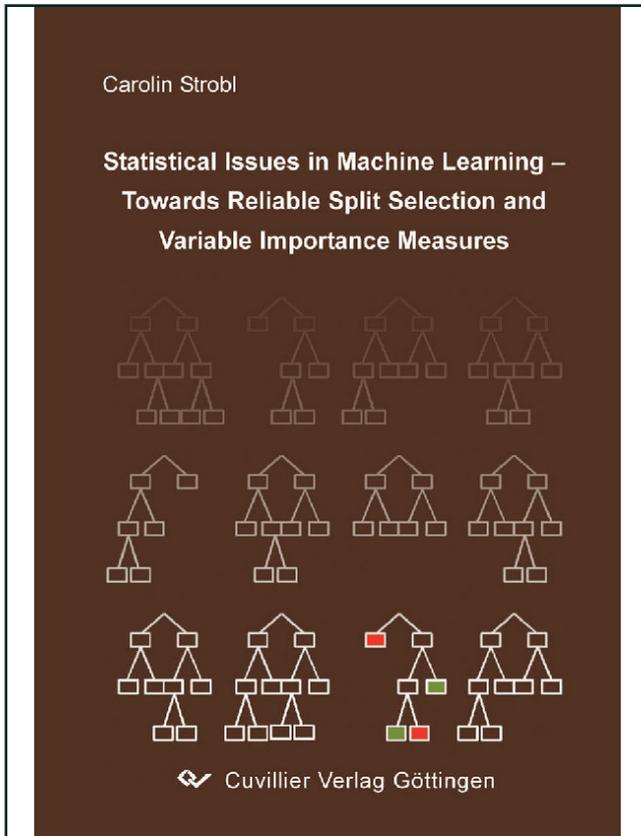




Carolin Strobl (Autor)

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



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

Copyright:

Cuvillier Verlag, Inhaberin Annette Jentzsch-Cuvillier, Nonnenstieg 8, 37075 Göttingen, Germany

Telefon: +49 (0)551 54724-0, E-Mail: info@cuvillier.de, Website: <https://cuvillier.de>

Contents

Scope of this work	vi
1. Introduction	1
1.1 Classification trees	5
1.1.1 Split selection and stopping rules	5
1.1.2 Prediction and interpretation	10
1.1.3 Variable selection bias and instability	13
1.2 Robust classification trees and ensemble methods	16
1.3 Characteristics and caveats	19
1.3.1 “Small n large p ” applicability	19
1.3.2 Out-of-bag error estimation	21
1.3.3 Missing value handling	22
1.3.4 Randomness and stability	22
2. Variable selection bias in classification trees	25
2.1 Entropy estimation	28
2.1.1 Binary splitting	28

2.1.2	k -ary splitting	32
2.2	Multiple comparisons in cutpoint selection	34
2.3	Summary	35
3.	Evaluation of an unbiased variable selection criterion	37
3.1	Optimally selected statistics	38
3.2	Simulation studies	40
3.2.1	Null case	41
3.2.2	Power case I	42
3.2.3	Power case II	43
3.3	Application to veterinary data	46
3.3.1	Variable selection ranking	47
3.3.2	Selected splitting variables	47
3.4	Summary	48
4.	Robust and unbiased variable selection in k-ary splitting	54
4.1	Classification trees based on imprecise probabilities	55
4.1.1	Total impurity criteria	57
4.1.2	Split selection procedure	59
4.1.3	Characteristics of the total impurity criterion TU2	60
4.2	Empirical entropy measures in split selection	64
4.2.1	Estimation bias for the empirical Shannon entropy	64
4.2.2	Effects in classification trees based on imprecise probabilities	65

4.2.3	Suggested corrections based on the IDM	67
4.3	Simulation study	68
4.4	Summary	69
5.	Adaptive cutpoint selection in TWIX ensembles	77
5.1	Building TWIX ensembles	79
5.1.1	Instability of cutpoint selection in recursive partitioning	80
5.1.2	Selecting extra cutpoints	81
5.2	A new, adaptive criterion for selecting extra cutpoints	83
5.2.1	Adding virtual observations	84
5.2.2	Recomputation of the split criterion	85
5.3	Behavior of the adaptive criterion	88
5.3.1	Application to olives data	89
5.3.2	Simulation study	91
5.4	Outlook on credal prediction and aggregation schemes	93
5.4.1	Credal prediction rules	93
5.4.2	Aggregation schemes	96
5.5	Summary	97
6.	Unbiased variable importance in random forests and bagging	99
6.1	Random forest variable importance measures	100
6.2	Simulation studies	102
6.2.1	Null case	105

6.2.2	Power case	107
6.3	Sources of variable importance bias	111
6.3.1	Variable selection bias in individual classification trees	112
6.3.2	Effects induced by bootstrapping	113
6.4	Application to C-to-U conversion data	115
6.5	Summary	118
7.	Statistical properties of Breiman and Cutler's test	130
7.1	Investigating the current test	131
7.1.1	The power	131
7.1.2	The construction of the z -score	133
7.1.3	Specifying the null hypothesis	134
7.2	Summary	135
8.	Conditional variable importance	138
8.1	Variable selection in random forests	143
8.1.1	Simulation design	144
8.1.2	Illustration of variable selection	145
8.2	A second look at the permutation importance	147
8.2.1	Background: Types of independence	147
8.2.2	A new, conditional permutation scheme	150
8.2.3	Simulation results	153
8.3	Application to peptide-binding data	156
8.4	Summary	158

9. Conclusion and outlook	159
Bibliography	165