon being studied, the researcher often needs to identify the subset of informative (predictive) variables and to set uninformative (noisy) variables apart. A preliminary variable selection is a fundamental and crucial step in model building, whatever the approach to model the phenomenon might be.

Many of the variable selection methods are directly or indirectly based on the assumption that, given the set $\mathbf{X} = \{X_1, \ldots, X_p\}$ of potential predictors for a response variable Y, an importance or relevance μ_i can be defined for each covariate X_i in terms of prediction/explanation of Y, and this measure can be evaluated from data using some variable importance estimator $\widehat{\mathrm{VI}}_i$. The notion of importance has been widely investigated in philosophical, AI, machine learning and statistical literatures. Several attempts have been proposed to formalize and quantify this notion. See Bell and Wang (2000) for a brief overview of the current lines of research and van der Laan (2005) for a novel approach. In the present paper, following Pearl (1988), we start by identifying unimportance with conditional independence of random variables and importance with the negation of unimportance.

This paper focuses on measures of variable importance developed in the area of tree-based ensemble methods. Decision trees model data by partitioning the feature space into a set of disjoint rectangles and then fitting a simple model (e.g. a constant) to each one. Classification and Regression Trees (CART) were introduced by Breiman, Friedman, Olshen and Stone in 1984 and are a milestone in this field. Since then, a great development has been proposed in several disciplines (Murthy 2004). One of the main problems of tree predictors is model instability, defined as the existence of many different models, distant in terms of form and interpretation, that have about the same training or test set error (Breiman 1996, 2001b).

M. Sandri · P. Zuccolotto (✉)
Department of Quantitative Methods, University of Brescia,
C.da Santa Chiara 50, 25122 Brescia, Italy
e-mail: zuk@eco.unibs.it

Ensemble learning is a class of methods developed for reducing model instability and improving the accuracy of a predictor through the aggregation of several similar predictors. Each ensemble member is constructed by a different function of the input covariates. Ensemble prediction is obtained by the linear combination of the predictions of ensemble members. Ensembles can be built using different prediction methods, i.e. using different base learners as ensemble members. An interesting proposal uses CART as base learners. Typically, such aggregation neutralizes the effects of tree instability and reaches greater accuracy by reducing either the bias or the variance of a single classifier (Bühlmann and Yu 2002). Popular examples of tree-based ensembles are Random Forests (RF, Breiman 2001a) and Gradient Boosting Machine (GBM, Friedman 2001).

The first approach to variable importance (VI, henceforth) measurement in tree-based predictors dates back to the cited book of Breiman et al. (1984), where an interesting and effective notion of variable importance was proposed. The importance μ_i of a covariate X_i is defined as the total decrease of heterogeneity of the response variable Y given by the knowledge of $\mathbf{X} = \{X_1, \ldots, X_p\}$ when the feature space is partitioned recursively. The VI measure originated by this notion is obtained by summing up all the decreases of the heterogeneity index in the nodes of the tree. This class of measures is called Total Decrease in Node Impurity (TDNI henceforth) and, with few changes, is used in many tree-based ensemble methods (see e.g. Breiman 2002; Friedman 2001). It is also available in many softwares 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).

In spite of their wide use, some TDNI measures are known to be biased. Breiman et al. (1984) noted that they tend to favor covariates having more values (i.e. less missing values, more categories or distinct numerical values) and thus offering more splits. White and Liu (1994), Kononenko (1995) and Dobra and Gehrke (2001) investigated in greater detail the nature of bias and elucidate the relation between bias and the number of values of covariates. Strobl (2005), Strobl et al. (2007a) and Sandri and Zuccolotto (2008) focused attention on the bias of the Gini variable importance measure (hereafter Gini TDNI), a measure frequently used in classification trees and based on the adoption of the Gini gain as a splitting criterion of tree nodes.

Several methods have been proposed in the last decade to eliminate bias from the Gini TDNI. Loh and Shih (1997) and Kim and Loh (2001) proposed modifying the algorithm for the construction of classification trees in order to avoid selection bias. The authors showed that bias can be eliminated by separating variable selection from split point selection at each node. In the work of Strobl (2005), an unbiased estimation of the Gini TDNI was found in Conditional Random Forests, a new class of RF developed by Hothorn et al. (2006). Strobl et al. (2007a) derived the exact distribution of the maximally selected Gini gain by means of a combinatorial approach and the resulting p-value is suggested as an unbiased split selection criterion in recursive partitioning algorithms. The heuristic correction strategy proposed by Sandri and Zuccolotto (2008) is based on the introduction of a set of random pseudocovariates in the $\mathbf{X}$ matrix. The authors showed that the algorithm can efficiently remove bias from Gini TDNI in RF and GBM.

The aim of this paper is twofold. Firstly, to investigate the source and the characteristics of bias in TDNI measures by the introduction of the notions of informative and uninformative splits, showing its connections with the level of covariates' measurement and with the number of uninformative splits. Secondly, to generalize and extend the domain of applicability of the correction algorithm of Sandri and Zuccolotto (2008) to the class of TDNI measures, evaluating its performance on simulated and real data in regression problems when the residual sum of squares is used as splitting criterion for the tree nodes.

The paper is organized as follows. In Sect. 2 we define the class of TDNI measures and the corresponding estimators for single trees and tree-based ensembles. In Sect. 3 we define the crucial notion of informative and uninformative splits and show that uninformative splits are the main source of bias for TDNI measures. Section 4 investigates the relationship existing between bias of TDNI measures and the level of covariates' measurement from a theoretical point of view and by means of some simulation experiments. In Sect. 5 the bias-correction strategy for classification problems proposed by Sandri and Zuccolotto (2008) is recalled and extended to the class of TDNI measures. The performance of the method is tested on simulated (Sect. 5) and real data (Sect. 6). Section 7 concludes.

2 Characterization of TDNI measures

Let $(Y, \mathbf{X}) : \Omega \to (D_Y \times D_{X_1} \times \cdots \times D_{X_p}) \equiv \mathcal{D}$ be a vector random variable defined on a probability space $(\Omega, \mathcal{F}, P)$, where $\mathbf{X} = \{X_1, \ldots, X_p\}$ is a set of covariates and Y a response variable. A tree-structured binary recursive partitioning algorithm yields a hierarchical partition of the domain $\mathcal{D}$ into J disjoint (hyper-)rectangles $R_j \subset \mathcal{D}$, $j = 1, 2, \ldots, J$. Each rectangle is generated in $\mathcal{D}$ by splitting a parent rectangle into two parts by a binary split of the domain of a covariate X_i. Therefore, a rectangle can be described by the set of covariates and splits used to generate it. If Y, X_1 and X_2 are three numerical real-valued random variables, a rectangle could be for example given by $R_j = \{(y, \mathbf{x}) \in \mathbb{R}^3 | x_1 > a \cap b \leq x_2 \leq c\})$, where $a \in D_{X_1} \subseteq \mathbb{R}$ and $b, c \in D_{X_2} \subseteq \mathbb{R}$, $b < c$.

Consider a value $s \in D_{X_i}|R_j$, where $D_{X_i}|R_j$ is the domain of X_i restricted to the rectangle R_j. The impurity reduction generated in the rectangle R_j by X_i at the cutpoint s, is given by:

$$d_{ij}^s = \Delta H_Y(X_i, R_j) = p_j \cdot \{H_Y - (p_{jL} H_{Y|X_i \le s} + p_{jR} H_{Y|X_i > s})\}, \quad (1)$$

where $p_j = P(R_j)$, $p_{jL} = P(X_i \le s|R_j)$ and $p_{jR} = P(X_i > s|R_j)$. H_Y, $H_{Y|X_i \le s}$ and $H_{Y|X_i > s}$ are the heterogeneity indexes of Y in the jth rectangle and in the left and right splits of R_j, respectively. Let d_{ij} be the maximum heterogeneity reduction allowed by covariate X_i in the jth rectangle, for all the possible cutpoints $s \in D_{X_i}|R_j$

$$d_{ij} = \max_{s \in D_{X_i}|R_j} d_{ij}^s. \quad (2)$$

The goal of partitioning algorithms is to maximally reduce the heterogeneity of Y within the rectangles. Therefore, for each R_j, the splitting variable X_i and the cutpoint s are those that maximize the impurity reduction in that subset. In other words, the partitioning variable X_i satisfies in R_j the condition $d_{ij} > d_{hj}$ for $h = 1, 2, \ldots, p$, $h \neq i$.

In this context, TDNI measures of variable importance are based on the following notion of importance μ_i of a covariate X_i: μ_i is the total decrease of the heterogeneity index H_Y attributable to X_i. In other words, μ_i is computed summing up all the decreases of heterogeneity d_{ij} obtained in the rectangles generated using X_i as the splitting variable:

$$\mu_i = \sum_{j \in J} d_{ij} \cdot I_{ij}, \quad (3)$$

where I_{ij} is the indicator function which equals 1 if the ith variable is used to split R_j and 0 otherwise.

Several impurity/heterogeneity indexes H have been proposed for the case of a categorical Y: the Pearson's chi-squared statistic, the Gini criterion, the entropy criterion, the families of splitting criteria of Shih (1999), etc. When Y is numerical, the most popular measure H is variance.

In the following example we show how, according to (3), the importance μ_i can be calculated using the joint probability distribution of $(Y, \mathbf{X})$. The data generating process described in this example will be used to produce the dataset of Example 2 in Sect. 3 and of Simulation (4) and (5) in Sect. 4.

