Statistical Issues in Machine Learning – Towards Reliable Split Selection and Variable Importance Measures

Dissertation

am

Institut für Statistik

der

Fakultät für Mathematik, Informatik und Statistik

der

Ludwig-Maximilians-Universität München

Vorgelegt von: Carolin Strobl

München, den 26. Mai 2008

Erstgutachter: Prof. Dr. Thomas Augustin

Zweitgutachter: Prof. Dr. Gerhard Tutz

Externer Gutachter: Prof. Dr. Kurt Hornik

Rigorosum: 2. Juli 2008

Abstract

Recursive partitioning methods from machine learning are being widely applied in many scientific fields such as, e.g., genetics and bioinformatics. The present work is concerned with the two main problems that arise in recursive partitioning, instability and biased variable selection, from a statistical point of view. With respect to the first issue, instability, the entire scope of methods from standard classification trees over robustified classification trees and ensemble methods such as TWIX, bagging and random forests is covered in this work. While ensemble methods prove to be much more stable than single trees, they also loose most of their interpretability. Therefore an adaptive cutpoint selection scheme is suggested with which a TWIX ensemble reduces to a single tree if the partition is sufficiently stable. With respect to the second issue, variable selection bias, the statistical sources of this artifact in single trees and a new form of bias inherent in ensemble methods based on bootstrap samples are investigated. For single trees, one unbiased split selection criterion is evaluated and another one newly introduced here. Based on the results for single trees and further findings on the effects of bootstrap sampling on association measures, it is shown that, in addition to using an unbiased split selection criterion, subsampling instead of bootstrap sampling should be employed in ensemble methods to be able to reliably compare the variable importance scores of predictor variables of different types. The statistical properties and the null hypothesis of a test for the random forest variable importance are critically investigated. Finally, a new, conditional importance measure is suggested that allows for a fair comparison in the case of correlated predictor variables and better reflects the null hypothesis of interest.

Zusammenfassung

Die Anwendung von Methoden des rekursiven Partitionierens aus dem maschinellen Lernen ist in vielen Forschungsgebieten, wie z.B. in der Genetik und Bioinformatik, weit verbreitet. Die vorliegende Arbeit setzt sich aus statistischer Sicht mit den zwei Hauptproblemen des rekursiven Partitionierens, Instabilität und verzerrter Variablenselektion, auseinander. Im Hinblick auf das erste Thema, die Instabilität, wird das gesamte Methodenspektrum von herkömmlichen Klassifikationsbäumen über robustifizierte Klassifikationsbäume und Ensemble Methoden wie TWIX, Bagging und Random Forests in dieser Arbeit abgedeckt. Ensemble Methoden erweisen sich im Vergleich zu einzelnen Klassifikationsbäumen als deutlich stabiler, verlieren aber auch größtenteils ihre Interpretierbarkeit. Deshalb wird ein adaptives Bruchpunkt-Selektionskriterium vorgeschlagen, mit dem ein TWIX-Ensemble auf einen einzelnen Klassifikationsbaum reduziert wird, falls die Partition stabil genug ist. Im Hinblick auf das zweite Thema, die verzerrte Variablenselektion, werden die statistischen Ursachen für dieses Artefakt in einzelnen Bäumen und eine neue Form von Verzerrung, die in Ensemble Methoden auftritt die auf Bootstrap-Stichproben beruhen, untersucht. Für einzelne Bäume wird ein unverzerrtes Selektionskriterien evaluiert und ein anderes hier neu eingeführt. Anhand der Ergebnisse für einzelne Bäume und weiteren Untersuchungen zu den Auswirkungen von Bootstrap-Stichprobenverfahren auf Assoziationsmaße wird gezeigt dass, neben der Verwendung von unverzerrten Selektionskriterien, Teilstichprobenverfahren anstelle von Bootstrap-Stichprobenverfahren in Ensemble Methoden verwendet werden sollten, um die Variable Importance-Werte von Prädiktorvariablen unterschiedlicher Art zuverlässig vergleichen zu können. Die statistischen Eigenschaften und die Nullhypothese eines Test für die Variable Importance von Random Forests werden kritisch untersucht. Abschliessend wird eine neue, bedingte Variable Importance vorgeschlagen, die im Fall von korrelierten Prädiktorvariablen einen fairen Vergleich erlaubt und die interessierende Nullhypothese besser widerspiegelt.

Contents

Scope of this work

This work is concerned with a selection of statistical methods based on the principle of recursive partitioning: classification and regression trees (termed classification trees in the following for brevity, while most results apply straightforwardly to regression trees), robust classification trees and ensemble methods based on classification trees.

From a practical point of view these methods have become extremely popular in many applied sciences, including genetics and bioinformatics, epidemiology, medicine in general, psychiatry, psychology and economics, within a short period of time – primarily because they "work so well". From a statistical point of view, on the other hand, recursive partitioning methods are rather unusual in many respects; for example they do not rely on any parametric distribution assumptions.

Leo Breiman, one of the most influential researchers in this field, has promoted "algorithmic models" like classification trees and ensembles methods in the late years of his career after he had left academia to work as a consultant and made the experience that current statistical practice has "Led to irrelevant theory and questionable scientific conclusions; Kept statisticians from using more suitable algorithmic models; Prevented statisticians from working on exciting new problems" (Breiman, 2001b, pp. 199–200).

Today, the scientific discussion about the legitimacy of algorithmic models in statistics continues, as illustrated by the contribution of Hand (2006) in *Statistical Science* with the provocative title "Classifier Technology and the Illusion of Progress" and the multitude of comments that were triggered by it. Of these comments, the most consensual one may be the reply of Jerome Friedman, another highly influential researcher in the field of statistical

learning, who states: "Whether or not a new method represents important progress is, at least initially, a value judgement upon which people can agree or disagree. Initial hype can be misleading and only with the passage of time can such controversies be resolved. It may well be too soon to draw conclusions concerning the precise value of recent developments, but to conclude that they represent very little progress is at best premature and, in my view, contrary to present evidence" (Friedman, 2006, p. 18).

The "evidence" that Friedman refers to can be found in several benchmark studies showing that the ensemble methods bagging and random forests, that are considered here, together with other computerintensive methods like boosting (Freund and Schapire, 1997) and support vector machines (Vapnik, 1995), belong to the top performing statistical learning tools that are currently available (Wu et al., 2003; Svetnik et al., 2004; Caruana and Niculescu-Mizil, 2006). They outperform traditional statistical modelling techniques in many situations – and in some situations traditional techniques may not even be applicable, as in the case of "small n large p" problems that arise, e.g., in genomics when the expression level of a multitude of genes is measured for only a handful of subjects. Another advantage of these methods, as compared to other recent approaches that can be applied to "small n large p" problems such as the LASSO (cf., e.g., Hastie et al., 2001), the elastic net (Zou and Hastie, 2005), and the recent approach of Candes and Tao (2007), is that no linearity or additivity assumptions have to be made.

Still, many statisticians feel uncomfortable with any method that offers no analytical way to describe beyond intuition why exactly it "works so well". In the meantime, Bühlmann and Yu (2002) have given a rather thorough statistical explanation of bagging, and Lin and Jeon (2006) have explored the properties of random forests by placing them in an adaptive nearest neighbors framework. However, both approaches are based on several simplifying assumptions (for example, linear models are partly used as base learners instead of classification trees in Bühlmann and Yu, 2002), that limit the generalizability of the results to the methods that are actually implemented and used by applied scientists.

In addition to these analytical approaches, several empirical studies have been conducted

to try to help our understanding of the functionality of algorithmic models. Most of these studies are based only on a few, real data sets that happen to be freely available in some machine learning repository. It is important to note, however, that these data sets are not a representative sample from the range of possible problems that the methods might be applied to, and that their characteristics are unknown and not testable (for example assumptions on the missing value generating mechanism). Therefore any conclusions drawn from this kind of empirical study may not be reliable.

A very prominent example for a premature conclusion resulting from this kind of research is the study referred to in Breiman (2001b), where it is stated (and has been extensively cited ever since) that random forests do not overfit. This statement – and especially the fact that it is based on a selection of a few real data sets with very particular features, that enhance the impression that random forests would not overfit – is heavily criticized by Segal (2004).

As opposed to such methodological "case studies", here we want to rely on analytical results as far as possible (that are available, e.g., for the optimally selected statistics and unbiased entropy estimates suggested as split selection criteria in some of the following chapters). When analytical results are impossible to derive for the actually used method (as in the case of ensemble methods based on classification trees), however, we follow the rationale that valid conclusions can only be drawn from well designed and controlled experiments – as in any empirical science.

Only such controlled simulation experiments allow us to test our hypotheses about the functionality of a method, because only in a controlled experiment do we know what is "the truth" and what is "supposed to happen" in each condition. Therefore, throughout the course of this work, analytical results will be presented in the early sections where feasible, while well planned simulation experiments will be applied in the later sections, where the functionality of complex ensemble methods is investigated and improved by promoting an alternative resampling scheme and suggesting a new measure for reliably assessing the importance of predictor variables.

As illustrated in the chart at the end of this section, the outline of this work follows two major issues, that have been shown to affect reliable prediction and interpretability in classification trees and their successor methods: instability and biased variable selection.

When focusing on variable selection we will see that in the standard implementations, variable selection in classification trees is unreliable in that predictor variables of certain types are preferred regardless of their information content. The reasons for this artefact are very fundamental statistical issues: biased estimation and multiple testing, as outlined in Chapter 2. In single classification trees these issues can be solved by means of adequate split selection criteria, that account for the sample differences in the size and the number of candidate cutpoints. The evaluation of such a split selection criterion is demonstrated in Chapter 3.

However, when the concepts inherent in classification trees are carried forward to robust classification trees or ensembles of classification trees, deficiencies in variable selection are carried forward, too, and new ones may arise. For robust classification trees this is illustrated, and an unbiased criterion is presented in Chapter 4.

From Chapter 5 we will focus on the second issue of instability, that can be addressed by means of robustifying the tree building process or by constructing different kinds of ensembles of classification trees. When abandoning the well interpretable single tree models for the more stable and thus better performing ensembles of trees, there is always a tradeoff between stability and performance on one hand and interpretability on the other hand.

A lack of interpretability can crucially affect the popularity of a method. The steep rise of some of the early so-called "black box" learners, such as neural networks (first introduced in the 1980s; cf, e.g., Ripley, 1996, for an introduction), seems to have been followed by a creeping recession – mainly because their decisions are not communicable, for example, to a customer whose application for a loan is rejected because some algorithms classifies him as "high risk".

As opposed to that, single classification trees owe part of their popularity to the fact

that the effect of each predictor variable can easily be read from the tree graph. Still, the interpretation of the effect might be severely wrong because the tree structure is so instable: due to the recursive construction and cutpoint selection, small changes in the learning sample can lead to a completely different tree. Ensembles of classification trees on the other hand are not directly interpretable, because the individual tree models are not nested in any way and thus cannot be integrated to one common presentable model.

In this tradeoff between stability and interpretability, it would be nice if the user himself could regulate the degree of stability he needs – and give up interpretability no more than necessary. This idea is followed in a fundamental modification of the TWIX ensemble method in Chapter 5: An ensemble is created only if necessary and reduces to a single tree if the partition is stable.

In situations where the partition really is instable, however, the other ensemble methods bagging and random forests usually outperform the TWIX method, because they not only manage to smooth instable decisions of the individual classification trees by means of averaging, but also additional variation is introduced by means of randomization, that promotes locally suboptimal but potentially globally beneficial splits. In addition to that – and as opposed to complete "black box" learners and dimension reduction techniques – they provide variable importance measures that have been acknowledged as valuable tools in many applied sciences, headed by genetics and bioinformatics where random forest variable importance measures are used, e.g., for screening large amounts of genes for candidates that are associated with a certain disease.

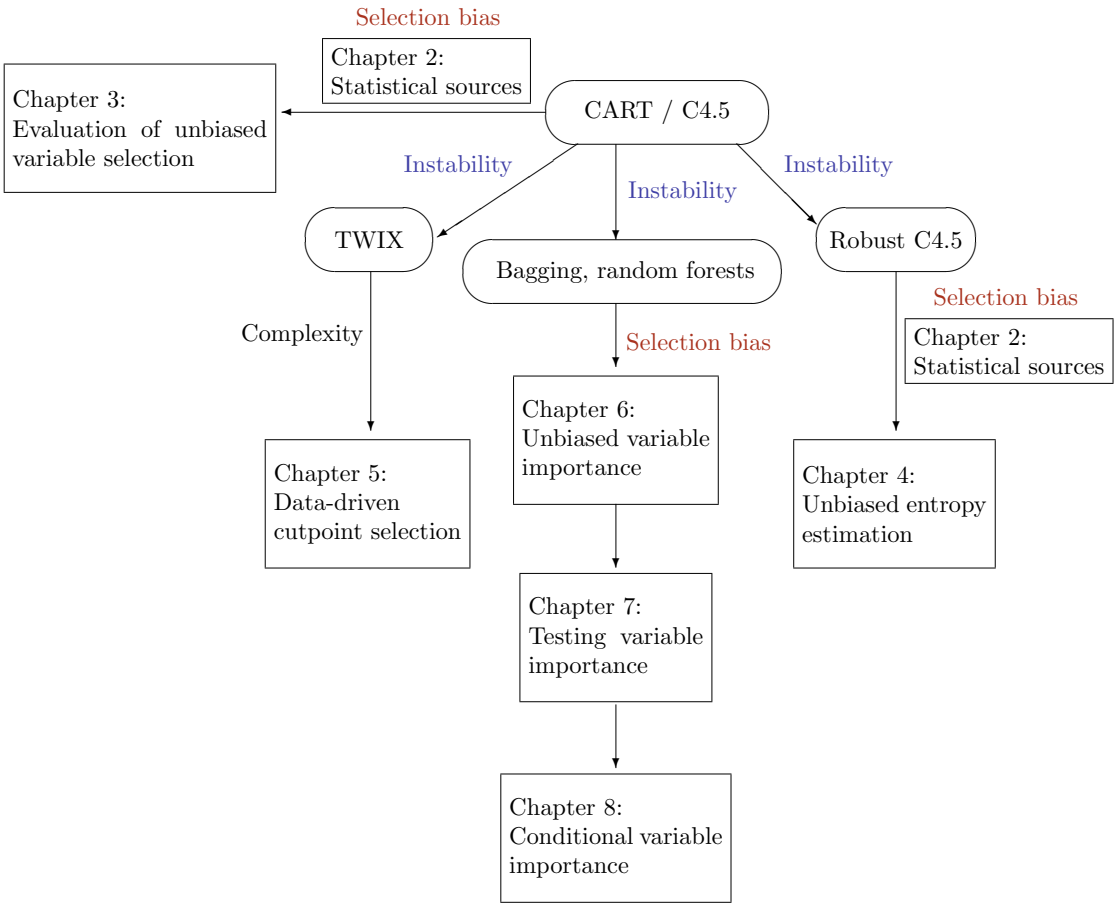
In such applications it is essential that variable importance measures are reliable. However, there are at least two situations where the originally proposed methods show undesired artifacts: the case of predictor variables of different types and the case of correlated predictor variables. In Chapter 6, a different resampling scheme is suggested to be used in combination with unbiased split selection criteria to guarantee that the variable importance is comparable for predictor variables of different types. The unbiased importance measures can then provide a fair means of comparison to decide which predictor variables are

most important and should be explored in further analysis. Additional variable selection schemes and tests for the variable importance have been suggested to aid this decision. The statistical properties of such a significance test are explored in Chapter 7.

Another aspect, that becomes relevant in the case of correlated predictor variables, as common in practical applications, is the distinction between marginal and conditional importance, that correspond to different null hypotheses. In Chapter 8 this distinction is facilitated and a new, conditional variable importance is suggested that allows for a fair comparison in the case of correlated predictor variables and better reflects the null hypothesis of interest. The theoretical reasoning and results presented in this chapter show that, only when the impact of each variable is considered conditionally on their covariates, it is possible to identify those predictor variables that are truly most important. Thus, the conditional importance forms a substantial improvement for applications of random forest variable importance measures in many scientific areas including genetics and bioinformatics, where algorithmic methods have effectively gained ground already, as well as new areas of application such as the empirical social and business sciences, for which some first applications are outlined in Chapter 1.

Parts of the work presented here are based on publications that were prepared in cooperation with coauthors named in the following:

Chapters	References
parts of 1	Strobl, Malley, and Tutz (2008) and
	Strobl, Boulesteix, Zeileis, and Hothorn (2007)
parts of 2 and 3	Strobl, Boulesteix, and Augustin (2007)
4	Strobl (2005)
parts of 5	Strobl and Augustin (2008)
6	Strobl, Boulesteix, Zeileis, and Hothorn (2007)
7	Strobl and Zeileis (2008)
8	Strobl, Boulesteix, Kneib, Augustin, and Zeileis (2008)


Selection bias
Chapter 2:
Statistical sources
Chapter 3:
Evaluation of unbiased
variable selection
CART / C4.5
Instability
Instability
Instability
TWIX
Bagging, random forests
Robust C4.5
Selection bias
Chapter 2:
Statistical sources
Complexity
Selection bias
Chapter 6:
Unbiased variable
importance
Chapter 5:
Data-driven
cutpoint selection
Chapter 4:
Unbiased entropy
estimation
Chapter 7:
Testing variable
importance
Chapter 8:
Conditional variable
importance

1. Introduction

After the early seminal work on automated interaction detection by Morgan and Sonquist (1963) the two most popular classification and regression tree algorithms were introduced by Breiman et al. (1984) and independently by Quinlan (1986, 1993). Their non-parametric approach and the straightforward interpretability of the results have added much to the popularity of classification trees, for example for psychiatric diagnoses from clinical or genetic data or for the prediction of therapy outcome (cf., e.g., Hannöver et al., 2002, for an application modelling the treatment effect in patients with eating disorders).

As an advancement of single classification trees, random forests (Breiman, 2001a), as well as its predecessor method bagging (Breiman, 1996a, 1998), are so-called "ensemble methods", where an ensemble (or committee) of classification and regression trees are aggregated for prediction. Ensemble methods show a high predictive performance and are applicable even in situations when there are many predictor variables. The individual classification or regression trees of an ensemble are built on bootstrap samples drawn from the original sample. Random forests take an important additional step, in that a subset of predictor variables is randomly preselected before each split. The next splitting variable is then selected only from the preselected subset. This additional randomization step has been shown to increase the predictive performance of random forests and enhances their applicability in situations when there are many predictor variables. In the following, some exemplary applications of ensemble methods – including the exploration of such high dimensional data sets – are outlined, before we return to take a closer look at the construction of classification trees and ensemble methods.

High dimensional problems, as well as problems involving correlated predictor variables and high-order interactions, are common in many scientific fields. As one important example, in genome studies often a very high number of genetic markers or SNPs (single nucleotide polymorphisms) are available, but only for a small number of subjects. Applications of random forests in genetics and bioinformatics include large-scale association studies for complex genetic diseases as in Lunetta et al. (2004) and Bureau et al. (2005), who detect SNP-SNP interactions in the case-control context by means of computing a random forest variable importance measure for each polymorphism. A comparison of the performance of random forests and other classification methods for the analysis of gene expression data is presented by Diaz-Uriarte and Alvarez de Andrés (2006), who propose a new gene selection method based on random forests for sample classification with microarray data. More applications of the random forest methodology to microarray data can be found in, e.g., Gunther et al. (2003), Huang et al. (2005) and Shih et al. (2005).

Prediction of phenotypes based on amino acid or DNA sequence is another important area of application of random forests, since possibly involving many interactions. For example, Segal et al. (2004) use random forests to predict the replication capacity of viruses, such as HIV-1, based on amino acid sequence from reverse transcriptase and protease. Cummings and Segal (2004) link the rifampin resistance in *Mycobacterium tuberculosis* to a few amino acid positions in rpoB, whereas Cummings and Myers (2004) predict C-to-U edited sites in plant mitochondrial RNA based on sequence regions flanking edited sites and a few other (continuous) parameters.

The random forest approach was shown to outperform six other methods in the prediction of protein interactions based on various biological features such as gene expression, gene ontology (GO) features and sequence data (Qi et al., 2006). Other applications of random forests can be found in fields as different as quantitative structure-activity relationship (QSAR) modeling (Guha and Jurs, 2003; Svetnik et al., 2003), nuclear magnetic resonance spectroscopy (Arun and Langmead, 2006), landscape epidemiology (Furlanello et al., 2003) and medicine in general (Ward et al., 2006).

Meanwhile, a few first applications of random forests in psychology have appeared, using the method for prediction or to obtain variable importance measures for selecting relevant predictor variables. For example, Oh et al. (2003) use random forests to measure the importance of the single components of neuronal ensemble spike trains collected from arrays of electrodes located in the motor and premotor cortex of a rat performing a reaction-time task. The advantages of random forests in this application are (i) that they can be easily applied to high dimensional and redundant data and (ii) as distinct from familiar dimension reduction methods such as principle components or factor analysis, in random forests the original input variables are not projected into a different set of components, so that the features selected are still identifiable and their importance is directly interpretable.

Other examples of applying random forests as a means for identifying relevant predictor variables in psychological and psychiatric studies are Rossi et al. (2005), who aim at identifying determinants of once-only contact in community mental health service, and Baca-Garcia et al. (2007), who employ random forests to identify variables associated with attempted suicide under consideration of the family history. Rossi et al. (2005) use random forest variable importance measures to support the stepwise variable selection approaches of logistic regression, that are known to be instable due to order effects. Baca-Garcia et al. (2007), despite a methodological weakness, combine the results of forward selection and random forests to identify the two predictor variables with the strongest impact on family history of attempted suicide and build a classification model with a high prediction accuracy.

In an application to the diagnosis of posttraumatic stress disorder (PTSD) Marinic et al. (2007) build several random forest models for predicting PTSD from structured psychiatric interviews, psychiatric scales or combinations of both. Different weightings of the response classes (PTSD or no PTSD) can be compared by means of random forests with respect to overall prediction accuracy, sensitivity and specificity. As pointed out by these authors, another advantage of random forests is that they generate realistic estimates of the prediction accuracy on a test set, as outlined below.

Luellen et al. (2005) point out another field of application in comparing the effects in an experimental and a quasi-experimental study on mathematics and vocabulary performance. Instead of predicting the actual response variable by means of classification trees and bagging, the methods are used here for estimating propensity scores: When the treatment assignment is chosen as a working response, classification trees and ensemble methods can be used to estimate the probability to be treated from the covariates, which can be used for stratification in the further analysis. The results of Luellen et al. (2005), even though somewhat inconsistent, indicate that bagging is well suited for propensity score estimation, and it is to be expected that there is even room for improvements that could be achieved by means of random forests.

These first applications of bagging and random forests in psychology point out several new potential areas of application in this field. In some applications random forests can add to the results or may even be preferable to standard methods. For example, their nonparametric approach does not require the specification of a sampling distribution or a certain functional form. In other applications, especially in high dimensional problems, or problems where the predictor variables are highly correlated or even subject to linear constraints, standard approaches such as logistic regression are simply not applicable and random forests provide a good alternative. On the other hand, random forests were not developed in a standard statistical framework so that their behavior is less predictable than that of standard parametric methods and some parts of random forests are still "under construction" (cf. also Polikar, 2006, for a brief history of ensemble methods, including fuzzy and Bayesian approaches).

The next section introduces the main concepts of classification trees, that are employed as the underlying so-called "base learners" in all following ensemble methods. The different ensemble methods themselves, that will be treated in detail in later chapters, are only shortly sketched in Section 1.2. Section 1.3 gives an overview over important features and advantages of classification trees and ensemble methods, as well as important caveats.

1.1 Classification trees

Classification and regression trees are a simple nonparametric method that recursively partitions the feature space into a set of rectangular areas and predicts a constant value within each area. Such a partition is illustrated in Figure 1.1. Here the first split is conducted in variable X_2 at cutpoint value 5. The left and right daughter nodes are then defined by all observations i with $x_{i2} \leq 5$ and $x_{i2} > 5$ respectively. Within the left daughter node the observations are again split up at cutpoint value 2.5 in variable X_1, so that all observations with $x_{i1} \leq 2.5$ proceed to the left daughter node and so forth. Note that it is possible to split again in the same variable. The splitting variable and cutpoint are chosen such as to reduce an impurity criterion as outlined in the following.

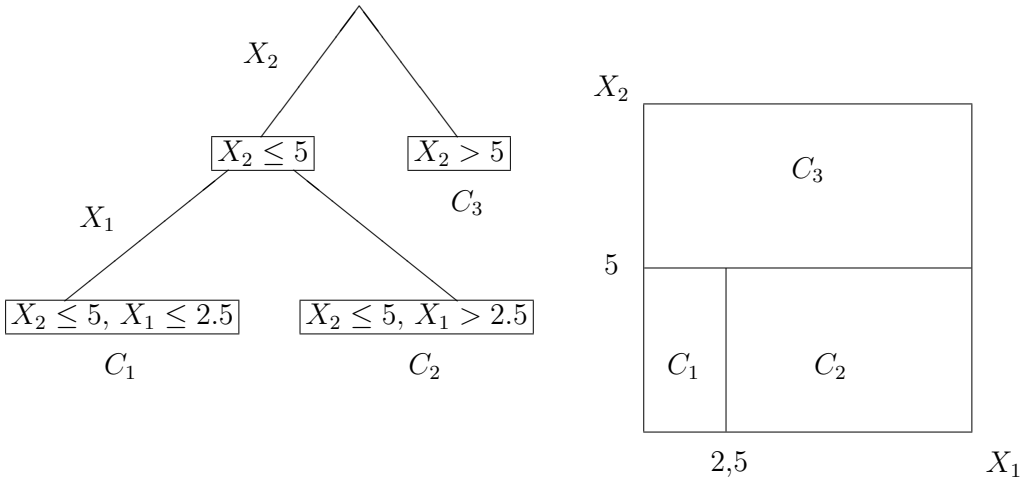


Fig. 1.1: Partition of a two dimensional feature space by means of a binary classification tree.

1.1.1 Split selection and stopping rules

Both the CART algorithm of Breiman et al. (1984) and the C4.5 algorithm (and its predecessor ID3) of Quinlan (1986, 1993) conduct binary splits in continuous predictor variables, as depicted in Figure 1.1. In categorical predictor variables (of nominal or ordinal scale

of measurement) C4.5 produces as many nodes as there are categories (often referred to as "k-ary" or "multiple" splitting), while CART again creates binary splits between the ordered or unordered categories.

For selecting the splitting variable and cutpoint in binary splitting, both CART and C4.5 follow the approach of impurity reduction (where the term "impurity" is used synonymously to the term "entropy" in the information technological sense) and use impurity criteria, such as the Gini index or the Shannon entropy or deviance, for variable and cutpoint selection: The impurity reduction that can be achieved by splitting a variable in a particular cutpoint into a left and right daughter node is computed for each variable and each cutpoint as the difference between the impurity before and after splitting. The predictor variable that, when split in its best cutpoint, produces the highest impurity reduction is then selected for splitting.

In every step of the recursive partitioning algorithm, this strategy can be expressed as a twofold optimization problem: From a response variable Y (that is considered to be categorical with categories $c \in \mathscr{C}$, including the easiest case of a binary response with $\mathscr{C} = \{1, 2\}$, throughout most of this work) and predictor variables $X_1, \ldots, X_p$ (of potentially different scales of measurement), a sample of n independent and identically distributed observations is used as a learning sample for tree construction.

For a starting node $\mathbf{C}$ and candidate daughter nodes $\mathbf{C}_{L,t_j}$ and $\mathbf{C}_{R,t_j}$ created by splitting a candidate variable X_j in cutpoint t_j, the steps are:

– Select the best cutpoint t_j^* within the range of predictor variable X_j with respect to the empirical impurity reduction $\widehat{\Delta\mathfrak{I}}$ (note that, throughout this work, empirical quantities will be denoted as estimators of the respective theoretical quantities by adding a hat to the symbol, because this notation facilitates our argumentation in Chapter 2):

$$t_j^* = \arg\max_{t_j} \widehat{\Delta\mathfrak{I}}\left(\mathbf{C}, \mathbf{C}_{L,t_j}, \mathbf{C}_{R,t_j}\right), \ \forall j = 1, \ldots, p. \tag{1.1}$$

– Out of all candidate variables choose the variable X_{j^*} that produces the highest impurity reduction in its best cutpoint t_j^*, i.e. consider X_{j^*} with

$$j^* = \arg\max_j \left\{ \widehat{\Delta\mathfrak{I}} \left(\mathbf{C}, \mathbf{C}_{L,t_j^*}, \mathbf{C}_{R,t_j^*} \right) \right\}. \tag{1.2}$$

The impurity reduction achieved by splitting in a candidate cutpoint t_j of a variable X_j is computed as the difference between the impurity in the starting node before splitting minus the weighted mean over the daughter node impurities after splitting

$$\widehat{\Delta\mathfrak{I}} \left(\mathbf{C}, \mathbf{C}_{L,t_j}, \mathbf{C}_{R,t_j} \right) = \mathfrak{I}(\mathbf{C}) - \left(\frac{n_{L,t_j}}{n} \mathfrak{I}(\mathbf{C}_{L,t_j}) + \frac{n_{R,t_j}}{n} \mathfrak{I}(\mathbf{C}_{R,t_j}) \right), \tag{1.3}$$

where n_{L,t_j} is the number of observations in $\mathbf{C}$ that are assigned to the left node and n_{R,t_j} to the right node, respectively. Note that the notation used here is limited to the first split of a classification tree, because this is sufficient to illustrate most arguments in the current and following chapters. However, the same principles apply to all subsequent splits and additional splits in the same variable, even though they are not covered by the notation so far.

Popular criteria that can be employed as the empirical impurity measure $\widehat{\mathfrak{I}}$ are the empirical Gini index $\widehat{G}$ used in CART and the empirical Shannon entropy $\widehat{S}$ used in C4.5. For the easiest case of two response classes the empirical Gini index (Breiman et al., 1984) for the starting node reduces to

$$\widehat{G}(\mathbf{C}) = 2\hat{\pi}(1 - \hat{\pi}), \tag{1.4}$$

where $\hat{\pi} = \frac{n_2}{n}$ is the relative frequency of response class $Y = 2$ within the node (the notation is, of course, exchangeable with respect to the two response classes), and the empirical Shannon entropy (Shannon, 1948) is

$$\widehat{S}(\mathbf{C}) = -\left\{ \hat{\pi} \log \hat{\pi} + (1 - \hat{\pi}) \log(1 - \hat{\pi}) \right\}. \tag{1.5}$$

Both functions have basically the same shape so that pure nodes, containing only observations of one class, have impurity zero and nodes with equal frequencies of observations

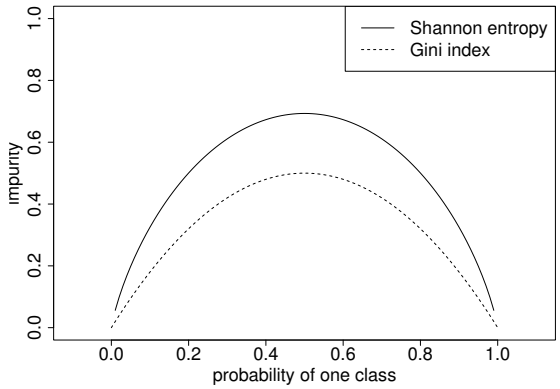


Fig. 1.2: Gini index and Shannon entropy as impurity functions for the two class case.

for each class have maximum impurity or entropy as illustrated in Figure 1.2.

In principle, any kind of criterion or statistic measuring the association between the predictor variable and the response (such as the χ^2-statistic or its p-value) can be used for split selection instead of the traditional impurity reduction approach. However, association statistics such as the χ^2-statistic can only be directly compared when the underlying degrees of freedom are equal (i.e., for contingency tables with equal dimensions or predictor variables with equal numbers of categories in recursive partitioning). When, on the other hand, p-values are used as split selection criteria, that account for different degrees of freedom of the underlying statistics, it is still important to adjust for the fact that each cutpoint t_j^* is chosen such as to maximize the association statistics. The more recent approach based on the p-values of optimally selected statistics treated in Chapter 3, for example, successfully addresses this issue. Note, however, that neither the traditional impurity reduction criteria nor the modern p-value based split selection approaches are designed to optimize the overall model fit or misclassification error of the final model. All recursive partitioning algorithms trade in global optimality for computational feasibility, as discussed further below.

In binary recursive partitioning, potential cutpoints for ordered and continuous variables

lie between any two successive values (resulting in $n-1$ possible cutpoints for n distinct values of a continuous predictor variable without ties, or $k-1$ possible cutpoints for k ordered categories), while for categorical predictors of nominal scale of measurement any binary partition of the categories can be used to determine the left and right daughter node (resulting in $2^{k-1}-1$ possible cutpoints for k unordered categories). Each split is represented by a binary partition of the feature space and the same variable can be used more than once in each branch to allow for flexible models.

In k-ary splitting on the other hand, for each categorical variable as many new nodes as categories are produced, and thus the variable can only be used once in each branch. Technically speaking, every k-ary tree can be represented as a binary tree. In this case the k-ary representation (for some $k > 2$) results in a wider tree, while the binary representation results in a deeper tree. However, truly binary splitting trees are more sparse than k-ary splitting trees in that they only branch when the distribution of the response variable actually differs in the nodes. As opposed to that k-ary splitting always produces k nodes, even if the distribution of the response variable in some nodes is very similar.

Another feature of the split selection strategy of recursive partitioning is that it makes the treatment of continuous, metrically scaled variables "robust" in the sense that they are treated as ordered. Technically speaking, classification trees are also invariant under monotone transformations of the predictor variables. In particular the scaling of continuous variables is irrelevant in tree-based models unlike, for example, in neural networks.

After a split is conducted in the first splitting variable, the observations in the learning sample are divided into different nodes defined by the split, and in each node splitting continues recursively, as illustrated in Figure 1.1, until some stop condition is reached. Common stop criteria are: Split until (i) all leaf nodes are pure (i.e. contain only observations of one class) (ii) a given threshold for the minimum number of observations left in a node is reached or (iii) a given threshold for the minimum change in the impurity measure is not succeeded any more by any variable. Recent classification tree algorithms also provide statistical stopping criteria that incorporate the distribution of the splitting

criterion (Hothorn et al., 2006), while other algorithms rely on pruning the complete tree to avoid overfitting.

1.1.2 Prediction and interpretation

Finally a response class or value is predicted in each terminal node of the tree (or each rectangular section in the partition respectively) by means of deriving from all observations in node $\mathbf{C}$ either the average response value $\hat{y}_{\mathbf{C}} = \text{ave}\,(y_i|\mathbf{x}_i \in \mathbf{C})$ in regression or the most frequent response class $\hat{y}_{\mathbf{C}} = \arg\max_{c\in\mathscr{C}} \left(\sum_i \mathrm{I}(y_i = c|\mathbf{x}_i \in \mathbf{C})\right)$ in classification trees. Note that this means that a regression tree creates a piecewise (or rectangle-wise for two dimensions as in Figure 1.1 and cuboid-wise in higher dimensions) constant prediction function.

We will see later that ensemble methods, by combining the predictions of many single trees, can approximate functions more smoothly. For classification problems it is also possible to predict an estimate of the class probabilities from the relative frequencies of each class in the terminal nodes. This kind of prediction more closely resembles the output of logistic regression models and can also be employed for estimating propensity scores as indicated in the introduction. The quality of probability estimates derived from random forests, both in comparison to logistic regression in problems where both methods are applicable and in high dimensional problems where logistic regression may not be applicable, is currently under research.

For the interpretation of a completed tree, prediction rules can be found by following down each branch and producing simple verbal interpretations such as "students that scored less than 50 points on a previous test and have a low motivation are likely to fail the final exam, while those that scored less than 50 points but have a high motivation are likely to pass". This easy interpretability has added much to the popularity of classification trees especially in the social and health sciences, where it is important, e.g., for both the clinician and the patient that the biological argument reflected by a model can be well

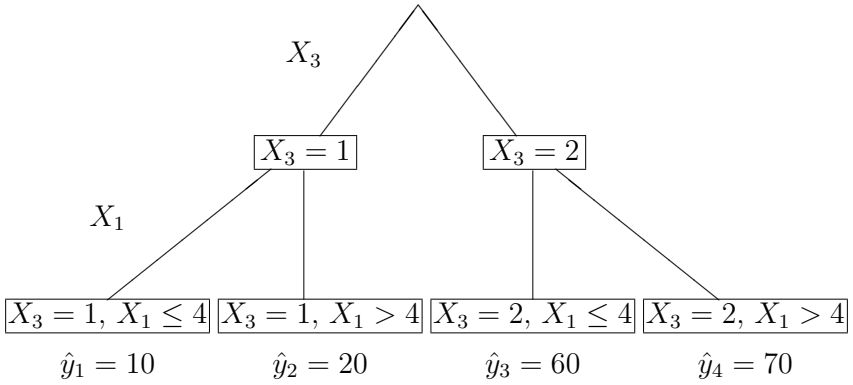


Fig. 1.3: Regression tree with two main effects.

understood. On the other hand this kind of visual interpretability might be tempting or even misguiding, because the actual statistical interpretation of a tree model is not entirely trivial. Especially the notions of main effects and interactions are often used rather incautiously in the literature, as seems to be the case, e.g., in Berk (2006): On p. 272 it is stated that a branch that is not split any further indicated a main effect. However, when in the other branch created by the same variable splitting continues, as is the case in the example of Berk (2006), this statement is not correct.

