Variable Selection in Classification Trees Based on Imprecise Probabilities

Carolin Strobl
LMU Munich, Germany
carolin.strobl@stat.uni-muenchen.de

Abstract

Classification trees are a popular statistical tool with multiple applications. Recent advancements of traditional classification trees, such as the approach of classification trees based on imprecise probabilities by Abellán and Moral (2004), effectively address their tendency to overfitting. However, another flaw inherent in traditional classification trees is not eliminated by the imprecise probability approach: Due to a systematic finite sample-bias in the estimator of the entropy criterion employed in variable selection, categorical predictor variables with low information content are preferred if they have a high number of categories. Mechanisms involved in variable selection in classification trees based on imprecise probabilities are outlined theoretically as well as by means of simulation studies. Corrected estimators are proposed, which prove to be capable of reducing estimation bias as a source of variable selection bias.

Keywords. Classification trees, credal classification, IDM, variable selection bias, 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 $X_1, \dots, X_p$. In comparison to other traditional 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 of the results. One major field of application of traditional classification trees and their advancements is the prediction of medical diagnoses from clinical and, most recently, genetical data (cp. e.g. Myles et al., 2004, for a review on applications in gene expression analysis).

A strong disadvantage of traditional classification trees, however, is their susceptibility to overfitting, which affects their robustness against outliers in the sample and necessitates terminal pruning. The extension of classification trees as credal classifiers based on imprecise probabilities by Abellán and Moral (2004) establishes a more sensitive means of classification, which is not as susceptible to overfitting and thus provides more reliable results.

Classification tree algorithms are specified by their split selection criterion, which controls variable selection, and the number of splits they produce in each node. In the approach of classification trees based on imprecise probabilities for categorical predictor variables by Abellán and Moral (2004), which is considered here, the number of nodes produced in each split is equal to the number of categories of the predictor variable chosen for the next split. Variable selection is conducted with respect to an upper entropy criterion.

Another serious problem in practical applications of classification trees is that split selection criteria can be biased in variable selection, preferring variables for features other than their information content (see Strobl, 2005, for a review on variable selection bias in traditional classification trees). We will show that variable selection bias does affect variable selection in the approach of Abellán and Moral (2004) if the predictor variables vary in their numbers of categories.

The source of this variable selection bias is the fact that the empirical Shannon entropy, a generalization of which is employed in the algorithm by Abellán and Moral (2004), is a negatively biased estimator of the true Shannon entropy. Even though the effect of this estimation bias on variable selection is moderated by the upper entropy-approach, it is still a relevant source of variable selection bias. Proposals for corrected entropy 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

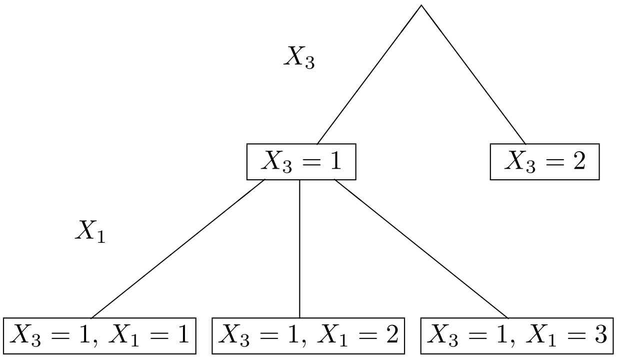

Figure 1: Example of a classification tree. Configurations of predictor values characterizing the observations in each node are displayed in boxes depicting the nodes.

probabilities.

The paper starts with an outline of the approach of classification trees based on imprecise probabilities in Section 2. Section 3 covers the problem of biased sample estimators of entropy measures in general and in application to classification trees based on imprecise probabilities, and introduces possible corrections, which are evaluated in a simulation study in Section 4. Section 5 gives a concluding summary of the results.

2 Classification trees based on imprecise probabilities

The rationale of classification trees based on imprecise probabilities for categorical predictor variables by Abellán and Moral (2004) is similar to the traditional classification tree approach of Quinlan (1993):

Starting with the set of all possible predictor variables the first splitting variable is selected such that it minimizes the value of a specified impurity criterion in the resulting nodes (the following considerations will focus on this part of the procedure). Once a predictor variable is selected for splitting as many nodes as categories of that predictor are produced. Each node is characterized by the configuration of predictor values that characterizes the observations in the node (cp. Figure 1). The splitting then proceeds in each node until the impurity reduction induced by splitting reaches a specified stopping criterion.

In an advancement of this traditional classification tree algorithm Abellán and Moral (2004) apply the Imprecise Dirichlet Model (IDM) (Walley, 1996) in the construction of the classification tree and in the credal classification of observations in the final nodes.

The split selection criteria and procedure of Abellán and Moral (2004) are introduced more formally in the following: At first, the split selection criteria are introduced for one arbitrary node in Section 2.1. Then the entire split selection procedure, starting from that node, is treated in Section 2.2.

2.1 Total impurity criteria

The predictor variable configuration that characterizes all observations in one node is denoted as σ (cp. again Figure 1: e.g. the lower left node is defined by the configuration $\sigma = (X_3 = 1, X_1 = 1)$).

