Optimal Screening Methods in Gene Expression Profiles Classification

Sandra Ramos[1], Antónia Amaral Turkman[2] and Marília Antunes[2]

[1] Instituto Superior de Engenharia do Porto - Instituto Politécnico do Porto
Rua Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal
sfr@isep.ipp.pt

[2] Faculdade de Ciências, Universidade de Lisboa
DEIO, Bloco C6, Campo Grande 1749-016 Liboa, Portugal
antonia.turkman@fc.ul.pt, marilia.antunes@fc.ul.pt

Abstract. We propose the application of a Bayesian Optimal Screening Method to classify an individual in one of two groups (presence/absence of disease) based on the observation of pairs of covariates, namely the expression level of pairs of genes. The method is general and can be applied to any correlated pair of covariates with bivariate normal distribution or that can be tranformed in a bivariate normal. In this case, the boundaries of the optimal screening region are approximated by a quadratic function of the screening variables. The classifier was evaluated on data from three gene expression studies - Leukemia, Prostate and Breast cancers - found in the literature. The classification error rates were calculated using the leave-one-out cross-validation approach.

Keywords: screening methods, DNA microarrys, classification

1 Introduction

Microarray technology is a powerful tool for genomic research, which allows the monitoring of expression profiles for tens of thousands of genes in parallel and is already producing huge amounts of data (Duggan et al. (1999)). However, the number of profile measurements per experimental study remains quite small, usually fewer than one hundred. The small-sample dilemma in the statistical methods for classification in micrroarray data is well documented in the literature (Dudoit et al. (2003)), with some simplifying assumptions appearing as necessary (such as the reduction of the dimensionality of the data). Geman et al. (2004) and Bo et al. (2002) propose the use of marker gene pairs for classification. In this paper, we propose the use of optimal screening methods applied to pairs of gene expression levels for classification purposes.

The screening method consists in the identification of successful individuals of the population, based on the observation, $\mathbf{x}$, of a feature vector $\mathbf{X}$ for each individual.

The purpose of screening is to find a region $C_{\mathbf{x}}$ such that if $\mathbf{x} \in C_{\mathbf{x}}$ the probability that the individual is considered a success is maximized (Turkman and Amaral Turkman (1989)). In section 2 we describe the fundamental concepts of the screening methodology, applied to classification based on the observation of expression levels of pairs of genes.

We demonstrate the usefulness of this methodology using several public data sets involving leukemia, breast and prostate cancers. The performance of the procedure will be evaluated using leave-one-out cross-validation and will be displayed for each data set. Results are presented in section 3 and conclusions and final remarks in section 4.

2 Method

We suggest an application of the screening methodology in supervised classification based on observation of pairs of genes. In this section we explain the main theoretical tools that are necessary to understand the methodology, and its application in classification problems.

2.1 Optimal screening methods in classification of gene pairs

Consider two genes whose expression levels $\mathbf{X} = (X_1, X_2)$ (measured using DNA microarrays) are regarded as random variables, each profile $\boldsymbol{X}$ having a true class label in $\{0, 1\}$. Let Y be a binary random variable that assumes value 1 (success) if the profile $\boldsymbol{X}$ has class 1 and assumes the value 0 otherwise. Suppose that we have a random sample of n individuals, $\mathcal{D} = \{(y_1, x_{11}, x_{21}), \cdots, (y_n, x_{1n}, x_{2n})\}$, for which the true classification label is known. The optimal screening problem has been stated by Turkman and Amaral Turkman (1989) and in this case the optimal region is

$$C_{\mathbf{x}} = \left\{\mathbf{x} \in \mathbb{R}^2 : P(Y = 1|\mathbf{x}, \mathcal{D}) \geq k\right\} \tag{1}$$

or equivalently

$$C_{\mathbf{x}} = \left\{\mathbf{x} \in \mathbb{R}^2 : \frac{P(Y = 1|\mathcal{D})\, p(\mathbf{x}|Y = 1, \mathcal{D})}{\sum_{i=0,1} P(Y = i|\mathcal{D})\, p(\mathbf{x}|Y = i, \mathcal{D})} \geq k\right\} \tag{2}$$

where k is such that

$$P(\mathbf{X} \in C_{\mathbf{x}}|\mathcal{D}) = \alpha. \tag{3}$$

We consider the case where $Y \sim$Ber(θ), $\theta \in (0, 1)$, and for $i = 0, 1$, $\log \mathbf{X}|Y = i \sim N_2\left(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i = \boldsymbol{\Sigma}_i^{-1}\right)$. The model parameters are $(\theta, \Theta_0, \Theta_1)$, where $\Theta_i = (\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)$, $i = 0, 1$. We assume that *a priori* the parameters θ, $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Theta}_1$ are independent.

