Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)
March 20–22, Vienna, Austria
http://www.ci.tuwien.ac.at/Conferences/DSC-2003/
K. Hornik & F. Leisch (eds.) *ISSN 1609-395X*

Revealing Predictive Gene Clusters with Supervised Algorithms

Marcel Dettling
Seminar für Statistik
ETH Zürich
CH-8092 Zürich
Switzerland

Abstract

Microarray technology allows the measurement of expression levels of thousands of genes simultaneously and is expected to contribute significantly to advances in fundamental questions of biology and medicine. While microarrays monitor thousands of genes, there is a lot of evidence that only a few underlying signature components of gene subsets account for nearly all of the outcome variation. Here, methodology for revealing these predictive gene clusters in microarray data is presented. For this task, we focus on supervised algorithms, defined as clustering techniques which utilize external information about the response variables for grouping the explanatory variables (genes). In studies where external response variables are available, our approach is often more effective than unsupervised techniques such as hierarchical clustering.

1 Introduction

Large-scale monitoring of gene expression by microarrays is considered to be the most promising technique to improve medical diagnostics and functional genomics. Given efficient statistical methods for exploiting the relevant information from large gene expression datasets, an accurate classification of tumor subtypes may become reality, allowing for specific therapies that maximize treatment efficacy and minimize toxicity. Moreover, gene expression data are an important resource to reconstruct gene regulatory networks, or more globally, to understand how the genome works.

Our goal is to reveal groups of genes which act together, for example in pathways, and are optimally predictive for a certain type of a disease. In other words, we are searching for rules such as "if in average, gene 534, gene 837 and gene 235 are overexpressed, as well as gene 2194, gene 1438, gene 931 and gene 694 are underexpressed, this is typical for cancer subtype A". These groups of genes can then be used as signature components to accurately predict the phenotypes of new individuals in medical diagnostics and to gain insights into biological and gene regulatory processes. However, finding such groups is difficult: we are facing computational problems due to the sheer amount of predictor variables (genes) which are present, and statistical difficulties due to the "small n, large p" phenomenon.

To tackle the search for co-regulated genes, unsupervised clustering algorithms are widely applied: mostly hierarchical clustering algorithms, but also k-means clustering, self-organizing maps and principal components, among other tools, are used. All these methods cluster genes according to unsupervised similarity measures. Since our goal is to reveal groups of co-regulated predictor variables with strong association to the response variable, we focus on supervised clustering algorithms. They are defined as grouping of predictor variables, controlled by external (supervised) information about the response variables, for example the tumor subtypes that are associated with the arrays. Because of the combinatorial complexity, we rely on a greedy clustering strategy, based on sequentially improving an empirical objective function that measures the strength for cancer type discrimination.

2 Methods

2.1 The partitioning problem

Given a thoroughly preprocessed gene expression profile $X \in \mathbb{R}^p$, which is standardized to zero mean and unit variance, as well as its associated response variable $Y \in \{0, 1\}$, coding for two different phenotypes, we assume that the conditional probability for class membership is given by

$$P[Y = 1|X] = f\left(X_{\mathcal{C}_1}, X_{\mathcal{C}_2}, \ldots, X_{\mathcal{C}_q}\right), \qquad (1)$$

where $f(\cdot)$ is a nonlinear function, $\mathcal{C}_1, \ldots, \mathcal{C}_q$ with $q \ll p$ are gene clusters and $X_{\mathcal{C}_i} \in \mathbb{R}$ are their representative values, defined as $X_{\mathcal{C}_i} = \frac{1}{|\mathcal{C}_i|} \sum_{g \in \mathcal{C}_i} s_g X_g$ with $s_g \in \{-1, 1\}$. This assumption reflects the fact that not all p genes individually, but rather a few underlying marker components of gene subsets determine most of the outcome variation. Even by using the simple arithmetic mean as a group value as we do, finding the optimal partition of thousands of genes into a few clusters is highly nontrivial and the design of a procedure that reveals the best partition from equation (1) is too ambitious. Thus, we suggest computationally intensive procedures that approximately solve the equality in (1) and which yield good empirical results.