Let Y be a categorical response variable with values $k = 1, 2, \ldots, |K|$ in a finite set K. The credal set $\mathscr{P}^\sigma$ is a convex set of probability distributions representing the available information on the unknown value of the response variable Y in the node defined by predictor variable configuration σ. The total impurity criterion $TU2(\mathscr{P}^\sigma)$ for the credal set $\mathscr{P}^\sigma$ consisting of probability distributions[1] p^σ on the set K

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

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

As an alternative the authors have previously suggested another total impurity criterion (which we will revisit later)

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

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

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

and $m_{\mathscr{P}^\sigma}$ is the Möbius inverse of the lower envelope $f_{\mathscr{P}^\sigma} = \inf_{p^\sigma \in \mathscr{P}^\sigma} p^\sigma(A)$

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

with $|A - B|$ denoting the cardinality of the set A excluding B. $IG(\mathscr{P}^\sigma)$ 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 $TU1(\mathscr{P}^\sigma)$ additively incorporates both uncertainty and non-specificity.

[1] To indicate classical precise probabilities they will be denoted as lower case $p(\cdot)$ throughout this paper.

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

The data are incorporated in estimating the value of $TU2(\mathscr{P}^\sigma)$ by means of applying the IDM locally within each node. For each node, 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 node:

$$[\underline{P}^\sigma_{\text{IDM}}(k), \overline{P}^\sigma_{\text{IDM}}(k)] = \left[\frac{n_k^\sigma}{N^\sigma + s}, \frac{n_k^\sigma + s}{N^\sigma + s}\right], \quad (3)$$

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 assigning 0 or s additional observations to class k. The credal set $\mathscr{P}^\sigma$ in $TU2(\mathscr{P}^\sigma)$ is then given by all probability distributions p^σ on the set K, for which $p^\sigma(k) \in [\underline{P}^\sigma_{\text{IDM}}(k), \overline{P}^\sigma_{\text{IDM}}(k)]$ for all k, as derived in Equation 3. The maximization in $TU2(\mathscr{P}^\sigma)$ is technically accomplished by means of the upper entropy algorithm introduced in Abellán and Moral (2003). The algorithm identifies the posteriori probability distribution on K with the upper entropy that is in accordance with the upper and lower probabilities for each class $k \in K$ derived from the IDM.

2.2 Split selection procedure

The complete process of variable selection in the classification tree algorithm of Abellán and Moral (2004) consists of the following successive tasks:

Let X_j be a categorical predictor variable with values $x_j = 1, 2, \ldots, |U_j|$ in a finite set U_j. Starting from the root node (or a subsequent node respectively), defined by predictor variable configuration σ, for each potential splitting variable X_j as many nodes as categories $x_j \in U_j$ are produced. Within each new node, defined by the previous configuration σ in combination with the value x_j of the potential splitting variable X_j by $\sigma \cup (X_j = x_j)$, the lower and upper probabilities $[\underline{P}^{\sigma\cup(X_j=x_j)}_{\text{IDM}}(k), \overline{P}^{\sigma\cup(X_j=x_j)}_{\text{IDM}}(k)]$ of each response class k are derived from the class counts $n_k^{\sigma\cup(X_j=x_j)}$ by means of the IDM. The interval width is determined by the number of observations per node $N^{\sigma\cup(X_j=x_j)}$ and the hyperparameter s of the IDM. The computation of the upper entropy criterion is then conducted in two steps:

1. From the credal set $\mathscr{P}^{\sigma\cup(X_j=x_j)}$ derived from the lower and upper probabilities $[\underline{P}^{\sigma\cup(X_j=x_j)}_{\text{IDM}}(k), \overline{P}^{\sigma\cup(X_j=x_j)}_{\text{IDM}}(k)]$ the posterior upper entropy distribution $p^{\sigma\cup(X_j=x_j)}_{\text{maxE}}$, i.e. the distribution closest to the uniform distribution over the response classes in the set K, is determined by the algorithm given in Abellán and Moral (2003).
2. The value of $TU2(\mathscr{P}^{\sigma\cup(X_j=x_j)})$ is then estimated[2] by applying the plug-in estimator of the Shannon entropy, indicated by $\widehat{H}(\cdot)$ (cp. Section 3), to the posterior upper entropy distribution.

$$TU2(\mathscr{P}^{\sigma\cup(X_j=x_j)}) = \widehat{H}\left(p^{\sigma\cup(X_j=x_j)}_{\text{maxE}}\right) = -\sum_{k=1}^{|K|} p^{\sigma\cup(X_j=x_j)}_{\text{maxE}}(k) \cdot ln\left[p^{\sigma\cup(X_j=x_j)}_{\text{maxE}}(k)\right] \quad (4)$$

The impurity reduction induced by splitting in variable X_j is measured by the weighted sum of total impurity measures over all new nodes

$$I(\sigma, X_j) = \sum_{x_j \in U_j} \frac{N^{\sigma\cup(X_j=x_j)}}{N^\sigma} TU2(\mathscr{P}^{\sigma\cup(X_j=x_j)}) \quad (5)$$

where $\frac{N^{\sigma\cup(X_j=x_j)}}{N^\sigma}$ is the relative frequency of observations assigned to each new node, with $\sum_{x_j \in U_j} N^{\sigma\cup(X_j=x_j)} = N^\sigma$. The variable X_j for which $I(\sigma, X_j)$ is minimal is selected for the next split.