If we assume a Beta prior distribution for θ ($\theta \sim$Be(a, b), $a > 0$, $b > 0$) and a conjugate prior for $(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)$ of the form $p(\boldsymbol{\mu}_i|\boldsymbol{\Lambda}_i)\, p(\boldsymbol{\Lambda}_i)$, where $p(\boldsymbol{\mu}_i|\boldsymbol{\Lambda}_i)$ is

$N_2\left(\boldsymbol{\mu}_{0i}, c_i \boldsymbol{\Lambda}_i\right)$ and $p\left(\boldsymbol{\Lambda}_i\right)$ is $\mathrm{Wi}_2\left(\alpha_i, \boldsymbol{\beta}_i\right)$, the predictive distribution of a future observation in class $Y = i$ is (Aitchison and Dunsmore, 1975)

$$\log \mathbf{X}|(Y = i, \boldsymbol{D}) \sim \mathrm{St}_2\left(\boldsymbol{\mu}_{ni}, \left(c_i + n_i + 1\right)^{-1}\left(c_i + n_i\right)\alpha_{ni}\boldsymbol{\beta}_{ni}^{-1}, 2\alpha_{ni}\right) \tag{4}$$

where

$$\alpha_{ni} = \alpha_i + \frac{1}{2}\left(n_i - 1\right),$$

$$\boldsymbol{\mu}_{ni}\left(c_i + n_i\right)^{-1}\left(c_i\boldsymbol{\mu}_{0i} + n_i\bar{\mathbf{x}}_i\right),$$

and

$$\boldsymbol{\beta}_{ni} = \boldsymbol{\beta}_i + \frac{1}{2}S_i + \frac{1}{2}\left(n_i + c_i\right)^{-1}\left(\boldsymbol{\mu}_{0i} - \bar{\mathbf{x}}_i\right)\left(\boldsymbol{\mu}_{0i} - \bar{\mathbf{x}}_i\right)^t.$$

The predictive probability of a future individual to be a success, $(Y = 1)$, is

$$\gamma = P\left(Y = 1|\mathcal{D}\right) = \frac{n_1 + a}{n + a + b}, \tag{5}$$

with $n = n_0 + n_1$, where n_i is the number of individuals in the sample for which $Y = i$.

The following predictive probabilities are called operating characteristics (OC) of the screening region,

1. $\alpha = P\left(\mathbf{X} \in C_{\mathbf{x}}|\mathcal{D}\right)$
2. $\gamma = P\left(Y = 1|\mathcal{D}\right)$
3. $\delta = P\left(Y = 1|\mathbf{X} \in C_{\mathbf{x}}, \mathcal{D}\right)$
4. $\varepsilon = P\left(Y = 1|\mathbf{X} \notin C_{\mathbf{x}}, \mathcal{D}\right)$

2.2 Classification

Prior to the classification procedure, the selection of differentially expressed genes, results in a family $\mathcal{P} = \{\mathbf{X}_j = \left(X_{j1}, X_{j2}\right), j = 1, \ldots, m\}$ of m distinct pairs. Usually, there are only a few pairs of genes good for discrimination purposes; for example, in two of the three experiments presented here there is only one such pair, and in the Leukemia study there are three pairs (Geman et al. (2004), Bo and Jonassen (2002)). We use $\mathcal{P}$ as input of our method. For each pair in $\mathcal{P}$, the classification rule and the operating characteristics are obtained for several values of k, defined in (1). The optimal k is the one which renders the best collection of operating characteristics and gives the smallest number of profiles incorrectly classified.

Consider a new individual, with family of profiles $\mathcal{P}$ with m pairs. Based on the $j - th$ pair, he is classified in $C_j = 1$ if the observed profile $\mathbf{x}_j \in C_{\mathbf{x}_j}$. Otherwise he is classified in $C_j = 0$. Let δ_j be the corresponding predictive probability of success given that his $j-th$ profile belongs to $C_{\mathbf{x}_j}$. Then, the final classification rule is given by

$$C = \text{Round}\left(\frac{C_1\delta_1 + C_2\delta_2 + \cdots + C_m\delta_m}{\delta_1 + \delta_2 + \cdots + \delta_m}\right). \tag{6}$$

where $C_j \in \{0,1\}$

2.3 Error Estimation

The classification performance, for all data sets, is assessed using leave-one-out cross validation procedure. The Leukemia study (Golub et al. (1999)) has one training and test data set, but in order to use the same method of error estimation on all studies, we combined these two data sets into one.

