Finding Predictive Gene Groups from Microarray Data and Combining Them with Clinical Predictor Variables

Marcel Dettling
Johns Hopkins University
550 Bldg., Suite 1131
Baltimore MD 21205, USA
dettling@jhu.edu

Introduction

Large-scale monitoring of gene expression by microarrays is considered to be one of the most promising techniques to improve medical diagnostics and functional genomics. A challenging task with these data is to reveal groups of genes whose collective expression is optimally predictive for a given response variable y. Such gene groups can be understood as molecular signatures and may be benficial for class prediction, as well as to enhance understanding about gene regulation. However, due to the huge combinatorial complexity and the usually small sample size, finding the groups is highly nontrivial.

This paper revisits *Pelora* (Dettling and Bühlmann, 2004), a novel algorithm for gene grouping, which is very different from popular distance-based clustering approaches like hierarchical or k-means clustering. Our method is based on a penalized maximum likelihood principle and operates via a simple but efficient greedy search heuristic. According to classification results on several real-world microarray datasets showing different cancer phenotypes, it indeed is able to identify highly predictive gene clusters.

Traditionally, cancer phenotyping is done on the basis of clinical factors. Many of them are easy to record, and it would be a waste of useful information if modern prognosis just relied on microarray data. We show how *Pelora* can serve as a prediction model selection tool in studies where clinical factors and gene expressions are available. The methodology is illustrated on the breast cancer study of Vant' Veer et al. (2003).

The Algorithm

We denote the ith out of n microarray experiments as a random pair $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i \in R^p$ is the gene expression profile, monitoring up to several thousands of genes. It can either originate from Affymetrix oligonucleotide chips or from two-color cDNA arrays, but it needs to be thoroughly preprocessed and log-transformed. $y_i \in \{0, 1\}$ is a dichotomous response, for example encoding two different cancer phenotypes.

Our algorithm *Pelora* builds on what we call the signature model: not all p genes on the array, but rather a few functional gene subsets $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_q\}$ determine most of the outcome variation. Under this assumption, the conditional probability can be written as

$$P[y = 1|\mathbf{x}] \approx p\left(\widetilde{\mathbf{x}}\right) \text{ with } \widetilde{\mathbf{x}} = \left(\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_q\right),$$

where $\widetilde{x}_j$ are the representative values for the $q \ll p$ unknown gene groups $\mathcal{G}_1, \ldots, \mathcal{G}_q$. Their representatives are chosen to be polarized centroids

$$\widetilde{x}_j = \frac{1}{|\mathcal{G}_j|} \sum_{g \in \mathcal{G}_j} \alpha_g x_g \text{ with } \alpha_g \in \{-1, 1\}.$$

The conditional probability estimate $p(\cdot)$ is obtained from a penalized logistic regression model, i.e.

$$\log\left(\frac{p_(\widetilde{\mathbf{x}})}{1-p_(\widetilde{\mathbf{x}})}\right) \quad = \quad \sum_{j=0}^{q}\theta_j\widetilde{x}_j = \widetilde{\mathbf{x}}\theta.$$

Fitting the signature model now encompasses both the estimation of the model parameters θ and the partition of the thousands of genes into a few signature components $\mathcal{G}_1,\ldots,\mathcal{G}_q$. We aim to tackle this by optimizing one of the most natural and popular goodness-of fit criteria for dichotomous problems, the ℓ_2-penalized log-likelihood function

$$S_{\mathbf{x},y}(\theta,\mathcal{G},\lambda) \quad = \quad -\sum_{i=1}^{n}(y_i\cdot\log p_\theta(\widetilde{\mathbf{x}}_i)+(1-y_i)\cdot\log(1-p_\theta(\widetilde{\mathbf{x}}_i)))+n\frac{\lambda}{2}\theta^TP\theta.$$

Note that λ is a tuning parameter that controls the penalization and P is a matrix that corrects for the generally non-equal variance of the predictors $\widetilde{x}_j$. The optimization problem is much too complex to be solved in full generality. We thus rely on a simplified iterative approach, where the search space with respect to $\mathcal{G}$ is limited to addition and removal of a single gene per iteration.

More specifically, *Pelora* starts from scratch with an empty model. Genes are added one after the other according to optimization of $S_{\mathbf{x},y}(\theta,\mathcal{G},\lambda)$. Regularly recurring pruning steps help to root out genes that were erroneously added to the group at earlier iterations. The first group $\mathcal{G}_1$ is terminated when the criterion cannot be improved any further. Once this happens, a new group is started, but the composition of $\mathcal{G}_1$ remains unchanged. The algorithm proceeds likewise and ends if a pre-specified number q_{final} of gene groups has been found. For further algorithmic details, we refer to Dettling and Bühlmann (2004). In summary, *Pelora* is a maximum-likelihood based procedure for variable selection, variable grouping and formation of new predictive features by averaging the gene expression withing a group, including potential sign-flipping. Software is available under GNU public license as an R-package called `supclust`.