2.3 Characteristics of the total impurity criterion TU2

The variable selection performance of a split selection criterion can be evaluated by means of simulation studies. In order to illustrate the variable selection characteristics of the total impurity criterion $TU2(\mathscr{P}^\sigma)$ the following standard simulation study design was chosen here:

Several predictor variables are generated such that they only differ in one feature, which is expected to affect variable selection. 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 for equally informative predictor variables if no selection bias occurs[3].

[2] In Section 3 the distinction between theoretical and empirical quantities will be emphasized, and the sample estimator will then be denoted by $\widehat{TU2}(\mathscr{P}^{\sigma\cup(X_j=x_j)})$.

[3] In this simulation design 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. In a tree building algorithm one variable has to be randomly chosen for splitting in this case.

X_1	X_2	Y
1	1 2	$Bin(0.5 + \text{relevance})$
2	3 4	$Bin(0.5 - \text{relevance})$

Table 1: Study design of simulation study on characteristics of the total impurity criterion TU2: For fixed predictor values the response is sampled from a Binomial distribution with sample size $\frac{n}{2}$ and different class probabilities.

The results displayed below are from a simulation study run with 1000 simulations and sample size $n = 120$. Two equally informative predictor variables were created, one of which had 2 and the other 4 equally frequent categories. The value of the hyperparameter s of the IDM was set equal to 1. The sampling distribution for the response variable was varied to manipulate the relevance of the predictor variables. As displayed in Table 1 the sampling distribution of the response variable differed in the categories of the predictor variables depending on the relevance parameter. All simulation studies were conducted with the software package R (Version 2.0.0).

Figures 2 through 4 depict the results of the simulation study as barplots with the bar height indicating the estimated selection probabilities for the two equally informative predictor variables and the crosses marking ± 2 empirical standard errors of the point estimates.

The results of the simulation studies show that two characteristics of the total impurity criterion $TU2(\mathscr{P}^{\sigma\cup(X_j=x_j)})$ have an impact when the categorical predictor variables competing for variable selection vary in their number of categories, and thus in the number of observations within each new node: When deriving the upper entropy distribution $p_{\text{maxE}}^{\sigma\cup(X_j=x_j)}$ (in step 1 of the computation of the upper entropy criterion outlined in Section 2.2) a smaller number of observations per node results in a wider interval of lower and upper probabilities $[\underline{P}_{\text{IDM}}^{\sigma\cup(X_j=x_j)}(k), \overline{P}_{\text{IDM}}^{\sigma\cup(X_j=x_j)}(k)]$. From a wider interval a more uninformative upper entropy distribution $p_{\text{maxE}}^{\sigma\cup(X_j=x_j)}$ can be derived. Thus, the total impurity criterion $TU2(\mathscr{P}^{\sigma} \cup (X_j = x_j))$ increases when the number of observations in the new node decreases, and variables with more distinct categories are penalized. This mechanism of variable selection bias is most prominent in highly informative variables, because their true information content differs strongly from the much less informative distribution $p_{\text{maxE}}^{\sigma\cup(X_j=x_j)}$, that is obtained from the wide intervals.

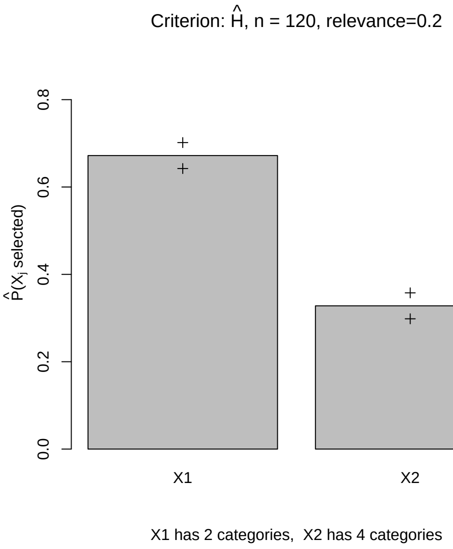

Figure 2: Estimated variable selection probabilities for the upper entropy-total impurity criterion TU2. Both predictors are informative with medium relevance, they only vary in their number of categories.

Figure 2 illustrates this mechanism for two equally informative predictor variables, showing that on average the predictor variable X_1 with 2 categories is preferred over X_2 with 4 categories.

However, when the relevance of the predictor variables decreases as in Figure 3 we see that the mechanism explained above is superposed by another, yet unaccounted, mechanism that affects variable selection in less relevant predictor variables. For uninformative predictor variables[4] this second mechanism is most prominent as shown in Figure 4. The mechanism obvious in Figures 3 and 4 induces a preference of the predictor variable X_2 with 4 categories over X_1 with 2 categories. We will show that the underlying mechanism is a bias in the estimation procedure of the total impurity criterion from the posterior upper entropy distribution (in step 2 of the computation of the upper entropy criterion outlined in Section 2.2). The statistical background of this estimation bias, as well a correction approach, is given in the next section.

The two mechanisms illustrated here counteract in their effect on variable selection: The tradeoff between the upper entropy-approach on one hand and estimation bias on the other hand depends on the data

[4] The paradigm employed in Figure 4 is also the standard paradigm used for the evaluation of variable selection bias (e.g. Kim and Loh, 2001; Strobl, 2005, for a summary). The underlying rationale is that uninformative predictor variables should be selected with equal (random choice) probability if no variable selection bias occurs.

