Detecting relevant gene structure through independent component analysis

Individuazione della struttura genica rilevante attraverso l'analisi in componenti indipendenti

Daniela G. Calò, Giuliano Galimberti, Marilena Pillati
Dipartimento di Scienze Statistiche
Università di Bologna
{calo,galimberti,pillati}@stat.unibo.it

Riassunto: Uno degli aspetti di maggior rilevanza nell'impiego di dati di espressione genica in problemi di classificazione è rappresentato dalla necessità di procedere preliminarmente ad una selezione dei predittori. In questo lavoro sono approfonditi alcuni aspetti della procedura proposta in Calò *et al.* (2003), basata su opportune trasformazioni lineari della matrice dei dati.

Keywords: Classification, Variable selection, Independent component analysis, Gene expression data.

1. Introduction

Among the numerous questions posed by the analysis of gene expression data, there is the problem of successfully distinguish different types of cells (for example, cells affected by different tumors). This is a classification problem, whose main feature is the very large number p of variables (the genes) relative to the number n of observations (the cells): the publicly available data sets currently contain gene expression data for thousands of genes on less than 100 cells.

As a matter of fact, in gene expression data many genes result to be not relevant to the discrimination problem. For this reason, a natural choice is to select the most discriminative genes, according to a measure of the association between the gene expression level and the variable denoting class membership. However, this univariate approach ignores gene interactions. On the other hand, traditional multivariate techniques for variable selection would lead to unstable results because of the particular condition $p \gg n$.

In Calò *et al.* (2003) an alternative approach to dimension reduction for this kind of data is proposed. Its starting point is the evidence (well known to the biological community) that the genes responsible for a given malignance show a behavior across the cells that differs from that of the other genes. Therefore, they suggest that a reasonable criterion for dimension reduction in such a diagnostic problem could consist in detecting and selecting these particular outlying genes. This implies that genes have to be considered as points in an n-dimensional space, thus the role of unit and that of variable are exchanged. In order to highlight these relevant genes, they employ linear transformations of the data, such as singular value decomposition and independent component analysis (ICA).

In this paper some aspects of the gene selection procedure proposed by Calò *et al.* (2003) are explored: section 2 deals with the chioce of the most adequate linear transformation method and section 3 addresses the problem of building a gene ranking w.r.t. outlyingness. Finally, in section 4 some applications to real data sets are presented.

2. The role of ICA in exploring gene expression data structure

When considering genes as units, the distribution of gene expression levels in the cells must be taken into account. Empirical evidence shows that these distributions are typically leptokurtic, with heavy tails and a pronounced peak in the middle. This implies that the observed variables (in this new perspective, the cells) have non-Gaussian distributions and hence the variance-covariance matrix does not suffice to describe the relations between them, but it is necessary to consider also higher-order moments. This is the fundamental reason why independent component analysis represents a useful tool in this context.

Independent component analysis is a recently developed linear transformation method, whose main feature is that it allows to minimize the statistical dependence of the projections. When statistical dependence is measured through mutual information, it is possible to prove that the less dependent projections are the most non-Gaussian ones (see Hyvärinen *et al.* (2001) for details). These directions should be able to catch the information enfolded in the moments of order higher than the second. Moreover, the ICA transformation seems to be the most suitable one for the purpose of detecting outlying genes. In fact, the "spiky" shape of the observed variable distributions should be emphasized in the distribution of gene scores along each component. In this way, it should be easier to distinguish between the bulk of the genes around the mode and the set of the interesting genes, lying in the tails of the distributions. Furthermore, the ICA methodology should allow to catch different aspects of the data structure, since the extracted components are not only uncorrelated but also as least statistically dependent as possible.

