Presentation write-up

Riz Noronha (rpw391), Mikkel Mødekjær (zkm716) & Jonas Pedersen (wlq622)

March 2023

1 Introduction

Several results in statistical bioinformatics are found to be over-optimistic: Models are often reported to be superior to the competition when that isn't necessarily the case. The article provides an example of how a new algorithm, despite having poor results in terms of the error rate, can be shown to be artificially superior to other algorithms in specific circumstances. The example provided is a classifier which uses Linear Discriminant Analysis (LDA) while taking prior knowledge of gene functional groups into consideration. This article therefore acts as a "wake-up call" to practitioners of bioinformatics to improve their modelling.

2 Review

The article lists 4 potential contributions to this overconfidence, which might artificially make your model seem better than it is:

1. **Optimization of the dataset**
   Concerns the behavior of searching for a dataset implying that your model is optimal.
2. **Optimization of the settings**
   Concerns the behavior of finding the optimal settings for your model considering a single dataset. Highly related to overfitting.
3. **Optimization of the competing methods**
   Concerns the behavior of only comparing your model with suboptimal competing models, instead of state-of-the-art models.
4. **Optimization of the method's characteristics**
   Concerns the behavior of optimizing their algorithm to the datasets they consider leading to non-general methodology.

The authors perform an empirical study of different recreations of these four pitfalls. They do this, by creating different kinds of algorithms, and optimize them to four different data sets in the ways shown above.

As mentioned, the "trial model" is LDA, which is a means of classifying higher dimensional data by generating a single classifier from a linear combination of the data (e.g: Fisher's Discriminant). LDA assumes that a random variable x of predictors follows a multivariate Gaussian distribution

$$x \,|\, (Y = r) \sim \mathcal{N}(\mu_r, \Sigma_r)$$

within each class, r. In order to calculate it, we typically use the inverse of the covariance matrix Σ (which we estimate with the inverse of the pooled estimator $\tilde{S}$ of the within-covariance matrix). In a high dimensional setting, $\tilde{S}$ is not necessarily invertible, and the problem is resolved with a *shrinkage* in regularized LDA (RLDA).

In the paper's example, they propose a covariance estimator $\hat{\Sigma}_{SHIP}$ that provides both shrinkage as well as incorporates prior knowledge, given by

$$\hat{\Sigma}_{SHIP} = \lambda \mathbf{T} + (1 - \lambda)\mathbf{S}$$

where $\lambda \in [0, 1]$, the optimal shrinkage intensity, can be computed analytically and $\mathbf{T}$ is the target matrix that incorporates prior biological information, defined as

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

where s_{ij} are entries from the unbiased covariance matrix, $\bar{r}$ is the average of sample correlations. The notation $i \sim j$ implies that genes corresponding to entries i and j are from the same gene functional group. This estimator is referred to as rlda.TG.

Four independent microarray datasets are used for the results: Golub's leukemia dataset[1], the CLL dataset[5], the Singh et al.'s prostate dataset[3] and the Wang et al.'s breast cancer dataset[4]. Each contains a binary outcome that needs to be predicted based on the gene expression data.

Eleven classifiers are used to classify the data: The classifier described above (rlda.TG) and ten variations of it (rlda.TG$^{(i)}$, $i = 1, 2..10$) which differ from each other either by how they treat "problematic genes" or how they redefine the target covariance matrix. For computational brevity, they train the classifiers on selections of datasets: Three different selectors are used (the t-test, the Limma procedure, and the Wilcoxin ranked test), and each used to select 4 different amounts of genes (100, 200, 500, 1000) which gives us 12 different combinations of selection procedures and the amount of selected genes.

The different classifiers (or methods) are run with different settings (i.e, the selection amounts and procedures) on different datasets, and the results are compared. Results (in terms of the covariance error rates) with settings that minimize the error can be seen in Figure 1 below (Note that the optimal settings for the Wang and Singh data are not unique):

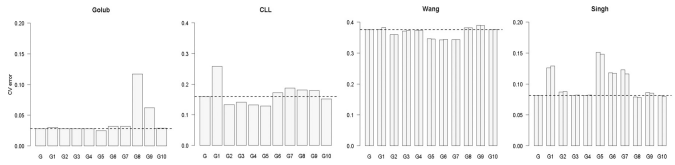

Figure 1: The CV error rates of the different classifiers, obtained for all datasets with 'optimal' settings: (200, Limma) for the Golub data, (200, Wilcoxon test) for the CLL data, (200, t-test) (left bar) and (200, Limma) (right bar) for the Wang data and (100, t-test) (left bar) and (100, Limma) (right bar) for the Singh data.[2]

For a particular method, a researcher who 'fishes for significance' will choose the settings which will minimize the error rates for the dataset they consider, which leads to an optimistic bias through optimization of the settings.