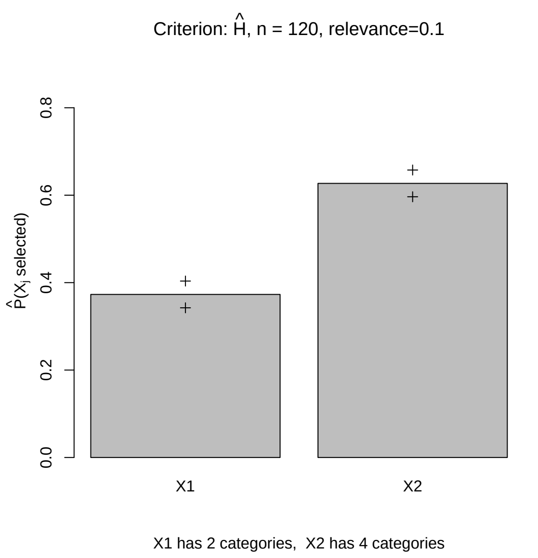

Figure 3: Estimated variable selection probabilities for the upper entropy-total impurity criterion TU2. Both predictors are informative with low relevance, they only vary in their number of categories.

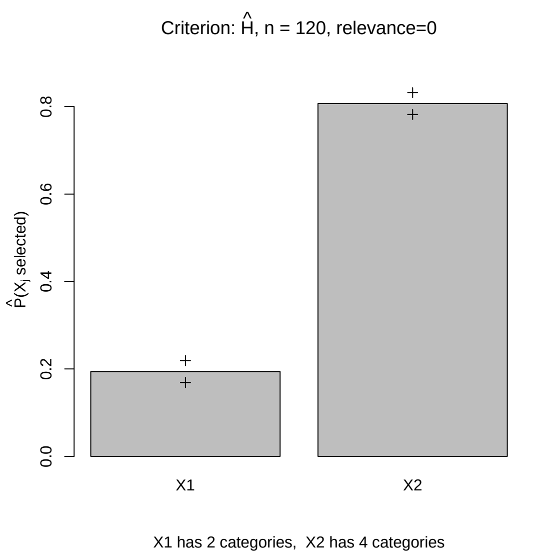

Figure 4: Estimated variable selection probabilities for the upper entropy-total impurity criterion TU2. Both predictors are uninformative, they only vary in their number of categories.

situation. In an extreme case, however, the effect of estimation bias can induce a preference of a less informative variable over a more informative variable in variable selection - merely due to different numbers of categories. Thus, the mechanism of estimation bias is elaborated in the following section.

3 Empirical entropy measures in split selection

As implied above the biased estimation of the splitting criterion can be identified as one source of variable selection bias in classification trees. In order to address this problem, we review the necessary statistical background on the estimation of entropy measures in a theoretical section and then apply the results to classification trees based on imprecise probabilities.

Since the central issue in this section is the distinction between theoretical quantities and their sample estimators, we will follow a common statistical notation, where estimators of theoretical quantities are indicated by adding a $\widehat{}$ to the symbol. E.g. estimators of classical probabilities are denoted as $\widehat{p}(\cdot)$, while the true probabilities are denoted as $p(\cdot)$.

3.1 Estimation bias for empirical entropy measures

The theoretical Shannon entropy

$$H(p) = -\sum_{k=1}^{|K|} p\,(k)\; ln[p\,(k)]$$

is a function of the true response class probabilities p(k). In order to estimate the Shannon entropy from empirical data the popular estimator $\widehat{H}$ is a plug-in estimator retaining the original function but replacing the true class probabilities by the observed relative class frequencies, i.e. by the maximum-likelihood estimators of the true class probabilities

$$\widehat{H}(\widehat{p}) = -\sum_{k=1}^{|K|} \widehat{p}\,(k)\; ln[\widehat{p}\,(k)].$$

However, this widely used estimator is biased for finite sample sizes, because with a decreasing number of observations the standard error of the estimators $\widehat{p}(k)$ increases, producing posterior class distributions misleadingly implying a higher information content.

Based on a statistical evaluation of the bias, possible correction strategies are derived in the following: From Jensen's inequality, $f\,(E_p(\widehat{p})) \geq E_p\,(f(\widehat{p}))$ for any concave function f, it is obvious that the unbiasedness of the maximum-likelihood estimators $\widehat{p}(k)$ is not necessarily transferred to the plug-in estimator $\widehat{H}$, which may be negatively biased. The extent of the bias can be evaluated from the expected value of the plug-in estimator $\widehat{H}$ for the true Shannon entropy H independently derived by Miller (1955) and Basharin (1959)

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

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. also Schürmann, 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

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

Due to the omission of the terms of order $\frac{1}{N^2}$ this correction provides a decent approximation of the true entropy value only for sufficiently large sample sizes, while for $N \to \infty$ the correction is negligible.

3.2 Relevance of estimation bias for classification trees based on imprecise probabilities

As described in the beginning of Section 2.3 small sample sizes result in wider intervals of lower and upper probabilities $[\underline{P}_{\text{IDM}}^{\sigma\cup(X_j=x_j)}(k), \overline{P}_{\text{IDM}}^{\sigma\cup(X_j=x_j)}(k)]$ in each new node, from which more uninformative posterior upper entropy distributions can be derived.