Empirical Results

We evaluated *Pelora* on several different cancer gene expression datasets.

- The leukemia dataset of Golub et al. (1999), describing gene expression levels of 47 patients suffering from ALL and 25 patients being affected by AML.
- The nodal dataset of West et al. (2001), which monitors the gene expression in 49 breast tumor samples. The response variable codes for positive and negative lymph node involvement.
- The colon cancer dataset of Alon et al. (1999) that shows expression levels of 40 tumor and 22 normal colon tissues.
- The prostate cancer dataset of Singh et al. (2002), which comprises the expression of 52 prostate tumor and 50 non-tumor prostate samples.

Extensive experimentation showed that $q_{final} = 10$ gene groups and $\lambda = \frac{1}{32}$ are reasonable default parameters. The typical group size with *Pelora* then lies between 10 and 20 genes and the group centroids clearly separate the phenotypes on training data. For testing the out-of-sample predictive performance of *Pelora*'s estimated conditional class probabilities, we run a comparison against three classifiers that work with 200 single genes chosen by the Wilcoxon statistic. These are the default 1-nearest-neighbor rule, diagonal linear discriminant analysis, and as the state-of-the-art in modern classification, a support vector machine with radial basis kernel.

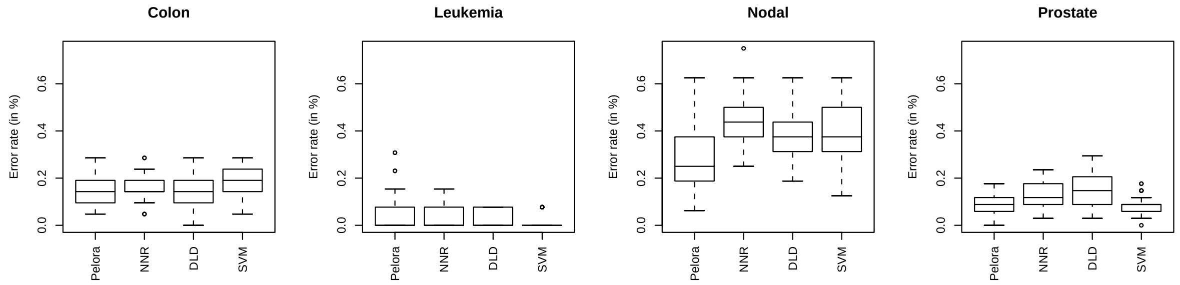

Figure 1: Boxplots, showing the variation of the misclassification rates over 50 random splits into learning sets (two thirds of data) and validations sets (one third of data) for 4 different classifiers: *Pelora*, based on $q = 10$ predictive gene groups and $\lambda = \frac{1}{32}$, as well as the 1-nearest-neighbor rule (NNR), diagonal linear discriminant analysis (DLDA) and a support vector machine (SVM), based on 200 single genes chosen by the Wilcoxon statistic.

Error rates are estimated by repeated random splits into learning sets comprising two thirds, and validation sets containing one-third of the data. According to Figure 1, the predictive potential of *Pelora* is convincing. We observe that it can at least keep up with the benchmark methods, and that is even has an edge on the nodal data. Moreover, *Pelora* is more than just a prediction tool. It also partitions thousands of genes into a few small groups that contain very useful information for explaining the outcome y. This dimensionality reduction is as well worthwile to explore for revealing functional gene groups or regulatory sub-networks.

Significance of Group Centroids and Clinical Variables

Here, we present a modeling and inference approach for combining microarray gene expression data with clinical covariates. Instead of the random pair $(\mathbf{x}, y)$, cancer prognosis shall be done by efficiently exploiting the random triple $(\mathbf{x}, \mathbf{u}, y)$, where $\mathbf{u} \in R^m$ contains the clinical covariates, which can either be continuous, polytomous or binary. The model selection is still done with *Pelora* by optimizing the penalized log-likelihood. The algorithm is slightly altered in that when a new predictor is added to the model, this can either be a gene group centroid or a clinical variable, depending on which yields better predictive value, see Dettling and Bühlmann (2004) for details.