Example 1 (Calculation of variable importance) Consider the variable $(Y, \mathbf{X}) = \{Y, X_1, X_2, X_3\}$, where X_1 and X_2 are two binary 0/1 independent covariates, Y is a continuous standard normal response variable generated by the following data generating process (see Fig. 1(a)): $P(X_1 = 0) = P(X_2 = 0) = 1/2$, $(Y|X_1 = 0) \sim N(-1/2, 3/4)$, $(Y|X_1 = 1) \sim N(1/2, 3/4)$, $(Y|X_2 = 0) \sim N(-1/3, 8/9)$, $(Y|X_2 = 1) \sim N(1/3, 8/9)$, $(Y|X_1 = 0 \cap X_2 = 0) = (Y|X_1 = 0 \cap X_2 = 1) \sim N(-1/2, 3/4)$, $(Y|X_1 = 1 \cap X_2 = 0) \sim N(0, 1/2)$, $(Y|X_1 = 1 \cap X_2 = 1) \sim N(1, 1/2)$.

Consider the following three cases for the uninformative variable X_3:

Case A: a binary 0/1 covariate independent on X_1 and X_2;
Case B: a continuous standard normal covariate independent on X_1 and X_2;
Case C: a continuous covariate, normally distributed conditionally to X_1: $(X_3|X_1 = 0) \sim N(0, 1)$ and $(X_3|X_1 = 1) \sim N(1, 1)$.

Variable importances can be calculated in the three cases A, B and C. Since Y is continuous, we can adopt variance as heterogeneity index and (1) can be expressed as:

$$d_{ij}^s = \Delta \sigma_Y^2(X_i, R_j) = p_j \cdot \{\sigma_Y^2 - (p_{jL} \sigma_{Y|X_i \le s}^2 + p_{jR} \sigma_{Y|X_i > s}^2)\}, \quad (4)$$

where σ_Y^2, $\sigma_{Y|X_i \le s}^2$ and $\sigma_{Y|X_i > s}^2$ are the variances of Y in the jth rectangle and in the left and right splits, respectively.

Cases A *and* B. The calculation of variable importance is the same in the two cases, because Y is stochastically independent on X_3. The different levels of measurement of X_3 in the two cases do not influence variable importance. When the whole sample space is considered (i.e. $R_1 = \mathcal{D}$), X_1 is the most effective variable in reducing the heterogeneity of Y by means of a binary split because $d_{11} = p_1 \cdot 1/4 = 1/4 > d_{21} = p_1 \cdot 1/9 = 1/9 > d_{31} = 0$. The sample space is then partitioned according to X_1. The sample space conditioned to $X_1 = 1$ (R_3), can be further partitioned by X_2 since $d_{23} = p_3 \cdot 1/4 = P(X_1 = 1) \cdot 1/4 = 1/8 > d_{33} = 0$. The sample space conditioned to $(X_1 = 1) \cap (X_2 = 0)$ (R_4) cannot be further partitioned because $d_{34} = 0$. The same is true in the sample space conditioned to $(X_1 = 1) \cap (X_2 = 1)$ ($R5$). Similarly, no further partitioning is possible in the sample space conditioned to $X_1 = 0$ (R_2) because $d_{22} = d_{32} = 0$. Hence, the VIs of the three covariates are $\mu_1 = d_{11} = 1/4$, $\mu_2 = d_{23} = 1/8$, $\mu_3 = 0$.

Case C. In this case X_3 is no longer independent on Y, due to the relationship existing between X_3 and X_1. Here X_3 is independent on Y, conditionally to X_1. Now we show that the VIs of the three covariates are the same as case A and B, because in $R_1 = \mathcal{D}X_1$ remains the most effective variable in reducing the heterogeneity of Y by means of a binary split. This can be proved as follows. Let s be a cutpoint for X_3 in R_1 and let $P(X_1 = 0|X_3 \le s) = p$ and $P(X_1 = 0|X_3 > s) = q$. The distributions of Y, conditionally on X_3 lower and greater than s, are mixtures of normal variables, that is $f(Y|X_3 \le s) = pf(Y|X_1 = 0) + (1 - p)f(Y|X_1 = 1)$ and $f(Y|X_3 > s) = qf(Y|X_1 = 0) + (1 - q)f(Y|X_1 = 1)$,

where $f(Y|X \in A)$ is the density function of y given that $X \in A$. It is easy to show that

$$\sigma^2_{Y|X_3 \le s} = 3/4 + p(1-p) > \sigma^2_{Y|X_1=0} = \sigma^2_{Y|X_1=1} = 3/4,$$
$$\sigma^2_{Y|X_3 > s} = 3/4 + q(1-q) > \sigma^2_{Y|X_1=0} = \sigma^2_{Y|X_1=1} = 3/4.$$

It follows that, for all s,

$$\sigma^2_Y - (P(X_1=0)\sigma^2_{Y|X_1=0} + P(X_1=1)\sigma^2_{Y|X_1=1}) > \sigma^2_Y - (P(X_3 \le s)\sigma^2_{Y|X_3 \le s} + P(X_3 > s)\sigma^2_{Y|X_3>s}),$$

thus $d_{11} > d_{31}$. When the sample space is partitioned according to X_1, X_3 is independent on Y in the two generated rectangles, so the domain partition is the same as in case A and B, as well as the resulting VI measures μ_1, μ_2 and μ_3.

We now consider the case of a tree t built using a sample with size N. The impurity reduction $\hat{d}_{ij}$ at node j attributable to covariate X_i with cutpoint s can be estimated by:

$$\hat{d}^s_{ij} = \widehat{\Delta H_Y}(X_i) = \frac{n_j}{N}\left\{\hat{H}_Y - \left(\frac{n_{jL}}{n_j}\hat{H}_{Y|X_i \le s} + \frac{n_{jR}}{n_j}\hat{H}_{Y|X_i > s}\right)\right\} \quad (5)$$

where $\hat{H}_Y$, $\hat{H}_{Y|X_i \le s}$ and $\hat{H}_{Y|X_i > s}$ are the estimated heterogeneities of Y in the jth rectangle and in the left and right splits, respectively. n_j, n_{jL}, n_{jR} are the sample sizes in node j and in the left and right splits. Similarly, d_{ij} is estimated by:

$$\hat{d}_{ij} = \max_{s \in S_{ij}} \hat{d}^s_{ij}, \quad (6)$$

where S_{ij} is the set of available cutpoints of variable X_i at node j.

The covariate X_i is selected at node j as splitting variable if $\hat{d}_{ij} > \hat{d}_{hj}$ for all $h = 1, \ldots, p$, $h \ne i$. The estimate of the TDNI importance μ_i using a tree t is given by the sum of the estimated impurity reductions attributable to covariate X_i over the set J of nonterminal nodes of the tree (Breiman et al. 1984), that is:

$$\widehat{\mathrm{VI}}_i(t) = \sum_{j \in J} \hat{d}_{ij} \cdot I_{ij}. \quad (7)$$

In the regression case we can use the sample variance $\hat{\sigma}^2$ as an estimator of the heterogeneity of node and splits. Hence, (5) becomes:

$$\hat{d}^s_{ij} = \widehat{\Delta\sigma^2_Y}(X_i) = \frac{n_j}{N}\left\{\hat{\sigma}^2_Y - \left(\frac{n_{jL}}{n_j}\hat{\sigma}^2_{Y|X_i \le s} + \frac{n_{jR}}{n_j}\hat{\sigma}^2_{Y|X_i > s}\right)\right\} \quad (8)$$

$$= \frac{n_j}{N}\left\{\frac{\mathrm{DEV}_{\mathrm{total}}(j)}{n_j} - \frac{\mathrm{DEV}_{\mathrm{within}}(jL, jR)}{n_j}\right\} = \frac{1}{N}\mathrm{DEV}_{\mathrm{between}}(jL, jR), \quad (9)$$

where $\mathrm{DEV}_{\mathrm{within}}$ and $\mathrm{DEV}_{\mathrm{between}}$ are the within-node and the between-node deviance, respectively. From (9) one can derive that, in the regression case, the TDNI measure (7) of a covariate X_i is equal to the total amount of $\mathrm{DEV}_{\mathrm{between}}$ imputable to that covariate in the tree.

For tree-based ensembles the VI measure is given by the average of $\widehat{\mathrm{VI}}_i$ over the set of T trees:

$$\widehat{\mathrm{VI}}_i = \frac{1}{T}\sum_{t=1}^{T} \widehat{\mathrm{VI}}_i(t). \quad (10)$$

The VI measure (10) has been proposed by Breiman (2002) in Random Forests and is called 'Measure 4' (M4), because the last of a set of four importance measures. With minor modifications, Friedman (2001) proposed an 'influence of input variables' for GBM, with $\hat{d}^2_{ij}$ in place of $\hat{d}_{ij}$ and $\widehat{\mathrm{VI}}_i$ rescaled by assigning a value of 100 to the most influential covariate.

3 Bias in TDNI measures

A crucial point for the analysis that follows is the notion of informative and uninformative splits. Suppose that $\mathcal{D}$ has been recursively partitioned into J rectangles $\{R_j\}_{j=1,2,\ldots,J}$. If X_i and Y are stochastically independent, they continue to be independent in each R_j. On the contrary, if some association between X_i and Y exists, X_i and Y could be dependent or conditionally independent in a given R_j. In other words, when predicting Y, uninformative covariates (i.e. stochastically independent on Y) always remain uninformative, in each subset of the sample space. Informative (i.e. somehow associated with Y) covariates can continue to be informative or can become uninformative in R_j.