3 Application

3.1 Data sets

Prostate study- The data is drawn from the study of prostate cancer reported in Singh et al. (2002). This study assigns profiles to either tumor or normal tissues classes based on expression values for 12600 genes. There are $n_1 = 52$ prostate tumor samples and $n_0 = 50$ non-tumor samples, selected from among several hundred radical prostatectomy patients. The top scoring gene pair used as input for the screening classifier is M84226 and M55914. The joint behaviour of this pair of genes, as we will see, is highly discriminative of prostate tumor versus non-tumor samples, yielding an error of 5.43%.

Leukemia study- This study (Golub et al, (1999)) compares two different types of leukemia (Acute Myeloid and Acute Lymphoplastic, ALL *vs* AML) with 7129 probes (6187 human genes) from 27 samples of ALL and 11 samples of AML. There is also a test set consisting of 34 samples (20 ALL and 14 AML). In order to use the same method of error estimation on all studies, we combined the two data sets into one of size $n = 72$ (47 ALL and 25 AML). Negative values due to normalization and/or background correction were eliminated in order to apply the logarithmic transformation and hence the final data set has size $n = 63$ with $n_1 = 38$ ALL samples and $n_0 = 25$ AML samples. The screening classifier uses three gene pairs (five genes) and classifies 60 samples correctly out of 63.

Breast Study- The data set (Huang et al. (2003)) consists of gene expression profiles measured in 52 women with breast cancer. $n_0 = 34$ women did not experience recurrence of the tumor during a 3 years time period and $n_1 = 18$ experienced the recurrence of the tumor. The screening classifier uses only one pair of genes ($38895 - i - at$ and $32625 - at$). The estimated error rate is 11.54%.

3.2 Classification Results

For each study, and for each gene pair in the corresponding $\mathcal{P}$ family, the approximated optimal screening region was computed together with the operating characteristics. All the procedure is automatically implemented in R.

For each study, we present the scatterplot of the log expression levels for two genes - the unique pair for Prostate (Fig. 1) and Breast data (Fig. 3) and one of the three pairs for the Leukemia data (Fig. 2).

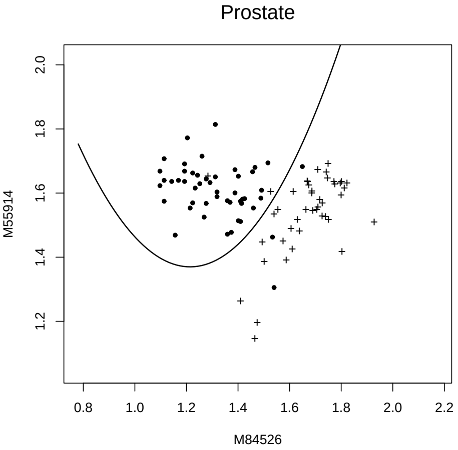

Fig. 1. Scatterplot for a pair of genes for Prostate study. Classes are represented using dots (C_1) and crosses (C_0). The axes represent the logarithm of the expression levels of the two genes. The curve, $x_2 = 4.3726 - 4.9457x_1 + 2.0364x_1^2$, represents the decision boundary.

Table 1 shows the operating characteristics of the optimal screening region for the represented gene pairs. The estimated prediction error rate of the classifier for each study is displayed in Table 2.

Problem	k	$P(Y=1\|\mathcal{D})$	$P(\mathbf{X}\in C_{\mathbf{x}}\|\mathcal{D})$	$P(Y=1\|\mathbf{X}\in C_{\mathbf{x}},\mathcal{D})$	$P(Y=1\|\mathbf{X}\notin C_{\mathbf{x}},\mathcal{D})$
Prostate	0.63	0.5319	0.5153	0.9399	0.0981
Leukemia	0.70	0.6000	0.5456	0.9814	0.1421
Breast	0.42	0.3519	0.3128	0.8458	0.1269

Table 1. Operating characteristics for the best value of k

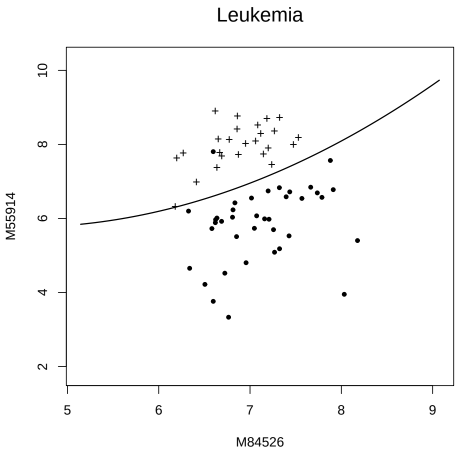

Fig. 2. Scatterplot for a pair of genes for Leukemia study. Classes are represented using dots (C_1) and crosses (C_0). The axes represent the logarithm of the expression levels of the two genes. The curve, $x_2 = 9.5632 - 1.6949x_1 + 0.1889x_1^2$, represents the decision boundary.