One could also consider trying the various classifiers (or methods) in order to find one which minimizes the error for a particular dataset: for instance, the CLL and Wang data have variants that lower the error more than the basic rlda.TG. Note that there is no universally good method: rlda.TG$^{(5)}$ is optimal for the Golub data while rlda.TG$^{(8)}$ is optimal for the Singh data. Once again, someone looking at only a particular dataset would be overoptimistic about the method being optimal, which is a case of optimization of the method's characteristics.

The performance of a new algorithm can often be seen by comparing it to an existing algorithm: if compared to a suboptimal algorithm, an arbitrary new algorithm could be seen as an improvement, which is a case of optimization of competing methods.

Some researchers may choose a dataset which has results that are favourable to the model (which might be linked to an optimization of the settings, as a dataset might look good because of the chosen optimal settings), which leads to an optimization of the dataset.

Figure 2 shows how often a particular method minimizes the error. For example, in the Wang dataset, the lowest error rate is reached by rlda.TG$^{(7)}$ in 9 of the 12 considered settings and by rlda.TG$^{(6)}$ in only three settings. It is clear that the 'optimal' variant depends on both the dataset and the settings, and thus there is no clear winner. If a researcher does not investigate multiple datasets and settings, one could have the impression that there is a clear winner.

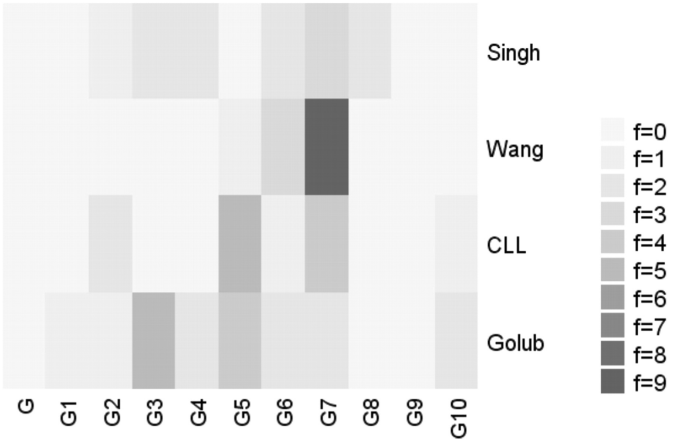

Figure 2: Frequency of selection of the 11 investigated variants of rlda.TG over the 12 settings: three selection methods (t-test, Limma, Wilcoxon test) and four numbers of genes (100, 200, 500, 1000). 'Selection' means that the variant yields the smallest error rate over the 11 variants. Note that the best variant may not be unique.[2]

3 Conclusion

The article concludes that the 4 pitfalls mentioned can highly affect the quality of the model when applied onto other datasets, resulting in over-confidence in their model. Practitioners of bioinformatics should therefore strive to have consider more things than just pure accuracy, and instead reconsider the validity of the model in question. An important thing to also include, is a comparison of the model to data not used in the development of the said model. Over-confidence should be more of a concern in bioinformatics.

References

[1] Todd Golub. *golubEsets: exprSets for golub leukemia data*. DOI: 10.18129/B9.bioc.golubEsets.

[2] Monika Jelizarow et al. "Over-optimism in bioinformatics: an illustration". In: *Bioinformatics* 26.16 (June 2010), pp. 1990–1998. DOI: 10.1093/bioinformatics/btq323.

[3] Dinesh Singh et al. "Gene expression correlates of clinical prostate cancer behavior". In: *Cancer Cell* 1.2 (2002), pp. 203–209. DOI: https://doi.org/10.1016/S1535-6108(02)00030-2.

[4] Yixin Wang et al. "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer". In: *The Lancet* 365.9460 (2005), pp. 671–679. DOI: https://doi.org/10.1016/S0140-6736(05)17947-1.

[5] Elizabeth Whalen. *CLL: A Package for CLL Gene Expression Data*. DOI: 10.18129/B9.bioc.CLL.