Unbiased split selection for classification trees based on the Gini Index

Carolin Strobl[a], Anne-Laure Boulesteix[b], Thomas Augustin[a]

[a]*Department of Statistics, University of Munich*
Ludwigstr. 33, 80539 Munich, Germany

[b]*Department of Medical Statistics and Epidemiology, Technical University of Munich, Ismaningerstr. 22, 81675 Munich, Germany*

Abstract

Classification trees are a popular tool in applied statistics because their heuristic search approach based on impurity reduction is easy to understand and the interpretation of the output is straightforward. However, all standard algorithms suffer from a major problem: variable selection based on standard impurity measures as the Gini Index is biased. The bias is such that, e.g., splitting variables with a high amount of missing values – even if missing completely at random – are artificially preferred. A new split selection criterion that avoids variable selection bias is introduced. The exact distribution of the maximally selected Gini gain is derived by means of a combinatorial approach and the resulting p-value is suggested as an unbiased split selection criterion in recursive partitioning algorithms. The efficiency of the method is demonstrated in simulation studies and a real data study from veterinary gynecology in the context of binary classification and continuous predictor variables with different numbers of missing values. The proposed method is extendible to categorical and ordinal predictor variables and to other split selection criteria such as the cross-entropy.

Key words: Classification trees, variable selection bias, Gini gain, missing values

1 Introduction

In many scientific fields recursive partitioning approaches like classification and regression trees (Breiman, Friedman, Olshen, and Stone, 1984; Bittencourt and Clarke, 2004, for an application in agricultural image recognition) and tree-based methods like emerging patterns (Dong and Li, 1999; Boulesteix and Tutz, 2006, for an application in microarray analysis) and random forests (Breiman, 2001; Jong et al., 2005, for an application in bioengineering) have become a popular tool for selecting relevant predictor variables even from large sets of candidates. In such applications, when trees and tree-based methods are used not only for prediction but also to (pre-)select relevant variables and thereby reduce the dimensionality of the sample space, it is particularly important that the variable selection is reliable and unbiased (cf. Strobl, Boulesteix, Zeileis, and Hothorn, 2007).

The traditional recursive partitioning approaches CART by Breiman et al. (1984) and C4.5 by Quinlan (1993) use empirical impurity reduction measures, such as the Gini gain derived from the Gini Index or the Information gain, as split selection criteria: the cutpoint and splitting variable that produce the highest impurity reduction are chosen for the next split. The intuitive approach of impurity reduction added to the popularity of recursive partitioning algorithms, and entropy based measures are still the default splitting criteria in most implementations of classification trees.

However, Breiman et al. (1984) already note that "variable selection is biased in favor of those variables having more values and thus offering more splits" (p.42) when the Gini gain is used as splitting criterion. For example, if the predictor variables are categorical variables of ordinal or nominal scale, variable selection is biased in favor of variables with a higher number of categories, which is a general problem not limited to the Gini gain (cf. Strobl, 2005). In addition, variable selection bias can also occur if the splitting variables vary in their number of missing values, even if the values are missing completely at random.

This is particularly remarkable since, in general, values missing completely at random (MCAR) can be discarded without producing a systematic bias in sample estimates (Little and Rubin, 1986, 2002). However, in the approach of classification trees even values missing completely at random can strongly affect the outcome and the evaluation of the variable importance. Again, this problem is not limited to the Gini gain criterion and affects both binary and multiway splitting recursive partitioning.

Possible strategies to deal with values missing completely at random (MCAR) include: (i) "Listwise" or "casewise deletion", where all observations or cases with the value of at least one variable missing are deleted. This strategy can result in a severe reduction of the sample size, if the missing values are spread over many observations and variables. (ii) "Pairwise deletion" or "available case" strategy, where only for the variables considered at each step of the analysis, e.g. for the two variables currently involved in a correlation, the observations with missing values in these variables are deleted for the current analysis, but are reconsidered in later analysis of different non-missing variables. With this strategy different sets of observations might be involved in different parts of the analysis or model building process. (iii) Various imputation methods, like, e.g., the simple "mean imputation" where the mean value in each variable is substituted to replace missing values. The naive "mean imputation" approach artificially reduces the variation of values of a variable, with the extent of the decrease depending on the number of missing values in each variable, and thus may change the strength of correlations, while more elaborate "multiple imputation" strategies overcome this problem.

The focus of this paper is to study from a theoretical point of view the variable selection bias occurring with the widely used Gini gain when missing values are treated in an available case strategy, as considered in the standard literature on valiable selection bias (e.g., Kim and Loh, 2001), and to propose an unbiased alternative splitting criterion based on the Gini gain for the case of continuous predictors and a binary response. In section 2, we identify and examine three components of variable selection bias, which are (i) estimation bias of the empirical Gini Index, (ii) variance of the empirical Gini Index and (iii) multiple comparison effects in cutpoint selection.

Section 3 presents our selection criterion that is based on the Gini gain and inspired by the theory of maximally selected statistics. Our criterion can be seen as the p-value computed from the distribution of the maximally selected Gini gain under the null hypothesis of no association between the response and the considered predictor variable. Our combinatorial method to derive the exact distribution of the maximally selected Gini gain under the null hypothesis is described in detail in section 3. The presentation is limited to the case of a binary response variable, which is the most common case in many applications such as medical studies, and to continuous predictor variables with different numbers of missing values. However, using the concepts of Boulesteix (2006a) and Boulesteix (2006b), our approach could be generalized to unbiased split selection from categorical and ordinal predictor variables with different numbers of categories, and to other entropy based measures.

Results from simulation studies documenting the performance of our split selection criterion are displayed in section 4. The relevance of our approach is illustrated by an application to veterinary data in section 5.

The rest of this section introduces the notation: Y denotes the binary response variable which takes the values $Y = 1$ and $Y = 2$, and $\mathbf{X}^T = (X_1, \ldots, X_q)$ denotes the random vector of continuous predictors. We consider a sample $(y_i, \mathbf{x}_i)_{i=1,\ldots,N}$ of N independent identically distributed observations of Y and $\mathbf{X}$. The variables $X_1, \ldots, X_q$ may have different numbers of missing values in the sample $(y_i, \mathbf{x}_i)_{i=1,\ldots,N}$. For $j = 1, \ldots, q$, let N_j denote the sample size obtained if observations with a missing value in variable X_j are eliminated in an available case or pairwise deletion strategy, where in each step of the recursive partitioning algorithm only the current splitting variable X_j containing missing values and the completely observed response variable are considered.

The following computations are implicitly conditional on these N_j available observations, of which there are N_{1j} observations with $Y = 1$ and N_{2j} with $Y = 2$.

Using machine learning terminology, $\mathbf{S}_j$, $j = 1, \dots, q$, denotes the starting set for variable X_j: $\mathbf{S}_j$ holds the N_j observations for which the predictor variable X_j is not missing. $(y_{(i)j}, x_{(i)j})_{i=1,\dots,N_j}$ denote the observed values of Y and X_j, where the sample is ordered with respect to the values of X_j ($x_{(1)j} \leq \dots \leq x_{(N_j)j}$). The subsets $\mathbf{S}_{Lj}(i)$ and $\mathbf{S}_{Rj}(i)$ are produced by splitting $\mathbf{S}_j$ at a cutpoint between $x_{(i)j}$ and $x_{(i+1)j}$, such that all observations with a value of $X_j \leq x_{(i)j}$ are assigned to $\mathbf{S}_{Lj}(i)$ and the remaining observations to $\mathbf{S}_{Rj}(i)$. These notations as well as the corresponding subset sizes are summarized in Table 1, where e.g. $n_{2j}(i)$ denotes the number of observations with $Y = 2$ in the subset defined by $X_j \leq x_{(i)j}$, i.e. by splitting after the i-th observation in the ordered sample. The function $n_{2j}(i)$ is thus defined as the number of observations with $Y = 2$ among the first i observations of variable X_j,

$$n_{2j}(i) = \sum_{k=1}^{i} I_{\{2\}}(y_{(k)j}), \quad \forall i = 1, \dots, N_j. \tag{1}$$

where $I_{\{2\}}(\cdot)$ is the indicator function; $n_{1j}(i)$ is defined in an analogous way.

Table 1
Contingency table obtained by splitting the predictor variable X_j at $x_{(i)j}$.

