Unsupervised Mining of Genes Classifying Leukemia

Diego Liberati
Politecnico di Milano, Italy

Sergio Bittanti
Politecnico di Milano, Italy

Simone Garatti
Politecnico di Milano, Italy

INTRODUCTION

Micro-arrays technology has marked a substantial improvement in making available a huge amount of data about gene expression in pathophysiological conditions; among the many papers and books recently devoted to the topic, see, for instance, Hardimann (2003) for a discussion on such a tool.

The availability of so many data attracted the attention of the scientific community on how to extract significant and directly understandable information in an easy and fast automatic way from such a big quantity of measurements. Many papers and books have been devoted as well to various ways to process micro-arrays data; Knudsen (2004) is a recent re-edition of a book pointing to some of the approaches of interest to the topic.

When such opportunity to have many measurements on several subjects arises, one of the typical goals one has in mind is to classify subjects on the basis of a hopefully reduced meaningful subset of the measured variables. The complexity of the problem makes it worthwhile to resort to automatic classification procedures. A quite general data-mining approach that proved to be useful also in this context is described elsewhere in this article (Liberati, 2004), where different techniques also are recalled as references, and where a clustering approach to piece-wise affine model identification also is reported. In this contribution, we will resort to a different recently developed unsupervised clustering approach, the PDDP algorithm, proposed in Boley (1998). According to the analysis provided in Savaresi & Boley (2004), PDDP is able to provide a significant improvement of the performances of a classical *k-means* approach (Hand et al., 2001; MacQueen, 1967), when PDDP is used to initialize the *k-means* clustering procedure. Such cascading of PDDP and *k-means* was, in fact, already successfully applied in a totally different context for analyzing the data regarding a large virtual community of Internet users (Garatti et al., 2004).

The approach taken herein may be summarized in the following four steps, the third of which is the core of the method, while the first two constitute a preprocessing phase useful to ease the following task, and the fourth one a post-processing designed to focus back on the original variables. This approach was found to be meaningful after the transforms operated in the previous steps:

1. A first pruning of genes not likely to be significant, for the final classification is performed on the basis of their small intersubject variance, thus reducing the size of the subsequently faced problem.
2. A principal component analysis defines a hierarchy in the remaining transformed orthogonal variables.
3. Finally, the clustering is obtained by means of the cascade of the principal direction divisive partitioning and the bisecting K-means algorithms. The classification is achieved without using a priori information on the patient's pathology (unsupervised learning). This approach presents the advantage that it automatically highlights the (possibly unknown) patient casuistry.
4. By analyzing the obtained results, the number of genes for the detection of pathologies is further reduced, so that the classification eventually is based on a few genes only.

The application of such classification procedure is quite general, even beyond micro-arrays data; many problems resemble this one for statistical structure, like prognostic factor in oncology or drug discovery, as described in Liberati (2004) but also, for instance, for risk management in finance in an apparently totally different framework.

Here, results will be shown in the paradigmatic case of automatically classifying two kinds of leukemia in a few patients whose thousands of gene expressions are publicly available on the Internet (Golub et al., 1999). Our approach seems to present some advantages with respect

to the one originally obtained by Golub, et al. (1999), with a different approach in that our classification eventually is based on a very limited number of genes without any type of a priori information. This encouraging result, together with the ones in Garatti et al. (2004) and with the theoretical considerations in Savaresi and Booley (2004) suggests that the methodology proposed in the present contribution, besides providing significant results in the presented example, is likely to be of help in (and beyond) the bioinformatics context.

BACKGROUND

Among the problems to which a bioinformatics approach to micro-arrays data is required, the classification problems are of paramount interest, as in almost every context in which one would resort to data mining; it often is needed to be able to discriminate among two (or more) classes of subjects on the basis of a small number of the many available measured variables.

For classification, a basic tool is provided by clustering procedures, which are the subject of many papers (Jain et al., 1999) and books (Duda & Hart, 1973; Hand et al., 2001; Jain & Dubes, 1998; Kaufman & Rousseeuw, 1990). As is well known, one can distinguish unsupervised procedures and supervised procedures; the former perform the classification on the sole basis of the intrinsic characteristics of the data by means of a suitable notion of distance; the latter makes use of additional information on the data classification available a priori. For applications illustrative of these two approaches, the interested reader is referred to Karayiannis and Bezdek (1997), Setnes (2000), Muselli and Liberati (2002), Ferrari-Trecate, et al. (2003), and Muselli and Liberati (2000).

