Correction for Tuning Bias in Resampling Based Error Rate Estimation

Christoph Bernau & Anne-Laure Boulesteix

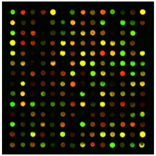
Institute of Medical Informatics, Biometry and Epidemiology (IBE),
Ludwig-Maximilians University, Munich, Germany

Workshop on Validation in Statistics and Machine Learning
Berlin
7.10.2010

Bias induced by tuning & classifier selection

Tuning & Classifier Selection

- a lot of different classifiers available for the data at hand
- no gold standard for classifier selection established in the case of highdimensional data (e.g. microarray data)
- selection of the classifier performed according to a specific performance measure, commonly obtained by resampling or bootstrap
- similar situation: optimization of tuning parameters, e.g. the cost parameter of support vector machines



Raw Data

Subsampling	SVM	PAM	...	5NN	LDA
Iter. 1	e_{11}	e_{12}	...	$e_{1,K-1}$	e_{1K}
Iter. 2	e_{12}	e_{22}	...	$e_{2,K-1}$	e_{2K}
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
Iter. $B-1$	$e_{B-1,1}$	$e_{B-1,2}$	...	$e_{B-1,K-1}$	$e_{B-1,K}$
Iter. B	e_{B1}	e_{B2}	...	$e_{B,K-1}$	e_{BK}
Average	$\bar{e}_1$	$\bar{e}_2$	...	$\bar{e}_{K-1}$	$\bar{e}_K$

- e_{bk}: test error of the kth clasifier in the bth resampling iteration
- $\bar{e}_k$: average test error of classifier k in the whole resampling procedure

Selection/Tuning Bias

- downward bias induced by selection/tuning process ([7])
- authors in biomedical research inclined to report best performance only
- information on performance of all classifiers needed in order to avoid overoptimism
- Is there a way to use the information on the performance of other candidate classifiers in order to estimate the actual performance of the optimal classifier on independent data?

Nested Cross-Validation [7]

$$MCR_{nest} = \frac{1}{B} \sum_{b=1}^{B} e_{bk_b^\sharp}. \tag{1}$$

- two nested loops: inner tuning/selection loop and outer performance estimation loop
- performs an extra cross-validation on each training set of the outer loop in order to find the most appropriate model for the specific training set ($k_b^\sharp$)
- mimicks the procedure that is actually applied to the whole data set
- computationally intensive

Alternative approaches for bias correction

Estimator by Tibshirani & Tibshirani [6]

$$Bias = \frac{1}{B}\sum_{b=1}^{B} Bias_b = \frac{1}{B}\sum_{b=1}^{B}\left(e_{bk^*} - e_{bk_b^*}\right) \qquad (2)$$

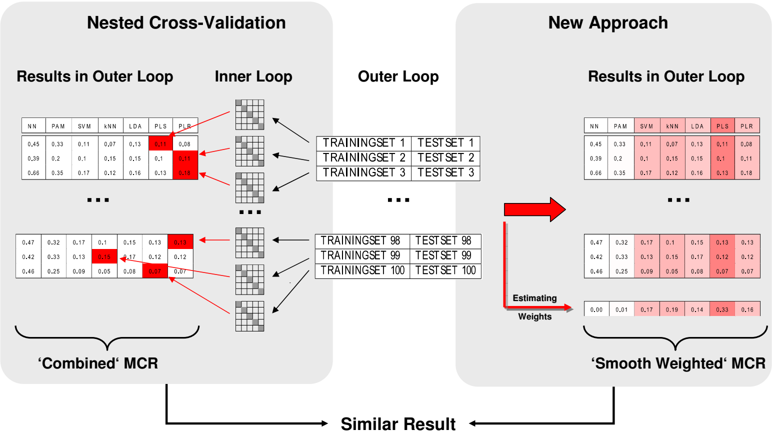
- k_b^* denotes the index corresponding to the classifier performing best on resampling test set b
- k^* denotes the index corresponding to the classifier performing best on the whole resampling procedure
- uses differences between locally and globally optimal classifiers on the different testsets

Our new weighted approach

$$MCR_{WM} = \sum_{k=1}^{K} w_k \bar{e}_k \tag{3}$$

- weighted mean of the resampling error rates of all classifiers
- sensible bounds (worst and optimal MCR)
- computationally less expensive than NCV
- theoretical motivation (see slide 12)
- uses all the information obtained in the classifier selection/tuning process

