Variable selection in cell classification problems: a strategy based on independent component analysis

Daniela G. Calò, Giuliano Galimberti, Marilena Pillati and Cinzia Viroli

Dipartimento di Scienze Statistiche,
Università di Bologna, Italy
{calo,galimberti,pillati,viroli}@stat.unibo.it

Abstract. In this paper the problem of cell classification using gene expression data is addressed. One of the main features of this kind of data is the very large number of variables (genes), relative to the number of observations (cells). This condition makes most of the standard statistical methods for classification difficult to employ. The proposed solution consists of building classification rules on subsets of genes showing a behavior across the cells that differs most from that of all the other ones. This variable selection procedure is based on suitable linear transformations of the observed data: a strategy resorting to independent component analysis is explored. Our proposal is compared with the nearest shrunken centroid method (Tibshirani *et al.* (2002)) on three publicly available data sets.

1 Introduction

The recent advances in biotechnology have yielded an ever increasing interest in genome research. The novel cDNA microarray technology allows for the monitoring of thousands of genes simultaneously and it is being currently applied in cancer research. The data from such experiments are usually in the form of large matrices of expression levels of p genes under n experimental conditions (different times, cells, tissues . . .), where n is usually less than 100 and p can easily be several thousands. Due to the large number of genes and to the complex relations between them, a reduction in dimensionality and redundancy is needed in order to allow for a biological interpretation of the results and for subsequent information processing.

In this paper the problem of supervised classification of cells is addressed. The particular condition $p \gg n$ makes most of the standard statistical methods difficult to employ from both analytical and interpretative points of view. For example, including too many variables may increase the error rate in classifying units outside the training set and make the classification rules difficult to interpret. The inclusion of irrelevant or noisy variables may also degrade the overall performances of the estimated classification rules. There is a vast literature on gene selection for cell classification; a comparative study of several discrimination methods in the context of cancer classification based on filtered sets of genes can be found in Dudoit *et al.* (2000). Variable selection in this context has also biological foundations: most of the abnormalities in cell behavior are due to irregular gene activities. It is then important to employ tools that allow to highlight these particular genes.

The proposed solution consists of building classification rules on genes selected by looking at the tails of the distributions of gene projections along suitable directions. Since gene expression profiles are typically non-gaussian, it seems relevant to catch not only the linear (second-order) aspects of the data structure but also the non-linear (higher-order) ones. For this reason, our proposal focuses on searching the less statistically dependent projections. These directions are obtained by independent component analysis (Hyvärinen *et al.* (2001)).

2 Independent component analysis

Independent component analysis is a recently developed method originally proposed in the field of signal processing, as a solution to the so called "blind source separation" problem. In this context the purpose is to recover some independent sources by the observation of different signals, that are assumed to be linear mixtures of these unknown sources.

Subsequently this method has been applied to image analysis, time series analysis and gene expression data analysis. In this latter context, much emphasis has been posed on the ability of ICA in finding so-called functional genomic units, each of which contains genes that work together to accomplish a certain biological function.

Denote by $x_1, x_2, ...x_m$ the m observed variables which are supposed to be modelled as linear combinations of k latent variables $s_1, s_2, ..., s_k$:

$$x_i = a_{i1}s_1 + a_{i2}s_2 + ... + a_{im}s_k \quad \text{for all} \quad i = 1, ..., m \tag{1}$$

where the a_{ij} $(j = 1, ..., k)$ are real coefficients. The s_j are assumed to be *mutually statistically independent.*

The ICA transformation can be put in the following compact notation:

$$\mathbf{X} = \mathbf{AS}. \tag{2}$$

Since it describes how the observed data are generated by a mixing process of hidden components, the matrix $\mathbf{A}$ is often called *mixing matrix.* The only requirement on $\mathbf{A}$ is that it is a full column rank matrix. However, it is easy to verify that if the data are supposed to be sphered, the mixing matrix must be an orthogonal one.

The estimation of the independent latent sources is performed by searching for a linear transformation of the observed variables

$$\hat{\mathbf{S}} = \mathbf{WX}. \tag{3}$$

such that the mutual statistical dependence between the estimated sources is minimized.