	$\mathbf{S}_{Lj}(i)$	$\mathbf{S}_{Rj}(i)$	
	$X_j \leq x_{j(i)}$	$X_j > x_{j(i)}$	Σ
$Y = 1$	$n_{1j}(i)$	$N_{1j} - n_{1j}(i)$	N_{1j}
$Y = 2$	$n_{2j}(i)$	$N_{2j} - n_{2j}(i)$	N_{2j}
Σ	$N_{Lj} = i$	$N_{Rj} = N_j - i$	N_j

For any subsequent split, the new node can be considered as the starting node. Thus, we are able to restrict the argumentation to the first root node for the sake of simplicity.

The empirical Gini Index $\widehat{G}_j$ of $\mathbf{S}_j$, defined by (Breiman et al., 1984) for the multi-class case, is a widely used impurity measure. For the considered variable X_j and in the special case of a binary response Y reduces to

$$\widehat{G}_j = 2\frac{N_{2j}}{N_j}\left(1 - \frac{N_{2j}}{N_j}\right). \tag{2}$$

The corresponding empirical Gini Indices in the nodes produced by splitting at the i-th cutpoint, $\widehat{G}_{Lj}(i)$ and $\widehat{G}_{Rj}(i)$, are defined analogously. The empirical

Gini gain, i.e. the impurity reduction produced by splitting at the i-th cutpoint of variable X_j, is based on the difference in impurity before and after splitting

$$\begin{aligned}\widehat{\Delta G}_j(i) &= \widehat{G}_j - \left(\frac{N_{Lj}}{N_j}\widehat{G}_{Lj}(i) + \frac{N_{Rj}}{N_j}\widehat{G}_{Rj}(i)\right) \\ &= \widehat{G}_j - \left(\frac{i}{N_j}\widehat{G}_{Lj}(i) + \frac{N_j - i}{N_j}\widehat{G}_{Rj}(i)\right).\end{aligned} \tag{3}$$

Obviously, the 'best' split according to the Gini gain criterion is the split with the largest Gini gain, i.e. with the largest impurity reduction, and so the most common approach for binary split and variable selection in classification trees consists of the following successive steps:

(1) for each variable X_j determine the maximal Gini gain $\widehat{\Delta G}_j^{max}$ over all possible cutpoints, which is defined as

$$\widehat{\Delta G}_j^{max} = \max_{i=1,\dots,N_j-1} \widehat{\Delta G}_j(i),$$

(2) select the variable X_{j^*} with the largest maximal Gini gain:

$$j^* = \arg \max_{j=1,\dots,q} \widehat{\Delta G}_j^{max}.$$

Variable selection bias occurring when the Gini Index is used as a selection criterion in this so called “greedy search” approach is studied in the next section.

2 Variable selection bias

In this section, empirical evidence for variable selection bias with the Gini gain criterion from the literature is briefly recalled. We then provide a comprehensive statistical explanation for variable selection bias in different settings by identifying three important sources of variable selection bias, namely estimation bias and variance effect and a multiple comparisons effect.

2.1 Empirical evidence for variable selection bias

Several simulation studies have provided empirical evidence for variable selection bias in different recursive partitioning algorithms (cp., e.g., White and Liu, 1994; Kononenko, 1995; Loh and Shih, 1997; Dobra and Gehrke, 2001). We will consider one exemplary study that covers the main aspects of variable selection bias:

In their simulation study Kim and Loh (2001) vary both the number of categories in categorical predictor variables and the number of missing values in continuous predictor variables in a binary splitting framework to compare the variable selection performance of the Gini gain to that of other splitting criteria. Their results show variable selection bias towards variables with many categories and variables with many missing values. However, the authors do not give a thorough statistical explanation for their findings.

In the next section, we address three important factors that can explain the selection bias occurring with the Gini gain in the different experimental settings.

2.2 Estimation effects

The first two sources of variable selection bias can be considered as estimation effects: the classical Gini index used in machine learning can be considered as an estimator of the true underlying entropy. The bias and the variance of this estimator tend to induce selection bias.

2.2.1 Bias

From a statistical point of view the empirical Gini Index (Equation (2)) used in machine learning can be rephrased as

$$\widehat{G}_j = 2\hat{p}_j(1 - \hat{p}_j)$$

with $\hat{p}_j$ denoting the relative class frequency $\frac{N_{2j}}{N_j}$ of $Y = 2$.

The relative frequency $\hat{p}_j$ is the maximum likelihood estimator, based on N_j observations as indicated by the index j, of the true class probability p of $Y = 2$.

The empirical Gini Index $\widehat{G}_j$ here is understood as the plug-in estimator of a true underlying Gini Index

$$G = 2p(1 - p)$$

which is a function of the true class probability p.

Since the empirical Gini Index $\widehat{G}_j$ is a strictly concave function of the maximum likelihood estimator $\hat{p}_j$, we expect from Jensen's inequality that the empirical Gini Index $\widehat{G}_j$ underestimates the true Gini Index G. Infact, we find for fixed N_j:

$$\begin{aligned} \mathrm{E}(\widehat{G}_j) &= \mathrm{E}\left(2\frac{N_{2j}}{N_j}\left(1-\frac{N_{2j}}{N_j}\right)\right), \text{ where } N_{2j} \sim \mathcal{B}(N_j, p) \\ &= 2p(1-p) - 2\frac{1}{N_j}p(1-p) \\ &= \frac{N_j - 1}{N_j} G. \end{aligned}$$

Thus, the empirical Gini Index $\hat{G}_j$ underestimates the true Gini Index G by the factor $\frac{N_j-1}{N_j}$, i.e. $\hat{G}_j$ is a negatively biased estimator:

$$\mathrm{Bias}(\hat{G}_j) = -G/N_j,$$

where the extent of the bias depends on the number of observations N_j that the estimation is based on. The same principle applies to the Gini Indices $\hat{G}_{Lj}$ and $\hat{G}_{Rj}$ obtained for the child nodes created by binary splitting.

We consider the null hypothesis that the considered predictor variable X_j is uninformative, i.e. that the distribution of the response Y does not depend on X_j. With respect to the child nodes created by binary splitting this null hypothesis means that the true class probability in the left node defined by X_j, denoted by $p_{Lj} = P(Y = 2|X_j \leq x_{j(i)})$, is equal to the true class probability in the right node $p_{Rj} = P(Y = 2|X_j > x_{j(i)})$ and thus equal to the overall class probability $p = P(Y = 2)$.

The expected value of the Gini gain $\widehat{\Delta G}_j$ (Equation (3)) for fixed N_{Lj} and N_{Rj} is then

$$\begin{aligned} \mathrm{E}(\widehat{\Delta G}_j) &= \mathrm{E}(\hat{G}_j - \tfrac{N_{Lj}}{N_j}\hat{G}_{Lj} - \tfrac{N_{Rj}}{N_j}\hat{G}_{Rj}) \\ &= G - \tfrac{G}{N_j} - \tfrac{N_{Lj}}{N_j}G + \tfrac{N_{Lj}}{N_j}\tfrac{G}{N_{Lj}} - \tfrac{N_{Rj}}{N_j}G + \tfrac{N_{Rj}}{N_j}\tfrac{G}{N_{Rj}} \\ &= \tfrac{G}{N_j}. \end{aligned}$$

Under the null hypothesis of an uninformative predictor variable, the true Gini gain ΔG_j equals 0. Thus, $\widehat{\Delta G}_j$ has a positive bias, that increases with decreasing sample size N_j and is most pronounced for large values of the true Gini Index G. When the predictor variables X_j, $j = 1, \ldots, q$, have different sample sizes N_j, this bias leads to a preference of variables with small N_j, i.e. variables with many missing values. Thus the criterion shows a systematic bias even if the values are missing completely at random (MCAR).

The result of the derivation of the expected value of the Gini gain corresponds to that of Dobra and Gehrke (2001) adopted for binary splits. However, the authors do not elaborate the interpretation as an estimation bias induced by the plug-in estimation based on a limited sample size, which we find crucial for understanding the bias mechanism.

2.2.2 Variance

After some computations (see Appendix), the variance of $\widehat{G}_j$ may be written as

$$\mathrm{Var}(\widehat{G}_j) = 4\frac{G}{N_j}\left(\frac{1}{2} - G\right) + O\left(\frac{1}{N_j^2}\right).$$

The variance of the empirical Gini Index $\widehat{G}$ again depends on the true Gini Index G and increases when G moves away from its maximum value $\frac{1}{2}$ or from its minimum value zero and for small sample sizes. The variance of $\widehat{\Delta G}_j$ also substantially increases with decreasing N_j (Dobra and Gehrke, 2001). Therefore, if the predictor variables have different numbers of missing values, $\widehat{\Delta G}_j^{max}$ can take more extreme values for variables with many missing values. This effect on the variance can again lead to a preference of variables with many missing values.

