technische universität dortmund

HEINRICH HEINE UNIVERSITÄT DÜSSELDORF

Identifikation von Risikofaktoren in der koronaren Herzchirurgie

Julia Schiffner[1] Erhard Godehardt[2] Stefanie Hillebrand[1]
Alexander Albert[2] Artur Lichtenberg[2] Claus Weihs[1]

[1]Fakultät Statistik, Technische Universität Dortmund

[2]Klinik für Kardiovaskuläre Chirurgie, Universitätsklinikum Düsseldorf, Heinrich-Heine Universität

12. November 2011

Outline

Quality Improvement in Medical Care

- German hospitals are obligated to release quality reports
- G-BA (Gemeinsamer Bundesausschuss) decides on service areas and quality indicators
- hospitals collect data
- data are submitted to external institutions for analysis
- results are reported to hospitals
  $\Rightarrow$ hospitals compare own quality with other hospitals
  $\Rightarrow$ hospitals develop strategies for quality improvement
- e. g. quality report 2010: data from 1800 hospitals, 30 service areas, 400 quality indicators

http://www.sqg.de/themen/qualitaetsreport,

http://www.g-ba.de/institution/presse/pressemitteilungen/399/

Quality Improvement in Coronary Bypass Surgery

- patients who undergo an isolated coronary bypass surgery
- different quality indicators: compliance with certain standards and postoperative complications
- risk adjustment in order to allow for comparability of different hospitals, logistic regression
- predictors: preoperative state of patient
- binary target variable: recovery state of patient
- ratio/difference of observed and expected amount of complications used to assess quality
- logistic regression model updated regularly:
  risk factors
  definition and number of categories for categorical predictors
- logistic KCH Score
  (KCH = Koronarchirurgie, coronary surgery)
  2005–2006: KCH Score 1.0
  2007: KCH Score 2.0
  2008–2010: KCH Score 3.0

Aims

Data for Clinic of Cardiovascular Surgery, Heinrich-Heine University, Düsseldorf

Identification of risk factors in coronary bypass surgery

- quality report: global model for all hospitals
  advantage: based on huge amount of data
- build individual prediction model for single hospitals
  most important: find individual risk factors
- compare prediction models / importance of predictors

Data

Data from 2007 and 2008 for Clinic of Cardiovascular Surgery, Heinrich-Heine University, Düsseldorf

Data preprocessing

- calculate from raw data all variables that are used in at least one of the KCH Scores 1.0, 2.0 or 3.0
- if meaning of a variable has changed: recent definition
- categorical predictors: maximum number of categories
- two continuous variables: age and body-mass-index (bmi), scaled to zero mean and unit variance
- merge data from 2007 and 2008

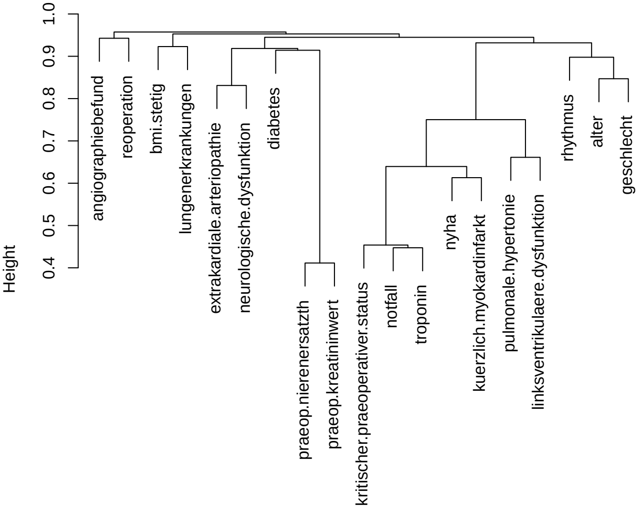
$\Rightarrow$ 21 possible predictor variables, 1163 observations,
1 target variable: postoperative (recovery) state of the patient