Let us now suppose to grow a tree using a sample of N units and suppose that, within a given node, there is at least one covariate having some association with Y. The node will be split by using the best covariate, that is the covariate that maximizes the heterogeneity reduction $\hat{d}_{ij}$. Hence, because the heterogeneity reductions of informative covariates will be typically greater than the heterogeneity reductions of the uninformative ones, an informative covariate will be chosen as splitting variable. We define this circumstance as an *informative split*. When within a node there are no informative covariates, only uninformative covariates and/or informative covariates which became uninformative can be chosen as splitting covariate. This is the case of an *uninformative split*. We can formalize the following definition.

Definition 1 (Informative and uninformative splits) Given a tree t grown from a sample, the split of a node j made by covariate X_i (i.e. $\hat{d}_{ij} > \hat{d}_{hj}$, $\forall h = 1, 2, \ldots, p$, $h \neq i$) is called *uninformative* if $d_{ij} = 0$ and *informative* otherwise. An uninformatively split node is a node where an uninformative split occurs.

In informative splits, the heterogeneity reduction $\hat{d}_{ij}$ of the splitting covariate is a direct consequence of its importance. Differently, in an uninformative split, $\hat{d}_{ij}$ is a product of chance. Therefore, when calculating the TDNI measure $\widehat{\mathrm{VI}}_i(t)$ of covariate X_i by (7), it is of fundamental importance to distinguish between impurity reductions attributable to informative splits and impurity reductions generated by uninformative splits. In other words, $\widehat{\mathrm{VI}}_i(t)$ can be expressed as the sum of two components:

$$\widehat{\mathrm{VI}}_i(t) = \sum_{j \in J_I} \hat{d}_{ij} \cdot I_{ij} + \sum_{j \in J_U} \hat{d}_{ij} \cdot I_{ij} = \hat{\mu}_i(t) + \varepsilon_i(t), \quad (11)$$

where J_I and J_U, $J_I \cup J_U = J$, are the set of nodes characterized by informative and uninformative splits, respectively. $\hat{\mu}_i(t)$ 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 $\varepsilon_i(t)$ is a noisy component associated with the selection of X_i within uninformative splits and is a source of bias for $\widehat{\mathrm{VI}}_i(t)$.

Example 2 (Informative and uninformative splits) Consider the sample of 16 units given in Table 1 and generated by the process of Example 1, Case A. Suppose to fit a fully-grown regression tree to these data. Applying (5), in the root node the impurity reductions associated to the three covariates are: $\hat{d}_{11} = 0.266$, $\hat{d}_{21} = 0.214$, $\hat{d}_{31} = 0.001$. Hence, X_1 is the best splitting variable at the root node and the split is informative because $d_{11} \neq 0$ (and $d_{11} > d_{21} > d_{31}$). In node 2 ($X_1 = 0$), $\hat{d}_{22} = 0.053$ and $\hat{d}_{32} = 0.041$. X_2 is used as splitting variable. This is a uninformative split because $d_{22} = 0$ (and $d_{32} = 0$). In node 3 ($X_1 = 1$), $\hat{d}_{23} = 0.180$ and $\hat{d}_{33} = 0.004$. The splitting variable is X_2 and the split is informative since $d_{23} \neq 0$ (and $d_{23} > d_{33}$). In nodes 4, 5, 6 and 7 the impurity reductions attributable to X_3 are $\hat{d}_{34} = 0.031$ and $\hat{d}_{35} = 0.012$, $\hat{d}_{36} = 0.006$, $\hat{d}_{37} = 0.028$. The splits are all uninformative. Nodes 8 through 15 are leaf nodes. The resulting regression tree is shown in Fig. 1(b). Uninformative splits have been marked by thick lines. The estimated VIs of the three covariates are: $\widehat{\mathrm{VI}}_1 = \hat{\mu}_1 = 0.266$,

Table 1 Sample data of Example 2

	1	2	3	4	5	6	7	8
X_1	0	0	0	0	0	0	0	0
X_2	0	0	0	0	1	1	1	1
X_3	0	1	0	1	1	0	0	1
Y	−1.894	−1.129	−0.581	0.069	−0.346	−0.878	−0.023	0.311
	9	10	11	12	13	14	15	16
X_1	1	1	1	1	1	1	1	1
X_2	1	1	1	1	0	0	0	0
X_3	0	1	0	0	0	0	0	1
Y	0.497	0.495	2.301	1.003	0.756	−1.505	0.104	0.137

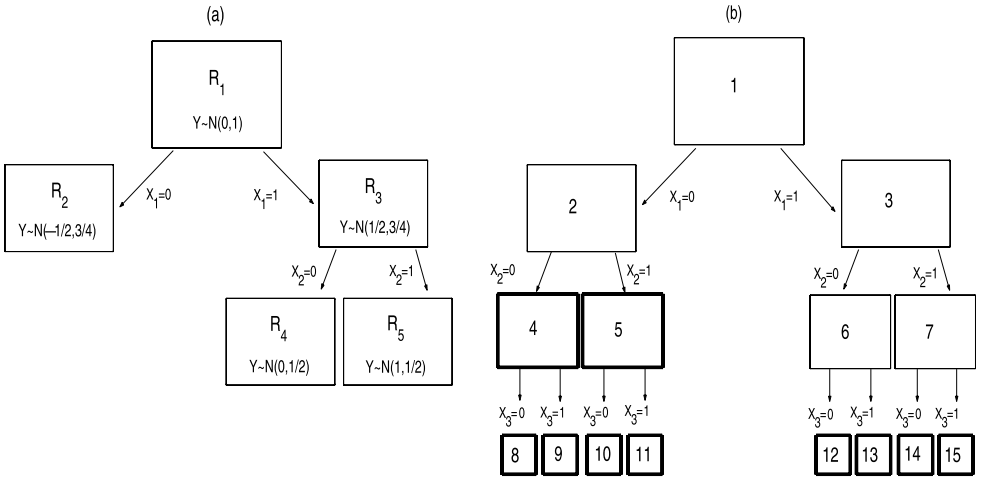

Fig. 1 Data generating process of Example 1(a) and the regression tree grown from the sample data of Example 2(b). Uninformative splits are marked by *thick boxes*

$\widehat{\mathrm{VI}}_2 = \hat{\mu}_2 + \varepsilon_2 = 0.180 + 0.053 = 0.233$ and $\widehat{\mathrm{VI}}_3 = \varepsilon_3 = 0.031 + 0.012 + 0.006 + 0.028 = 0.077$.

Two remarks are worth pointing out. Firstly, the sample data used to grow the tree are also used to calculate TDNI measures. In other words, TDNI are a class of *in-sample* measures of variable importance. This is the crucial difference between TDNI measures and the mean decrease in prediction accuracy, a popular permutation-based VI measure. This measure is defined in RF as follows: for each tree, the algorithm randomly rearranges the values of the ith variable for the out-of-bag set (i.e. the subset of the bootstrap sample not used in the construction of the tree), puts this permuted set down the tree, and gets new predictions from the forest. The importance of the ith variable is defined as the difference between the original out-of-bag error rate and the out-of-bag error rate for the randomly permuted ith covariate. Hence, mean decrease in accuracy is fundamentally an out-of-sample VI measure, in contrast to the in-sample character of TDNI measures.

Secondly, the notion of informative and uninformative splits is intimately related to the notion of *overfitting*. It is well known that an overfitted model shows a high in-sample accuracy, but does not validate, that is, does not provide accurate predictions for out-of-sample observations. Base learners of RF are fully-grown trees. They have a serious risk of overfitting (Berk 2006). When building a CART, the model starts learning the underlying structure of data and typically the first splits of the tree are informative splits. Subsequently, after an adequate number of informative splits, informative variables become uninformative, uninformative splits take place and the model learns the fine structure of data that is generated by noise. In other words, overfitting and uninformative splits of tree-based models are synonymous.

The in-sample character of TDNI measures and the effects of overfitting can lead to a seemingly paradox. Consider the case where only one covariate X is available, X and Y are continuous, a sample of N units has been observed and $\hat{d}_{ij}$ as defined in (8) is used. In a fully grown tree the importance of X is equal to the variance of Y, no matter what relationship between X and Y exists. Consider the two limiting cases: the deterministic case $Y = f(X)$ and the null case (i.e. Y and X are stochastically independent). The importance (7) of X is the same in the two cases, but in the first case the (maximal) tree contains only informative splits and $\widehat{\mathrm{VI}} = \hat{\mu}$, while in the second case only uninformative splits are present and $\widehat{\mathrm{VI}} = \varepsilon$. This problem vanishes if the mean decrease in prediction accuracy is used instead of TDNI measures or if one avoids uninformative splits when building the tree.

Pruning techniques are effective methods for controlling overfitting in CART. The aim of pruning is to remove uninformative splits: in well-pruned trees the number of uninformative splits is minimized. Hence, bias of TDNI measures is minimized, too. In the context of RF unpruned trees are typically used as base learners because the RF prediction is obtained averaging the predictions of the single trees and this neutralizes the problem of overfitting. Pruning seems to be unnecessary, but we know that this is only partially true. There is no need of pruning for improving the accuracy of prediction, but when we use RF for variable selection we cannot forget that fully grown unpruned trees can generate a substantial amount of bias in TDNI measures.