The term "interaction" commonly describes the fact that the effect of one predictor variable, say X_1, on the response variable Y depends on the value of another predictor variables, say X_3. For classification and regression trees this means that, if in one branch created by X_3 it is not necessary to split in X_1, while in the other branch created by X_3 it is necessary, an interaction between X_1 and X_3 is present. We will illustrate this important issue and source of misinterpretations by means of stylized regression trees given in Figures 1.3 through 1.5.

Only Figure 1.3, where the effect of X_1 is the same in both branches created by X_3, represents two main effects of X_1 and X_3 without an interaction. Both Figures 1.4 and 1.5

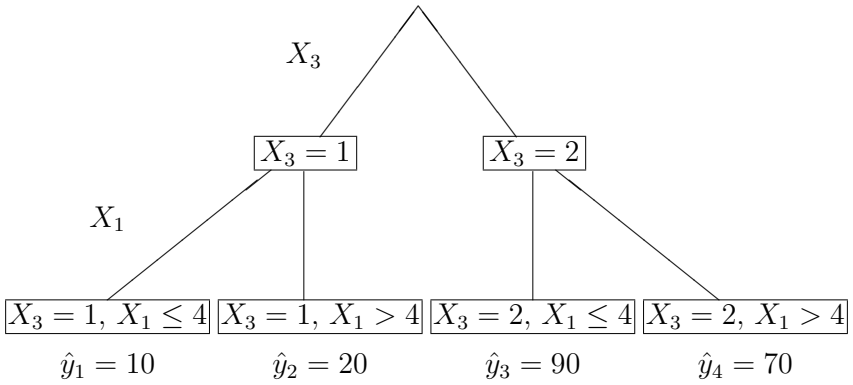


Fig. 1.4: Regression tree with an interaction.

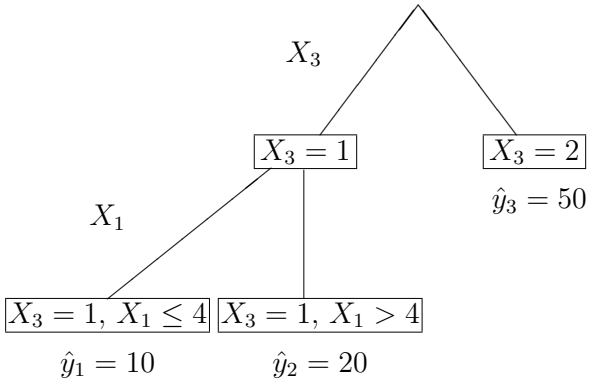


Fig. 1.5: Regression tree with an interaction.

represent interactions, because the effect of X_1 is different in both branches created by X_3. In Figure 1.4 the same split in X_1 is conducted in every branch and only the effect on the predicted response is different in both branches created by X_3. In Figure 1.5 on the other hand the effect of X_1 is different in both branches created by X_3: X_1 does have an effect in the left branch but it does not have an effect in the right branch.

However, in trees built on real data, it is extremely unlikely to actually discover a pattern as that in Figure 1.3. The reason is that, even if the true distribution of the data in both branches created by X_3 was very similar, due to random variations in the sample and the deterministic cutpoint selection strategy of classification trees it is extremely unlikely that the exact same cutpoint would be found in both partitions. Even a different cutpoint in the same variable would, however, strictly speaking represent an interaction. Therefore it is stated in the literature that classification trees cannot (or rather, are extremely unlikely to) represent additive functions that consist only of main effects, while they are perfectly well suited for representing multiplicative functions that consist of interactions. This implies that, if it is known from subject matter that the underlying problem can only be additive, recursive partitioning methods are not a good choice.

If, on the other hand, one suspects that the problem contains interactions of possibly high order, classification trees are more flexible than parametric models, where interactions of order higher than two can hardly ever be considered. However, in principle any decision boundary, including linear ones, can be approximated by a tree given enough data.

1.1.3 Variable selection bias and instability

In the following we now want to treat two statistical issues that have not only caused serious problems in the application of classification trees but have led to important insights and advancements of the method: biased variable selection on one hand and instability due to deterministic splitting on the other hand. We will follow and revisit several aspects of these two issues throughout this work, and provide a deeper statistical understanding as

well as solutions for theoretical and practical problems that arise from them.

The term "variable selection bias" describes the fact that the standard classification tree algorithms are known to artificially prefer variables with many categories or many missing values (cf., e.g., White and Liu, 1994; Kim and Loh, 2001). The sources of this bias are multiple testing effects in binary splitting and an estimation bias of empirical entropy measures, such as the Gini index or the Shannon entropy, as will be illustrated in detail in Chapter 2. We will see later that this kind of bias can also affect variable selection in ensemble methods.

There are different approaches to eliminate variable selection bias: For k-ary splitting Dobra and Gehrke (2001) introduce an unbiased p-value criterion based on the Gini index for split selection, while for binary splitting it is necessary to account for multiple testing as well. This is conducted, e.g., by means of the p-value criterion based on the optimally selected Gini gain introduced by Boulesteix in Strobl et al. (2007), for which an evaluation study is conducted in Chapter 3.

A different approach to eliminate variable selection bias in either case is to separate the issue of variable selection from the cutpoint selection procedure, as proposed by Loh and Shih (1997). This can be conducted by first selecting the next splitting variable by means of some association test, and then selecting the best cutpoint within the chosen predictor variable. In their technically advanced approach Hothorn et al. (2006) introduce an unbiased tree algorithm based on conditional inference tests that provides p-values as split selection criteria for predictor and response variables of any scale of measurement. Here the p-values can serve not only as split selection criteria but also as a stop criteria. An implementation of random forests based on this approach forms the basis for some of our later simulation studies in Chapters 6 through 8.

The other flaw of the standard classification trees is their instability to small changes in the learning data: In binary splitting algorithms the best cutpoint within one predictor variable determines both which variable is chosen for splitting, and how the observations

are split up in two new nodes – in which splitting continues recursively. Thus, as an undesired side effect, the entire tree structure could be altered if the first cutpoint was chosen differently and one can imagine that the tendency to meticulously adapt to small changes in the learning data can lead to severe changes in the tree structure and even overfitting when trees are grown extensively.

The term overfitting refers to the fact that a classifier that adapts too closely to the learning sample will not only discover the systematic components of the structure that is present in the population, but also the random variation from this structure that is present in the learning data due to random sampling. When such an overfitted model is later applied to a new test sample from the same population, its performance will be poor because it does not generalize well. For a more thorough introduction on the issue of performance estimation based on different sampling and resampling schemes cf. Boulesteix et al. (2008).

The classic strategy to cope with overfitting in recursive partitioning is to prune the classification trees after growing them, which means that branches that do not add to the prediction accuracy in cross validation are eliminated. Pruning is not discussed in detail here, because the unbiased classification tree algorithm of Hothorn et al. (2006), that is used in most parts of this work, employs p-values for variable selection and as a stopping criterion and therefore does not rely on pruning, and the robust classification tree approach of Abellán and Moral (2005) that forms the basis for Chapter 4 avoids overfitting by means of an upper entropy approach. Moreover, ensemble methods usually employ unpruned trees.

We will see in the next section that ensemble methods have been introduced to not only overcome but even utilize the instability of single trees as a source overfitting and therefore can achieve much better performance on test data.

1.2 Robust classification trees and ensemble methods

One possible extension of classification trees is that of credal classifiers based on imprecise probabilities by Abellán and Moral (2005), that is not as susceptible to overfitting as the original classification trees and thus provides more reliable results. Abellán and Moral (2005) employ a k-ary splitting approach inspired by Quinlan (1993). Variable selection is conducted with respect to an upper entropy criterion in this approach and is investigated with respect to variable selection bias in Chapter 4.

The ensemble methods bagging and random forests (Breiman, 1996a, 2001a) on the other hand, that will be described in more detail shortly, employ sets of classification trees and thus provide more stable predictions – but at the expense of completely giving up the interpretability of the single tree model. Therefore, variable importance measures for ensemble methods are discussed in Chapters 6 through 8.

The TWIX method, introduced by Potapov (2006) (see also Potapov et al., 2006; Potapov, 2007), that is the basis for the modification suggested in Chapter 5, resides somewhere in between single classification trees and fully parallel ensemble methods like bagging and random forests: It begins with a single starting node but branches to a set of trees at each decision by means of splitting not only in the best cutpoint but also in reasonable extra cutpoints. With respect to prediction accuracy, TWIX outperforms single trees and can even reach the performance of bagging and random forests on some data sets, but in general it cannot compete with them because it becomes computationally infeasible.

The rationale behind ensemble methods is that they use a whole set of classification trees rather than a single tree for prediction. The prediction of all trees in the set is combined by voting (in classification) or averaging (in regression). This approach leads to a significant increase in predictive performance on a test sample as compared to the performance of a single tree. TWIX shares this feature with the ensemble methods bagging and ran-

dom forests even though the sets of trees are created differently, as described in detail in Chapter 5.

In bagging and random forests this set of trees is built on random samples of the learning sample: In each step of the algorithm, a bootstrap sample or a subsample of the learning sample is drawn randomly, and an individual tree is grown on each sample. Each random sample reflects the same data generating process, but differs slightly from the original learning sample due to random variation. Keeping in mind that each individual classification tree depends highly on the learning sample as outlined above, the resulting trees can differ substantially. The prediction of the ensemble is then the average or vote over the single trees' prediction. The term "voting" can be taken literally here: Each subject with given values of the predictor variables is "dropped down" every tree in the ensemble. Each single tree returns a predicted class for the subject and the class that most trees "voted" for is returned as the prediction of the ensemble. This democratic voting process is the reason why ensemble methods are also called "committee" methods. Note, however, that there is no diagnostic for the unanimity of the vote. A summary over several aggregation schemes is given in Gatnar (2008).

By combining the prediction of a diverse set of trees bagging utilizes the fact that classification trees are instable but in average produce a good prediction, which has been supported by several empirical as well as simulation studies (cf., e.g., Breiman, 1996a, 1998; Bauer and Kohavi, 1999; Dietterich, 2000) and especially the theoretical results of Bühlmann and Yu (2002), that show the superiority in prediction accuracy of bagging over single classification or regression trees: Bühlmann and Yu (2002) conclude from their asymptotic results that the improvement in the prediction is achieved by means of smoothing the hard cut decision boundaries created by splitting in single classification trees, which in return reduces the variance of the prediction. The smoothing of hard decision boundaries also makes ensembles more flexible than single trees in approximating functional forms that are smooth rather than piecewise constant. Grandvalet (2004) also points out that the key effect of bagging is that it equalizes the influence of particular observations – which

is beneficial in the case of "bad" leverage points but may be harmful when "good" leverage points, that could improve the model fit, are downweighted. The same effect can be achieved not only by means of bootstrap sampling as in standard bagging, but also by means of subsampling (Grandvalet, 2004). Ensemble construction can also be viewed in the context of Bayesian model averaging (cf., e.g., Domingos, 1997; Hoeting et al., 1999, for an introduction). For random forests, Breiman (2001a) states that they may also be viewed as a Bayesian procedure (and continues: "Although I doubt that this is a fruitful line of exploration, if it could explain the bias reduction, I might become more of a Bayesian.").

In random forests another source of diversity is introduced when the set of predictor variables to select from is randomly restricted in each split, producing even more diverse trees. The number of randomly preselected splitting variables, as well as the overall number of trees, are parameters of random forests that affect the stability of their results. Obviously random forests include bagging as the special case where the number of randomly preselected splitting variables is equal to the overall number of variables.

Intuitively speaking random forests can improve the predictive performance even further with respect to bagging, because they employ even more diverse single trees in the ensemble: In addition to the smoothing of hard decision boundaries the random selection of splitting variables in random forests allows predictor variables that were otherwise outplayed by other predictors to enter the ensemble – which may reveal interaction effects that otherwise would have been missed.

To understand why such apparently suboptimal splits can improve the prediction accuracy of an ensemble, it is helpful to recall that the split selection process in regular classification trees is only locally optimal at each node: A variable and cutpoint are chosen with respect to the impurity reduction they can achieve in a given node defined by all previous splits, but regardless of all splits yet to come. This approach does not necessarily (or rather hardly ever) lead to the globally optimal tree over all possible combinations of cutpoints in all variables. However, searching for a globally optimal tree is computationally infeasible (a first approach involving dynamic programming was introduced by van Os and Meulman,

2005, but is currently limited to problems with very few categorical predictor variables). Randomization in ensemble construction has the side effect that a randomly chosen and locally suboptimal split may improve the global performance.

1.3 Characteristics and caveats of classification trees and ensemble methods

The way classification trees and ensembles are constructed induces some special characteristics of these methods that distinguish them from other (even other nonparametric) regression approaches.

1.3.1 "Small n large p" applicability

The fact that variable selection can be limited to random subsets in random forests make them particularly well applicable in "small n large p" problems with many more variables than observations, and has added much to the popularity of random forests. However, even if the set of candidate predictor variables is not restricted as in random forests, but covers all predictor variables as in bagging and single trees, the search is only a question of computational effort: Unlike logistic regression models, e.g., where parameter estimation is instable if not impossible when there are too many predictor variables and too few observations, tree-based methods only consider one predictor variable at a time and can thus deal with high numbers of variables sequentially. Therefore Bureau et al. (2005) and Heidema et al. (2006) point out that the recursive partitioning strategy is a clear advantage of random forests as opposed to more common methods like logistic regression. While other statistical methods directly include variable selection as part of the modeling process in linear or additive models, random forests can be used in a combined strategy to identify predictors relevant in potentially complex functions and then further explore

this smaller set of predictors with a simpler, for example linear, model if the prediction accuracy indicates that it is sufficient to reflect the underlying problem.

A restriction imposed by recursive partitioning is that in some situations a variable that is only relevant in an interaction might be missed out by the marginal sequential search strategy: The so-called "XOR problem" represents such a case, where two variables have no main effect but a perfect interaction effect. In this case none of the variables might be selected in the first split, and the interaction might never be discovered, due to the lack of a marginally detectable main effect. In a perfectly symmetric artificial "XOR problem", a tree would indeed not find a cutpoint to start with – but a logistic regression model would not be able to identify a main effect in any of the variables either. Only if the interaction is explicitly included in the logistic regression model it will be able to discover it – and in that case a tree model, where an interaction effect of two variables can also be explicitly added as a potential predictor variable, would do equally well. In addition to this, a tree, and even better an ensemble of trees, is able to approximate the "XOR problem" by means of a sequence of cutpoints driven by random fluctuations that are present in the learning data sets. In addition to this, the random preselection of splitting variables in random forests again increases the chance that a variable with a weak marginal effect is still selected, at least in some trees, because some of its competitors are not available.

A similar argument applies to order effects when comparing stepwise variable selection in regression models with the variable selection that can be conducted on the basis of random forest variable importance measures: In both stepwise variable selection and single trees order effects are present, because only one variable at a time is considered – in the context of the variables that were already selected but regardless of all variables yet to come. However, in ensemble methods, that employ several parallel tree models, the order effects of all individual trees counterbalance and the importance of a variable reflects its impact in different contexts.

1.3.2 Out-of-bag error estimation

Another key advantage of bagging and random forests over standard regression and classification approaches is that they come with their own "built-in" test sample for error estimation. In model validation when the (misclassification or mean squared) error is computed from the learning data, the estimation is far too optimistic (cf., e.g., Boulesteix et al., 2008). This is especially so for methods that tend to overfit, i.e., that adapt too closely to the learning data and thus do not generalize well to new test data.

The usual procedure when evaluating model performance is to build the model on learning data and evaluate it on a new test set, that was not used in model construction. Random forests and bagging on the other hand bring their own test set for every tree of the ensemble: Every tree is learned on a bootstrap sample (or subsample) of the original sample – and for each bootstrap sample (or subsample) there are some observations of the original sample that are not in it. These leftover observations are called "out-of-bag" (often abbreviated as "oob") observations, and can be used to correctly evaluate the predictive performance by measuring the misclassification error of each tree applied to the out-of-bag observations that were not used to build that tree (Breiman, 1996b).

Of course similar validation strategies, based either on sample splitting or resampling techniques (cf., e.g., Hothorn et al., 2005; Boulesteix et al., 2008), can and should be applied to any statistical method. König et al. (2007), for example, state that random forests can be considered to be "internally validated" but for other classification methods employ cross-validation for error estimation. However, in many disciplines intensive model validation is not common practice. Therefore a method that comes with a built-in test sample like random forests may help sensitize for the issue and relieve the user of the decision for an appropriate validation scheme.

1.3.3 Missing value handling

Tree based methods such as bagging and random forests come with an intuitive strategy for missing value handling that does not involve cancelation of observations with missing values as a whole, which would result in heavy data loss, or imputation.

In the variable selection step of the tree building process the so-called "available case" strategy is applied: Observations that have missing values in the variable that is currently evaluated are ignored in the computation of the impurity reduction for this variable, while the same observations are included in all other computations. However, we will show in Chapter 2 that this strategy can cause variable selection bias.

Another problem is that in the next step, after a splitting variable is selected, it would be unclear to what daughter node the observations that have a missing values in this variable should be assigned. To solve this problem a so-called "surrogate variable" is selected, that best predicts the values of the originally chosen splitting variable. By means of this surrogate variable the observations can then be assigned to the left or right daughter node (cf., e.g., Hastie et al., 2001). Another flaw of this approach is, however, that currently it is not clear how variable importance values can be computed for variables with missing values.

1.3.4 Randomness and stability

In random forests two sources of randomness are evident: The bootstrap samples (or subsamples) are randomly drawn and a random preselection of predictor variables is conducted. Due to these two random processes a random forest is only exactly reproducible when the random seed, determining the internal random number generation of the computer that is used for modelling, is fixed. Otherwise, the randomness involved will induce differences in the results. These differences are, however, negligible as long as the parameters of a random forest have been chosen such as to guarantee stable results:

– The number of trees highly affects the stability of the model. In general, the higher the number of trees the more reliable is the prediction and the interpretability of the variable importance.

– The number of randomly preselected predictor variables, termed `mtry` in most implementations of random forests, also affects the stability of the model, particularly the reliability of the variable importance: It can be chosen by means of cross validation, but it is often found in empirical studies (cf., e.g., Svetnik et al., 2003) that the default value `mtry`= $\sqrt{p}$ is optimal with respect to prediction accuracy. Our recent results displayed in Chapter 8, however, indicate that in the case of correlated predictor variables different values of `mtry` should be considered.

  Note that both parameters also interact: For a high number of predictor variables a high number of trees or a high number of preselected variables, or ideally both, are needed so that each variable has a chance to occur in enough trees. Only then its average variable importance measure is based on enough trials to actually reflect the importance of the variable and not just a random fluctuation.

  In summary this means: If one observes that, for a different random seed, the results for prediction and variable importance differ notably, one should not interpret the results but adjust the number of trees and preselected predictor variables.

– Another user defined parameter in building ensemble methods is the tree size. Most previous publications have argued that in an ensemble each individual tree should be grown as large as possible and that trees should not be pruned. However, the recent results of Lin and Jeon (2006) point out that creating large trees is not necessarily the optimal strategy: In problems with a high number of observations and few variables a better convergence rate (of the mean squared error as a measure of prediction accuracy) can be achieved when the terminal node size increases with the sample size (i.e. when smaller trees are grown for larger samples). On the other hand, for problems with small sample sizes or even "small n large p" problems growing large

trees often does lead to the best performance.

Besides these fundamental characteristics of recursive partitioning methods in general and ensemble methods in particular, we now address the first of the two issues that we will follow throughout this work: variable selection bias in individual classification trees. Later we will return to this issue and investigate implications and new sources of bias in ensemble methods.

2. Variable selection bias in binary and k-ary classification trees

The traditional recursive partitioning approaches use empirical impurity reduction measures, such as the Gini gain derived from the Gini index, 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. 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 k-ary splitting recursive partitioning.

Common 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 may 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 "available case" strategy is used in standard classification tree algorithms in the variable selection step. To investigate the effect of missing values in this setting, 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 in a simulation study. 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.

Here we want 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 in Kim and Loh (2001). Moreover, we want to address and clarify previous misperceptions of variable selection bias in the literature, that seem to be due to a lack of differentiation between binary and k-ary splitting and the mechanisms of variable selection

bias inherent in each setting.

For example, Jensen and Cohen (2000) misleadingly state that variable selection bias for categorical predictor variables with many categories was due to multiple comparisons when defining the left and right nodes of a classification tree, and explicitly cite the algorithm of Quinlan (1986) (the predecessor publication of Quinlan (1993), that describes the C4.5 algorithm) as an example. However, the algorithms of Quinlan perform k-ary splitting for categorical predictor variables, so that the intuition of a left and right node is not valid here. We will see later that the multiple testing argument does apply to binary splitting, but not to k-ary splitting, where the reasons for the preference for categorical variables with many categories are different.

Dobra and Gehrke (2001), on the other hand, do correctly accredit their findings of variable selection bias in a simulation study to the distribution of the split selection criterion (see below). However, they also explicitly state that variable selection bias with the Gini index, which was introduced by Breiman et al. (1984) and is usually associated with binary splitting, was not at all due to multiple testing. The reason for this is that they used the Gini index for k-ary splitting, where their argument is valid, while the literature they were citing referred to binary splitting, where their argument does not apply. By ignoring results for binary splitting Dobra and Gehrke (2001) missed the statistical aspects relevant for both k-ary and binary splitting explained below.

Kim and Loh (2001) themselves claim to have found a statistical explanation for the preference for variables with missing values, but as an explanation give only a special case that can easily be refuted. Finally Shih (2004) gives a sound statistical explanation, that, however, again only refers to the multiple testing problem in choosing the optimal cut-point in binary splitting, and can neither account for the bias in k-ary splitting, nor for the preference for variables with many missing values.

Therefore, in the following we provide a statistical explanation for variable selection bias in binary splitting with missing values and show that the same statistical source, but through

a very different mechanism, is responsible for variable selection bias in k-ary splitting.

2.1 Entropy estimation

The main source of variable selection bias is an estimation effect: The classical Gini index used in machine learning can be considered as an estimator of the true underlying entropy. The bias of this estimator – aggravated by its variance – induces variable selection bias.

We concentrate on the Gini index in the following sections, while the same principles hold for the Shannon entropy as illustrated in Chapter 4.

2.1.1 Binary splitting

We again consider a sample of n independent and identically distributed observations of a binary response Y and predictors $X_1, \ldots, X_p$, where the different $X_1, \ldots, X_p$ may have different numbers of missing values in the sample: For $j = 1, \ldots, p$, 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 $Y = 2$.

For illustrating the effects of biased entropy estimation in split selection in a situation with continuous predictor variables containing different numbers of missing values as in Kim and Loh (2001), let us slightly simplify the notation from Chapter 1: In binary splitting of continuous variables a cutpoint t_j can be any value $x_{(i)j}$ within the range of variable X_j. The index (i) here refers to the sample that is ordered with respect to X_j, so that a binary split in $x_{(i)j}$ discriminates between values smaller than (or equal to) and greater than $x_{(i)j}$, as illustrated in Table 2.1.1

Let $\mathbf{C}_j$, $j = 1, \ldots, p$, now denote the starting set for variable X_j: $\mathbf{C}_j$ holds the n_j observations for which the predictor variable X_j is not missing. The subsets $\mathbf{C}_{Lj}(i)$ and $\mathbf{C}_{Rj}(i)$ are produced by splitting $\mathbf{C}_j$ at a cutpoint between $x_{(i)j}$ and $x_{(i+1)j}$ in the sample ordered with respect to the values of X_j ($x_{(1)j} \leq \ldots \leq x_{(n_j)j}$): All observations with a value of $X_j \leq x_{(i)j}$ are assigned to $\mathbf{C}_{Lj}(i)$ and the remaining observations to $\mathbf{C}_{Rj}(i)$.

In Table 2.1.1, $n_{2j}(i)$, for example, 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,

$$n_{2j}(i) = \sum_{l=1}^{i} I_{\{2\}}(y_{(l)j}), \quad \forall i = 1, \ldots, n_j. \tag{2.1}$$

where $I_{\{2\}}(\cdot)$ is the indicator function for response $y_{(l)j} = 2$; $n_{1j}(i)$ is defined in an analogous way. 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 again for the sake of simplicity.

Tab. 2.1: Contingency table obtained by splitting the predictor variable X_j at $x_{(i)j}$.

	$\mathbf{C}_{Lj}(i)$	$\mathbf{C}_{Rj}(i)$	
	$X_j \leq x_{j(i)}$	$X_j > x_{j(i)}$	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

The empirical Gini index from Equation 1.4 can then be denoted as

$$\widehat{G}\left(\mathbf{C}_j\right) =: \widehat{G}_j = 2\frac{n_{2j}}{n_j}\left(1 - \frac{n_{2j}}{n_j}\right). \tag{2.2}$$

The corresponding empirical Gini Indices in the nodes produced by splitting at the i-th cutpoint, $\widehat{G}(\mathbf{C}_{Lj}(i)) =: \widehat{G}_{Lj}(i)$ and $\widehat{G}(\mathbf{C}_{Rj}(i)) =: \widehat{G}_{Rj}(i)$, are defined analogously. The empirical Gini gain, i.e. the impurity reduction produced by splitting at the i-th cutpoint of variable X_j that corresponds to Equation 1.3 with the Gini index as impurity measure $\widehat{\mathfrak{I}}$, can also be displayed as a function of i and is based on the difference in impurity before and after splitting

$$\begin{aligned} \widehat{\Delta G}_j(i) &= \widehat{G}_j - \left(\frac{n_{Lj}}{n_j} \widehat{G}_{Lj}(i) + \frac{n_{Rj}}{n_j} \widehat{G}_{Rj}(i) \right) \\ &= \widehat{G}_j - \left(\frac{i}{n_j} \widehat{G}_{Lj}(i) + \frac{n_j - i}{n_j} \widehat{G}_{Rj}(i) \right). \end{aligned} \tag{2.3}$$

From a statistical point of view the empirical Gini index can be rephrased as

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

with $\hat{\pi}_j$ abbreviating the relative class frequency $\frac{n_{2j}}{n_j}$ of $Y = 2$.

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

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

$$G = 2\pi(1 - \pi)$$

which is a function of the true class probability π.

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

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

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

$$\text{Bias}(\widehat{G}_j) = -G/n_j,$$

where the extent of the bias depends on the true value of the Gini index and the number of observations n_j, that depends on the number of missing values in variable X_j. 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 $\pi_{Lj} = P(Y = 2|X_j \leq x_{j(i)})$, is equal to the true class probability in the right node $\pi_{Rj} = P(Y = 2|X_j > x_{j(i)})$ and thus equal to the overall class probability $\pi = P(Y = 2)$.

The expected value of the Gini gain $\widehat{\Delta G}_j$ (Equation 2.3) for fixed n_{Lj} and n_{Rj}, i.e. for a given cutpoint, is then

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

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, even if the cutpoint is not optimally chosen. The issue of optimal cutpoint selection and the multiple comparisons problem it induces is treated below. Estimation effects and multiple testing interact as sources of variable selection bias in binary splitting of variables with missing values. However, we will see in the simulation results in Chapter 3 that the estimation effect is predominant.

Our result of the derivation of the expected value of the Gini gain corresponds to that of Dobra and Gehrke (2001) when 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, and do not investigate the dependence on the sample size that is necessary to understand the preference for variables with many missing values in the study of Kim and Loh (2001).

The bias in favor of variables with many missing values 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, $j = 1, \ldots, p$, have different sample sizes n_j, this bias leads to a preference for 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).

2.1.2 k-ary splitting

When we consider k-ary splitting, the notation can be simplified even further, because no mutable cutpoint is selected, but the nodes are defined deterministically by the numbers of categories of a variable once it is selected: Let X_j, $j = 1, \ldots, p$, denote categorical predictor variables. For the categorical predictors let m_j, with $m_j \in \{1, \ldots, k_j\}$, denote the category. The starting set of all observations in the root node is again denoted by $\mathbf{C}$. The subsets $\mathbf{C}_{1,j}$ through $\mathbf{C}_{k_j,j}$ are produced by splitting $\mathbf{C}$ into k_j subsets defined by the categories of predictor X_j.

The empirical impurity reduction induced by splitting in the variable X_j is the following function (that corresponds to Equation 1.3 extended to k_j nodes).

$$\widehat{\Delta\Im}(\mathbf{C}, \mathbf{C}_{1,j}, \ldots, \mathbf{C}_{jk_j}) = \widehat{\Im}(\mathbf{C}) - \sum_{m_j=1}^{k_j} \frac{n_{m_j,j}}{n} \cdot \widehat{\Im}(\mathbf{C}_{m_j,j}), \tag{2.4}$$

where $\widehat{\Im}(\mathbf{C})$ is again the empirical impurity measure for the set $\mathbf{C}$ before splitting, while $\widehat{\Im}(\mathbf{C}_{m_j,j})$ is the empirical impurity measure for the subset $\mathbf{C}_{m_j,j}$. The proportion of observations assigned to subset $\mathbf{C}_{m_j,j}$ is denoted as $\frac{n_{m_j,j}}{n}$. If the variables vary in their number of missing values, the number of available observations of X_j could again be indicated by

using n_j instead of the overall number of observations n. When the Gini index is used as the impurity measure $\widehat{\mathfrak{I}}$ the empirical Gini gain results as

$$\Delta\widehat{G}(\mathbf{C}, \mathbf{C}_{1,j}, \ldots, \mathbf{C}_{jk_j}) = \widehat{G}(\mathbf{C}) - \sum_{m_j=1}^{k_j} \frac{n_{m_j,j}}{n} \cdot \widehat{G}(\mathbf{C}_{m_j,j}). \tag{2.5}$$

In this notation, the expected value for the plug-in estimator of the Gini index in one node is

$$E\left(\widehat{G}(\mathbf{C}_{m_j,j})\right) = G(\mathbf{C}_{m_j,j}) - \frac{G(\mathbf{C}_{m_j,j})}{n_{m_j,j}}. \tag{2.6}$$

Obviously this quantity again underestimates the true node impurity $\widehat{G}(\mathbf{C}_{m_j,j})$ by the quantity $\frac{G(\mathbf{C}_{m_j,j})}{n_{m_j,j}}$ depending on the true Gini index and inversely on the sample size of the node $n_{m_j,j}$. It is again well interpretable that the estimation of $\widehat{G}(\mathbf{C}_{m_j,j})$ is less reliable and the bias increases when the estimation is based on a smaller number of observations.

Under the null hypothesis of an uninformative predictor variable X_j, the true Gini index is equal in each node (i.e., $G(\mathbf{C}_{m_j,j}) = G(\mathbf{C}_{m'_j,j}) = G(\mathbf{C})$) and can again be denoted as an overall G. The expected value of the Gini gain over all nodes is again supposed to be 0 in this case, because splitting in a meaningless variable should produce no systematic impurity reduction. However, we find for k-ary splitting that

$$\begin{aligned} E\left(\Delta\widehat{G}(\mathbf{C}, \mathbf{C}_{1,j}, \ldots, \mathbf{C}_{jk_j})\right) &= G - \frac{G}{n} - \sum_{m_j=1}^{k_j} \frac{n_{m_j,j}}{n} \cdot G - \frac{G}{n} \\ &= \sum_{m_j=1}^{k_j-1} \frac{G}{n}. \end{aligned} \tag{2.7}$$

This quantity obviously depends on the number of categories k_j such that variables with more categories are likely to produce a higher Gini gain in average. The reason for this is that, when the original sample size is split up in more different nodes, the number of observations in each node decreases and the entropy estimation is less reliable as described

above. This effect is added up over all nodes and aggravated by the number of nodes that the sample size is divided into. The same principle holds for the Shannon entropy used as a split selection criterion in C4.5 and related algorithms, as illustrated in Chapter 4.

The variance of the empirical Gini index can be shown to depend on the true Gini index and to increase with decreasing sample size (Strobl et al., 2007). The variance of the empirical Gini gain also depends on the number of categories and increases with decreasing sample size (Dobra and Gehrke, 2001). Thus, not only does the bias result in a higher average, but also the variance may induce more extreme values – in principle both positive and negative but shifted by the estimation bias in favor of variables with many categories.

The other mechanism responsible for variable selection bias, namely the effect of multiple comparisons, that 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 but not in k-ary splitting, is outlined in the next section.

2.2 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 $n - 1$. For categorical and ordinal predictor variables, the number of cutpoints depends on the number of categories. The multiple comparisons effect results in a preference for 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 k-ary 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.

2.3 Summary

The simulation results obtained by Kim and Loh (2001) and Dobra and Gehrke (2001) in different settings may be explained by the three partially counteracting effects:

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 continuous 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, as is also illustrated in the simulation study in Chapter 3.

In the k-ary splitting case of Dobra and Gehrke (2001) on the other hand, the bias towards variables with large number of categories is due to the bias and variance effects, and not due to multiple comparisons.

In practice, the number of categories in categorical variables of nominal and ordinal scales of measurement often depends on arbitrary choices (in particular in the design of ques-

tionnaires) 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 a historical note: The reason why Breiman et al. (1984) did notice the multiple comparisons effect evident when categorical predictors vary in their number of categories, but did not notice the bias in favor of variables with many missing values (and even claim that the CART approach can deal particularly well with missing values, because it provides surrogate splits when predictor values are missing in the test sample), was that in their simulation studies Breiman et al. (1984) only spread missing values randomly over all predictor variables – instead of varying the sample sizes between variables, which induces the bias.

3. Evaluation of an unbiased variable selection criterion for binary splitting

The different approaches that have been suggested to eliminate variable selection bias in classification trees can be roughly divided in three categories: (i) those that are only applicable to k-ary splitting because they do not account for multiple testing, (ii) those that avoid the problem by means of separating the issue of variable and cutpoint selection and (iii) those that account for optimal cutpoint selection within the framework of combined variable and cutpoint selection. Representatives of these groups are (i) Dobra and Gehrke (2001), (ii) Loh and Shih (1997) and Hothorn et al. (2006), and (iii) Shih (2004), Lausen et al. (2004) and Strobl et al. (2007). Note also that an unbiased or even uniformly minimum variance unbiased (UMVU) estimator for an empirical impurity measure would not be sufficient as an unbiased split selection criterion in binary splitting, because it does not account for optimal cutpoint selection. Only in k-ary splitting an unbiased impurity estimator can guarantee that variable selection is unbiased (cf. Chapter 4), while the variance of an unbiased and even an UMVU estimator can differ for variables with different numbers of categories.

The idea to employ the p-values of optimally selected statistics for split selection, that is shared by the representatives of the third group of criteria, is very straightforward and intuitive because the well-known structure of combined variable and cutpoint selection in classification tree algorithms can be retained and only the statistic used for split selection has to be replaced by a p-value that is computed such that it accounts for the optimal selection of the cutpoint. Therefore, this approach and the evaluation of a p-value criterion

is described in detail in this chapter.

3.1 Optimally selected statistics

To illustrate the effect of optimal cutpoint selection, let us consider the familiar χ^2-statistic for a given 2×2 table, that results from a binary predictor variable X_j and a binary response Y. Under the null hypothesis of no association between X_j and Y, the distribution of the χ^2-statistic for this table with 1 degree of freedom is depicted as the dashed line in Figure 3.1.

If, however, the variable X_j originally comes with 4 categories, but instead of considering the resulting 2×4 table we select the best binary partition of X_j such as to optimize the χ^2-statistic of the resulting 2×2 table – like we would in binary splitting – the corresponding distribution is shifted to the right as illustrated by the solid line in Figure 3.1: When the binary partition is not given, but is selected such as to optimize the statistic, it is easier to produce large values. Thus, if the nominal χ^2 distribution was used to evaluate the optimally selected statistic, a value that may well be produced under the null hypothesis may seem highly significant.

Therefore, a fair comparison of variables that provide different numbers of cutpoints, as for example our variables X_j with either two or four categories, is only possible when the respective distribution of the optimally selected statistic (i.e. of the maximum over all statistics resulting from the different cutpoints) is considered. The p-values derived from this distribution are a fair means of comparison, because both sample size and multiple testing effects are accounted for.

Technically, the distributions of various optimally selected statistics can be derived by means of asymptotic or exact combinatorial arguments. For recursive partitioning, in principle the p-value of any association statistic can serve as a split selection criterion. However, since many applicants of the standard procedures CART and C4.5 are more

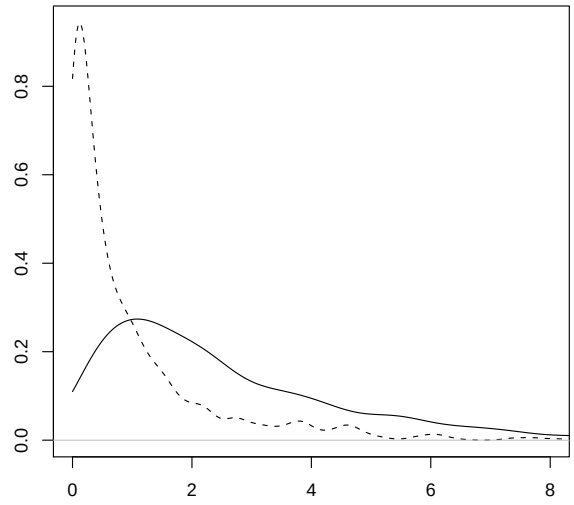


Fig. 3.1: Distribution (kernel density estimates) of the χ^2-statistic under the null hypothesis of no association between X_j and Y for a given 2×2 table (dashed) and for an optimally selected binary partition from a 2×4 table (solid).

familiar with the impurity reduction approach based on empirical entropy measures, a very intuitive approach is to use the p-value of the optimally selected Gini gain as the criterion. A way to derive the exact distribution of the optimally selected Gini gain was suggested by Boulesteix in Strobl et al. (2007). Here we will illustrate how this criterion was evaluated in a series of simulation studies and an application to veterinary data in order to support and complement to the theoretical results from the previous chapter.

