Strobl, Carolin

Working Paper

Variable selection bias in classification trees based on imprecise probabilities

Discussion Paper, No. 419

Provided in Cooperation with:
Collaborative Research Center (SFB) 386: Statistical Analysis of discrete structures - Applications in Biometrics and Econometrics, University of Munich (LMU)

Suggested Citation: Strobl, Carolin (2005) : Variable selection bias in classification trees based on imprecise probabilities, Discussion Paper, No. 419, Ludwig-Maximilians-Universität München, Sonderforschungsbereich 386 - Statistische Analyse diskreter Strukturen, München, https://doi.org/10.5282/ubm/epub.1788

This Version is available at:
https://hdl.handle.net/10419/31010

Variable Selection Bias in Classification Trees Based on Imprecise Probabilities

Carolin Strobl
Department of Statistics
University of Munich
Akademiestrasse 1, 80799 Munich, Germany
carolin.strobl@stat.uni-muenchen.de

Abstract

Classification trees based on imprecise probabilities provide an advancement of classical classification trees. The Gini Index is the default splitting criterion in classical classification trees, while in classification trees based on imprecise probabilities, an extension of the Shannon entropy has been introduced as the splitting criterion. However, the use of these empirical entropy measures as split selection criteria can lead to a bias in variable selection, such that variables are preferred for features other than their information content. This bias is not eliminated by the imprecise probability approach. The source of variable selection bias for the estimated Shannon entropy, as well as possible corrections, are outlined. The variable selection performance of the biased and corrected estimators are evaluated in a simulation study. Additional results from research on variable selection bias in classical classification trees are incorporated, implying further investigation of alternative split selection criteria in classification trees based on imprecise probabilities.

Keywords. Classification trees, credal classification, variable selection bias, attribute selection error, Shannon entropy, entropy estimation.

1 Introduction

Classification trees are a means of non-parametric regression analysis for predicting the value of a categorical response variable Y from the values of categorical or continuous predictor variables $X_1, \ldots, X_p$. In comparison to other classical classification procedures such as the linear discriminant analysis or logistic regression the prominent advantages of classification trees are the nonparametric and nonlinear approach and the straightforward interpretability and applicability of the results.

The extension of classification trees as credal classifiers based on imprecise probabilities by Abellán and Moral (2004) establishes a more sensitive means of classification, that is not as susceptible to overfitting as classical classification trees requiring terminal pruning (Section 2).

Classification tree algorithms are specified by their split selection criterion, which controls variable selection, and the number of splits they produce in each node. Some authors favor binary splits (e.g. Breiman, Friedman, Olshen, and Stone, 1984, as implemented in the statistical programming tools `CART`® and `R`), while others favor multiway splits (e.g. Abellán and Moral, 2004; Quinlan, 1993, as implemented in `C4.5`). In case of binary splits, i.e. if not as many nodes as categories of the categorical predictor used for splitting are created in each split, and generally for metric predictor variables, the split selection criterion also determines the cutpoint selection.

For classical classification and regression trees (CART) research has revealed that split selection criteria can be biased in variable selection, preferring variables for features other than their information content (Section 3).

Sources of variable selection bias, also termed attribute selection error in the literature, are firstly multiple testing effects in cutpoint selection in binary splitting algorithms, and secondly effects of sample size in both binary and multiway splitting. This work concentrates on the effects of sample sizes on variable selection bias.

Differences in sample size between two predictor variables can be caused either by different numbers of categories, leaving smaller numbers of observations in each node, or by missing values. While the latter problem of missing values can be handled within the imprecise probabilities framework (Zaffalon, 2002; de Cooman and Zaffalon, 2004), a solution for the for-

mer problem of different numbers of categories is yet to be found.

The problem of different numbers of categories affects variable selection when empirical measures of entropy serve as split selection criteria. This is the case both in classical classification tree approaches (Breiman et al., 1984), where the empirical Gini Index is used to evaluate the amount of entropy in each node, and in recent approaches employing imprecise probabilities (Abellán and Moral, 2004), where the empirical Shannon entropy is employed.

We will show that the source of variable selection bias for the empirical Shannon entropy is an estimation bias. The bias is due to the fact that the empirical Shannon entropy is a plug-in estimator based on relative frequencies of observations as estimators for the class probabilities (Sections 2 and 4).[1]

Corrected estimators will be discussed and evaluated in simulation studies investigating the variable selection performance of the biased and corrected estimators in classification trees based on imprecise probabilities (Section 4).

Additional results of simulation studies comparing the variable selection performance of the Gini Index to the performance of alternative split selection criteria based on p-values in classical classification trees are displayed (Section 5), and an outlook on transferring the results on unbiased split selection in classical classification trees to classification trees based on imprecise probabilities is given (Section 6).

2 Split selection in classification trees based on imprecise probabilities

Abellán and Moral (2004) present a measure of entropy for credal sets as a split selection criterion in classification trees based on imprecise probabilities. Their impurity criterion for the credal set $\mathscr{P}$ defined on the finite set K of values $k = 1, 2, \ldots, |K|$ of the response variable Y with $p(k) := p(Y = k)$

$$TU2(\mathscr{P}) = \max_{p \in \mathscr{P}} \left\{ -\sum_{k=1}^{|K|} p(k)\ ln[p(k)] \right\} \qquad (1)$$

is a generalization of the popular Shannon entropy (Shannon, 1948) for classical probabilities.

