Variable Selection Bias in Classification Trees and Ensemble Methods

Carolin Strobl, Achim Zeileis,
Anne-Laure Boulesteix, Torsten Hothorn

useR! 2006

Variable selection bias
Random forests
Implication
References

Variable selection bias

- variable selection in classification trees is affected by characteristics other than information content, e.g. variables with more categories are preferred

e.g. Breiman, Friedman, Olshen, and Stone (1984), Kim and Loh (2001), Dobra and Gehrke (2001)

Variable selection bias
rpart
Random forests
Implication
References

Standard simulation design

- binary response Y
- uninformative predictor variables $X_1, \ldots, X_p$
- with different numbers of categories
- record relative frequency (e.g. out of 1000 iterations) for each variable to be selected for the first split

Variable selection bias
rpart
Random forests
Implication
References

Standard simulation design

$$Y \in \{1, 2\}$$

$$\begin{aligned} X_1 &\in \{1, \ldots, 20\} \\ X_2 &\in \{1, \ldots, 10\} \\ X_3 &\in \{1, \ldots, 4\} \\ X_4 &\in \{1, 2\} \end{aligned}$$

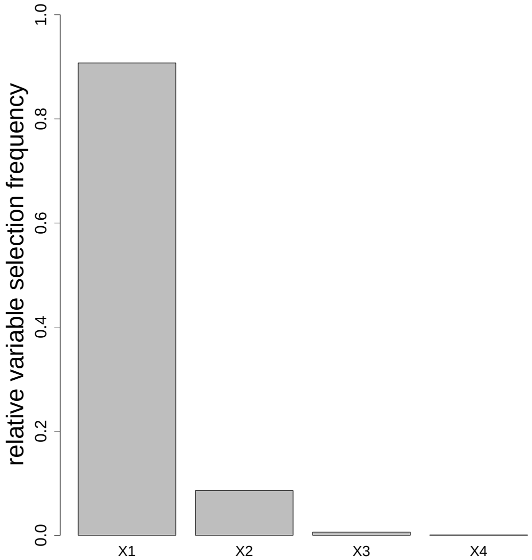
sampled independently

Variable selection bias
rpart
Random forests
Implication
References

Variable selection bias in classification trees

package: rpart

function: rpart



Sources of variable selection bias

- estimation bias and variance of empirical entropy measures (Strobl, Boulesteix, and Augustin, 2005)
- in binary splitting: multiple testing (combined variable and cutpoint selection)

Random forests

package: randomForest

functions: randomForest, importance

variable importance measure:
permutation accuracy importance

"In every tree grown in the forest, put down the oob cases and count the number of votes cast for the correct class. Now randomly permute the values of variable X_j in the oob cases and put these cases down the tree. Subtract the number of votes for the correct class in the variable-j-permuted oob data from the number of votes for the correct class in the untouched oob data. The average of this number over all trees in the forest is the raw importance score for variable X_j."

Permutation accuracy importance

- informative variables produce a systematic decrease in accuracy when permuted
- uninformative variables produce a random decrease or increase in accuracy when permuted

Permutation accuracy importance

employed as a criterion for variable selection in many recent publications in biochemistry, neurology, forestry, etc., e.g. by

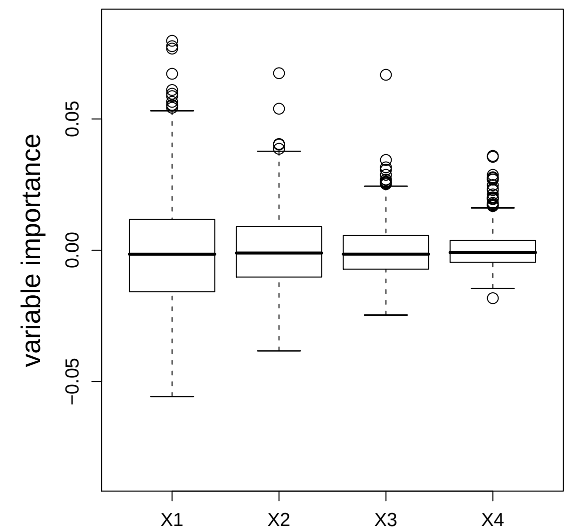
Bureau et al. (2005), Chen and Lin (2005), Cummings and Segal (2004), Diaz-Uriarte and de Andrés (2006), Furlanello et al. (2003), Guha and Jurs (2003), Jong et al. (2005), Lunetta et al. (2003), Lunetta et al. (2004), Ward et al. (2006) etc.

