Statistical Sources of Variable Selection Bias in Classification Tree Algorithms Based on the Gini Index

Carolin Strobl⋆

Department of Statistics
University of Munich
Ludwigstr. 33, 80799 Munich, Germany
carolin.strobl@stat.uni-muenchen.de

Abstract. Evidence for variable selection bias in classification tree algorithms based on the Gini Index is reviewed from the literature and embedded into a broader explanatory scheme: Variable selection bias in classification tree algorithms based on the Gini Index can be caused not only by the statistical effect of multiple comparisons, but also by an increasing estimation bias and variance of the splitting criterion when plug-in estimates of entropy measures like the Gini Index are employed. The relevance of these sources of variable selection bias in the different simulation study designs is examined. Variable selection bias due to the explored sources applies to all classification tree algorithms based on empirical entropy measures like the Gini Index, Deviance and Information Gain, and to both binary and multiway splitting algorithms.

1 Introduction

The aim of this paper is to review previous publications on empirical evidence of variable selection bias from simulation studies by Kim and Loh (2001) and Dobra and Gehrke (2001), and to give a more broad statistical explanation of variable selection bias than those publications, integrating all sources of variable selection bias in tree algorithms based on the Gini Index.

Since the Gini Index is still the default splitting criterion in all classification tree algorithms based on the CART approach, i.e. the commercial version of `CART` by Salford Systems, the `tree` function in `S-Plus` and the `tree` and `rpart` functions in `R`, a thorough investigation of the properties of this splitting criterion is crucial. Additionally, mechanisms corresponding to the ones described in the following sections also hold for the Deviance and Information Gain criteria employed in the above and other classification tree algorithms (cp. Breiman et al., 1984; Quinlan, 1993). However, due to space constraints, these results have to be omitted here.

In the course of this paper Section 2 will provide the necessary background on splitting rules in classification tree algorithms. Section 3 will review empirical

⋆ I would like to thank Anne-Laure Boulestix for the stimulating discussion.

evidence of variable selection bias from simulation studies. These results will be explained by means of the statistical mechanisms underlying variable selection bias in Section 4.

2 Splitting rules in classification tree algorithms

Classification tree algorithms can be categorized by their splitting rules determined firstly by the choice of the number of nodes produced in each split and secondly by their split selection criterion. The number of nodes produced in each split can follow three rationales: An algorithm can produce binary splits (for continuous and categorical predictors), produce as many nodes as categories in the predictor selected for the current split (for categorical/categorized predictors) or produce as many nodes as categories in the response. Different split selection criteria include approaches of impurity reduction (based on empirical entropy measures) and statistical measures of association strength.

Here we will restrict the considered range of splitting rules to those with binary splits or multiway splits with as many nodes as categories in the predictor, and to those employing empirical entropy measures as split selection criteria. The standard classification tree algorithms work on this basis.

2.1 Binary vs. multiway splitting

All classification tree algorithms based on the CART (Breiman et al., 1984) approach produce binary splits in both categorical and continuous predictors. The split selection in these algorithms is performed in two steps:

1. Within the observed range of each predictor the cutpoint that minimizes the criterion value is selected.
2. The predictor variable for which the minimally selected criterion value is lowest is selected for the next split.

Other tree algorithms like `C4.5` by Quinlan (1993) perform multiway splits with as many nodes as categories in the predictor for categorical predictors.

It is important to note that in any splitting rule that produces less nodes than the number of values of the splitting variable, i.e. in binary splitting of categorical predictors with more than two categories and in splitting of continuous predictors, cutpoint selection is critical in split selection.

2.2 Impurity reduction

The approach of impurity reduction chosen in CART for split selection is based on the following idea: The nodes produced by a split should be more pure (in a yet to be defined manner) than the preceding node. The splitting variable that produces the highest impurity reduction with respect to the current node is selected for the next split.

In the following, we will treat the case of a binary response Y, for which $Y = c$, with $c \in \{1, 2\}$, denotes the class membership. Let X_j, $j = 1, \ldots, p$, denote categorical or continuous predictor variables. For the categorical predictors let $X_j = k$, with $k \in \{1, \ldots, K\}$, denote the category.