4 Level of bias and level of covariate measurement

In a tree grown from a sample of N units, each covariate X_i can be used as splitting variable only in a finite number of nodes. At node j, X_i is characterized by a finite number nps_{ij} of possible binary splits which depends on the level of measurement of the covariate. A nominal covariate with k categories within a given node has $\mathrm{nps}_{ij} = 2^{k-1} - 1$ possible splits, while an ordinal covariate with k categories has $\mathrm{nps}_{ij} = k - 1$ possible splits. A numerical (continuous) covariate with n_j distinct values within node j can be viewed as the limiting case of an ordinal covariate with as many categories as the number of sample units in the node. Thus, it has $\mathrm{nps}_{ij} = n_j - 1$. Of course, $\mathrm{nps}_{ij} \leq \mathrm{nps}_{ik}$ for all parent nodes k of a node j because each child node contains only a subset of the original sample and each covariate in a node has a number of distinct values (or categories) lower than or equal to the number of its distinct values (or categories) in the parent nodes.

Consider a node where all the covariates X_i are conditionally independent on Y. By definition, only an uninformative split can take place in this node. All the binary partitions of all the covariates have the same probability of being the best one and the selection of the splitting variable is only a product of chance. Therefore, the covariates with the highest number nps_{ij} of possible splits are more likely to be chosen as splitting variables. Recalling the decomposition given in (11), the above considerations imply that the expected values of the noisy component ε_i of the estimated VIs are not equal but depend on the level of measurement of covariates.

Let $\mathcal{J}_{\mathcal{U}}$ be the set of all the possible uninformatively split nodes, that is the set of all the nodes where an uninformative split occurs, for all the trees grown on all the possible N-size samples. We have:

$$E(\varepsilon_i) = E\left(\sum_{j \in J_U} \hat{d}_{ij} \cdot I_{ij}\right) = \sum_{j \in \mathcal{J}_{\mathcal{U}}} E((\hat{d}_{ij} \cdot I_{ij}) \cdot I_j),$$

where I_j is the indicator function which equals 1 if the uninformatively split node j occurs in the tree t and 0 otherwise.

Since the occurrence of a given node j depends only on former splits, I_j and $\hat{d}_{ij} \cdot I_{ij}$ are independent and we can write:

$$E(\varepsilon_i) = \sum_{j \in \mathcal{J}_{\mathcal{U}}} E(\hat{d}_{ij} \cdot I_{ij}) \cdot E(I_j) = \sum_{j \in \mathcal{J}_{\mathcal{U}}} E(\hat{d}_{ij} \cdot I_{ij}) \cdot q_j,$$

where q_j is the probability of occurrence of node j in a tree. Finally, applying the law of iterated expectation, we obtain:

$$E(\varepsilon_i) = \sum_{j \in \mathcal{J}_{\mathcal{U}}} E(\hat{d}_{ij} | I_{ij} = 1) \cdot p_{ij} \cdot q_j \tag{12}$$

where p_{ij} is the probability of selecting covariate X_i at node j. By definition, in uninformatively split nodes all the covariates are independent on the response variable and $E(\hat{d}_{ij} | I_{ij} = 1) = \bar{d}_j$ for all $i = 1, 2, \ldots, p$. On the contrary, p_{ij} depends on the number nps_{ij} of possible splits of X_i at node j:

$$p_{ij} = \frac{\text{nps}_{ij}}{\sum_{i=1}^{p} \text{nps}_{ij}}. \tag{13}$$

Equation (12) proves two fundamental facts: (a) covariate importances estimated by TDNI measures can show different levels of bias according to different levels of measurement of covariates and (b) the source of bias is very closely connected to the selection mechanism of the splitting variable in uninformatively split nodes.

In the following two subsections, using some simulation experiments, we investigate these characteristics of bias in further detail: its dependence on the number of uninformative splits and on the covariates' level of measurement.

4.1 Simulation studies with a single regression tree

The main goal of the present study is to investigate the problem of bias in TDNI VI measures estimated by ensemble methods with fully-grown trees as base learners, like RF. Equations (10) and (11) show that bias in tree-based ensembles can be expressed as an average of the bias originated in base learners. Hence, it is convenient to start our investigation about the source and the characteristics of bias considering a single unpruned regression tree.

Two simulation studies are performed. We consider a data generating process with 9 independent covariates: 1 binary variable (B), 4 ordinal variables (O4, O8, O16 and O32) with 4, 8, 16 and 32 categories and 4 nominal variables (N4, N8, N16 and N32) with 4, 8, 16 and 32 categories. The continuous outcome variable Y is independent on these covariates (null case). The null case guarantees that only uninformative splits take place and estimated VIs are entirely dominated by bias.

Simulation (1) In the first numerical experiment we investigate the relationship between bias and the number of uninformative splits. We generate 100 random samples with size $N = 3000$. For each sample, VIs are estimated by a single unpruned tree varying the `nodesize` parameter (the minimum size of terminal nodes) in the set $\{5, 6, 7, 8, 9, 10, 13, 15, 20, 25, 30, 60\}$. The number of leaf nodes of the trees together with the estimated VIs of the 9 covariates are collected in a dataset. For each covariate, we estimate a quadratic regression model between number of leaf nodes and VI.

Figure 2(a) shows the relationship existing between the mean level of bias predicted by quadratic models and the number of uninformative splits (R^2 values ranges from 0.702 to 0.949, the lower value has been observed for the binary variable and the higher for N32). These results substantially agree with the information contained in (12): bias levels appear to be a non-decreasing function of the number of uninformative splits. The estimated functions show a statistically significant concavity. This fact can be explained considering (5) and (7): growing trees with increasing levels of ramification generates nodes with a decreasing size n_j and $\hat{d}_{ij}$ is directly proportional to the node sample size n_j. Hence, a progressive increase in the number of leaf nodes produces in (7) the addition of terms that do not grow at the same speed and increases of VI that are less than proportional.

Simulation (2) In this simulation we study the relationship between bias and the number of categories in ordinal and nominal covariates. We generate a set of 100 random samples with size $N = 3000$. For each sample, VIs are estimated by a single unpruned tree with the `nodesize` parameter fixed to 5. The estimated VIs of the 9 covariates are collected in a dataset and the relationship between bias and number of categories is analyzed estimating two quadratic regression models: one for the set of nominal covariates and one for ordinal covariates.

The mean levels of bias predicted by quadratic models are plotted in Fig. 2(b) with respect to the logarithm of the number of covariate categories (the value of R^2 of the two models is 0.997). According to (12) and (13), these curves show that bias levels grow monotonically with the number of categories. In addiction, nominal variables have higher level of bias compared to ordinal variables because at each node they allow a greater number of possible splits nps_{ij} than ordinal variables with the same number of categories.

4.2 Simulation studies with a tree-based learning ensemble

In this subsection, we continue our investigation about the characteristics of bias of TDNI measures taking into account tree-based ensembles. We present four numerical ex-

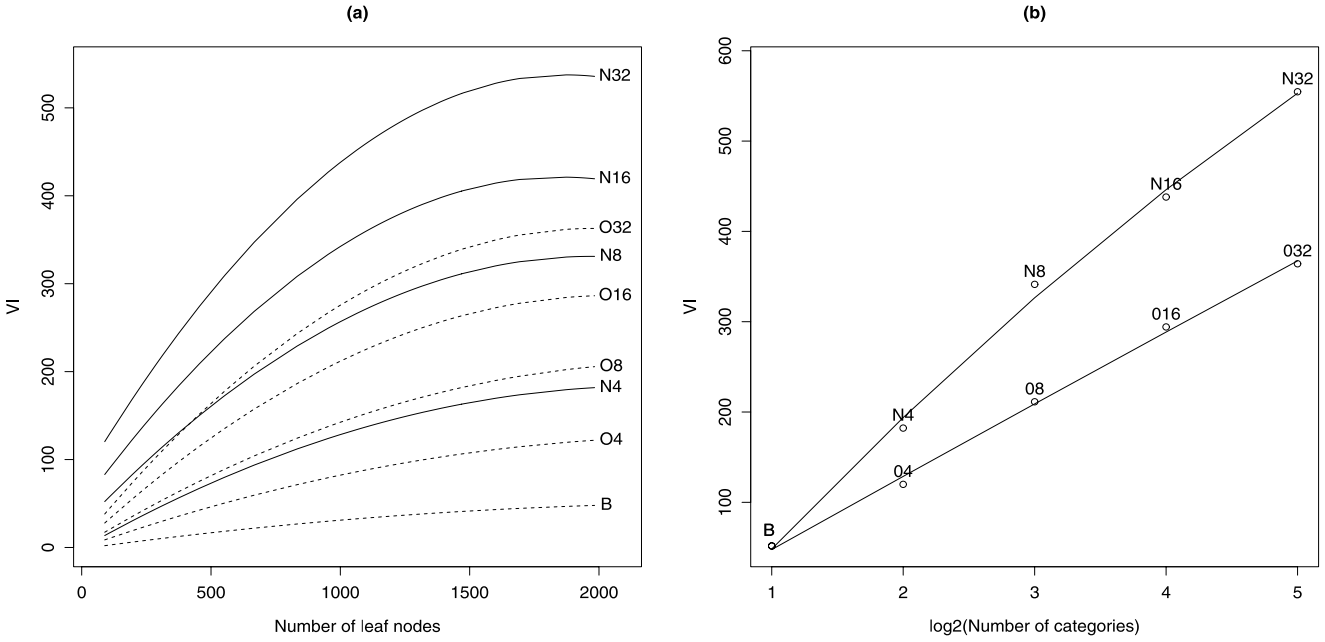

Fig. 2 (**a**) Dependence of bias on the number of uninformative splits. (**b**) Dependence of bias on the level of covariate measurement (number of categories of ordinal and nominal variables)

periments where VIs are estimated using RF. We devote special attention to the effect of bias on the ranking of covariates sorted by decreasing levels of estimated importance.