3.2 Simulation studies

In this section, simulation studies are conducted to compare the variable selection performance of the p-value of the optimally selected Gini gain 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 in variable X_1 are sampled completely at random 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 the 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 theoretical results in the previous

chapter, we expect the Gini gain criterion to be biased towards the predictor variable with missing values, regardless of its information content.

3.2.1 Null case

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

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

For each percentage of missing values (MCAR), the obtained frequencies of selection of X_1, X_2, X_3, X_4 and X_5 over the 1000 simulation runs are given in Table 3.1 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, random choice frequency $\frac{1}{5}$.

However, 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.

Tab. 3.1: 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

3.2.2 Power case I

In the first power case study, the four uninformative predictor variables X_2, X_3, X_4 and 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

$$\begin{aligned} X_1|Y=1 &\sim N(0,1) \\ X_1|Y=2 &\sim N(0.5,1). \end{aligned}$$

(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. 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 3.2 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.

Tab. 3.2: 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

3.2.3 Power case II

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

$$\begin{aligned} X_2|Y=1 &\sim N(0,1) \\ X_2|Y=2 &\sim N(0.5,1). \end{aligned}$$

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. 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 3.3 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 3.2. In contrast, the p-value criterion shows no variable selection bias.

Tab. 3.3: 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

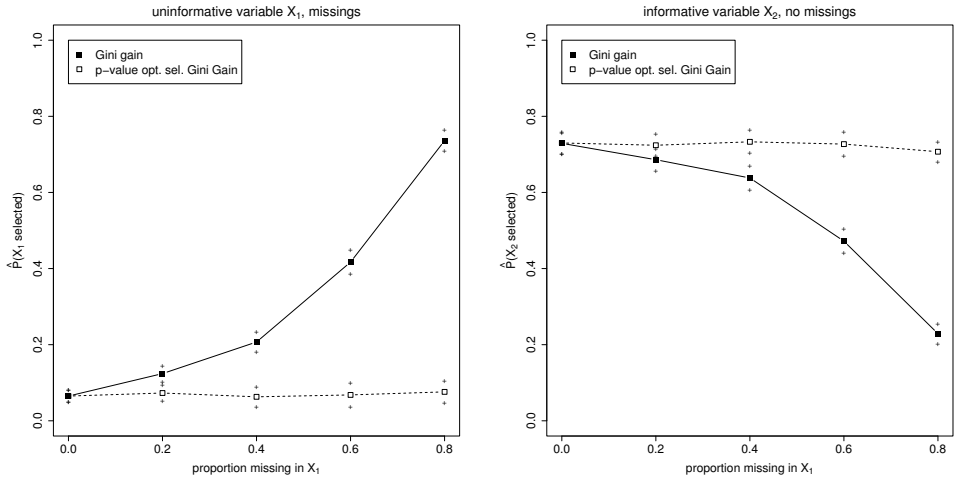


Fig. 3.2: Power case II: Variable selection frequencies for the uninformative variable X_1 containing missing values (left) and the informative variable X_2 containing no missing values (right).

3.3 Application to veterinary data

In addition to the simulation studies, the two split selection criteria were also applied to a real data set from veterinary gynecology. 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 3.4.

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

For the first exemplary analysis presented here we treat the missing values as if they were missing completely at random within each variable, even though this assumption is debatable for the data at hand. Meanwhile we know from the results of Svejdar et al. (2008) that even for many non-random missing data generating mechanisms the p-value approach behaves conservatively and underrates the information content of a variable with

many missing values rather than preferring it.

3.3.1 Variable selection ranking

The Gini gain criterion and the 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 gain criterion and with the 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 3.3 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 (i) the spots deviate noticeably from the bisector and (ii) 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 the p-value criterion. Thus, it is of high practical relevance to use the unbiased p-value criterion instead of the biased Gini gain for variable selection, because the variable ranked highest by the chosen criterion would be selected for further splitting.

3.3.2 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 the 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 3.5.

We find again in Table 3.5 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 the 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 the 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 the p-value criterion in this data set.

Moreover, the 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 at, e.g., the 95%-level, corresponding to a 5%-level of significance. The considered node is split only if the criterion value of the selected variable exceeds this threshold, i.e. if the corresponding p-value is ≤ 0.05. In this example the split in the selected variable would be conducted for weeks 0 through 3; only in week 4 the split does not produce enough impurity reduction and is omitted if the threshold is fixed at the 5%-level. If the threshold was fixed at the 99%-level, corresponding to a 1%-level of significance, the split would be conducted in weeks 0 through 2. This way to proceed is compatible with the insignificant results of the logistic regression models in weeks 3 and 4.

3.4 Summary

Using p-values of optimally selected statistics as split selection criteria avoids all sources of variable selection bias examined in Chapter 2. In simulation and real data studies, the approach has proved to deal effectively with different amounts of missing values in the predictor variables. While the results presented here focus on the case of values missing completely at random, the subsequent studies of Svejdar et al. (2008) have shown that even if values are missing not completely at random the p-value criterion guarantees conservative variable selection, that does not favor variables with many missing values.

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, the approach presented here has the advantage that it detects the information drop in informative variables caused by an increasing number of missing values.

Another advantage is that the approach 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 (in combination with, e.g., Bonferroni adjustment for multiple testing or optimally selected versions of them) as split selection criteria. The p-value of the Gini gain can replace the original 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.

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. In a more consistent approach, Hothorn et al. (2006) propose a unifying conditional inference framework to separately select the splitting variable and cutpoint. Here, p-values from the Monte-Carlo estimate or 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.

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. By using a p-value criterion based on the Gini index, one can efficiently address the problem

of selection bias – and at the same time preserve the simplicity of traditional recursive partitioning methods. In addition, the p-value can provide a statistically sound stopping criterion.

However, the exact derivation of the distribution of the optimally selected Gini gain, that was suggested by Boulesteix in Strobl et al. (2007), comes with computational strings attached: The advantage of an exact approach is that it is well suited for small samples sizes, that frequently occur in the bottom nodes of classification trees. On the other hand, any exact approach is computationally expensive. Therefore, an implementation of the asymptotic approach for unbiased variable and split selection described in Hothorn et al. (2006) will be used for the large ensemble studies in the later chapters of this work.

Tab. 3.4: Potential predictor variables from the cow data set. All variables are continuous, 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)

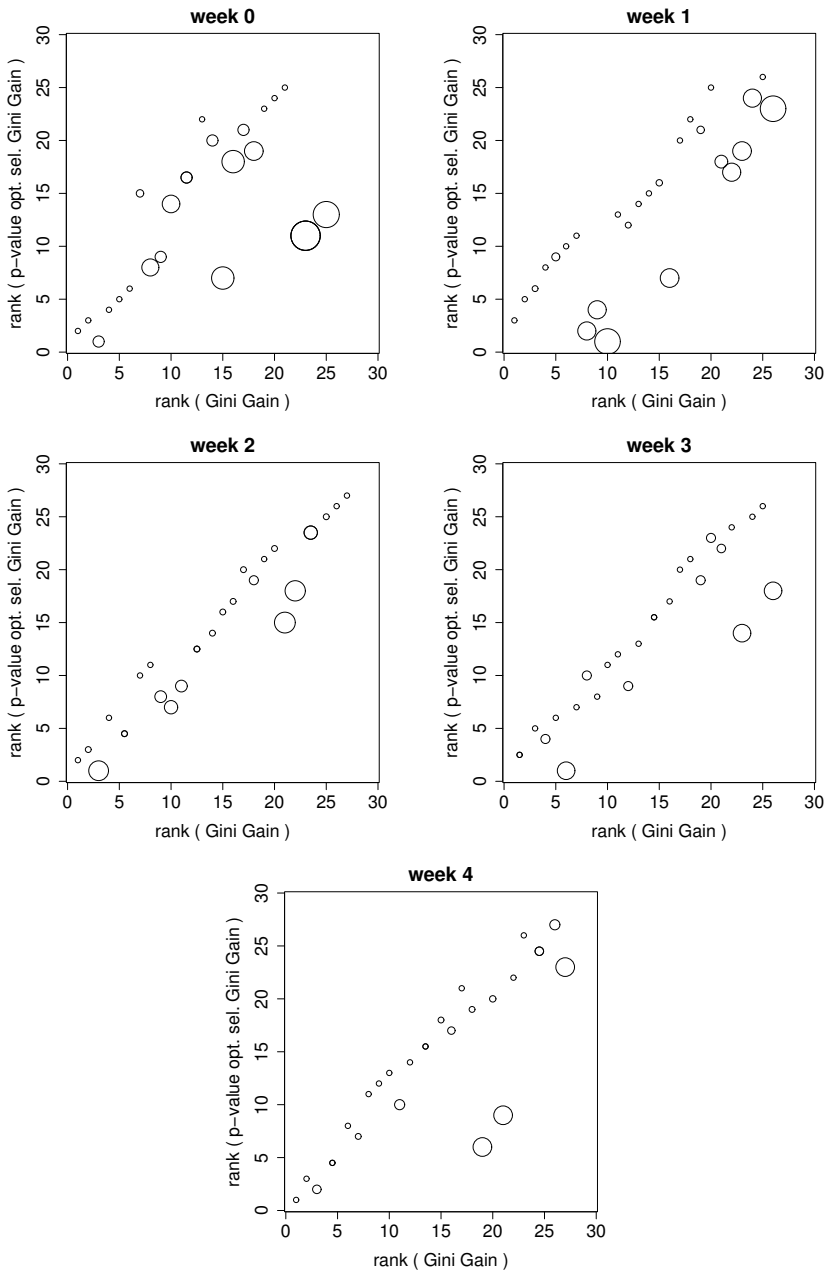


Fig. 3.3: Rank obtained with the new p-value criterion vs. rank obtained with the Gini gain. The circumference of each point is proportional to number of missing values in the predictor.

Tab. 3.5: Variables selected for the first split using the standard Gini gain (top) and the p-value criterion (bottom), as well as p-values from the logistic regression model (LRM) corresponding to model likelihood ratio tests, with the 5%-level of significance indicated by the * and the 1%-level of significance 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**

4. Robust and unbiased variable selection in k-ary splitting

A strong disadvantage of traditional classification trees is their instability and susceptibility to overfitting, that affects their robustness against outliers in the sample and necessitates terminal pruning. The extension of classification trees as so-called "credal classifiers" based on imprecise probabilities by Abellán and Moral (2005) establishes a more robust means of classification, that is not as susceptible to overfitting and thus provides more reliable results.

The approach of classification trees based on imprecise probabilities for categorical predictor variables by Abellán and Moral (2005), that is considered here, is inspired by the k-ary splitting C4.5 algorithm. Variable selection is conducted with respect to an upper entropy criterion based on the Shannon entropy.

As outlined in Chapter 2, a 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. We will show that variable selection bias affects variable selection in the approach of Abellán and Moral (2005), too, if the predictor variables vary in their numbers of categories.

The main 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 (2005), is a negatively biased estimator of the true Shannon entropy. In this respect, the same problem of biased entropy estimation that affected the empirical Gini index in

standard classification trees in Chapter 2 now applies to the empirical Shannon entropy in classification trees based on imprecise probabilities. However, in the context of imprecise probabilities, that are processed by means of an upper entropy approach in the work of Abellán and Moral (2005), a new, counteracting effect is induced, that depends on the true information content of the variables.

An unbiased entropy estimator is suggested and discussed as a split selection criterion in this context, and is evaluated in simulation studies investigating the variable selection performance of the biased and corrected estimators.

This chapter starts with an outline of the approach of classification trees based on imprecise probabilities in Section 4.1. Section 4.2 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 4.3.

4.1 Classification trees based on imprecise probabilities

The rationale of classification trees based on imprecise probabilities for categorical predictor variables by Abellán and Moral (2005) is similar to the traditional classification tree approach C4.5 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. 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 (cf. Figure 4.1). 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

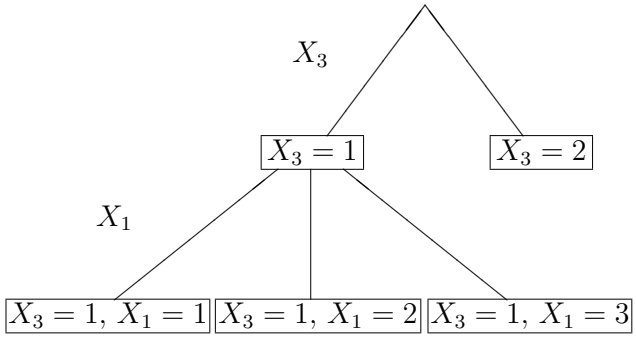


Fig. 4.1: Example of a k-ary splitting classification tree. Configurations of predictor values characterizing the observations in each node are displayed in boxes depicting the nodes.

(2005) apply the Imprecise Dirichlet Model (abbreviated by IDM in the following; see Walley (1996) for the introduction of the model and Bernard (2004), as well as Bernard (2008), for an overview of further developments) in the construction of the classification tree.

The IDM was developed as a means of predictive inference for modeling prior and posterior uncertainty about the class probabilities in learning from multinomial data. It was proposed in the framework of imprecise probabilities, where sets of prior distributions, rather than single ones, are processed to account for uncertainty about model parameters.

In the application of the IDM in classification trees, that is considered here, this means that instead of using the original class frequencies as estimates for the class probabilities in the computation of the impurity criterion, upper and lower bounds for the class probabilities are derived by means of the IDM. These upper and lower bounds enclose a set of class probabilities, from which Abellán and Moral (2005) proceed with the one that produces the most conservative estimation of the impurity criterion in order to robustify the split selection. The resulting classification trees are called "credal classifiers" because, in the spirit of imprecise probabilities, a set of response classes, rather than a single one, is returned whenever there is no clearly dominating class.

The split selection criteria and procedure of Abellán and Moral (2005) are introduced more formally in the following: At first, the split selection criteria are introduced for one arbitrary node in Section 4.1.1. Then the entire split selection procedure, starting from that node, is treated in Section 4.1.2.

As opposed to the previous chapters, we now need additional notation to encode split selection not only in an exemplary first node, but also in later nodes of the tree. As outlined in the introduction, in these later nodes split selection is conditional on the configuration of previously selected splitting variables in the same branch. Therefore, the predictor variable configuration that characterizes all observations in one node is now denoted as γ (cf. again Figure 4.1: for example, the lower leftmost node is defined by the configuration $\gamma = (X_3 = 1, X_1 = 1)$).

4.1.1 Total impurity criteria

Let Y again be a categorical response variable with values $c \in \{1, 2, \ldots, |\mathscr{C}|\}$ in a finite set $\mathscr{C}$. The credal set $\mathscr{P}^\gamma$ is a convex set of classical probability distributions π^γ on the set $\mathscr{C}$, representing the available information on the unknown value of the response variable Y in the node defined by predictor variable configuration γ.

The total impurity criterion $TU2(\mathscr{P}^\gamma)$ for the credal set $\mathscr{P}^\gamma$

$$TU2(\mathscr{P}^\gamma) = \max_{\pi^\gamma \in \mathscr{P}^\gamma} \left\{ -\sum_{c=1}^{|\mathscr{C}|} \pi^\gamma(c) \ln[\pi^\gamma(c)] \right\} \tag{4.1}$$

is a generalization of the popular Shannon entropy for classical probabilities.

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

$$TU1(\mathscr{P}^\gamma) = TU2(\mathscr{P}^\gamma) + IG(\mathscr{P}^\gamma), \tag{4.2}$$

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

$$IG(\mathscr{P}^\gamma) = \sum_{A \subseteq \mathscr{C}} \mathfrak{m}_{\mathscr{P}^\gamma}(A) \ln(|A|)$$

and $\mathfrak{m}_{\mathscr{P}^\gamma}$ is the Möbius inverse of the lower envelope $f_{\mathscr{P}^\gamma} = \inf_{\pi^\gamma \in \mathscr{P}^\gamma} \pi^\gamma(A)$

$$\mathfrak{m}_{\mathscr{P}^\gamma}(A) = \sum_{B \subseteq A} (-1)^{|A-B|} f_{\mathscr{P}^\gamma}(B),$$

with $|A-B|$ denoting the cardinality of the set A excluding B. $IG(\mathscr{P}^\gamma)$ 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 (cf. Klir, 1999, 2003).

The total impurity measure $TU1(\mathscr{P}^\gamma)$ additively incorporates both uncertainty and non-specificity. Abellán and Moral (2005) argue that adding a measure of non-specificity as in $TU1(\mathscr{P}^\gamma)$ overweighs non-specificity in the total impurity criterion, because $TU2(\mathscr{P}^\gamma)$ also increases with non-specificity. The authors thus settle for $TU2(\mathscr{P}^\gamma)$ as a measure of total uncertainty.

The data are incorporated in estimating the value of $TU2(\mathscr{P}^\gamma)$ by means of applying the IDM locally within each node. For each node, defined by predictor variable configuration γ, the calculation of the lower and upper probabilities with the IDM is based on counts of n_c^γ class c objects out of n^γ objects in total in the node:

$$[\underline{\pi}^\gamma(c), \overline{\pi}^\gamma(c)] = \left[\frac{n_c^\gamma}{n^\gamma + s} \, , \, \frac{n_c^\gamma + s}{n^\gamma + s} \right], \tag{4.3}$$

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 c. The credal set $\mathscr{P}^\gamma$ in $TU2(\mathscr{P}^\gamma)$ is then given by all probability distributions π^γ on the set $\mathscr{C}$, for which $\pi^\gamma(c) \in [\underline{\pi}^\gamma(c), \overline{\pi}^\gamma(c)]$ for all c, as derived in Equation 4.3. The maximization in $TU2(\mathscr{P}^\gamma)$ is technically accomplished by means of the upper entropy algorithm introduced in Abellán and Moral (2003). The algorithm identifies the posteriori probability distribution on $\mathscr{C}$ with the upper entropy, that is in accordance with the upper and lower probabilities for each class $c \in \mathscr{C}$ derived from the IDM.

4.1.2 Split selection procedure

The complete process of variable selection in the classification tree algorithm of Abellán and Moral (2005) consists of the following successive tasks:

Let X_j again be a categorical predictor variable with values $m_j \in \{1, 2, \ldots, k_j\}$ in a finite set K_j, with $k_j = |K_j|$. Starting from a node defined by predictor variable configuration γ, for each potential splitting variable X_j as many nodes as categories k_j are produced. Within each new node, defined by the previous configuration γ in combination with the value m_j of the potential splitting variable X_j by $\gamma \cup (X_j = m_j)$, the lower and upper probabilities $[\underline{\pi}^{\gamma \cup (X_j = m_j)}(c), \overline{\pi}^{\gamma \cup (X_j = m_j)}(c)]$ of each response class c are derived from the class counts $n_c^{\gamma \cup (X_j = m_j)}$ by means of the IDM. The interval width is determined by the number of observations per node $n^{\gamma \cup (X_j = m_j)}$ and the hyperparameter s of the IDM. The computation of the upper entropy criterion is then conducted in two steps:

- From the credal set $\mathscr{P}^{\gamma \cup (X_j = m_j)}$ derived from the lower and upper probabilities $[\underline{\pi}^{\gamma \cup (X_j = m_j)}(c),\ \overline{\pi}^{\gamma \cup (X_j = m_j)}(c)]$ the posterior upper entropy distribution $\pi_{\text{maxE}}^{\gamma \cup (X_j = m_j)}$, i.e., the distribution closest to the uniform distribution over the response classes in the set $\mathscr{C}$, is determined by the algorithm given in Abellán and Moral (2003).
- The value of $TU2(\mathscr{P}^{\gamma \cup (X_j = m_j)})$ is then estimated by applying the plug-in estimator of the Shannon entropy, indicated by $\widehat{H}(\cdot)$, to the posterior upper entropy distribution.

$$\begin{aligned} \widehat{TU2}(\mathscr{P}^{\gamma \cup (X_j = m_j)}) &= \widehat{H}\left(\pi_{\text{maxE}}^{\gamma \cup (X_j = m_j)}\right) \\ &= -\sum_{c=1}^{|\mathscr{C}|} \pi_{\text{maxE}}^{\gamma \cup (X_j = m_j)}(c) \cdot \ln\left[\pi_{\text{maxE}}^{\gamma \cup (X_j = m_j)}(c)\right] \end{aligned} \tag{4.4}$$

The impurity that remains after splitting in variable X_j is empirically measured by the weighted sum of total impurity measures over all new nodes

$$\widehat{\mathfrak{I}}(\gamma, X_j) = \sum_{m_j \in K_j} \frac{n^{\gamma \cup (X_j = m_j)}}{n^{\gamma}} \widehat{TU2}(\mathscr{P}^{\gamma \cup (X_j = m_j)}), \tag{4.5}$$

where $\frac{n^{\gamma \cup (X_j = m_j)}}{n^{\gamma}}$ is the relative frequency of observations assigned to each new node. The variable X_j for which $\widehat{\mathfrak{I}}(\gamma, X_j)$ is minimal is selected for the next split. This approach is equivalent to selecting the variable that produces the maximal empirical impurity reduction $\widehat{\Delta\mathfrak{I}}$ as in the previous chapters, because the impurity of the starting node is equal for all candidate splits.

4.1.3 Characteristics of the total impurity criterion TU2

In order to illustrate the variable selection characteristics of the total impurity criterion $TU2(\mathscr{P}^{\gamma})$ 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. Note that 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.

The results displayed below are from a simulation study run with 1000 iterations 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 4.1 the sampling distribution of the response variable differed in the categories of the predictor variables depending on the relevance parameter.

Figures 4.2 through 4.4 depict the results of the simulation study as barplots with the bar height indicating the estimated selection probabilities for the two equally informative

Tab. 4.1: 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.

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

predictor variables and the crosses marking $\pm$ 2 empirical standard errors of the point estimates.

The results of the simulation studies show that two characteristics of the total impurity criterion $TU2(\mathscr{P}^{\gamma \cup (X_j = m_j)})$ 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 $\pi_{\text{maxE}}^{\gamma \cup (X_j = m_j)}$ (in step 1 of the computation of the upper entropy criterion outlined in Section 4.1.2) a smaller number of observations per node results in a wider interval of lower and upper probabilities $[\underline{\pi}^{\gamma \cup (X_j = m_j)}(c), \overline{\pi}^{\gamma \cup (X_j = m_j)}(c)]$. From a wider interval a more uninformative upper entropy distribution $\pi_{\text{maxE}}^{\gamma \cup (X_j = m_j)}$ can be derived. Thus, the total impurity criterion $TU2(\mathscr{P}^{\gamma} \cup (X_j = m_j))$ 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 $\pi_{\text{maxE}}^{\gamma \cup (X_j = m_j)}$, that is obtained from the wide intervals. Figure 4.2 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 4.3 we see

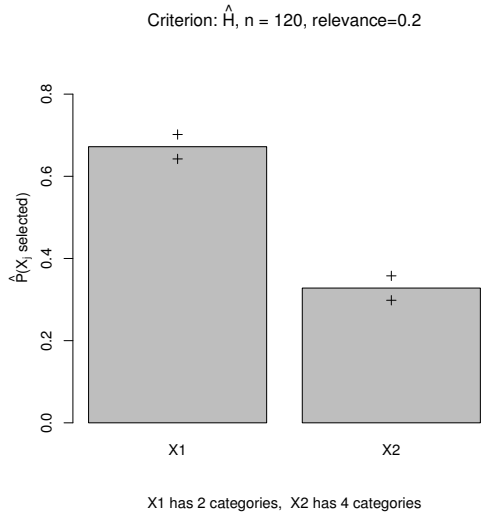


Fig. 4.2: Estimated variable selection probabilities for the upper entropy-total impurity criterion TU2. Both predictors are informative with medium relevance, they only vary in their number of categories.

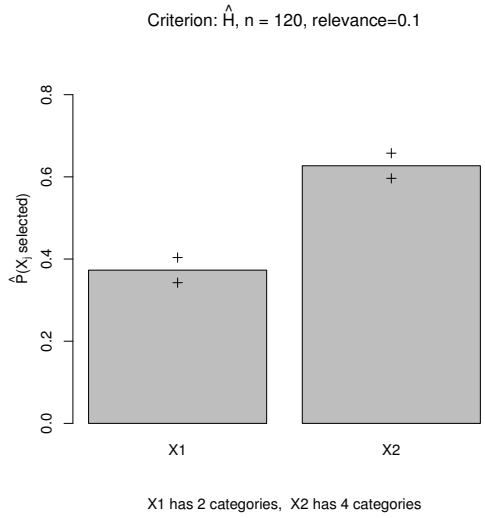


Fig. 4.3: Estimated variable selection probabilities for the upper entropy-total impurity criterion TU2. Both predictors are informative with low relevance, they only vary in their number of categories.

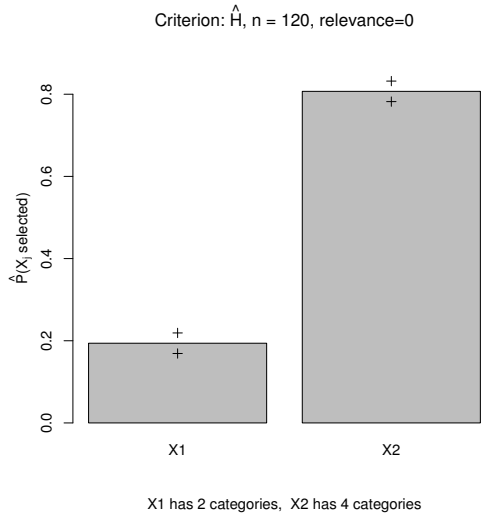


Fig. 4.4: Estimated variable selection probabilities for the upper entropy-total impurity criterion TU2. Both predictors are uninformative, they only vary in their number of categories.

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 this second mechanism is most prominent as shown in Figure 4.4. The mechanism obvious in Figures 4.3 and 4.4 induces a preference for 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 4.1.2). 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 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.

4.2 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 shortly review the necessary statistical background on the estimation of the Shannon entropy and then apply the results to classification trees based on imprecise probabilities.

4.2.1 Estimation bias for the empirical Shannon entropy

The theoretical Shannon entropy

$$H(\pi) = -\sum_{c=1}^{|\mathscr{C}|} \pi\left(c\right) \ln[\pi\left(c\right)]$$

is a function of the true response class probabilities $\pi(c)$. 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

$$\widehat{H}(\widehat{\pi}) = -\sum_{c=1}^{|\mathscr{C}|} \widehat{\pi}\left(c\right) \ln[\widehat{\pi}\left(c\right)].$$

However, this widely used estimator is biased for finite sample sizes, because with a decreasing number of observations the standard error of the estimators $\widehat{\pi}(c)$ 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, $f\left(E_{\pi}(\widehat{\pi})\right) \geq E_{\pi}\left(f(\widehat{\pi})\right)$ for any concave function f, it is obvious that the unbiasedness of the maximum-likelihood estimators $\widehat{\pi}\left(c\right)$ is not

necessarily transferred to the plug-in estimator $\widehat{H}$, which may be negatively biased. The same principle was illustrated for the empirical Gini index in Chapter 2.

The extent of the bias for the empirical Shannon entropy can be evaluated from the expected value of the plug-in estimator $\widehat{H}$ for the true Shannon entropy H, that was derived independently by Miller (1955) and Basharin (1959)

$$\begin{aligned} E_\pi\left(\widehat{H}(\widehat{\pi})\right) &= E_\pi\left(-\sum_{c=1}^{|\mathscr{C}|}\widehat{\pi}\,(c)\,\ln[\widehat{\pi}\,(c)]\right) \\ &= E_\pi\left(-\sum_{c=1}^{|\mathscr{C}|}\frac{n_c}{n}\,\ln\left[\frac{n_c}{n}\right]\right) \\ &= H(\pi)-\frac{|\mathscr{C}|-1}{2n}+O\left(\frac{1}{n^2}\right), \end{aligned}$$

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 $\pi(c)$ (cf. 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

$$\widehat{H}_{\text{Miller}}(\widehat{\pi}) = \widehat{H}(\widehat{\pi}) + \frac{|\mathscr{C}|-1}{2n}.$$

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.

4.2.2 Effects in classification trees based on imprecise probabilities

As described in the beginning of Section 4.1.3, small sample sizes result in wider intervals of lower and upper probabilities $[\underline{\pi}^{\gamma\cup(X_j=m_j)}(c), \overline{\pi}^{\gamma\cup(X_j=m_j)}(c)]$ 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 4.3 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 $\widehat{TU2}(\mathscr{P}^{\gamma \cup (X_j = m_j)})$ 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 $\mathscr{C}$, the negative estimation bias occurring in each node is carried forward to the estimated criterion value $\widehat{\mathfrak{I}}(\gamma, X_j)$, on which the final decision in the variable selection procedure is based.

For an uninformative predictor variable, with the true class distribution $\pi^* := \pi_{\text{maxE}}^{\gamma \cup (X_j = m_j)} = U(1, |\mathscr{C}|)$ discretely uniform on support $[1, |\mathscr{C}|]$, the true entropy value $H^* := \sum_{c=1}^{|\mathscr{C}|} \pi^*(c) \cdot \cdot \ln[\pi^*(c)]$ is maximal and equal in each node. The approximated expected value of $\widehat{\mathfrak{I}}(\gamma, X_j)$ under the null hypothesis is then

$$\begin{aligned} E_{\pi^*}\left(\widehat{\mathfrak{I}}(\gamma, X_j)\right) &\approx \sum_{m_j \in K_j} \frac{n^{\gamma \cup (X_j = m_j)}}{n^{\gamma}} \left\{ H^* - \frac{|\mathscr{C}| - 1}{2\left(n^{\gamma \cup (X_j = m_j)} + s\right)} \right\} \\ &\approx H^* - k_j \cdot \frac{|\mathscr{C}| - 1}{2\, n^{\gamma}} \end{aligned}$$

where the number of response categories $|\mathscr{C}|$ is fixed, while the number of categories k_j

differs between the predictor variables X_j. Thus, the number of categories of the predictor variable X_j crucially affects its selection chance.

4.2.3 Suggested corrections based on the IDM

With $\widehat{H}\left(\pi_{\text{maxE}}^{\gamma\cup(X_j=m_j)}\right)$ denoting the standard plug-in estimator of the Shannon entropy applied to the posterior upper entropy distribution we suggest to use

$$\widehat{H}_{\text{Miller}}\left(\pi_{\text{maxE}}^{\gamma\cup(X_j=m_j)}\right) = \widehat{H}\left(\pi_{\text{maxE}}^{\gamma\cup(X_j=m_j)}\right) + \frac{|\mathscr{C}|-1}{2(n^{\gamma\cup(X_j=m_j)}+s)} \tag{4.6}$$

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 $n^{\gamma\cup(X_j=m_j)}$. This correction is again appropriate for medium $n^{\gamma\cup(X_j=m_j)}$, while it over-penalizes for small $n^{\gamma\cup(X_j=m_j)}$ with respect to the number of categoies $|\mathscr{C}|$, which is supported by the numerical results in Section 4.3.

In another correction approach we are revisiting the empirical measure $\widehat{IG}$, the theoretical analogy of which was employed by Abellán and Moral (2005) as a measure of non-specificity in the total impurity criterion $TU1(\mathscr{P}^{\gamma\cup(X_j=m_j)})$. Like the correction term in the above approach, the term $\widehat{IG}(\mathscr{P}^{\gamma\cup(X_j=m_j)})$ in Equation 4.2 is a function of the sample size $n^{\gamma\cup(X_j=m_j)}$ and the number of categories $|\mathscr{C}|$. In the special case where the lower probabilities used in the computation of the Möbius inverses in $\widehat{IG}(\mathscr{P}^{\gamma\cup(X_j=m_j)})$ are derived from the IDM, the Möbius inverses of all subsets of the power set of $\mathscr{C}$, besides the singletons $c \in \mathscr{C}$ and the complete set $\mathscr{C}$, 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 $\mathscr{C}$ collapses to the width $\frac{s}{n^{\gamma\cup(X_j=m_j)}+s}$ of the intervals of lower and upper probabilities on $\mathscr{C}$ computed from the IDM with hyperparameter s, and the empirical non-specificity measure $\widehat{IG}(\mathscr{P}^{\gamma\cup(X_j=m_j)})$ depends only on the sample size $n^{\gamma\cup(X_j=m_j)}$ through the interval width,

and on the number of categories $|\mathscr{C}|$ through the factor $\ln(|\mathscr{C}|)$. We thus suggest

$$\widehat{H}\left(\pi_{\max E}^{\gamma\cup(X_j=m_j)}\right) + \widehat{IG}\left(\mathscr{P}^{\gamma\cup(X_j=m_j)}\right) = \widehat{H}\left(\pi_{\max E}^{\gamma\cup(X_j=m_j)}\right) + \widehat{m}_{\mathscr{P}^{\gamma\cup(X_j=m_j)}}(\mathscr{C})\ln(|\mathscr{C}|) \tag{4.7}$$

i.e. $\widehat{TU1}(\mathscr{P}^{\gamma\cup(X_j=m_j)})$, as another corrected estimator, where $\widehat{m}_{\mathscr{P}^{\gamma\cup(X_j=m_j)}}(\mathscr{C})$ is the Möbius inverse computed from the posterior lower class probabilities derived from the IDM. We will again see in Section 4.3 that this correction is only reliable for sufficiently large $n^{\gamma\cup(X_j=m_j)}$ and small $|\mathscr{C}|$, while otherwise it is overcautious.

4.3 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 frequency 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 4.2 the response values in the simulation were fixed, while the uninformative predictors were sampled from discrete uniform distributions 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.

In this study, the behavior of the plug-in estimator $\widehat{H}$ for the Shannon entropy (cf. Equation 4.4) is compared to the behavior of the corrected estimators $\widehat{H}_{\text{Miller}}$ (Equation 4.6)

Tab. 4.2: 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.

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

and $\widehat{H} + \widehat{IG}$ (Equation 4.7). Figures 4.5 through 4.8 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 4.9 through 4.12 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 4.13 through 4.16 show that the reverse bias for small sample sizes and large difference in the number of categories is even stronger than for $\widehat{H}_{\text{Miller}}$.

4.4 Summary

The split selection criterion TU2 introduced for classification trees based on imprecise probabilities for categorical predictor variables by Abellán and Moral (2005) 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 informative 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 $\widehat{H}_{\text{Miller}}$ (Equation 4.6) shows even better variable selection performance than the corrected estimator $\widehat{H} + \widehat{IG}$ (Equation 4.7). 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 (2005). 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 may be considered for split selection in future research.

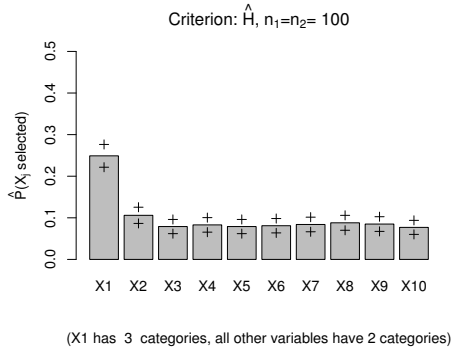


Fig. 4.5: Estimated variable selection probabilities for the plug-in estimator of the Shannon entropy for 3 vs. 2 categories in the predictor variables and medium sample sizes.

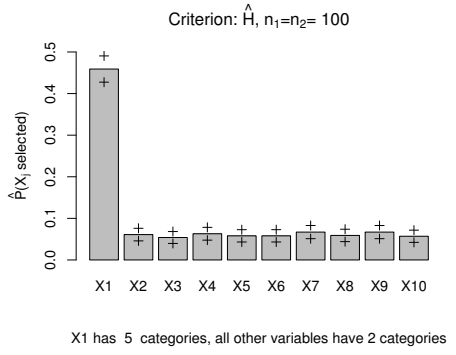


Fig. 4.7: Estimated variable selection probabilities for the plug-in estimator of the Shannon entropy for 5 vs. 2 categories in the predictor variables and medium sample sizes.

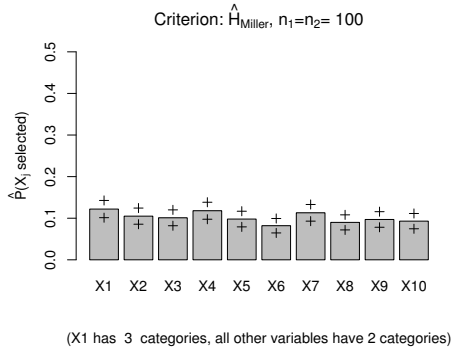


Fig. 4.9: Estimated variable selection probabilities for the corrected estimator of the Shannon entropy $\widehat{H}_{\text{Miller}}$, for 3 vs. 2 categories in the predictor variables and medium sample sizes.

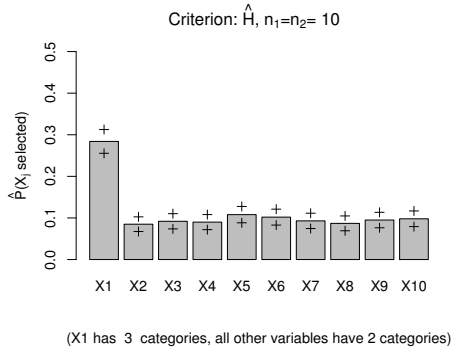


Fig. 4.6: Estimated variable selection probabilities for the plug-in estimator of the Shannon entropy for 3 vs. 2 categories in the predictor variables and small sample sizes.