The following notation applies to the first split of the root node: The starting set is denoted by S_j, holding n_j observations not missing in the predictor variable X_j currently evaluated in split selection. The subsets S_{jL} and S_{jR} are produced by splitting S_j into two subsets at a cutpoint in the range of predictor X_j. For continuous predictors the subset S_{jL} results by means of splitting in cutpoint t_j and assigning all observations with $x_{ij} \leq t_j$ to S_{jL}, and the remaining to S_{jR}. For categorical predictors S_{jL} and S_{jR} characterize any binary partition of the categories. The empirical impurity reduction[1] is a function of these quantities:

$$\Delta\widehat{\mathfrak{I}}(S_j, S_{jL}, S_{jR}) = \widehat{\mathfrak{I}}(S_j) - \left[\frac{n_{jL}}{n_j} \cdot \widehat{\mathfrak{I}}_j(S_{jL}) + \frac{n_{jR}}{n_j} \cdot \widehat{\mathfrak{I}}_j(S_{jR})\right], \tag{1}$$

where $\widehat{\mathfrak{I}}(S_j)$ is the empirical impurity measure for the set S_j before splitting, while $\widehat{\mathfrak{I}}(S_{jL})$ is the empirical impurity measure for the subset S_{jL}. The proportion of observations assigned to subset S_{jL} is denoted as $\frac{n_{jL}}{n_j}$ (and respectively for the other subset S_{jR}).

As a possible empirical impurity measure $\widehat{\mathfrak{I}}(\cdot)$ Breiman et al. (1984) introduce the Gini Index. For two response classes the Gini Index (denoted here exemplarily for the subset S_{jL} to exploit the notation) reduces to

$$\widehat{G}(S_{jL}) = 2 \cdot \frac{n_{jL1}}{n_{jL}} \left(1 - \frac{n_{jL1}}{n_{jL}}\right) \tag{2}$$

where the relative frequency $\frac{n_{jL1}}{n_{jL}}$ denotes the proportion of class 1 individuals in the subset S_{jL}, which holds n_{jL} observations.

The empirical Gini Gain, analogous to the empirical impurity reduction in Equation 1, is

$$\Delta\widehat{G}(S_j, S_{jL}, S_{jR}) = \widehat{G}(S_j) - \left[\frac{n_{jL}}{n_j} \cdot \widehat{G}_j(S_{jL}) + \frac{n_{jR}}{n_j} \cdot \widehat{G}_j(S_{jR})\right]. \tag{3}$$

3 Empirical evidence of variable selection bias with the Gini Index

Recent publications on variable selection bias in classification tree algorithms based on the CART approach with the Gini Index provide empirical evidence from simulation studies documenting that the selection probability of a predictor variable is affected by features other than its discriminatory power.

[1] Note that throughout this paper all empirical quantities will be denoted as estimators of theoretical quantities by adding a hat to the symbol.

3.1 Study designs

Features found relevant for variable selection bias with empirical entropy measures like the Gini Index are:

1. The number of possible cutpoints for binary splits, which is determined by the number of categories (abbreviated as #categories in the following schemes) in categorical predictors or the number of distinct, non-missing observations (inverse to #missing) in continuous predictors. The number of cutpoints determines the number of comparisons of criterion values to be conducted (#tests).
2. The sample size (#sample), which is determined by the number of missing values (#missing) in each predictor.
3. The number of nodes (#nodes) produced in each split in the case of multiway splits in categorical predictors with different numbers of categories (#categories).

Table 1 gives an overview over the dependencies between these features. The number of categories in categorical predictors (#categories) and the number of missing values in continuous predictors (#missing) were experimentally varied between the available splitting variables in simulation studies on variable selection bias with the Gini Index. The corresponding study designs by Kim and Loh (2001) and Dobra and Gehrke (2001)[2] are resumed in Table 2.

	binary splits	multiway splits
# categories ↑ (categorical predictor)	#tests ↑ #nodes = #sample =	#tests = #nodes ↑ #sample =
# missing ↑ (continuous predictor)	#tests ↓ #nodes = #sample ↓	

Table 1: Dependencies between features found relevant for variable selection bias.

The variable selection performance of a split selection criterion can be evaluated by means of the following simulation study design: Several uninformative predictor variables are generated by random sampling. The predictor variables are sampled such that they only differ in one feature, which is expected to generate variable selection bias. The relative frequencies of simulations in which each variable is selected by the split selection criterion, out of the number of all simulations, are estimates for the selection probabilities, which should be equal (at

[2] Precursors of the study design of Dobra and Gehrke (2001) can be found in White and Liu (1994) and Kononenko (1995). However, Dobra and Gehrke (2001) give an additional statistical background.

	binary splits	multiway splits
# categories (categorical predictor)	Kim and Loh (2001)	Dobra and Gehrke (2001)
# missing (continuous predictor)	Kim and Loh (2001) Strobl (2004)	

Table 2: Study designs of simulation studies on variable selection bias.

random choice probability 1/number of variables) for uninformative predictor variables if no selection bias occurs.

