Graph-Constrained Discriminant Analysis of functional genomics data

Vincent Guillemot [†] [‡], Laurent Le Brusquet [‡], Arthur Tenenhaus [†] , Vincent Frouin [†]

[†] CEA, iRCM, Laboratoire d'Exploration Fonctionnelle des Génomes, F-91000, France.
[‡] Supélec, Department of Signal and Electronic Systems, F-91190, France.

Abstract

Classification studies from microarray data have proved useful in tasks like predicting patient class. At the same time, more and more biological information about gene regulation networks has been gathered mainly in the form of graph. Incorporating the a priori *biological information encoded by graphs turns out to be a very important issue to increase classification performance. We present a method to integrate information from a network topology into a classification algorithm: the graph-Constrained Discriminant Analysis (gCDA). We applied our algorithm to simulated and real data and show that it performs better than a linear Support Vector Machines classifier.*

1 Introduction

In functional genomics, the issue of integrating information about gene-coregulation arises more and more frequently in the analysis of microarray data: this information may consist in *a priori* or results from other experiments. For example, some differential analysis procedures, incorporating gene networks, were recently proposed [14]. Several methods can be found in the literature to achieve the integration of the *a priori* information one have on the relations between genes into the classification task [9], [11].

In [9], the authors proposed to fit a linear model $y = \beta X$ on the dataset X to predict the binary variable y. Their approach follows regularization ideas that consist in optimizing the following criterion

$$\begin{aligned} J(\beta, c_1, c_2) &= (y - \beta X)^\top (y - \beta X) + c_1 \left|\beta\right| \\ &+ c_2 \beta^\top \mathcal{L}_{\mathcal{G}} \beta \end{aligned}$$

with respect to β, where $\mathcal{L}_{\mathcal{G}}$ is the Laplacian of the *a priori* graph $\mathcal{G}$ (see paragraph 2.1). The solution of this optimization problem is achieved through a Lasso procedure [5] and leads to a sparse model ; it is used to characterize the subnetworks from the reference network that are "expressed" in the dataset.

In [11], a spectral transformation is applied to the graph $\mathcal{G}$ and then a kernel Support Vector Machines classification is performed. The new metric between two expression profiles f and g is:

$$d_\phi(f;g) = f^\top K_\phi g \tag{1}$$

, where K_ϕ is the positive semidefinite matrix obtained after the spectral transformation of graph $\mathcal{G}$.

These procedures lead to a better graph-interpretability of the resulting classifiers. But the goal of these integration algorithms is not to improve classification performance.

We propose a method, based on the Discriminant Analysis introduced by Fisher [6], that takes into account an *a priori* and improves classification performance when compared to competitive classification algorithm (here Support vector Machines [4]). We considered first simulated data to demonstrate our point; the parameters of the simulated dataset generator were tuned in order to mimic any given real dataset while following some interaction constraints. Then, we applied our method on data from the literature [1]. We focused on binary classification problems, and left multiclass problems as future work.

The remainder of the paper is structured as follows: in the second part gCDA is presented and the last part is devoted to the validation of gCDA and its comparison with the state-of-the-art methods on simulated and real datasets.

2 Graph-Constrained Discriminant Analysis

A constrained version of Discriminant Analysis is developed to compute, from a learning dataset, a classifier able to predict the class of a new sample.

2.1 Interaction between genes

A graph $\mathcal{G}$ is defined by the set of its edges $\mathcal{E}$ and the set of its vertices $\mathcal{V}$. Let w be the function $w : \mathcal{V} \times \mathcal{V} \to \{0,1\}$ such that $w(i,j)$ is 1 if there is an edge between vertices i and j and 0 otherwise. For each vertex i of $\mathcal{V}$, the connectivity degree d_i is defined as the cardinality of the set of the vertices in $\mathcal{V}$ connected to i. In the following, $\mathcal{G}$ is a finite graph with p vertices (genes) and m edges.

We take a graph of independence as equivalent to a graph of gene regulations (which is only a partial view of the biological concept of a regulation network). Each vertex represents a variable and the absence of an edge between two vertices means that the two variables are independent knowing the remaining variables. For example, there is no edge between variables 4 and 5 (denoted respectively by x_4 and x_5) in the graph depicted on figure 1: x_4 and x_5 are independent knowing the remaining variables. The precision matrix of the multivariate Gaussian variable $X = [x_1, ..., x_n]$, which is also its inverse covariance matrix Σ^{-1}, has the following property [15]:

$$\begin{array}{rcr} & x_4 \perp\!\!\!\perp x_5 | \left\{x_j, j \in \{1, ..., p\} \setminus \{4,5\}\right\} & (2) \\ \Leftrightarrow & \left[\Sigma^{-1}\right]_{4,5} = 0 & \end{array}$$