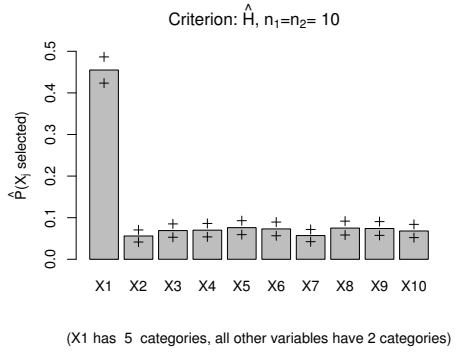


Fig. 4.8: Estimated variable selection probabilities for the plug-in estimator of the Shannon entropy for 5 vs. 2 categories in the predictor variables and small sample sizes.

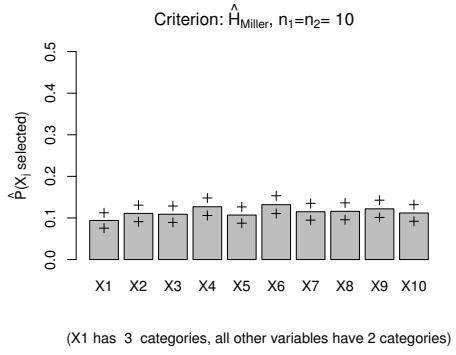


Fig. 4.10: Estimated variable selection probabilities for the corrected estimator of the Shannon entropy $\widehat{H}_{\text{Miller}}$, for 3 vs. 2 categories in the predictor variables and small sample sizes.

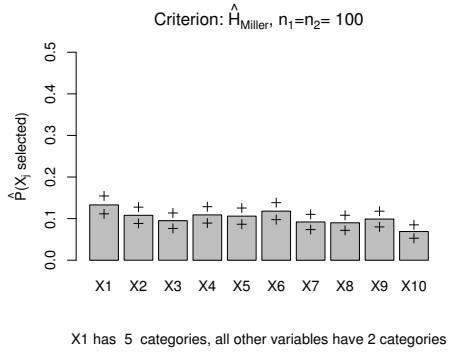


Fig. 4.11: Estimated variable selection probabilities for the corrected estimator of the Shannon entropy $\widehat{H}_{\text{Miller}}$, for 5 vs. 2 categories in the predictor variables and medium sample sizes.

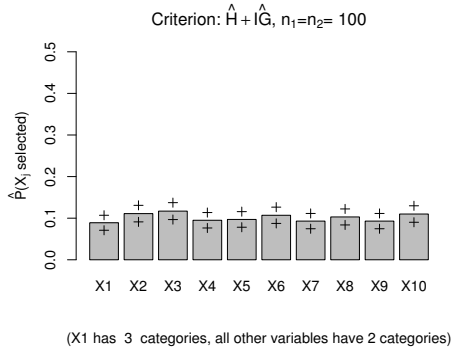


Fig. 4.13: Estimated variable selection probabilities for the corrected estimator of the Shannon entropy $\widehat{H} + \widehat{IG}$, for 3 vs. 2 categories in the predictor variables and medium sample sizes.

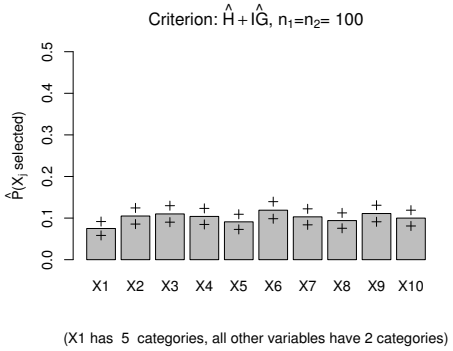


Fig. 4.15: Estimated variable selection probabilities for the corrected estimator of the Shannon entropy $\widehat{H} + \widehat{IG}$, for 5 vs. 2 categories in the predictor variables and medium sample sizes.

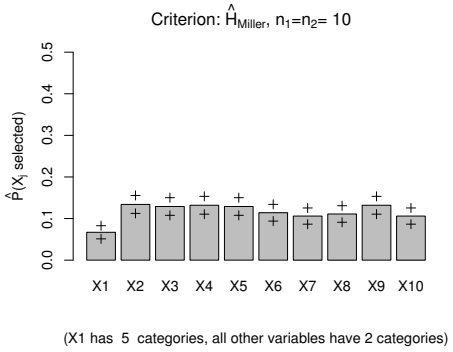


Fig. 4.12: Estimated variable selection probabilities for the corrected estimator of the Shannon entropy $\widehat{H}_{\text{Miller}}$, for 5 vs. 2 categories in the predictor variables and small sample sizes.

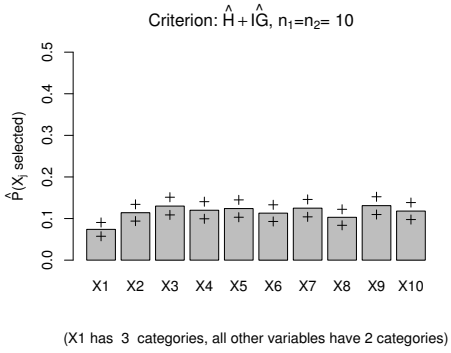


Fig. 4.14: Estimated variable selection probabilities for the corrected estimator of the Shannon entropy $\widehat{H} + \widehat{IG}$, for 3 vs. 2 categories in the predictor variables and small sample sizes.

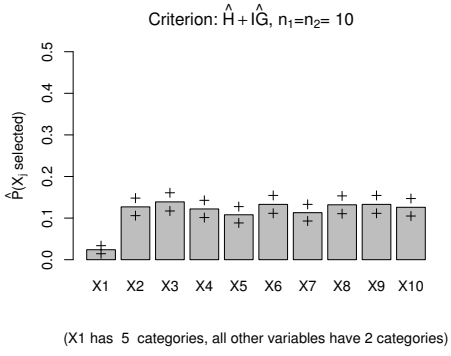


Fig. 4.16: Estimated variable selection probabilities for the corrected estimator of the Shannon entropy $\widehat{H} + \widehat{IG}$, for 5 vs. 2 categories in the predictor variables and small sample sizes.

5. Adaptive cutpoint selection in TWIX ensembles

The ensemble methods bagging and random forests, that will be treated in later chapters, employ sets of classification trees as a means to provide more stable predictions – but at the expense of completely giving up the interpretability of a single tree model. The TWIX method on the other hand, that was introduced by Potapov (2006) (see also Potapov et al., 2006; Potapov, 2007) and forms the basis for this chapter, resides somewhere in between single trees and usual ensemble methods: It starts with a single starting node, but branches to a set of trees at each decision by means of splitting not only in the best cutpoint (note that we are now returning to the case of binary splitting), but also in reasonable extra cutpoints.

When considering the prediction accuracy of tree-based models, TWIX has been shown to reliably outperform single trees and even to reach the predictive performance of ensemble methods like bagging and random forests on some data sets. However, in general it cannot compete with them because – in particular in the currently available version where the number of extra cutpoints has to be predetermined by the user and remains fixed – it becomes computationally infeasible for large sets of trees. In addition to this, the TWIX approach is limited to locally optimal variable selection on the original data set, while bagging and random forests induce variation by means of random sampling from the original data set and the set of predictor variables. This may reveal interaction effects that otherwise remain unnoticed, as outlined in the introduction.

When considering the overall value of computer intensive statistical learning methods, however, it is important to be aware that there is a tradeoff between two rivaling interests: prediction accuracy on one hand and interpretability on the other hand. With respect to interpretability, TWIX trees have an advantage over parallelized ensemble methods like bagging and random forests: While in random forest and bagging any form of direct interpretability is lost, TWIX trees can be considered as an expansion of a single tree model – since a TWIX ensemble forms a set of nested trees, that is derived from one single starting node.

In addition to this, we will show that by means of introducing a new, adaptive cutpoint selection strategy the size of the TWIX ensemble can be regulated in a data driven way. This approach combines two attractive features:

- Firstly, the robustification is parsimonious in the sense that additional cutpoints are considered only if the cutpoint under consideration proves to be unstable. This saves extra splits and thus makes the resulting TWIX ensemble more concise – a fact that adds not only to interpretability, but may also considerably reduce the computational expense of the TWIX method.

- Secondly, as a quite welcome by-product, it provides a diagnostic for the robustness of a single tree model: In an extreme case a TWIX ensemble with an adaptively chosen number of extra cutpoints can reduce to a single tree model, when one clearly dominant cutpoint is found in each split. The resulting tree has then proved to be stable with respect to small changes in the data set. The other extreme case is a widespread TWIX ensemble that indicates high instability of cutpoint selection. Such a large ensemble is not interpretable by any means, and a black box method like bagging or random forests with a higher prediction accuracy may be better suited for the particular data set.

To achieve this adaptive cutpoint selection, our approach takes one key problem of classification trees literally: the fact that they are so instable that a completely different tree

may have resulted if a few different observations had been observed. We formalize and utilize this very aspect and introduce additional virtual, yet unobserved, observations in the analysis. The values of these observations are, of course, unknown so that we consider all possible values. We then construct a TWIX ensemble resulting from all splits that are optimal under some constellation of these virtual observations.

In the following we will first shortly review the instability issue for the current cutpoint selection approaches and introduce the original suggestions for selecting extra cutpoints in TWIX in Section 5.1, before the adaptive cutpoint selection criterion is derived in Section 5.2. The behavior of the criterion is explored in Section 5.3. Section 5.4 gives a short outlook on the aggregation of predictions from ensembles of trees in the spirit of credal classification, before the results are summarized in Section 5.5.

5.1 Building TWIX ensembles

The rationale behind all ensemble methods is that they use a whole set of classification trees rather than a single tree for prediction. The prediction of all trees in the set is combined by voting or averaging. This approach leads to a significant increase in prediction accuracy on a test sample as compared to the performance of a single tree. TWIX shares this feature with the ensemble methods bagging and random forests, even though the sets of trees are created differently.

A question that arises with respect to sets of trees generated on random bootstrap or subsamples in bagging and random forests is: "Why use randomly generated and thus sub-optimal models?" (Potapov, 2006; Potapov et al., 2006). The TWIX response to this question is to start with a single tree built on the original learning sample, but to proceed in each split not only with the best cutpoint, but also with reasonable extra cutpoints, such as the second and third best cutpoint. In this approach a set of trees is created that start with the same root node but diverge further and further, whenever more than one cutpoint is considered worthwile for splitting. From this nested set of trees either the best individual

tree is selected, e.g., by means of a cross validation criterion, or trees are aggregated for prediction as in bagging and random forests. Potential strategies for aggregating credal predictions from sets of classification trees are outlined in Section 5.4.

On some data sets aggregated TWIX ensembles can even outperform standard ensemble methods (Potapov, 2006; Potapov et al., 2006). However, the TWIX approach is – at least in the originally proposed non-adaptive form – computationally expensive because the ensemble grows exponentially in the number of extra cutpoints. Depending on the number of extra cutpoints and the depth of the trees, the approach can soon become computationally infeasible and in general cannot compete with other ensemble methods that employ large sets of trees, because by default a moderate but fixed number of extra cutpoints in the current splitting variable are selected, that leads to an exponential growth of the TWIX ensemble. In this context it is helpful that our method not only improves interpretability but also lowers the computational load by restricting the number of extra cutpoints to those that are reasonable alternatives, and thereby reducing the complexity of the TWIX ensemble.

5.1.1 Instability of cutpoint selection in recursive partitioning

In standard binary splitting classification tree algorithms the cutpoint that produces the highest value of some split selection criterion, like the Gini gain, is selected. In a learning sample of size n, there are $n - 1$ potential cutpoints in each continuous variable without ties. Each of these candidate cutpoints defines two new daughter nodes. For cutpoint selection, within each daughter node an empirical entropy measure is computed. From these two individual node impurities of the daughter nodes an average impurity is derived and compared to the impurity in the mother node before splitting to assess the impurity reduction that can be achieved by splitting in each candidate cutpoint, as was described in detail in Chapter 1.

The distribution of the Gini gain split selection criterion over the range of the predictor

variable may, however, show that other cutpoints have a criterion value that is very similar to that of the best cutpoint, and thus might be equally well suited for splitting. This makes the trees very sensitive to small changes in the data set, because one of the other candidate cutpoints might have been chosen, if slightly different data had been observed. To judge whether such an instable situation is indeed present, a simple graphical visualization, the so-called "mountain plots" (Potapov, 2006; Potapov et al., 2006), can be very helpful: Mountain plots can be used to visualize the distribution of any split selection criterion over the range of the predictor variable as illustrated in Figure 5.1. The variables presented here are measurements of three different fatty acids from a data set on olives from different regions in Italy that comes with the `TWIX` add-on package (Potapov, 2007) to the R system for statistical computing (R Development Core Team, 2008).

The solid lines in the three plots illustrate the distribution of the Gini gain for a binary response over the range of the three predictor variables. Peaks in the mountain plots indicate good candidates for cutpoints. The $n-1$ potential cutpoints on the abscissa are ordered with respect to their value of the predictor variable X_j.

The first plot in Figure 5.1 shows a variable that produces one clear cutpoint, while the second and third plots show variables in the range of which several cutpoints are similarly well suited for splitting. The distributions of the binary response variable over the range of the predictor variable are displayed as circles in Figure 5.1. A clear distinction between the response classes leads to one clearly best cutpoint, while a high overlap between the response classes produces several similarly well suited cutpoints.

5.1.2 Currently implemented methods for selecting extra cutpoints in TWIX ensembles

From the mountain plots in Figure 5.1 it is obvious that in many cases the best cutpoint is only slightly better than the second best and so forth, and that small changes in the learning data may reverse the ranking of the cutpoints. Therefore it is reasonable to select a

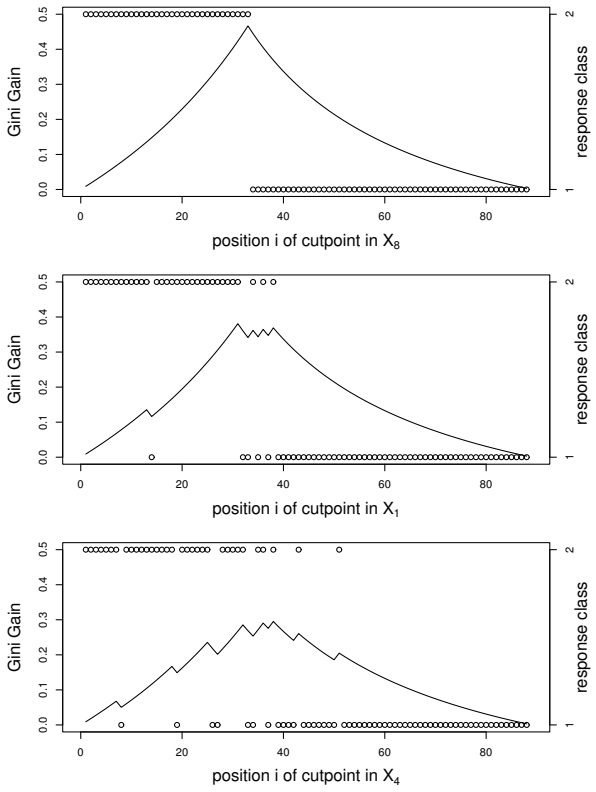


Fig. 5.1: Mountain plots: Distribution of the Gini gain and the binary response over the range of three predictor variables from the olives data set.

subset of extra cutpoints that appear similarly well suited for splitting as the best cutpoint for further branching, rather than to rely only on the best cutpoint in the learning sample. Different selection principles for this subset of suited cutpoints were outlined in Potapov (2006) and Potapov et al. (2006). The most obvious selection principles proposed by the authors are: (i) Select the best m cutpoints. (ii) Select the best m cutpoints that are local maxima.

Both selection principles select a fixed number of cutpoints m in each level of the tree. The implementation of TWIX in the R system for statistical computing also allows for different numbers of cutpoints m at different levels of the tree to account for the fact that at lower

levels of the tree fewer observations are left in each node producing less possible cutpoints.

A different approach for cutpoint selection, that aims at a different issue, is the grid selection principle. Here m cutpoints are selected at a given grid on the range of the predictor variable regardless of the distribution of the criterion. The rationale of the grid selection is that cutpoints that are optimal with respect to the current node (i.e. locally optimal) may not produce the globally most optimal tree. Therefore a wide range of possible cutpoints is used to produce a wide range of trees, from which the globally most optimal tree can then be selected. However, in continuous predictor variables this procedure is computationally extremely expensive and only manageable by means of parallel computing (Potapov, 2006; Potapov et al., 2006). In addition to this, even if cutpoints are selected by means of a grid search, variable selection in TWIX – as opposed to, e.g., random forests – is still limited to locally optimal choices.

Another issue in the TWIX cutpoint selection process is whether the best m cutpoints should be selected only within a previously chosen predictor variable or, inducing automatic variable selection, over all variables. Both options are available in TWIX. However, since the selection over all variables may produce variable selection bias for the reasons outlined in the previous chapters, here we will consider only the case that, in a first step, a predictor variable X_j is preselected for splitting by means of an unbiased association measure (like the ones suggested by, e.g., Strobl, 2005; Hothorn et al., 2006; Strobl et al., 2007, cp. Chapters 3 and 4), and in a second step suited cutpoints are selected only within the previously chosen variable.

5.2 A new, adaptive criterion for selecting extra cutpoints in TWIX ensembles

The currently implemented TWIX cutpoint selection principles usually select a given number m of cutpoints regardless of the underlying data. The first criterion selects, e.g., the

$m = 5$ cutpoints with the highest criterion values even in a clear cut situation as that of the first plot in Figure 5.1. The second criterion selects the $m = 5$ local maxima (or, if the number of local maxima is smaller than m, the number of extra cutpoints may be reduced to this number implicitly). This would allow the identification of the one clearly optimal cutpoint in the ideal situation depicted in the first plot in Figure 5.1, but would also enforce the selection of extra cutpoints in all $m = 5$ local maxima in the second plot in Figure 5.1, even if one of them is as far off the others as the one on the left hand side.

We will now suggest a new cutpoint selection criterion that directly operationalizes the possible instability to small changes in the learning sample. It adaptively selects a data driven number of cutpoints, namely only those that actually turn out to be (or remain) optimal when the original data set is exposed to such small changes.

5.2.1 Adding virtual observations

We start again by computing the Gini gain, or another split selection criterion like the Information gain based on the Shannon entropy, for each potential cutpoint $x_{(i)j}$ in the range of the preselected predictor variable X_j (as in Equation 1.3 in Chapter 1 or Equation 2.3 in Chapter 2). This gives us one optimal cutpoint with the highest criterion value – and maybe a few others that are similarly well suited.

Now imagine that we expose the original data set to small changes by introducing virtual, yet unobserved observations. The crucial question then is: Will the cutpoint that performed best on the original data set still perform best on the slightly changed data set, or will another cutpoint outperform the previously best one?

We assess the robustness of the cutpoint by means of successively adding virtual observations. Adding only one new observation might already lead to a different optimal cutpoint, but usually more than one virtual observation is necessary to actually induce a change. Therefore we successively add one, two and more new observations at each step of our algorithm: the current number of virtual observations s_{max} runs from 1 to an upper bound

s_{MAX}, as summarized in Table 5.1. An intuitive interpretation of the upper bound s_{MAX} is the number of unknown values of s_{MAX} subjects that, for some reason, could not participate in the original study even though they were supposed to. We will discuss the choice and meaning of s_{MAX} in more detail in Section 5.3.

In each step of the algorithm, the current number of s_{max} virtual observations are assigned either to the left or right node after the following rationale: For each yet unobserved observation (let us think of only one observation for a start) we know neither its value of the predictor X_j, nor of the response Y. However, in order to assign the observations to the left or right node, it is not even necessary to know the exact value of X_j of our unobserved observation. With respect to a given cutpoint $x_{(i)j}$ it is sufficient to discriminate between values of X_j smaller than or equal to $x_{(i)j}$, that would be assigned to the left node, and values of X_j greater than $x_{(i)j}$, that would be assigned to the right node.

Since the true value of X_j of our unobserved observation is unknown, however, we proceed with both options in our approach: For every potential cutpoint the unobserved observation is first assigned to the left node and the split criterion is recalculated; then the observation is assigned to the right node and the split criterion is recalculated again. This gives two best cutpoints, that are each either the same as in the original data set or new reasonable candidates for an extra split. For more than one virtual observation, this means that in each step of the algorithm we study the effect of s_{max} yet unobserved observations of which s_L are assigned to the left and $s_{max} - s_L$ to the right node, where s_L runs from 0 to s_{max}.

5.2.2 Recomputation of the split criterion

The unknown response class Y of our virtual observations, that is needed in the recomputation of the split criterion, is again incorporated by considering each possible response class. This leads to a set of class frequencies whose envelopes turn out to produce lower and upper possible class frequencies that coincide with the lower and upper class probabilities of a locally applied Imprecise Dirichlet Model (again abbreviated by IDM; Walley (1996)).

When, e.g., one virtual observation is assigned to the left node, the relative class frequency of response class 1 in this node is either $\frac{n_1(i)+1}{i+1}$, if the observation was of class 1, or $\frac{n_1(i)}{i+1}$, if the observation was of the other class. For a general number s_L of observations assigned to the left node we receive the lower and upper probabilities for class 1 given by the IDM with hyperparameter s_L, where

$$[\underline{\pi}_L(i, s_L), \overline{\pi}_L(i, s_L)] = \left[\frac{n_1(i)}{i + s_L} \, , \, \frac{n_1(i) + s_L}{i + s_L} \right]. \tag{5.1}$$

Note, however, that here the IDM is not understood as a meta model assumed in advance, but simply as a mathematical device that directly results from our method of robustification.

The interval-valued class probabilities derived in Equation 5.1 produce a set of Gini gains, and different criteria seem reasonable to select cutpoints from them. Here we will follow the rationale of a worst case scenario as in the minimax approach in decision theory: A single cutpoint is selected; namely the one that corresponds to the most conservative evaluation of the Gini gain. The lowest, and thus most conservative, evaluation of the Gini gain is produced by the distribution that is closest to the uniform distribution over the classes. This most conservative distribution of all distributions covered by the interval-valued class probabilities, $\pi_L^*(i, s_L)$, will be called upper entropy distribution (cp. Abellán and Moral, 2003, 2005) in the sequel. It can be derived by means of the upper entropy algorithm of Abellán and Moral (2003), that was originally developed for handling the Shannon entropy.

The criterion value for the left node is then calculated with the conservative upper entropy class probability $\pi_L^*(i, s_L)$ instead of the relative class frequency from the original sample. The criterion value for the right node is the same as for the original data set as long as no unobserved observation is assigned to the right node. In general, as shown by in the pseudo code in Table 5.1, zero or more, namely $s_{max} - s_L$, unobserved observations are assigned to the right node. From these $s_{max} - s_L$ observations a probability interval

$$[\underline{\pi}_R(i, s_{max} - s_L), \overline{\pi}_R(i, s_{max} - s_L)] =$$

$$\left[\frac{n_1 - n_1(i)}{i + (s_{max} - s_L)}\,,\ \frac{n_1 - n_1(i) + (s_{max} - s_L)}{i + (s_{max} - s_L)}\right]$$

and an upper entropy distribution $\pi_R^*(i, s_{max} - s_L)$ are derived as above. The Gini gain over both nodes is again computed as in Equation 2.3.

Overall up to s_{MAX} yet unobserved observations are assigned to either node to compute a conservative evaluation of the split criterion, and for each configuration the best cutpoint is evaluated. Some configurations will produce the same best cutpoint as the original sample, some may produce different but also well suited cutpoints, that can be used for further splitting. The algorithm is summarized in pseudo code in Table 5.1. It returns a vector $\widehat{\Delta G}$ of Gini gain values for each cutpoint in each configuration of s_{max} and s_L. In each configuration the best cutpoint is the one producing the highest Gini gain based on the upper entropy distributions. If the cutpoint of a certain configuration differs from previous optimal cutpoints, it is added to the list of cutpoints used for further splitting.

Tab. 5.1: Pseudo code for adaptive cutpoint selection.

for (s_{max} in $1 : s_{MAX}$){
 for (s_L in $0 : s_{max}$){
  for (i in $1 : (n-1)$){
   determine $\pi_L^*(i, s_L)$ and $\pi_R^*(i, s_{max} - s_L)$
   $\widehat{G_L^*}(i) = 2\ \pi_L^*(i, s_L)\,(1 - \pi_L^*(i, s_L))$
   $\widehat{G_R^*}(i) = 2\ \pi_R^*(i, s_{max} - s_L)\,(1 - \pi_R^*(i, s_{max} - s_L))$
   $\widehat{\Delta G}(i, s_L, s_{max}) = \widehat{G} - \left(\frac{i}{n}\widehat{G_L^*}(i) + \frac{n-i}{n}\widehat{G_R^*}(i)\right)$ } } }

It may be helpful to summarize explicitly that our robustification is applied only locally. In the last line of Table 5.1 this becomes most evident: The virtual observations are employed only in the conservative evaluation of the Gini indices $\widehat{G_L^*}(i)$ and $\widehat{G_R^*}(i)$ in the left and right nodes for a given cutpoint position i. The conservativeness of this evaluation is determined by the current value of s_{max}: the higher the number of virtual observations that can be

assigned to the left and right node, the wider the probability intervals produced by the IDM – and the more entropy is possible in the upper entropy distributions $\pi^*_L(i, s_L)$ and $\pi^*_R(i, s_{max} - s_L)$.

The new, virtual observations are introduced to locally manipulate the evaluation of a given cutpoint – but they are not supposed to be processed further through the tree, as the original observations. Consequently the weights $\frac{i}{n}$ for $\widehat{G^*_L}(i)$ and $\frac{n-i}{n}$ for $\widehat{G^*_R}(i)$, that represent the distribution of the original observations to the left and right of the potential cutpoint, are not altered by this "thought experiment".

Our procedure relies on the idea that the more observations can be newly assigned, the more likely it is that a cutpoint different from the one in the original sample will be optimal in some configuration. However, we will confirm in the simulation studies below that in situations where one cutpoint is clearly superior, as in the top plot in Figure 5.1, this cutpoint will remain superior, as desirable for a sensible data driven cutpoint selection method.

As a quite welcome by-product, our reasoning provides us directly with a robustness measure: The minimum number of newly assigned observations that is necessary to produce an optimal cutpoint different from that in the original sample can be used as a diagnostic of the robustness of the original split and will be referred to as s^* in the following. If s^* is small, i.e., if only few newly assigned observation are necessary to produce a different cutpoint, then the original cutpoint was not robust and most likely produced by random variations in the learning sample. This will also be illustrated in the next section with the olives data and the simulated data.

5.3 Behavior of the adaptive criterion

In this section we will show some applications of our cutpoint selection criterion to illustrate on one hand the interpretability of s^* as a robustness measure and on the other hand the

effects of different choices of s_{MAX} on the selection of extra cutpoints. First we will revisit the data set from Section 5.1.1 and after that we will use a simulation study to explore the behavior of the criterion for different choices of s_{MAX} more systematically.

5.3.1 Application to olives data

First of all let us revisit the three variables from the olives data set that were used in Figure 5.1. The data set consists of data on 89 olives from one out of two regions $Y = 1$ or $Y = 2$, that are supposed to be predicted from the measurements of three different fatty acids, X_8, X_1 and X_4, that produce the characteristic mountain plots in Figure 5.1. From the mountain plots we expect that the clear distinction of the response classes in X_8 results in one stable cutpoint, that is not easily changed by adding extra observations, while the less clear distinctions in X_1 and X_4 produce less stable cutpoints that may be easily affected by small changes in the learning sample.

For a first illustration, Table 5.2 gives the cutpoints found to be suited for further splitting by our robust selection criterion. The bold face typed cutpoints are the best ones in the original sample without assigning any new virtual observations. We find in the right column that the minimum number of newly assigned observations s^* necessary to produce a different cutpoint differs for the three variables exactly in the way that we expect from Figure 5.1: For X_1 only 3 newly assigned observations are enough to produce a different cutpoint for the first time, and only 5 newly assigned observations are enough to produce a different cutpoint in X_4. For X_8, however, it would be necessary to newly assign 58 observations, which would be 65% of the sample size, before a different cutpoint would be produced – which indicates that the cutpoint found in X_8 is very clear and robust against changes in the data.

The results in Table 5.2 can be compared to the mountain plots in Figure 5.1 by finding the positions i of the suited cutpoints on the abscissa.

From the top plot in Figure 5.1 it is obvious that the original cutpoint in X_8 can discrim-

Tab. 5.2: Cutpoints for the olives data. Position of original cutpoint (bold), position of the next cutpoint produced by adding virtual observation and robustness measure s^* (minimal number of additional observations needed to produce a different best cutpoint).

variable	position i of cutpoints	necessary for change s^*
X_8	**33**	
	65	58
X_1	**31**	
	38	3
X_4	**38**	
	32	5

inate perfectly between class 1 and 2, so that a great change in the data is necessary for producing another cutpoint to outperform it. Other examples of situations where cutpoints are more and less robust to data changes are given below in a simulation study.

The number of cutpoints found suited for splitting obviously depends on the strength of the association between the predictor variable and the response, and on the number of newly assigned observations in comparison to the original sample size. Thus, for a predictor variable such as X_8, that offers a single cutpoint that clearly dominates all other cutpoints in the range of that variable, this results indeed in a single split (and eventually in a single tree if other variables share this property of X_8) rather than a set of extra splits that are clearly suboptimal here and would start a new branch of the ensemble for any reasonable choice of s_{MAX}. This illustrates nicely how – determined by the characteristics of the underlying data set rather than the arbitrary choice of a hyperparameter as in the original version of TWIX – the application of our adaptive cutpoint selection criterion takes into account properly the extent of instability and produces a single tree as a special case of a TWIX ensemble, if the partition is sufficiently stable.

5.3.2 Simulation study

We now explore the interaction between the choice of s_{MAX} and characteristics of the data set more systematically in a simulation experiment with a fixed realistic sample size of $n = 200$ and a varying number of newly assigned observations s_{MAX}, as well as varying strength of association between the predictor variable and the response. The latter is achieved by generating a continuous predictor variable sampled from a standard normal distribution and determining the sampling probability of each response class of every observation by means of a logistic regression model with coefficient β varied. Overall, the logistic regression model was chosen such that it produces about 50% observations of class 1 and class 2 respectively. For large values of β we expect the two response classes to be clearly discriminated over the range of the predictor variable similar to the situation in the top plot of Figure 5.1, while for decreasing values of β we expect an increasing overlap as in the bottom plot of Figure 5.1, producing several similarly well suited cutpoints.

The average number of different cutpoints (over 100 simulation runs) found to be suited for splitting by our new criterion is displayed as a function of the logistic regression coefficient β and s_{MAX} in Table 5.3. We find again that the number of cutpoints can be regulated by the choice of the robustness parameter s_{MAX} and reflects the underlying distribution of the response on the range of the predictor variable, that is determined by means of the coefficient β.

This is quite a desirable property and distinguishes our data driven approach from the previous suggestions for finding extra splits, where the user is forced to determine an absolute number of cutpoints to proceed with in every level of the classification tree without being led by the data. In our approach the user only has to set the hyperparameter s_{MAX}, and the number of extra cutpoints for splitting is regulated depending on the underlying data. This means especially that when only few cutpoints, or even one as in variable X_8 of our example, are found reasonable for splitting, the branching will only proceed in these few cutpoints and the resulting ensemble will be much less complex than with

Tab. 5.3: Average number of different cutpoints for the simulated data based on a logistic model with parameter β varied and different maximal number s_{MAX} of virtual observations and two different sample sizes n.

		s_{MAX}				
n	β	2	5	10	20	50
100	0.5	1.65	2.21	3.16	6.29	13.21
	1.0	1.52	2.09	2.76	4.11	13.15
	1.5	1.36	1.89	2.66	3.47	12.61
	2.0	1.36	1.95	2.27	3.29	12.18
200	0.5	1.39	1.85	2.55	4.40	12.37
	1.0	1.33	1.84	2.33	3.00	9.45
	1.5	1.38	1.72	2.27	2.91	6.31
	2.0	1.29	1.79	2.29	3.04	4.66

previous approaches. Besides the considerable gain with respect to interpretation, the computational effort necessary for handling our criterion is also by far outweighted by the computational effort saved by not letting the ensemble grow exponentially in a fixed number of extra cutpoints.

In applications of our criterion the user only has to choose a reasonable value for s_{MAX}; for example, a certain percentage of the sample size. Intuitively we would suggest to choose an s_{MAX} of 5% to 10% of the original sample size n as a rule of thumb. This suggestion is led by the idea that in robust statistics it is often argued that about 5% of the original data set might consist of faulty observations, erroneous measurements and the like (cf. e.g. Hampel, 1980, who even cites historical data with up to 40% severe errors). These numbers justify an equal percentage of newly assigned observations for robustification. Another line of reasoning is that s_{MAX} should be chosen such as to represent the number of subjects who did not participate in the study even though they were supposed to, or otherwise lost observations in any sense. Different sizes of s_{MAX} might be reasonable and adequate

for different applications. In classification trees, where the number of observations available in successive nodes decreases rapidly due to splitting, the number of newly assigned observations s_{MAX} can also be adapted to this thinning process, i.e., chosen relative to the sample size in the current node. In this case we suggest that the user could set as a hyperparameter a certain percentage to compute s_{MAX} from the current sample size, that would then regulate the number of extra cutpoints.

5.4 Outlook on credal prediction and aggregation schemes

After suited cutpoints have been chosen for further splitting, a separate tree is grown with each of the cutpoints. Then either the best tree, with respect to some cross validation criterion, is chosen or the individual trees must be aggregated for a prediction in order to increase prediction accuracy.

In the following, we will shortly outline possible prediction and aggregation methods in the spirit of credal classification, that was already employed in the credal classification trees of Abellán and Moral (2005) treated in Chapter 4. For the sake of simplicity, we will only consider one-level trees, so-called stumps, here. The principles are, however, equally applicable to larger trees and the beneficial effects of selecting more than one cutpoint to the stability of the prediction will be even more pronounced for larger trees (Potapov, 2006; Potapov et al., 2006).

5.4.1 Credal prediction rules

In the standard classification tree algorithms C4.5 and CART the response class in each final node is predicted by a majority vote of all observations in that node: If the number of observations with response class 1 is higher than the number of observations with response

class 2, class 1 is predicted and vice versa.

In order to produce a cautious prediction, another option is the credal classification approach put forward by Zaffalon (2002b) and Abellán and Moral (2005). Credal classification does not necessarily return one predicted class, but may return a set of possible classes if the data do not contain enough information to justify a precise prediction.

For credal predictions in single trees, in each terminal node interval-valued probabilities for each class can be produced by means of, e.g., an IDM or the method of Coolen and Augustin (2008) based on nonparametric predictive inference, that forms a promising alternative to the IDM for processing relative frequencies for prediction.

When, as in the following, the IDM is employed to create probability intervals for each class, the width of the intervals – and thus the precision of the prediction – depends on the choice of the hyperparameter s of the IDM (cf. Abellán and Moral, 2005), that can again be interpreted as the number of yet unobserved observations. We arbitrarily chose $s = 1$ here, just to give an impression of the structure of the results and to illustrate how to proceed with aggregating in the next section.

For the olives data example the predictions for the three considered variables are displayed in Table 5.4 for each position i of cutpoints found suited for further splitting. The bold face typed cutpoint is again the best one in the original sample for each predictor variable. For the majority prediction the predicted response class is given, for the prediction based on the IDM the lower and upper probabilities for response class 1 is given. In the case of two response classes and probability intervals produced by the IDM considered here, the lower and upper probabilities for response class 2 follow directly by conjugacy. For more than two classes it would be more concise to report the set of plausible predictions.

We find that the coarse predictions of the majority votes are clear and do not differ in this example, even though they may in general differ for the different cutpoints. The more sensitive IDM predictions differ noticeably. For example the cutpoint at position 31 in predictor variable X_1 produces a very low upper probability for class 1 in the left node,

Tab. 5.4: Predictions from the olives data.

			prediction	
	i	node	majority	IDM
X_8	**33**	L	2	[0.0000, 0.0294]
		R	1	[0.9825, 1.0000]
X_1	**31**	L	2	[0.0312, 0.0625]
		R	1	[0.9322, 0.9492]
	38	L	2	[0.1282, 0.1538]
		R	1	[0.9808, 1.0000]
X_4	**38**	L	2	[0.1795, 0.2051]
		R	1	[0.9423, 0.9615]
	32	L	2	[0.1212, 0.1515]
		R	1	[0.8966, 0.9138]

while the cutpoint at position 38 produces a slightly higher upper probability for class 1 in the left node. This is an indicator that some observations with response class 1 may be situated to the right of position 31 but to the left of position 38.

From the probability intervals a set of predictions can be generated – either by means of the strong dominance criterion as in Abellán and Moral (2005), or more generally by any other criterion inducing a partial interval ordering (cf., e.g., Chapter 2.6 of Weichselberger, 2001; Troffaes, 2007). When no single dominant class can be identified, the set of all non-dominated classes is returned. Therefore, the credal classification strategy, that provides a set of plausible response classes when the available information does not justify a clear decision for one class, is especially beneficial in problems with more than two response classes.