However, another general effect of small sample sizes is that small changes in the data result in high changes of relative class frequencies computed from the data. This limited sample-effect also affects the intervals of lower and upper probabilities for the response classes in the approach of classification trees based on imprecise probabilities. The interval-bounds in Equation 3 can be naively considered as artificial relative class frequencies, where imprecision is incorporated by means of the s yet unobserved observations the class of which is not yet determined. The hyperparameter s is often set to a value of the magnitude 1 or 2. Thus, the artificial relative frequencies derived from the IDM suffer from the same weakness as classical relative frequencies, namely that for small sample sizes small changes in the data produce crucial changes in the relative frequencies, misleadingly implying class distributions with a higher information content. The estimation bias for empirical entropy measures outlined in the previous section therefore applies to the estimation of the total impurity criterion $\widehat{TU2}(\mathscr{P}^{\sigma\cup(X_j=x_j)})$ from the data.

When a predictor variable is highly informative, the effect of the estimation bias is compensated by the upper entropy-approach. However, for less or uninformative predictor variables the effect of estimation bias influences variable selection in favor of variables with more categories: For less informative or uninformative variables, where the posterior upper entropy distribution is a uniform distribution over the set of response classes K, the negative estimation bias occurring in each node is carried forward to the estimated criterion value $\widehat{I}(\sigma, X_j)$ (cp. Equation 5), on which the final decision in the variable selection procedure is based.

For an uninformative predictor variable, with the true class distribution $p^* := p_{\text{maxE}}^{\sigma\cup(X_j=x_j)} = U(1, |K|)$ discretely uniform on support $[1, |K|]$, the true entropy value $H^* := \sum_{k=1}^{|K|} p^*(k) ln[p^*(k)]$ is maximal and equal in each node. The approximated expected value of $\widehat{I}(\sigma, X_j)$ is then

$$\begin{aligned} & E_{p^*}\left(\widehat{I}(\sigma, X_j)\right) = \\ & \approx \sum_{x_j \in U_j} \frac{N^{\sigma\cup(X_j=x_j)}}{N^{\sigma}} \left\{ H^* - \frac{|K|-1}{2\left(N^{\sigma\cup(X_j=x_j)} + s\right)} \right\} \\ & \approx H^* \; - \; |U_j| \cdot \frac{|K|-1}{2\, N^{\sigma}} \end{aligned}$$

where the number of response categories $|K|$ is fixed, while the number of categories $|U_j|$ differs between the predictor variables X_j. Thus, the number of categories of the predictor variable X_j crucially affects its selection chance.

3.3 Suggested corrections based on the IDM

With $\widehat{H}\left(p_{\text{maxE}}^{\sigma\cup(X_j=x_j)}\right)$ denoting the plug-in estimator of the Shannon entropy applied to the posterior upper entropy distribution (cp. Equation 4) we suggest

$$\begin{aligned} & \widehat{H}_{\text{Miller}}\left(p_{\text{maxE}}^{\sigma\cup(X_j=x_j)}\right) = \\ & \widehat{H}\left(p_{\text{maxE}}^{\sigma\cup(X_j=x_j)}\right) + \frac{|K|-1}{2(N^{\sigma\cup(X_j=x_j)} + s)} \qquad (6) \end{aligned}$$

as the empirical entropy estimator in every new node of a classification tree based on imprecise probabilities. This correction accounts for the derivation of the posterior upper entropy distribution, to which the entropy estimator is applied, from the posterior lower and upper probabilities computed with respect to the IDM with hyperparameter s and sample size $N^{\sigma\cup(X_j=x_j)}$. This correction is again appropriate for medium $N^{\sigma\cup(X_j=x_j)}$, while it over-penalizes for small $N^{\sigma\cup(X_j=x_j)}$ with respect to the number of categoies $|K|$, which is supported by the numerical results in Section 4.

In another correction approach we are revisiting the empirical measure $\widehat{IG}(\mathscr{P}^{\sigma\cup(\mathscr{X}=)})$, the theoretical analogy of which was employed by Abellán and Moral (2004) as a measure of non-specificity in the total impurity criterion $TU1(\mathscr{P}^{\sigma\cup(X_j=x_j)})$ (cp. Equation 2). Like the correction term in the above approach $\widehat{IG}(\mathscr{P}^{\sigma\cup(X_j=x_j)})$ is a function of the sample size $N^{\sigma\cup(X_j=x_j)}$ and the number of categories $|K|$. In the special case where the lower probabilities used in the computation of the Möbius inverses in $\widehat{IG}(\mathscr{P}^{\sigma\cup(X_j=x_j)})$ 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. Because the logarithm of the cardinality of the singletons is zero, the Möbius inverse for the set K collapses to the width $\frac{s}{N^{\sigma\cup(X_j=x_j)}+s}$ of the intervals of lower and upper probabilities on K computed from the IDM with hyperparameter s, and the empirical non-specificity measure $\widehat{IG}(\mathscr{P}^{\sigma\cup(X_j=x_j)})$ depends only on the sample size $N^{\sigma\cup(X_j=x_j)}$ through the interval width, and on the number of categories $|K|$ through the factor $ln(|K|)$. We thus suggest

$$\widehat{H}\left(p_{\text{maxE}}^{\sigma\cup(X_j=x_j)}\right) + \widehat{IG}\left(\mathscr{P}^{\sigma\cup(X_j=x_j)}\right) = \widehat{H}\left(p_{\text{maxE}}^{\sigma\cup(X_j=x_j)}\right) + \widehat{m}_{\mathscr{P}^{\sigma\cup(X_j=x_j)}}(K)\; ln(|K|) \quad (7)$$

i.e. $\widehat{TU1}(\mathscr{P}^{\sigma\cup(X_j=x_j)})$, as another corrected estimator, where $\widehat{m}_{\mathscr{P}^{\sigma\cup(X_j=x_j)}}(K)$ is the Möbius inverse computed from the posterior lower class probabilities derived from the IDM. We will again see in Section 4 that this correction is only reliable for sufficiently large $N^{\sigma\cup(X_j=x_j)}$ and small $|K|$, while otherwise it is overcautious.