In this section, we outlined two possible sources of selection bias affecting binary splitting with categorical or continuous predictor variables with different numbers of missing values. It can be shown that similar mechanisms apply in multiway splitting (Strobl, 2005). However, there is another mechanism responsible for variable selection bias: the effect of multiple comparisons, which is relevant only if the number of nodes produced in each split is smaller than the number of distinct observations or categories, as in binary splitting.

2.3 Multiple comparisons in cutpoint selection

The common problem of multiple comparisons refers to an increasing type I error-rate in multiple testing situations: When multiple statistical tests are conducted for the same data set, the chance to make a type I error for at least one of the tests increases with the number of performed tests. In the context of split selection, a type I error occurs when a variable is selected for splitting even though it is not informative.

In the case of binary splitting, the number of conducted comparisons for a given predictor variable increases with the number of possible binary partitions, i.e. with the number of possible cutpoints. In continuous predictors without ties the number of possible cutpoints to be evaluated is $N_j - 1$. For categorical and ordinal predictor variables the number of cutpoints depends on the number of categories. The ‘multiple comparisons effect’ results in a preference of predictor variables with many possible partitions: with few missing values or few ties (for continuous variables) or many categories (for categorical and ordinal variables).

This finding is not in contradiction to Dobra and Gehrke (2001), who state explicitly that variable selection bias for categorical predictor variables was not due to multiple comparisons, since the authors use the Gini gain for multiway splits with as many nodes as categories in the predictor rather than for binary splits, which does not correspond to the standard CART algorithm usually associated with the Gini criterion and obviously does not induce multiple testing effects.

The next section gives a summary of all three effects.

2.4 Resume and practical relevance

The simulation results obtained by Kim and Loh (2001) reported in section 2.1 in different settings may be explained by the three partially counteracting effects outlined in sections 2.2 and 2.3.

In the binary splitting task of Kim and Loh (2001), the bias towards predictor variables with many categories is mainly due to the multiple comparison effect: variables with more categories have more possible binary partitions to be evaluated. In contrast, the bias towards variables with many missing values observed for the metric variables may be explained by the bias and variance effects: variables with small sample sizes, for which the Gini gain is overestimated and has large variance, tend to be favored. In this case the reverse multiple comparisons effect seems to be outweighed.

In the standard simulation designs in the literature on variables selection bias all predictor variables are uninformative and thus there is no reason to prefer variables with more categories or even more missing values. These scenarios are artificial but necessary to understand the sources of variables selection bias and to evaluate the split selection criteria. In practice, the number of categories in categorical variables of nominal and ordinal scales often depends on arbitrary choices (in particular in the design of questionnaires) and randomly missing values in categorical and metric variables are common (if, e.g., questions are skipped by accident in automated data input). In such a scenario a reasonable split selection criterion should be able to identify relevant variables without being mislead by the number of categories, that may be related to - but is not in itself an indicator of - the relevance of the variable, or the number of missing values, that is inversely related to its information content.

As cited in the introduction, Breiman et al. (1984) noted the multiple comparisons effect evident when categorical predictors vary in their number of categories. In addition, they claim that their CART approach can deal particularly well with missing values, because it provides surrogate splits when predictor values are missing in the test sample. However, for missing predictor values in the learning sample, the CART algorithm applies an available case strategy when evaluating the variables in split selection, leading to the bias outlined above. This went unnoticed by Breiman et al. (1984) though, because they only spread missing values randomly over all predictor variables, instead of varying the sample sizes between variables.

In the next section, we suggest an alternative p-value selection criterion based on the Gini Index that corrects simultaneously for all three types of bias described above.

3 The distribution of the maximally selected Gini gain

3.1 A p-value based variable and split selection approach

For the case of binary splits, we introduce the p-value from the exact distribution of the maximally selected Gini gain over all possible splits as a new unbiased splitting criterion. Note that the variable index j will be dropped in most of this section focusing only on one predictor variable X with sample size N.

Beside the classification context considered here maximally selected statistics, e.g. the maximally selected χ^2- statistic or maximally selected rank statistics, have been the subject of a few tens of papers published mainly in the journal *Biometrics* in the last decades, headed by Miller and Siegmund (1982). They are based on the following idea: Suppose one computes an association measure $T(i)$ (e.g. the Gini gain or the χ^2- statistic) for all the $i = 1, \ldots, N-1$ possible cutpoints of one considered continuous predictor X and selects the cutpoint yielding the maximal association measure $T^{max} = \max_{i=1,\ldots,N-1} T(i)$. The distribution of the resulting "maximally selected" association measure T^{max} is different from the distribution of the original association measure. In particular, the distribution of the maximally selected measure may depend on the sample size N of X, causing the selection bias observed in the case of predictors with different numbers of missing values, and has to account for the deliberate choice of the cutpoint.

Possible penalizations for the choice of the optimal cutpoint in multiple comparisons are Bonferroni adjustments, which tend to overpenalize (Benjamini and Hochberg, 1995; Hawkins, 1997; Loh, 2002, for a review), and the approach of maximally selected statistics applied here.

Dobra and Gehrke (2001) on the other hand claim that p-value based criteria in general reduce the selection bias in classification trees, and derive an approximation of the distribution of the Gini gain in the case of multiway splits. Their approach does not aim at providing an unbiased split selection criterion for binary splitting, however, because it does not account for the multiple comparisons effect in cutpoint selection.

Previous applications of p-values of maximally selected statistics as unbiased split selection criteria in binary recursive partitioning are Shih and Tsai (2004), who employ the p-values of exact and approximated distributions of maximally selected split selection criteria in regression trees, and Shih (2004), who introduces the p-value of the maximally selected χ^2-statistic as an unbiased split selection criterion for classification trees. Shih (2004) explicitly states that for other criteria, e.g. for entropy criteria like the Gini Index, "the exact methods are yet to be found" (p. 465).

In the present paper, we accept the challenge posed by Shih (2004) and propose to correct the variable selection bias occurring with the Gini gain in binary splitting by using a criterion based on the exact distribution of the maximally selected Gini gain rather than the Gini gain itself.

In the rest of the paper, F denotes the distribution function of the maximally selected Gini gain under the null hypothesis of no association between the predictor and the response, given N_1 and N_2. We use the notation

$$F(d) = P_{H_0}\left(\widehat{\Delta G}^{max} \leq d\right).$$

In a nutshell, our variable and split selection approach consists of the following steps:

(1) Determine $\widehat{\Delta G}_j^{max}$ for each of the predictor variables X_j, $j = 1, \ldots, q$,

(2) compute the criterion $F\left(\widehat{\Delta G}_j^{max}\right)$ (which is equivalent to 1 - the p-value of $\widehat{\Delta G}_j^{max}$) for each variable X_j and

(3) select the variable X_{j^*} with the largest $F\left(\widehat{\Delta G}_j^{max}\right)$. The split of X_{j^*} maximizing $\widehat{\Delta G}_{j^*}(i)$ is then selected.

The rest of this section presents our method to determine the distribution function F for one predictor variable X with N non-missing independent and identically distributed observations.

3.2 Outline of the method

Our aim is to derive the distribution of the maximally selected Gini gain over the possible cutpoints of X, i.e. over all the possible partitions $\{\mathbf{S}_L, \mathbf{S}_R\}$ of the sample, under the null hypothesis of no association between X and Y. In accordance with section 1 the term $(y_{(i)}, x_{(i)})_{i=1,\dots,N}$ denotes the ordered sample. The function $n_2(i)$ again denotes the number of observations with $Y = 2$ among the i first observations (cf. equation (1)).

Obviously, we have $n_2(0) = 0$ and $n_2(N) = N_2$. Our approach to derive the exact distribution function of the maximally selected Gini gain consists of two independent steps:

(i) First, we show that the maximally selected Gini gain $\widehat{\Delta G}^{max}$ exceeds a given threshold if and only if the graph $(i, n_2(i))$ crosses the boundaries of a zone located around the line of equation with slope N_2/N and intercept 0. The coordinates of these boundaries are derived in section 3.3.