The leukemia dataset, chosen as a paradigmatic example to illustrate the classification performances of the algorithm proposed here, is often used as a test bed in bioinformatics. For example, it was treated in Golub, et al. (1999) by resorting to a supervised approach and in De Moor, et al. (2003) by the *k-means* technique alone; in the last paper, no final results are available in order to make a direct comparison; this may be due to the fact that *k-means* alone is sensitive to initialization, while our preprocessing via PDDP provides unique initialization to *k-means*, as shown in Savaresi and Boley (2004), where it is also discussed that the cascade of the two algorithms outperforms each one alone.

MAIN THRUST

Our four-step data analysis can be outlined as follows:

1. **Variance Analysis:** The variance of the expression value is computed for each gene across the patients in order to have a first indicator of the relative intersubject expression variability and to reject those genes whose variability is below a defined threshold. The idea behind this is that if the variability of a gene expression over the subjects is small, then that gene does not detect any variability and, hence, is not useful for classification.
2. **Principal Component Analysis:** Principal Component Analysis (O'Connel, 1974; Hand et al., 2001) is a multivariate analysis designed to select the linear combinations of variables with higher intersubject covariances; such combinations are the most useful for classification. More precisely, PCA returns a new set of orthogonal coordinates of the data space,

Table 1. PDDP clustering algorithm

Step 1. Compute the centroid w of S.

Step 2. Compute an auxiliary matrix $\tilde{S}$ as:

$$\tilde{S} = S - ew,$$

where e is the N-dimensional vector of ones (i.e., $e = [1,1,1,1,1,...1]^T$).

Step 3. Compute the Singular Value Decompositions (SVD) of $\tilde{S}$:

$$\tilde{S} = U\Sigma V^T,$$

where Σ is a diagonal $N \times p$ matrix, and U and V are orthonormal unitary square matrices whose dimensions are $N \times N$ and $p \times p$, respectively (Golub & van Loan, 1996).

Step 4. Take the first column vector of V (i.e., $v = V_1$), and divide $S = [x_1, x_2, \ldots, x_N]^T$ into two subclusters, S_L and S_R, according to the following rule:

$$\begin{cases} x_i \in S_L & if \quad v^T(x_i - w) \le 0 \\ x_i \in S_R & if \quad v^T(x_i - w) > 0 \end{cases}$$

Table 2. Bisecting K-means algorithm

Step 1. (*Initialization*). Select two points in the data domain space (i.e., $c_L, c_R \in \Re^p$).

Step 2. Divide $S = [x_1, x_2, \ldots, x_N]^T$ into two subclusters, S_L and S_R, according to the following rule:

$$\begin{cases} x_i \in S_L & \text{if} \quad \|x_i - c_L\| \leq \|x_i - c_R\| \\ x_i \in S_R & \text{if} \quad \|x_i - c_L\| > \|x_i - c_R\| \end{cases}$$

Step 3. Compute the centroids w_L and w_R of S_L and S_R,.

Step 4. If $w_L = c_L$ and $w_R = c_R$, stop. Otherwise, let $c_L := w_L$, $c_R := w_R$ and go back to Step 2.

where such coordinates are ordered in decreasing order of intersubject covariance.

3. **Clustering:** Unsupervised clustering is performed via the cascade of a non-iterative technique—the Principal Direction Divisive Partitioning (PDDP) (Booley, 1998*)* based upon singular value decomposition (Golub & van Loan, 1996) and the iterative centroid-based divisive algorithm *K-means* (Mac Queen, 1967). Such a cascade, with the clusters obtained via PDDP used to initialize K-means centroids, is shown to achieve best performances in terms of both quality of the partition and computational effort (Savaresi & Boley, 2004). The whole dataset thus is bisected into two clusters, with the objective of maximizing the distance between the two clusters and, at the same time, minimizing the distance among the data points lying in the same clusters. These two algorithms are recalled in Tables 1 and 2. In both tables, the input is a $N \times p$ matrix *S*, where the data for each subject are the rows of the matrix, and the outputs are the two matrices S_L and S_R each one representing a cluster. Both algorithms are based on the following quantity:

   $w = \frac{1}{N}\sum_{i=1}^{N} x_i$, where x_i's are the rows of *S*,

   and where *w* is the average of data samples and is called the centroid of *S*.