Tab. 5.5: Aggregated interval-valued predictions from the olives data.

		aggregation rule	
	node	outer	mean
X_1	L	[0.0312, 0.1538]	[0.0797, 0.1082]
	R	[0.9322, 1.0000]	[0.9565, 0.9746]
X_4	L	[0.1212, 0.2051]	[0.1503, 0.1783]
	R	[0.8966, 0.9615]	[0.9194, 0.9377]

5.4.2 Aggregation schemes

For the aggregation of the predictions of several classification trees majority voting is used in standard ensemble methods, as was outlined in the introduction. For credal predictions based on imprecise probabilities, aggregation rules are less obvious and, motivated by different applications, several authors have made different suggestions (cf., e.g., Moral and del Sagrado, 1987; Walley, 1991; Weichselberger, 2001; de Cooman and Troffaes, 2004; Bronevich, 2005; Troffaes, 2006), that could fruitfully be transferred to ensemble methods.

Reasonable first approaches to be considered here are conjunction, disjunction and the mean lower and upper probabilities. We found that different cutpoints often produce conflicting probability intervals, so that mere conjunction is often not possible. Therefore we only display results for the disjunction and mean approaches here. However, in the case of conflicting information from the individual trees it could be reasonable to place (imprecise) weights on the individual trees in a way comparable to the approach of Troffaes (2006), who assigns different imprecise "trust" values to conflicting experts. Regarding ensembles of classification trees, the trust value of each tree could be chosen according to some cross validated performance measure.

From our first results it looks like the disjunction approach might be too conservative for our purpose, because it does not reflect the property of other ensemble methods that the aggregated predictions from sets of trees show less deviation than the predictions of

individual trees. The mean approach is a naive but sound first attempt. Evidently further research on adequate aggregation rules for predictions from sets of classification trees is needed.

5.5 Summary

Our aim here was to give a first impression of the potential of an adaptive cutpoint selection approach. Based on the general idea to address robustness issues by studying the effect of adding some virtual observations, we proposed a new adaptive cutpoint selection criterion in this chapter. Its main advantages are (i) that instead of a fixed number of extra cutpoints, which deterministically lead to an exponential growth of the ensemble, an adaptive number of cutpoints is selected for further splitting, (ii) the approach is data driven so that (iii) the user does not have to fix a certain number of cutpoints in advance, but only an intuitively interpretable hyperparameter, that implicitly regulates the number of cutpoints. Finally, (iv) the size of a TWIX ensemble resulting from adaptive cutpoint selection for a particular data set can be used as a diagnostic when considering the tradeoff between the interpretability of a single tree model against the high prediction accuracy of black box ensemble methods.

The general idea to introduce virtual observations in order to robustify cutpoint selection could be transferred to a variety of other applications, including optimally selected thresholds in diagnostic tests.

With respect to computational complexity the use of the adaptive split selection criterion is computationally expensive, but can drastically reduce the number of extra cutpoints to those that are robust to small changes in the learning data and therefore reasonably suited for further splitting. Thus, its expense is outweighted by the computational complexity saved by avoiding an exponential growth of the entire ensemble in an exponent m that is fixed a priori.

For credal predictions from sets of classification trees, our first results show that different aggregation rules may prove beneficial and should be further investigated.

The presentation was limited to continuous predictor variables and a binary response here, but the method is generalizable straightforwardly to deal with ordinal and categorical predictor variables as well as problems with more than two response classes. In the latter case the impact of credal classification would be even more beneficial, as outlined above. The cautious treatment of missing values or coarse data in the spirit of Zaffalon (2002a) (see also de Cooman and Zaffalon, 2004; Zaffalon, 2005) as well as, with an emphasis on the IDM, Utkin and Augustin (2007), could also be embedded directly in our approach.

With respect to interpretability, the main advantage of the adaptive approach is that the size of the adaptive TWIX ensemble is determined by the underlying data set and can be taken into account when judging whether an interpretable single classification tree is sufficient, or if a complex ensemble method is necessary for analyzing the particular data set: In extreme cases with clearly dominating cutpoints in each split the adaptive TWIX ensemble will collapse to a single tree model that can be interpreted without hesitation, because it has proven its stability to changes in the learning data. On the other hand, an adaptive TWIX ensemble that branches very widely indicates that the data set cannot be analyzed adequately with a single interpretable tree model, because several competing cutpoints are employed in the branching process to compensate a high level of instability. In this case parallelized ensemble methods like bagging and random forests, that allow for more diverse sets of trees, are likely to provide better prediction accuracy. For these methods, that offer no straightforward means of interpretation, measures for evaluating the importance of each predictor variable are investigated in the following chapters.

6. Unbiased variable importance in random forests and bagging

In the remaining chapters we will turn to the ensemble methods bagging and random forests, where a non-nested set of classification trees is constructed, usually from bootstrap samples. The resulting set of trees cannot be combined into one interpretable model. Therefore, in order to be able to assess the impact of each predictor variable in the model, different variable importance measures have been suggested, that may also be employed to discriminate the subset of relevant predictors from the remaining noise variables.

However, we will find not only that the variable selection bias that is inherent in standard single classification trees based on impurity criteria is carried over to ensembles of trees and their variable importance measures, but also that new sources of bias in favor of variables of certain types are induced by the resampling scheme employed in tree construction and the permutation scheme employed in the computation of one popular variable importance measure.

The scope of this chapter is to show that the variable importance measures of the original random forest method (Breiman, 2001a), based on CART classification trees (Breiman et al., 1984), are a sensible means for variable selection in many applications, but are not reliable in situations where potential predictor variables vary in their scale of measurement or their number of categories – as is often the case in genomics, bioinformatics and related disciplines, where both genetic and environmental variables, individually and in interactions, are considered as potential predictors or predictor variables of the same type vary

in the number of categories present in a certain sample.

We will illustrate in the following simulation studies that variable selection with the variable importance measure of the original random forest method bears the risk that suboptimal predictor variables are artificially preferred in such scenarios and provide statistical explanations for this deficiency of the variable importance measures of the original random forest method.

Based on these statistical explanations we propose to employ an alternative, unbiased random forest method, and to build ensembles of trees based on subsamples, rather than bootstrap samples. The performance of this approach is compared to that of the original random forest method in simulation studies, and is illustrated by an application to the prediction of C-to-U edited sites in plant mitochondrial RNA, re-analyzing the data of Cummings and Myers (2004) that were previously analyzed with the original random forest method.

6.1 Random forest variable importance measures

A naive variable importance measure to use in tree-based ensemble methods would be to merely count the number of times each variable is selected by all individual trees in the ensemble.

More elaborate variable importance measures incorporate a (weighted) mean of the individual trees' improvement in the splitting criterion produced by each variable (Friedman, 2001). An example for such a measure in classification is the "Gini importance" available in random forest implementations. The Gini importance describes the improvement in the Gini gain splitting criterion.

The most advanced variable importance measure available in random forests is the "permutation accuracy importance" measure (termed "permutation importance" hereafter). Its rationale is the following: By means of randomly permuting the predictor variable X_j

by some permutation ψ_j, its original association with the response Y is broken. When the permuted variable X_j, together with the remaining non-permuted predictor variables, is used to predict the response for the out-of-bag observations, the prediction accuracy (i.e. the number of observations classified correctly) decreases substantially if the original variable X_j was associated with the response. Thus, Breiman (2001a) suggests the difference in prediction accuracy before and after permuting X_j, averaged over all trees, as a measure for variable importance, that we formalize as follows: Let $\overline{\mathfrak{B}}^{(t)}$ be the out-of-bag sample for a tree t, with $t \in \{1, \ldots, ntree\}$. Then the variable importance of variable X_j in tree t is

$$VI^{(t)}(\mathbf{X}_j) = \frac{\sum_{i \in \overline{\mathfrak{B}}^{(t)}} I\left(y_i = \hat{y}_i^{(t)}\right)}{\left|\overline{\mathfrak{B}}^{(t)}\right|} - \frac{\sum_{i \in \overline{\mathfrak{B}}^{(t)}} I\left(y_i = \hat{y}_{i,\psi_j}^{(t)}\right)}{\left|\overline{\mathfrak{B}}^{(t)}\right|} \tag{6.1}$$

where $\hat{y}_i^{(t)} = f^{(t)}(\mathbf{x}_i)$ is the predicted class for observation i before and $\hat{y}_{i,\psi_j}^{(t)} = f^{(t)}(\mathbf{x}_{i,\psi_j})$ is the predicted class for observation i after permuting its value of variable X_j, i.e. with $\mathbf{x}_{i,\psi_j} = \left(x_{i,1}, \ldots, x_{i,j-1}, x_{\psi_j(i),j}, x_{i,j+1}, \ldots, x_{i,p}\right)$. (Note that $VI^{(t)}(\mathbf{X}_j) = 0$ by definition, if variable X_j is not in tree t.) The raw variable importance score for each variable is then computed as the average importance over all trees

$$VI(\mathbf{X}_j) = \frac{\sum_{t=1}^{ntree} VI^{(t)}(\mathbf{X}_j)}{ntree}. \tag{6.2}$$

From this raw importance score a standardized importance score, also called "z-score", can be computed with the following rationale: The individual importance scores $VI^{(t)}(\mathbf{x}_j)$ are computed from *ntree* bootstrap samples, that are independent given the original sample, and are identically distributed. Thus, if each individual variable importance $VI^{(t)}$ has standard deviation σ, the average importance from *ntree* replications has standard error $\sigma/\sqrt{ntree}$. The standardized or scaled importance is then computed as

$$\widetilde{VI}(\mathbf{x}_j) = \frac{VI(\mathbf{x}_j)}{\frac{\hat{\sigma}}{\sqrt{ntree}}}. \tag{6.3}$$

When the central limit theorem is applied to the mean importance $VI(\mathbf{x}_j)$, Breiman and Cutler (2008) argue that the z-score is asymptotically normal. This property will be used

explicitly for the statistical test that is critically investigated in Chapter 7. In this current chapter, however, we focus on the properties of the importance scores as purely descriptive measures of variable importance. In the simulation studies presented in the next section, we compare the selection frequency, the Gini importance and the permutation importance for different base learners and different resampling schemes.

6.2 Simulation studies

The reference implementation of the original random forest method of Breiman (2001a) is available in the R system for statistical computing (R Development Core Team, 2008) via the `randomForest` add-on package by Breiman et al. (2006) (cf. Liaw and Wiener, 2002, for an introduction). The behavior of the selection frequency, the Gini importance and the permutation importance of the `randomForest` function is explored in a simulation design where potential predictor variables vary in their scale of measurement and number of categories, because we know from the previous chapters that this setting induces variable selection bias in the individual trees.

As an alternative, we propose to use the alternative random forest function `cforest` available in the R add-on package `party` (Hothorn et al., 2008, 2006) in such scenarios. In contrast to `randomForest`, the `cforest` function creates random forests not from CART classification trees based on the Gini split criterion, that are known to prefer variables with, e.g., more categories in variable selection (cf. Breiman et al., 1984; Kononenko, 1995; Kim and Loh, 2001; Boulesteix, 2006b,a; Strobl et al., 2007, and Chapters 2 and 3), but from unbiased classification trees based on the conditional inference framework of Hothorn et al. (2006).

Since the `cforest` function does not employ the Gini criterion, we investigate the behavior of the Gini importance for the `randomForest` function only. The selection frequency and the permutation importance is studied for both functions `randomForest` and `cforest` in two ways: Either the individual trees are built on bootstrap samples of the original sample

size n drawn with replacement, as suggested by Breiman (2001a), or on subsamples drawn without replacement.

Subsampling as an alternative to bootstrap sampling in aggregating, e.g., individual classification trees is investigated further by Bühlmann and Yu (2002), who also coin the term "subagging" as an abbreviation for "subsample aggregating" as opposed to "bagging" for "bootstrap aggregating". Politis et al. (1999) show that, for statistical inference in general, subsampling works under weaker assumptions than bootstrap sampling and even in situations when bootstrap sampling fails. The subsample size here is set to 0.632 times the original sample size n, because in bootstrap sampling with replacement about 63.2% of the data end up in the bootstrap sample. Other fractions for the subsample size are possible, as discussed in the end of this chapter.

Tab. 6.1: The predictor variables are sampled independently from the following distributions. $N(0,1)$ stands for the standard normal distribution, $M(k)$ stands for the multinomial distribution with values in $\{0, \ldots, k-1\}$ and equal probabilities (discrete uniform distribution on $\{0, \ldots, k-1\}$), $B(\pi)$ stands for the binomial (Bernoulli) distribution with probability π, thus $M(2)$ equals $B(0.5)$.

Predictor variables	
X_1	$\sim N(0,1)$
X_2	$\sim M(2)$
X_3	$\sim M(4)$
X_4	$\sim M(10)$
X_5	$\sim M(20)$

The simulation design used throughout this chapter represents a scenario where a binary response variable Y is supposed to be predicted from a set of potential predictor variables that vary in their scale of measurement and number of categories. The first predictor variable X_1 is continuous, while the other predictor variables $X_2, \ldots, X_5$ are categorical

Tab. 6.2: The response variable is sampled from binomial (Bernoulli) distributions. The degree of dependence between the response Y and X_2 is regulated by the probability π of the binomial distribution $B(\pi)$ of Y conditional on X_2, with the *relevance* parameter taking values in $\{0.05, 0.1, 0.15, 0.2\}$ to model different degrees of dependence.

	Response variable	
null case	Y	$\sim B(0.5)$
power case	$Y\|X_2 = 1$	$\sim B(0.5 - \textit{relevance})$
	$Y\|X_2 = 2$	$\sim B(0.5 + \textit{relevance})$

(on a nominal scale of measurement) with their number of categories between two and up to twenty. The simulation designs of both studies are summarized in Table 6.1 and 6.2. The sample size for all simulation studies was set to $n = 120$.

In the first simulation study, the so-called null case, none of the predictor variables is informative for the response, i.e., all predictor variables and the response are sampled independently. In this situation a sensible variable importance measure should not prefer any one predictor variable over any other.

In the second simulation study, the so-called power case, the predictor variable X_2 is informative for the response, i.e., the distribution of the response depends on the value of this predictor variable. The degree of dependence between the informative predictor variable X_2 and the response Y is regulated by the *relevance* parameter of the conditional distribution of Y given X_2 (cf. Table 6.2). We will later display results for different values of the *relevance* parameter indicating different degrees of dependence between X_2 and Y.

In the power case, a sensible variable importance measure should be able to distinguish the informative predictor variable from its uninformative competitors, and even more so with increasing degree of dependence.

Our simulation studies show that for the `randomForest` function all three variable impor-

tance measures are unreliable, and the Gini importance is most strongly biased. For the `cforest` function reliable results can be achieved both with the selection frequency and the permutation importance if the function is used together with subsampling without replacement. Otherwise the measures are biased as well.

6.2.1 Results of the null case simulation study

In the null case, when all predictor variables are equally uninformative, the selection frequencies as well as the Gini importance and the permutation importance of all predictor variables are supposed to be equal.

However, as presented in Figure 6.1, the average selection frequencies (over 1000 simulation runs) of the predictor variables differ substantially when the `randomForest` function (cf. top row in Figure 6.1) or the `cforest` function with bootstrap sampling (cf. bottom row, left plot in Figure 6.1) are used. Variables with more categories are obviously preferred. Only when the `cforest` function is used together with subsampling without replacement (cf. bottom row, right plot in Figure 6.1) are the variable selection frequencies for the uninformative predictor variables equally low as desired.

It is obvious that variable importance cannot be represented reliably by the selection frequencies, that can be considered as very basic variable importance measures, if the potential predictor variables vary in their scale of measurement or number of categories when the `randomForest` function or the `cforest` function with bootstrap sampling is used.

The average Gini importance (over 1000 simulation runs), that is displayed in Figure 6.2, is biased even stronger. Like the selection frequencies for the `randomForest` function (cf. top row in Figure 6.1) the Gini importance shows a strong preference for variables with many categories and the continuous variable, as expected from the bias in favor of variables offering many cutpoints in single trees, that was explored in previous chapters. We conclude that the Gini importance cannot be used to reliably measure variable importance in this situation either.

We now consider the more advanced permutation importance measure. We find that here an effect of the scale of measurement or number of categories of the potential predictor variables is less obvious but still severely affects the reliability and interpretability of the variable importance measure.

Figure 6.3 shows boxplots of the distributions (over 1000 simulation runs) of the permutation importance measures of both functions for the null case. The plots in the top row again display the distribution when the `randomForest` function is used, the bottom row when the `cforest` function is used. The left column of plots displays the distributions when bootstrap sampling is conducted with replacement, while the right column displays the distributions when subsampling is conducted without replacement.

Figure 6.4 shows boxplots of the distributions of the scaled version of the permutation importance measures of both functions, incorporating the standard deviation of the measures.

The scaled variable importance is the default output of the `randomForest` function. However, it has been noted, e.g., by Diaz-Uriarte and Alvarez de Andrés (2006) in their supplementary material, that the magnitude of the scaled variable importance of the `randomForest` function depends on the number of trees grown in the random forest. This issue is further investigated in the next chapter. Meanwhile, we suggest not to interpret the absolute magnitude of the scaled variable importance of the `randomForest` function.

The plots show that for the `randomForest` function (cf. top row in Figures 6.3 and 6.4) and, less pronounced, for the `cforest` function with bootstrap sampling (cf. bottom row, left plot in Figures 6.3 and 6.4), the deviation of the permutation importance measure over the simulation runs is highest for the variable X_5 with the highest number of categories, and decreases for the variables with less categories and the continuous variable. This effect is weakened but not substantially altered by scaling the measure (cf. Figure 6.3 vs. Figure 6.4).

As opposed to the obvious effect in the selection frequencies and the Gini importance, there

is no effect in the mean values of the distributions of the permutation importance measures, which are in average close to zero as expected for uninformative variables. However, the notable differences in the variance of the distributions for predictor variables with different scale of measurement or number of categories seriously affect the expressiveness of the variable importance measure.

In a single trial this effect may lead to a severe over- or underestimation of the variable importance of variables that have more categories as an artefact of the method, even though they are no more or less informative than the other variables.

Only when the `cforest` function is used together with subsampling without replacement (cf. bottom row, right plot in Figures 6.3 and 6.4) does the deviation of the permutation importance measure over the simulation runs not increase substantially with the number of categories or scale of measurement of the predictor variables.

Thus, only the variable importance measure available in `cforest`, and only when used together with sampling without replacement, reliably reflects the true importance of potential predictor variables in a scenario where the potential predictor variables vary in their scale of measurement or number of categories.

6.2.2 Results of the power case simulation study

In the power case, where only the predictor variable X_2 is informative, a sensible variable importance measure should be able to distinguish the informative predictor variable.

The following figures display the results of the power case with the highest value 0.2 of the *relevance* parameter, indicating a high degree of dependence between X_2 and the response. In this setting, each of the variable importance measures should clearly prefer X_2, while the respective values for the remaining predictor variables should be equally low.

Figure 6.5 shows that the average selection frequencies (again over 1000 simulation runs) of the predictor variables again differ substantially when the `randomForest` function (cf.

top row in Figure 6.5) is used, and the relevant predictor variable X_2 cannot be identified. With the `cforest` function with bootstrap sampling (cf. bottom row, left plot in Figure 6.5) there is still bias obvious in the selection frequencies of the categorical predictor variables with many categories. Only when the `cforest` function is used together with subsampling without replacement (cf. bottom row, right plot in Figure 6.5), the variable selection frequencies for the uninformative predictor variables are equally low as desired, and the value for the relevant predictor variable X_2 sticks out.

The average Gini importance, that is displayed in Figure 6.6, again shows a strong bias towards variables with many categories and the continuous variable. It completely fails to identify the relevant predictor variable, with the mean value for the relevant variable X_2 only slightly higher than in the null case.

Figures 6.7 and Figure 6.8 show boxplots of the distributions of the unscaled and scaled permutation importance measures of both functions. Again for the `randomForest` function (cf. top row in Figures 6.7 and Figure 6.8) and, less pronounced, for the `cforest` function with bootstrap sampling (cf. bottom row, left plot in Figures 6.7 and Figure 6.8), the deviation of the permutation importance measure over the simulation runs is highest for the variable X_5 with the highest number of categories, and decreases for the variables with less categories and the continuous variable. Again, this effect is weakened but not substantially altered by scaling the measure (cf. Figure 6.7 vs. Figure 6.8).

As expected, the mean value of the permutation importance measure for the informative predictor variable X_2 is higher than for the uninformative variables. However, the deviation of the variable importance measure for the uninformative variables with many categories X_4 and X_5 is so high that in a single trial these uninformative variables may outperform the informative variable as an artefact of the method.

Thus, only the variable importance measure computed with the `cforest` function, and only when used together with sampling without replacement, is able to reliably detect the informative variable out of a set of uninformative competitors, even if the degree of

dependence between X_2 and the response is high.

Tab. 6.3: Rates of correct identifications of the informative variable with the scaled and unscaled permutation importance of the `randomForest` method, applied with sampling with and without replacement, as compared to those of the `cforest` method, applied with sampling with and without replacement, as a function of the degree of dependence (indicated by the *relevance* parameter, cf. Table 2) between the informative variable X_2 and the response. (Standard errors of the rates of correct identifications r over 1000 iterations can easily be computed by $se = \sqrt{r \cdot (1 - r) \,/\, 1000}$.)

			Degree of dependence			
	Method	Repl.	0.05	0.1	0.15	0.2
Scaled	`randomForest`	true	0.234	0.497	0.770	0.956
		false	0.237	0.489	0.760	0.949
	`cforest`	true	0.338	0.672	0.923	0.991
		false	0.365	0.728	0.943	0.994
Unscaled	`randomForest`	true	0.194	0.413	0.701	0.928
		false	0.186	0.400	0.710	0.919
	`cforest`	true	0.324	0.648	0.910	0.989
		false	0.370	0.729	0.943	0.994

The rate at which the informative predictor variable is correctly identified (by producing the highest value of the permutation importance measure) increases with the degree of dependence between X_2 and the response. In Table 6.3 the rates of correct identifications (over 1000 simulation runs) for four different degrees of dependence between X_2 and the response are summarized for the `randomForest` and `cforest` functions with different options.

For all degrees of dependence between X_2 and the response Y the `cforest` function detects the informative variable more reliably than the `randomForest` function, and the `cforest`

function used with subsampling without replacement outperforms the `cforest` function with bootstrap sampling with replacement.

For the `randomForest` function scaling the permutation importance measure can slightly increase the rates of correct identifications because, as shown in Figures 6.4 and 6.8, scaling weakens the differences in variance of the permutation importance measure for variables of different scale of measurement and number of categories. For the `cforest` function, that is not affected by the scale of measurement and number of categories of the predictor variables, both the unscaled and the scaled permutation importance perform equally well.

In addition to its superiority in the assessment of variable importance the `cforest` method, especially when used together with subsampling without replacement, can also be superior to the `randomForest` method with respect to classification accuracy in situations like that of the power case simulation study, where uninformative predictor variables with many categories "fool" the `randomForest` function.

Tab. 6.4: Average misclassification rates of the `randomForest` method, applied with sampling with and without replacement, as compared to those of the `cforest` method, applied with sampling with and without replacement, as a function of the degree of dependence (indicated by the *relevance* parameter, cf. Table 2) between the informative variable X_2 and the response. (Standard errors of the average misclassification rates are given in parentheses.)

		Degree of dependence			
Method	Repl.	0.05	0.1	0.15	0.2
`randomForest`	true	0.4945 (0.0014)	0.4819 (0.0015)	0.4510 (0.0016)	0.4028 (0.0017)
	false	0.4942 (0.0014)	0.4814 (0.0015)	0.4496 (0.0016)	0.4026 (0.0017)
`cforest`	true	0.4910 (0.0014)	0.4660 (0.0016)	0.4169 (0.0019)	0.3491 (0.0019)
	false	0.4879 (0.0014)	0.4581 (0.0017)	0.4022 (0.0019)	0.3384 (0.0019)

Due to its artificial preference for uninformative predictor variables with many categories the `randomForest` function can produce a higher average misclassification rate than the `cforest` function. The average misclassification rates (again over 1000 simulation runs) for the `randomForest` and `cforest` function, again for four different degrees of dependence and used with sampling with and without replacement, are displayed in Table 6.4.

Each method was applied to the same simulated test set in each simulation run. The test sets were generated from the same data generating process as the learning sets. We find that for all degrees of dependence between X_2 and the response Y the `cforest` function, especially with sampling without replacement, outperforms the other methods. A similar result is obtained in the application to C-to-U conversion data presented in the next section.

The differences in classification accuracy are moderate in the latter case. However, one could think of more extreme situations that would produce even greater differences. This shows that the same mechanisms underlying the variable importance bias can also affect the classification accuracy, e.g. when suboptimal predictor variables, that do not add to the classification accuracy, are artificially preferred in variable selection merely because they have more categories.

6.3 Sources of variable importance bias

The main difference between the `randomForest` function, based on CART trees (Breiman et al., 1984), and `cforest` function, based on conditional inference trees (Hothorn et al., 2006), is that in `randomForest` the variable selection in the individual CART trees is biased, so that ,e.g., variables that offer more potential cutpoints are preferred, as described in the earlier chapters of this work. Consequences of the variable selection bias, that is inherent in each single tree, on the variable importance measures of the entire ensemble are pointed out in the next section.

However, even if the individual trees select variables in an unbiased way as in the `cforest`

function, we find that the variable importance measures, as well as the selection frequencies of the variables, are affected by the bootstrap sampling with replacement. This is explained in the section on effects induced by bootstrapping.

6.3.1 Variable selection bias in individual classification trees

The variable selection bias that occurs in every individual tree of an ensemble produced with the `randomForest` function has a direct effect on its variable importance measures: Predictor variables with more categories are artificially preferred in variable selection in each splitting decision. Thus, they are selected in more individual classification trees and tend to be situated closer to the root node in each tree.

This affects the variable importance measures in two respects. Firstly, the variable selection frequencies over all trees are directly affected by the variable selection bias in each individual tree. Secondly, an effect on the permutation importance occurs, that is less obvious but just as severe: When permuting the variables to compute their permutation importance measure, the variables that appear in more trees and are situated closer to the root node can affect the prediction accuracy of a larger set of observations, while variables that appear in fewer trees and are situated closer to the bottom nodes affect only small subsets of observations. Thus, the range of possible changes in prediction accuracy in the random forest, i.e., the deviation of the variable importance measure, is higher for variables that are preferred by the individual trees due to variable selection bias.

We found in Figures 1 through 9, that the effects induced by the different types of predictor variables were more pronounced for the `randomForest` function, where variable selection in the individual trees is biased, than for the `cforest` function, where the individual trees are unbiased. However, we also found that when the `cforest` function is used with bootstrap sampling, the variable selection frequencies of the categorical predictors still depend on their number of categories (cf., e.g., bottom row, left plot in Figure 6.1), and also the deviation of the permutation importance measure is still affected by the number

of categories (cf., e.g., bottom row, left plot in Figures 3 and 4).

Thus, there must be another source of bias, besides the variable selection bias in the individual trees, that affects the selection frequencies and the deviation of the permutation importance measure.

We show in the next section that this additional effect is due to bootstrap sampling with replacement, that is traditionally employed in random forests.

6.3.2 Effects induced by bootstrapping

In the comparison of left and right columns (representing sampling with and without replacement) in Figures 1 and 5 we could illustrate that the variable selection frequencies in random forest functions are affected by the resampling scheme.

Even when the `cforest` function based on unbiased classification trees is used, variables with more categories are preferred when bootstrap sampling is conducted with replacement, while no bias occurs when subsampling is conducted without replacement, as displayed in the bottom right plot in Figures 1 and 5. Thus, the bootstrap sampling induces an effect that is more pronounced for predictor variables with more categories.

For a better understanding of the underlying mechanism let us consider only the categorical predictor variables X_2 through X_5 with different numbers of categories from the null case simulation study design. Rather than trying to explain the effect of bootstrap sampling in the complex framework of random forests, we use a much simpler and more familiar independence test for the explanation.

We consider the p-values of χ^2-tests (computed from 1000 simulated data sets). In each simulation run, a χ^2-test is computed for each predictor variable and the binary response Y. Remember that the variables in the null case are not informative, i.e., the response is independent of all variables.

For independent variables it follows from reversing the inversion method that the distrib-

ution of the p-values of the χ^2-test forms a uniform distribution.

The left plot in Figure 6.11 displays the distribution of the p-values of χ^2-tests from each predictor variable and the response Y as boxplots. We find that the boxplots range from 0 to 1 with median 0.5, because the p-values of the χ^2-test form a uniform distribution when computed before bootstrapping, as expected under the null hypothesis.

However, if in each simulation run we draw a bootstrap sample from the original sample and then again compute the p-values based on the bootstrap sample, we find that the distribution of the p-values is shifted towards zero as displayed in the right plot in Figure 6.11.

Obviously, the bootstrap sampling artificially induces an association between the variables. This effect is always present when statistical inference, such as an association test, is carried out on bootstrap samples: Bickel and Ren (2001) point out that bootstrap hypothesis testing fails whenever the distribution of any statistic in the bootstrap sample, rather than the distribution of the statistic under the null hypothesis, is used for statistical inference. We found that this issue directly affects variable selection in random forests, because the deviation from the null hypothesis is more pronounced for variables that have more categories.

The reason for the shift in the distribution of the p-values displayed in Figure 6.11 is that each original sample, even if sampled from theoretically independent distributions, may show some minor variations from the null hypothesis of independence. These minor variations are aggravated by bootstrap sampling with replacement, because the cell counts in the contingency table are affected by observations that are either not included or are doubled or tripled in the bootstrap sample, and therefore the bootsrap sample deviates notably from the null hypothesis – even if the original sample was generated under the null hypothesis.

This effect is more pronounced for variables with more categories, because in larger tables (such as the 4×2 table from the cross-tabulation of X_3 and the binary response Y), the

absolute cell counts are smaller than in smaller tables (such as the 2×2 table from the cross-tabulation of X_2 and the binary response Y). With respect to the smaller absolute cell counts, excluding or duplicating an observation produces more severe variations from the null hypothesis.

This effect is not eliminated if the sample size is increased, because in bootstrap sampling the size n of the original sample and the bootstrap sample size n increase simultaneously. However, if subsamples are drawn without replacement the effect disappears.

The apparent association that is induced by bootstrap sampling, and that is more pronounced for predictor variables with many categories, affects both variable importance measures: The selection frequency is again directly affected, and the permutation importance is affected because variables with many categories are selected more often and gain positions closer to the root node in the individual trees. Together with the mechanisms described in the previous section, this explains our findings.

From our simulation results we can see, however, that the effect of bootstrap sampling is mostly superposed by the much stronger effect of variable selection bias when comparing the conditions of sampling with and without replacement for the `randomForest` function only (cf. Figures 1 through 9, top row). Only when variable selection bias is removed by the `cforest` function, the differences between the conditions of sampling with and without replacement are obvious (cf. Figures 1 through 9, bottom row).

We therefore conclude that in order to be able to reliably interpret the variable importance measures of a random forest, the forest must be built from unbiased classification trees, and sampling must be conducted without replacement.

6.4 Application to C-to-U conversion data

RNA editing is the process whereby RNA is modified from the sequence of the corresponding DNA template (Cummings and Myers, 2004). For instance, cytidine-to-uridine conver-

sion (abbreviated C-to-U conversion) is common in plant mitochondria. The mechanisms of this conversion remain largely unknown, although the role of neighboring nucleotides is emphasized. Cummings and Myers (2004) suggest to use information from sequence regions flanking the sites of interest to predict editing in *Arabidopsis thaliana*, *Brassica napus* and *Oryza sativa* based on random forests. The *Arabidopsis thaliana* data of Cummings and Myers (2004) can be loaded from the journal's homepage. For each of the 876 observations, the data set gives

- the response at the site of interest (binary: edited/not edited)

and as potential predictor variables

- the 40 nucleotides at positions -20 to 20, relative to the edited site (4 categories),
- the codon position (4 categories),
- the estimated folding energy (continuous) and
- the difference in estimated folding energy between pre-edited and edited sequences (continuous).

We first derive the permutation importance measure for each of the 43 potential predictor variables with each method. As can be seen from the barplot in Figure 6.9, the (scaled) variable importance measures largely reflect the results of Cummings and Myers (2004) based on the Gini importance measure, but differ slightly for the `randomForest` and `cforest` function and the different resampling schemes. In particular, the variable importance measure of the `randomForest` function seems to produce more "noise" than that of the `cforest` function: the contrast of amplitudes between irrelevant and relevant predictors is more pronounced when the `cforest` function is used.

Note, however, that the permutation importance values for one predictor variable can vary between two computations, because each computation is based on a different random

permutation of the variable. Therefore, before interpreting random forest permutation importance values, the analysis should be repeated (with several different random seeds) to test the stability of the results.

Tab. 6.5: Average misclassification rates of the `randomForest` method applied with sampling with and without replacement as compared to those of the `cforest` method applied with sampling with and without replacement. (Standard errors of the average misclassification rates are given in parentheses.)

Method	Repl.	
`randomForest`	true	0.2896 (0.0022)
	false	0.2879 (0.0026)
`cforest`	true	0.2807 (0.0024)
	false	0.2788 (0.0025)

Similarly to the simulation study, we also compared the prediction accuracy of the four approaches for this data set. To do so, we split the original data set into learning and test sets with size ratio 2:1 in a standard split-sample validation scheme. A random forest is grown based on the learning set and subsequently used to predict the observations in the test set. This procedure is repeated 100 times, and the average misclassification rates over the 100 runs are reported in Table 6.5. Again we find a slight superiority of the `cforest` function, especially when sampling is conducted without replacement. (Differences to the accuracy values reported by Cummings and Myers (2004) are most likely due to their use of a different validation scheme, that is not reported in detail by Cummings and Myers (2004).)

6.5 Summary

The popularity of random forests, especially in bioinformatics and related fields, where identifying a subset of relevant predictor variables from very large sets of candidates is the major challenge, is largely due to the variable importance measures they provide. However, when a method is used for interpretation and variable selection purposes, rather than prediction only, it is particularly important that it actually depicts the importance of the variable and is not affected by any other characteristics.

For the original random forest method we have argued theoretically and shown in simulation studies that the variable importance measures are affected by the number of categories and scale of measurement of the predictor variables, which are no direct indicators of the true importance of the variable.

As long as, e.g., only continuous predictor variables, as in most gene expression studies, or only variables with the same number of categories are considered in the sample, variable selection with random forest variable importance measures is not affected by our findings. However, in studies where continuous variables, such as the folding energy, are used in combination with categorical information from the neighboring nucleotides, or when categorical predictors, as in amino acid sequence data, vary in their number of categories present in the sample, variable selection with random forest variable importance measures is unreliable and may even be misleading.

Especially information on clinical and environmental variables are often gathered by means of questionnaires, where the number of categories can vary between questions. The number of categories is typically determined by many different factors, but is not necessarily an indicator of variable importance. Similarly, the number of different categories of a predictor actually available in a certain sample is not an indicator of its relevance for predicting the response. Hence, the number of categories of a variable should not influence its estimated importance – otherwise the results of a study could easily be distorted when an irrelevant variable with many categories is included in the study design.

We showed that, due to variable selection bias in the individual classification trees and effects induced by bootstrap sampling, the variable importance measures of the `randomForest` function are not reliable in many scenarios relevant in applied research.

As an alternative random forest method we propose to use the `cforest` function, that provides unbiased variable selection in the individual classification trees. When this method is applied with subsampling without replacement as suggested here, the resulting variable importance measure can be used reliably for variable selection even in situations where the potential predictor variables vary in their scale of measurement or their number of categories.

The subsampling size was set to 0.632 in our first approach. This sample size reflects the number of observations that, in average, end up in a bootstrap sample: The probability for one observation not to be included in one draw is $1 - \frac{1}{n}$ and thus its probability not be included in any one of the n draws for large n tends to $\lim_{n\to\infty} \left(1 - \frac{1}{n}\right)^n = e^{-1} \approx 0.368 = 1 - 0.632$. Respectively each observation has a 63.2% chance to end up in the bootstrap sample of size n, and in average 63.2% of the n observations are included.

The reasoning and results of Friedman and Hall (1999) and Buja and Stuetzle (2006) on the other hand indicate that half-sampling (drawing a sample half the size of the original sample) might be more appropriate, because it shares some theoretical characteristics with bootstrap sampling. Therefore the effect of different subsample sizes was evaluated in an additional simulation study (Wösthoff, 2008, supervised by Strobl and Augustin). The results support our previous findings and indicate that the choice of the subsample size is not critical: for example the results for size 0.5 and 0.632 are virtually equivalent. Only for extreme sample size there is a tradeoff between the subsample size that is actually used to fit the model and the size of the remaining out-of-bag sample, that is used to compute the permutation importance: If the subsample size is so large that the remaining out-of-bag sample becomes very small, the variability of the importance measure increases, while if the subsample size used for model fitting is is too small, the relevant predictor variables may not always be detected.

With respect to computation time it should be noted that the `cforest` function is more expensive than the `randomForest` function, because in order to be unbiased split decisions and stopping rely on time-consuming conditional inference. To give an impression, the computation times of the application to C-to-U conversion data, with 876 observations and 44 predictor variables, as stated in the supplementary file for the `cforest` function used with bootstrap sampling with replacement are in the range of 8.38 sec., while subsampling without replacement is computationally less expensive and in the range of 4.82.

Since we saw that only subsampling without replacement guarantees reliable variable selection and produces unbiased variable importance measures, the faster version without replacement should be preferred anyway. The computation time for the `randomForest` function is in the range of 0.24 sec. with and 0.18 sec. without replacement. However, we saw that the `randomForest` function should not be used when the potential predictor variables vary in their scale of measurement or their number of categories.