Predictor Variables

(preoperative) variable	# values	KCH Score
alter	6	3.0
alter.stetig		
angiographiebefund	2	3.0
bmi	3	3.0
bmi.stetig		
diabetes	2	3.0
extrakardiale.arteriopathie	2	3.0
geschlecht	2	3.0
kritischer.praeoperativer.status	2	3.0
kuerzlich.myokardinfarkt	2	3.0
linksventrikulaere.dysfunktion	3	3.0
lungenerkrankungen	3	3.0
neurologische.dysfunktion	2	3.0
notfall	2	3.0
nyha (severity of cardiac insufficiency)	3	2.0
praeop.kreatininwert	2	1.0
praeop.nierenersatztherapie	2	3.0
pulmonale.hypertonie	2	3.0
reoperation	2	3.0
rhythmus	3	2.0
troponin	2	2.0

Relationship Between Predictor Variables

Cluster dendrogram based on Cramér's V, average linkage



Variable Selection (Guyon et al., 2006; Guyon & Elisseeff, 2003)

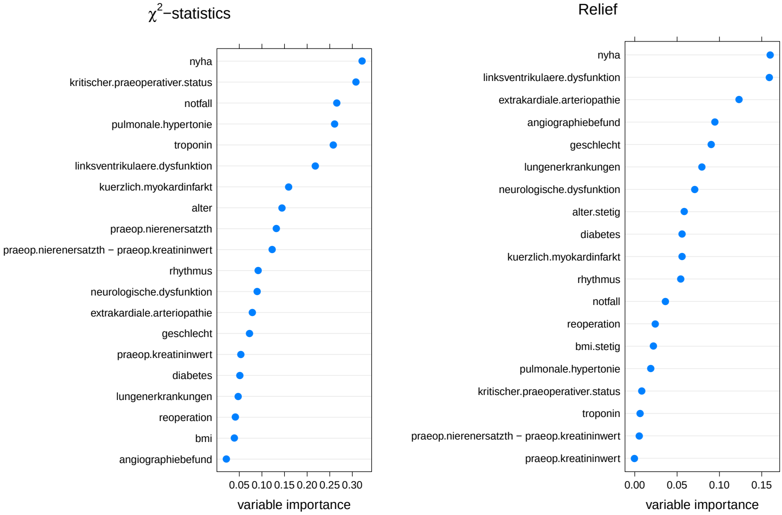
General aims

- understand the data
- improve prediction performance
- data reduction, provide faster and more cost-effective predictors

Approaches

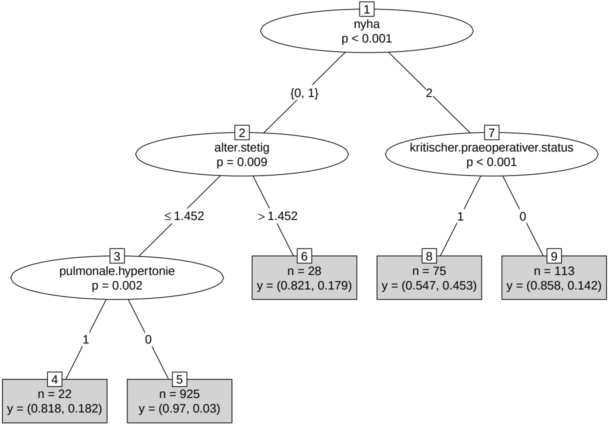
- filter methods: preprocessor, independent of the choice of the predictor
- embedded: variable selection is part of the training process
- wrapper: use classification method as a "black box" to assess goodness of a variable set

Filter Approach: Results



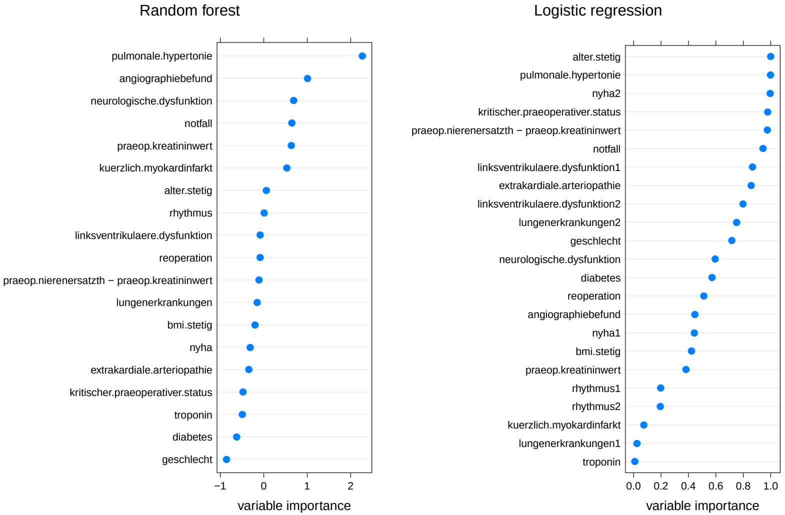
- R version 2.13.1 (R Development Core Team, 2011)
- R package `FSelector` (Romanski, 2009),
  Relief with sample size 600 and 5 nearest neighbors