[1] Estimators for classical probabilities will be denoted as $\hat{p}(\cdot)$ in the following, while the true probabilities will be denoted as $p(\cdot)$. To distinguish between classical probabilities and interval-valued probabilities, the latter will be denoted as capital $P(\cdot)$.

The authors have previously suggested the total impurity criterion

$$TU1(\mathscr{P}) = TU2(\mathscr{P}) + IG(\mathscr{P}), \qquad (2)$$

where $IG(\mathscr{P})$ is a measure of non-specificity with

$$IG(\mathscr{P}) = \sum_{A \subseteq K} m_{\mathscr{P}}(A)\ ln(|A|)$$

where $m_{\mathscr{P}}$ is the Möbius inverse of the lower probability function $f_{\mathscr{P}}$

$$m_{\mathscr{P}}(A) = \sum_{B \subseteq A} (-1)^{|A-B|}\ f_{\mathscr{P}}(B),$$

and $|A - B|$ is the cardinality of the set A excluding B.

$IG(\mathscr{P})$ is a generalization of the Hartley measure of non-specificity $I(A) = log_2(|A|)$ (in bits). Here, the finite set A includes all possible candidates for a true class. Thus, the non-specificity of the characterization increases with the cardinality of the set of possible alternatives (cp. Klir, 1999, 2003).

The total impurity measure $TU1(\mathscr{P})$ additively incorporates both uncertainty and non-specificity. However, Abellán and Moral (2004) settle for $TU2(\mathscr{P})$ as a measure of total uncertainty, arguing that $TU2(\mathscr{P})$ also increases with non-specificity. The authors thus conclude that adding a measure of non-specificity as in $TU1(\mathscr{P})$ overweighs non-specificity in the total impurity criterion.

Technically, the maximization in $TU2(\mathscr{P})$ is accomplished by means of the maximum entropy algorithm introduced in Abellán and Moral (2003). The algorithm identifies the posteriori probability distribution on K with the maximum entropy that is in accordance with the upper and lower probabilities for each class $k \in K$, which are derived from the Imprecise Dirichlet Model (IDM) (Walley, 1996). The Shannon entropy is then applied to the posteriori maximum entropy distribution.

Abellán and Moral (2004) apply the IDM locally to subsets of the data defined by configurations of the predictor variables. For each subset, defined by predictor variable configuration σ, the calculation of the lower and upper probabilities with the IDM is based on counts of n_k^σ class k objects out of N^σ objects in total in the subset defined by σ:

$$P(k) = [\underline{P}(k), \overline{P}(k)] = \left[\frac{n_k^\sigma}{N^\sigma + s}\ ,\ \frac{n_k^\sigma + s}{N^\sigma + s} \right],$$

where s denotes the hyperparameter of the IDM, interpretable as the number of yet unobserved observations. Taking this interpretation of s literally, the calculation of the lower and upper probabilities is based on relative frequencies excluding, and respectively including, s additional observations of class k.

3 Experiences from split selection in classical classification trees

Recent publications on classical CART address the problem of variable selection bias, indicating that for some split selection criteria the selection probability of a predictor variable is affected by features other than its discriminatory power.

Features relevant for variable selection bias are the number of observations assigned to subsequent nodes, from which the criterion value is calculated, for binary as well as for multiway splits, and the number of possible cutpoints for binary splits. Both features are affected by the number of categories in each categorical predictor and by the number of missing values (or ties) in each metric predictor.

In categorical predictors a higher number of categories leaves less observations in each node, and provides more possible cutpoints, while in metric predictor variables missing values also leave less observations in each node, but provide less possible cutpoints. A higher number of possible cutpoints can produce multiple testing effects, while a lower number of observations in each node affects the quality of criterion estimates.

Loh and Shih (1997) present numerical evidence for variable selection bias with the Pearson χ^2- (metric predictor with bisecting cutpoint, binary response; see also (Shih, 2004)) and F-statistic (categorical predictor with bisecting cutpoint, metric response). In both cases predictor variables with more distinct values or classes are preferred, while predictor variables with less distinct values are penalized in variable selection.

The authors accredit their findings of selection bias to an increasing type I error-rate in multiple testing situations: for the search algorithms used in CART 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.

An inverse effect of variable selection bias has been reported for the Gini Index as early as Breiman et al. (1984). The numerical evidence confirms that for the Gini Index split selection in classification trees is biased toward selecting variables with a lower number of distinct values (caused by different numbers of categories in in categorical predictors and by missing values in metric predictors in Kim and Loh (2001), by missing values and different numbers of categories in categorical predictors in Dobra and Gehrke (2001) and by missing values in metric predictors in Strobl (2004)).

However, the Gini index is still the default split selection criterion for frequentist classification trees in statistical programming tools such as `CART`® and `R`.

4 Entropy measures in split selection

The Gini Index, as Shannon's entropy, is a theoretical entropy measure that suffices the following desirable properties (depicted in Figure 1 for the two-class case):

1. Pure sets (with all but one class probability equal to zero) have minimum entropy.
2. Maximally impure sets (with all class probabilities equal) have maximum entropy.
3. The impurity function is continuous and concave, with the slope increasing with the distance from the equilibrium point.