4 Simulation study: performance of entropy estimators in split selection

Again the variable selection performance of each split selection criterion can be evaluated by means of the following simulation study design: Several uninformative predictor variables are generated such that they only differ in the number of categories. 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 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 again set equal to 1. As displayed in Table 2 the response values in the simulation were fixed, while the uninformative predictors were sampled from discrete uniform distributions[5] on support [1,3] (respectively [1,5]) and [1,2]. The frequencies of the two response classes were set equal at $n_1 = n_2 = 100$ for medium sample size and $n_1 = n_2 = 10$ for small sample size.

Y	X_1	$X_2 \ldots X_{10}$
1	$U(1,3)$ or $U(1,5)$	$U(1,2)$
2		

Table 2: Study design of simulation study on entropy estimators: For fixed response values (n_1 class 1 observations and n_2 class 2 observations, set equal) the uninformative predictors were sampled from discrete uniform distributions with sample sizes $n = n_1 + n_2$ and different ranges.

In this study, the behavior of the plug-in estimator $\widehat{H}$ for the Shannon entropy (cp. Equation 4) is compared to the behavior of the corrected estimators $\widehat{H}_{\text{Miller}}$ (Equation 6) and $\widehat{H} + \widehat{IG}$ (Equation 7). Figures 5 through 8 display that, with the plug-in estimator $\widehat{H}$ 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 estimator $\widehat{H}_{\text{Miller}}$, Figures 9 through 12 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 estimator $\widehat{H} + \widehat{IG}$, Figures 13 through 16 show that the reverse bias for small sample sizes and large difference in the number of categories is even stronger than for $\widehat{H}_{\text{Miller}}$.

[5] This simulation design is equivalent to the standard paradigm displayed in Figure 4 in Section 2.3.

5 Discussion and perspective

The split selection criterion TU2 introduced for classification trees based on imprecise probabilities for categorical predictor variables by Abellán and Moral (2004) is affected by two mechanisms relevant in variable selection when predictors differ in their number of categories:

The first mechanism, relying on the selection of the posterior upper entropy distribution, penalizes highly informative predictor variables with many categories. The second counteracting mechanism, relying on the biased estimation of the total impurity criterion, favors less or uninformative predictor variables with many categories. In a tradeoff the combination of both mechanisms can lead to unwanted variable selection bias depending on the data situation.

In a first approach employing corrected estimators of the total impurity criterion in variable selection our results imply that the corrections accomplish to eliminate part of the variable selection bias induced by estimation bias. Both corrected estimators perform better than the TU2 criterion in the standard paradigm with uninformative predictor variables. The corrected estimator $\widehat{H}_{\text{Miller}}$ in Equation 6 shows even better variable selection performance than the corrected estimator $\widehat{H+IG}$ in Equation 7. The corrected estimators are less reliable for small sample sizes and large numbers of categories of the predictor variables, where they react overcautious. However, for application in a classification tree this effect can be accounted for by incorporating the tolerable minimum number of observations per node in the stopping criterion. The corrected estimators can be easily applied to the posterior upper entropy distribution derived from the lower and upper probabilities computed with the IDM as suggested by (Abellán and Moral, 2004). The correction so far incorporates only the deviation of the expected value of the estimator of the Shannon entropy. Another relevant factor, which could be integrated in further corrections, is the variance of the estimator derived e.g. in Roulston (1999). More elaborate entropy estimators will be considered for split selection in future research.

Acknowledgements

I would like to thank Thomas Augustin and three anonymous referees for their helpful comments that have essentially added to the readability of the paper.

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 39*, 235–255.

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.

Kim, H. and W. 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 M. W. de Cooman, Cozman (Ed.), *Proceedings of the First International Symposium on Imprecise Probabilities and their Applications.*

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

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

Myles, A., R. Feudale, Y. Liu, N. Woody, and S. Brown (2004). An introduction to decision tree modelling. *Journal of Chemometrics 18*, 275–285.

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

Roulston, M. (1999). Estimating the errors on measured entropy and mutual information. *Physica D 125*, 285–294.

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.

Strobl, C. (2005). Statistical sources of variable selection bias in classification tree algorithms based on the gini index. *http://www.stat.uni-muenchen.de/sfb386/papers/dsp/paper420.ps*, (SFB–Discussion Paper).

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

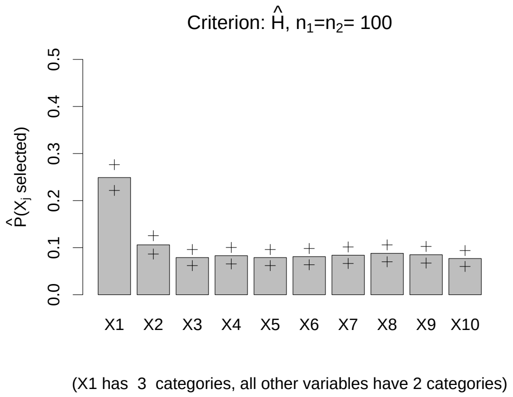

Figure 5: Estimated variable selection probabilities for the plug-in estimator of the Shannon entropy for 3 vs. 2 categories in the predictor variables and medium sample sizes.