The statistical dependence between k variables $s_1, \ldots, s_k$ can be quantified by the mutual information $I(s_1, \ldots, s_k)$. Restricting the attention to sphered data, minimizing $I(s_1, \ldots, s_k)$ is equivalent to maximizing the sum of the marginal negentropies $J(s_j)$ since:

$$I(s_1, \ldots, s_k) = J(s_1, \ldots, s_k) - \sum_{j=1}^{k} J(s_j), \tag{4}$$

and the joint negentropy $J(s_1, \ldots, s_k)$ is a constant (see Hyvärinen *et al.* (2001) for details). As negentropy is the Kullback-Leibler divergence between a probability density function and a gaussian one with the same covariance matrix, the less dependent directions are the most non-gaussian ones. This implies that it is not sufficient to take into account the information in the covariance matrix, but it is necessary to consider also higher-order relations. For this reason, ICA allows to explore not only the linear structure of the data, but also the non-linear one. It should be stressed that the core assumption of ICA is the existence of mutual independent components. However, it is interesting to note that even if this assumption does not hold, the method can be interpreted as a particular projection pursuit solution.

3 Gene selection in cell classification: a solution based on ICA

As already mentioned above, the aim of this paper is to propose a method to select subsets of genes that could be relevant for cell classification. This selection is performed by projecting the genes onto the directions obtained by ICA: thus, the p genes are considered as units and the n cells as variables. In practice, any other linear transformation method, such as singular value decomposition (SVD) (Wall *et al.* (2003)), could be employed. The use of ICA is consistent with the fact that gene expression profiles usually exhibit non-gaussianity. In particular, the distribution of gene expression levels on a cell is "approximately sparse", with heavy tails and a pronounced peak in the middle. Due to this particular feature, the projections obtained by ICA should emphasize this sparseness. Highly induced or repressed genes, that may be useful in cell classification should lie on the tails of the distributions of $\hat{s}_j$ $(j = 1, \ldots, k)$. Since these directions are as less dependent as possible, they may catch different aspects of the data structure that could be useful for classification tasks.

The proposed solution is based on a ranking of the p genes. This ranking is obtained as follows:

- k independent components $\hat{s}_1, \ldots, \hat{s}_k$ with zero mean and unit variance are extracted from the training set;

- for gene l ($l = 1, \ldots, p$), the absolute score on each component $|\hat{s}_{lj}|$ is computed. These k scores are synthesized by retaining the maximum one, denoted by g_l: $g_l = \max_j |\hat{s}_{lj}|$;
- the p genes are sorted in increasing order according to the maximum absolute scores $\{h_1, \ldots, h_p\}$ and for each gene the rank $r(l)$ is computed.

We suggest to use the subset of genes located in the last m positions of this ranking (with $m \ll p$) to build any classification rule, that is $\{l : r(l) \geq m\}$. The rationale behind this is that these m genes show, with respect to at least one of the components, a behavior across the cells that differs most from that of the bulk of the genes.

The proposed strategy relies on the choice of suitable values for both the number of components and the number of genes. If the goal is to build a classification rule on a manageable set of genes that accurately classifies the cells, a plausible criterion to select the optimal number of components consists in selecting the value of k which yields the smallest estimated error rate with the smallest number of genes (in this way the value of m is chosen implicitly). In practice, this criterion may be implemented by considering several values for the number k of components. For each value of k the ranking is computed and a sequence of classification rules is built for several values of m (with $m \ll p$). For each of these classification rules the error rate is estimated, and the minimum is determined. Finally, the value of k is chosen such that it achieves this minimum rate with the smallest number of genes. (in case that more than one value is selected, the smallest one is obviously preferred).

4 Applications to real data sets

In this section the proposed strategy is applied to three publicly available data sets: the lymphoma data set of Alizadeth *et al.* (2000), the small round blue cell tumor data set of Khan *et al.* (2001) and the leukemia data set of Golub *et al.* (1999).

We run our gene selection procedure (both ICA and SVD based ones) for k ranging from 1 to 10. For each value of k, we tried 30 different values of the number m of selected genes, ranging from p to 1.