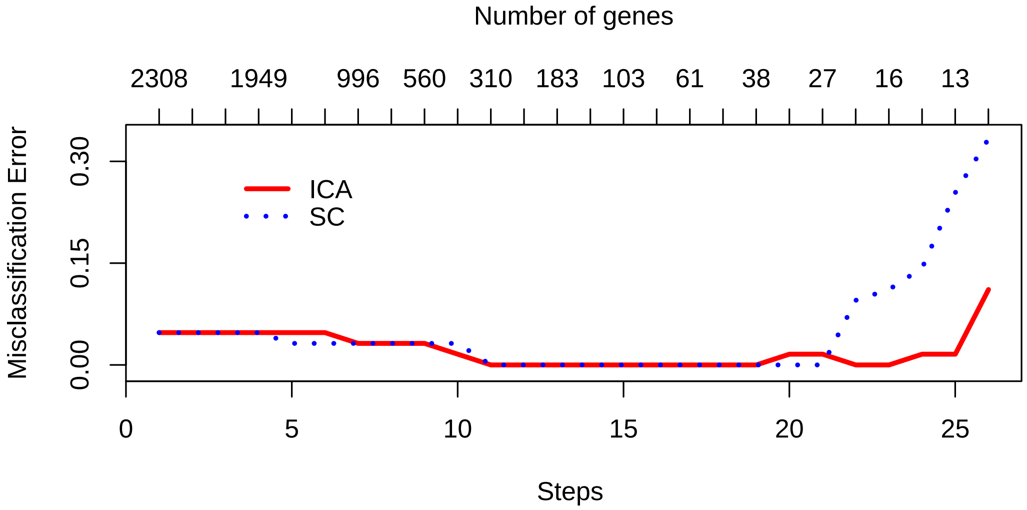
3. The problem of detecting the relevant genes

As already mentioned in section 1, the dimension reduction procedure we are dealing with is based on detecting and selecting the genes showing a behavior across the cells that differs most from that of the bulk of the genes. After k (with $k < n$) independent components with zero mean and unit variance have been extracted from the training set, in Calò *et al.* (2003) genes are ranked according to their maximum absolute score across the components. The genes located in the last m positions of this ranking (with $m \ll p$) should be used to build any classification rule. This is equivalent to rank the genes according to their distance from the mean vector in the space of the components in terms of the Minkowski metric with parameter $l \to \infty$. Therefore, as alternatives, we considered other distance measures, such as those obtained from the Minkowski metric for $l = 1$ (Manhattan distance), $l = 2$ (Euclidean distance) and $l = 3$ (cubic distance), in order to evaluate the sensitivity of the procedure to the choice of the measure. Given the distance measure, it could be also interesting to study the robustness of the gene ranking with respect to slightly perturbations of the training set. For example, this could be addressed by comparing the ranking associated to the whole data set with those obtained at each step of a cross-validation procedure.

4. Some applications on real data sets

The proposed strategy is applied to two data sets: the small round blue cell tumor data set of Khan *et al.* (2001) and the leukemia data set of Golub *et al.* (1999). The performances are compared with those obtained by the nearest shrunken centroid (SC) method (Tibshi-

Figure 1: *Small round blue cell tumor data set: cross-validated misclassification rates. The axis at the top of the plot indicates the number of genes retained at each step.*

