THRESHOLDING METHODS FOR FEATURE SELECTION IN GENOMICS: HIGHER CRITICISM VERSUS FALSE NON-DISCOVERY RATES

Bernd Klaus and Korbinian Strimmer

Institute for Medical Informatics, Statistics and Epidemiology (IMISE),
University of Leipzig, Härtelstr. 16–18, D-04107 Leipzig, Germany
Contact: bernd.klaus@uni-leipzig.de

ABSTRACT

In high-dimensional genomic analysis it is often necessary to conduct feature selection, in order to improve prediction accuracy and to obtain interpretable classifiers. Traditionally, feature selection relies on computer-intensive procedures such as cross-validation. However, recently two approaches have been advocated that both are computationally more efficient: False Non-Discovery Rates (FNDR) and Higher Criticism (HC). Here, we describe the rationale behind the two approaches, conduct an empirical comparison based on synthetic and real data, and discuss the respective merits of HC-based and FNDR-based feature selection.

1. INTRODUCTION

Feature selection is an integrative part of many genomic analyses, e.g., in classification of cancer tissues using microarray data. Without variable selection the prediction functions are unstable and can neither be properly estimated nor interpreted.

In the majority of currently employed algorithms feature selection is based on minimizing prediction error, which is typically estimated by a variant of cross-validation. As resampling can be quite demanding in terms of required computing time, alternative computationally much more efficient criteria have been proposed. In this note we focus on feature selection in linear classification using two specific selection methods,"Higher Criticism" (HC) and "False Non-Discovery Rates" (FNDR).

Originally, HC refers to an approach to multiple testing [1]. It has been rediscovered in the context of sparse signal detection [2] and subsequently employed for feature selection in high-dimensional classification [3; 4]. "False Discovery Rates" (FDR) provide an alternative criterion to multiple testing [5; 6; 7] and has become standard in large-scale statistical analysis [8]. Feature selection in classification based on "False Non-Discovery Rates" (FNDR) has been suggested recently in [9].

Both FDR/FNDR and HC-based feature selection assume a null model for the observed test statistics is known, e.g., a normal distribution. Subsequently, a threshold separating null from non-null features is determined.

Our analysis extends the one of [9] in looking more closely at the rationale behind HC and offering some results in the context of the so-called "Rare and Weak" feature model [3; 4].

2. METHODS

2.1. Linear classification rules

Many classification algorithms can be put in the framework of linear decision rules $d_k(\boldsymbol{x})$, which assign data $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ to the class k maximizing $d_k(\boldsymbol{x})$. For example, if we represent each group k by a multivariate normal distribution $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ with mean $\boldsymbol{\mu}_k$ and common covariance matrix $\boldsymbol{\Sigma}$, and assume a prior π_k we arrive at standard multi-class linear discriminant analysis (LDA) with

$$d_k^{\text{LDA}}(\boldsymbol{x}) = \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - \frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log(\pi_k)\,. \quad (1)$$

Note that $d_k^{\text{LDA}}(\boldsymbol{x})$ is linear in $\boldsymbol{x}$. LDA is optimal under the normal assumptions and forms the basis of most classification algorithms currently employed in high-dimensional data analysis, including Nearest Shrunken Centroids (NSC) or Shrinkage Discriminant Analysis (SDA), cf. [9; 10].

For a binary prediction ($K = 2$) it is useful to consider as decision criterion the difference between the discriminant scores of the two classes $L(\boldsymbol{x}) = d_1^{\text{LDA}}(\boldsymbol{x}) - d_2^{\text{LDA}}(\boldsymbol{x})$, which can be written as [10]

$$L(\boldsymbol{x}) = \boldsymbol{\omega}^T \boldsymbol{\delta}(\boldsymbol{x}) + \log\left(\frac{\pi_1}{\pi_2}\right) \tag{2}$$

with feature weights

$$\boldsymbol{\omega} = \boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\,, \tag{3}$$

and decorrelated predictors

$$\boldsymbol{\delta}(\boldsymbol{x}) = \boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}\left(\boldsymbol{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}\right), \tag{4}$$

The matrix $\boldsymbol{P}$ contains the correlations among the predictors, and $\boldsymbol{V}$ is a diagonal matrix with the variances (thus, $\Sigma = \boldsymbol{V}^{1/2}\boldsymbol{P}\boldsymbol{V}^{1/2}$). If $L(\boldsymbol{x}) \geq 0$ then the test sample $\boldsymbol{x}$ is assigned to class 1, and otherwise to class 2.