2.2 A generic strategy for supervised clustering

Here, we present a heuristic for finding gene clusters, each consisting of a few genes whose mean expression profile is optimally predictive for tissue discrimination. Because of the combinatorial complexity due to the presence of thousands of genes, we rely on a greedy strategy. This is:

1) We start from scratch and grow the clusters incrementally by adding one gene after the other. In each step, try all genes and add the one which improves the cluster most, according to a well-defined clustering criterion S. Repeat the growing until S worsens.

2) Subsequent stepwise pruning helps to remove spurious genes that were incorrectly added to the cluster. In each step, try all clustered genes and exclude the one whose removal improves the cluster most, according to the clustering criterion S. Repeat the pruning until S worsens.

3) If the clustering criterion S cannot be improved any longer by adding or removing genes, the current cluster is terminated and a new cluster is started. From now on, the current and previous clusters remain unchanged.

The very important difference between our and most other clustering algorithms is, that we do not augment (or shorten) the cluster by the gene that suits best (or least) into the current cluster in terms of an unsupervised similarity measure, but base our strategy for supervised clustering of genes on adding (or removing) the gene that improves the differential expression of the current cluster most. Thus, our clustering criterion S is a (possibly penalized) goodness-of-fit measure, which is used to find groups of genes that separate two different tissue types as accurately as possible.

2.3 Wilma - a first implementation

Our first implementation of a supervised clustering algorithm, called *Wilma*, follows exactly the generic strategy described above and was published under the heading "Supervised Clustering of Genes", see Dettling and Bühlmann (2002). The first priority clustering criterion S that measures the strength of differential expression for the two tissue types is the statistic of Wilcoxon's test for two unpaired samples. The criterion is refined with a second priority margin function M, measuring the size of the gap (in standardized gene expression units) between the two response classes. Hence the name Wilma, as an acronym for the *Wil*coxon and *ma*rgin criteria. In Wilma, if a cluster is terminated, all the clustered genes are removed from the expression matrix before the search for the next cluster can begin.

This implementation of supervised clustering yields very good empirical results in terms of the predictive potential, the stability and the relevance of the gene groups it identifies. As an example, figure 1 impressively shows how well the mean expression of the first two clusters separate the 3 response classes of a dataset describing the gene expression of 62 patients suffering from one of the 3 prevalent

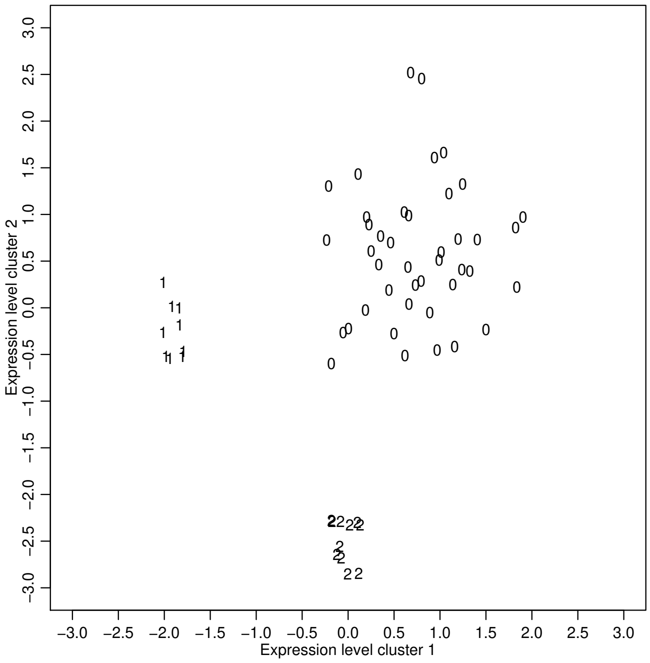

Figure 1: 2-dimensional projection of the lymphoma dataset from Alizadeh et al. (2000) into the space of the first two supervised gene clusters. The data describe the gene expression of 62 patients suffering from one of the 3 prevalent adult lymphoid malignancies.