Simulation (3) A set of $r = 100$ samples with $N = 250$ observations are generated using the null-case mechanism described in the previous subsection: 9 independent covariates with different levels of measurement and a continuous outcome variable independent on these covariates.

Simulation (4) In this experiment the data generating process of Example 1 (case B) is used, with the addition of 6 random independent variables having different levels of measurement: a binary covariate (B), two ordinal (O6 and O11) and two nominal (N6 and N11) covariates and 1 continuous covariate (C2). The covariates are all mutually independent and only X_1 and X_2 are informative. We consider a set of $r = 100$ samples with $N = 400$ observations.

Simulation (5) Here we consider the case of correlated predictor variables. This simulation is similar to simulation (4). The only difference relates the two continuous covariates. X_3 and C2 are normally distributed conditionally to X_1 and B, respectively: $(X_3|X_1 = 0) \sim N(0, 1)$, $(X_3|X_1 = 1) \sim N(1, 1)$ and $(\text{C2}|\text{B} = 0) \sim N(0, 1)$, $(\text{C2}|\text{B} = 1) \sim N(1, 1)$.

Simulation (6) In the last simulation experiment we consider the case of a regression problem with $p > N$. The sample size is $N = 50$ and the number of covariates is $p = 200$. The covariates have been divided into 20 groups, each consisting of 10 ordinal covariates with the following number of categories: 2, 2, 3, 3, 4, 4, 6, 6, 8, 8. The first group $\mathbf{X_1} = (X_1, X_2, \ldots, X_{10})$ contains mutually correlated covariates and X_1, X_3, X_5, X_7 and X_9 are correlated to the outcome. In the second group $\mathbf{X_2}$, covariates are mutually independent and X_{11}, X_{13}, X_{15}, X_{17} and X_{19} are correlated to the outcome. The remaining 18 groups have 180 covariates mutually independent and independent on outcome.

More specifically, $r = 1000$ repetitions have been generated by the following data generating process:

(1) N observations $(y, x_1^c, \ldots, x_p^c)'$ are randomly drawn from a multivariate $(p + 1)$-dimensional Gaussian distribution with mean vector $\mu = (0, \ldots, 0)'$ and covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & \rho'_{Y\mathbf{X_1}} & \rho'_{Y\mathbf{X_2}} & \cdots & \rho'_{Y\mathbf{X_{20}}} \\ \rho_{Y\mathbf{X_1}} & \mathbf{\Sigma_{X_1}} & \mathbf{0} & \ldots & \mathbf{0} \\ \rho_{Y\mathbf{X_2}} & \mathbf{0} & \mathbf{I} & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{Y\mathbf{X_{20}}} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{I} \end{bmatrix}$$

where $\rho_{Y\mathbf{X_i}}$ is the vector of 10 correlations between the covariates of the ith group and the outcome Y, $\rho_{Y\mathbf{X_1}} = \rho_{Y\mathbf{X_2}} \approx (0.31, 0, 0.28, 0, 0.26, 0, 0.26, 0, 0.25, 0)'$ and $\rho_{Y\mathbf{X_3}} = \cdots = \rho_{Y\mathbf{X_{20}}} = (0, \ldots, 0)'$. The symbol $\mathbf{0}$ denotes a 10-dimensional square matrix of zeros, $\mathbf{I}$ is the 10-dimensional identity matrix, and the generic element of $\mathbf{\Sigma_{X_1}}$ is given by $s_{ij} = 0.4^{|i-j|}$.

(2) Each covariate generated at step (1) is transformed into a categorical variable X_i, with a number k_i of cat-

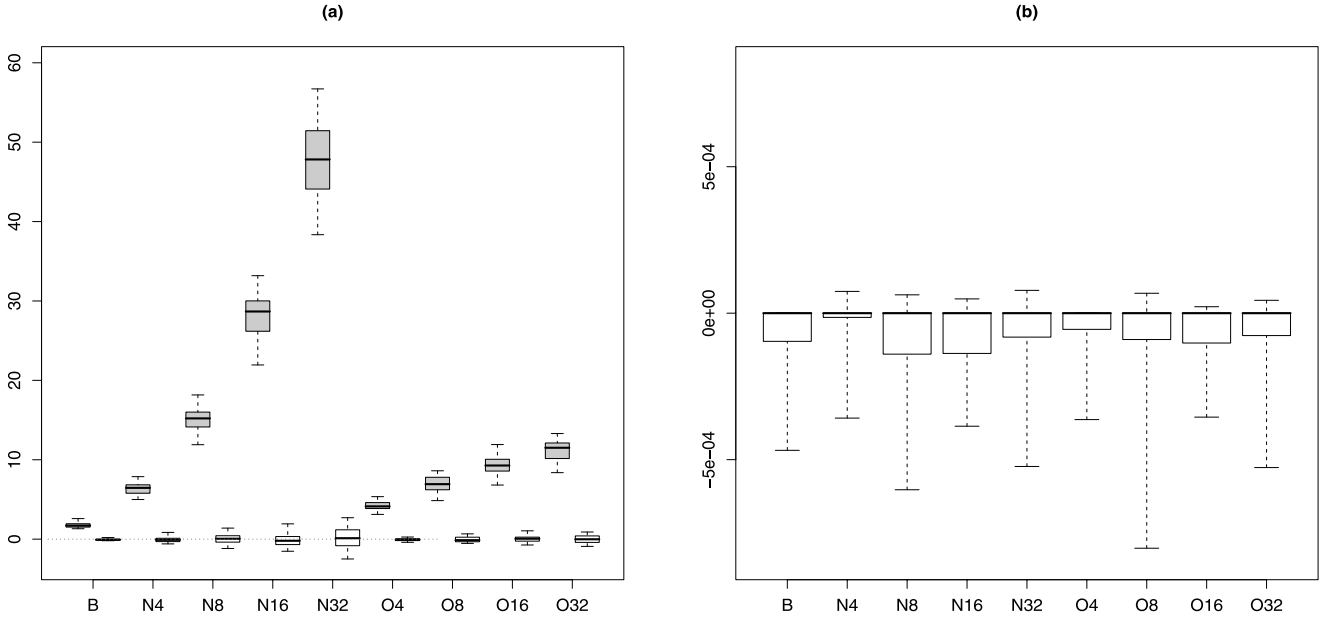

Fig. 3 Box-plots of VIs for simulation (3): (**a**) biased (*gray boxes*) and bias-corrected (*white boxes*) VIs and (**b**) permutation importance estimated by Conditional Random Forests

egories as given above. Categories are defined according to quantiles of the normal distribution: $X_i = k$ if $X_i \in (q_{(k-1)/k_i}, q_{k/k_i}]$, $k = 1, 2, \ldots, k_i$, where q_h is the hth quantile of a standard normal distribution.

The resulting data generating process has the following features:

- Outcome Y is associated to covariates $X_1, X_3, X_5, \ldots, X_{19}$ with constant correlation ratios $\eta^2_{Y|X_i} = 0.05$.
- Covariates $X_1, X_2, \ldots, X_{10}$ are mutually associated with Cramer's v indexes approximately given by $v_{i,j} \approx 0.4^{j-i-1} \cdot v_{i,i+1}$, $j > i + 1$, where $\{v_{i,i+1}\}_{i=1,2,\ldots,9} \approx \{0.26, 0.29, 0.23, 0.24, 0.20, 0.21, 0.17, 0.18, 0.15\}$.
- Covariates $X_{11}, X_{12}, \ldots, X_{200}$ are mutually independent and independent on covariates $X_1, X_2, \ldots, X_{10}$.

The gray boxes of Figs. 3(a), 4(a) and 5(a) show the distribution of the uncorrected TDNI measures estimated by means of RF for the three experiments (3), (4) and (5), respectively. Random Forests are implemented in the `randomForest` package (Liaw and Wiener 2002) of the R language (R Development Core Team 2008). In the simulation experiments of this section we have trained RFs with `ntree` $= 1000$ regression trees, with `mtry` $= 5$ variables randomly sampled as candidates at each split and with a minimum size of terminal nodes `nodesize` $= 5$.

The results of simulation (3) substantially confirm what was already observed in Fig. 2(b): higher numbers of covariate categories are generally associated to higher levels of bias. A byproduct of this relationship is an artificial ranking of covariates according to the number of categories, instead of variable importance. In this simulation, covariates are equally important because all uninformative but bias generates an erroneous ranking where the highest positions in the ranking are achieved by variables with the highest number of categories.

The negative effects of VI bias on ranking are more evident in simulations (4) and (5), where only X_1 and X_2 are informative. In these experiments the ranking of covariates using uncorrected VIs is clearly wrong. Uninformative covariates X_3, C2 and N11 are erroneously more important than the true informative variable X_2.

The average TDNI measures of simulation (6) estimated using RF are visualized in Fig. 6(a). A great amount of bias hides informative covariates. Their uncorrected importance is less than the importance of many uninformative variables. Bias strongly distort the ranking of variables.