For $\boldsymbol{X}_k$ a random vector drawn from the multivariate normal $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ associated with class k this implies

$$\boldsymbol{\delta}(\boldsymbol{X}_k) \sim N\left(\pm\frac{\boldsymbol{\omega}}{2}, \boldsymbol{I}\right) \tag{5}$$

with the plus sign corresponding to $k = 1$, and hence

$$L(\boldsymbol{X}_k) \sim N\left(\pm\frac{\boldsymbol{\omega}^T\boldsymbol{\omega}}{2} + \log\left(\frac{\pi_1}{\pi_2}\right), \boldsymbol{\omega}^T\boldsymbol{\omega}\right). \tag{6}$$

Knowing the distribution of $L(\boldsymbol{X}_k)$ we can compute the probability $\alpha_1 = \Pr(L(\boldsymbol{X}_1) < 0)$ which is the nominal misclassification error of incorrectly assigning an element from group 1 to group 2. (likewise, $\alpha_2 = \Pr(L(\boldsymbol{X}_2) > 0)$). Both kinds of errors can be obtained as

$$\alpha_k = \Phi\left(-\frac{\boldsymbol{\omega}^T\boldsymbol{\omega}/2 \pm \log\left(\frac{\pi_2}{\pi_1}\right)}{\sqrt{\boldsymbol{\omega}^T\boldsymbol{\omega}}}\right). \tag{7}$$

2.2. Variable selection

By construction of rule Eq. **2**, variables x_i with large corresponding feature weights ω_i are most influential in class prediction, and in addition contribute most to decrease the prediction error in Eq. **7**. Therefore, it is sensible to conduct feature selection by thresholding features according to the magnitude of ω_i^2. Note that $c_0\,\omega_i$ (where $c_0 = \sqrt{n_1 n_2/(n_1+n_2)}$ is a sample size dependent constant) are the correlation-adjusted t-scores (also called "cat" scores) introduced in [10].

A key problem in feature selection is that we do not know the true value of each weight ω_i but instead have to estimate it from data. In turn, this implies that the prediction error Eq. **7** is also an estimate (and one that is typically heavily biased). The nominal error is trivially minimized by including all predictors. However, if the coefficients ω_i are estimated with error, then it is no longer beneficial to include all predictors. In contrast, if there is a large number of candidate predictors but the actual number of true predictors is small, it is necessary to remove all the random null predictors to improve prediction accuracy.

Several computationally efficient strategies have been proposed recently:

- The SDA algorithm of [9] uses control of False Non-Discovery Rate (FNDR) to identify the null genes to be eliminated from the classifier.
- Ebay approach of [11] uses an empirical Bayes model for the mean difference to estimate the prediction error and uses t-scores for gene ranking (hence it assumes $\boldsymbol{P} = \boldsymbol{I}$). Features are added until the prediction error falls below a given threshold.
- The Higher-Criticism (HC) approach [3] outlined in more detailed below uses a proxy of the estimated prediction error. The number of included features is determined by maximizing the HC criterion, which in turn implies that the estimated prediction error is minimized.

Note that the FNDR and the Ebay approach rely on a prespecified error threshold in order to determine the cut-off, whereas in the HC approach no such threshold is needed.

2.3. The rationale behind HC

As with the FNDR and Ebay methods the HC approach starts by arranging features in decreasing order of magnitude $\omega_{(1)}, \ldots, \omega_p$ so that $\omega^2_{(i)} > \omega^2_{(i+1)}$, and with the aim to include the top t features. In the same order we arrange the corresponding p-values $\pi_{(1)}, \ldots, \pi_{(p)}$.

If the number of non-null features is small and $\pi_1 = \pi_2$ then can be shown [3; 4] that minimizing the expected empirical prediction error $E(\hat{\alpha}_k|t)$ for t included predictors is equivalent to maximizing the expression

$$\frac{TP(t)}{\sqrt{TP(t) + FP(t)}},$$

where $TP(t)$ and $FP(t)$ denote the expected number of true and false positives at threshold t. An empirical estimate of this quantity is given by the HC criterion

$$HC(t) = \arg\max_{\pi_i} \frac{t/p - \pi_{(t)}}{\sqrt{t/p \cdot (1 - t/p)}}. \qquad (8)$$

The optimal number of features to be included according to Higher Criticism is then given by $t^* := \arg\max_t HC(t)$.

3. RESULTS

3.1. Simulation study

The rare-weak (RW) model [3] is a simple simulation setup that allows to study the performance of HC and related approaches in a situation where the number of features is very large but only a few are relevant for prediction.

Specifically, we repeatedly simulated the observed values of $p = 10000$ scores $\hat{\omega}_i \sim N(\theta_i, 1)$ where $\theta_i = \tau$ for 25 features and $\theta_i = 0$ for the remaining 9975 noise features. Varying the value of τ between 0 and 5 we counted the number of false and true discoveries using HC and FNDR thresholding (the latter using local FNDR with a threshold of 0.2).

The results are shown in Figure 1. If τ is small and thus the rare features are very weak HC selects many more features than FNDR, with most of those features being false positives. As τ increases both methods yield very similar results. In terms of selection stability the standard deviation of the number of features chosen by FNDR (S.E.=10) was an order of magnitude lower than that of HC (S.E.=60).