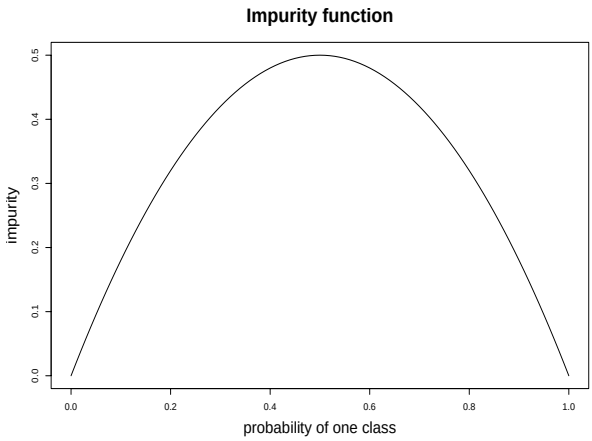

Figure 1: Desirable shape of an impurity function in the two class case.

However, the properties of these theoretical entropy measures are not self-evidently passed over to their empirical equivalents. Therefore, the quality of potential estimators for the entropy measures has to be assessed before using them in applications such as split selection in classification trees.

4.1 Estimation bias for empirical entropy measures

To evaluate a split in the course of variable selection in a classification tree, an estimate of the empirical entropy measure is computed in each node. A weighted sum of these estimates then describes the empirical entropy induced by the split, with the relative frequencies of observations per node as the weights. A variable is selected for splitting if the empirical entropy induced by the split is sufficiently small as compared to the empirical entropy before splitting.

The estimator for the Gini Index in classical classification trees, as well as the estimator for the Shannon entropy, used by Abellán and Moral (2004) in classification trees based on imprecise probabilities and Chiang and Hsu (2002) in fuzzy classification trees, are plug-in estimators based on the relative class-frequencies as maximum-likelihood estimators of the class probabilities.

When using these plug-in estimators for the Gini Index and the Shannon entropy for variable selection in classification trees, variable selection is biased towards variables with less observations per node. The variable selection bias is due to the biased small-sample estimator for the empirical entropy measures: With a decreasing number of observations per node the standard error of the estimators increases, producing posterior class distributions misleadingly implying a higher amount of information.

The same mechanism takes effect in the approach of Abellán and Moral (2004), where the biased Shannon entropy estimator is applied to the posterior maximum entropy distribution derived from the IDM, the fluctuation of which is also due to statistical deviation in the random variables for the class counts in small samples.

The bias becomes relevant for variables with more categories and hence less observations per category, or for variables with missing data. A solution for the latter case can be derived from the approach of Zaffalon (2002) (q.v. de Cooman and Zaffalon, 2004) within the imprecise probabilities framework. However, the remaining problem of different numbers of categories has to be addressed. Based on a statistical evaluation of the bias, possible correction strategies are derived in the following:

Both Miller (1955) and Basharin (1959) independently derived the expected value of the plug-in estimate $\widehat{H}$ for the Shannon entropy H

$$\begin{aligned} E_p\left(\widehat{H}\right) &= E_p\left(-\sum_{k=1}^{|K|}\widehat{p}\left(k\right)\, ln[\widehat{p}\left(k\right)]\right) \\ &= E_p\left(-\sum_{k=1}^{|K|}\ \frac{n_k}{N}\ ln\left[\frac{n_k}{N}\right]\right) \\ &= H - \frac{k-1}{2N} + O\left(\frac{1}{N^2}\right), \end{aligned}$$

where $O(\frac{1}{N^2})$ includes terms of order $\frac{1}{N^2}$, which are suppressed in the following naive correction approach because they depend on the true class probabilities $p(k)$ (cp. Schürmann, 2004).

Since this estimation bias applies to any classical probability distribution, it applies analogously to the posterior maximum entropy distribution derived from a credal set by means of the maximum entropy algorithm (Abellán and Moral, 2003) employed in Abellán and Moral (2004).

According to the above assessment of the estimation bias a naive correction approach for an unbiased estimate $\widehat{H}_{\text{Miller}}$ as suggested by Miller (1955) is

$$\widehat{H}_{\text{Miller}} = \widehat{H} + \frac{|K|-1}{2N}.$$

4.2 Suggested corrections based on the IDM

As an empirical entropy estimator in every node of a classification tree based on imprecise probabilities in accordance to Abellán and Moral (2004), i.e. for every predictor value configuration σ, we suggest

$$\widehat{H}_{\text{Miller}} = \widehat{H} + \frac{|K|-1}{2(N^\sigma + s)}, \tag{3}$$

as a corrected estimator of the Shannon entropy. This correction accounts for the derivation of the posterior maximum entropy distribution, to which the entropy estimator is applied, from the posterior lower and upper probabilities computed with respect to the imprecise Dirichlet model with hyperparameter s (cp. Section 2). The correction seems appropriate for large N^σ, while it over-penalizes for small N^σ with respect to the number of categoies $|K|$, which is supported by the numerical results in Section 4.3.