In this paper, we propose to use property (2) to derive an estimation of the covariance matrix $\Sigma_{\mathcal{G}}$ from the Laplacian matrix $\mathcal{L}_{\mathcal{G}}$ of an *a priori* graph $\mathcal{G}$. Indeed $\mathcal{L}_{\mathcal{G}}$, is a semi definite positive $p \times p$ matrix which coefficients are:

$$[\mathcal{L}_{\mathcal{G}}]_{i,j} = \begin{cases} -w(i,j) & \text{, if } i \neq j \\ d_i & \text{, if } i = j \end{cases}$$

and in which each null term corresponds to an absence of edge in $\mathcal{G}$. Thus, after the addition of a $\delta > 0$ on the diagonal of $\mathcal{L}_{\mathcal{G}}$, we have a good positive definite candidate for the precision matrix of a multivariate Gaussian variable of which $\mathcal{G}$ is an independence graph:

$$\Sigma_{\mathcal{G}}^{-1} = \mathcal{L}_{\mathcal{G}} + \delta I \qquad (3)$$

with I the identity matrix of size p. In the gCDA algorithm, the property 3 is used to compute covariance matrices from both microarray data and an *a priori* graph.

2.2 Discriminant Analysis

Let $X_1 = (x_1^{(i)})_{i \in \{1,...,n_1\}}$ (resp. $X_2 = (x_1^{(i)})_{i \in \{1,...,n_2\}}$) be an n_1-sample (resp. n_2) of a Gaussian multivariate variable with mean μ_1 (resp. μ_2) and covariance matrix Σ, and let $X = [X_1, X_2]$ be the corresponding 2 classes dataset. With these notations, we can define the between and within classes covariances as

$$S_b = n_1(\overline{x}_1 - \overline{x})(\overline{x}_1 - \overline{x})^\top + n_2(\overline{x}_2 - \overline{x})(\overline{x}_2 - \overline{x})^\top \qquad (4)$$

and

$$S_w = \frac{1}{n-2} \sum_{k=1}^{2} \frac{1}{n_k} \sum_{j=1}^{n_k} \left(x_k^{(j)} - \overline{x}_k \right) \left(x_k^{(j)} - \overline{x}_k \right)^\top , \tag{5}$$

where $\overline{x}_k$ is the empirical mean of X_k and $\overline{x}$ the empirical mean of X.

The goal of the Discriminant Analysis is to determine a linear transformation W of dataset X that minimizes the within class variance to between class variance ratio. Provided that S_w^{-1} exists, this problem is equivalent to perform the eigenvalue decomposition of the matrix $S_b S_w^{-1}$:

$$S_b S_w^{-1} = V D V^\top$$

and to choose $W = V$ as the linear transformation.

Once W is computed, the prediction of the class of a new individual x^{new} is based on the estimation of the posterior probability of the class ω_k knowing the sample x^{new}:

$$\begin{aligned} \log \Pr(\omega_k | z^{new} = W x^{new}) &= \\ -\frac{1}{2}(z^{new} - \mu_k)^\top \widehat{\Sigma}_k^{-1} (z^{new} - \mu_k) & \\ - \log \det \widehat{\Sigma}_k + \log \pi_k + a & \end{aligned}$$

where π_k is the prior probability of each class, $\widehat{\Sigma}_k$ is the estimation of the covariance matrix of class k after the transformation by W and $a \in \mathbb{R}$ is a constant which does not depend on the class k. x^{new} will be attributed the class k^{new} maximizing the posterior probabilities according to two different strategies. The so-called Linear Discriminant Analysis assumes that $\widehat{\Sigma}_k = \widehat{\Sigma}$ does not depend on the class k. On the contrary, the Quadratic Discriminant Analysis assumes that the estimation of $\widehat{\Sigma}_k$ has to be different from one class to another.

2.3 Taking into account the *a priori* graph

In the $n \ll p$ case, the Maximum Likelihood Estimators (MLE) presented in equations 4 and 5 are unbiased but show poor performances as regards their variance. In order to overcome this problem several approaches are proposed in the literature. In [8], the authors present a method that consists in inferring a regularized version of the covariance matrix:

$$\widetilde{\Sigma} = \alpha \widehat{\Sigma} + (1 - \alpha) I \tag{6}$$

and in [12], a similar strategy of regularization is used to shrink the estimator of the covariance matrix by replacing I in equation 6 by sparse positive definite matrices, and the parameter α is analytically computed.

Our method is inspired by the previously reported ideas: in the graph-Constrained Discriminant Analysis, I is replaced by the covariance matrix obtained from the Laplacian of the *a priori* graph:

$$\widetilde{\Sigma} = \alpha \widehat{\Sigma} + (1 - \alpha) \left(\mathcal{L}_{\mathcal{G}} + \delta I \right)^{-1}$$

The α parameter is determined with a cross-validation procedure.