Comparison of NCV and the weighted MCR approach



Theoretical Motivation

- estimator for $\mathbf{E}_{P_n}\left[\varepsilon(k^*(S) \parallel S)\right]$ rather than for a specific classifier
- S: whole sample, ϵ: true generalization error
- decompose the mean $\mathbf{E}_{P_n}\left[\varepsilon(k^*(S) \parallel S)\right]$ into:

$$\sum_{k=1}^{K} P\left(k^*(S) = k\right) \times E_{P_n}\left[\varepsilon(k \parallel S) | k^*(S) = k\right]$$

$$\approx \sum_{k=1}^{K} P\left(k^*(S) = k\right) \times \mathbf{E}_{P_n}\left(\varepsilon(k \parallel S)\right)$$

- crucial assumption: $\varepsilon(k \parallel S) \perp k^*(S)$ for each classifier

Estimating $P(k^*(S) = k)$ using a parametric approach

- approximate the probabilities by a Monte Carlo simulation with normality assumption:

$$(\bar{e}_1, \ldots, \bar{e}_K) \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{4}$$

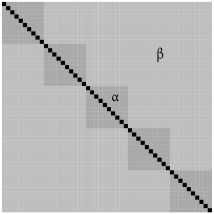
  where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are estimated from the matrix $(e_{bk})_{b=1\ldots B}^{k=1\ldots K}$

- use these probablities as weights in the new estimator
- vector of mean MCRs ($\bar{\mathbf{e}}$) plugged in as $\boldsymbol{\mu}$

The problem of variance estimation

- proof of nonexistence of an unbiased estimator ([1],[3])
- several low biased variance estimators in the literature
- problem of dependencies between testsets
- good estimator for $\rho(\alpha, \beta) = Cor(\bar{e}_{b_1 k}, \bar{e}_{b_2 k})$ required
- sensible estimator ([3]), if $\hat{\rho}(\alpha, \beta)$ is provided:

$$\left(\frac{1}{B} + \frac{\rho}{1-\rho}\right) \times \frac{1}{B-1} \sum_{b=1}^{B} (e_{kb} - \bar{e}_k)^2$$



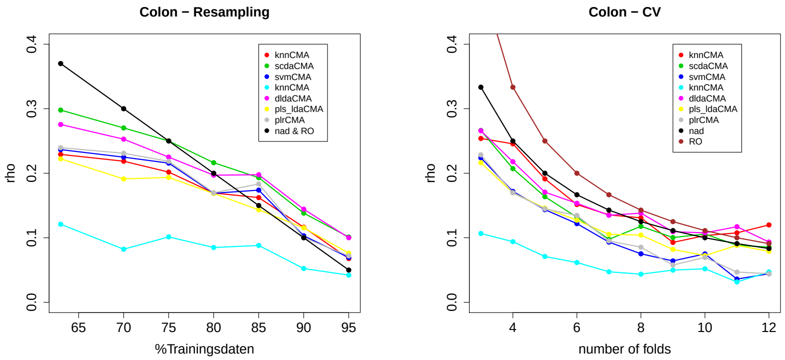
A new approach for estimating ρ

- simulation of several independent non-informative response vectors
- conditional and unconditional error rates known to be 0.5
- simulation of R replicates (e.g. 1000) of the response vector with $y_i \sim B(1, 0.5)$
- for each r: computation of the average test errors for two resampling steps ($\bar{e}_{r1}$ and $\bar{e}_{r2}$)
- estimation of ρ by:

$$\hat{\rho} = \frac{\widehat{\mathbf{Cov}}(\bar{e}_1, \bar{e}_2)}{\sqrt{\widehat{\mathbf{Var}}(\bar{e}_1)}\sqrt{\widehat{\mathbf{Var}}(\bar{e}_2)}} \qquad (5)$$

Estimating ρ

Results:



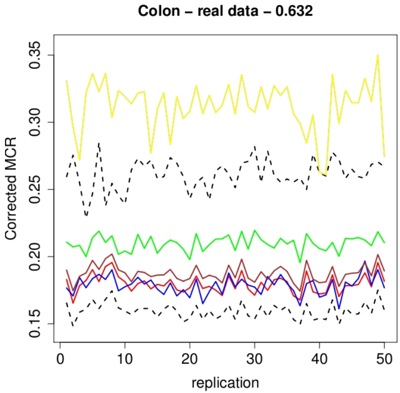
- $\frac{n_t}{n}$ ([3]) as a good approximation which ignores the differences between classifiers