Figure 1 displays estimated variable selection probabilities for the Gini Index. In this simulation study design the percentage of missing values in one of ten predictor variables is varied, while the rest of the variables remain complete. The variables are all uninformative.

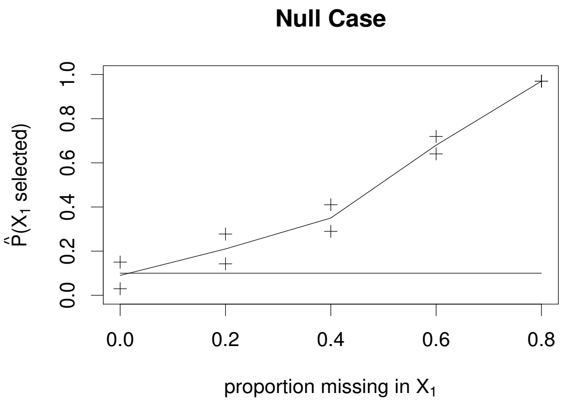

Fig. 1: Estimated variable selection probabilities for the Gini Index, the p-value adjusted risk criterion and the p-value adjusted Fisher criterion. All variables are uninformative.

3.2 Results

The results of the simulation studies were the following: Dobra and Gehrke (2001) show for multiway splits that variables with a higher number of categories are preferred in split selection. Kim and Loh (2001) report for binary splits that categorical predictors with a higher number of categories and continuous predictors with a higher number of misisng values are preferred in variable selection.

The latter results were replicated by Strobl (2004) (cp. again Figure 1): For uninformative predictor variables the estimated selection probability increases with the number of missing values in the regarded variable (and thus decreases in all other variables due to competition) when using the Gini Index for split selection, indicating variable selection bias. With an unbiased criterion the the estimated selection probability is supposed to remain at chance level.

4 Sources of variable selection bias

The publications on the empirical evidence presented in the previous section either lack a satisfactory statistical explanation of the mechanisms underlying the variable selection bias, or leave some results unnoticed.

Kim and Loh (2001) give a vague explanation for their finding that continuous predictors with a higher number of misisng values are preferred in variable selection: the authors state that if missing values randomly replace some of the observations, the Gini Index automatically decreases, because in the most extreme case where only one observation per node is not missing, the criterion takes on the minimum value 0, guaranteeing that the corresponding predictor variable is chosen in variable selection. However, in less extreme cases there can be both situations in which the criterion value decreases with randomly missing values, and situations in which the criterion value increases with randomly missing values. We choose a more appropriate probabilistic approach to explain this effect in Section 4.2.

Dobra and Gehrke (2001) do accurately accredit their findings to the statistical fluctuation in empirical entropy measures used in split selection. However, they do not interpret their computational results with respect to the statistical theory of estimation, and ignore results for binary splitting relevant for a wide range of classification tree algorithms.

In the following we want to point out that two mechanisms interact in variable selection bias in all CART-like classification tree algorithms: the effects of multiple comparisons in cutpoint selection and the effects of limited-sample plug-in estimation of the entropy measures.

4.1 Multiple comparisons

The common problem of multiple comparisons refers to an increasing type I error-rate in multiple testing situations. For variables with a higher number of possible cutpoints the probability of choosing an uninformative split by chance increases - as in any multiple testing situation, where multiple statistical tests are conducted for the same data set. In the context of split selection a type I error occurs when a variable is selected for splitting even though it is not informative, and the number of tests conducted increases with the number of distinct values of the predictor variable, which determines the number of possible cutpoints to be evaluated.

The problem of multiple comparisons is relevant in cutpoint selection, i.e. in any splitting rule that produces less nodes than the number of values of the splitting variable. Note here that Dobra and Gehrke (2001) state that the variable selection bias for categorical predictor variables was not due to multiple comparisons. However, the authors use the Gini Gain for multiway splits with as many nodes as categories in the predictor, which is not employed in any of the standard classification tree algorithms.