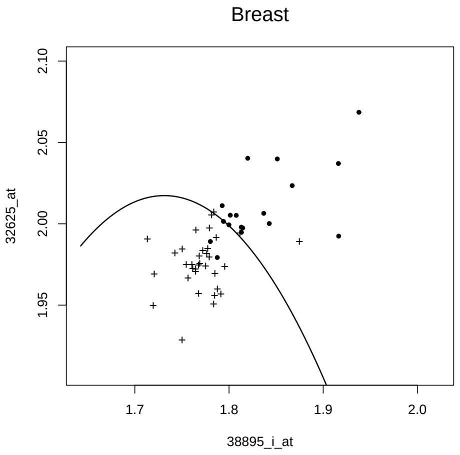

Fig. 3. Scatterplot for a pair of genes for Breast study. Classes are represented using dots (C_1) and crosses (C_0). The axes represent the logarithm of the expression levels of the two genes. The curve, $x_2 = -9.7506 + 13.5948x_1 - 3.9263x_1^2$, represents the decision boundary.

4 Conclusions and further work

We have introduced a new classification methodology for microarray data based entirely on expression levels of pairs of genes. In bivariate normal case,

Problem	Sample Size	# genes	Error (%)
Prostate	102	2	5.43%
Leukemia	63	5	4.76%
Breast	52	2	11.54%

Table 2. Classification error rate for a pair of genes for each study. The results are based on leave-one-out cross-validation.

the optimal screening region is approximated by a quadratic function of the screening variables. We have chosen leave-one-out cross-validation to estimate the error rate of the classifier. For the three data sets presented here the estimated prediction rate is very satisfactory.

The computer code used to obtain the optimal screening regions, compute the operating characteristics and perform the final classification has been written in R.

It is our aim to make the programs fully automatic so that it can be generally used and made available to the R community.

5 Acknowledges

The authors acknowledge the support given by CEAUL (Centro de Estatística e Aplicações da Universidade de Lisboa) and FCT (Fundação da Ciência e Tecnologia) - projects FCT/POCI/2010 and FCT/PTDC/MAT/64353/2006.

References

AITCHISON,J. and DUNSMORE, I.R. (1975): *Statistical Prediction Analysis. Cambdridge University Press.*

Bo, T. H. and JONASSEN, I. (2002): New feature subset selection procedures for classification of expression profiles. *Genome Biology, 3(4): research0017.1-0017.11.*

DUDOIT and FRIDLYAND, J. (2003): *Classification in microarrays experiments. In T. Speed, editor, Statistical Analysis of Gene Expression Microarray Data. Chapman and Hall.*

DUGGAN, D. J., BITTNER, M., CHEN, Y., MELTZER, P. and TRENT, J. M. (1999): Expression profiling using cDNA microarrays. *Nature Genetics Supplement, 21:10-14.*

GEMAN, D., d'AVIGNON, C., NAIMAN, D. and WINSLOW, R. (2004): Classification Gene Expression Profiles from Pairwise mRNA Comparisons. *Statistical Applications in Genetics and Molecular Biology, 3 (1).*

GOLUB, T. R., SLOMIN, D. K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J. P., COLLIER, H., LOH, M. L. and et al (1999): Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science, 286, 531-537.*

HUANG, E., CHENG, S. H., DRESSMAN, H., PITTMAN, J., TSOU, M., HORNG, C., BILD, A., INVERSEN, E., LIAO, M. and CHEN, C. (2003): Gene expression predictors of breast cancer outcomes. *The Lancet, 361 (9369), 1590-1596.*

SINGH, D., FEBBO, P. G., JACKSON, D. G., MANOLA, J., LADD, C., TAMAYO, P., RENSHAW, A. A., D'AMICO, A. V., RICHIE, J. P., LANDER, E. S., LODA, M., KANTOFF, P. W., GOLUB, T. R. and SELLERS, W. R. (2002): Gene expression correlates of clinical prostate cancer behaviour. *Cancer Cell, 1(2):203-209.*

TIBSHIRANI, R., HASTIE, T., NARASIMHAN, B. and CHU, G. (2003): Class prediction by nearest shrunken centroids, with applications to DNA microarray. *Statistical Science, 18, 104-117.*

TURKMAN, K. F. and AMARAL TURKMAN, M. A. (1989): Optimal Screening Methods. *Journal of the Royal Statistical Society, B 51, 287-295.*