The performances of classification rules based on subsets of genes selected according to our proposal are compared with those obtained by the nearest shrunken centroid (SC) method (Tibshirani *et al.* (2002)). This method is based on an enhancement of the nearest centroid classifier and its main feature is that the class centroids are shrunken toward the overall centroid in order to reduce the effect of noisy genes. Classification is made to the nearest shrunken centroid. This shrinkage procedure performs automatic gene selection. In particular, if a gene is shrunken to zero for all classes, then it is dropped from the prediction rule. In order to compare the results of our gene selection procedure with those obtained through the shrunken centroids, the nearest centroid method is used in class prediction, but any other postprocessing classifier could be applied.

We implemented our procedure in `R` code, resorting to the libraries `pamr` and `fastICA` to perform nearest shrunken centroid classification and independent component analysis, respectively.

Given the small number of cells in each data set, the classification error rates are estimated by balanced cross-validation for each of the compared procedures. However, when comparing the estimated error rate curves, the following difference should be taken into account. In our procedure (both the ICA and the SVD version) each cross-validation training set is used to extract k components, according to which the sequence of nested gene subsets of given sizes is created (therefore, this sequence may vary from one training set to another); the cells in the corresponding cross-validation test set are finally classified on the basis of these subsets of variables. Differently, in `pamr` the sequence of nested gene subsets is unique, being based on the whole training set, and each cross-validation training set differs from the others only with respect to the class centroids. Therefore, the variability in gene ranking due to training set perturbations is not taken into account when evaluating the SC method performances.

4.1 Lymphoma data set

The data set contains gene expression levels for p =4026 genes in 62 cells and consists of 11 cases of B-cell chronic lymphocytic leukemia (B-CLL), 9 cases of follicular lymphoma (FL) and 42 cases of diffuse large B-cell lymphoma (DLBCL). The gene expression data are summarized by a 4026×62 matrix. Missing data were imputed by a 15 nearest-neighbors algorithm.

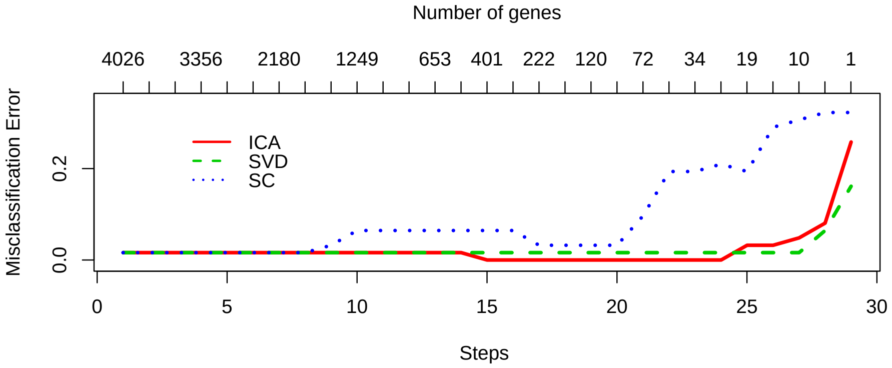

Fig. 1. Lymphoma data set: cross-validated misclassification rates. The axis at the top of the plot indicates the number of genes retained at each step.

	4026	*1817*	*1522*	*1249*	*519*	*401*	*72*	*26*	*19*	*10*	*5*
ICA	0.016	0.016	0.016	0.016	0.016	0.000	0.000	**0.000**	0.032	0.048	0.081
SVD	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016	**0.016**	0.065
SC	0.016	**0.016**	0.032	0.065	0.065	0.065	0.097	0.210	0.194	0.306	0.323

Table 1. Lymphoma data set: cross-validated misclassification rates for different values of m ($k = 7$ for ICA, $k = 2$ for SVD).

Figure 1 displays the results obtained with k=7 components for ICA and for k=2 components for SVD. The graph shows that gene selection by suitable projections gives better performances than those achievable by the nearest shrunken centroid method, which is based on marginal gene selection. Moreover, the ICA-based procedure performs better than the SVD one, since it allows to achieve a zero cross-validated error rate by reducing the number of genes from 4026 to just 26. For this number of genes the shrunken centroids estimated error rate is dramatically higher (0.210, as shown in Table 1).