The aim of this chapter was to explore the limits of different measures of variable importance currently provided by random forests and to guarantee for the permutation importance that variable importance scores are unbiased and reliable for predictor variables of different types. So far the variable importance scores were considered as merely descriptive statistics. However, when variable importance scores are supposed to be used for variable selection, it would be nice to have a statistical test to guide the decision on which and how many predictor variables to select in a certain problem. One such statistical test, that was suggested for the purpose of identifying "significantly important" predictor variables, is critically investigated in the next chapter.

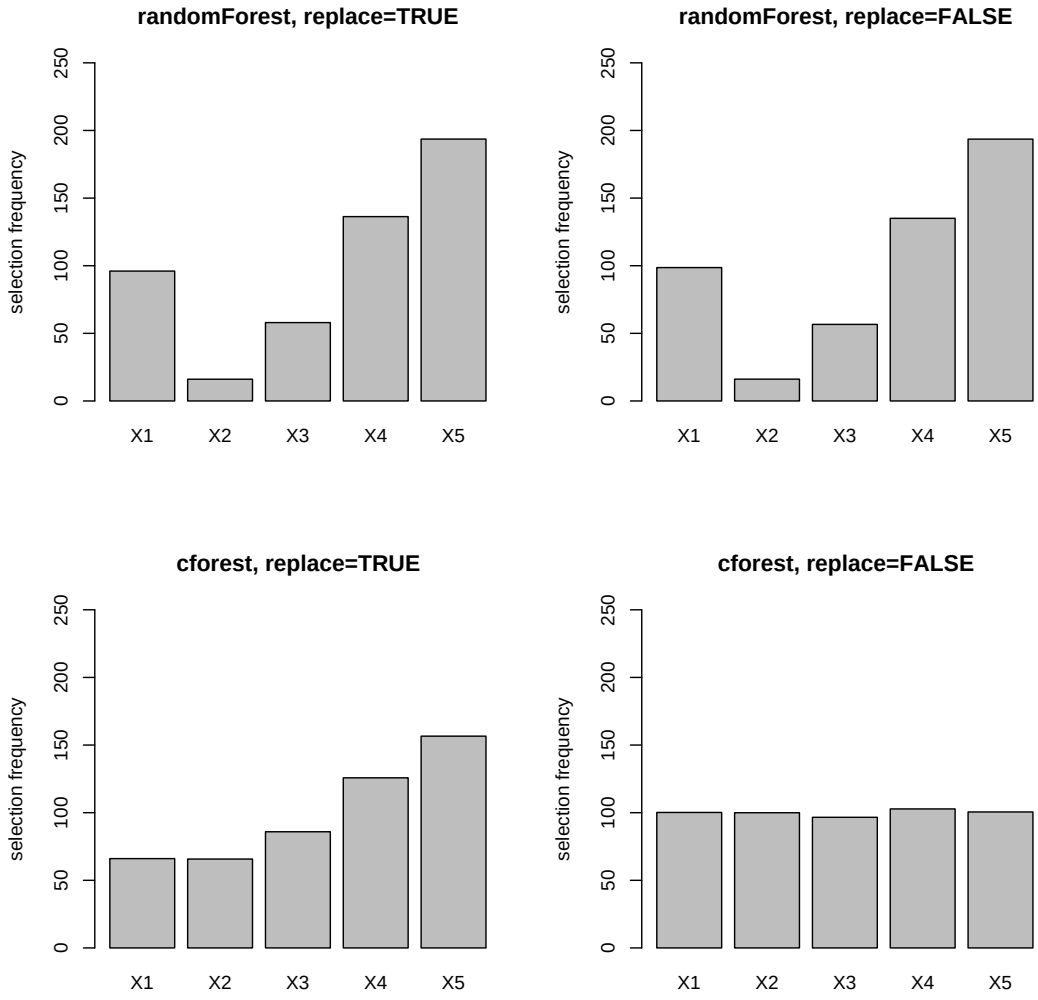


Fig. 6.1: Average variable selection frequencies for the null case, where none of the predictor variables is informative. The plots in the top row display the frequencies when the `randomForest` function is used, the bottom row when the `cforest` function is used. The left column corresponds to bootstrap sampling with replacement, the right column to subsampling without replacement.

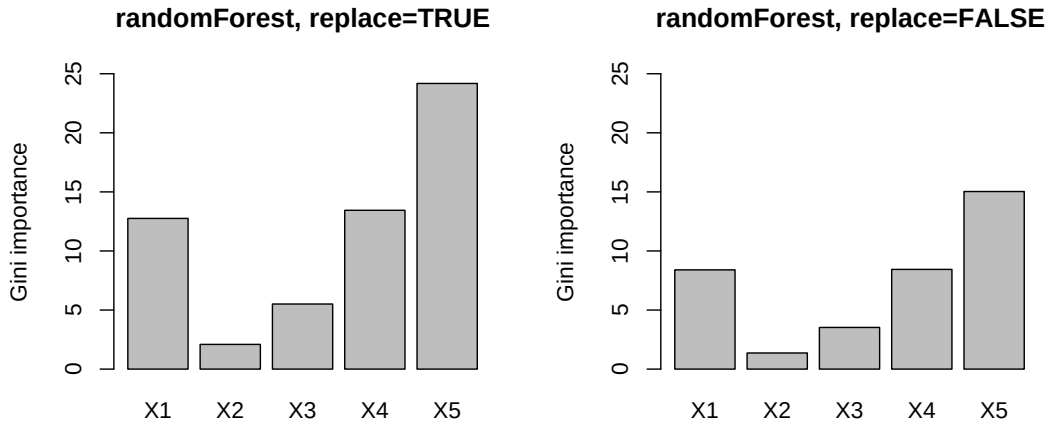


Fig. 6.2: Average Gini importance for the null case, where none of the predictor variables is informative. The left plot corresponds to bootstrap sampling with replacement, the right plot to subsampling without replacement.

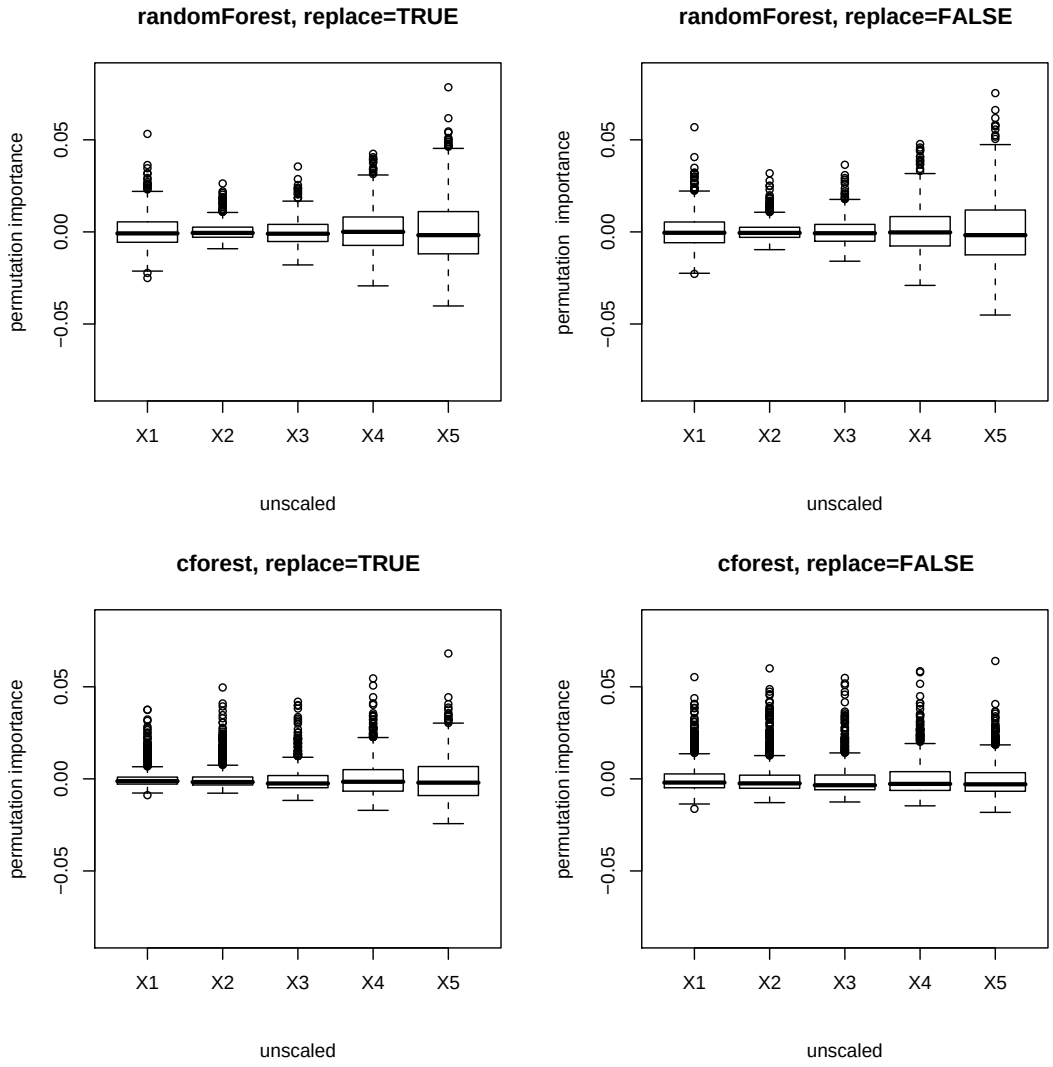


Fig. 6.3: Distributions of the unscaled permutation importance measures for the null case, where none of the predictor variables is informative. The plots in the top row display the distributions when the `randomForest` function is used, the bottom row when the `cforest` function is used. The left column corresponds to bootstrap sampling with replacement, the right column to subsampling without replacement.

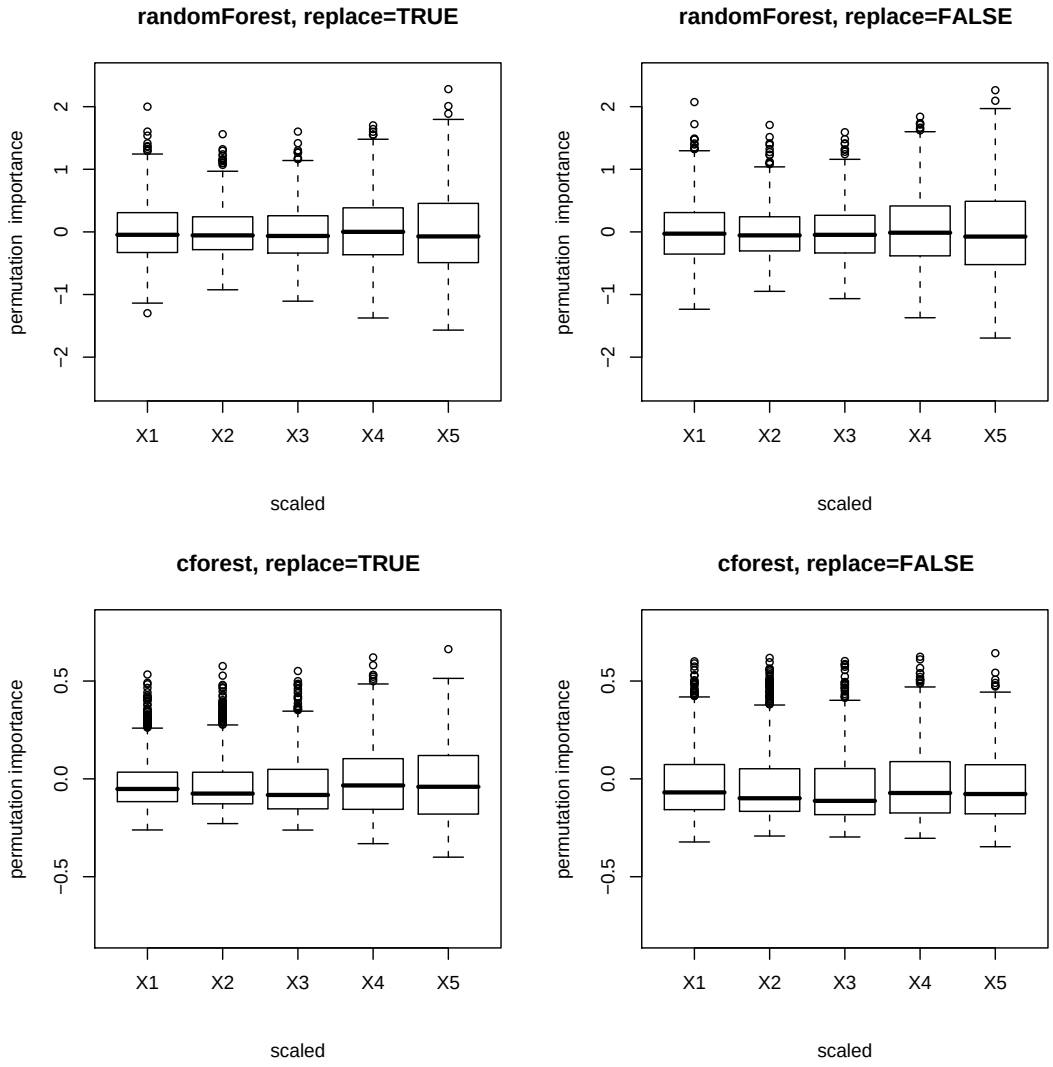


Fig. 6.4: Distributions of the scaled permutation importance measures for the null case, where none of the predictor variables is informative. The plots in the top row display the distributions when the `randomForest` function is used, the bottom row when the `cforest` function is used. The left column corresponds to bootstrap sampling with replacement, the right column to subsampling without replacement.

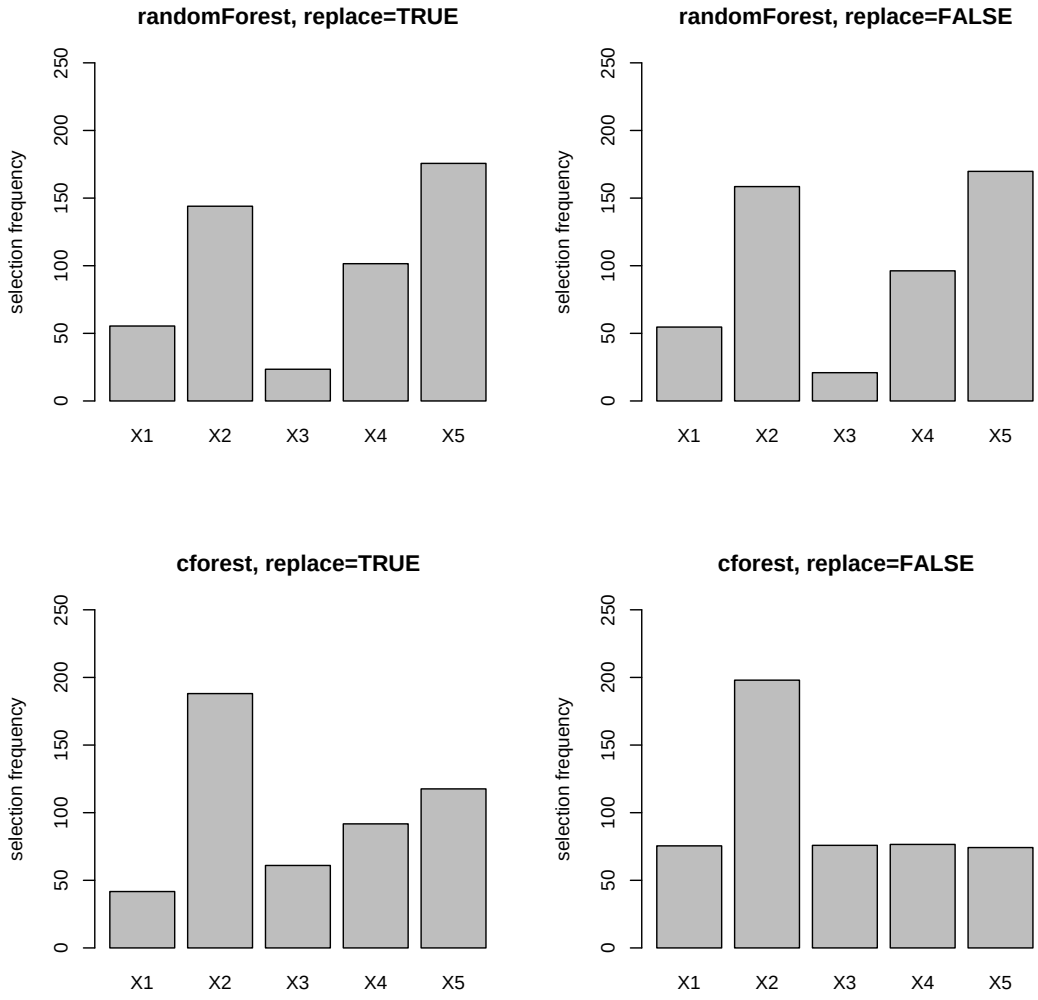


Fig. 6.5: Average variable selection frequencies for the power case, where only the second predictor variable is informative. The plots in the top row display the frequencies when the `randomForest` function is used, the bottom row when the `cforest` function is used. The left column corresponds to bootstrap sampling with replacement, the right column to subsampling without replacement.

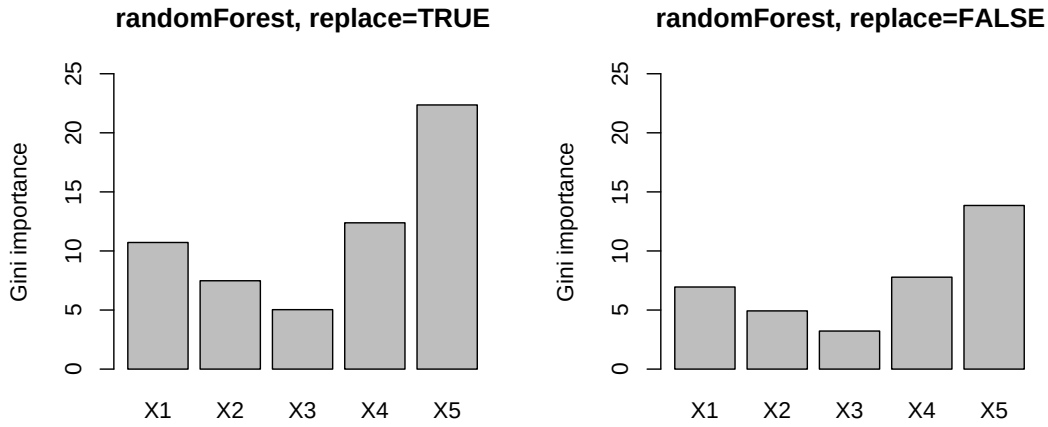


Fig. 6.6: Average Gini importance for the power case, where only the second predictor variable is informative. The left plot corresponds to bootstrap sampling with replacement, the right plot to subsampling without replacement.

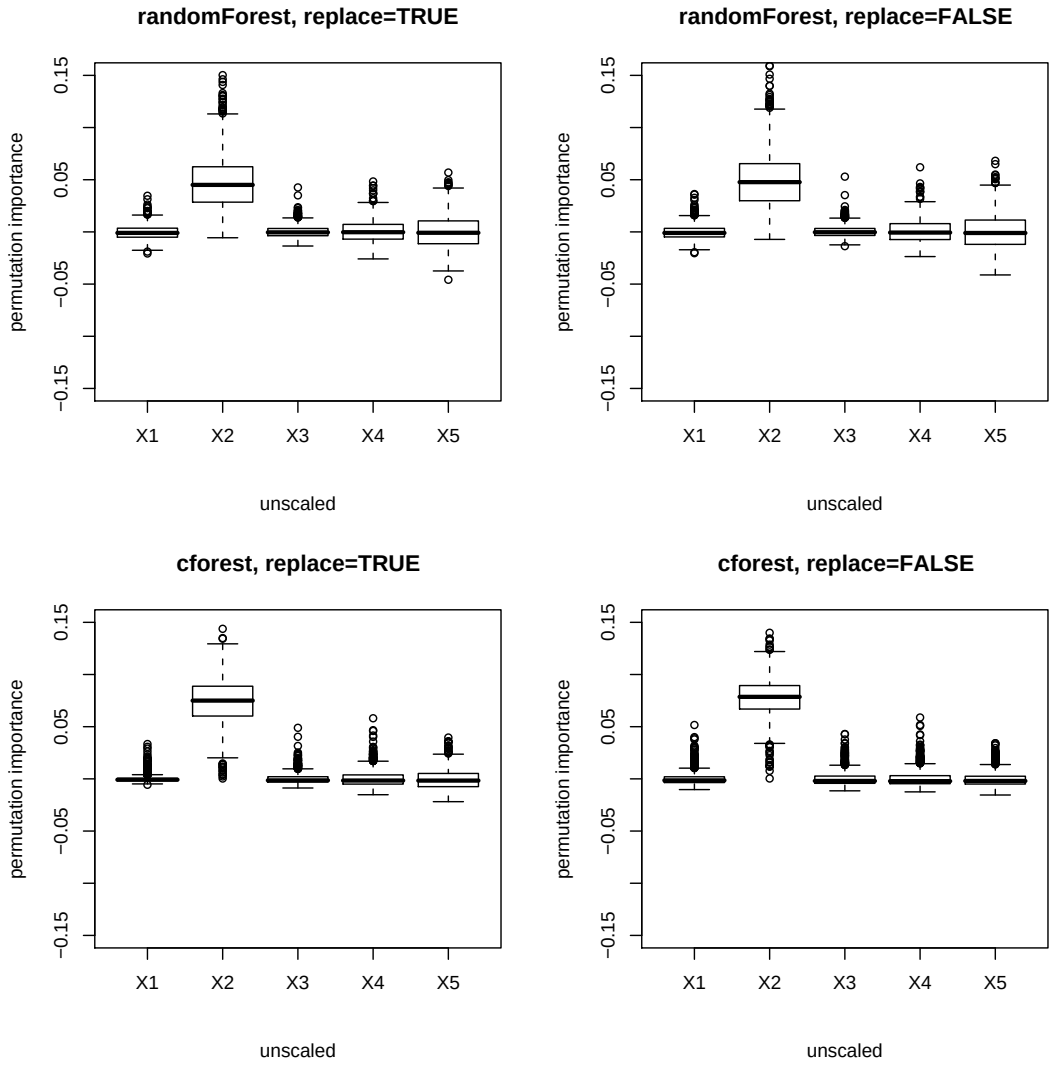


Fig. 6.7: Distributions of the unscaled permutation importance measures for the power case, where only the second predictor variable is informative. The plots in the top row display the distributions when the `randomForest` function is used, the bottom row when the `cforest` function is used. The left column corresponds to bootstrap sampling with replacement, the right column to subsampling without replacement.

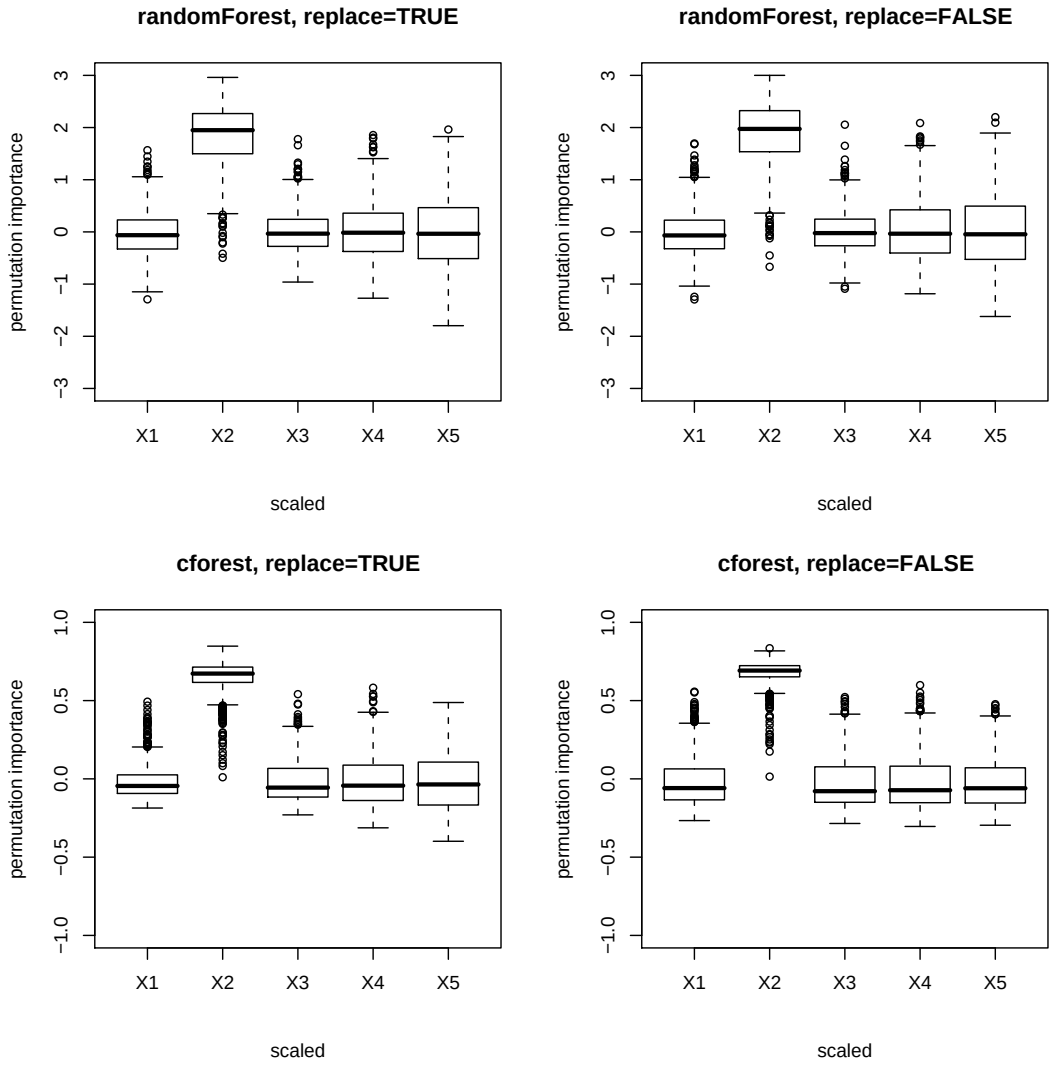


Fig. 6.8: Distributions of the scaled permutation importance measures for the power case, where only the second predictor variable is informative. The plots in the top row display the distributions when the `randomForest` function is used, the bottom row when the `cforest` function is used. The left column corresponds to bootstrap sampling with replacement, the right column to subsampling without replacement.

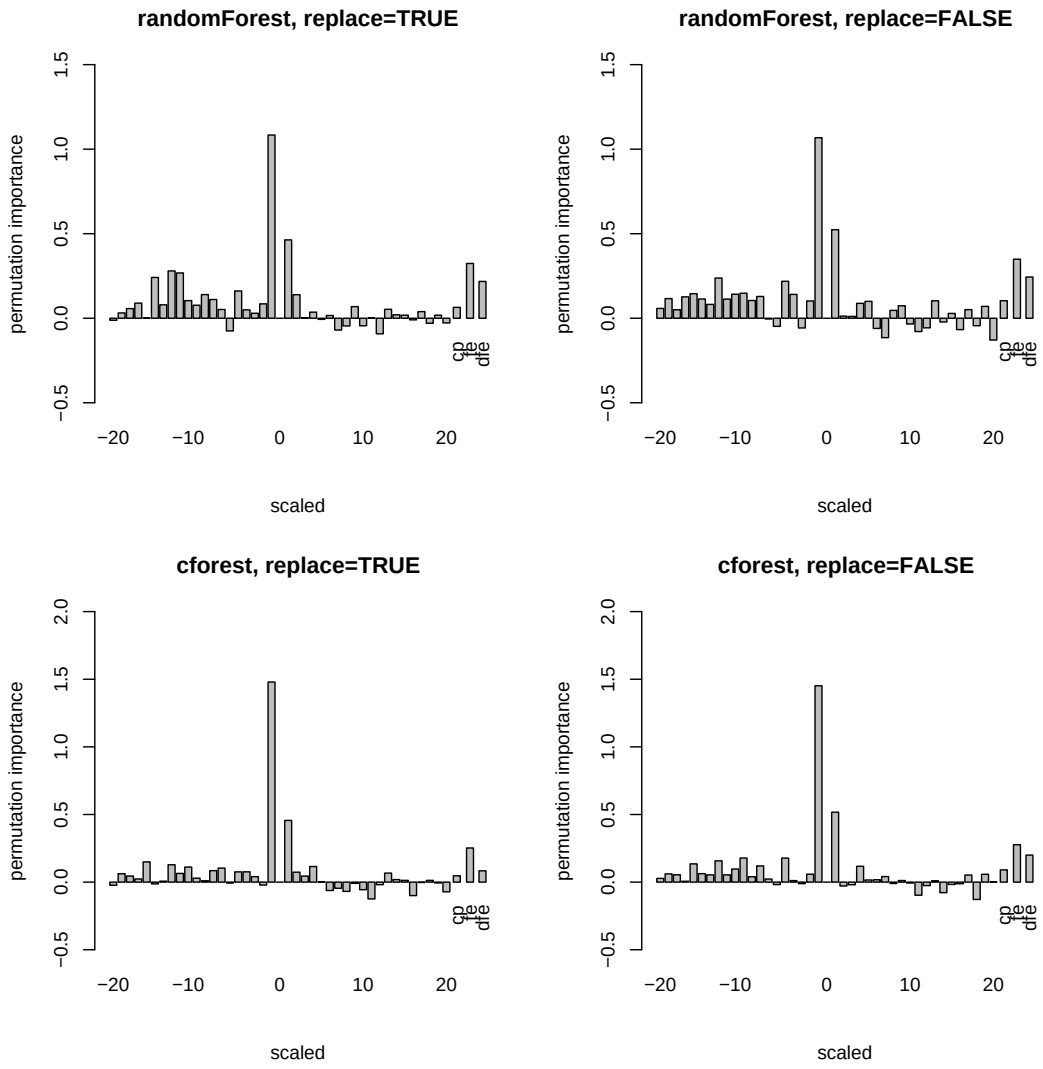


Fig. 6.9: Scaled variable importance measures for the C-to-U conversion data. The plots in the top row display the measures when the `randomForest` function is used, the bottom row when the `cforest` function is used. The left column corresponds to bootstrap sampling with replacement, the right column to subsampling without replacement. In each plot the positions -20 through 20 indicate the nucleotides flanking the site of interest, and the last three bars on the right refer to the codon position (`cp`), the estimated folding energy (`fe`) and the difference in estimated folding energy (`dfe`).

7. Statistical properties of Breiman and Cutler's test for variable importance

Currently, most applications of the random forest permutation importance rely on a merely descriptive ranking of the potential predictor variables with respect to their importance: The few top-ranked predictors are selected for further exploration, where the number of selected variables is chosen arbitrarily or with respect to subject matter. A different approach for variable selection with random forests is introduced by Diaz-Uriarte and Alvarez de Andrés (2006), who suggest a backward elimination strategy based on the variable importance scores that takes under consideration the prediction accuracy: The underlying rationale is that the prediction accuracy will remain almost constant when irrelevant predictor variables are excluded, while it drops when relevant ones are excluded.

While in statistical modelling the aim may often be to select a model as sparse as possible, it is of equal interest in many applied sciences to be able to identify *all* predictor variables that are associated with the response, even if some of them are correlated. The question of interest here is to decide for each variable whether or not its importance is significantly greater than zero. A statistical test for this question is suggested by Breiman and Cutler (2008). It has been promoted on the official random forests website for some time, and thus has been applied in a variety of studies – ranging from the investigation of predictors of attempted suicide (Baca-Garcia et al., 2007) to the monitoring of a large area space telescope on board of a satellite (Paneque et al., 2007) – since its publication.

At first sight it looks like this test could aid the decision which or how many of the top-

ranked variables have significant importance and can be considered relevant. However, in the following we will present statistical reasoning and simulation results illustrating that the suggested test is not appropriate for statements of significance. Moreover, we will explore the unclear null hypothesis of the suggested test and suggest a new permutation scheme that better represents the desired null hypothesis in the next chapter.

7.1 Investigating the current test

The rationale of the random forest permutation importance and the computation of the raw importance score as well as the scaled z-score was already outlined in the previous chapter. If we assume that, under the null hypothesis of zero variable importance, the asymptotic distribution of the z-score is standard normal, a simple test for the permutation importance can be constructed: When the z-score $\widetilde{VI}(\mathbf{x}_j)$ from Equation 6.3 exceeds the α-quantile of the standard normal distribution, the null hypothesis of zero importance for variable X_j is rejected. This approach has been suggested by Breiman and Cutler (2008) for testing the variable importance. Note, however, that in the computation of the z-score averaging and scaling is not conducted with respect to the sample size n but to the number of trees in the ensemble *ntree* (cf. also Lunetta et al., 2004).

7.1.1 The power

To investigate the power of the test suggested by Breiman and Cutler (2008), that is outlined in the previous section, a simulation study was conducted. The experimental parameters that were varied are (i) the relevance of the predictor variable, (ii) the sample size, and (iii) the number of trees in the forest. For each combination of experimental parameters, 1000 replications were run. In each replication, a data set with the respective relevance and sample size was generated, a random forest with the respective number of trees was fit to the data and the z-score was computed as described in the previous

section. The test decision, i.e., whether or not the null hypothesis was rejected, was stored in every replication. The relative frequency of rejections of the null hypothesis (out of the 1000 replications) serves as an estimator for the power of the test in each combination of experimental parameters. In Figure 7.1 the empirical power is displayed as a function of the experimental parameters. For a deeper understanding of the underlying mechanism we also display the curves for the unstandardized average importance VI, the standard error of the mean and the z-score $\widetilde{VI}$ (all averaged over 1000 replications).

In each iteration, a data set of sample size $n = 100, 200$ or 500 was generated, that included five predictor variables of which only one binary variable was relevant. Within the categories of this variable the binary response class was sampled from a binomial distribution with class probability $0.5 \pm relevance$ (with $relevance = 0, 0.05, \ldots, 0.5$) as indicated on the abscissas of Figure 7.1. The parameter settings for the random forests were given by the varying number of trees ($ntree = 100, 200$ or 500) and a fixed number of two preselected variables per split. The simulation was conducted with the function `randomForest` (Breiman et al., 2006; Liaw and Wiener, 2002), the reference implementation of Breiman and Cutler's random forests in the R system for statistical computing (R Development Core Team, 2008).

As depicted in the bottom row of Figure 7.1 the power of the test against the null hypothesis of zero importance shows the following irritating behavior: The power does increase with the relevance of the predictor variable as expected for any reasonable power curve. However, the power also does increase with the number of trees in the forest (the curves are shifted to the left, resulting in higher power for low relevance values), meaning that the power here depends on a tuning parameter that can be arbitrarily chosen by the user. This effect is due to the construction of the test statistic where, unlike in the standard test for the mean under normality, averaging and scaling is not with respect to a given sample size n but to the number of trees as outlined above. Even more dramatically, we find that the power does depend on the sample size – however not as expected for any reasonable test, where the power is supposed to increase with increasing sample size, but to the contrary:

For large sample sizes (as compared to the number of trees) the power is zero.

7.1.2 The construction of the z-score

To explore in more detail the mechanism responsible for this odd behavior, we will follow the construction of the z-score, that is derived from the average importance by division through the standard error of the mean. The top row of Figure 7.1 shows that the unstandardized average importance VI for one predictor variable increases with the relevance of the predictor variable and with the sample size as expected. There is no effect of the number of trees on the average importance – at least not when the number of trees is chosen sufficiently large to guarantee a stable estimate of the mean importance. This increase in the relevance and the sample size is desirable and exactly what we would have expected for any statistic to be employed in a test against the null hypothesis of zero importance. Therefore, the standard error of the mean, which is used for scaling, must be responsible for the odd behavior of the z-scores: The numerator of the fraction for the standard error of the mean, the standard deviation, also increases with the relevance and with the sample size, and does not depend on the number of trees either.

The increase in the sample size is due to the resulting increase in the out-of-bag sample size, that again extends the range of possible changes in the prediction accuracy induced by permuting the predictor variable. The dependence on the relevance is caused by a mechanism in the tree-building process: In many trees of the ensemble a variable with a low relevance may not be included at all, and produce an importance score of exactly zero, which diminishes the variation of the importance. As a result of the division by the square root of the number of trees, however, an additional dependence on the number of trees is induced in the standard error of the mean, such that it decreases in the number of trees as depicted in the second row of Figure 7.1. Note also that the curves for the different sample sizes vary more strongly for the standard error of the mean than for the average importance.

When finally the z-score is computed by means of standardizing the average importance with the standard error of the mean, the rationale of this standardization is to account for the fact that the average importance is an average over all trees in the ensemble – it does, however, not account for the effect of the sample size. The fact that the dependence of the average importance on the sample size is less pronounced than that of its standard error causes an inversion of the importance pattern with respect to the sample size in the z-scores: We find in the third row of Figure 7.1 that the z-score decreases in the sample size but increases with the number of trees. This finally leads to the pattern for the power curves that we found in the bottom row of Figure 7.1: Only for high numbers of trees the overall level of the scaled importance is high enough for all sample sizes to ever reject the null hypothesis, while for lower numbers of trees the curves for the high sample sizes never exceed the threshold for rejecting the null hypothesis and result in a power of zero.