All these results undoubtedly show that an effective method for bias correction is necessary when using TDNI measure of variable importance.

5 A bias-correction strategy

This Section starts with a brief recall of the bias-correction strategy recently proposed by Sandri and Zuccolotto (2008).

Let $\mathbf{X}$ be the $(N \times p)$ matrix containing the N observed values of the p covariates $\{X_1, \ldots, X_p\}$. A set of matrices $\{\mathbf{Z}_s\}_{s=1}^{S}$ is generated by randomly permuting S times the N rows of $\mathbf{X}$. We call the columns of $\mathbf{Z}$ 'pseudocovariates'. Row permutation destroys the association existing between the response variable Y and each pseudocovariate Z_i. On the contrary, the association between two pseudocovariates Z_i

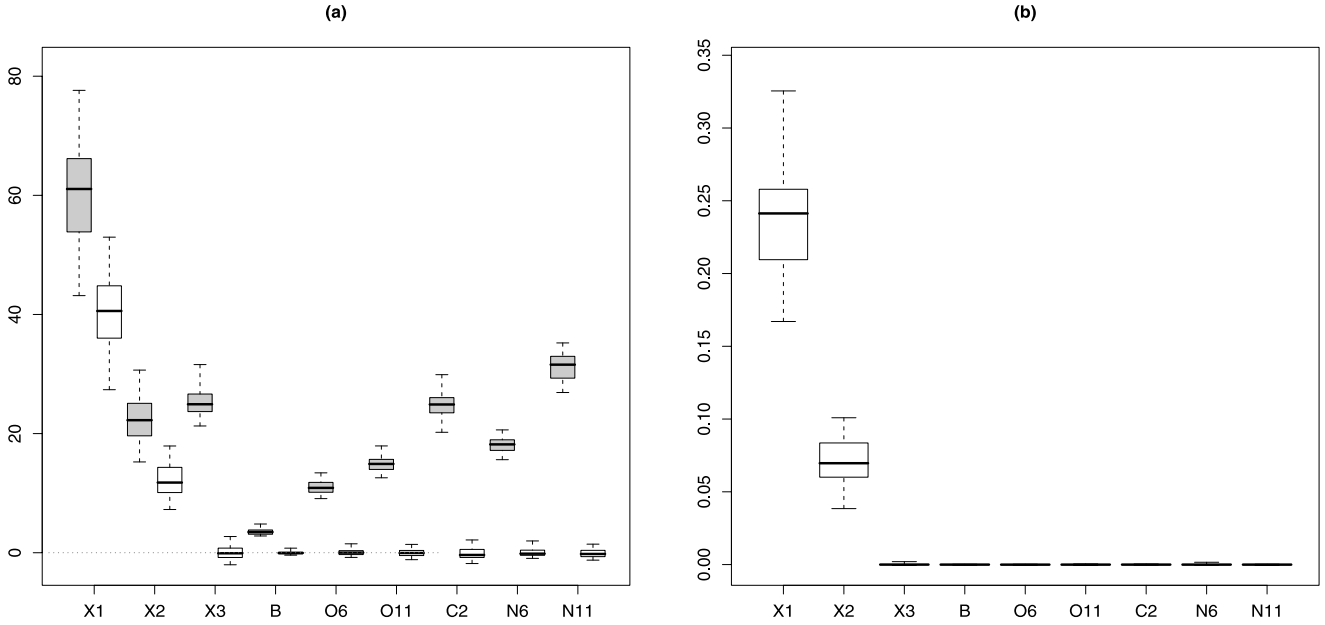

Fig. 4 Box-plots of VIs for simulation (4): (**a**) biased (*gray boxes*) and bias-corrected (*white boxes*) VIs and (**b**) permutation importance estimated by Conditional Random Forests

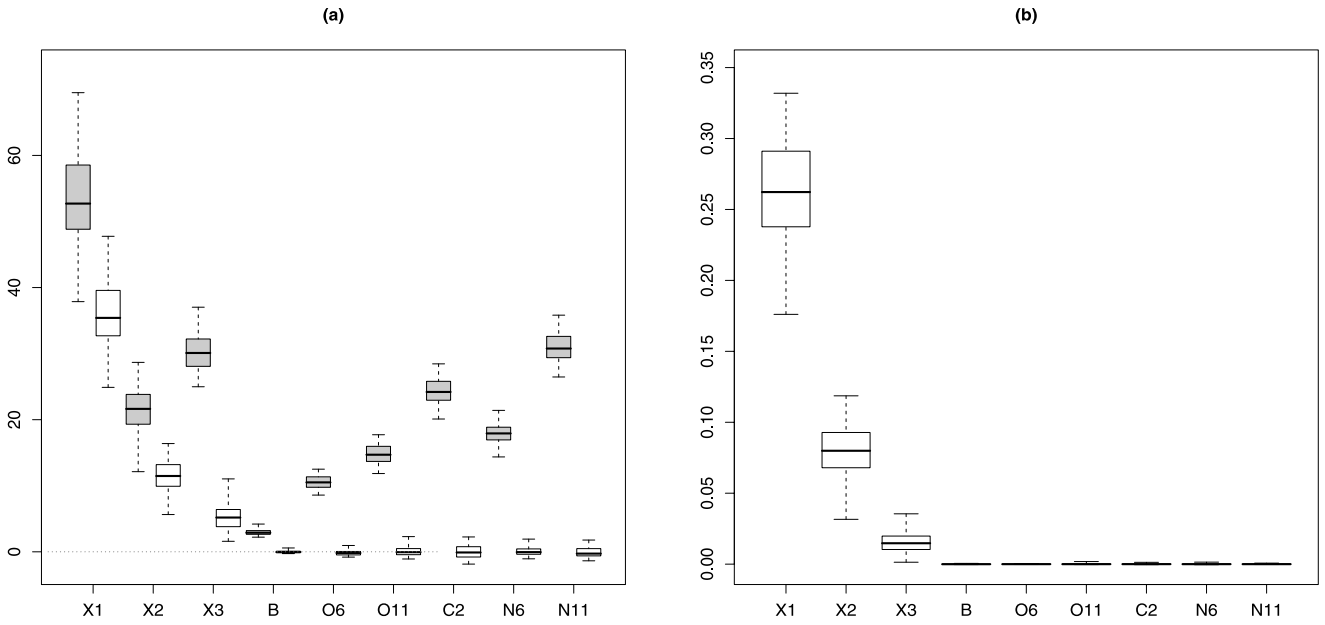

Fig. 5 Box-plots of VIs for simulation (5): (**a**) biased (*gray boxes*) and bias-corrected (*white boxes*) VIs and (**b**) permutation importance estimated by Conditional Random Forests

and Z_j is preserved, i.e. it is equal to the association existing between X_i and X_j.

Horizontally concatenating matrices $\mathbf{X}$ and $\mathbf{Z}_s$ generates a set of $(N \times 2p)$ matrices $\tilde{\mathbf{X}}_s \equiv [\mathbf{X}, \mathbf{Z}_s]$, $s = 1, 2, \ldots, S$. The augmented matrices $\tilde{\mathbf{X}}$ are repeatedly used to predict Y using an ensemble predictor and S importance measures are then computed for each covariate X_i and for the corresponding pseudocovariate Z_i. Let $\widehat{\mathrm{VI}}^s_{X_i}$ and $\widehat{\mathrm{VI}}^s_{Z_i}$ be the sth importance measures of X_i and Z_i, respectively. The adjusted variable importance measure $\overline{\mathrm{VI}}_i$ is computed considering the average of $\widehat{\mathrm{VI}}^s_{X_i} - \widehat{\mathrm{VI}}^s_{Z_i}$ differences, that is:

$$\overline{\mathrm{VI}}_i = \frac{1}{S}\sum_{s=1}^{S}(\widehat{\mathrm{VI}}^s_{X_i} - \widehat{\mathrm{VI}}^s_{Z_i}). \tag{14}$$