In order to understand the reason why the SC method is outperformed, we focused our attention on the last 5 genes surviving the elimination procedure. As far as the SC method is concerned (Figure 2), it seems that this procedure may not always be able to identify genes that discriminate between all of the classes; it also tends to select genes that are highly correlated (correlations between these genes range between 0.70 and 0.98). On the other hand, the ICA based solution in this case selects genes that make the class structure more evident (Figure 3). It is interesting to remind that the information about class membership is not taken into account in extracting the components (and hence in building the gene ranking).

4.2 Small round blue cell tumor data set

The data set contains gene expression levels for p =2038 genes in 63 cells and consists of 8 cases of Burkitt lymphoma, 23 cases of Ewing sarcoma, 12 cases of neuroblastoma and 20 cases of rhabdomyosarcoma.

	2308	*761*	*423*	*310*	*38*	*33*	*21*	*16*	*15*	*13*	*9*
ICA	0.048	0.032	0.016	0.000	0.000	0.016	0.000	**0.000**	0.016	0.016	0.111
SVD	0.048	0.032	0.016	0.016	0.000	0.000	0.000	0.000	**0.000**	0.016	0.032
SC	0.048	0.032	0.032	0.000	0.000	**0.000**	0.095	0.111	0.143	0.254	0.333

Table 2. Small round blue cell tumor data set: cross-validated misclassification rates for different values of m ($k = 6$ for both ICA and SVD).

As Table 2 shows, all the three methods are able to accurately predict the classes, but the ones based on ICA and SVD achieve this result with a lower number of genes (16 and 15 respectively, against 33 for SC method).

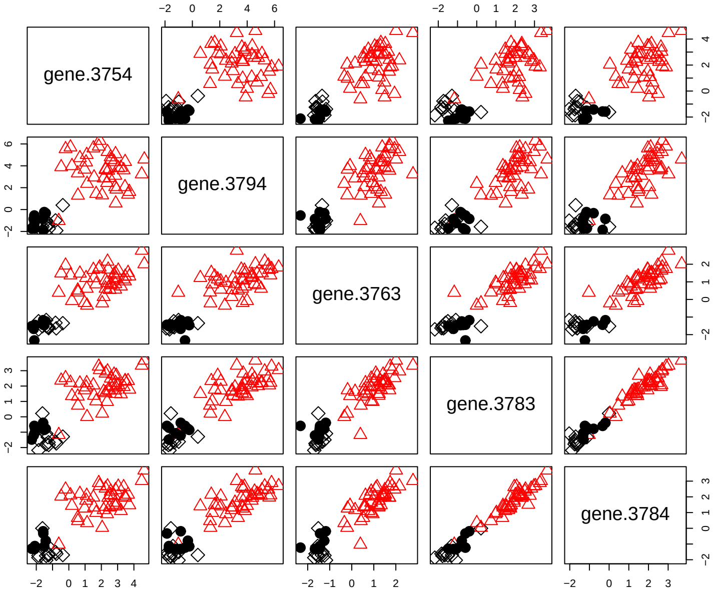

Fig. 2. Lymphoma data set: scatter plot matrix of the last 5 genes surviving the shrinkage procedure (△=DLCL, ●=FL, ◇=CLL).

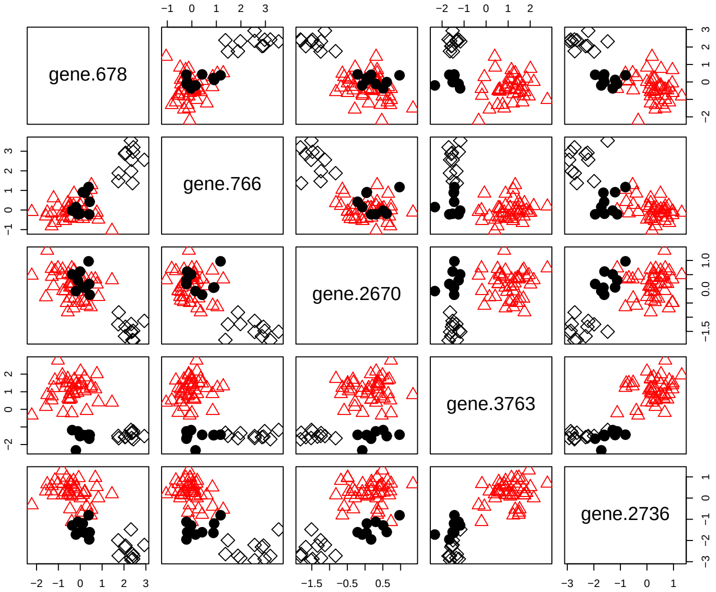

Fig. 3. Lymphoma data set: scatter plot matrix of the last 5 genes of the ranking obtained by ICA (△=DLCL, ●=FL, ◇=CLL).