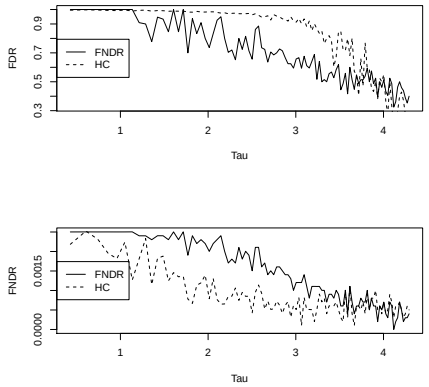

Figure 1. False discovery and false non-discovery rates under the RW model for HC and FNDR thresholding.

3.2. Gene expression data

In application to genomic data and LDA/DDA classification FNDR-based and HC-based thresholding feature selection show similar performance [9]. In the original HC paper [3] a simplified three weights classifier using only the signs of the estimated feature weights is advocated as an alternative to LDA/DDA. In Tab. **1** and Tab. **2** we compared this with the LDA and DDA and found that this approach is competitive on the colon and prostate cancer data sets considered in [12] but is clearly outperformed by DDA analysis in the lymphoma data from [13]. In order to infer the LDA and DDA parameters we used the shrinkage approach described in [9]. The error rates shown in Tab. **1** were obtained by 10-fold cross-validation with 20 repetitions.

4. DISCUSSION AND CONCLUSION

HC thresholding is an effective feature selection approach if individual features are rare and weak, as it is often the case in genomics. However, the selection variability of HC is an important disadvantage compared to using other approaches such as FNDR. Furthermore, we showed the HC

Method	Colon [12]	Prostate [12]
FNDR-DDA	0.128 (0.009)	0.068 (0.005)
HC-DDA	0.1304193 (0.008)	0.073 (0.05)
HC-3W	0.157 (0.147)	0.103 (0.008)
Method	Hummel-Lymphoma [13]	
lfdr-DDA	0.004 (0.001)	
HC-DDA	0.002 (0.0007)	
HC-3W	0.256 (0.002)	

Table 1. Optimal estimated prediction errors for various data sets and the DDA/3W classifiers in combination with HC and FNDR thresholding. The respective associated error is given in brackets.

Method	Colon [12]	Prostate [12]
FNDR	176	153
HC	200	129
Method	Hummel-Lymphoma [13]	
FNDR	589	
HC	588	

Table 2. Number of features selected by the FNDR and HC feature selection methodologies.

method performs best in a full classification model (LDA/DDA) rather than with the overly simplistic HC classifier employed in the original HC paper.

Acknowledgments

We thank Verena Zuber for helpful discussion. Part of this work was supported by BMBF grant no. 0315452A (HaematoSys project).

References

[1] J. W. Tukey, "Higher criticism for individual significances in several tables or part of tables," Working Paper 89–9, Princeton University, 1977.

[2] D. Donoho and J. Jin, "Higher criticism for detecting sparse heterogeneous mixtures," *Annals of Statistics*, vol. 32, pp. 962–994, 2004.

[3] D. Donoho and J. Jin, "Higher criticism thresholding: optimal feature selection when useful features are rare and weak," *Proc. Natl. Acad. Sci. USA*, vol. 105, pp. 14790–15795, 2008.

[4] D. Donoho and J. Jin, "Feature selection by higher criticism thresholding achieves the optimal phase diagram," *Phil. Trans. R. Soc. A*, vol. 367, pp. 4449–4470, 2009.

[5] T. Schweder and E. Spjøtvoll, "Plots of p-values to evaluate many tests simultaneously," *Biometrika*, vol. 69, pp. 493–502, 1982.

[6] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. R. Statist. Soc. B*, vol. 57, pp. 289–300, 1995.

[7] K. Strimmer, "A unified approach to false discovery rate estimation," *BMC Bioinformatics*, vol. 9, pp. 303, 2008.

[8] B. Efron, "Microarrays, empirical Bayes, and the two-groups model," *Statist. Sci.*, vol. 23, pp. 1–22, 2008.

[9] M. Ahdesmäki and K. Strimmer, "Feature selection in omics prediction problems using cat scores and false non-discovery rate control," *Ann. Appl. Statist.*, vol. 4, 503-519, 2010.

[10] V. Zuber and K. Strimmer, "Gene ranking and biomarker discovery under correlation," *Bioinformatics*, vol. 25, pp. 2700–2707, 2009.

[11] B. Efron, "Empirical Bayes estimates for large-scale prediction problems," *J. Amer. Statist. Assoc.*, vol. 104, pp. 1015–1028, 2009.

[12] M. Dettling, "BagBoosting for tumor classification with gene expression data," *Bioinformatics*, vol. 20, pp. 3583–3593, 2004.

[13] M. Hummel et al. and Molecular Mechanisms in Malignant Lymphomas Network Project of the Deutsche Krebshilfe, "A biologic definition of burkitt's lymphoma from transcriptional and genomic profiling," *N. Engl. J. Med.*, vol. 354, no. 23, pp. 2419–2430, Jun 2006.