4. **Gene Pruning:** The previous procedure is complemented with an effective gene-pruning technique in order to detect a few genes responsible for each pathology. In fact, from each identified principal component, many genes may be involved. Only the one(s) influencing each selected principal component more are kept.

A Paradigmatic Example

The Leukemia Classification: Data are taken from a public repository often adopted as a reference benchmark (Golub et al., 1999) in order to test new classification techniques and compare the various methodology to each other. Such database is constituted by gene expression data over 72 subjects, relying on 7,129 genes. Of the 72 subjects, 47 are cases of acute lymphoblastic leukemia (ALL), while the remaining 25 are cases of acute myeloid leukemia (AML).

An experimental bottleneck in this kind of experiment is the difficulty in collecting a high number of homogeneous subjects in each class of interest, making the classification problem even harder; not only a big matrix is involved, but such matrix has a huge number of variables (7,129 genes) with only a very poor number of samples (72 subjects). The cutoff on lower inter-subject gene variance thus is implemented in order to limit the number of genes in the subsequent procedure.

The result of the variance analysis for the 7,129 genes shows that the variance is small for thousands of genes. Having selected a suitable threshold, 6,591 genes were pruned from the very beginning. So, attention has been focused on 178 genes only. Of course, the choice of the cutoff level is a tuning parameter of the algorithm. The adopted level may be decided on the basis of a combination of biological considerations, if it is known under which level the variance should be considered of little significance; technological knowledge, when assessing how confident the micro-arrays measurements can be; empirical considerations, by imposing either a maximum number of residual variables or a minimum fraction of variance with respect to the maximum one.

Then, the remaining phases of the outlined procedure have been applied. In this way, the set of 72 subjects has been subdivided into two subsets containing 23 and 49

Table 3. The seven genes discriminating AML from ALL

1.	FTL Ferritin, light polypeptide	M11147_at
2.	MPO Myeloperoxidase	M19507_at
3.	CST3 Cystatin C	M27892_at
4.	Azurocidin gene	M96326_ rna1_at
5.	GPX1 Glutathione peroxidase 1	Y00433_at
6.	INTERLEUKIN-8 PRECURSOR	Y00787_s_at
7.	VIM Vimentin	Z19554_s_at

patients, respectively. As already said, this portioning has been obtained without exploiting a priori information on the pathology of the patients (i.e., ALL or AML).

Interestingly, all 23 subjects of the smaller cluster turn out to be affected by the AML pathology. Thus, the only error of our unsupervised procedure consists in the misclassification of two AML patients, erroneously grouped in the bigger cluster, together with the remaining 47 subjects affected by the ALL pathology. Thus, the misclassification percentage is 2/72=8%.

In addition, it should be pointed out that the final gene-pruning step leads to a very small number of significant genes; namely, only seven genes, as listed in Table 3.

Our results outperform the original one of Golub, et al. (1999), using a supervised tool, and, thus, splitting the 72 patients into 38 training samples and 34 testing samples, they correctly classified 29 (about 85%) of the 34 test subjects with as much as 50 genes, three of which (Cystatin C, Azurocidin, and Interleukin-8 precursor) also are among the seven sufficient in our approach. A possible interpretation is that the three genes within the intersection of the two subsets probably are really determinant, while the complementing four genes identified by the procedure proposed in this article discriminate better than the complementing 43 in the subset of Golub, et al. (1999).

FUTURE TRENDS

The proposed approach is now under application in other similar contexts. The fact that a combination of different approaches, taken from partially complementary disciplines, proves to be effective may indicate a fruitful direction in combining in different ways classical and new approaches to improve classification.