In another correction approach we are revisiting the empirical measure $\widehat{IG}(\mathscr{P})$, the theoretical analogy of which was employed by Abellán and Moral (2004) as a measure of non-specificity in the total impurity criterion $TU1(\mathscr{P})$ (cp. Equation 2 in Section 2). Like

the correction term in the above approach $\widehat{IG}(\mathscr{P})$ is a function of the sample size N^σ and the number of categories $|K|$.

In the special case where the lower probabilities used in the computation of the Möbius inverses in $\widehat{IG}(\mathscr{P})$ are derived from the IDM, the Möbius inverses of all subsets of the power set of K, besides the sigletons $k \in K$ and the complete set K, are equal to zero due to the additivity induced by the IDM. Again taking the interpretation of s as the number of yet unobserved observations literally, the basic probability assignment $m_{\mathscr{P}}(\cdot)$ is greater than zero only for the sigletons, due to the N^σ out of $N^\sigma + s$ observations for which one class $k \in K$ was observed, and for the complete set K, due to the s out of $N^\sigma + s$ yet unobserved observations for which any class $k \in K$ can be observed.

Because the logarithm of the cardinality of the singletons is zero, the Möbius inverse for the set K collapses to the width $\frac{s}{N^\sigma+s}$ of the lower and upper probabilities on K computed from the IDM with hyperparameter s, and the empirical non-specificity measure $\widehat{IG}(\mathscr{P})$ depends only on the sample size N^σ through the interval width, and on the number of categories $|K|$ through the factor $ln(|K|)$. We thus suggest

$$\widehat{H} + \widehat{IG} = \widehat{H} + \widehat{m}_{\mathscr{P}}(K)\ ln(|K|) = \widehat{TU1}(\mathscr{P}) \qquad (4)$$

as another corrected estimator, where $\widehat{m}_{\mathscr{P}}(K)$ is the Möbius inverse computed from the posterior lower class probabilities derived from the IDM. We will again see in Section 4.3 that this correction is only reliable for large N^σ and small $|K|$, while otherwise it is overcautious.

4.3 Simulation study: performance of entropy estimators in split selection

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

The relative frequencies can sum up to values greater than 1 if more than one variable reaches the minimum criterion value, i.e. if more than one variable is equally appropriate to be selected, in one simulation, which is more likely for small sample sizes. (In a tree building algorithm one variable has to be randomly chosen for splitting in this case.)

The following results are from a simulation study run with 1000 simulations and 10 uninformative predictor variables, one of which has 3 (respectively 5) distinct categories, while the rest have 2 distinct categories. The value of the hyperparameter s of the IDM was set equal to 1.

In this study, the behavior of the plug-in estimator $\widehat{H}$ for the Shannon entropy (cp. Equation 1) is compared to the behavior of the corrected estimators $\widehat{H}_{\text{Miller}}$ (Equation 3) and $\widehat{H}+\widehat{IG}$ (Equation 4 or cp. Equation 2) for medium sample sizes N ($n_1 = n_2 = 100$ class 1 and 2 observations) and small sample sizes N ($n_1 = n_2 = 10$ class 1 and 2 observations).

Figures 2 through 5 display that, with the plug-in estimate $\widehat{H}$ for the Shannon entropy, variable selection bias affects the estimated selection probabilities even if the variables differ in their number of categories only by 1. This effect is strongly aggravated if the variables differ more in their number of categories.

For the corrected estimate $\widehat{H}_{\text{Miller}}$, Figures 6 through 9 document that the variable selection bias caused by the estimation bias of the entropy estimate can be fairly compensated by the correction. Only for small sample sizes, aggravated by a large difference in the number of categories of the predictor variables, the correction is overly cautious, resulting in a reverse variable selection bias.

For the corrected estimate $\widehat{H}+\widehat{IG}$, Figures 10 through 13 show that the reverse bias for small sample sizes and large difference in the number of categories is even stronger than for $\widehat{H}_{\text{Miller}}$.

4.4 Alternative entropy estimators

Alternative estimators for the Shannon entropy have been suggested e.g. by Pöschel et al. (2003), who employ rank ordered probabilities in entropy estimation, and Grassberger (2003), who's approach is based on the assumption of Poisson-distributed frequencies in small samples.

Another approach that might provide the opportunity to be extended to split selection in classification trees based on imprecise probabilities is the Bayesian entropy estimator introduced by Holste, Grosse, and Herzel (1998). The Bayesian estimator of the Shan-

non entropy (in bits) is

$$\widehat{H}_{\text{Bayes}} = \frac{1}{ln(2)} \sum_{k=1}^{K} \frac{n_k + s\, t_k}{N+s} \left(\sum_{j=n_k+st_k+1}^{N+s} \frac{1}{j} \right)$$

where the weights t_k are the parameters of a Dirichlet prior distribution with hyperparameter s for the $k = 1, 2, \ldots, |K|$ class probabilities. The Bayesian estimate is least biased for a uniform prior.

5 P-value adjusted split selection in classical CART

A different approach to prevent variable selection bias promoted for classical CART is the use of exact or approximated p-values of association measures as split selection criteria, e.g. of Fisher's exact test statistic, where the sample size is incorporated in the degrees of freedom of the hypergeometric distribution.