lymphoid malignancies from Alizadeh et al. (2000). However, there are some limitations. Because the clusters are disjoint, this first implementation cannot capture genes which possibly operate in multiple pathways. Next, each cluster is (up to the disjointness to the former clusters) built independently of all the others. So, it might happen that the clusters are not sufficiently orthogonal. Then, the clustering criterion was non-penalized which might lead to overfitting, although we did not observe this in practice, probably because of the wiggly and very rigorous margin criterion. Moreover, the second priority margin criterion is highly non-robust and results in very hard supervision. The latter has been especially successful in "easy" classification problems, but we expect that some milder form of supervision may lead to better empirical results in problems in difficult, inhomogeneous classification problems with substantial Bayes risk.

2.4 Pelora - a second implementation

Our more refined second proposal of a supervised clustering algorithm is called Pelora, as it is based on *pe*nalized *lo*gistic *r*egression *a*nalysis. It is described in Dettling and Bühlmann (2004), addresses all the limitations of the first implementation and still follows exactly the generic strategy described in section 2.2. It differs in the criterion S the clustering is based on. We work with the penalized negative log-likelihood function

$$\begin{aligned} S &= -\sum_{i=1}^{n} (y_i \cdot \log p_\theta(x_i) + (1-y_i) \cdot \log(1-p_\theta(x_i))) + \lambda P \\ &= -\ell(\theta) + \lambda P, \end{aligned} \tag{2}$$

where $y_i \in \{0,1\}$ are the class labels and x_i are the cluster-dependent predictor variables for experiments $i = 1, \ldots, n$. Furthermore, $p_\theta(x) = P_\theta[Y = 1|X = x]$ are estimates of the conditional class probabilities from a parametric model, λ is a tuning parameter that controls the amount of penalization and P is the penalty term. We rely on the the ℓ_2-penalty, that is, we penalize by $\theta^T\theta$, the dot product of the model parameters.

Our parametric model for estimation of the conditional class probabilities was chosen to be penalized logistic regression analysis (Le Cessie and Van Houwelingen, 1990; Eilers et al., 2001). The classical logistic model is given by

$$\log\left(\frac{p_\theta(x_i)}{1-p_\theta(x_i)}\right) = \sum_{j=0}^{p} \theta_j x_{ij}, \text{ for all } i = 1, \ldots, n.$$

The idea of penalized logistic regression is now to estimate the parameter vector θ by a penalized maximum likelihood principle. We minimize

$$Q(\theta) = -\ell(\theta) + \lambda P \tag{3}$$

with respect to the parameter vector θ. As in equation (2), λ is the tuning parameter that controls the amount of penalization and P is the ℓ_2-Penalty $\theta^T\theta$. Taking derivatives in equation (3) leads to $(p+1)$ nonlinear equations, whose solution needs to be approximated. We do this iteratively by Newton-Raphson stepping. Instead of iterating until convergence, we restrict to 2 full iterations. This saves much computing time and already yields an accurate solution, which from a practical viewpoint can be judged as precise enough. We observed that the clustering decisions did hardly ever change if we ran the algorithm until convergence, instead of doing only 2 iterations.

In summary, the forward step in Pelora works as follows. Assume that clusters $\mathcal{C}_1, \ldots, \mathcal{C}_p$ with predictor variables $x_1, \ldots, x_p$ are already found. We try to augment $\mathcal{C}_p$ and thus repeat for all genes j:

a) construct candidate clusters $\mathcal{C}_p^j$ and corresponding predictor variables x_p^j.

b) by Newton-Raphson stepping, estimate θ^j and use it to compute the negative penalized log-likelihood criterion S_j. The gene $j^* = \arg\min_j S_j$ is the winner. If S_{j^*} is smaller than the best criterion value from the previous round, gene j^* enters the clusters and $\mathcal{C}_p$ as well as x_p are updated, before the search of the next gene is started.

Full details about technical issues and about the clustering procedure can be found in Dettling and Bühlmann (2004).

