Over-optimism in biostatistics and bioinformatics

Anne-Laure Boulesteix
joint with M. Jelizarow, V. Guillemot, A. Tenenhaus, K. Strimmer

Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie
Ludwig-Maximilians-Universität München

Paris, 23. August 2010

Bias in reporting error rates: An empirical study

- **Setup:** supervised classification based on high-dimensional data like microarray data
- Many available methods (SVM, lasso, etc) but no consensus
- Cross-validation is often used to estimate error rates.
- Choosing the classification method *a posteriori* based on the estimated error rates yields a strongly optimistic estimate: **the minimal error rate was as low as 31%** (!!) with **permuted class labels** for a colon cancer data set in our empirical study.

A.-L. Boulesteix, C. Strobl, 2009. Optimal classifier selection and negative bias in error rate estimation: An empirical study on high-dimensional prediction. *BMC Medical Research Methodology* 9:85.

Bias in methodological research

- When developing statistical methods, researchers often think of several possible variants (called “methods’ characteristics” here).
- If they choose the methods’ characteristics a posteriori (i.e. because they obtain nice results with these characteristics), the results of the new method are also optimistically biased!

Here we present an empirical study to illustrate this bias and the need for validation with independent data.

A “promising” method

Discriminant function in linear discriminant analysis:

$$d_r(\mathbf{x}) = \mathbf{x}^\top \Sigma^{-1} \boldsymbol{\mu}_r - \frac{1}{2} \boldsymbol{\mu}_r^\top \Sigma^{-1} \boldsymbol{\mu}_r + \log(\pi_r),$$

Problem: The sample estimator $\hat{\Sigma}$ of the covariance matrix Σ is not invertible when $n \ll p$!

Solution: Use a regularized estimator of Σ instead of the $\hat{\Sigma}$, for instance the shrinkage estimator by Schäfer and Strimmer (2005):

$$\hat{\Sigma}^* \; = \; \lambda \hat{\Sigma} \; + \; (1 - \lambda) T,$$

where T is an adequately chosen target and λ a shrinkage parameter.

A “promising” method

Idea: Define T using priori knowledge on the gene function groups (GFG):

Target D

$$t_{ij} = \begin{cases} s_{ii} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

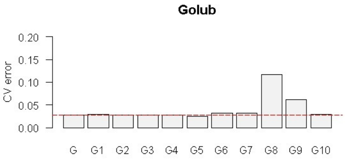
Target G

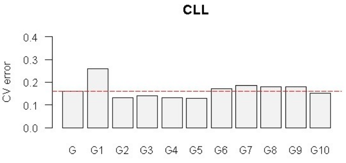
$$t_{ij} = \begin{cases} s_{ii} & \text{if } i = j \\ \bar{r}\sqrt{s_{ii}s_{jj}} & \text{if } i \neq j, i \sim j \\ 0 & \text{otherwise} \end{cases}$$

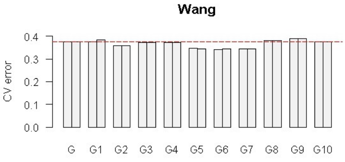
Problem: How should we deal with genes that are in no GFG, genes that are in several GFG, negative correlations within GCG, non-significant correlations?

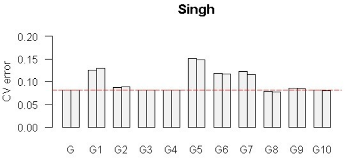
→ 10 candidate variants

Selecting the methods' characteristics optimally









The error rate can be decreased by optimizing the "methods' characteristics" (i.e. by choosing the optimal variant for a particular data set).

Selecting the methods' characteristics optimally

	M_{opt}	s_{opt}	Golub	CLL	Wang	Singh
Golub	rlda.TG$^{(5)}$	s_{opt} = (200, Limma)	**0.025**	0.180	0.345	0.152
CLL	rlda.TG$^{(5)}$	s_{opt} = (200, Wilcoxon test)	0.079	**0.129**	0.363	0.141
Wang	rlda.TG$^{(6)}$	s_{opt} = (200, t-test)	0.029	0.221	**0.342**	0.115
Singh	rlda.TG$^{(8)}$	s_{opt} = (100, Limma)	0.033	0.274	0.384	**0.078**

- Seemingly good results are obtained by "fishing for significance" (i.e. optimizing the variable selection setting and the methods' characteristics).
- These seemingly good results cannot be validated based on other data sets.

Sources of the problems

Results presented in statistical bioinformatics papers are sometimes the product of intense optimization: optimization of the settings and optimization of the methods characteristics.

- **Problem 1:** Error rate estimators have high variance in $n \ll p$ settings, hence the opportunity for optimization.
- **Problem 2:** In methodological research we are interested in the *unconditional* error rate of the method. Since variability between data sets is high, several data sets are needed.

Some (partial) solutions

- Internal cross-validation?
  - → not for the methods' characteristics
  - → would not address the (most important) variability between data sets
- Check the superiority of the new method using other "validation" data sets. ... But the unbiased selection of appropriate data sets is a non-trivial task!
- Pay more attention to the substantive context.
- Publish negative results?

Jelizarow et al, 2010. Over-optimism in bioinformatics: an illustration. Bioinformatics 26:1990–1998.

Boulesteix, 2010. Over-optimism in bioinformatics research (letter to the editor). Bioinformatics 26:437–439.

Thanks for your attention!

Thanks to V. Guillemot, M. Jelizarow, K. Strimmer (University Leipzig), C. Strobl, A. Tenenhaus (Ecole Supélec).

The papers:

- M. Jelizarow, V. Guillemot, A. Tenenhaus, K. Strimmer, A.-L. Boulesteix, 2010. Over-optimism in bioinformatics: an illustration. *Bioinformatics* 26:1990–1998.
- A.-L. Boulesteix, 2010. Over-optimism in bioinformatics research. *Bioinformatics* 26:437–439.
- A.-L. Boulesteix and C. Strobl, 2009. Optimal classifier selection and negative bias in error rate estimation: An empirical study on high-dimensional prediction. *BMC Medical Research Methodology* 9:85.