Permutation accuracy importance

Variable Selection Bias in Ensemble Methods
Variable selection bias
Random forests
randomForest
cforest
Implication
References

function: importance

option: scale=FALSE

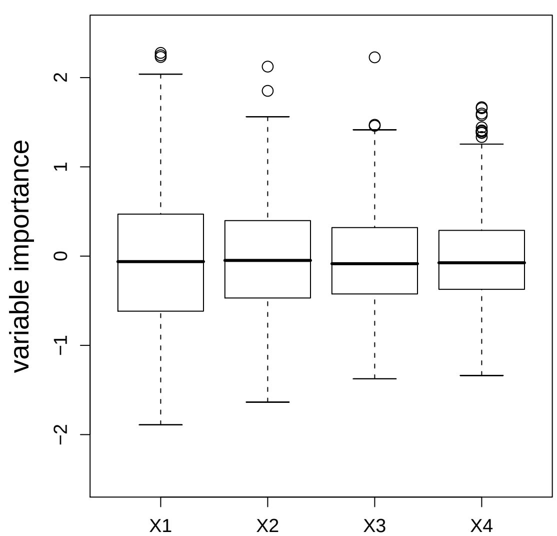


Permutation accuracy importance

Variable Selection Bias in Ensemble Methods
Variable selection bias
Random forests
randomForest
cforest
Implication
References

function: importance

option: scale=TRUE



Permutation accuracy importance

Variable Selection Bias in Ensemble Methods
Variable selection bias
Random forests
randomForest
cforest
Implication
References

- due to variable selection bias in individual trees
  $\Rightarrow$ variables with more categories end up closer to root node of individual tree
- potential change in accuracy is more pronounced for variables closer to root node
  $\Rightarrow$ variable importance of variables with more categories shows higher deviation

Expectation

random forests built from unbiased trees

do not produce biased variable selection measures

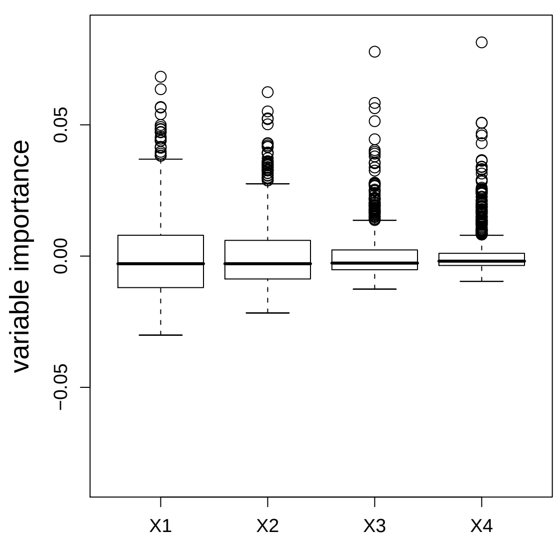
Unbiased variable selection criteria for classification trees

- Strobl, Boulesteix, and Augustin (2005)
  exact p-value of maximally selected Gini gain
  package: exactmaxsel
  function: maxsel.test
- Hothorn, Hornik, and Zeileis (2006)
  p-value of independence test in conditional inference framework
  package: party
  functions: ctree, cforest
  internal: party:::varimp

Variable Selection Bias in Ensemble Methods
Variable selection bias
Random forests
randomForest
cforest
Implication
References

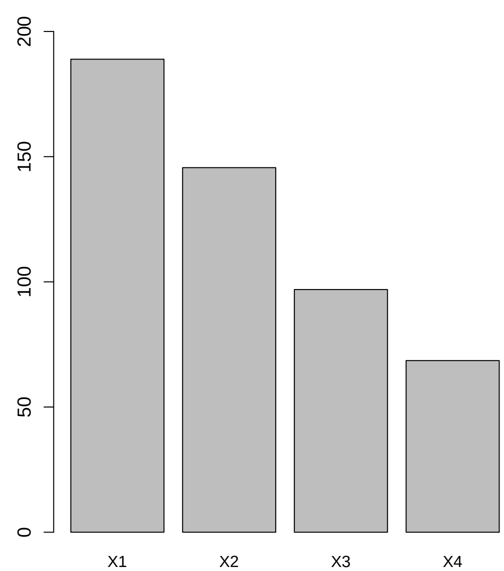
Permutation accuracy importance

internal: party:::varimp



Variable Selection Bias in Ensemble Methods
Variable selection bias
Random forests
randomForest
cforest
Implication
References