Some results of a simulation study on such a p-value approach for split selection in classification trees are outlined in the following. In this study the variable selection performance of the Gini Index was compared to the performance of the exact p-value of a risk statistic derived from statistical decision theory, which was designed to account for asymmetric misclassification costs, and the statistic of Fisher's exact test as split selection criteria in classification trees (for details see Strobl, 2004).

5.1 Simulation study: performance of Gini Index and p-value criteria in split selection

Figures 2 and 3 display estimated variable selection probabilities for the Gini Index and the two p-value adjusted split selection criteria. 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 either all uninformative (Figure 2) or one variable with no missing values and one variable with missing values is informative (Figure 3).

The results support the expected behavior for the Gini Index and p-value adjusted split selection criteria:

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 the p-value adjusted split selection criteria the estimated selection probability does not exceed random choice level.

For informative predictor variables the estimated selection probability again increases with the number of missing values in the regarded variable when using the Gini Index. With the p-value adjusted split selection criteria the estimated selection probability decreases with the number of missing values in the regarded variable (and thus increases in all other variables due to a lack of competition), because of the decrease of information inherent in the sample.

Thus, in Strobl (2004) we conclude that the use of the p-value adjusted split selection criteria is advisable in classical classification trees in order to avoid variable selection bias.

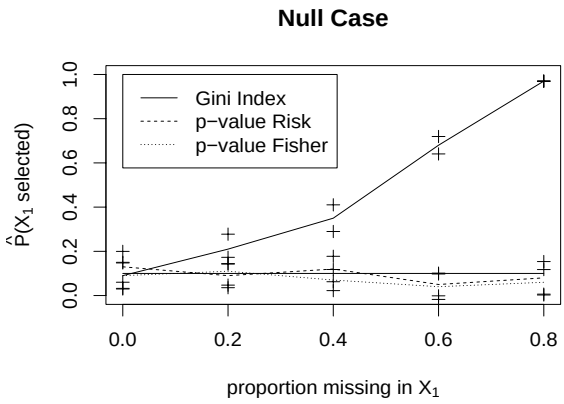

Figure 2: Estimated variable selection probabilities for the Gini Index, the p-value adjusted risk criterion and the p-value adjusted Fisher criterion. All variables are uninformative.

5.2 Detour: minimally or maximally selected statistics in cutpoint selection

For binary splits and for metric predictor variables, variable selection is conducted by comparing the criterion value for the cutpoint that minimizes or respectively maximizes the criterion value. The split selection therefore consists firstly of cutpoint selection and secondly of variable selection on the basis of the optimally selected cutpoint.

The distribution of a minimally or maximally selected statistic used as criterion value in split selection is not equivalent to the statistic's original null-distribution, because the cutpoint was not set a priori but chosen a posteriori so as to minimize or maximize the statistic. Adequate significance levels have to be derived from

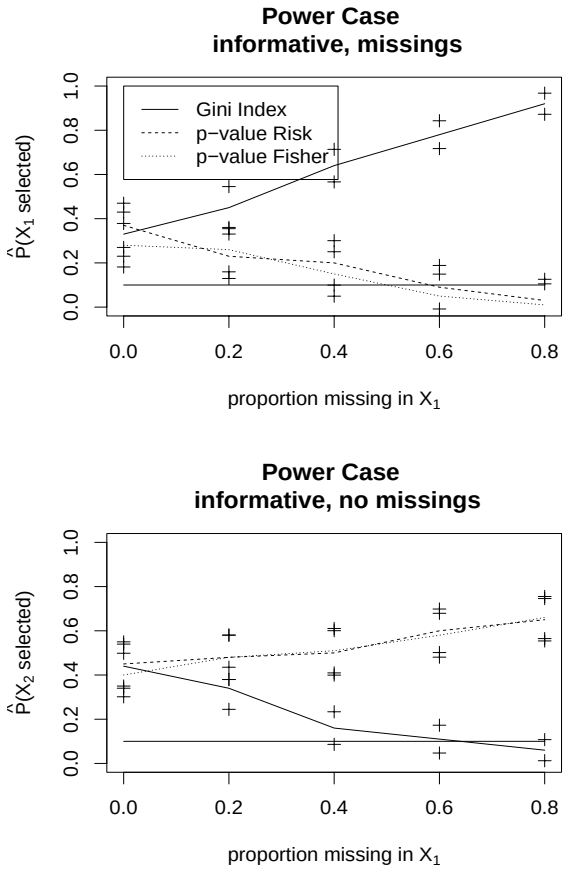

Figure 3: Estimated variable selection probabilities for the Gini Index, the p-value adjusted risk criterion and the p-value adjusted Fisher criterion. One variable with missing values and one variable with no missing values is informative.

an exact or approximated distribution (cp. Miller and Siegmund, 1982).

The approach of using the p-value of a maximally or minimally selected statistic in split selection avoids the multiple testing effects relevant in cutpoint selection. Due to the fact that the derived distributions of the maximally or minimally selected statistic incorporate the sample size, this approach also prevents sample size effects.

Consider e.g. the χ^2-statistic, for which variable selection bias has been documented in Loh and Shih (1997): neither the statistic nor the original p-value take the sample size into account, because the degrees of freedom of the original χ^2-distribution rely only on the number of cells in the contingency table. However, the exact distribution of the maximally selected χ^2-statistic derived by Koziol (1991) does depend on the sample size. The exact p-value of the distribution of the maximally selected χ^2-statistic is employed for split selection in classical classification trees by (Shih, 2004).