This behavior is undesired and is an artefact of the scaling, that induces a dependence on the number of trees but at the same time inverts the dependence on the sample size. We therefore summarize the results of our simulation study that the original, unscaled average variable importance VI shows the increase in the relevance and sample size that would be desired for a test for the null hypothesis of zero importance, while the scaled variable importance and the resulting test behave oddly.

7.1.3 Specifying the null hypothesis

Another issue when considering the test for the random forest permutation importance suggested by Breiman and Cutler (2008) is the very fundamental question: Exactly what null hypothesis is being tested? In the previous sections we referred to the null hypothesis as "importance equal to zero" for simplicity . This implies some kind of independence between the predictor variable whose importance is being tested and the response. However, it is unclear what kind of independence is being tested. The currently employed permutation scheme, where only the values of the variable of interest are permuted while the values of the

response variable and the other predictors are held constant, does mimic the elimination of the predictor variable when predicting the response – however, at the same time it destroys all correlations between the variable of interest and the other covariates. Unlike standard permutation test of the global null hypothesis that the response is not correlated with any of the predictor variables, where the response is permuted against the complete predictor matrix and all associations within the predictor matrix are retained, the current random forest approach tests the rather unintuitive null hypothesis that the predictor of interest is not correlated with either one of the response or covariates. In cases where predictor variables may be correlated, this permutation scheme may not reflect the actual null hypothesis of interest. This topic is investigated in more detail and a new, conditional permutation importance measure is suggested in the next chapter.

7.2 Summary

We conclude that, in principle, a test for the random forest permutation importance could help identify relevant predictor variables. However, the results of our simulation studies also show that, in its current form, the test of Breiman and Cutler (2008) has prohibitively undesirable properties: The power of the test does not increase with the sample size, as would be expected for any reasonable statistical test, but rather remains zero for large sample sizes as compared to the number of trees. On the other hand the power does increase with the number of trees, which is a parameter that can be arbitrarily chosen by the user. This means that any statement of significance made with the current test for random forest variable importance is nullified.

Another issue, that is relevant in the context of correlated predictor variables, is the question whether the null hypothesis that is being tested in the current test is the one that reflects our understanding of the impact of a predictor variable on the response. A conditional permutation scheme that better reflects the null hypothesis of interest is suggested in the next chapter.

Further research will address the issue of an adequate test statistic and rejection area for this null hypothesis. For high numbers of variables, multiple testing issues should also be taken into consideration.

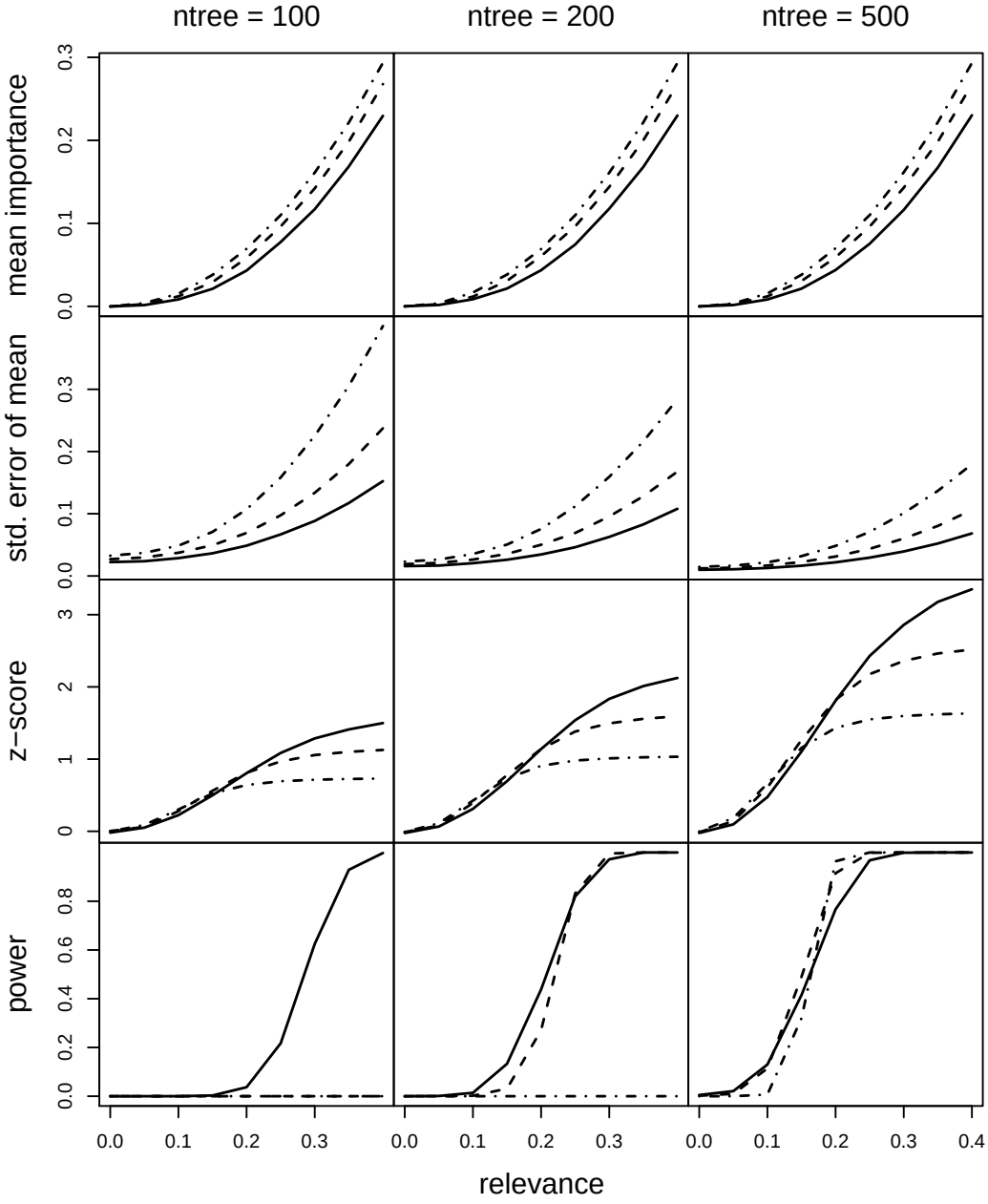


Fig. 7.1: Average variable importance , standard error of mean, z-score and power as functions of relevance for sample size 100 (solid), 200 (dashed), and 500 (dash-dotted) and different numbers of trees.

8. Conditional variable importance

Identifying relevant predictor variables, rather than only predicting the response by means of some black box model, is of interest in many applications. By means of variable importance measures the candidate predictor variables can be compared with respect to their impact in predicting the response or even their causal effect (cf. van der Laan, 2006, for theoretical assumptions necessary for interpreting the importance of a variable as a causal effect). In this case, a key advantage of random forest variable importance measures, as compared to univariate screening methods, is that they cover the impact of each predictor variable individually as well as in multivariate interactions with other predictor variables.

This feature of random forests has made them particularly popular in the field of genomics. For example, Lunetta et al. (2004) find that genetic markers relevant in interactions with other markers or environmental variables can be detected more efficiently by means of random forests than by means of univariate screening methods like Fisher's exact test. In the analysis of amino acid sequence data Segal et al. (2001) also point out the necessity to consider interactions between sequence positions. Tree-based methods like random forests can help identify relevant predictor variables even in such high dimensional settings involving complex interactions. Therefore, the impact of different amino acid properties, some of which have been shown to be relevant in DNA and protein evolution by Xia and Li (1998), for predicting peptide binding is investigated in our application example in Section 8.3. However, we will find in this application example, as often in practical problems, that many predictor variables are highly correlated.

The issue of correlated predictor variables is prominent in, but not limited to, applications

in genomics and other high-dimensional problems. Therefore, it is important to note that in any non-experimental scientific study, where the predictor variable settings cannot be manipulated independently by the investigator, the distinction between the marginal and the conditional effect of a variable is crucial – otherwise we might be mislead to, e.g., consider the shoe size of school children a valuable predictor for their reading skills.

In this obvious case of a spurious correlation the age of the children, that is associated with both their shoe size and reading skills, can be easily identified as a background variable – and it is clear that once their age is used as a predictor variable for the reading skills of the children, knowing their shoe size has no additional benefit. From a statistical point of view, however, this distinction can only be made by a conditional importance measure, while a marginal importance measure would consider the shoe size an equally valuable predictor variable.

We will point out throughout this chapter that correlations between predictor variables – regardless of whether they arise from the proximity of genetic loci or more obviously related characteristics of the subjects, such as their age and shoe size – severely affect the original random forest variable importance measures, because they can be considered as measures of marginal importance, even though what is of interest in most applications is the conditional effect of each variable.

In parametric models, such as (generalized) linear models, variable importance is usually associated with standardized coefficient estimates or the change in a fit index when one predictor variable at a time is excluded as in nested models. In nonparamteric black box learners like neural networks, bagging, random forests and boosting, on the other hand, it is not at all obvious how to assess variable importance – but there are various suggestions that in principle share two rationales: Either the change in the response variable is considered when the value of the predictor variable of interest is changed in the sense: "if X_j is increased by one entity, how will Y change?" (as, e.g., in the "partial dependence plots" of Breiman, 2001b), or the exclusion of one variable from the model is mimiced by means of a random permutation of the predictor variable (as, e.g., in the permutation accuracy

importance of Breiman, 2001a, the rational of which was outlined in Chapter 6).

An advantage of the permutation approach, that is described in more detail below, is that the effect of excluding a predictor variable can be evaluated without actually having to refit the model without that variable. Refitting a model with one variable left out is an adequate strategy in parametric regression, where the resulting models would be nested and the restricted model can be tested by means of computational "shortcuts" like the Wald- and score-test. As opposed to that, in a computerintensive procedure with random components like bagging and random forests, refitting is computationally expensive and the resulting models would not be nested in the common sense, so that model comparison is not straightforward.

Another interesting aspect of the permutation approach in random forests is that it is the prediction accuracy, rather than an information criterion, that is compared before and after permuting the predictor variable. This rationale is comparable to that of another group of descriptive association measures termed "PRE (proportional reduction in error)-measures", that are particularly popular in the social sciences and also compare the prediction error with and without incorporating a predictor (cf, e.g., Liebetrau, 1983, for an introduction).

Let us now shortly review previous suggestions from the literature for measuring or illustrating variable importance in classification trees and random forests, with an emphasis on the distinction between marginal and conditional approaches: As for graphical assessment of variable importance, Breiman (2001b) and Feraud and Clerot (2002) display the change in the predictor over the range of one predictor variable in plots. Feraud and Clerot (2002) later define variable importance as the change in the prediction for different variations of the original value of that variable. In their definition Feraud and Clerot (2002) employ the distribution of the predictor variable and suggest the use of prior distributions to model the possible variation in the distribution of the predictor variable. Lemaire and Clerot (2006) point out that the measure of Feraud and Clerot (2002) is closely related to that of the permutation accuracy importance of Breiman (2001a), with the difference that Breiman's approach does not rely on a prior distribution because he uses bootstrap sampling for

reflecting the distribution of values in the sample.

The interpretation of the "partial dependence plots" of Breiman (2001b) may remind of the interpretation of model coefficients in linear models. However, whether the effect of a variable is interpretable as conditional on all other variables, as in linear models, may not be guaranteed in other models – and we will point out explicitly below that this is not the case in classification trees or random forests.

With regard to measures of variable importance, the permutation accuracy importance follows the rationale that a random permutation of the values of the predictor variable is supposed to mimic the absence of the variable from the model (cp. Chapter 6), while the alternative Gini importance, is based on the principle of impurity reduction that is followed in most traditional classification tree algorithms: A split in a certain variable is considered good when it leads to a reduction in the impurity between the response classes. The Gini importance of a random forest is an average over the impurity reductions a variable can achieve in all trees in the forest. However, it has been shown to be biased when predictor variables vary in their number of categories or scale of measurement in Chapter 6, because the underlying Gini gain splitting criterion is a biased estimator and can be affected by multiple testing effects, as described in the earlier chapters. Therefore, we will focus on the permutation importance in the following, for which we have already shown that it is reliable when subsampling without replacement – instead of bootstrap sampling – is used in the construction of the forest.

Based on the permutation importance, schemes for variable selection and for providing statements of the "significance" of a predictor variable (instead of a merely descriptive ranking of the variables w.r.t. their importance scores) have been derived: Breiman and Cutler (2008) suggest a simple significance test that, however, shows poor statistical properties as illustrated in the previous chapter. An approach for variable selection in large scale screening studies is introduced by Diaz-Uriarte and Alvarez de Andrés (2006), who suggest a backward elimination strategy. This approach has been shown to provide a reasonable selection of genes in many situations and is freely available in an R package

(Diaz-Uriarte, 2007), that also provides different plots for comparing the performance on the original data set to those on a data set with randomly permuted values of the response variable. The latter mimics the overall null hypothesis that none of the predictor variables is relevant and may serve as a baseline for significance statements. A similar approach is followed by Rodenburg et al. (2008).

However, some recent simulation studies indicate that the performance of the variable importance measures may not be reliable when predictor variables are correlated: Even though Archer and Kimes (2008) show in their extensive simulation study that the Gini importance can identify influential predictor variables out of sets of correlated covariates in many settings, the preliminary results of the simulation study of Nicodemus and Shugart (2007) indicate that the ability of the permutation importance to detect influential predictor variables in sets of correlated covariates is less reliable than that of alternative machine learning methods and highly depends on the number of previously selected splitting variables `mtry`. These studies, as well as our simulation results, indicate that random forests show a preference for correlated predictor variables, that is also carried forward to any significance test or variable selection scheme constructed from the importance measures.

In this chapter we aim to provide an understanding of the underlying mechanisms responsible for the observations of Archer and Kimes (2008) and Nicodemus and Shugart (2007). In addition to this, we want to broaden the scope of considered problems to the comparison of the influence of correlated and uncorrelated predictor variables. For this type of problem we introduce a new, conditional permutation importance for random forests, that better reflects the true importance of predictor variables. Our approach is motivated by the visual means of illustration introduced by Nason et al. (2004): In their "CARTscans" plots, Nason et al. (2004) not only display the marginal influence of a predictor variable, like the partial dependence plots of Breiman (2001b), but the influence of continuous predictor variables separately for the levels of two other, categorical predictor variables, namely a conditional influence plot. In the case of correlated predictor variables it is important to distinguish between conditional and marginal influence of a variable, because a variable that may ap-

pear influential marginally might actually be independent of the response when considered conditional on another variable, as pointed out in the spurious correlation example above. Thus the approach of Nason et al. (2004) is an important improvement, but in its current form is only applicable for categorical covariates. Therefore our aim in this chapter is to provide a general scheme that can be used both for illustrating the effect of a variable and for computing its permutation importance conditional on relevant covariates of any type. While the conditioning scheme of Nason et al. (2004) can be considered as a full-factorial cross-tabulation based on two categorical predictor variables, our conditioning scheme is based on a partition of the entire feature space that is determined directly by the fitted random forest model.

In the following Section 8.1 we will shortly review the particular variable selection approach employed in recursive partitioning and illustrate in a simulation study why correlated predictor variables tend to be overselected with this approach. In Section 8.2 we will question the construction of the original permutation importance in more detail, before we introduce a new permutation scheme that we suggest for the construction of a conditional permutation importance measure. The advantage of this measure over the currently-used one is illustrated in the second part of our simulation study and in the application to peptide-binding data in Section 8.3.

8.1 Variable selection in random forests

As already outlined in Chapter 1, classification trees are built recursively in that the next splitting variable is selected by means of locally optimizing a criterion (such as the Gini gain in the traditional CART algorithm of Breiman et al., 1984) within the current node. This current node is defined by a configuration of predictor values, that is determined by all previous splits in the same branch of the tree. In this respect, the evaluation of the next splitting variable can be considered conditional on the previously selected predictor variables, but regardless of any other predictor variable. In particular, the selection of

the first splitting variable involves only the marginal, univariate association between that predictor variable and the response, regardless of all other predictor variables. However, this search strategy leads to a variable selection pattern where a predictor variable that is per se only weakly or not at all associated with the response, but is highly correlated with another influential predictor variable, may appear equally well suited for splitting as the truly influential predictor variable. We will illustrate this point in more detail in the following simulation study.

8.1.1 Simulation design

A simulation study was set up in order to illustrate the treatment of correlated predictor variables in ensemble methods based on classification trees. Data sets were generated according to a linear model with twelve predictor variables $y_i = \beta_1 \cdot x_{i,1} + \ldots + \beta_{12} \cdot x_{i,12} + \varepsilon_i$, with $\varepsilon_i \overset{i.i.d.}{\sim} N(0, 0.5)$. The predictor variables were sampled from a multivariate normal distribution $X_1, \ldots, X_{12} \sim N(0, \boldsymbol{\Sigma})$ where the covariance structure $\boldsymbol{\Sigma}$ was chosen such that all variables have unit variance $\sigma_{j,j} = 1$ and only the first four predictor variables are block-correlated with $\sigma_{j,j'} = 0.9$ for $j \neq j' \leq 4$, while the rest were independent with $\sigma_{j,j'} = 0$. Of the twelve predictor variables only six were influential, as indicated by their coefficients in Table 8.1. A covariance structure of this type was already used for illustrating the effect of correlations by Archer and Kimes (2008). However, while their study mainly aimed at identifying one influential predictor out of a correlated set, here we also want to compare the importance scores of predictor variables with equally large coefficients, while some of the predictor variables are correlated and others are not: $X_1, \ldots, X_4$ and $X_5, \ldots, X_8$ share the same coefficient pattern, while only $X_1, \ldots, X_4$ are correlated. From the generated data sets, random forests were built with the `cforest` function from the `party` package (Hothorn et al., 2008, 2006) in the R system for statistical computing (R Development Core Team, 2008). Different values for the parameter `mtry`, that regulates the number of randomly preselected splitting variables, were considered to be able to investigate the mechanisms responsible for the results of Nicodemus and Shugart (2007). Default settings

Tab. 8.1: Regression coefficients of the data generating process.

X_j	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	$\cdots$	X_{12}
β_j	5	5	2	0	-5	-5	-2	0	$\cdots$	0

were used for all other parameters.

8.1.2 Illustration of variable selection

We find in the panel on the left hand side of Figure 8.1 that in the first splits of all trees, where the variables are considered only marginally with respect to their association to the response, those variables (X_3 and X_4) correlated with highly influential predictors are selected equally often as the highly influential predictor variables themselves (X_1 and X_2 as well as X_5 and X_6) for `mtry`$= 1$, where no competitors are available and the correlated predictors can serve as replacements of the influential ones. The fact that the non-influential predictor variables X_8 through X_{12} are selected almost equally often is only due to the lax choice of the stop criterion: With a lax stop criterion a split is conducted whenever a variable is selected for splitting, which is equally likely for each variable in the case of `mtry`$= 1$, even if a split in the variable is not worthwhile. If the stop criterion was chosen more strictly, the variables X_8 through X_{12} would still be selected with equal probabilities, but would not actually be used for splitting.

When `mtry` increases and the highly influential variables may be available as predominant competitors in some splits, those variables (X_3 and X_4) correlated with highly influential predictors are selected less often than the highly influential correlated ones themselves (X_1 and X_2), but more often than even the highly influential uncorrelated ones (X_5 and X_6). When we consider all splits of all trees in the panel on the right hand side of Figure 8.1, the correlated predictors loose most of their advantage because variable selection is now conditional on the previously chosen variables in the same branch of the tree, that may include the truly influential correlated predictors. However, since variable selection is not

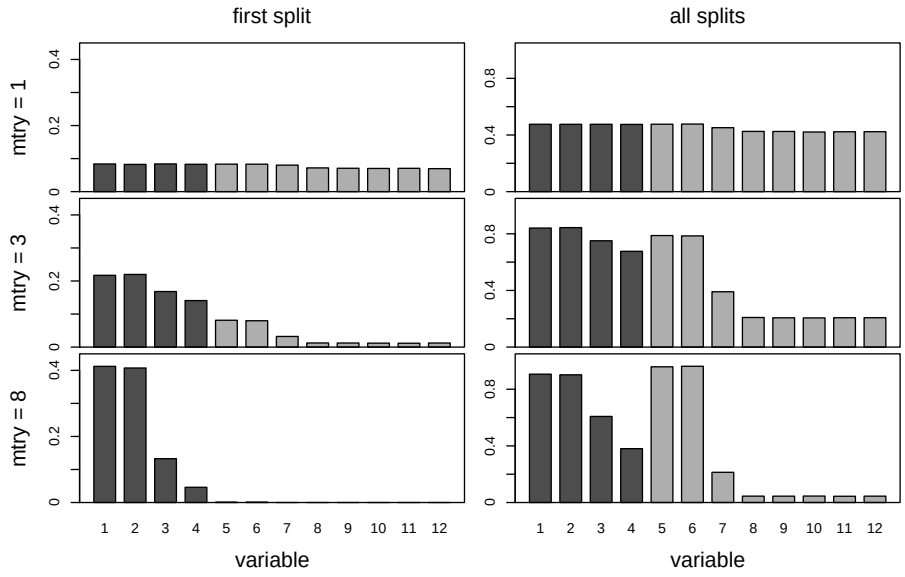


Fig. 8.1: Relative selection rates for twelve variables in the first splits (left) and in all splits (right) of all trees in random forests built with different values for `mtry`.

conditional on all (or at least all correlated) variables, there is still a preference for the correlated variables with low and zero coefficients (X_3 and X_4 over X_7 and X_8), with a similar dependency on `mtry`.

This selection pattern is due to the locally optimal variable selection scheme used in recursive partitioning, that considers only one variable at a time and conditional only on the current branch. However, since this characteristic of tree-based methods is a crucial means of reducing computational complexity (and any attempts to produce globally optimal partitions are strictly limited to low dimensional problems at the moment, cf. van Os and Meulman, 2005), it shall remain untouched here.

8.2 A second look at the permutation importance

We again consider the raw random forest permutation importance $VI(\mathbf{X}_j)$, that is given in Equation 6.2 in Chapter 6. In standard implementations of random forests, the z-score, that is achieved by dividing the raw importance by its standard error, is also provided. However, since the results in the previous chapter indicate that the raw importance has far better statistical properties, we will only consider the unscaled version here.

8.2.1 Background: Types of independence

We know that the original permutation importance overestimates the importance of correlated predictor variables. Part of this artefact may be due to the preference for correlated predictor variables in early splits as illustrated in Section 8.1. However, we also have to take into account the permutation scheme that is employed in the computation of the permutation importance. In the following we will first outline what notion of independence corresponds to the current permutation scheme of the random forest permutation importance. Then we will introduce a more sensible permutation scheme that better reflects the true impact of predictor variables.

It can help our understanding to consider the permutation scheme in the context of permutation tests (cf., e.g., Good, 2005): Usually a null hypothesis is considered that implies the independence of particular (sets of) variables. Under this null hypothesis, some permutations of the data are permitted because they preserve the structure determined by the null hypothesis. If, for example, the response variable Y is independent from all predictor variables (global null hypothesis), a permutation of the (observed) values of Y affects neither the marginal distribution of Y nor the joint distribution of $X_1, \ldots, X_p$ and Y, because the joint distribution can be factorized as $P(Y, X_1, \ldots, X_p) = P(Y) \cdot P(X_1, \ldots, X_p)$ under the null hypothesis. (Note that – in an obvious misuse of notation, but for the sake of comprehensibility of the argument – the form $P(Y, X_1, \ldots)$ is used here not only as an

abbreviation for $P(Y = y, X_1 = x_1, \ldots)$ in the discrete case, but is also meant to cover the case of continuous distributions accordingly.)

If, however, the null hypothesis is not true, the same permutation will lead to a deviation in the joint distribution or some reasonable test statistic computed from it. Therefore, a change in the distribution or test statistic caused by the permutation can serve as an indicator that the data do not follow the independence structure we would expect under the null hypothesis.

With this framework in mind, we can now take a second look at the random forest permutation importance and ask: Under which null hypothesis would this permutation scheme be permitted? If the data are actually generated under this null hypothesis the permutation importance will be (a random value from a distribution with mean) zero, while any deviation from the null hypothesis will lead to a change in the prediction accuracy, that is used as a test statistic here, and thus will be detectable as an increase in the value of the permutation importance.

We find that the original permutation importance, where one predictor variable X_j is permuted against both the response Y and the remaining (one or more) predictor variables $Z = X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p$ as illustrated in the left panel of Table 8.2, corresponds to a null hypothesis of independence between X_j and both Y and Z:

$$H_0 : X_j \perp Y, Z \text{ or equivalently } X_j \perp Y \wedge X_j \perp Z \tag{8.1}$$

Under this null hypothesis the joint distribution can be factorized as

$$P(Y, X_j, Z) \overset{H_0}{=} P(Y, Z) \cdot P(X_j). \tag{8.2}$$

What is crucial when we want to understand why correlated predictor variables are preferred by the original random forest permutation importance is that a positive value of the importance corresponds to a deviation from this null hypothesis – that can be caused by a violation of either part: the independence of X_j and Y, or the independence of X_j and Z.

Tab. 8.2: Permutation scheme for the current and for the conditional permutation importance.

Y	X_j	Z
y_1	$x_{\psi_j(1),j}$	z_1
$\vdots$	$\vdots$	$\vdots$
y_i	$x_{\psi_j(i),j}$	z_i
$\vdots$	$\vdots$	$\vdots$
y_n	$x_{\psi_j(n),j}$	z_n

Y	X_j	Z
y_1	$x_{\psi_{j\|Z=a}(1),j}$	$z_1 = a$
y_3	$x_{\psi_{j\|Z=a}(3),j}$	$z_3 = a$
y_{27}	$x_{\psi_{j\|Z=a}(27),j}$	$z_{27} = a$
y_6	$x_{\psi_{j\|Z=b}(6),j}$	$z_6 = b$
y_{14}	$x_{\psi_{j\|Z=b}(14),j}$	$z_{14} = b$
y_{21}	$x_{\psi_{j\|Z=b}(21),j}$	$z_{21} = b$
$\vdots$	$\vdots$	$\vdots$

However, from these two aspects only one is of interest when we want to assess the impact of X_j to help predict Y, namely the question if X_j and Y are independent.

This aim, to measure only the impact of X_j on Y, would be better reflected if we could create a measure of deviation from the null hypothesis that X_j and Y are independent under a given correlation structure between X_j and the other predictor variables, that is determined by our data set. To meet this aim we suggest a conditional permutation scheme, where X_j is permuted only within groups of observations with $Z = z$, to preserve the correlation structure between X_j and the other predictor variables as illustrated in the right panel of Table 8.2.

We denote the permutation corresponding to this scheme by $\psi_{j|Z}$. By means of conditioning on the variables in Z, the possible permutations of the values of X_j are restricted to those that exchange only the indices of observations within sets of the form $\mathcal{I}_a = \{i \,|\, z_i = a\}$, so that $\psi_{j|Z}(i) \in \mathcal{I}_a \; \forall \, i \in \mathcal{I}_a$ in all constellations a.

The conditional permutation corresponds to the following null hypothesis

$$H_0 : (X_j \perp Y) \,|Z, \tag{8.3}$$

where the conditional distribution can be factorized under the null hypothesis as

$$\begin{aligned} P(Y, X_j|Z) &\overset{H_0}{=} P(Y|Z) \cdot P(X_j|Z) \\ \text{or } P(Y|X_j, Z) &\overset{H_0}{=} P(Y|Z), \end{aligned} \tag{8.4}$$

which is the definition of conditional independence.

In the special case where X_j and Z are independent, both permutation schemes will give the same result, as illustrated by our simulation results below. When X_j and Z are correlated, however, the original permutation scheme will lead to an apparent increase in the importance of correlated predictor variables, that is due to deviations from the uninteresting null hypothesis of independence between X_j and Z.

8.2.2 A new, conditional permutation scheme

Technically, any kind of conditional assessment of the importance of one variable conditional on another one is straightforward whenever the variables to be conditioned on, Z, are categorical (cf., e.g., Nason et al., 2004). However, for our aim to conditionally permute the values of X_j within groups of $Z = z$, where Z can contain potentially large sets of covariates of different scales of measurement, we want to supply a grid that (i) is applicable to variables of different types, (ii) is as parsimonious as possible, but (iii) is also computationally feasible to generate. Our suggestion is to define the grid within which the values of X_j are permuted for each tree by means of the partition of the feature space induced by that tree. The main advantages of this approach are that this partition was already learned from the data during model fitting, contains splits in categorical, ordered and continuous predictor variables and can thus serve as an internally available means for discretizing the feature space.

In principle, any partition derived from a classification tree can be used to define the permutation grid. Here we used partitions produced by the unbiased conditional inference trees of Hothorn et al. (2006), that employ binary splitting as in the standard CART

algorithm of Breiman et al. (1984). This means that, if k is the number of categories of an unordered or ordered categorical variable, up to k, but potentially less than k, subsets of the data are separated. Continuous variables are treated in the same way: Every binary split in a variable provides one or more cutpoints, that can induce a more or less fine graded grid on this variable. By using the grid resulting from the current tree we are able to condition in a straightforward way not only on categorical, but also on continuous variables and create a grid that may be more parsimonious than the full factorial approach of Nason et al. (2004). Only in one aspect we suggest to leave the recursive partition induced by a tree: Within a tree structure, each cutpoint refers to a split in a variable only within the current node (i.e. a split in a variable may not bisect the entire sample space but only partial planes of it). However, for ease of computation, we suggest that the conditional permutation grid uses all cutpoints as bisectors of the sample space (the same approach is followed by Nason et al., 2004). This leads to a more fine graded grid, and may in some cases result in small cell frequencies inducing greater variation (even though our simulation results indicate that in practice this is not a critical issue). From a theoretical point of view, however, conditioning too strictly has no negative effect, while a lack of conditioning produces artifacts as observed for the unconditional permutation importance.

In summary, the conditional permutation importance is derived as follows:

- In each tree compute the oob-prediction accuracy before the permutation as in Equation 6.1: $\frac{\sum_{i \in \overline{\mathfrak{B}}^{(t)}} I\left(y_i = \hat{y}_i^{(t)}\right)}{\left|\overline{\mathfrak{B}}^{(t)}\right|}$.
- For all variables Z to be conditioned on: Extract the cutpoints that split this variable in the current tree and create a grid by means of bisecting the sample space in each cutpoint.
- Within this grid permute the values of X_j and compute the oob-prediction accuracy after permutation: $\frac{\sum_{i \in \overline{\mathfrak{B}}^{(t)}} I\left(y_i = \hat{y}_{i,\psi_j|Z}^{(t)}\right)}{\left|\overline{\mathfrak{B}}^{(t)}\right|}$, where $\hat{y}_{i,\psi_j|Z}^{(t)} = f^{(t)}(\mathbf{x}_{i,\psi_j|Z})$ is the predicted classes for observation i after permuting its value of variable X_j within the grid defined by the variables Z.

– The difference between the prediction accuracy before and after the permutation accuracy again gives the importance of X_j for one tree (cf. Equation 6.1). The importance of X_j for the forest is again computed as an average over all trees.

To determine the variables Z to be conditioned on, the most conservative – or rather overcautious – strategy would be to include all other variables as conditioning variables, as was indicated by our initial notation. A more intuitive choice is to include only those variables whose empirical correlation with the variable of interest X_j exceeds a certain moderate threshold, as we do with the Pearson correlation coefficient for continuous variables in the following simulation study and application example. For the more general case of predictor variables of different scales of measurement the framework promoted by Hothorn et al. Hothorn et al. (2006) provides p-values of conditional inference tests as measures of association. The p-values have the advantage that they are comparable for variables of all types and can serve as an intuitive and objective means for selecting the variables Z to be conditioned on in any problem. Another option is to let the user himself select certain variables to condition on, if, e.g., a hypothesis of interest includes certain independencies.

Note however, that neither a high number of conditioning variables nor a high overall number of variables in the data set poses a problem for the conditional permutation approach: The permutation importance is computed individually for each tree and then averaged over all trees. Correspondingly, the conditioning grid for each tree is determined by the partition of that particular tree only. Thus, even if in principle the stability of the permutation may be affected by small cell counts in the grid, practically the complexity of the grid is limited by the depth of each tree.

The depth of the tree, however, does not depend on the overall number of predictor variables, but on various other characteristics of the data set (most importantly the ratio of relevant vs. noise variables, that is usually low, for example in genomics) in combination with tuning parameter settings (including the number of randomly preselected predictor variables, the split selection criterion, the use of stopping criteria and so forth). Lin and Jeon Lin and Jeon (2006) even point out that limiting the depth of the trees in random

forests may prove beneficial w.r.t. prediction accuracy in certain situations.

Another important aspect is that the conditioning variables, especially if there are many, may not necessarily appear all together with the variable of interest in each individual tree, but different combinations may be represented in different trees if the forest is large enough.

8.2.3 Simulation results

For the simulation design introduced in Section 8.1.1, Figure 8.2 shows the median and interquartile range (over 500 iterations) of the importance scores of each variable for the different permutation schemes: the original marginal permutation and the newly suggested conditional permutation scheme. The set of variables Z to be conditioned on was chosen here to include all variables with a Pearson correlation coefficient $r \geq 0.2$.

We find that the pattern of the coefficients induced in the data generating process is not reflected by the importance values computed with the ordinary permutation scheme. With this scheme the importance scores of the correlated predictor variables are highly overestimated. This effect is most pronounced for small values of `mtry`, because correlated variables have a higher chance to enter a tree when their correlated competitors are not available.

For the conditional permutation scheme the importance scores better reflect the true pattern: The correlated variables X_1 and X_2 with the same coefficient show an almost equal level of importance as the uncorrelated variables X_5 and X_6, while the importance of X_3 and X_4, that are correlated but have a lower or zero coefficient, decrease. For the variables with small and zero coefficients we still find a difference between the correlated and uncorrelated variables, such that for the correlated variables the importance values are still overestimated – however to a much lesser extent than with the unconditional permutation scheme. This remaining disadvantage of the uncorrelated predictor variables, especially those with low coefficients, may be due to the fact that within the individual trees these

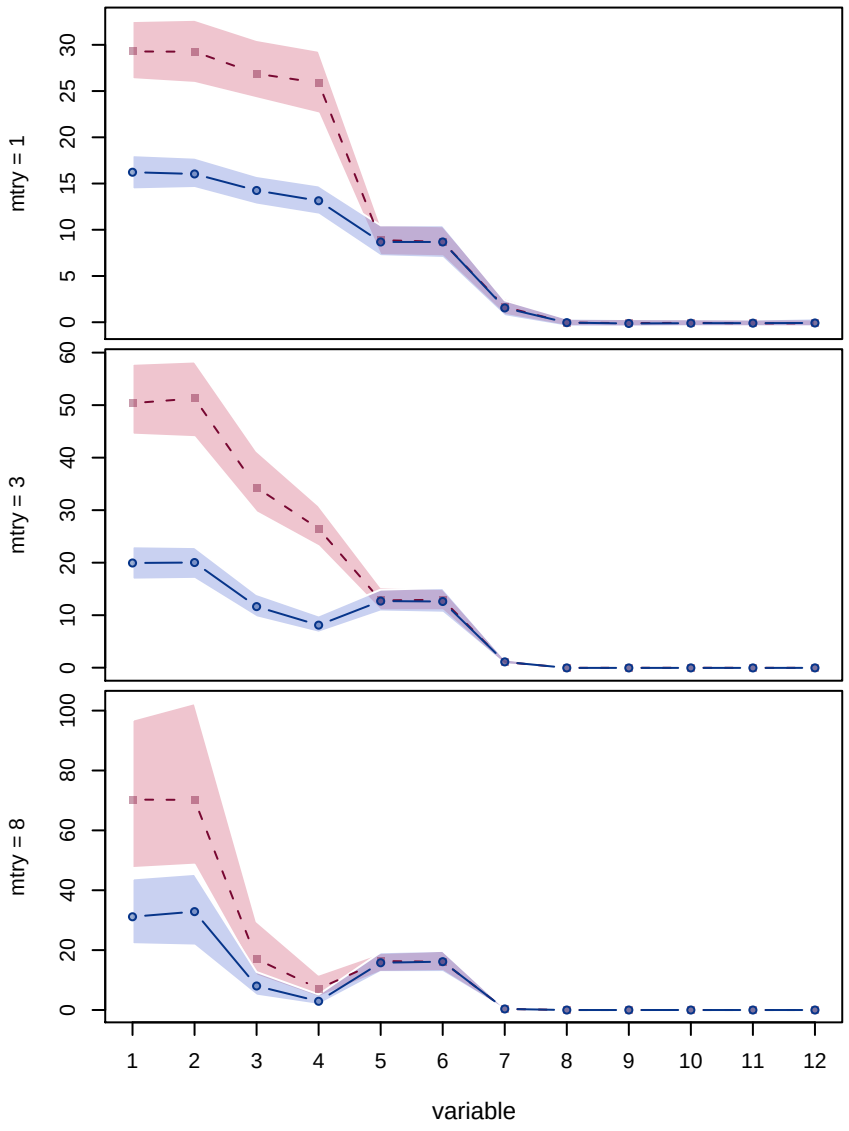


Fig. 8.2: Median permutation importance for unconditional (dashed) and conditional (solid) permutation scheme along with inter-quartile range. Note that the ordering of variables in the plot is arbitrary.

variables are selected less often and in lower positions than their correlated counterparts (cf. Figure 8.1), which results in a lower chance to produce a high importance value for

these variables.

