Over-optimism in bioinformatics: an illustration

Article by:
M. Jelizarow, V. Guillemot, A. Tenenhaus, K. Strimmer and A. Boulesteix

Presentation by:
Jonas Pedersen, Mikkel Mødekjær and Riz Noronha

Date: 09-03-2023

UNIVERSITY OF COPENHAGEN

Motivation

Statistical bioinformatics research papers often include optimization methods, that might lead to an over-confidence in resulting models.

The article is an empirical study of several pitfalls in model optimization using a high-dimensional example.

The motivation is therefore to call out practitioners of bioinformatics on their "malpractice".

Four independent microarray datasets are used: Golub's leukemia dataset, the CLL dataset, the Singh et al.'s prostate dataset and the Wang et al.'s breast cancer dataset. Each contains a binary outcome that needs to be predicted based on the gene expression data.

Common errors found in Bioinformatics

1. Optimization of datasets
   - Choosing the dataset that best fits your model
2. Optimizations of settings
   - Choosing settings for your model leading to maximal accuracy
   - Highly related to overfitting
3. Optimizations of competing method
   - Comparing to suboptimal methods
4. Optimization of the method's characteristics
   - Specializing the algorithm to the concerned dataset

Linear Discriminate Analysis (LDA)

LDA is an algorithm to classify multidimensional data, by defining regions of parameter space as one class.

Assumes all classes are distributed according to multivariate Gaussians, and results in a division of parameter space using hyperplanes.

RLDA is a regularized version, which redefines the covariance matrix to:

$$\Sigma_{reg} = (1 - \lambda)\Sigma + \lambda I$$

Where λ is the *regularization parameter*.

Method

Prior biological information is incorporated by replacing I with target matrix T:

$$\Sigma_{SHIP} = (1-\lambda)\Sigma + \lambda T$$

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

11 classifiers are used: rlda.TG, and 10 variants of it.

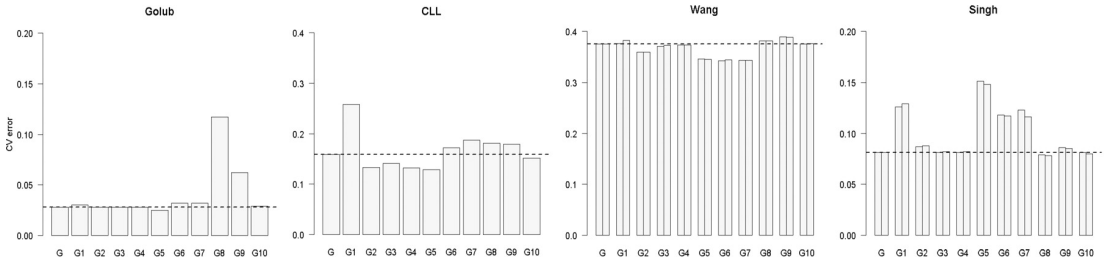
The classifiers are trained on selections of each dataset: Three different selectors are used (t-test, Wilcoxon rank test, Limma procedure), and each is used to select 4 different amounts of genes (100, 200, 500, 1000): 3×12 total combinations.

Prediction accuracy is estimated with a 10 times 5-fold Cross Validation evaluation scheme.

Results

Settings are not the same for every dataset! There are two columns on the right datasets because there are two optimal setting configurations.

A method can also be compared to an inferior method.



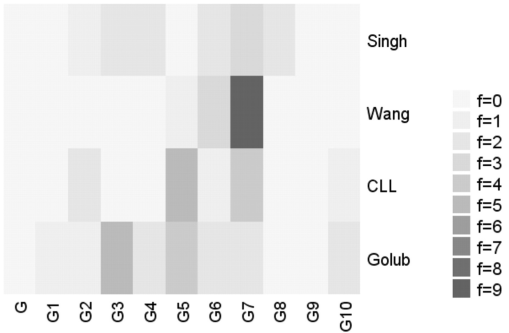
Results

Finding the 'optimal' method in each of the 12 settings, for every dataset.

The optimal algorithm changes across datasets.

No clear winner, but without investigating different datasets and settings one would see a clear winner.

Independent data is key!



Discussion

Other points briefly mentioned:

- The pitfalls often occur in combination with one another.
- Other pitfalls are also possible.
  - Problems in the proposed model would likely be a priority to solve, whereas problems with competing models would not.
  - Pitfalls involving other evaluation criteria than error rates.
- Many pitfalls could be avoided with a higher sample size.
  - Simulated data could be a solution, although difficult in practice.

Conclusion/Summary

Model development is a very involved subject.

Pitfalls are abundant.

Tests using independent data should be more common in bioinformatics.

Jelizarow et. al. like throwing shade :)