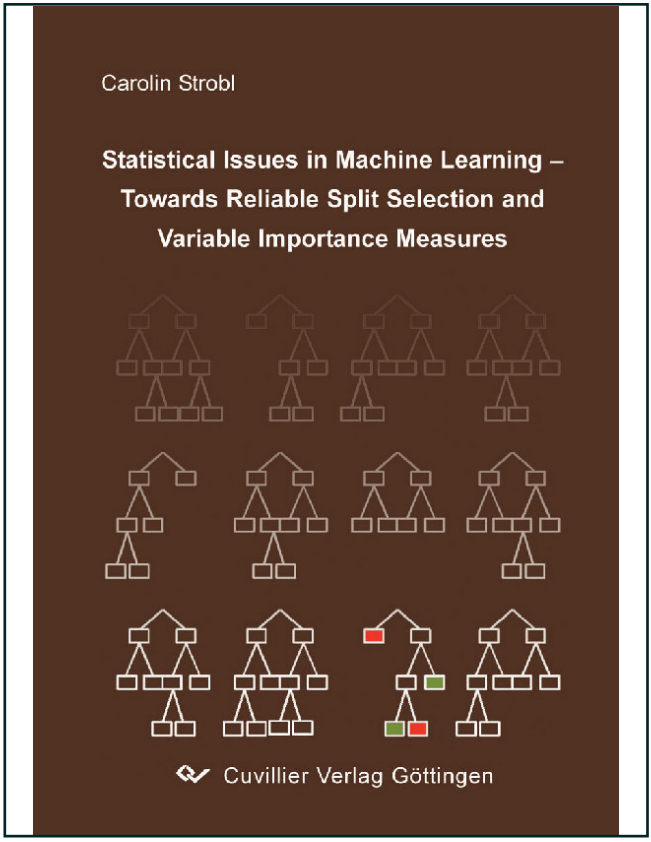
Carolin Strobl (Autor)

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



https://cuvillier.de/de/shop/publications/1400

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.

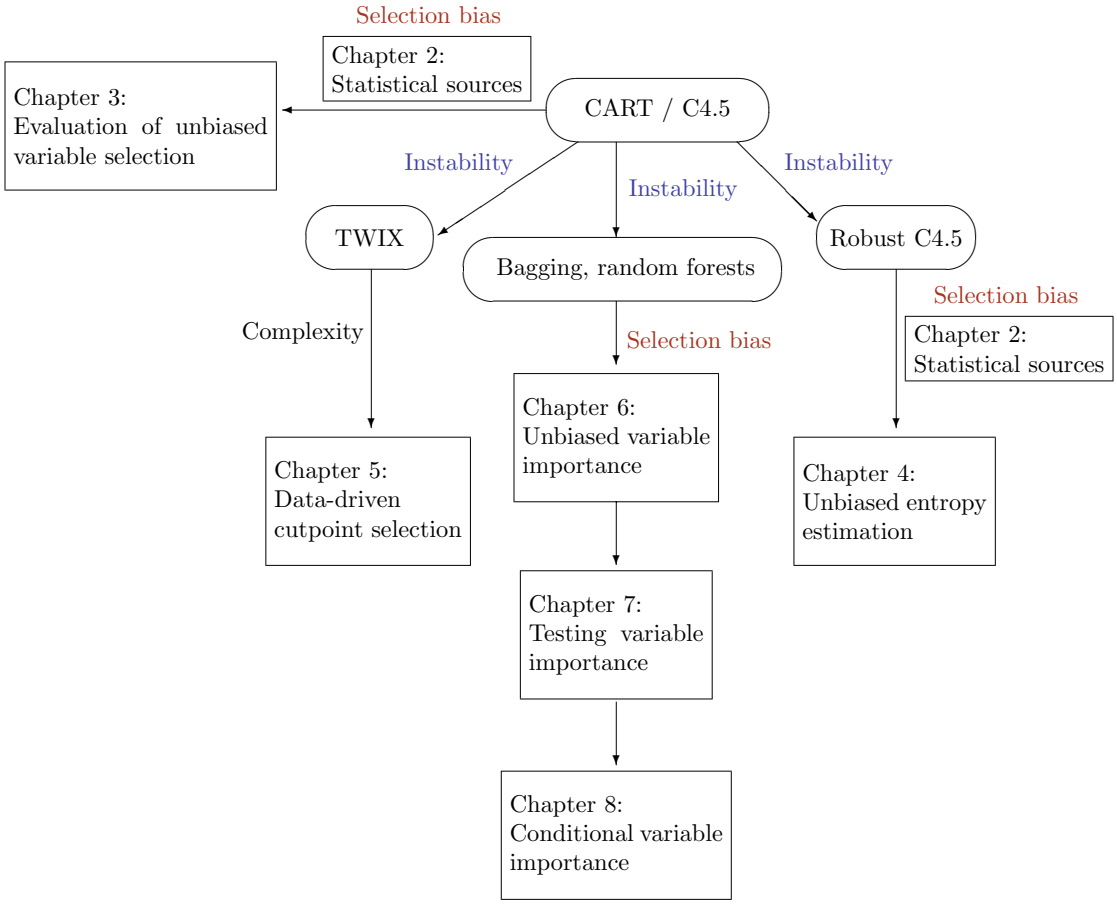
In such applications it is essential that variable importance measures are reliable. However, there are at least two situations where the originally proposed methods show undesired artifacts: the case of predictor variables of different types and the case of correlated predictor variables. In Chapter 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