In this context, the within class variability estimation can be redefined:

$$\widetilde{\Sigma}_w(\alpha) = \frac{n_1}{n} \widetilde{\Sigma}_1(\alpha) + \frac{n_2}{n} \widetilde{\Sigma}_2(\alpha) \tag{7}$$

to perform a Linear or a Quadratic graph-Constrained Discriminant Analysis.

In the Linear gCDA, each class is supposed to have the same covariance and there is only one *a priori* graph:

$$\widetilde{\Sigma}_w(\alpha) = \alpha S_w + (1 - \alpha) \left(\mathcal{L}_{\mathcal{G}} + \delta I \right)^{-1} \tag{8}$$

In the Quadratic gCDA, each class is characterized by a different *a priori* graph:

$$\begin{aligned} \widetilde{\Sigma}_w(\alpha) &= \alpha S_w + (1 - \alpha) \Big(\left(\mathcal{L}_{\mathcal{G}_1} + \delta I \right)^{-1} \\ &+ \left(\mathcal{L}_{\mathcal{G}_2} + \delta I \right)^{-1} \Big) \end{aligned} \tag{9}$$

3 Results

In this section the two datasets we used to challenge gCDA are presented. The first one is a simulated dataset and the second one is a microarray dataset [1]. In what follows, n_1 and n_2 are the number of individuals in classes 1 and 2, p is the number variables, $\mathcal{G}$ the graph of independence, μ the mean and Σ the covariance matrix. We compare gCDA to Support Vector Machines [4].

The characteristics of the two datasets are represented in table 1, the number of iterations used to compute the test error rate with a Monte Carlo Cross Validation algorithm [3] is fixed to $B = 50$.

	p	n_1	n_2
Simulated	100	20	20
Alon	100	26	14

Table 1. Characteristics of the datasets

3.1 Data simulation

Since we want to evaluate the integration of a graph into a classification, we have to design a dataset with a given independence graph structure. Classically, microarray data are simulated by sampling the covariance matrix of a real dataset ; the underlying structure of interactions remains unknown in those simulations. Other simulation strategies consider the generation of a random graph, followed by the solution of a system of Ordinary Differential Equations [10], which implies the choice of numerous parameters for a graph with a realistic size. We formulate the graph constraint as follows: the dataset is a multivariate Gaussian sample with known mean, and with a covariance matrix built from the given graph of independence between the variables.

We make the assumption that the graph of independence underlying each class is the same and is characterized by a scale free distribution of the node degrees [2]. Figure 1 shows a graphical representation of a random graph according to the algorithm we used. We notice that no loop is generated in such graphs, which is probably unrealistic from a biological point of view, but does not impact the classification procedure.

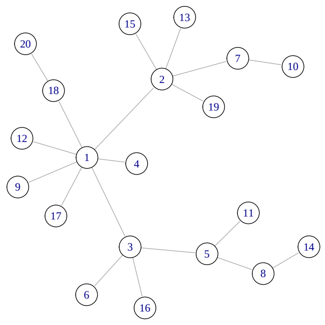

Figure 1. An example of a random graph.

For each class k, our simulation algorithm can be summed up as follows:

a) random generation of network $\mathcal{G}$

b) the mean vector μ_k of the expression profiles mimics the distribution of mean expressions in real datasets [7]

c) the covariance matrix $\Sigma_{\mathcal{G}}$ is computed from the graph $\mathcal{G}$ with respect to equation 3

d) an expression profile is a multivariate Gaussian random vector of mean μ_k and covariance matrix $\Sigma_{\mathcal{G}}$

For a given graph, we compute a corresponding covariance matrix and generate two samples of Multivariate Gaussian Variables X_1 and X_2. We tested gCDA with the "real" graph and with another random graph. The results are presented on figure 2.

It can be seen that constraining the classification with a random graph is not better than

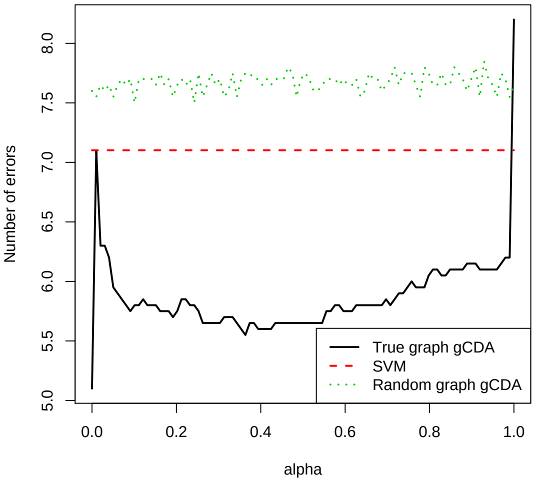

Figure 2. Test error rate against α.

performing SVM, while gCDA shows an improvement of the error rate of approximately 4 % (1.6 errors out of 40 individuals).