2.5 Multiclass and continuous response problems

Our clustering procedures Wilma and Pelora can also deal with multiclass problems, which are handled by formulating them as multiple binary problems. This approach has been successful in a variety of problems, and with microarray data, according to our experience from Dettling and Bühlmann (2003), it often works better than simultaneous multiclass versions, especially when variable selection is involved. Various approaches for reduction of multiclass to multiple binary problems exist, see Allwein et al. (2000). We already observed very good empirical results with the most simple solution, the one-against-all approach, which we also used for the lymphoma dataset in this paper.

Pelora, our second implementation of supervised clustering, can easily be adapted to continuous response variables. Instead of the penalized log-likelihood criterion S from equation (2), we recommend the use of the ℓ_2-penalized sum of squared residuals.

$$S = -\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \lambda P$$

This relies on model based fitted response values $\widehat{y}_i$, which we recommend to compute from ridge regression (Hoerl and Kennard, 1970; Ghosh, 2003).

3 Numerical Results

We evaluated our supervised clustering algorithms broadly on several different datasets, all describing the gene expression of cancer patients. The full results can be found in our original papers, see Dettling and Bühlmann (2002, 2004). The output of the supervised algorithms was very promising throughout, since the cluster expression x_C always discriminated the cancer classes very clearly on the training data, as in figure 1. The average cluster size was (depending on the dataset) between 5-7 genes for Wilma and between 15-30 genes for Pelora. The number of clusters can be set according to previous knowledge, can be chosen data-adaptively by cross validation or can be estimated by techniques such as proposed in Dudoit and Fridlyand (2002) or Tibshirani et al. (2000).

10-fold cv	Leukemia	Estrogen	Nodal	Colon	Prostate	Lymphoma
Wilma	2.78%	4.08%	36.79%	11.29%	10.78%	0.00%
Pelora	4.17%	2.04%	14.28%	12.90%	7.84%	0.00%
1-NN	1.29%	12.24%	34.69%	19.35%	11.76%	0.00%
SVM	2.78%	6.12%	28.57%	20.97%	8.82%	1.61%

Table 1: Error rates from 10-fold cross validation with 5 supervised clusters from Wilma and Pelora as predictors in a 1-nearest-neighbor classifier, and with the best 100 single genes according to Wilcoxon's test statistic as input for a 1-nearest-neighbors and support vector machines.

To see whether the output of Wilma and Pelora could successfully reveal functionally relevant groups of genes with good predictive potential, we report the classification results for 5 binary datasets, the famous AML/ALL leukemia dataset of Golub et al. (1999), the two breast cancer datasets with estrogen and nodal response of West et al. (2001), the colon cancer dataset of Alon et al. (1999), the prostate cancer dataset of Singh et al. (2002), and a 3-class problem, the lymphoma dataset of Alizadeh et al. (2000). For further information about the availability of these data and our preprocessing, see Dettling and Bühlmann (2002, 2004).

Because, except for the leukemia data, no genuine test sets are available, we base our empirical study of the predictive potential on 10-fold cross validation. This means that we partition the data into 10 blocks, set aside one block and use the remaining 9 to carry out cluster building and classifier fitting. We then honestly predict the class labels of the left-out block and cycle through all blocks. The test-set error can be determined by calculating the fraction of misclassified observations.

In table 1, the results for our supervised clustering procedures Wilma and Pelora were obtained with $q = 5$ clusters (varying on each block) as predictor variables in a 1-nearest-neighbor classifier. The penalty parameter λ was optimally chosen and varying across the datasets. We compare the results to classification with single genes. For this, we selected the 100 most predictive genes according to Wilcoxon's test statistic on each block. We used them as predictor values for the 1-nearest-neighbor method. Note that the number of genes which are used in the $q = 5$ supervised clusters from Pelora is around 100, too. Moreover, we compare to a support vector machine with linear kernel, a state-of-the-art machine learning method for tumor classification.