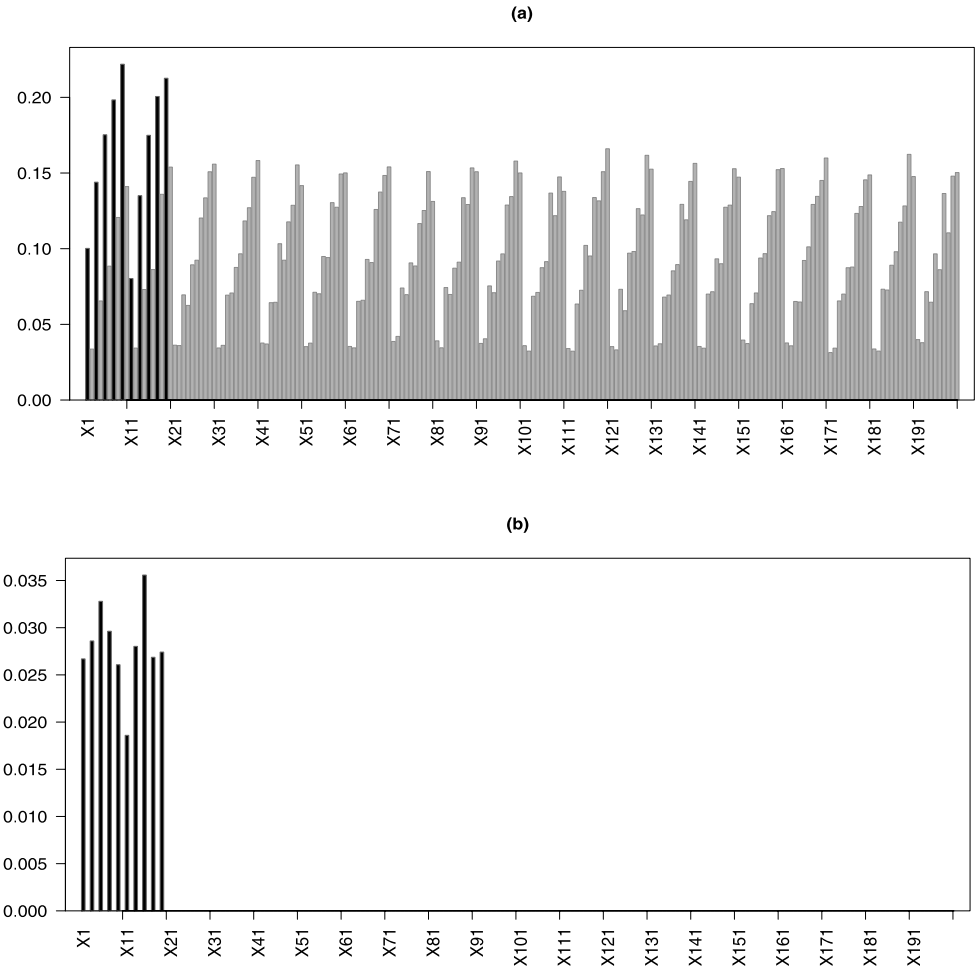

Fig. 6 Bar-plots of mean VIs for simulation (6): (**a**) biased and (**b**) bias-corrected mean VIs estimated by RF (*black* and *gray* bars for informative and uninformative variables, respectively)

For more details about the algorithm and the principles that support and guide the procedure, see Sandri and Zuccolotto (2008).

This method has been originally developed for classification tree-based ensembles when the Gini gain is used as splitting criterion, but it can be extended to the entire class of TDNI measures.

In Sect. 3, we have shown that for this class of measures the decomposition of (11) holds. The proposed bias-correction method is crucially based on this decomposition: in the set J_U of uninformatively-split nodes, each pseudocovariate Z_i shares approximately the same properties (number of possible splits nps_{ij}, independence on Y) of the corresponding covariate X_i and consequently has approximately the same probability p_{ij} of being selected as splitting variable. Hence, the average importance of Z_i calculated for S random permutations is an approximation of the sum given in (12): $av[\widehat{\text{VI}}^s_{Z_i}] \approx E[\varepsilon_{X_i}]$. In other words, $\sum_{s=1}^{S} \widehat{\text{VI}}^s_{Z_i}$ can be used as an approximation of the bias affecting the importance of X_i.

A direct consequence of the generalization of the bias-correction strategy to TDNI measures is the extension of its domain of applicability to regression problems.

We investigate the effectiveness of the above bias-correction algorithm in the 4 regression problems related to the data generating processes described in Sect. 4.2. The white boxes of Figs. 3(a), 4(a), 5(a) and 6(a) visualize the distribution of the corrected VIs (`ntree` $= 1000$, `mtry` $= 3$, `nodesize` $= 5$ and number of random permutation `S` $= 25$). These experiments clearly show that bias-correction using pseudocovariates has the power to reduce bias, even when using a small number S of random permutations. The technique generates right rankings of covariates according to their importance and informative variables can be correctly discriminated from uninformative ones. The method shows good performances also when estimating VIs of a large number of (mutually correlated, mutually independent, informative and uninformative) variables using a very limited number of sample units. All the uninformative covariates correctly have a null mean importance and all the mean VIs of informative variables are greater than zero.

The optimal number of variables randomly selected in each node, i.e. the value of the parameter `mtry` that minimizes the out of bag prediction error rate of the RF, is typically a function of the number of covariates of the dataset. Extensive simulations (not reported here) show that, when

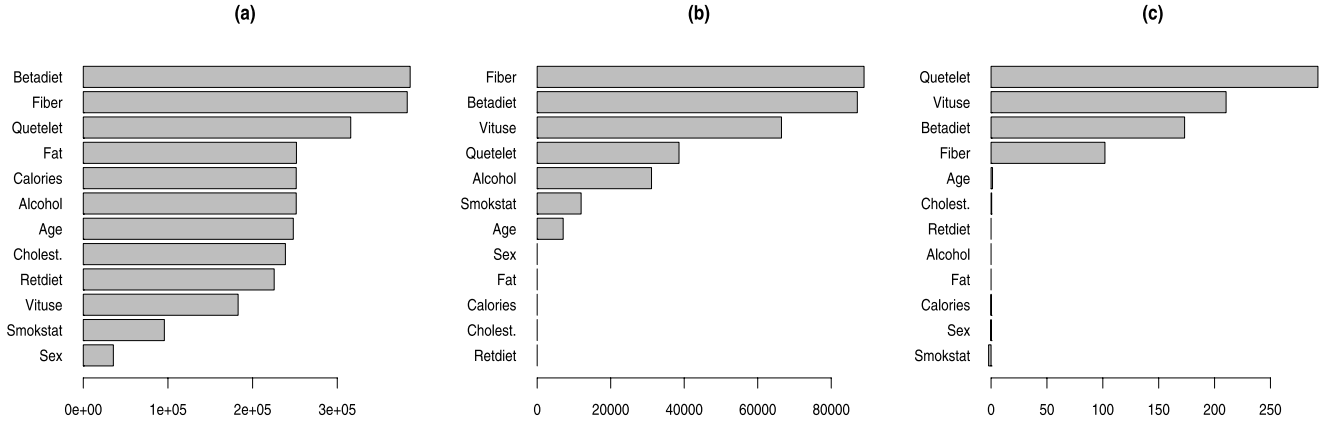

Fig. 7 Plasma beta-carotene outcome: (**a**) biased and (**b**) bias-corrected TDNI importance estimated by RF and (**c**) permutation importance estimated by Conditional Random Forests

adding the p pseudocovariates, it is not necessary to increase `mtry` and the optimal value chosen for the original dataset can still be used. This fact can be explained by considering that pseudocovariates are not additional potentially predictive variables but only 'competitors' of the original covariates. For each covariate, our method generates a 'twin' uninformative pseudocovariate that, only in uninformative splits, participates in the competition for the best split, with approximately the same probability of the selection of X_i.

For the sake of comparison, Figs. 3(b), 4(b) and 5(b) show the distribution of the (unbiased) permutation measures computed by Conditional Random Forests (CRF). This choice is motivated by two considerations: (a) the unified framework for recursive partitioning used by CRF to grow single trees strongly reduces the number of uninformative splits and (b) permutation measures are not affected by the kind of bias described in this paper. We have estimated CRF using the `cforest` command of the `party` package for R, with the following settings: 1000 trees, 3 randomly picked input variables in each node, `minsplit` $= 5$, a quadratic test statistics and Bonferroni-adjusted p-values (see Hothorn et al. 2006 and Strobl et al. 2007b). The results clearly show that these measures and the proposed bias-corrected TDNI measures are qualitatively very similar. Hence, the two methods can be used alternatively when one needs to calculate VIs.

In simulation study (5) the procedure is not able to completely remove the bias due to the association between X_3 and X_1. Bias in favor of X_3 is still present. This effect can also be observed when using the permutation measure calculated by CRF (see Fig. 5(b)). A new conditional permutation scheme for the computation of VI in case of correlated predictor variables has been recently proposed by Strobl et al. (2008).

6 Case study

The Plasma-Retinol dataset is available at the StatLib Datasets Archive and contains 315 observations of 14 variables aiming at investigating the relationship between personal characteristics and dietary factors, and plasma concentrations of retinol, beta-carotene and other carotenoids. The identification of determinants of low plasma concentration of retinol and beta-carotene is important because observational studies have suggested that this circumstance might be associated with increased risk of developing certain types of cancer.

Previous studies showed that plasma retinol levels tend to vary by age and sex, while the only dietary predictor seems to be alcohol consumption. For plasma beta-carotene, dietary intake, regular use of vitamins, and fiber intake are associated with higher plasma concentrations, while Quetelet Index and cholesterol intake are associated with lower plasma levels (Nierenberg et al. 1989).

We used the RF algorithm to predict BETAPLASMA and RETPLASMA as a function of the 12 covariates described in Table 2. The hyperparameters of the tree-based ensembles are `ntree` $= 3000$, `mtry` $= 4$, `nodesize` $= 5$ and a number of random permutations $S = 300$. We computed biased $\widehat{\mathrm{VI}}_{X_i}$ (Figs. 7(a) and 8(a)) and bias-corrected importance measures $\overline{\mathrm{VI}}_{X_i}$ (Figs. 7(b) and 8(b)) for BETAPLASMA and RETPLASMA, respectively. We set negative values of VI to zero.

For BETAPLASMA the corrected measures allow to identify 5 mainly predictive covariates (FIBER, BETADIET, VITUSE, QUETELET, ALCOHOL). Similarly, the most important predictors of RETPLASMA seem to be ALCOHOL, SEX and AGE. Our results largely confirm the findings of the preceding analyses except for the importance of CHOLESTEROL and ALCOHOL in predicting BETAPLASMA.