(ii) The probability that the graph $(i, n_2(i))$ crosses the boundaries under the null hypothesis of no association between X and Y is computed via a combinatorial method in the spirit of Koziol (1991), to determine the distribution of the maximally selected χ^2- statistic.

Our two-step approach can be seen as an extension of Koziol's method. We use the same combinatorial method, but with new boundaries corresponding to the Gini gain instead of the χ^2- statistic. This approach could be generalized to other splitting criteria for which a condition of the type of (5) (see section 3.3) can be formulated. In the rest of this section, we derive the new boundaries corresponding to the Gini gain (section 3.3) and adapt Koziol's combinatorial computation method (section 3.4).

3.3 Definition of the boundaries

The Gini gain $\widehat{\Delta G}(i)$ obtained by cutting between $x_{(i)}$ and $x_{(i+1)}$ may be rewritten (cf. again table 1 for the notation) as a quadratic function of $n_2(i)$:

$$\begin{aligned}
\widehat{\Delta G}(i) &= \widehat{G} - \tfrac{i}{N}\left[2\tfrac{n_2(i)}{i}\left(1-\tfrac{n_2(i)}{i}\right)\right] - \tfrac{N-i}{N}\left[2\tfrac{(N_2-n_2(i))}{N-i}\left(1-\tfrac{N_2-n_2(i)}{N-i}\right)\right]\\
&= 2\tfrac{N_2}{N}\left(1-\tfrac{N_2}{N}\right) - 2\tfrac{N_2}{N} + 2\tfrac{n_2(i)^2}{iN} + 2\tfrac{(N_2-n_2(i))^2}{N(N-i)}\\
&= n_2(i)^2\left(\tfrac{2}{iN} + \tfrac{2}{N(N-i)}\right) - n_2(i)\tfrac{4N_2}{N(N-i)} - 2\tfrac{N_2^2}{N^2} + 2\tfrac{N_2^2}{N(N-i)}\\
&= n_2(i)^2\tfrac{2}{i(N-i)} - n_2(i)\tfrac{4N_2}{N(N-i)} + \tfrac{2iN_2^2}{N^2(N-i)}.
\end{aligned}$$

For $d \geq 0$, we have:

$$\widehat{\Delta G}(i) \leq d \quad \Leftrightarrow \quad n_2(i)^2 \frac{2}{i(N-i)} - n_2(i)\frac{4N_2}{N(N-i)} + \frac{2iN_2^2}{N^2(N-i)} - d \leq 0 \quad (4)$$

With the notations

$$\begin{aligned} a_i &= \tfrac{2}{i(N-i)}, \\ b_i &= -\tfrac{4N_2}{N(N-i)}, \end{aligned}$$

we obtain after simple computations that

$$\widehat{\Delta G}(i) \leq d \quad \Leftrightarrow \quad n_2(i) \in \left[\frac{-b_i - \sqrt{\frac{8d}{i(N-i)}}}{2a_i}, \frac{-b_i + \sqrt{\frac{8d}{i(N-i)}}}{2a_i} \right]. \qquad (5)$$

We want to derive the distribution function of

$$\widehat{\Delta G}^{max} = \max_{i=1,\ldots,N-1} \widehat{\Delta G}(i)$$

under the null hypothesis of no association between X and Y, i.e. $P_{H_0}\left(\widehat{\Delta G}^{max} \leq d\right)$ for any $d \geq 0$. We have $\widehat{\Delta G}^{max} \leq d$ if and only if condition (5) holds for all i in $1, \ldots, N-1$, i.e. if and only if the path $(i, n_2(i))$ remains on or above the graph of the function

$$\text{lower}_d(i) = \frac{-b_i - \sqrt{\frac{8d}{i(N-i)}}}{2a_i}$$

and on or under the graph of the function

$$\text{upper}_d(i) = \frac{-b_i + \sqrt{\frac{8d}{i(N-i)}}}{2a_i}.$$

A sufficient and necessary condition for $\widehat{\Delta G}^{max} \leq d$ is that the graph $(i, n_2(i))$ does not pass through any point of integer coordinates (i_0, j_0) with $i_0 = 1, \ldots, N-1$ and

$$\text{lower}_d(i_0) - 1 \leq j_0 < \text{lower}_d(i_0),$$

or

$$\text{upper}_d(i_0) < j_0 \leq \text{upper}_d(i_0) + 1.$$

Let us denote these points as $B_1, \ldots, B_k$ and their coordinates as $(i_1, j_1), \ldots, (i_k, j_k)$, where $B_1, \ldots, B_k$ are labeled in order of increasing abscissa and increasing ordinate for each value of the abscissa. The exact computation of the probability that the graph $(i, n_2(i))$ passes through at least one of the points $B_1, \ldots, B_k$ (i.e. that it leaves the boundaries defined above) under the null hypothesis of no association between X and Y is described in the next section. Exemplary boundaries are displayed in Figure 1.

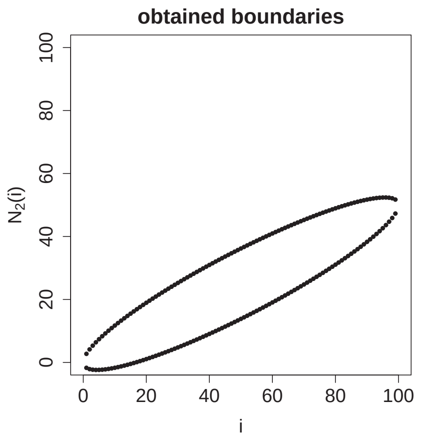


Fig. 1. Boundaries as defined in section 3.3 for an example with $N_1 = N_2 = 50$ and d=0.1

3.4 Koziol's combinatorial approach

Under the null hypothesis of no association between X and Y, all the possible paths $(i, n_2(i))$ have equal probability $1/\binom{N}{N_2}$. Thus, the probability that the path $(i, n_2(i))$ passes through at least one of the points $B_1, \ldots, B_k$ can be computed using the combinatorial approach of Koziol (1991). This approach is based on a Markov representation of $n_2(i)$ as the path of a binomial process with constant probability of success and with unit jumps, conditional on $n_2(N) = N_2$. Let $\mathcal{P}_s$ denote the set of the paths from $(0, 0)$ to B_s that do not pass through points $B_1, \ldots, B_{s-1}$ and b_s the number of paths in $\mathcal{P}_s$. Since the sets $\mathcal{P}_s$, $s = 1, \ldots, k$, are mutually disjoint, b_s, $s = 1, \ldots, k$, can be computed recursively as

$$\begin{aligned} b_1 &= \tbinom{i_1}{j_1} \\ b_s &= \tbinom{i_s}{j_s} - \textstyle\sum_{r=1}^{s-1} \tbinom{i_s - i_r}{j_s - j_r} b_r, \; s = 2, \ldots, k. \end{aligned}$$

The above formula can also be derived by means of simple combinatorial considerations: The number of paths from $(0, 0)$ to B_s is given by $\binom{i_s}{j_s}$. To obtain the number of paths from $(0, 0)$ to B_s that do not pass through any of the $B_1, \ldots, B_{s-1}$, one has to subtract from $\binom{i_s}{j_s}$ the sum over $r = 1, \ldots, s-1$ of the numbers of paths from $(0, 0)$ to B_s that pass through B_r but not through $B_1, \ldots, B_{r-1}$. For a given r $(r < s)$, the number of paths from $(0, 0)$ to B_r that do not pass through $B_1, \ldots, B_{r-1}$ is b_r and the number of paths from B_r to B_s is $\binom{i_s - i_r}{j_s - j_r}$, which gives the product $\binom{i_s - i_r}{j_s - j_r} b_r$ in the sum in the above formula.

The number of paths from $(0, 0)$ to (N, N_2) that pass through B_s, $s = 1, \ldots, k$, but not through $B_1, \dots, B_{s-1}$ is then given as

$$\binom{N - i_s}{N_2 - j_s} b_s.$$

Since all the possible paths are equally likely under the null hypothesis, the probability that the graph $(i, n_2(i))$ passes through at least one of the points $B_1, \dots, B_k$ is simply obtained as

$$P_{H_0}(\widehat{\Delta G}^{max} > d) = \binom{N}{N_2}^{-1} \sum_{s=1}^{k} \binom{N - i_s}{N_2 - j_s} b_s. \tag{6}$$

It follows

$$F(d) \;=\; P_{H_0}\left(\widehat{\Delta G}^{max} \leq d\right) \;=\; 1 - \binom{N}{N_2}^{-1} \sum_{s=1}^{k} \binom{N - i_s}{N_2 - j_s} b_s. \tag{7}$$

We implemented the computation of the boundaries (step (i)) as well as the result of the combinatorial derivation of $F(d) = P_{H_0}\left(\widehat{\Delta G}^{max} \leq d\right)$ (step (ii)) in the R system for statistical computing (R Development Core Team, 2006). The boundaries depicted in Figure 1 are obtained for $N_1 = N_2 = 50$ and $d = 0.1$.