Embedded Approach: Results I



- R package `party` (Hothorn et al., 2006),
  AUC of 0.73 ± 0.07 (25 subsampling iterations with 4/5 splits)

Embedded Approach: Results II



- random forest with 15000 trees: mean decrease in accuracy
- logistic regression: $1 -$ p-value of Wald test

Wrapper Approach

Three important choices to be made

1. classification method(s)
2. search strategy
3. selection criterion

Wrapper Approach

Classification methods of different complexity

- logistic regression (logreg)
- linear discriminant analysis (lda, `MASS`)
- kernel k nearest neighbors (kknn, `kknn`)
- support vector machine with polynomial kernel (svm.poly, `kernlab`)
- support vector machine with radial kernel (svm.radial, `kernlab`)
- random forests (rF, `randomForest`)
- gradient boosting machine (gbm, `gbm`)

Search strategy

forward search (until AUC cannot be improved by at least 0.001)

Wrapper Approach

Selection criterion

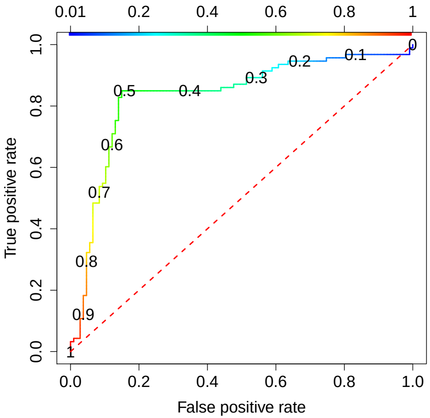
error rate usual criterion in classification, but

- imbalanced problem
- certainly unequal, but unknown misclassification costs: choice of threshold?

$\Rightarrow$ area under ROC curve (AUC), ROC = receiver operating characteristic (Fawcett, 2006)

positive class: 1

- R package `ROCR` (Sing et al., 2005)



Wrapper Approach

Resampling strategy

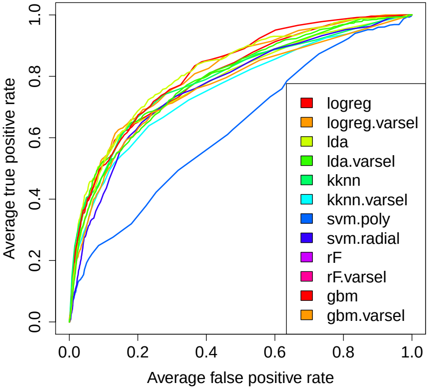
1. parameter tuning: 5-fold stratified cross-validated AUC
2. variable selection: nested resampling strategy
   outer loop: 25-fold subsampling with 4/5 splits
   (Bi et al., 2003)
   - accurate estimates of prediction performance
   - use selection frequency as importance measure
   - stabilize selection results
   - analyze behavior across subsamples

   inner loop: 3-fold stratified cross-validation

- ideally: adapt hyperparameters to variable set and vice versa
- for comparison: train all classification methods without variable selection on the same 25 subsamples
- R package `mlr` (Bischl, 2010)