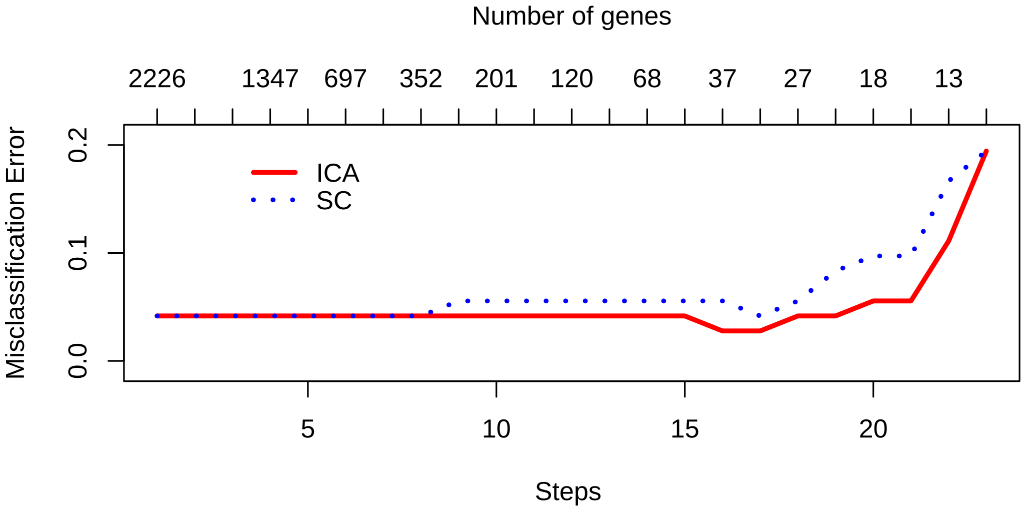


rani *et al.* (2002)). In order to allow a fair comparison, data are standardized by the within standard deviation and the nearest centroid method is used to build the classification rules based on the subsets of genes selected according to our proposal. Our procedure is implemented in `R` code, resorting to the libraries `pamr` and `fastICA` to perform nearest shrunken centroid classification and independent component analysis. In `fastICA`, the option `exp` is used to approximate neg-entropy as a measure of non-gaussianity. The number of components k is selected so that the smallest estimated error rate with the smallest number of genes is achieved. Given the small number of cells, classification error rates are estimated by balanced cross-validation. The results shown below were obtained using the maximum absolute score across the components as distance measure. In fact, the comparison among the different choices of the metrics mentioned in section 3 revealed a substantial stability of the classification performances. This is due to the strong agreement between the last positions of the rankings associated to the considered metrics. A possible explanation is that genes with high absolute scores along one component are likely to have low absolute scores along the other ones, since the components extracted through ICA are uncorrelated (by construction) and leptokurtic (due to the typical shape of the observed variable distributions). Hence, the choice of the distance measure seems not to be crucial.

Khan data set contains gene expression levels for $p = 2038$ genes in 63 cells and consists of 4 classes of $n_1 = 8$, $n_2 = 23$, $n_3 = 12$ and $n_4 = 20$ cases, respectively. Figure 1 displays the results obtained with $k = 6$ components: the ICA-based selection procedure gives better performances than those achievable by the SC method, based on marginal gene selection. In particular, both the methods are able to predict the classes without errors, but the one based on ICA achieves this result with a lower number of genes ($m = 16$ for ICA against $m = 33$ for SC method). In this data set it is evident that the use of suitable subsets of genes instead of the whole set yields better classification performances: the estimated misclassification rate for $m = 2308$ is equal to 0.048.

The leukemia data set contains gene expression levels for $p = 6817$ genes in 72 cells and consists of 3 classes of $n_1 = 38$, $n_2 = 9$ and $n_3 = 25$ cases, respectively. According to Dudoit *et al.* (2002) , three preprocessing steps were applied: (a) thresholding, (b) filtering and (c) base 10 logarithmic transformation. Step (b) has been slightly strengthen

Figure 2: *Leukemia data set: cross-validated misclassification rates. The axis at the top of the plot indicates the number of genes retained at each step.*



in order to make stricter the exclusion criterion for genes with low variability across the cells, by using for each gene the 90^{th} and the 10^{th} percentile instead of its maximum and minimum values respectively. The number of genes retained is 2226. Both the methods are not successful in accurately predict the class membership (Figure 2). However, our strategy allows to achieve a smaller minimum error rate (0.028 for $m = 29$ and $k = 7$) than that of the SC method (0.042 for the same number of genes).

In conclusion, also the results obtained in these two data sets seem to confirm the usefulness of the ICA-based strategy as a tool to perform dimension reduction in supervised cell classification based on gene expression data. However, some aspects deserve further research; in particular, given the noisy nature of the data, it could be interesting to evaluate also the potentialities of the noisy ICA model.

References

Calò D.G., Galimberti G., Pillati M. and Viroli C. (2003) Variable selection in classification problems: a strategy based on independent component analysis, *CLADAG 2003 Book of Short Papers*, 87–90.

Dudoit S., Fridlyand J. and Speed T.P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association*, 457, 77–87.

Golub T.R., Slonim D.K. and Tamayo P. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, 286, 531–537.

Hyvärinen A., Karhunen J. and Oja E. (2001) *Independent Component Analysis*, Wiley, New York.

Khan J., Wei J. and Ringner M. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nature Medicine*, 7, 673–679.

Tibshirani R., Hastie T., Narasimhan B. and Chu G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression, *Proceedings of the National Accademy of Sciences*, 99, 6567–6572.