4.2 Estimation bias and variance

Empirical entropy measures as the Gini Index used in classification tree algorithms are naive plug-in estimators of the respective theoretical entropy measures. The plug in estimators are based on the relative class-frequencies as maximum-likelihood estimators of the class probabilities. The quality of an estimator $\hat{H}$ for the theoretical entropy measure H can be evaluated by its estimation bias and variance. The bias of an estimator $Bias_H(\hat{H}) = E_H(\hat{H}) - H$ is the deviation of the expected value of the estimator $\hat{H}$ from the true value H. We will show in the following, that the plug-in estimator of the Gini Index and the derived Gini Gain is biased.

For variable selection bias in classification tree algorithms it is relevant that the sample size and number of nodes produced in each split affect both the expected value and variance of the splitting criterion.

Bias of the empirical Gini Index

We derive the expected value of the empirical Gini Index with respect to the true class 1 probability p_1 for the exemplary subset S_{jL} of size n_{jL}:

$$\begin{aligned} E_{p_1}\left(\widehat{G}(S_{jL})\right) &= G - 2\,\frac{p_1(1-p_1)}{n_{jL}} \\ &= G\,\frac{n_{jL}-1}{n_{jL}}. \end{aligned}$$

The empirical Gini Index underestimates the true Gini Index by factor $\frac{n_{jL}-1}{n_{jL}}$. From the line before the last it also becomes obvious that the estimation bias is $Bias_G(\widehat{G}) = E(\widehat{G}) - G = -\frac{G}{n_{jL}}$. The estimation bias increases for small sample sizes n_{jL}. The effect is most pronounced for $p_1 \rightarrow 0.5$, which is reasonable because in this case the true impurity can only be underestimated. Note however, that the true response class probabilities do not vary between predictors, and thus variable selection bias does not rely on the class probabilities.

Bias of the empirical Gini Gain

Under the the null hypothesis of an uninformative predictor X_j the true Gini Index $G = 2p_1(1-p_1)$, depending on the true class 1 probability p_1, is supposed to be equal in each subset. Thus, the true Gini Gain $\Delta G(S_j, S_{jL}, S_{jR})$ abbreviated in the following by ΔG, which is the Gini Index of the complete set minus a weighted sum of the true Gini Indices in the subsets (which is equal to the Gini Index of the complete set under the null hypothesis, cp. Equation 3) is equal to 0. However, the expected value of the empirical Gini Gain $\Delta\widehat{G}(S_j, S_{jL}, S_{jR})$ abbreviated by $\Delta\widehat{G}$ is

$$
\begin{aligned}
E_{p_1}\left(\Delta\widehat{G}\right) &= G - \frac{2p_1(1-p_1)}{n_j} - \\
&- \left[\frac{n_{jL}}{n_j}\cdot\left(G - \frac{2p_1(1-p_1)}{n_{jL}}\right) + \frac{n_{jR}}{n_j}\cdot\left(G - \frac{2p_1(1-p_1)}{n_{jR}}\right)\right] \\
&= 2\,\frac{p_1(1-p_1)}{n_j}
\end{aligned}
$$

overestimating the true Gini Gain under the null hypothesis by the value of $Bias_{\Delta G}(\Delta\widehat{G}) = E(\Delta\widehat{G}) - \Delta G = \frac{G}{n_j}$ independent of the partition.

It is plain to see that the same principle applies in classification tree algorithms with multiway splits, where the bias increases with the number of categories of the splitting variable determining the number of nodes produced in each split: for each additionally created node the bias increases by adding $\frac{G}{n_j}$.

Our results on the expected value of the empirical Gini Gain correspond to those of Dobra and Gehrke (2001) adopted for binary splits. However, the authors do not elaborate on the interpretation as an estimation bias induced by the plug-in estimation based on a limited sample size.

We have shown above that the estimation bias for the empirical Gini Gain is a relevant source of variable selection bias in multiway splits (cp. the results of Dobra and Gehrke (2001) reviewed in Section 3), when variables differ in their number of categories. However, the effect of estimation bias for the empirical Gini Gain applies in binary splitting only if the overall sample size n_j varies between variables, i.e. in simulation study designs with missing values (cp. the results of Kim and Loh (2001) and Strobl (2004) also reviewed in Section 3).