These optimal solutions have 7 genes in common. With this data set it is particularly evident that the use of suitable subsets of genes instead of the whole set yields better classification performances.

4.3 Leukemia data set

The data set contains gene expression levels for p =6817 genes in 72 cells and consists of 38 cases of B-cell acute lymphoblastic leukemia, 9 cases of T-cell acute lymphoblastic leukemia and 25 cases of acute myeloid leukemia. 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 in order to make stricter the exclusion criterion for genes with low variability across the cells, by using for each gene the 90^{th} percentile and the 10^{th} percentile instead of its maximum and minimum values respectively. The number of genes retained for the analysis is 2226. As shown in Table 3, none of the three methods is successful in accurately predict the class membership. However, our strategy allows to achieve a smaller minimum error rate (0.028) than that of the SC method (0.042). It is worth noting that these minimum values are referred to approximately the same number of selected genes.

	2226	*262*	*37*	*29*	*27*	*20*	*17*	*13*	*10*	*6*	*3*
ICA	0.042	0.042	0.028	**0.028**	0.042	0.042	0.056	0.111	0.194	0.333	0.528
SVD	0.042	0.042	0.042	0.042	**0.028**	0.056	0.083	0.153	0.153	0.306	0.403
SC	0.042	0.056	0.056	**0.042**	0.056	0.083	0.097	0.167	0.194	0.194	0.333

Table 3. Leukemia data set: cross-validated misclassification rates for different values of m ($k = 7$ for ICA and $k = 5$ for SVD).

5 Conclusions and open issues

As the preliminary results on these real data sets show, the proposed strategy seems to represent a useful tool to detect subsets of relevant genes for supervised cell classification based on microarray data. However, some aspects deserve further research.

For example, some alternatives to the proposed criterion for building the ranking could be investigated.

Firstly, in the proposed strategy all the k estimated components are assumed to be equally important, since the definition of ICA implies no ordering of the independent components. It is possible, however, to introduce an order among them: Hyvärinen *et al.* (2001) suggest as ordering criteria the norm of the columns of the mixing matrix or the value of suitable non-gaussianity measures on the estimated components. These criteria could be adopted to weight each component during the construction of the gene ranking (for example, by increasing the importance of the most non-gaussian ones).

Secondly, it can be noted that g_l is equivalent to the distance of gene l from the (zero) mean vector in the space of the k components in terms of the Minkowski metric

$$g_l = \left\{ \sum_{j=1}^{k} |\hat{s}_{lj}|^{\lambda} \right\}^{1/\lambda} \tag{5}$$

with $\lambda \rightarrow \infty$. It could be interesting to evaluate the sensitivity of the procedure and the robustness of the gene ranking to the choice of different values for λ or of different distance measures.

Moreover, the issues concerning the choice of both the number k of the components and the number m of retained genes should be examined in more depth.

Finally, the interaction between the proposed selection method and other classifiers could be explored.

References

ALIZADEH, A.A., EISEN, M.B., DAVIS, R.E. *et al.* (2000): Distinct Types of Diffuse Large B-cell Lymphoma Identified by Gene Expression Profiling. *Nature*, 403, 503-511.

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., TAMAYO, P. *et al.* (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., RINGNER, M. *et al.* (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.

VIROLI, C. (2003): Reflections on a Supervised Approach to Independent Component Analysis, *Between Data Science and Applied Data Analysis*, (M. Schader, W. Gaul e M. Vichi eds.), Studies in Classification, Data Analysis, and Knowledge Organization, Springer Berlin, 501-509.

WALL, M.E., RECHTSTEINER, A. and ROCHA, L.M. (2003): Singular Value Decomposition and Principal Component Analysis, in: *A Practical Approach to Microarray Data Analysis*, Berrar D.P., Dubitzky W. and Granzow M. (Eds.), Kluwer, Norwell, 91-109.