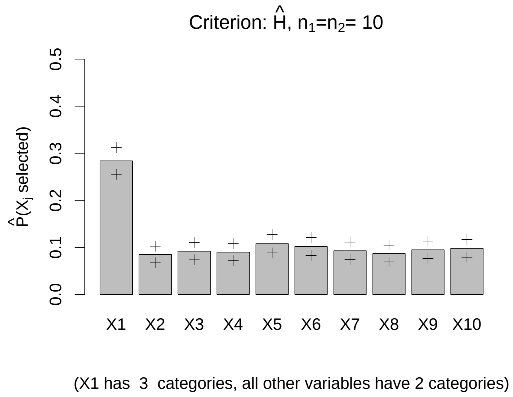

Figure 6: Estimated variable selection probabilities for the plug-in estimator of the Shannon entropy for 3 vs. 2 categories in the predictor variables and small sample sizes.

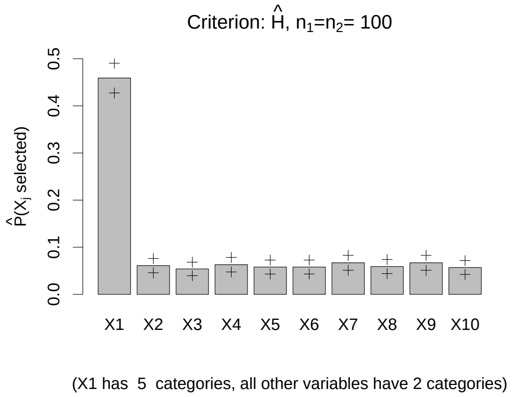

Figure 7: Estimated variable selection probabilities for the plug-in estimator of the Shannon entropy for 5 vs. 2 categories in the predictor variables and medium sample sizes.

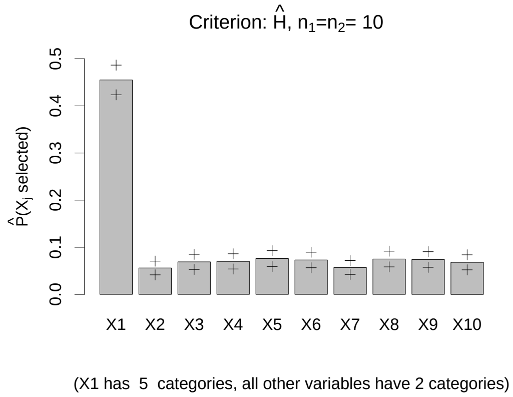

Figure 8: Estimated variable selection probabilities for the plug-in estimator of the Shannon entropy for 5 vs. 2 categories in the predictor variables and small sample sizes.

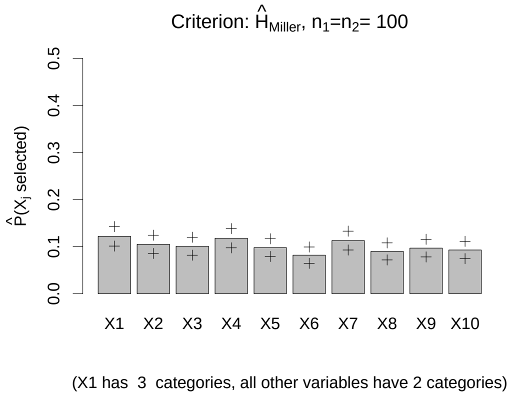

Figure 9: Estimated variable selection probabilities for the corrected estimator of the Shannon entropy $\widehat{H}_{\text{Miller}}$, for 3 vs. 2 categories in the predictor variables and medium sample sizes.

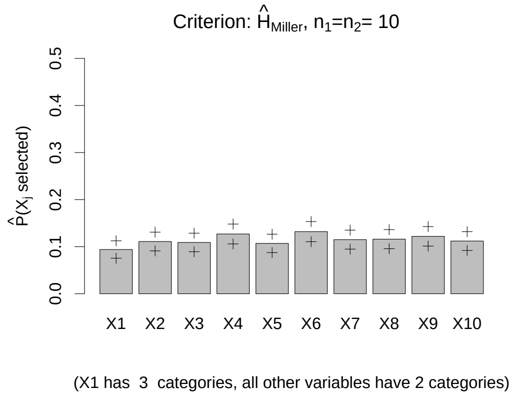

Figure 10: Estimated variable selection probabilities for the corrected estimator of the Shannon entropy $\widehat{H}_{\text{Miller}}$, for 3 vs. 2 categories in the predictor variables and small sample sizes.