However, the degree of the preference for correlated predictor variables also strongly depends on the choice of `mtry` – both for the unconditional and the conditional permutation scheme – and is most pronounced for small values of `mtry`. The key to understanding this effect is that conditioning – both in recursive tree building and in computing the conditional permutation importance – is effective only when an influential covariate, that is correlated with variable of interest, is already available in the model. In this case, the remaining, conditional effect of the variable of interest can be realistically assessed, because the effect of the influential covariate can be accounted for. However, in trees built with values of `mtry` as low as 1, variable selection is conducted completely at random, so that the influential covariates may not be included in the model at all, and thus cannot be conditioned on. In this context, high values of `mtry` appear more favorable with respect to conditioning.

On the other hand, we find in Figure 8.2 that the variability of the importance increases for large values of `mtry` – and the prediction accuracy of random forests is in general expected to be higher for smaller values of `mtry`. In any case it is interesting that the variability of the conditional importance is lower than that of the unconditional importance within each level of `mtry`.

With respect to the identifiability of few influential predictors from a set of correlated and other noise variables (which was the task in Nicodemus and Shugart (2007) and Archer and Kimes (2008)), we can see from the importance scores for $X_1, \ldots, X_3$ in comparison to that of X_4 that the conditional importance reflects the same pattern as the unconditional importance, however with a notably smaller variation that may improve the identifiability. In the comparison of potentially influential correlated and uncorrelated predictor variables on the other hand, the conditional importance is much better suited as a means of comparison than the original importance.

8.3 Application to peptide-binding data

As an application example we consider peptide-binding data that were previously analysed with recursive partitioning techniques by Segal et al. (2001). The data set includes 105 variables for a total of $n = 310$ amino acid sequences. The response to be predicted is a binding property that can be coded as a binary variable (binding/no binding). The remaining variables available in this data set correspond to 13 amino acid properties for each of the eight considered amino acid positions. These 13 properties include, e.g. volume, polarity, bulkiness, flexibility, aromaticity, and charge, yielding in total 104 continuous predictor variables.

A random forest with 1000 trees and `mtry` $= 104$ (which is equal to bagging) was fit to the data set and the permutation importance was computed either with the unconditional or the conditional permutation scheme. The resulting importance scores are displayed in Figure 8.3 (note that the absolute values of the scores should not be interpreted). The few predictor variables whose importance scores reach highest or even exceed the plotting area would be selected for further analysis by any means. However, for some of the variables with the next smaller importance scores the ranking strongly depends on the permutation scheme.

We will focus our illustration on the ranking of three exemplary predictor variables, "h2y8", "flex8" and "pol3", that are highlighted in Figure 8.3: We find in the unconditional view in the top panel of Figure 8.3 that "h2y8" and "flex8" appear to be of higher importance than "pol3" (ranks "h2y8": 8, "flex8": 9, "pol3": 11). However, in the conditional view in the bottom panel of Figure 8.3 their order is reversed and it turns out that "pol3" is really more important than "h2y8" and "flex8"(ranks "h2y8": 9, "flex8": 8, "pol3": 7). This change in the ranks of the predictor variables is most pronounced for large `mtry` as expected, but similar effects can be observed for smaller values.

When exploring the reason why the importances of "h2y8" and "flex8" are moderated by conditioning, while the importance of "pol3" remains almost constant, we find that "h2y8"

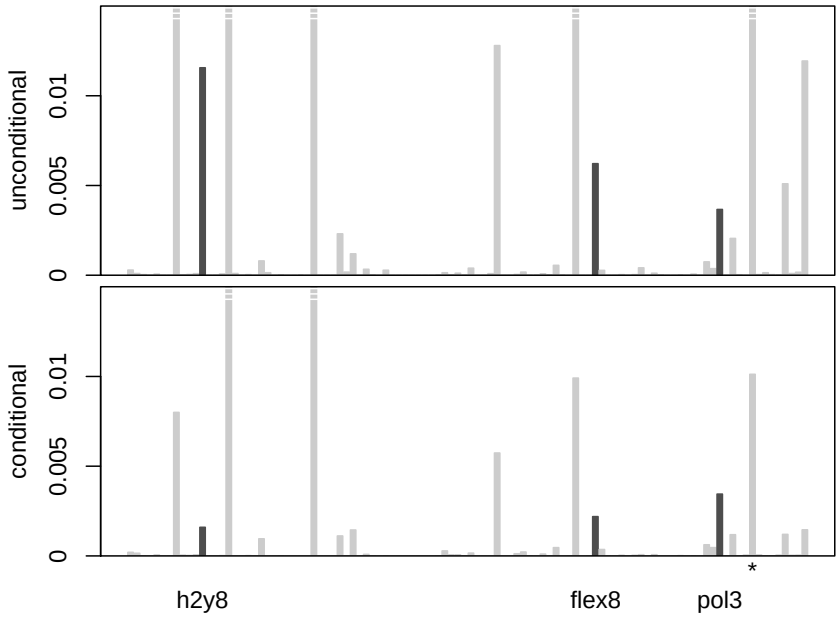


Fig. 8.3: Unconditional and conditional permutation importance of 104 predictors of peptide-binding.

and "flex8" are correlated with influential covariates, while "pol3" is only correlated with non-influential covariates. For example, "h2y8" is highly correlated with the polarity at position eight "pol8", that is indicated by the $*$ symbol in in Figure 8.3. The variable "pol8" shows a high importance (that is however also moderated by conditioning) and was already found to be influential by Segal et al. (2001), who note that it may approximate an effect of the eighth position in the original sequence data, while the results of Xia and Li (1998) indicate an effect of the amino acid property polarity itself.

This shows that importance rankings in data sets that contain complex correlations between predictor variables can be severely affected by the underlying permutation scheme: When the conditional permutation is used, the importance scores of correlated predictor are moderated such that the truly influential predictor variables have a higher chance to be detected.

8.4 Summary

We have investigated the sources of preferences in the variable importance measures of random forests in favor of correlated predictor variables and suggested a new, conditional permutation scheme for the computation of the variable importance measure. This new, conditional permutation scheme uses the partition that is automatically provided by the fitted model as a conditioning grid and reflects the true impact of each predictor variable better than the original, marginal approach. Even though the conditional permutation scheme cannot entirely eliminate the preference for correlated predictor variables, it has been shown to provide a more fair means of comparison that can help identify truly relevant predictor variables.

Our simulation results also illustrate the impact of the choice of the random forest tuning parameter `mtry`: While the default value `mtry`$= \sqrt{p}$ is often found to be optimal with respect to prediction accuracy in empirical studies (cf., e.g., Svetnik et al., 2003), our findings indicate that, in the case of correlated predictor variables, different values of `mtry` should be considered. However, it should also be noted that any interpretation of random forest variable importance scores can only be sensible when the number of trees is chosen sufficiently large such that the results produced with different random seeds do not vary systematically. Only then it is assured that the differences between, e.g., unconditional and conditional importance are not only due to random variation.

9. Conclusion and outlook

The aim of this work was to provide a statistical understanding of the sources of biased variable selection and variable importance measures in recursive partitioning methods, and to improve these measures such that they can be used to reliably assess and compare the relevance of predictor variables of different types.

For single classification trees employing empirical entropy measures as split selection criteria, we found that biased variable selection can be attributed to two very fundamental statistical issues, namely biased estimation and multiple testing effects. While the latter mechanism has been known in the machine learning community for some time, its negative carry over effects to variable importance measures in ensemble methods were not accounted for, and the former source of biased entropy estimation went unnoticed in previous studies. The criteria for unbiased split selection that were introduced and evaluated here, unbiased entropy estimates in robust k-ary splitting and the p-values of optimally selected statistics in binary splitting, amongst others, can effectively solve the problem of variable selection bias in single trees.

When, in order to produce more stable predictors, we leave the well interpretable single trees and turn towards ensembles of trees, the advantage of the TWIX approach – as opposed to the ensemble methods based on random resampling – is that its individual trees are nested and thus preserve some interpretability. For this method, an adaptive cutpoint selection criterion was suggested here, that can serve as a diagnostic of the stability of a split decision. Moreover, when this criterion is employed in the construction of a TWIX ensemble, the nested set of trees may reduce to a single, interpretable tree if the underlying

partition is sufficiently stable.

Otherwise, the ensemble methods bagging and random forests are more efficient with respect to prediction accuracy, but offer no means of interpreting the exact form and direction of the effect of a predictor variable. Therefore, different variable importance measures for ensemble methods have been suggested and are widely applied, for example in high dimensional problems with many noise variables. As opposed to univariate screening methods, that have been suggested previously for use in high dimensional problems, variable importance measures can reflect the impact of each predictor variable in both main effects and interactions.

When variable importance measures are used as a means of interpretation or variable selection, however, it is particularly important that these measures are reliable and comparable. As revealed by recent empirical studies, and the systematic simulation experiments and statistical reasoning presented here, this was not the case for the originally proposed measures, that show artificial preferences for variables of certain types.

Part of these artifacts can be attributed to the effects of variable selection bias that were already investigated for single classification trees in the early chapters of this work. Other effects are newly induced either by the bootstrap resampling scheme usually employed in ensemble methods, or by the construction of the variable importance measure itself. Solutions to both problems were presented: Our results indicate that bootstrap sampling induces artifacts in association measures used as split selection criteria in ensemble methods, and should be discarded in favor of subsampling. Another question that was emphasized was: What kind of importance is measured by the current variable importance scores in the first place – and is that what is desired?

In the context of recursive partitioning, where models cannot be derived in a closed form with all predictor variables processed simultaneously as, e.g, in generalized linear models, there are no coefficient estimates available that could serve as indicators of the relevance of a variable conditional on all covariates. This can be considered as the price one has

to pay for a model that is flexible, computationally feasible and applicable even to high dimensional data, because only one variable is processed at a time.

In order to be able to provide a conditional measure of variable importance from a model built with recursive partitioning, an additional effort is necessary in the computation of the importance measure. Only a conditional variable importance measure is capable of distinguishing between the marginal effect of a variable and its effect after potential correlations between the predictors are accounted for, as we have pointed out in the last chapter. The conditional permutation importance introduced here addresses one of the key remaining issues in the practical application of variable importance measures in genomics and proteomics, where predictor variables are often highly correlated, but also has the potential to clear the way for further applications, e.g., in the social sciences.

In summary, the descriptive variable importance measures introduced in this work provide a fair means of comparison for assessing the impact of predictor variables of different types in high dimensional problems involving interactions and even correlations between predictor variables. Besides their application as a merely descriptive tool, different schemes for deriving significance statements can be applied to aid the decision which and how many candidate predictors should be selected for further analysis. Some of these approaches were critically discussed here: While the significance test suggested by Breiman and Cutler (2008) should be discarded due to its poor statistical properties, the approaches of Diaz-Uriarte and Alvarez de Andrés (2006), Diaz-Uriarte (2007) and Rodenburg et al. (2008) in their original form show the same undesired preference as the marginal variable importance. This artifact can only be avoided when these approaches are used together with the conditional importance suggested here.

However, it may also be worth discussing if the overall null hypothesis that all predictor variables are irrelevant, that is implicitly presumed when permuting the response variable against the predictor matrix as in the approach of Rodenburg et al. (2008), is a desirable baseline for significance statements for the importance of individual variables. In addition to this, any approach that derives the distribution of the importance measure in real time

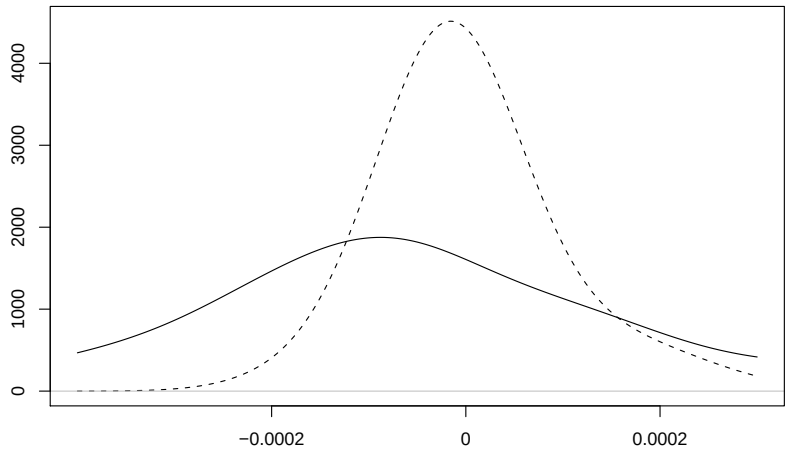


Fig. 9.1: Distribution (kernel density estimates) of permutation importance scores for balanced samples (dashed) and for strongly unbalanced samples with minority class probability 5% (solid) under the null hypothesis.

for the present model and data set is computationally expensive – if not prohibitive for large samples. Therefore, a better understanding of the distribution of variable importance measures and the parameters it depends on is a crucial field for further research.

We have seen here that the scale of importance measure may depend on model parameters, as well as characteristics of the data set itself. A related issue is the case of unbalanced response classes: When strongly unbalanced data are processed in random forests, the distribution of the variable importance has a higher variance, but is also systematically shifted to the left, so that negative importance values appear more frequently under the null hypothesis, as illustrated in Figure 9.1. Rather unintuitively, this indicates that, in average, the prediction accuracy for the oob-observations of a random forest with a randomly permuted random noise variable is higher than with the original random noise variable. The investigation of this effect may reveal additional insights into the characteristics of the permutation importance measure.

In the literature it is suggested that unbalanced class frequencies should be counterbalanced

either by means of incorporating class weights or loss functions as in a decision theoretic framework, or by means of "undersampling" (or "down sampling", i.e., sampling from the majority class as few observations as there are of the minority class; Chen et al., 2004). Accordingly, the results presented in this work apply to the case of balanced samples or unbalanced samples that were balanced by means of undersampling.

Besides the issues of variable selection and interpretability, that were treated here, open research questions in the area of ensemble methods include the effect of different model parameters and settings on the prediction accuracy. For example, we are currently investigating potential benefits of advanced aggregation schemes for ensemble methods: While the commonly implemented majority voting approach has been shown to give excellent prediction results in a multitude of standard settings, the evaluation of, e.g., weighted aggregation schemes is especially interesting in sensitive cases such as highly unbalanced samples.

An important issue in this context, that should be taken into account when evaluating the predictive performance of a classifier, is the distinction between the prediction accuracy with respect to the percentage of correctly classified observations (which is a rather coarse criterion for comparing aggregation schemes) or with respect to probability estimates (that may allow for a more fine graded comparison). In general, a high prediction accuracy of a classifier does not guarantee that the corresponding class probabilities are being estimated "(even remotely) accurately" (Friedman, 1997, p. 76). Therefore, future research should investigate in particular whether ensemble methods are merely good classifiers, as indicated by their excellent performance in a wide variety of simulation and real-data studies, or if they also make good probability estimators.

The latter would make them an attractive alternative to logistic regression – not only in classification problems but also, for example, in the estimation of propensity scores (i.e., the probability to receive treatment in quasi-experimental trials) in high dimensional problems. A first application of bagging for propensity score estimation (Luellen et al., 2005) has also fueled the discussion if random forests may overfit.

Most applied publications on random forests state somewhere in their introduction that random forests do not overfit. However, all these publications rely on one, apparently biased source: Breiman himself, who made this claim based on a non-representative data basis, as outlined in the introduction. The results of Lin and Jeon (2006) imply that the depth of the trees in a random forest, rather than the number of trees as suspected by Luellen et al. (2005), might induce overfitting. Thus, the impact of this model parameter on the prediction accuracy of ensemble methods, especially for predicting probabilities, will be further investigated in future research.

Bibliography

Abellán, J. and S. Moral (2003). Maximum of entropy for credal sets. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 11*(5), 587–597.

Abellán, J. and S. Moral (2005). Upper entropy of credal sets. Applications to credal classification. *International Journal of Approximate Reasoning 39*(2–3), 235–255.

Archer, K. J. and R. V. Kimes (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis 52*(4), 2249–2260.

Arun, K. and C. J. Langmead (2006). Structure based chemical shift prediction using random forests non-linear regression. In T. Jiang, U.-C. Yang, Y.-P. P. Chen, and L. Wong (Eds.), *Proceedings of the Fourth Asia-Pacific Bioinformatics Conference, Taipei, Taiwan*, pp. 317–326.

Baca-Garcia, E., M. M. Perez-Rodriguez, D. Saiz-Gonzalez, I. Basurte-Villamor, J. Saiz-Ruiz, J. M. Leiva-Murillo, M. de Prado-Cumplido, R. Santiago-Mozos, A. Artes-Rodriguez, and J. de Leon (2007). Variables associated with familial suicide attempts in a sample of suicide attempters. *Progress in Neuro-Psychopharmacology & Biological Psychiatry 31*(6), 1312–1316.

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*(3), 333–336.

Bauer, E. and R. Kohavi (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning 36*(1-2), 105–139.

Berk, R. A. (2006). An introduction to ensemble methods for data analysis. *Sociological Methods & Research 34*(3), 263–295.

Bernard, J.-M. (2004). An introduction to the Imprecise Dirichlet Model for multinomial data. *International Journal of Approximate Reasoning 39*(2–3), 123–150.

Bernard, J.-M. (2008). Special issue on the Imprecise Dirichlet Model. *International Journal of Approximate Reasoning*, to appear.

Bickel, P. J. and J.-J. Ren (2001). The bootstrap in hypothesis testing. In M. de Gunst, C. Klaassen, and A. van der Vaart (Eds.), *State of the Art in Probability and Statistics, Festschrift for Willem R. van Zwet, IMS Lecture Notes Monograph Series, Volume 36, Beachwood, OH, USA*, pp. 91–112.

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., C. Strobl, T. Augustin, and M. Daumer (2008). Evaluating microarray-based classifiers: An overview. *Cancer Informatics 4 4*, 77–97.

Breiman, L. (1996a). Bagging predictors. *Machine Learning 24*(2), 123–140.

Breiman, L. (1996b). Out-of-bag estimation. Technical report, Department of Statistics, University of California at Berkeley, CA, USA.

Breiman, L. (1998). Arcing classifiers. *The Annals of Statistics 26*(3), 801–849.

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

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science 16*(3), 199–231.

Breiman, L. and A. Cutler (2008). Random forests – Classification manual. Website accessed in 1/2008; http://www.math.usu.edu/∼adele/forests.

Breiman, L., A. Cutler, A. Liaw, and M. Wiener (2006). *Breiman and Cutler's Random Forests for Classification and Regression*. R package version 4.5-16.

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

Bronevich, A. G. (2005). On eventwise aggregation of coherent lower probabilities. In F. Cozman, R. Nau, and T. Seidenfeld (Eds.), *Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications, Carnegie Mellon University, Pittsburgh, PA, USA*, pp. 340–348. SIPTA, Manno.

Bühlmann, P. and B. Yu (2002). Analyzing bagging. *The Annals of Statistics 30*(4), 927–961.

Buja, A. and W. Stuetzle (2006). Observations on bagging. *Statistica Sinica 16*(2), 323–351.

Bureau, A., J. Dupuis, K. Falls, K. L. Lunetta, B. Hayward, T. P. Keith, and P. V. Eerdewegh (2005). Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology 28*(2), 171–182.

Candes, E. and T. Tao (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics 35*(6), 2313–2351.

Caruana, R. and A. Niculescu-Mizil (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning (ICML 2006)*, New York, NY, USA, pp. 161–168. ACM Press.

Chen, C., A. Liaw, and L. Breiman (2004). Using random forest to learn imbalanced data. Technical Report 666, Department of Statistics, University of California, Berkeley, CA, USA.

Coolen, F. and T. Augustin (2008). A nonparametric predictive alternative to the Imprecise Dirichlet Model: the case of a known number of categories. *International Journal of Approximate Reasoning*, to appear.

Cummings, M. P. and D. S. Myers (2004). Simple statistical models predict C-to-U edited sites in plant mitochondrial RNA. *BMC Bioinformatics 5:132*.

Cummings, M. P. and M. R. Segal (2004). Few amino acid positions in rpoB are associated with most of the rifampin resistance in Mycobacterium tuberculosis. *BMC Bioinformatics 5:137*.

de Cooman, G. and M. Troffaes (2004). Coherent lower previsions in systems modelling: products and aggregation rules. *Reliability Engineering and System Safety 85*(1–3), 113–134.

de Cooman, G. and M. Zaffalon (2004). Updating beliefs with incomplete observations. *Artificial Intelligence 159*(1–2), 75–125.

Diaz-Uriarte, R. (2007). `GeneSrF` and `varselrf`: A web-based tool and R package for gene selection and classification using random forest. *BMC Bioinformatics 8:328*.

Diaz-Uriarte, R. and S. Alvarez de Andrés (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics 7:3*.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning 40*(2), 139–157.

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

Domingos, P. (1997). Why does bagging work? A bayesian account and its implications. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, Newport Beach, CA, USA*, pp. 155–158. AAAI Press.

Feraud, R. and F. Clerot (2002). A methodology to explain neural network classification. *Neural Networks 15*(2), 237–246.

Freund, Y. and R. E. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences 55*(1), 119–139.

Friedman, J. (1997). On bias, variance, 0/1–loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery 1*(1), 55–77.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics 29*(5), 1189–1232.

Friedman, J. and P. Hall (1999). On bagging and nonlinear estimation. Technical report, Department of Statistics, Stanford University, CA, USA.

Friedman, J. H. (2006). Comment: Classifier technology and the illusion of progress. *Statistical Science 21*(1), 15–18.

Furlanello, C., M. Neteler, S. Merler, S. Menegon, S. Fontanari, D. Donini, A. Rizzoli, and C. Chemini (2003). `GIS` and the `randomForest` predictor: Integration in `R` for Tick-Borne desease risk assessment. In F. L. K. Hornik and A. Zeileis (Eds.), *Proceedings of the 3rd international workshop on distributed statistical computing DSC*.

Gatnar, E. (2008). Fusion of multiple statistical classifers. In H. H. Bock, W. Gaul, and M. Vichi (Eds.), *Proceedings of the 31st Annual Conference of the German Classification Society (GfKl), Freiburg i. Br., Germany,* to appear.

Good, P. (2005). *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. New York: Springer Series in Statistics, 3rd Edition.

Grandvalet, Y. (2004). Bagging equalizes influence. *Machine Learning 55*(3), 251–270.

Guha, R. and P. C. Jurs (2003). Development of linear, ensemble, and nonlinear models for the prediction and interpretation of the biological activity of a set of PDGFR inhibitors. *Journal of Chemical Information and Computer Sciences 44*(6), 2179–2189.

Gunther, E. C., D. J. Stone, R. W. Gerwien, P. Bento, and M. P. Heyes (2003). Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro. *Proceedings of the National Academy of Sciences 100*(16), 9608–9613.

Hampel, F. (1980). Robuste Schätzungen: Ein anwendungsorientierter Überblick. *Biometrical Journal 22*(1), 3–21.

Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science 21*(1), 1–14.

Hannöver, W., M. Richard, N. B. Hansen, Z. Martinovich, and H. Kordy (2002). A classification tree model for decision-making in clinical practice: An application based on the data of the German multicenter study on eating disorders, project TR-EAT. *Psychotherapy Research 12*(4), 445–461.

Hastie, T., R. Tibshirani, and J. H. Friedman (2001). *The Elements of Statistical Learning*. New York: Springer.

Heidema, A. G., J. M. A. Boer, N. Nagelkerke, E. C. M. Mariman, D. L. van der A, and E. J. M. Feskens (2006). The challenge for genetic epidemiologists: How to analyze large numbers of SNPs in relation to complex diseases. *BMC Genetics 7:23*.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: A tutorial. *Statistical Science 14*(4), 382–417.

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

Hothorn, T., K. Hornik, and A. Zeileis (2008). `party`: A laboratory for recursive part(y)itioning. R package version 0.9-96.

Hothorn, T., F. Leisch, A. Zeileis, and K. Hornik (2005). The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics 14*(3), 675–699.

Huang, X., W. Pan, S. Grindle, X. Han, Y. Chen, S. J. Park, L. W. Miller, and J. Hall (2005). A comparative study of discriminating human heart failure etiology using gene expression profiles. *BMC Bioinformatics 6:205*.

Jensen, D. D. and P. R. Cohen (2000). Multiple comparisons in induction algorithms. *Machine Learning 38*(3), 309–338.

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

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

Klir, G. J. (1999). Uncertainty and information measures for imprecise probabilities: An overview. In de Cooman, Cozman, Moral, and Walley (Eds.), *Proceedings of the First International Symposium on Imprecise Probabilities and their Applications, Ghent, Belgium*.

Klir, G. J. (2003). An update on generalized information theory. In Z. Bernard, Seidenfeld (Ed.), *Proceedings of the Third International Symposium on Imprecise Probabilities and their Applications, Lugano, Switzerland*, pp. 321–334. Carleton Scientific.

König, I., J. D. Malley, C. Weimar, H.-C. Diener, and A. Ziegler (2007). Practical experiences on the necessity of external validation. *Statistics in Medicine 26*(30), 5499 – 5511.

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

Lausen, B., T. Hothorn, F. Bretz, and M. Schumacher (2004). Assessment of optimal selected prognostic factors. *Biometrical Journal 46*(3), 364–374.

Lemaire, V. and F. Clerot (2006). An input variable importance definition based on empirical data probability distribution. In I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh (Eds.), *Feature Extraction, Foundations and Applications,* Series Studies in Fuzziness and Soft Computing, Number 26, Part II. Springer.

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

Liebetrau, A. M. (1983). *Measures of Association.* Newbury Park: Sage.

Lin, Y. and Y. Jeon (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association 101*(474), 578–590.

Little, R. and D. Rubin (1986). *Statistical Analysis with Missing Data.* New York, NY, USA: John Wiley & Sons, Inc.

Little, R. and D. Rubin (2002). *Statistical Analysis with Missing Data, 2nd Edition.* Hoboken, NJ, USA: John Wiley & Sons, Inc.

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

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

Luellen, J. K., W. R. Shadish, and M. H. Clark (2005). Propensity scores: An introduction and experimental test. *Evaluation Review 29*(6), 530–558.

Lunetta, K. L., L. B. Hayward, J. Segal, and P. V. Eerdewegh (2004). Screening large-scale association study data: Exploiting interactions using random forests. *BMC Genetics 5:32*.

Marinic, I., F. Supek, Z. Kovacic, L. Rukavina, T. Jendricko, and D. Kozaric-Kovacic (2007). Posttraumatic stress disorder: Diagnostic data analysis by data mining methodology. *Croatian Medical Journal 48*(2), 185–197.

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

Moral, S. and J. del Sagrado (1987). Aggregation of imprecise probabilities. In B. Bouchon-Meunier (Ed.), *Aggregation and Fusion of Imperfect Information*, pp. 162–188. Physica Verlag.

Morgan, J. N. and J. A. Sonquist (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association 58*(302), 415–434.

Nason, M., S. Emerson, and M. Leblanc (2004). CARTscans: A tool for visualizing complex models. *Journal of Computational and Graphical Statistics 13*(4), 1–19.

Nicodemus, K. and Y. Y. Shugart (2007). Impact of linkage disequilibrium and effect size on the ability of machine learning methods to detect epistasis in case-control studies. In *Abstract volume of the Sixteenth Annual Meeting of the International Genetic Epidemiology Society, North Yorkshire, UK*, Volume 31 (6), pp. 611.

Oh, J., M. Laubach, and A. Luczak (2003). Estimating neuronal variable importance with random forest. In *Proceedings of the 29th Annual IEEE Bioengineering Conference, New Jersey Institute of Technology, Newark, NJ, USA*, pp. 33–34.

Paneque, D., A. Borgland, A. Bovier, E. Bloom, Y. Edmonds, S. Funk, G. Godfrey, R. Rando, L. Wai, and P. Wang (2007). Novel technique for monitoring the performance of the LAT instrument on board the GLAST satellite. In *Proceedings of the First*

GLAST Symposium, Stanford, CA, USA, Volume 921, pp. 562–563. American Institute of Physics.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine 6*(3), 21–45.

Politis, D. N., J. P. Romano, and M. Wolf (1999). *Subsampling*. New York: Springer.

Potapov, S. (2006). Analyse von Bäumen. Unpublished diploma-thesis, Department of Mathematics, Universität Augsburg, Germany.

Potapov, S. (2007). *TWIX: Trees WIth eXtra Splits*. R package version 0.2.4.

Potapov, S., M. Theus, and S. Urbanek (2006). *TWIX: Trees WIth eXtra Splits*. Presentation slides from the Third Ensemble Workshop of the Statistical Computing task group of the German Section of the International Biometric Society, Munich, Germany; http://www.rosuda.org/TWIX.

Qi, Y., Z. Bar-Joseph, and J. Klein-Seetharaman (2006). Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins 63*(3), 490–500.

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

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

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge, UK: Cambridge University Press.

Rodenburg, W., A. G. Heidema, J. M. Boer, I. M. Bovee-Oudenhoven, E. J. Feskens, E. C. Mariman, and J. Keijer (2008). A framework to identify physiological responses

in microarray based gene expression studies: Selection and interpretation of biologically relevant genes. *Physiological Genomics 33*(1), 78–90.

Rossi, A., F. Amaddeo, M. Sandri, and M. Tansella (2005). Determinants of once-only contact in a community-based psychiatric service. *Social Psychiatry and Psychiatric Epidemiology 40*(1), 50–56.

Roulston, M. (1999). Estimating the errors on measured entropy and mutual information. *Physica D 125*(3), 285–294.

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, Ludwig-Maximilians-Universität München, Germany.

Schürmann, T. (2004). Bias analysis in entropy estimation. *Journal of Physics A 37*(27), 295–301.

Segal, M. R. (2004). Machine learning benchmarks and random forest regression. Technical Report, Center for Bioinformatics & Molecular Biostatistics papers, University of California, San Francisco, CA, USA.

Segal, M. R., J. D. Barbour, and R. M. Grant (2004). Relating HIV-1 sequence variation to replication capacity via trees and forests. *Statistical Applications in Genetics and Molecular Biology 3*(1). Article 2.

Segal, M. R., M. P. Cummings, and A. E. Hubbard (2001). Relating amino acid sequence to phenotype: Analysis of peptide-binding data. *Biometrics 57*(2), 632–643.

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

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

Shih, Y.-S., D. Seligson, A. S. Belldegrun, A. Palotie, and S. Horvath (2005). Tumor classification by tissue microarray profiling: Random forest clustering applied to renal cell carcinoma. *Modern Pathology 18*(4), 547–557.

Shih, Y.-S. and H. Tsai (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 F. Cozman, R. Nau, and T. Seidenfeld (Eds.), *Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications, Pittsburgh, PA, USA*, pp. 340–348. SIPTA, Manno.

Strobl, C. and T. Augustin (2008). Adaptive selection of extra cutpoints – an approach towards reconciling robustness and interpretability in classification trees. *Journal of Statistical Theory and Practice*, to appear.

Strobl, C., A.-L. Boulesteix, and T. Augustin (2007). Unbiased split selection for classification trees based on the Gini index. *Computational Statistics & Data Analysis 52*(1), 483–501.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis (2008). Conditional variable importance for random forests. *BMC Bioinformatics 9:307*.

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

Strobl, C., J. Malley, and G. Tutz (2008). Random forests for regression, classification and assessment of variable importance in psychological research. Manuscript, Department of Statistics, Ludwig-Maximilians-Universität München, Germany.

Strobl, C. and A. Zeileis (2008). Danger: High power! – exploring the statistical properties of a test for random forest variable importance. In *Proceedings of the 18th International Conference on Computational Statistics, Porto, Portugal*.

Svejdar, V., T. Augustin, and C. Strobl (2008). Variable and split selection in classification trees based on the Gini index – What if values are missing not at random? In L. A. Hothorn, U. Mansmann, G. Tutz, U. Burger, and S. Mejza (Eds.), *Abstract book of the lifestat2008 conference of the German Biometrical Society, Munich, Germany*, Berlin, pp. 163. Lehmanns Media.

Svetnik, V., A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston (2003). Random forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences 43*(6), 1947–1958.

Svetnik, V., A. Liaw, C. Tong, and T. Wang (2004). Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. In F. Roli, J. Kittler, and T. Windeatt (Eds.), *Lecture Notes in Computer Science: Multiple Classifier systems*, Berlin/Heidelberg, pp. 334–343. Springer.

Troffaes, M. (2006). Generalising the conjunction rule for aggregating conflicting expert opinions. *International Journal of Intelligent Systems 21*(3), 361–380.

Troffaes, M. (2007). Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning 45*(1), 17–29.

Utkin, L. and T. Augustin (2007). Decision making under incomplete data using the Imprecise Dirichlet Model. *International Journal of Approximate Reasoning 44*(3), 322–338.

van der Laan, M. (2006). Statistical inference for variable importance. *International Journal of Biostatistics 2*(1). Article 2.

van Os, B. J. and J. Meulman (2005). Globally optimal tree models. In S. Azen, E. Kontoghiorghes, and J. C. Lee (Eds.), *Abstract Book of the 3rd World Conference on Computational Statistics & Data Analysis of the International Association for Statistical Computing, Cyprus, Greece*, pp. 79. Matrix Computations and Statistics Group.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. New York: Springer.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall.

Walley, P. (1996). Inferences from multinomial data: Learning from a bag of marbles. *Journal of the Royal Statistical Society B 58*(1), 3–57.

Ward, M. M., S. Pajevic, J. Dreyfuss, and J. D. Malley (2006). Short-term prediction of mortality in patients with systemic lupus erythematosus: Classification of outcomes using random forests. *Arthritis and Rheumatism 55*(1), 74–80.

Weichselberger, K. (2001). *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I: Intervallwahrscheinlichkeit als umfassendes Konzept*. Berlin: Springer.

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

Wösthoff, J. L. (2008). Moderne Klassifikationsverfahren in der Biometrie – Einfluss der Stichprobengrösse beim resampling in random forests. Unpublished bachelor-thesis under the supervision of T. Augustin and C. Strobl, Department of Statistics, Ludwig-Maximilians-Universität München, Germany.

Wu, B., T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao (2003). Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics 19*(13), 1636–1643.

Xia, X. and W.-H. Li (1998). What amino acid properties affect protein evolution? *Journal of Molecular Evolution 47*(5), 557–564.

Zaffalon, M. (2002a). Exact credal treatment of missing data. *Journal of Statistical Planning and Inference 105*(1), 105–122.

Zaffalon, M. (2002b). The naive credal classifier. *Journal of Statistical Planning and Inference 105*(1), 5–21.

Zaffalon, M. (2005). Conservative rules for predictive inference with incomplete data. In F. Cozman, R. Nau, and T. Seidenfeld (Eds.), *Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications, Pittsburgh, PA, USA*, pp. 406–415. SIPTA, Manno.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B 67*(2), 301–320.

Curriculum Vitae

Carolin Strobl

born November 6th, 1978, in Gräfelfing, Germany

Education and employment

since 2/2005	**Ludwig-Maximilians-Universität Munich, Germany** PhD student and research assistant at the Department of Statistics advisor: Prof. Dr. T. Augustin
10/2002-1/2005	**Ludwig-Maximilians-Universität Munich, Germany** graduate student in statistics, degree: M. Sc. research assistant at the Department for Medical Information Processing, Biometrics and Epidemiology (IBE) and at the Department of Statistics
10/1998-9/2002	**University of Regensburg, Germany** student in psychology, degree: Dipl.-Psych.
4/2000-9/2002	student assistant at the Department of Experimental Psychology
3/2001-4/2001	**Max-Planck-Institute for Psychological Research** research internship
9/1996-7/1998	**Theresien-Gymnasium, Munich, Germany** degree: Abitur (high school diploma)
8/1995-7/1996	**Northwood School, Lake Placid, NY, USA** exchange student, ASSIST-scholarship holder
9/1988-8/1995	**Viscardi-Gymnasium, Fürstenfeldbruck, Germany**

Bernd-Streitberg laureate	of the German Section of the International Biometric Society 2007
Women's representative	of the Faculty for Mathematics, Computer Sciences and Statistics, Ludwig-Maximilians-Universität Munich, Germany

Curriculum Vitae

Carolin Strobl

geboren am 6. November 1978, in Gräfelfing

Ausbildung und Beschäftigung

seit 2/2005	**Ludwig-Maximilians-Universität München** Promotionsstudentin und wissenschaftliche Mitarbeiterin am Institut für Statistik, Betreuung: Prof. Dr. T. Augustin
10/2002-1/2005	**Ludwig-Maximilians-Universität München** Aufbaustudium der Statistik, Abschluß: M. Sc. Hilfskraft am Institut für medizinische Informationsverarbeitung, Biometrie und Epidemiologie (IBE) und am Institut für Statistik
10/1998-9/2002	**Universität Regensburg** Studium der Psychologie, Abschluß: Dipl.-Psych.
4/2000-9/2002	Hilfskraft am Institut für Experimentelle Psychologie
3/2001-4/2001	**Max-Planck-Institut für Psychologische Forschung** Forschungspraktikum
9/1996-7/1998	**Theresien-Gymnasium, München** Abschluß: Abitur
8/1995-7/1996	**Northwood School, Lake Placid, NY, USA** Austauschschülerin, ASSIST - Stipendiatin
9/1988-8/1995	**Viscardi-Gymnasium, Fürstenfeldbruck**

Bernd-Streitberg Preisträgerin	der Deutschen Sektion der Internationalen Biometrischen Gesellschaft 2007
Frauenbeauftragte	der Fakultät für Mathematik, Informatik und Statistik, Ludwig-Maximilians-Universität München