4 Simulation studies

In this section, simulation studies are conducted to compare the variable selection performance of the p-value criterion derived in section 3 to that of the standard Gini gain criterion. We consider a binary response variable Y and 5 mutually independent continuous predictor variables X_1, X_2, X_3, X_4, X_5. In the whole simulation study, the binary response Y is sampled from a Bernoulli distribution with probability of success 0.5. The manipulated parameter is the percentage of missing values in the predictor variable X_1, set successively to 0%, 20%, 40%, 60% and 80%. The missing values are sampled completely at random from variable X_1 in each setting. The sample size is set to $N = 100$. Three cases are investigated:

- **Null case:** all the predictor variables X_1, X_2, X_3, X_4, X_5 are uninformative, i.e. independent of the response variable.
- **Power case I:** X_1 is informative and X_2, X_3, X_4, X_5 are uninformative.
- **Power case II:** X_2 is informative and X_1, X_3, X_4, X_5 are uninformative.

For each parameter setting 1000 data sets are generated. For each data set, variable selection is performed using successively the standard Gini gain and

Table 2
Null case: Variable selection frequencies. The symbol ∘ indicates a varying number of missing values in the marked variable with the percentage of missing values displayed in the left column.

	Gini gain					**p-value criterion**				
	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5
	∘					∘				
0%	0.20	0.21	0.20	0.20	0.19	0.20	0.21	0.20	0.20	0.19
20%	0.28	0.19	0.18	0.18	0.17	0.18	0.21	0.21	0.21	0.20
40%	0.50	0.14	0.13	0.12	0.12	0.24	0.22	0.21	0.17	0.19
60%	0.67	0.09	0.07	0.07	0.09	0.22	0.20	0.20	0.19	0.21
80%	0.91	0.02	0.03	0.03	0.02	0.23	0.18	0.19	0.20	0.21

our p-value criterion. For both criteria, the obtained relative frequencies of selection out of the 1000 simulation runs for all variables are given in tables. Based on the reviewed literature and our theoretical results in section 2, we expect the Gini gain criterion to be biased towards the predictor variable with missing values, regardless of its information content.

4.1 Null case

In the null case study, X_1, X_2, X_3, X_4, X_5 are sampled from the standard normal distribution

$$X_j \sim \mathcal{N}(0,1), \text{ for } j = 1, \ldots, 5.$$

For each percentage of missing values (MCAR), the obtained frequencies of selection of X_1, X_2, X_3, X_4, X_5 over the 1000 simulation runs are given in Table 2 for the Gini gain (left) and the p-value criterion (right). Since the predictor variables are all independent of the response Y, a good criterion is supposed to select X_1, X_2, X_3, X_4 and X_5 with equal probability $\frac{1}{5}$.

We find that for the Gini gain criterion the selection frequency of X_1 increases with the amount of missing values, while it decreases for all other variables. In contrast, the p-value criterion shows almost no variable selection bias.

4.2 Power case I

In the first power case study, the four uninformative predictor variables X_2, X_3, X_4, X_5 are sampled from the standard normal distribution, while the predictor variable X_1 is informative now and still contains missing values. X_1 is sampled from

$$X_1|Y = 1 \sim \mathcal{N}(0, 1)$$
$$X_1|Y = 2 \sim \mathcal{N}(0.5, 1).$$

(We sampled $X_1|Y$ rather than $Y|X_1$ only to be able to control the number of class 1 and 2 observations in each iteration. The reverse sampling scheme produces the same effect.)

The manipulated parameter is again the percentage of missing values (MCAR) in the now informative predictor variable X_1, with successively 0%, 20%, 40%, 60% and 80% of the original sample size missing completely at random. All other predictors contain no missing values. With a sensible selection criterion, the selection frequency of the informative predictor variable X_1 is supposed to decrease when the number of randomly missing values increases, because the information contained in the observed values of the variable actually decreases (cf. Shih, 2004; Shih and Tsai, 2004).

Table 3 summarizes the variable selection frequencies for all variables in the power case I design with X_1 being informative and containing missing values. We find that for the Gini gain criterion the selection frequency of X_1 increases with its amount of missing values, despite the loss of information content. In contrast, the p-value criterion selects X_1 less often when it has many missing values. This dependence of the selection frequency on the number of available cases of the informative predictor variable corresponds to the findings of Shih (2004) for the p-value of the maximally selected χ^2-statistic, and is a desirable property for a split selection criterion.

If the underlying missing mechanism is known to be missing not at random, however, the missing mechanism should be modeled accordingly. Otherwise our approach will behave conservatively and underrate the information content of the variable.

4.3 Power case II

In the second power case study, the four uninformative predictor variables X_1, X_3, X_4, X_5 are sampled from standard normal distributions, while now X_2

Table 3
Power case I: Variable selection frequencies. The ○ symbol indicates a varying number of missing values in the marked variable with the percentage of missing values displayed in the rows of the table. The ● symbol indicates that the marked variable is also an informative predictor.

	Gini gain					**p-value criterion**				
	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5
	●					●				
	○					○				
0%	0.71	0.07	0.08	0.06	0.08	0.71	0.07	0.08	0.06	0.08
20%	0.77	0.06	0.06	0.06	0.06	0.66	0.08	0.08	0.09	0.09
40%	0.79	0.05	0.06	0.05	0.05	0.58	0.12	0.12	0.11	0.09
60%	0.84	0.06	0.03	0.04	0.03	0.45	0.16	0.13	0.14	0.13
80%	0.94	0.01	0.01	0.02	0.01	0.35	0.16	0.17	0.16	0.15

is the informative predictor variable sampled from

$$X_2|Y = 1 \sim \mathcal{N}(0, 1)$$
$$X_2|Y = 2 \sim \mathcal{N}(0.5, 1).$$

X_1 now is not informative but still contains missing values. The manipulated variable is again the percentage of missing values (MCAR) in the uninformative predictor variable X_1 with successively 0%, 20%, 40%, 60% and 80% of the original sample size missing completely at random. The other predictors contain no missing values. We expect the estimated probability of X_1 being selected as splitting variable to increase with the percentage of missing values in X_1 for the Gini gain, despite the higher information content of X_2, but not for the p-value criterion.

Table 4 summarizes the variable selection frequencies for all variables in the power case II design. We find again that the selection frequency of X_1 indeed increases with its amount of missing values for the Gini gain criterion, outweighing the higher information content of X_2. This effect is also depicted in Figure 2. In contrast, the p-value criterion shows no variable selection bias.

Table 4

Power case II: Variable selection frequencies. The ○ symbol indicates a varying number of missing values in the marked variable with the percentage of missing values displayed in the left column. The symbol • indicates that the marked variable is an informative predictor.