The methodology will be illustrated with the breast cancer dataset of van't Veer et al. (2002). It contains expression values for 78 patients: 34 who developed metastases within 5 years, and 44 who remained disease-free during this period. Furthermore, information about 6 covariates is provided: tumor grade, estrogen receptor status, progesteron receptor status, tumor size, patient age and angioinvasion. When using *Pelora* on the combined breast cancer expression and clinical data, we observe that none of the clinical variables entered the model, even if the number of predictors was raised to $q_{final} = 30$. This can be interpreted that the clinical covariates, compared to the expression profile, do not contain much useful information for class prediction. To exemplify how one can determine which predictors contribute significantly to sample classification, we artificially reduced the breast cancer dataset to 1141 arbitrarily chosen genes. Then, among the first 10 predictors *Pelora* selected, are the intercept, six gene groups and 3 clinical variables. In order of selection, the latter are tumor grade, patient age and angioinvasion.

To answer the question whether some of these clinical covariates, and which of the group centroids, contribute significantly to sample classification, we do bootstrap-based statistical inference on Van't Veers independent breast cancer test dataset, which contains the expression values and clinical data of 19 additional patients: 7 who remained metastasis-free for 5 years and 12 who experienced disease progression. By using only the model-structure from the training data, we fitted penalized logistic regression on the test dataset and obtained the parameter vector $\widehat{\theta}^{test} = (\widehat{\theta}_0^{test}, \ldots, \widehat{\theta}_q^{test})$. To get an impression about the distribution and variability of these coefficients, we generate 1,000 non-parametric

predictor	0	1	2	3	4	5	6	7	8	9
variable	intercept	clinical	group	clinical	group	group	group	group	clinical	group
p-value	0.012	0.000	0.000	0.000	0.136	0.084	0.008	0.146	0.024	0.022

Table 1: Bootstrap p-values for the coefficients of *Pelora's* prediction model on the breast cancer data with 1141 arbitrarily chosen genes. Variables 2, 4, 5-7 and 9 are group centroids, variable 1 is the tumor grade, variable 3 is the patient age and variable 8 is angioinvasion.

bootstrap samples from the test data by drawing with replacement: every run $b \in \{1, \ldots, 1000\}$ yields an estimated parameter vector $\widehat{\theta}^{(b)} = (\widehat{\theta}_0^{(b)}, \ldots, \widehat{\theta}_q^{(b)})$. For quantifying the significance of each predictor variable, we computed the $(1-\alpha)$-bootstrap confidence intervals

$$[2 \cdot \widehat{\theta}_j^{test} - q_{j,(1-\frac{\alpha}{2})}; 2 \cdot \widehat{\theta}_j^{test} - q_{j,\frac{\alpha}{2}}],$$

where $q_{j,\alpha}$ is the α-quantile of the bootstrap distribution. Inverting these intervals leads to the p-values reported in table 1. For the reduced breast cancer dataset with 1141 genes, all fitted predictor variables except for 3 group centroids turned out to be significant at the 5%-level.

REFERENCES

Alon, U., Barkai, N., Notterdam, D., Gish, K., Ybarra, S., Mack, D. and Levine, A. (1999). *Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays.* PNAS **96**, 6745–6750.

Dettling, M. and Bühlmann, P. (2004). *Finding Predictive Gene Groups from Microarray Data.* Journal of Multivariate Analysis **90**, 106–131.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Bloomfield, C. and Lander, E. (1999). *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring.* Science **286**, 531–537.

Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. D'Amico, A., Richie, J., Lander, E., Loda, M., Kantoff, P., Golub, T. and Sellers, W. (2002). *Gene Expression Correlates of Clinical Prostate Cancer Behavior.* Cancer Cell **1**, 203–209.

Van't Veer, L., Dai, H., Van de Vijver, M., He, Y., Hart, A., Mao, M. Peterse, H., Van der Kooy, K., Marton, M., Witteveen, A., Schreiber, G., Kerkhoven, R., Bernards, R. and Friend, S. (2002). *Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer.* Nature **415**, 530–535.

West, M., Blanchette, C., Dressman, H. Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J., Marks, J. and Nevins, J. (2001). *Predicting the Clinical Status of Human Breast Cancer by Using Gene Expression Profiles.* PNAS **98**, 11462–11467.

RÉSUMÉ

Avec des dates de la téchnologie des microarrays, il est un grand défi de chercher des groupes de gènes qui réagissent et qui sont intéressant pour une discrimination des different phènotypes. Pour la recherche de ces groupes, on utilise Pelora*, un algorithme, qui réunit la sélection de gène, le groupement de gène et la classification.* Pelora *fournit des résultats de classification qui sont plein de promesse et il peut aussi manier avec différentes sources de dates.*