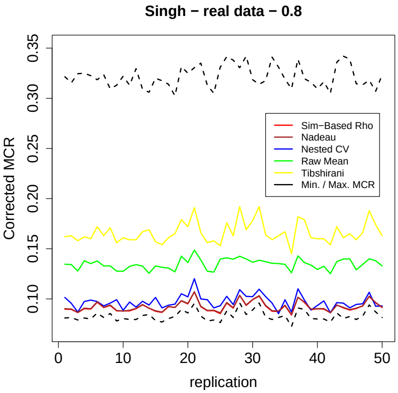
Simulation study

Setup

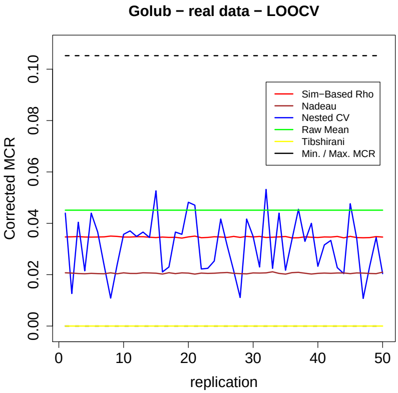
- seven different classifier algorithms (PAM ($\Delta = 0.5$), linear SVM ($cost = 50$), kNN ($k = 1$), kNN ($k = 18$), DLDA, PLSLDA (3 components) and PLR ($\lambda = 0.01$))
- feature selection according to t-Test
- three different real data sets (Golub, Colon, Singh)
- 100 resampling iterations with 0.63% or 0.8% and LOOCV
- 50 replications of the whole procedure in order to asses variability of the different bias correction methods

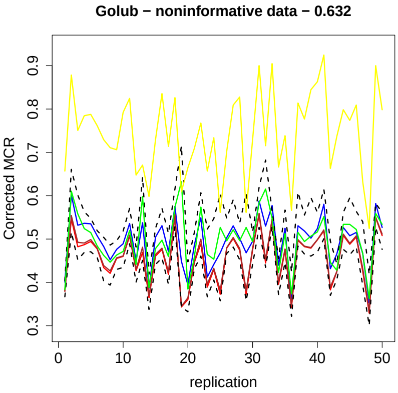
Results


Colon – real data – 0.632
Corrected MCR
0.15
0.20
0.25
0.30
0.35
0
10
20
30
40
50
replication


Singh – real data – 0.8
Corrected MCR
0.10
0.15
0.20
0.25
0.30
0.35
0
10
20
30
40
50
replication
Sim−Based Rho
Nadeau
Nested CV
Raw Mean
Tibshirani
Min. / Max. MCR

Results


Golub − real data − LOOCV
Corrected MCR
replication
Sim−Based Rho
Nadeau
Nested CV
Raw Mean
Tibshirani
Min. / Max. MCR
0.00
0.02
0.04
0.06
0.08
0.10
0
10
20
30
40
50


Golub − noninformative data − 0.632
Corrected MCR
replication
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0
10
20
30
40
50

Outlook

Outlook

- improvement of the estimate of $\mathbf{\Sigma}$ in MC-simulation
  - better estimator for correlation between test sets (ρ)
  - better estimator for correlations between classifiers
- evaluation of weighted mean approach on independent real data sets and further analysis on simulated data
- alternative approach: Generalized Degrees of Freedom [2,4,8]
  $\longrightarrow$ tries to correct apparent error [5]
  $\longrightarrow$ provides information on prediction stability for individual observations

Thank you for your attention

References

1. Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of K-Fold Cross-Validation. J. Mach. Learn. Res., 5:1089-1105,2004
2. Bradley Efron. The estimation of prediction error. Journal of the American Statistical Association, 99(467):619-632, 2004.
3. Claude Nadeau and Yoshua Bengio. Inference for the generalization error. Mach. Learn., 52(3):239-281, 2003.
4. Xiaotong Shen and Jianming Ye. Adaptive model selection. Journal of the American Statistical Association, 97(457):210-221, 2002.
5. Robert Tibshirani and Keith Knight. The covariance inflation criterion for adaptive model selection. J. ROY. STATIST. SOC. B, 55:757-796, 1999.
6. Ryan J Tibshirani and Robert Tibshirani. A bias correction for the minimum error rate in cross-validation. Annals of Applied Statistics 2009, Vol. 3, No. 2, 822-829, 2009.
7. Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics, 7(1):91, 2006.
8. Jianming Ye. On measuring and correcting the effects of data mining and model selection. Journal of the American Statistical Association, 93(441):120-131, March 1998.