	Gini gain					**p-value criterion**				
	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5
		•					•			
	○					○				
0%	0.07	0.73	0.07	0.07	0.07	0.07	0.73	0.07	0.07	0.07
20%	0.12	0.69	0.07	0.07	0.06	0.07	0.72	0.07	0.07	0.06
40%	0.21	0.64	0.05	0.04	0.06	0.06	0.73	0.07	0.06	0.08
60%	0.42	0.47	0.03	0.03	0.05	0.07	0.73	0.06	0.06	0.09
80%	0.74	0.23	0.01	0.01	0.01	0.08	0.71	0.07	0.07	0.09

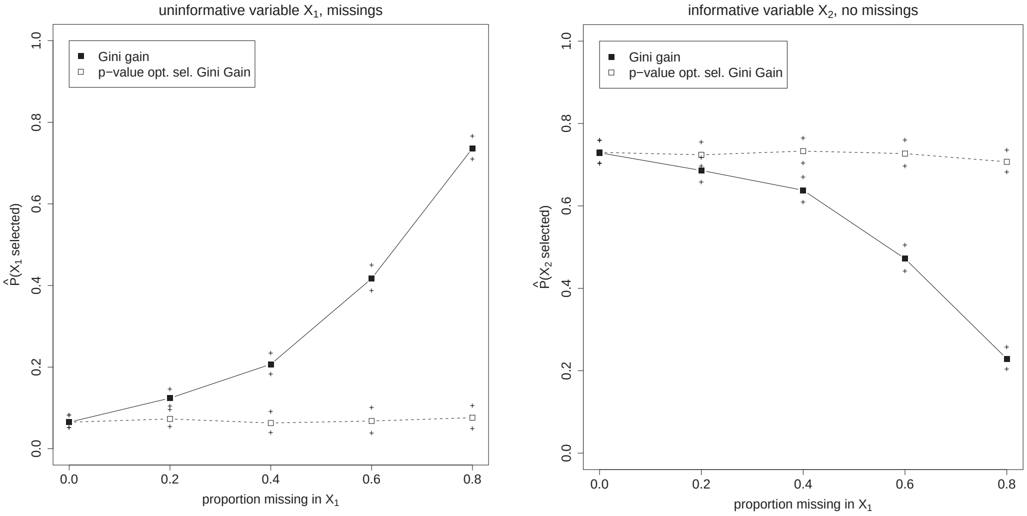


Fig. 2. Power case II: Variable selection frequencies for the uninformative variable X_1 containing missing values (left) and the informative variable X_2 containing no missing values (right).

5 Application to veterinary data

5.1 Data set

The data were collected in 2004 at a research farm in the area of Munich, Germany (Schmaußer, 2005). They contain various measurements recorded for 51 cows from the week of their first delivery (week 0) until the fourth week post partum (week 4). The binary response variable of interest takes value $Y = 1$ if the cow shows no signs of genital infection or signs of a minor genital infection only and $Y = 2$ if it shows signs of a major genital infection or even puerperal sepsis (childbed fever) and pyometra (uterine suppuration). The potential predictor variables are measures of body condition, various parameters of the hemogram, milk production, energy consumption and gynecological indicators that are displayed in Table 5.

The predictor variables vary strongly in their numbers of missing values, e.g., between 0 and 50 in week 0 and between 0 and 25 in week 4. Some variables contain less than three observations for some of the weeks, which is obviously not a reasonable sample size in a binary classification task. These variables were excluded from the analysis for the considered week (week 0: USHR, USHL; week 1: FFS; week 3: FFS).

With this application we want to point out that in practice the Gini gain and the p-value criterion rank predictor variables substantially differently with respect to their number of missing values as we have expected from our theoretical and simulation results. In addition, we explore the explanatory power of the variables that would be selected for the first split with each criterion.

For the following exemplary analysis we treat the missing values as if they were missing completely at random within each variable, even though this assumption is debatable in this context. As stated above, our approach will behave conservatively and underrate the information content of the variable if the true underlying missing mechanism was informative.

The analysis is carried out for each week separately, because the longitudinal structure is not in focus here.

5.2 Variable selection ranking

The Gini gain criterion and our novel p-value criterion may be used to rank the variables: the least informative variable is assigned rank 1, and so on. In this section, the rankings of the predictor variables obtained by the Gini

Table 5
Potential predictor variables from the cow data set. All variables are measured on a metric scale but contain strongly varying numbers of missing values.

body condition	BCS	body condition score
	RFD	backfat thickness (mm)
	MD	muscle thickness (mm)
hemogram	FFS	free fatty acids (μmol/l)
	Caro	carotene (μg/l)
	Bili	bilirubin (μmol/l)
	AST	aspartate aminotransferase (U/l)
	CK	creatine kinase (U/l)
	AP	alkaline phosphatase (U/l)
	GLDH	glutamate dehydrogenase (U/l)
	GGT	gamma glutamiltransferase (U/l)
	BHB	beta hydroxybutyric acid (mmol/l)
	IGF1	insulin growth factor 1 (nmol/l)
milk production	Milch	milk yield (kg)
	FettM	milk fat (week mean; %)
	EiM	milk protein (week mean; %)
	FEQ	fat-protein-ratio
	LaktM	milk lactose (week mean; %)
	FLQ	fat-lactose-ratio
	HarnM	milk carbamide (week mean; mmol/l)
energy consumption	TMGes	dry matter intake total (kg)
	Eauf	energy intake (MJ NEL)
	EbedM	energy requirement (MJ NEL)
	EbilM	energy balance (MJ NEL)
gynecology	UZD	cervix diameter (cm)
	USHR	uterine horn diameter right (cm)
	USHL	uterine horn diameter left (cm)

gain criterion and with our p-value criterion are compared. Due to selection bias of the Gini gain towards variables with many missing values, the two rankings are expected to diverge substantially. The scatterplots of the two rankings are displayed in Figure 3 for each week. The number of missing values is represented by the circumference of the corresponding spot. It can be observed from the scatterplots that indeed

- the spots deviate noticeably from the bisector,
- the deviation from the bisector is linked to the number of missing values.

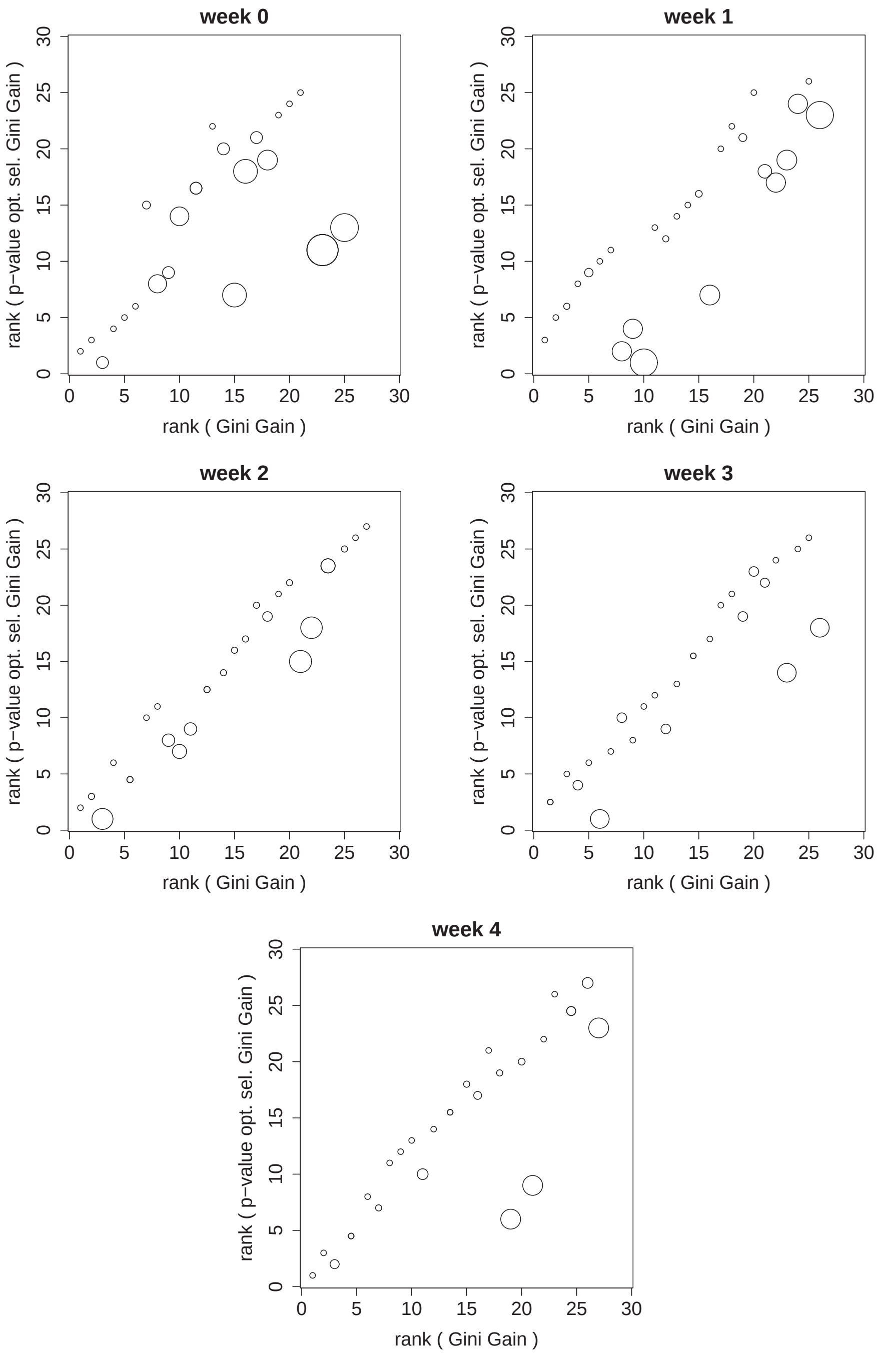
Variables with more missing values tend to be ranked higher by the Gini gain criterion than with our p-value criterion. Considering the results of this and the previous sections it is thus practically relevant to use the unbiased p-value criterion instead of the biased Gini gain for variable selection. In classification trees, the variable ranked highest by the chosen criterion is then selected for splitting.