3.2 Real microarray data

100 genes were randomly chosen among the 2000 genes of the *Alon* dataset [1]. Since we do not have any *a priori* graph of genetic regulation for this dataset, we applied a new algorithm using Partial Least Squares regressions [13] to infer the independence graph on an independent set of microarrays. The *Alon* dataset is then split into two independent datasets: the first one is used to infer a regulation graph for each class and on the second one is applied the Quadratic procedure described in section 2: the Linear procedure does not work, hypothetically because the independence graphs are different from one class to another (which is confirmed by the graph inference step: there are only 7 % common edges between the two graphs). Results are given in table 2. To compare the performance, we also report the results on the simulated data

data	gCDA	SVM
simulated	13.75%	17.75%
Alon	3.75%	4.5%

Table 2. Results on real data

On figure 3, we see that there is an optimal $\alpha = 0.21$ for which the test error rate is significantly inferior to the test error rate obtained with SVM (a Wilcoxon test applied on the two sets of B error rates gave a p-value of 1 %).

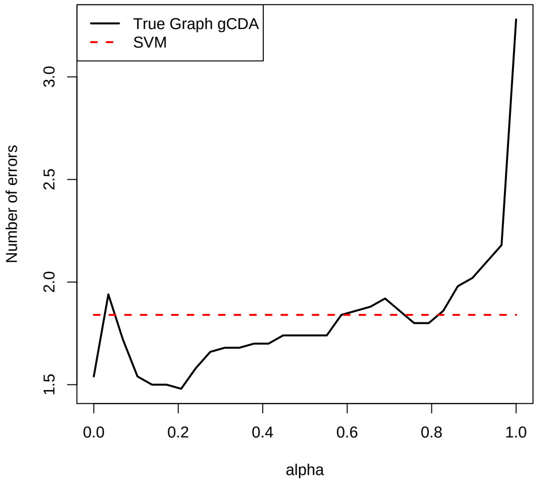

Figure 3. Test error rate against α.

4 Conclusion

We show a significant improvement in classification performance when the underlying graph of regulation is known in the case of simulated data or when the underlying graph of regulation is inferred in the case of real microarray data.

5 Perspective

The graph-constrained estimator of the covariance matrix will be studied in the spirit of [12] to characterize its bias and variance. We will also study the differences on simulated data between the Linear and the Quadratic versions of gCDA. And finally we will work on an implementation of gCDA able to cope with thousands of variables (for the moment it is possible to run gCDA with only hundreds of variables).

References

[1] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.*, Proc Natl Acad Sci U S A **96** (1999), no. 12, 6745–6750.

[2] A. L. Barabasi and R. Albert, *Emergence of scaling in random networks*, Science **286** (1999), no. 5439, 509–512.

[3] A.-L. Boulesteix, C. Strobl, T. Augustin, and M. Daumer, *Evaluating microarray-based classifiers: an overview*, Cancer Informatics **6** (2008), 53–61.

[4] C. Cortes and V. Vapnik, *Support-vector networks*, Machine Learning **20** (1995), 273–297.

[5] B. Efron, T. Hastie, L. Johnstone, and R. Tibshirani, *Least angle regression*, (2002).

[6] R. A. Fisher, *The use of multiple measurements in taxonomic problems*, Annals of Eugenics **7** (1936), 179–188.

[7] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.*, Science **286** (1999), no. 5439, 531–537.

[8] Y. Guo, T. Hastie, and R. Tibshirani, *Regularized linear discriminant analysis and its application in microarrays.*, Biostatistics **8** (2007), no. 1, 86–100.

[9] C.Y. Li and H.Z. Li, *Network-constrained regularization and variable selection for analysis of genomic data.*, Bioinformatics **24** (2008), no. 9, 1175–1182.

[10] P. Mendes, W. Sha, and K. Ye, *Artificial gene networks for objective comparison of analysis algorithms.*, Bioinformatics **19 Suppl 2** (2003), ii122–ii129.

[11] F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot, and J.P. Vert, *Classification of microarray data using gene networks.*, BMC Bioinformatics **8** (2007), 35.

[12] J. Schäfer and K. Strimmer, *A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics.*, Stat Appl Genet Mol Biol **4** (2005), Article32.

[13] A. Tenenhaus, V. Guillemot, X. Gidrol, and V. Frouin, *Gene association networks from microarray data using a regularized estimation of partial correlation based on pls regression*, IEEE/ACM Transactions on Computational Biology and Bioinformatics (accepted).

[14] P. Wei and W. Pan, *Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model.*, Bioinformatics **24** (2008), no. 3, 404–411.

[15] J. Whittaker, *Graphical Models in Applied Multivariate Statistics*, New York, Wiley, 1990.