Variance of the empirical Gini Index

The variance of the empirical Gini Index can be approximated by means of the delta method (cp. e.g. Rice, 1995): In the following notation the empirical Gini Index, again computed for the exemplary subset S_{jL} of size n_{jL}, is considered as a function of the unbiased estimator $\hat{p}_1 = \frac{n_{jL1}}{n_{jL}}$ of the true class 1 probability p_1 with $E(\hat{p}_1) = p_1$. The first derivative $G'(p_1)$ (from a truncated Taylor series expansion of $G(\hat{p}_1)$ about p_1 for a linear approximation) applied to $\hat{p}_1$ is used in the delta method to approximate the variance of of the empirical Gini Index by $\widehat{Var}(G(\hat{p}_1)) \cong [G'(\hat{p}_1)]^2\, Var(\hat{p}_1)$. Thus, under the null hypothesis the variance of the empirical Gini Index in each node increases with decreasing sample size:

$$\widehat{Var}(G(\hat{p}_1)) \cong \left[4p_1 - 16p_1^2 + 16p_1^3\right]\frac{(1-p_1)}{n_{jL}}.$$

Dobra and Gehrke (2001) derived the variance of the empirical Gini Gain. From their results we can see that the variance of the empirical Gini Gain depends on the overall sample size n_j in the denominator. Note that in the multiway split case derived in Dobra and Gehrke (2001) the variance depends also on the number of nodes produced in multiway splits.

5 Conclusions

The empirical Gini Gain is derived from the emprirical Gini Index as a splitting criterion in classification tree algorithms based on the CART approach. The true Gini Index is underestimated by the empirical Gini Index, which is a plug-in estimator based on the relative class frequencies as estimators of the class probabilities. Under the null hypothesis that the predictor variable is uninformative for the response, we could show that the empirical Gini Gain overestimates the true Gini Gain. The estimation bias increases with a decreasing sample size for each splitting variable, and a higher number of nodes produced in each split. The effect of estimation bias is thus relevant for variable selection bias in classification tree algorithms in the case of different amounts of missing values in the predictors or for multiway splits when the predictors vary in their number of categories. In addition we have seen that the variance of the empirical Gini Index increases with decreasing sample size. The variance of the empirical Gini Gain also increases with decreasing sample size and increasing number of nodes created in each split for multiway splits (Dobra and Gehrke, 2001). For binary splits multiple comparisons are another relevant source of variable selection bias.

Considering again Tables 1 and 2 we conclude that the results on variable selection bias displayed there can be explained in the following way: For multiway splits producing as many nodes as categories of the predictor, and the categorical predictors varying in their number of categories as simulated in Dobra and Gehrke (2001), sources of variable selection bias are the estimation bias and variance of the splitting criterion. Both increase with the number of nodes produced in each split. For binary splits, with the categorical predictors varying in their number of categories as simulated in Kim and Loh (2001), the source of variable selection bias is exclusively due to the effect of multiple comparisons. For binary splits, with the metric predictors varying in their number of missing values as also simulated in Kim and Loh (2001), two possible sources of variable selection bias seem to counteract: For binary splits the multiple comparisons effect in cutpoint selection is supposed to penalize variables with more missing values and thus less possible cutpoints. On the other hand, a decrease in sample size due to the missing values promotes the estimation bias and variance of the splitting criterion.

The results of Kim and Loh (2001) and Strobl (2004) document that variables with more missing values are preferred rather than punished. We conclude that the overestimation of the splitting criterion and the greater variance of the criterion values in variables with more missing values outbalance the disadvantage in multiple comparisons. Under the null hypothesis extreme criterion values are still more likely to be detected by chance in variables with more missing values due to their increase in bias and variance.

Bibliography

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). *Classification and Regression Trees*. New York: Chapman and Hall.

Dobra, A. and J. Gehrke (2001). Bias correction in classification tree construction. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 90–97. Morgan Kaufmann.

Kim, H. and W. Loh (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association 96*, 589–604.

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

Quinlan, R. (1993). *C4.5: Programms for Machine Learning*. San Francisco: Morgan Kaufmann Publishers Inc.

Rice, J. (1995). *Mathematical statistics and data analysis*. Belmont: Duxbury Press.

Strobl, C. (2004). Variable selection bias in classification trees. *www.stat.uni-muenchen.de/∼ carolin/MA_homepage.ps*, (unpublished Master–Thesis).

White, A. and W. Liu (1994). Bias in information based measures in decision tree induction. *Machine Learning 15*, 321–329.