In the simulation study introduced above (Strobl, 2004), the main focus was on sample size effects. The results show that sample size effects can be eliminated by means of the p-value from the distribution of the minimally selected risk statistic derived by Gail and Green (1976), incorporating both the minimally selected character of the statistic and the sample size, or by means of the p-value of Fisher's exact test (cp. Martin, 1997), the original exact distribution of which incorporates sample size as well. Halpern (1999) also derived the exact p-value of the minimally selected statistic of Fisher's exact test.

For multiway splits in categorical predictors as used in Abellán and Moral (2004) the outlined effects of cutpoint selection are not relevant. However, the topic of minimally and maximally selected statistics, as the potential bias induced by multiple testing in general, has to be considered when the classification trees are extended to splits in metric predictor variables, or if not as many nodes as categories of the categorical predictor used for splitting are created in each split.

6 Discussion and perspective

We have seen that the use of biased estimators for entropy measures as the Gini Index and the Shannon entropy in the tradition of classification trees, both classical and based on imprecise probabilities, leads to variable selection bias.

Our results imply the use of corrected estimators for the Shannon entropy as split selection criterion in classification trees based on imprecise probabilities.

The corrected estimator $\widehat{H}_{\text{Miller}}$ in Equation 3 shows even better variable selection performance than the corrected estimator $\widehat{H} + \widehat{IG}$ in Equation 4. Both corrected estimators are less reliable for small sample sizes and large numbers of categories of the predictor variables, where they react overcautious. The corrected estimators can be easily applied to the posterior maximum entropy distribution derived from the lower and upper probabilities computed with the IDM as suggested by (Abellán and Moral, 2004). More elaborate entropy estimators can be considered for split selection in future research.

For classification trees based on imprecise probabilities another notion for research on unbiased split selection could evolve from the results of Bernard

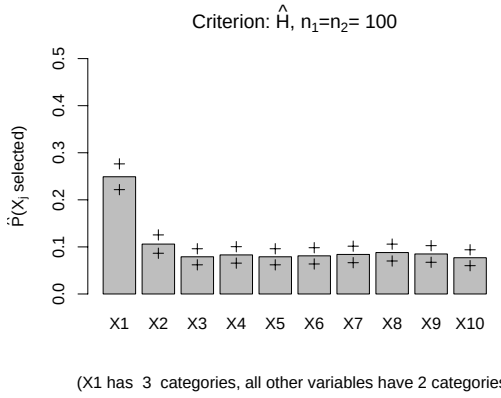

Figure 2: Estimated variable selection probabilities for the plug-in estimator of the Shannon entropy for 3 vs. 2 categories in the predictor variables and medium sample sizes.

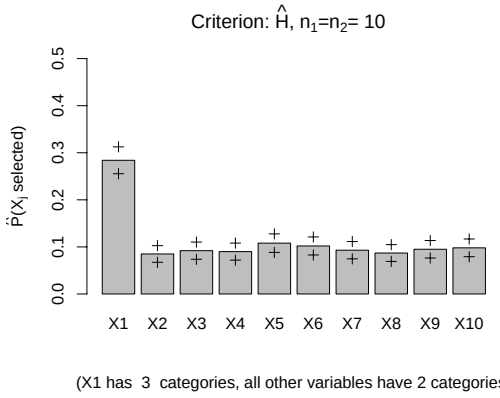

Figure 3: Estimated variable selection probabilities for the plug-in estimator of the Shannon entropy for 3 vs. 2 categories in the predictor variables and small sample sizes.

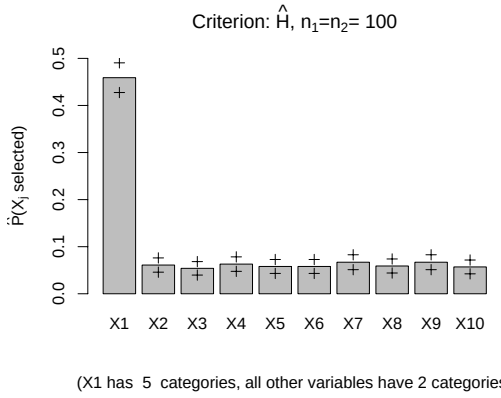

Figure 4: Estimated variable selection probabilities for the plug-in estimator of the Shannon entropy for 5 vs. 2 categories in the predictor variables and medium sample sizes.

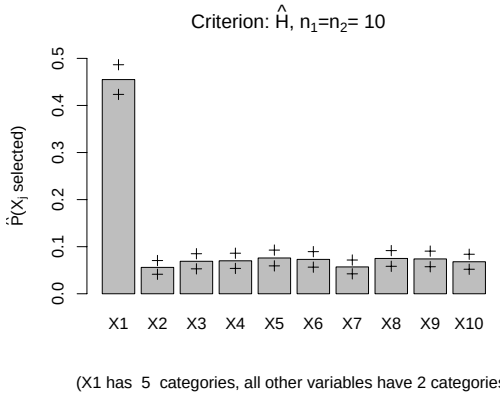

Figure 5: Estimated variable selection probabilities for the plug-in estimator of the Shannon entropy for 5 vs. 2 categories in the predictor variables and small sample sizes.