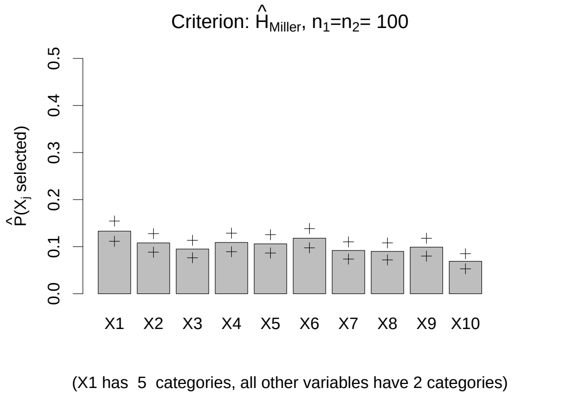

Figure 11: Estimated variable selection probabilities for the corrected estimator of the Shannon entropy $\widehat{H}_{\text{Miller}}$, for 5 vs. 2 categories in the predictor variables and medium sample sizes.

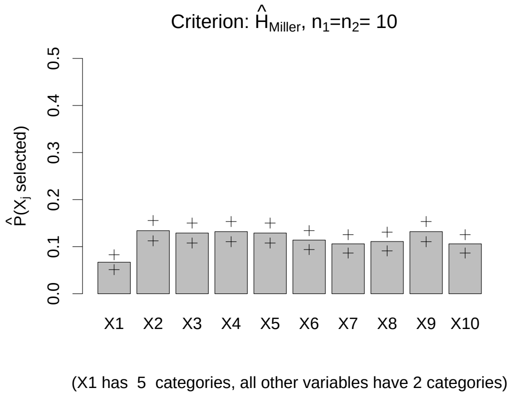

Figure 12: Estimated variable selection probabilities for the corrected estimator of the Shannon entropy $\widehat{H}_{\text{Miller}}$, for 5 vs. 2 categories in the predictor variables and small sample sizes.

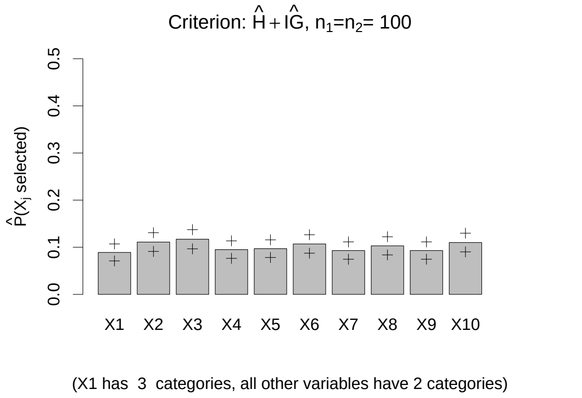

Figure 13: Estimated variable selection probabilities for the corrected estimator of the Shannon entropy $\widehat{H}+\widehat{IG}$, for 3 vs. 2 categories in the predictor variables and medium sample sizes.

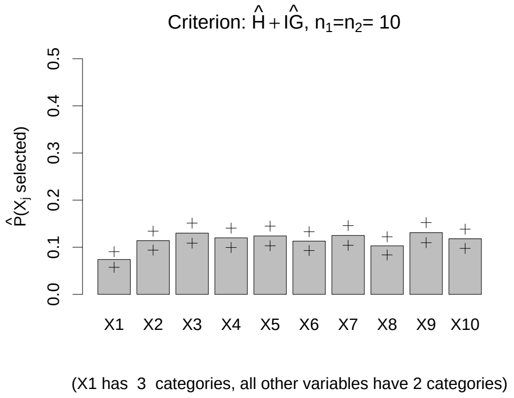

Figure 14: Estimated variable selection probabilities for the corrected estimator of the Shannon entropy $\widehat{H}+\widehat{IG}$, for 3 vs. 2 categories in the predictor variables and small sample sizes.

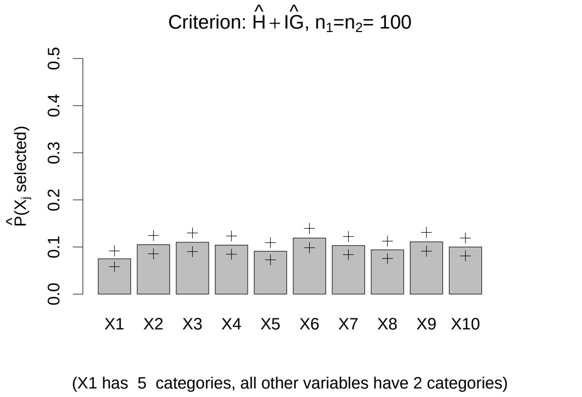

Figure 15: Estimated variable selection probabilities for the corrected estimator of the Shannon entropy $\widehat{H}+\widehat{IG}$, for 5 vs. 2 categories in the predictor variables and medium sample sizes.

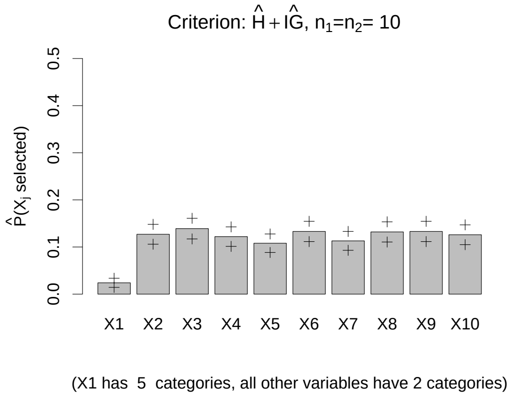

Figure 16: Estimated variable selection probabilities for the corrected estimator of the Shannon entropy $\widehat{H}+\widehat{IG}$, for 5 vs. 2 categories in the predictor variables and small sample sizes.