The results are in favor of our supervised clustering procedures. Pelora seems to have an edge over Wilma and as expected, the difference is the biggest on the difficult nodal response problem. Classification based on single genes sometimes can, but often cannot keep up with Wilma and Pelora. We never observed that our clusters totally failed or yielded much worse results than single genes. We take this as evidence that our supervised clustering procedures really identify predictive and functionally relevant groups of genes.

4 Conclusions

We have suggested methodology for supervised clustering of genes from microarray experiments. Our procedures are potentially useful in the context of medical diagnostics, as they identify groups of interacting genes which can be used as signatures for tumor classification. At the same time, the clusters may give insight into gene regulation and function.

Our goal in supervised gene clustering is to find gene groups whose average expression renders the discrimination of different tissue types as simple as possible. We solve this by building the clusters incrementally with a stepwise forward and backward strategy. In empirical studies, this yielded excellent classification results, superior to state-of-the-art methods with single genes.

References

Ash Alizadeh, Mike Eisen, Eric Davis, Chi Ma, Izidore Lossos, Andreas Rosenwald, Jennifer Boldrick, Hajeer Sabet, Truc Tran, Xin Yu, John Powell, Liming Yang, Gerald Marti, Troy Moore, James Hudson, Lisheng Lu, David Lewis, Robert Tibshirani, Gavin Sherlock, Wing Chan, Timothy Greiner, Dennis Weisenburger, James Armitage, Roger Warnke, Ronald Levy, Wyndham Wilson, Michael Grever, John Byrd, David Botstein, Patrick Brown, and Louis Staudt. Distinct types of diffuse large b-cell-lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

Erin Allwein, Robert Schapire, and Yoav Freund. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1:113–141, 2000.

U Alon, N Barkai, D Notterdam, K Gish, S Ybarra, D Mack, and A Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Science*, 96:6745–6750, 1999.

Marcel Dettling and Peter Bühlmann. Supervised clustering of genes. *Genome Biology*, 3:research 0069.1–0069.15, 2002.

Marcel Dettling and Peter Bühlmann. Boosting for tumor classification with microarray data. *Bioinformatics*, 19:1061–1069, 2003.

Marcel Dettling and Peter Bühlmann. Finding predictive gene groups from microarray data. *Journal of Multivariate Analysis*, 2004. To appear.

Sandrine Dudoit and Jane Fridlyand. A prediction-based resampling method to estimate the number of clusters in a dataset. *Genome Biology*, 3:research 0036.1–0036.21, 2002.

Paul Eilers, Judith Boer, Gert-Jan Van Ommen, and Hans Van Houwelingen. Classification of microarray data with penalized logistic regression. In *Proceedings of SPIE: Progress in Biomedical Optics and Imaging*, volume 2, pages 187–198, 2001.

Debashis Ghosh. Penalized discriminant methods for the classification of tumors from gene expression data. *Biometrics*, 59, 2003.

Todd Golub, Donna Slonim, Pablo Tamayo, Carine Huard, Michelle Gassenbeek, Jill Mesirov, H Coller, M Loh, J Downing, M Caliguri, C Bloomfield, and Eric Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–538, 1999.

Arthur Hoerl and Robert Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

Saskia Le Cessie and J Van Houwelingen. Ridge estimators in logistic regression. *Applied Statistics*, 41:191–201, 1990.

Dinesh Singh, Philipp Febbo, Kenneth Ross, Donald Jackson, Judith Manola, Christine Ladd, Pablo Tamayo, Andrew Renshaw, Anthony D'Amico, Jerome Richie, Eric Lander, Massimo Loda, Philipp Kantoff, Todd Golub, and William Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1: 203–209, 2002.

Robert Tibshirani, Gunther Walther, and Trevor Hastie. Estimating the number of clusters in a dataset via the gap statistic. Technical Report 208, Department of Statistics, University of Stanford, 2000.

Mike West, Carrie Blanchette, Holly Dressman, Erich Huang, Seiichi Ishida, Rainer Spang, Harry Zuzan, John Olson, Jeffrey Marks, and Joseph Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Science*, 98:11462–11467, 2001.