5.3 Selected splitting variables

In this section, we examine the variables selected for the first split in each week with the standard Gini gain and with our p-value criterion. When comparing the variables we take into account the number of missing values, and additionally compute logistic regression models for the binary response and each selected variable individually. The p-value of the likelihood ratio χ^2- test of logistic regression models does not strictly match with the deterministic bisection approach of classification trees, but can serve as another indicator of the explanatory power of the selected variables. The results are summarized in Table 6.

We find again in Table 6 that the Gini gain criterion systematically prefers variables with high numbers of missing values. For example, the variable UZD selected by the Gini gain in week 0 has 39 missing values and only 12 observed values. It should thus be treated with caution. In contrast, the variables selected by our p-value criterion do not have any or have only few missing values. Through all weeks the p-values of the logistic regression model (abbreviated by LRM) are lower for the variables selected by our p-value criterion than for those selected by the Gini gain criterion in each week. This indicates a higher explanatory power of the variables selected by our p-value criterion in this data set.

Moreover, our p-value criterion may be used as a stopping rule when constructing a classification tree: We suggest to fix a threshold for the p-value criterion, e.g. 0.95. The considered node is split only if the criterion value of the selected variable exceeds this threshold, i.e. if the corresponding p-value


week 0

week 1

week 2

week 3

week 4

Fig. 3. Rank obtained with the new p-value criterion vs. rank obtained with the Gini gain. The circumference of each point is proportional to number of missing values in the predictor.

is ≤ 0.05. In this example the split with the selected variable would be conducted for weeks 0 through 3 (with the level of significance indicated by the * and ** symbols); only in week 4 the split does not produce enough impurity reduction and is omitted if the threshold is fixed at 0.95. If the threshold was set to .99 the split would be conducted in weeks 0 through 2 (indicated by **).

Table 6
Variables selected for the first split using the standard Gini gain (top) and our p-value criterion (bottom). The p-values from the logistic regression model (LRM) that correspond to model likelihood ratio tests significant on a 5%-level are indicated by the * symbol, those significant on a 1%-level by the ** symbol.

	week 0	week 1	week 2	week 3	week 4
Gini gain					
selected variable	**UZD**	**UZD**	**Bili**	**BCS**	**BCS**
missing values	39	38	0	23	25
p-value LRM	0.094	0.028*	0.001**	0.305	0.121
p-value criterion					
selected variable	**Bili**	**GLDH**	**Bili**	**Caro**	**USHL**
missing values	0	0	0	0	9
p-value LRM	0.007**	0.003**	0.001**	0.207	0.059
criterion value	**0.990****	**0.999****	**0.994****	**0.983***	**0.927**

This way to proceed is compatible with the insignificant results of the logistic regression models in weeks 3 and 4.

6 Discussion and conclusion

In this paper, we derived the exact distribution of the maximally selected Gini gain under the null hypothesis of no association between the binary response variable and a continuous predictor. The resulting p-value can be applied as a split selection criterion in recursive partitioning algorithms, as well as as an information measure in 2 × 2 tables where the cutpoint is preselected such as to optimize the separation of the response classes.

Our p-value based approach for split and variable selection avoids all sources of variable selection bias examined in section 2. The estimation bias and variance effects as well as the multiple comparisons effects are overcome by considering the maximally selected Gini gain given the class sizes N_{1j} and N_{2j}. In simulation and real data studies, our approach has proved to deal effectively with different amounts of randomly missing values in the predictor variables. The implementation of our method in the R system for statistical computing, that was used in this work, is freely available in the package exactmaxsel downloadable from *www.r-project.org*.

Other strategies to cope with randomly missing values in classification tree induction have been proposed in the machine learning literature. Most of them are imputation methods (see e.g. Quinlan, 1986; Liu et al., 1997, for a comprehensive review). Apart from any skepticism against imputation methods our approach has the advantage that it detects the information drop in informative variables caused by an increasing number of missing values.

Our p-value based approach may be applied to other common selection criteria such as the deviance (also called cross-entropy). In future research we are working on a generalization to categorical and ordinal predictors using the boundaries defined in Boulesteix (2006a) and Boulesteix (2006b) for use in classification trees. In this context, our p-value criterion would address the problem of missing values and the problem of different numbers of categories simultaneously.

Another advantage of our method is that it is based on the popular Gini index, with possible extensions to other impurity measures. The easily tangible impurity measures may attract applied scientists without a strong statistical background more than classical association test statistics (e.g. in combination with Bonferroni adjustment for multiple testing) as split selection criteria. Our criterion can replace the Gini gain criterion in the traditional greedy search approach of CART, the intuitiveness of which has played a crucial role in making classification trees understandable and attractive to a broad scientific community.

Our method is well suited for medium and small data sets like the ones presented in our simulation and real data studies, because as an exact combinatorial approach it is applicable and valid even for small sample sizes. This is particularly important in recursive partitioning, where the starting sample size is divided in (at least) two nodes in every split, leading to a rapid decrease of the sample size underlying the next split decision. Therefore only an exact method can guarantee a valid computation of the p-value used as split selection criterion even in the bottom nodes, that may hold only very small sample sizes. Of course our method is also applicable to larger samples with some computational expense: The computation time increases quadratically with increasing sample size, because the number of binomial coefficients to be evaluated increases quadratically in N. For $N = 100$ the computation time (user cpu time on a 64 bit unix machine) for the exact p-value is in the region of 0.35 seconds and for $N = 1000$ in the region of 35 seconds correspondingly.

At this point, our approach is limited to the case of a binary response in order to limit the complexity of the exact combinatorial method. An extension of our approach to more than two classes would be a challenging but interesting field for further studies. However, the binary case we focus on here is most common in applications, as e.g. in medical studies, and well suited to illustrate how unbiased variable selection can be accomplished by means of the exact p-value of a maximally selected statistic.

Different authors argue along the lines of Kass (1980) and Loh and Shih (1997), who state that the key to avoiding variable selection bias is to separate the process of variable selection from that of cutpoint selection. The unbiased algorithms QUEST (Loh and Shih, 1997) and CRUISE (Kim and Loh, 2001) e.g. employ association test statistics (of the ANOVA F-test for metric predictors and of the χ^2-test for categorical predictors) for variable selection. The split is selected subsequently using discriminant analysis techniques.

Most recently Hothorn, Hornik, and Zeileis (2006) propose a unifying conditional inference approach that also separates variable selection from cutpoint selection. Here, p-values from an asymptotic distribution of linear association test statistics are used for unbiased variable selection; the cutpoint in the selected variable is then derived within the same framework.

However, we argue that, in order to achieve unbiased variable selection in classification trees, it is neither necessary to give up the popular impurity measures, nor to give up the greedy search approach that attracted such a diverse group of applicants with different statistical background. Giving up the greedy search approach of the traditional recursive partitioning algorithms for an advanced statistical modeling approach might, as an unwanted side effect, result in leaving those applicants with a weaker statistical background behind - with easy to handle but biased classification trees.

Using a p-value criterion based on the Gini index, we address efficiently the problem of selection bias but preserve the simplicity of traditional classification trees with binary splits. In addition, the p-value can provide a statistically sound stopping criterion. Our exact procedure is able to handle small sample sizes, as e.g. in the bottom nodes of a classification tree, more relyably than asymptotic approaches, but, as all exact procedures, is computationally intensive for large samples. The p-value criterion can be integrated into any traditional recursive partitioning algorithm and might thus prove both manageable and useful for applied scientists, as demonstrated in the veterinary example.

Acknowledgements

The authors would like to thank R. Mansfeld and M. Schmaußer from the Veterinary Faculty of the Ludwig-Maximilians-University Munich for providing the veterinary data, and F. Barth, C. Heumann and G. Tutz for fruitful discussions. Financial support from the German Research Foundation (DFG), Collaborative Research Center (SFB) 386 "Statistical Analysis of Discrete Structures", is gratefully acknowledged.