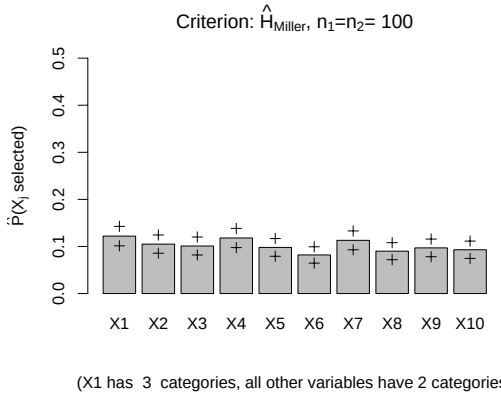

Figure 6: Estimated variable selection probabilities for the corrected estimator of the Shannon entropy $\widehat{H}_{\text{Miller}}$, for 3 vs. 2 categories in the predictor variables and medium sample sizes.

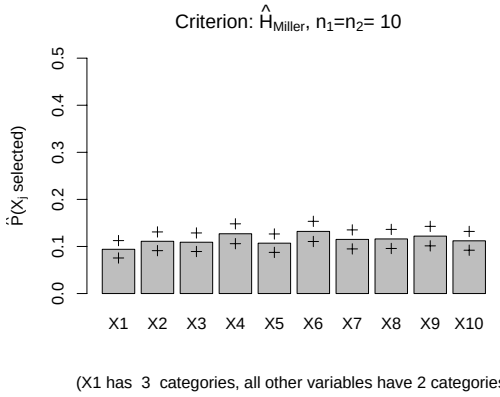

Figure 7: Estimated variable selection probabilities for the corrected estimator of the Shannon entropy $\widehat{H}_{\text{Miller}}$, for 3 vs. 2 categories in the predictor variables and small sample sizes.

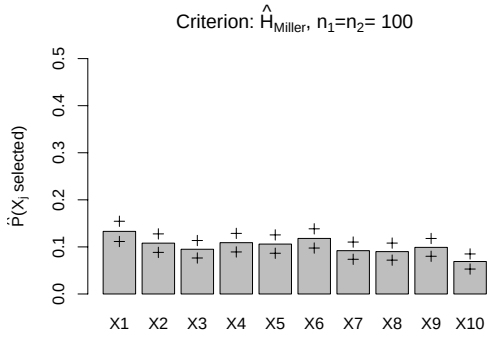

Figure 8: Estimated variable selection probabilities for the corrected estimator of the Shannon entropy $\widehat{H}_{\text{Miller}}$, for 5 vs. 2 categories in the predictor variables and medium sample sizes.

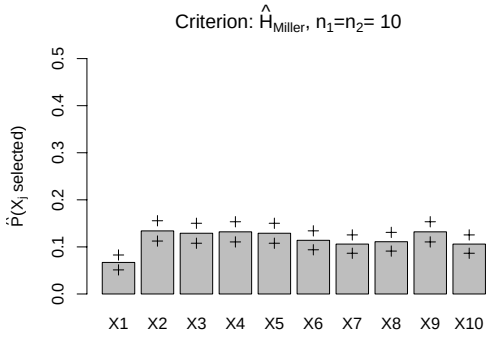

Figure 9: Estimated variable selection probabilities for the corrected estimator of the Shannon entropy $\widehat{H}_{\text{Miller}}$, for 5 vs. 2 categories in the predictor variables and small sample sizes.

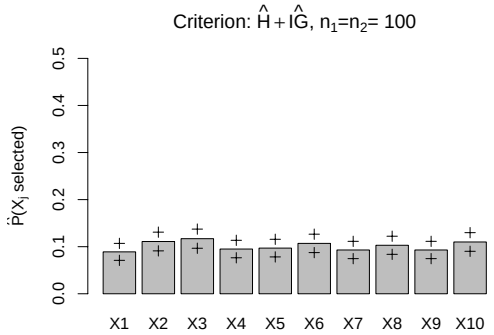

Figure 10: Estimated variable selection probabilities for the corrected estimator of the Shannon entropy $\widehat{H}+\widehat{IG}$, for 3 vs. 2 categories in the predictor variables and medium sample sizes.

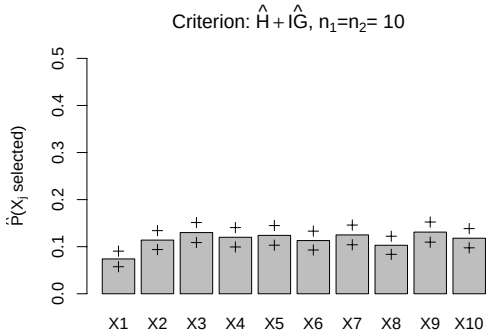

Figure 11: Estimated variable selection probabilities for the corrected estimator of the Shannon entropy $\widehat{H}+\widehat{IG}$, for 3 vs. 2 categories in the predictor variables and small sample sizes.