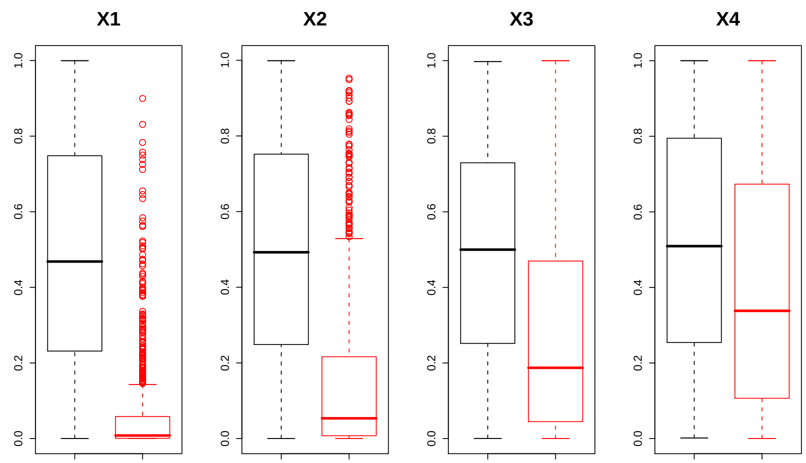
Number of times variable is selected in individual trees



Variable Selection Bias in Ensemble Methods
Variable selection bias
Random forests
randomForest
cforest
Implication
References

Bootstrap bias

distribution of the p-values of a χ^2-test before and after bootstrapping (1000 iterations, each n = 10 000)



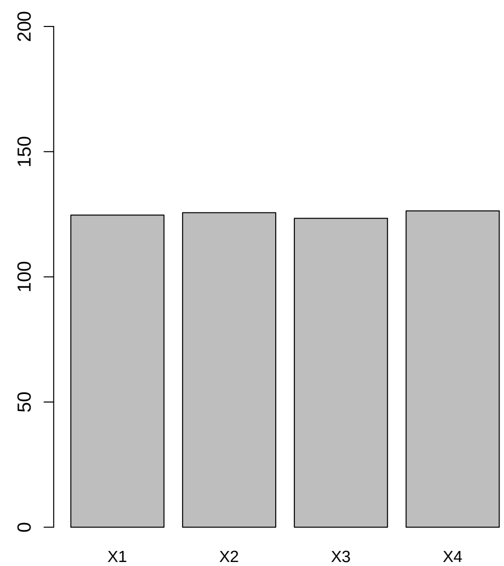
Bootstrap bias

- bootstrap sampling with replacement artificially induces an association
- the effect is more pronounced for contingency tables with more cells and more df

Expectation

when samples (e.g. of the size $0.632 \cdot n$) are drawn without replacement the bias is eliminated

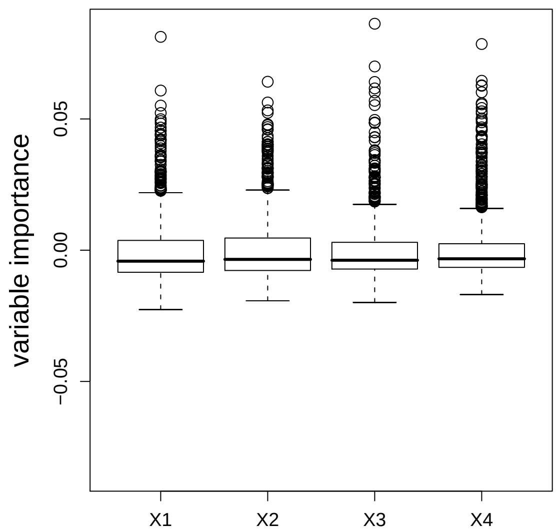
Number of times variable is selected in individual trees



Permutation accuracy importance

internal: party:::varimp

option: replace=FALSE



Implication

if your potential predictors vary in their number of categories or scale level

- use variable importance of unbiased **cforest**
- with option **replace=FALSE**

for the evaluation of variable importance and for variable selection

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

Chen, Y.-W. and C.-J. Lin (2005). Combining SVMs with various feature selection strategies. In M. N. I. Guyon, S. Gunn and L. Zadeh (Eds.), *Feature extraction, Foundations and Applications*.

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

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

Guha, R. and P. 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*, 2179–2189.

Hothorn, T., K. Hornik, and A. Zeileis (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics (to appear)*.

Jong, O., M. Laubach, and A. Luczak (2005). Estimating neuronal variable importance with random forest. In *Proceedings of 29th Annual Northeast Bioengineering Conference*, pp. 33–34.

Lunetta, K., L. Hayward, J. Segal, and P. V. Eerdewegh (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences 43*, 1947–1958.

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

Strobl, C., A.-L. Boulesteix, and T. Augustin (2005). Unbiased split selection for classification trees based on the Gini Index. *SFB-Discussion Paper 464, Department of Statistics, University of Munich LMU*.

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