Wrapper Approach: Results I

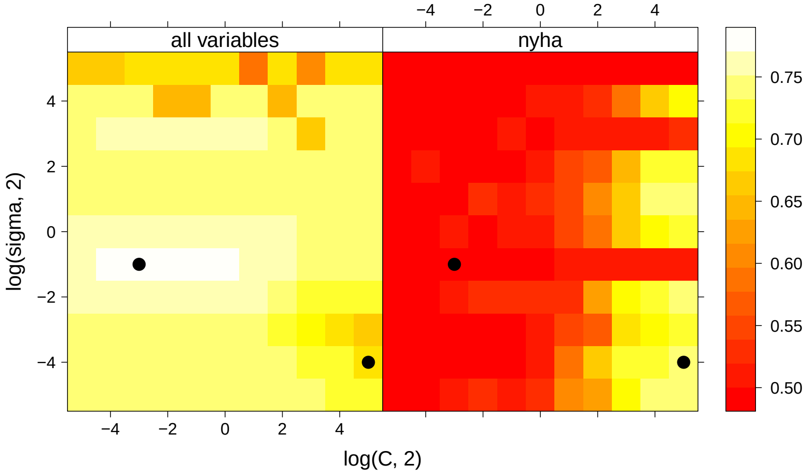
AUC



method	AUC	stand. dev.
logreg	0.81	0.07
logreg.varsel	0.82	0.07
lda	0.80	0.08
lda.varsel	0.81	0.07
kknn	0.81	0.06
kknn.varsel	0.78	0.06
svm.poly	0.72	0.08
svm.poly.varsel	0.55	0.12
svm.radial	0.78	0.06
svm.radial.varsel	0.55	0.13
rF	0.82	0.05
rF.varsel	0.77	0.07
gbm	0.82	0.06
gbm.varsel	0.79	0.07

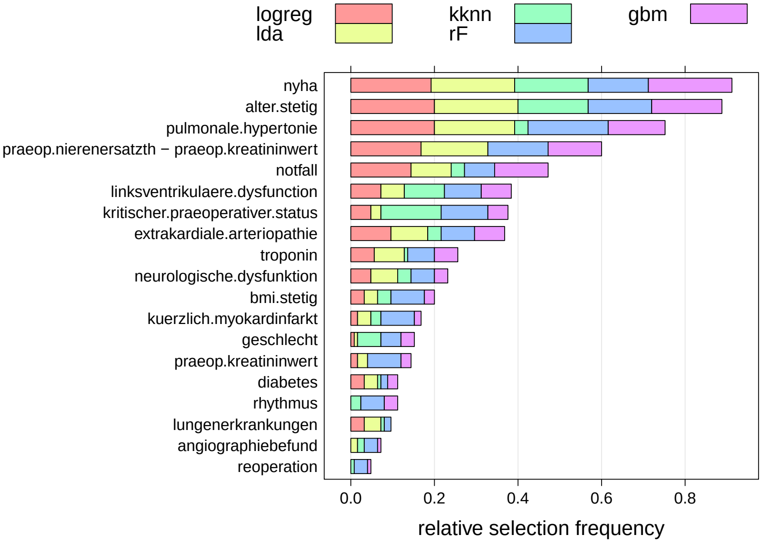
Wrapper Approach: Results II Problem with Support Vector Machines

AUC depending on hyperparameter values for svm.radial using all variables and using one variable (nyha)