References

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B 57 (1), 289 – 300.

Bittencourt, H. R., Clarke, R. T., 2004. Feature selection by using classification and regression trees. In: Altan, O. (Ed.), The International Archives of the Photogrammetry, Proceedings of the 20th Congress of the International Society for Photogrammetry and Remote Sensing (ISPRS 2004), Istanbul, Turkey. pp. 66–70.

Boulesteix, A.-L., 2006a. Maximally selected chi-square statistics and binary splits of nominal variables. Biometrical Journal 48 (5), 838–848.

Boulesteix, A.-L., 2006b. Maximally selected chi-square statistics for ordinal variables. Biometrical Journal 48 (3), 451–462.

Boulesteix, A.-L., Tutz, G., 2006. Identification of interaction patterns and classification with applications to microarray data. Computational Statistics and Data Analysis 50 (3), 783–802.

Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Chapman and Hall, New York, NY, USA.

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

Dong, G., Li, J., 1999. Efficient mining of emerging patterns: Discovering trends and differences. In: Proceedings of the 5th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD 1999), San Diego, CA, USA. ACM Press, pp. 43–52.

Evans, M., Hastings, N., Peacock, B., 1993. Statistical Distributions. John Wiley & Sons, Inc., New York, NY, USA.

Hawkins, D. M., 1997. Firm: Formal inference-based recursive modeling. Release 2.1, Technical Report 546, School of Statistics, University of Minnesota, MN, USA.

Hothorn, T., Hornik, K., Zeileis, A., 2006. Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics 15 (3), 651–674.

Jong, O., Laubach, M., Luczak, A., 2005. Estimating neuronal variable importance with random forest. In: Reisman, S., Foulds, R. (Eds.), Proceedings of the 29th Annual Northeast Bioengineering Conference, Newark, NJ, USA. pp. 33–34.

Kass, G., 1980. An exploratory technique for investigating large quantities of categorical data. Applied Statistics 29 (2), 119–127.

Kim, H., Loh, W., 2001. Classification trees with unbiased multiway splits. Journal of the American Statistical Association 96 (454), 589–604.

Kononenko, I., 1995. On biases in estimating multi-valued attributes. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI 1995), Montral, Quebec, Canada. Morgan Kaufmann, pp. 1034–1040.

Koziol, J. A., 1991. On maximally selected chi-square statistics. Biometrics 47 (4), 1557–1561.

Little, R., Rubin, D., 1986. Statistical analysis with missing data. John Wiley & Sons, Inc., New York, NY, USA.

Little, R., Rubin, D., 2002. Statistical analysis with missing data, 2nd edition. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Liu, W., White, A., Thompson, S., Bramer, M., 1997. Techniques for dealing with missing values in classificaton. In: Liu, X., Cohen, P., Berthold, M. (Eds.), Advances in Intelligent Data Analysis (IDA 1997), London, UK. pp. 527–536.

Loh, W., 2002. Regression trees with unbiased variable selection and interaction detection. Statistica Sinica 12 (2), 361–386.

Loh, W., Shih, Y., 1997. Split selection methods for classification trees. Statistica Sinica 7 (4), 815–840.

Miller, R., Siegmund, D., 1982. Maximally selected Chi square statistics. Biometrics 38 (4), 1011–1016.

Quinlan, J. R., 1986. Induction of decision trees. Machine Learning 1 (1), 81–106.

Quinlan, R., 1993. C4.5: Programms for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

R Development Core Team, 2006. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, (http://www.R-project.org).

Schmaußer, M., 2005. Auswirkungen verschiedener Stoffwechsellagen auf die Fertilität beim Milchrind unter besonderer Berücksichtigung der individuellen Futteraufnahme und unter Berücksichtigung verschiedener Melksysteme. PhD Thesis, Faculty of Veterinary Medicine, University of Munich LMU, Munich, Germany.

Shih, Y., 2004. A note on split selection bias in classification trees. Computational Statistics and Data Analysis 45 (3), 457–466.

Shih, Y., Tsai, H., 2004. Variable selection bias in regression trees with constant fits. Computational Statistics and Data Analysis 45 (3), 595–607.

Strobl, C., 2005. Variable selection in classification trees based on imprecise probabilities. In: Cozman, F., Nau, R., Seidenfeld, T. (Eds.), Proceedings of the Fourth International Symposium on Imprecise Probabilities and their Applications, Carnegy Mellon University, Pittsburgh, PA, USA. SIPTA, Manno, pp. 340–348.

Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8:25.

White, A., Liu, W., 1994. Bias in information based measures in decision tree induction. Machine Learning 15 (3), 321–329.

Appendix

Derivation of the variance of the empirical Gini Index $Var(\widehat{G}_j)$ displayed in section 2.2. Note that the index j for variable X_j is suppressed in the following.

$$\begin{aligned}
Var(\widehat{G}) &= Var(2\widehat{p}(1-\widehat{p})) \\
&= 4\ Var(\widehat{p}(1-\widehat{p})) \\
Var(\widehat{p}(1-\widehat{p})) &= E(\widehat{p}^2(1-\widehat{p})^2) - E(\widehat{p}(1-\widehat{p}))^2 \\
&= E(\widehat{p}^2) - 2E(\widehat{p}^3) + E(\widehat{p}^4) - \tfrac{1}{4}E(\widehat{G})^2
\end{aligned}$$

We compute and approximate the four terms successively (e.g., by means of the moment generating function of the Binomial distribution, cf. Evans et al.

(1993), p.38), and obtain:

$$\begin{aligned}
E(\widehat{p}^2) &= \tfrac{1}{N^2}E(Z^2),\ \text{where } Z \sim \mathcal{B}(N,p) \\
&= \tfrac{p}{N} + p^2 - \tfrac{p^2}{N} \\
&= p^2 + \tfrac{p(1-p)}{N} \\
-2E(\widehat{p}^3) &= -2(\tfrac{1}{N^3}E(Z^3)),\ \text{where } Z \sim \mathcal{B}(N,p) \\
&= -2(\tfrac{3p^2}{N} + p^3 - \tfrac{3p^3}{N} + O(\tfrac{1}{N^2})) \\
&= -\tfrac{6p^2}{N} - 2p^3 + \tfrac{6p^3}{N} + O(\tfrac{1}{N^2}) \\
E(\widehat{p}^4) &= \tfrac{1}{N^4}E(Z^4),\ \text{where } Z \sim \mathcal{B}(N,p) \\
&= \tfrac{6p^3}{N} + p^4 - \tfrac{6p^4}{N} + O(\tfrac{1}{N^2}) \\
-\tfrac{1}{4}E(\widehat{G})^2 &= -\tfrac{1}{4}\tfrac{(N-1)^2}{N^2}G^2 \\
&= -G^2(\tfrac{1}{4} - \tfrac{1}{2N}) + O(\tfrac{1}{N^2})
\end{aligned}$$

Finally,

$$\begin{aligned}
Var(\widehat{p}(1-\widehat{p})) &= p^2 + \tfrac{p(1-p)}{N} - \tfrac{6p^2}{N} - 2p^3 + \tfrac{6p^3}{N} + \tfrac{6p^3}{N} + p^4 - \tfrac{6p^4}{N} - G^2(\tfrac{1}{4} - \tfrac{1}{2N}) + O(\tfrac{1}{N^2}) \\
&= (p^2 - 2p^3 + p^4)(1 - \tfrac{6}{N}) + \tfrac{G}{2N} - G^2(\tfrac{1}{4} - \tfrac{1}{2N}) + O(\tfrac{1}{N^2}) \\
&= \tfrac{G2}{4}(1 - \tfrac{6}{N}) + \tfrac{G}{2N} - G2(\tfrac{1}{4} - \tfrac{1}{2N}) + O(\tfrac{1}{N^2}) \\
&= \tfrac{G2}{N}(-\tfrac{6}{4} + \tfrac{1}{2}) + \tfrac{G}{2N} + O(\tfrac{1}{N^2}) \\
&= \tfrac{G}{N}(\tfrac{1}{2} - G) + O(\tfrac{1}{N^2}) \\
Var(2\widehat{p}(1-\widehat{p})) &= 4\tfrac{G}{N}(\tfrac{1}{2} - G) + O(\tfrac{1}{N^2}).
\end{aligned}$$