The influence of bias correction on ranking by importance is apparent. On one side, it allows to discard predictors

Table 2 Variables in the Plasma-Retinol dataset

Response variables
BETAPLASMA: Plasma beta-carotene (ng/ml)
RETPLASMA: Plasma Retinol (ng/ml)
Covariates
AGE: Age (years)
SEX: Sex (1 = Male, 2 = Female)
SMOKSTAT: Smoking status (1 = Never, 2 = Former, 3 = Current smoker)
QUETELET: Quetelet (weight/(height^2))
VITUSE: Vitamin Use (1 = Yes, fairly often, 2 = Yes, not often, 3 = No)
CALORIES: Number of calories consumed per day
FAT: Grams of fat consumed per day
FIBER: Grams of fiber consumed per day
ALCOHOL: Number of alcoholic drinks consumed per week
CHOLESTEROL: Cholesterol consumed (mg per day)
BETADIET: Dietary beta-carotene consumed (mcg per day)
RETDIET: Dietary retinol consumed (mcg per day)

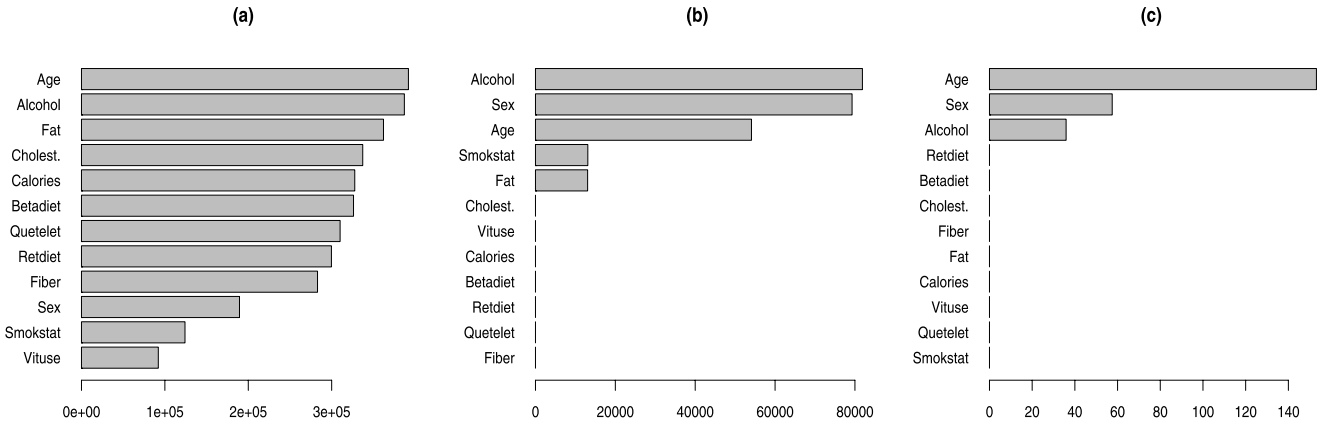

Fig. 8 Plasma retinol outcome: (**a**) biased and (**b**) bias-corrected TDNI importance estimated by RF and (**c**) permutation importance estimated by Conditional Random Forests

whose importance is artificially amplified by bias. In fact, on the basis of biased VIs, FAT and CALORIES seem informative covariates of BETAPLASMA, but after the removal of bias their importance vanishes. On the other side, bias correction can reveal informative predictors, hidden by bias in VI of other predictors. In Fig. 8(a), SEX seems scarcely influential on RETPLASMA, but the correction procedure shows that this variable is one of the three most important covariates.

Figures 7(c) and 8(c) show the permutation-based VIs calculated using CRF. These estimates are very similar to VIs obtained by the proposed algorithm.

7 Concluding remarks

It is well-known that the Gini VI measure, computed by means of tree-based learning ensembles, is affected by different kinds of bias (Strobl 2005). The existence of bias in VI is potentially dangerous especially in a variable selection perspective, since it can dramatically alter the ranking of predictors.

The main source of bias for the class of TDNI variable importance measures is intimately connected to the tree-construction mechanism: covariates X_i with the same importance in a node j can have different probabilities p_{ij} of being selected as splitting variables. Using theoretical considerations and simulation experiments, the present paper

shows that the levels of this bias depend on the characteristics of covariates, i.e. on their measurement levels. In addiction, the analysis indicates that this kind of bias is generated by uninformative splits. These splits are binary partitions of sample units that are completely driven by chance where the association between response variable and covariates have been entirely captured by earlier splits (the informative splits).

The heuristic bias-correction strategy of Sandri and Zuccolotto (2008) is based on the introduction of a set of pseudocovariates in the original dataset, in the spirit of Wu et al. (2007). Pseudocovariates are noise variables that are independent on the response variable and have the same correlation structure of the original covariates. Working on simulated and real life regression problems,[1] the paper shows that pseudocovariates have the capacity to approximate the component of the estimated VI attributable to bias. Hence, the proposed permutation method can reduce bias due to different measurement levels of covariates and can yield correct ranking of variables according to their importance.

Of course, the drawback of repeatedly generating pseudocovariates is an additional computational burden. This is a problem common to all permutation-based procedures and it cannot be ignored when the method is applied to large-scale datasets. However, it is fundamental to take into account that the use of the bias-correction method is necessary only when covariates with different levels of measurement are present. For example, genome-wide association studies typically collect data about thousands to hundreds of thousands single nucleotide polymorphisms that all have the same characteristics (e.g. are all real-valued variables). In these cases bias in TDNI measure does not affect covariates' ranking and therefore one can avoid to apply any correction. Anyway, simulation experiments show that the number S of sets $\{Z_s\}$ of pseudocovariates required for a satisfactory bias correction is often small and the extra computational effort is generally moderate even in medium-size problems.

References

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

Berk, R.A.: An introduction to ensemble methods for data analysis. Sociol. Methods Res. **34**(3), 263–295 (2006)

Breiman, L.: The heuristic of instability in model selection. Ann. Stat. **24**, 2350–2383 (1996)

Breiman, L.: Random Forests. Mach. Learn. **45**, 5–32 (2001a)

Breiman, L.: Statistical modeling: the two cultures. Stat. Sci. **16**, 199–231 (2001b)

Breiman, L.: Manual on setting up, using, and understanding Random Forests v3.1. Technical report (2002). http://oz.berkeley.edu/users/breiman

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

Breiman, L., Cutler, A., Liaw, A., Wiener, M.: Breiman and Cutler's Random Forests for classification and regression. R package version 4.5-18 (2006). http://cran.r-project.org/doc/packages/randomForest.pdf

Bühlmann, P., Yu, B.: Analyzing bagging. Ann. Stat. **30**(4), 927–961 (2002)

Dobra, A., Gehrke, J.: Bias correction in classification tree construction. In: Brodley, C.E., Danyluk, A.P. (eds.) Proceedings of the Seventeenth International Conference on Machine Learning, Williams College, Williamstown, MA, USA, pp. 90–97 (2001)

Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**, 1189–1232 (2001)

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

Hothorn, T., Hornik, K., Zeileis, A.: Unbiased recursive partitioning: a conditional inference framework. J. Comput. Graph. Stat. **15**(3), 651–674 (2006)

Kim, H., Loh, W.: Classification trees with unbiased multiway splits. J. Am. Stat. Assoc. **96**, 589–604 (2001)

Kononenko, I.: On biases in estimating multi-valued attributes. In: Mellish, C. (ed.) Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Montréal, Canada, pp. 1034–1040 (1995)

Liaw, A., Wiener, M.: Classification and regression by Random Forest. R News **2**(3), 18–22 (2002)

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

Murthy, K.: Automatic construction of decision trees from data: a multi-disciplinary survey. Data Min. Knowl. Discov. **2**(4), 1384–5810 (2004)

Nierenberg, D.W., Stukel, T.A., Baron, J.A., Dain, B.J., Greenberg, E.R.: Determinants of plasma levels of beta-carotene and retinol. Am. J. Epidemiol. **130**, 511–521 (1989)

Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Francisco (1988)

R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org. (2008)

Ridgeway, G.: Generalized boosted models: a guide to the gbm package. http://i-pensieri.com/gregr/papers/gbm-vignette.pdf (2007)

Sandri, M., Zuccolotto, P.: A bias correction algorithm for the Gini variable importance measure in classification trees. J. Comput. Graph. Stat. **17**(3), 1–18 (2008)

Schonlau, M.: Boosted regression (boosting): a tutorial and a stata plugin. Stata J. **5**(3), 330–354 (2005)

Shih, Y.-S.: Families of splitting criteria for classification trees. Stat. Comput. **9**, 309–315 (1999)

Strobl, C.: 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 (2005)

Strobl, C., Boulesteix, A.-L., Augustin, T.: Unbiased split selection for classification trees based on the Gini index. Comput. Stat. Data Anal. (2007a). doi:10.1016/j.csda.2006.12.030

Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T.: Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinf. **8**, 25 (2007b). doi:10.1186/1471-2105-8-25

[1] R codes of simulation experiments and real data analyses are downloadable from http://www.msandri.it/soft/TDNI.zip.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A.: Conditional variable importance for random forests. BMC Bioinf. **9**, 307 (2008). doi:10.1186/1471-2105-9-307

van der Laan, M.J.: Statistical inference for variable importance. Int. J. Biostat. **2**(1), 1–30 (2005)

White, A.P., Liu, W.Z.: Bias in information-based measures in decision tree induction. Mach. Learn. **15**, 321–329 (1994)

Wu, Y., Boos, D.D., Stefanski, L.A.: Controlling variable selection by the addition of pseudovariables. J. Am. Stat. Assoc. **102**(477), 235–243 (2007)