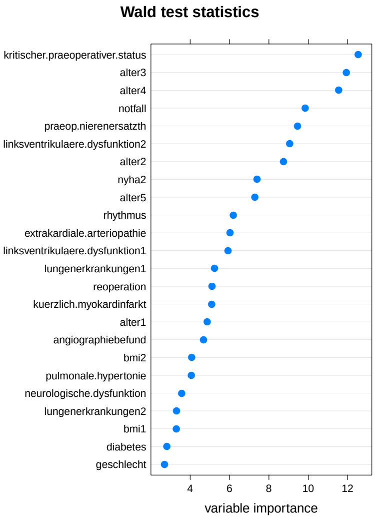
Wrapper Approach: Results III

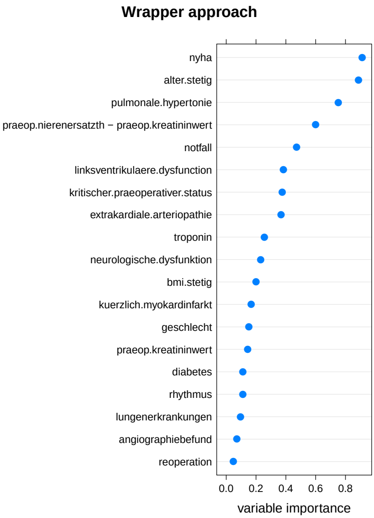
Selected Variables



Wrapper Approach: Results IV

Comparison with KCH Score 3.0



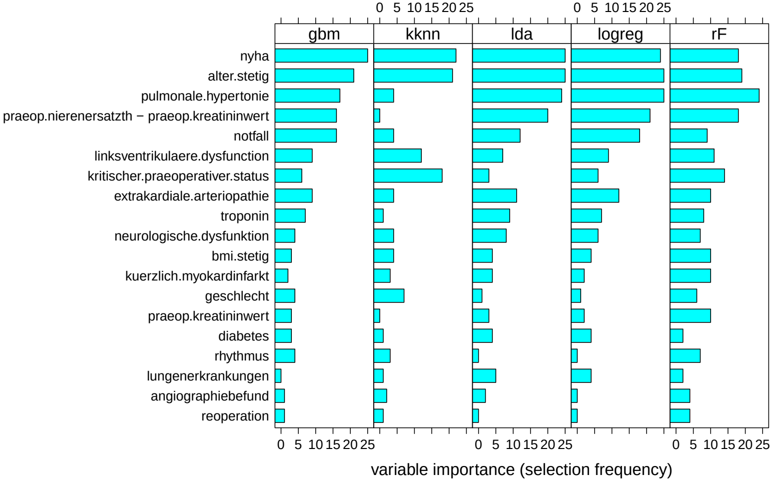


Wrapper Approach: Results V

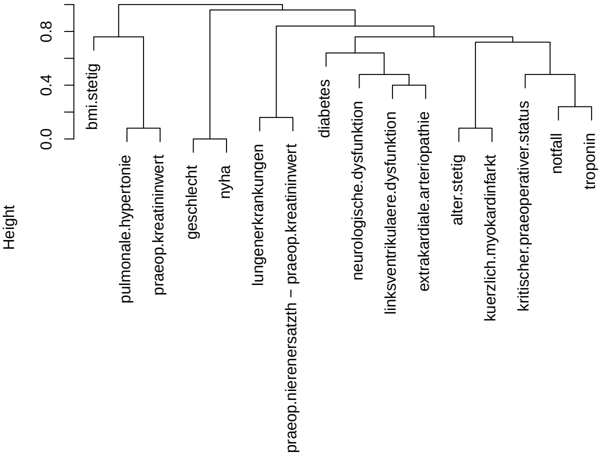
Number of selected variabes

method	# selected variables	stand. dev.
logreg	6.8	1.3
lda	6.7	1.3
kknn	4.5	1.6
svm.poly	3.3	2.7
svm.radial	3.0	3.6
rF	7.7	1.8
gbm	6.0	2.0

Wrapper Approach: Results VI Comparison of Classification Methods



Wrapper Approach: Results VII Behavior Across Subsampling Iterations



Summary & Outlook

- linear classification methods work well on this problem, only small improvements by more complex classification methods like SVMs, kknn, random forest and gbm
- variable selection does not result in much smaller AUCs
- most important variables (for individual clinic): severity of cardiac insufficiency (nyha), age (alter), pulmonary hypertension (pulmonale.hypertonie), preoperative renal replacement therapy (praeop.nierenersatzth)
- unimportant variables (for individual clinic): angiography findings, reoperation

References I

J. Bi, K. P. Bennett, M. Embrechts, C. M. Breneman, and M. Song.
Dimensionality reduction via sparse support vector machines.
Journal of Machine Learning Research, 3:1229–1243, 2003.

B. Bischl.
mlr: Machine learning in R, 2010.

T. Fawcett.
An introduction to ROC analysis.
Pattern Recognition Letters, 27:861–874, 2006.

I. Guyon and A. Elisseeff.
An introduction to variable and feature selection.
Journal of Machine Learning Research, 3:1157–1182, 2003.

I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, editors.
Feature Extraction, Foundations and Applications.
Studies in Fuzziness and Soft Computing. Springer, Berlin Heidelberg, 2006.

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

References II

R Development Core Team.
R: A Language and Environment for Statistical Computing.
R Foundation for Statistical Computing, Vienna, Austria, 2011.
URL `http://www.R-project.org`.
ISBN 3-900051-07-0.

P. Romanski.
FSelector: Selecting attributes, 2009.
URL `http://CRAN.R-project.org/package=FSelector`.
R package version 0.18.

T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer.
ROCR: visualizing classifier performance in R.
Bioinformatics, 21(20):3940–3941, 2005.