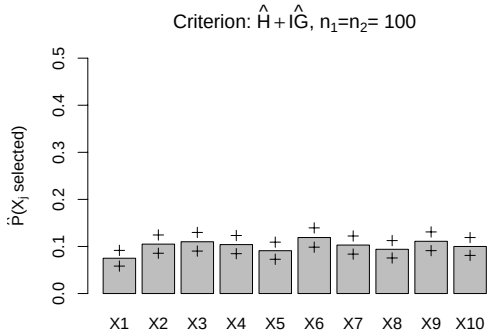

Figure 12: Estimated variable selection probabilities for the corrected estimator of the Shannon entropy $\widehat{H}+\widehat{IG}$, for 5 vs. 2 categories in the predictor variables and medium sample sizes.

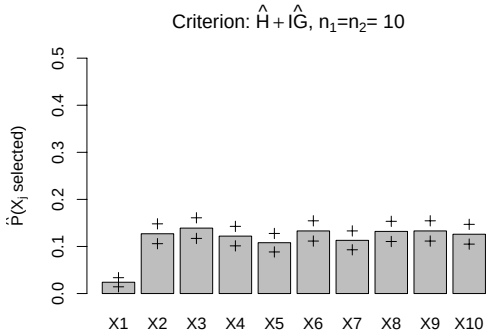

Figure 13: Estimated variable selection probabilities for the corrected estimator of the Shannon entropy $\widehat{H}+\widehat{IG}$, for 5 vs. 2 categories in the predictor variables and small sample sizes.

(2003) and Bernard (2005) on association measures in contingency tables based on the IDM. In this way, the p-value approach successfully applied in classical CART, e.g. based on the p-value of Fisher's exact test, could be extended towards imprecise probabilities. The posterior upper probability of H_0: "The response class is independent from the category of the predictor variable" could serve as the split selection criterion in classification trees based on imprecise probabilities.

Our perspective is to establish unbiased criteria for classification trees based on imprecise probabilities, as well as classical classification trees, in order to abolish the widespread use of biased criteria that overshadows the advantages of classification trees.

Acknowledgements

I would like to thank Thomas Augustin for his helpful comments.

References

Abellán, J. and S. Moral (2003). Maximum of entropy for credal sets. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 11*, 587–597.

Abellán, J. and S. Moral (2004). Upper entropy of credal sets. Applications to credal classification. *International Journal of Approximate Reasoning*.

Basharin, G. P. (1959). On a statistical estimate for the entropy of a sequence of independent random variables. *Theory of Probability and its Applications 4*, 333–336.

Bernard, J.-M. (2003). Analysis of local or asymmetric dependencies in contingency tables using the imprecise Dirichlet model. *Proceedings of the Third International Symposium on Imprecise Probabilities and their Applications*.

Bernard, J.-M. (2005). An introduction to the imprecise Dirichlet model for multinomial data. *International Journal of Approximate Reasoning, to appear*.

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

Chiang, I.-J. and J. Y. Hsu (2002). Fuzzy classification trees for data analysis. *Fuzzy Sets and Systems 130*, 87–99.

de Cooman, G. and M. Zaffalon (2004). Updating beliefs with incomplete observations. *Artificial Intelligence 159*, 75–125.

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.

Gail, M. H. and S. B. Green (1976). A generalization of the one-sided two-sample Kolmogorov-Smirnov statistic for evaluating diagnostc tests. *Biometrics 32*, 561–570.

Grassberger, P. (2003). Entropy estimates from insufficient sampling. *http://www.arxiv.org, physics/0307138*.

Halpern, A. L. (1999). Minimally selected p and other tests for a single abrupt changepoint in a binary sequence. *Biometrics 55*, 1044–1050.

Holste, D., I. Grosse, and H. Herzel (1998). Bayes' estimators of generalized entropies. *Journal of Physics A 31*, 2551–2566.

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

Klir, G. J. (1999). Uncertainty and information measures for imprecise probabilities: An overview. In *Proceedings of the First International Symposium on Imprecise Probabilities and their Applications*.

Klir, G. J. (2003). An update on generalize information theory. In *Proceedings of the Third International Symposium on Imprecise Probabilities and their Applications*, pp. 321–334.

Koziol, J. A. (1991). On maximally selected chi-square statistics. *Biometrics 4*, 1557–1561.

Loh, W. and Y. Shih (1997). Split selection methods for classification trees. *Statistica Sinica 7*, 815–840.

Martin, J. K. (1997). An exact probability metric for decision tree splitting and stopping. *Machine Learning 28*, 257–291.

Miller, G. (1955). Note on the bias of information estimates. In *Information Theory in Psychology*, pp. 95–100. Free Press: Glencoe, IL.

Miller, R. and D. Siegmund (1982). Maximally selected rank statistics. *Biometrics 38*, 1011–1016.

Pöschel, T., W. Eberling, C. Frömmel, and R. Ramirez (2003). Correction algorithm for finite sample statistics. *European Physical Journal E 12*, 531–541.

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

Schürmann, T. (2004). Bias analysis in entropy estimation. *Journal of Physics A 37*, 295–301.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal 27*, 379–423.

Shih, Y.-S. (2004). A note on split selection bias in classification trees. *Computational Statistics and Data Analysis 45*, 457–466.

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

Walley, P. (1996). Inferences from multinomial data: learning from a bag of marbles. *Journal of the Royal Statistical Society B 58*, 3–57.

Zaffalon, M. (2002). Exact credal treatment of missing data. *Journal of Statistical Planning and Inference 105*, 105–122.