CONCLUSION

The proposed clustering algorithm is effective for the discrimination of the two kinds of leukemia of the considered dataset on the basis of an extremely limited number of genes. The unsupervised nature of the presented approach enables the classification without any knowledge on the pathologies of the patients. Also, it does not require the subdivision of the data into a training set and a testing set. The proposed approach is very general and is not limited to the bioinformatics field. For instance, it already was used successfully for analyzing the data regarding a large virtual community of Internet users (Garatti et al., 2004).

ACKNOWLEDGEMENTS

This article was supported by CNR-IEIIT. The authors would also like to thank Andrea Maffezzoli, who was in charge of the computation in fulfillment of his master's thesis at Milan Institute of Technology. Three anonymous reviewers are also gratefully acknowledged for indicating how to improve the writing.

REFERENCES

Boley, D.L. (1998). Principal direction divisive partitioning. *Data Mining and Knowledge Discovery, 2*(4), 325-344.

De Moor, B., Marchal, K., Mathys, J., & Moreau, Y. (2003). Bioinformatics: Organism from Venus, technology from Jupiter, algorithms from Mars. *European Journal of Control, 9*(2-3).

Duda, R.O., & Hart, P.E. (1973). *Pattern classification and scene analysis*. New York: Wiley.

Ferrari-Trecate, G., Muselli, M., Liberati, D., Morari, M. (2003). A clustering technique for the identification of piecewise affine systems. *Automatica, 39*, 205-217.

Garatti, S., Savaresi, S., & Bittanti, S. (2004). On the relationships between user profiles and navigation ses-

sions in virtual communities: A data-mining approach. *Intelligent Data Analysis.*

Golub, G.H., & van Loan, C.F. (1996). *Matrix computations*. Johns Hopkins University Press.

Golub, T.R., et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression. *Science, 286*, 531-537.

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data-mining*. Cambridge, MA: MIT Press.

Hardimann, G. (2003). Microarrays methods and applications: Nuts & bolts. DNA Press.

Jain, A., & Dubes, R. (1998). *Algorithms for clustering data*. London: Sage Publications.

Jain, A.K., Murty, M.N., & Flynn, P.J. (1999). Data clustering: A review. *ACM Computing Surveys*, *31*, 264-323.

Karayiannis, N.B,. & Bezdek, J.C. (1997). An integrated approach to fuzzy learning vector quantization and fuzzy C-means clustering. *IEEE Trans. Fuzzy Systems*, *5*, 622-628.

Kaufman, L., & Rousseeuw, P.J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: Wiley.

Knudsen, S. (2004). Guide to analysis of DNA microarray data. John Wiley & Sons.

Liberati, D. (2004). Data mining for model identification.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceeedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, California.

Muselli, M., & Liberati, D. (2000). Training digital circuits with Hamming clustering. *IEEE Transactions on Circuits and Systems—I: Fundamental Theory and Applications, 47,* 513-527.

Muselli, M., & Liberati, D. (2002). Binary rule generation via Hamming clustering. *IEEE Transactions on Knowledge and Data Engineering, 14***,** 1258-1268.

O'Connel, M.J. (1974). Search program for significant variables. *Comp. Phys. Comm., 8***,** 49.

Savaresi, S.M., & Boley, D.L. (2004). A comparative analysis on the bisecting K-means and the PDDP clustering algorithms. *International Journal on Intelligent Data Analysis,*

Setnes, M. (2000). Supervised fuzzy clustering for rule extraction. *IEEE Trans. Fuzzy Systems, 8*, 416-424.

KEY TERMS

Bioinformatics: The processing of the huge amount of information pertaining to biology.

Discriminant Variables: The information that really matters among the many apparently involved in the true core of a complex set of features.

DNA: Nucleic acid, constituting the genes, codifying proteins.

Gene Expression: The proteins actually produced in the specific cell by the individual.

K-Means: Iterative clustering technique subdividing the data in such a way to maximize the distance among centroids of different clusters, while minimizing the distance among data within each cluster. It is sensitive to initialization.

Leukemia: Blood disease affected by genetic factors.

Micro-Arrays: Chips where thousands of gene expressions may be obtained from the same biological cell material.

PDDP (Principal Direction Divisive Partitioning): One-shot clustering technique based on principal component analysis and singular value decomposition of the data, thus partitioning the dataset according to the direction of maximum variance of the data. It is used here in order to initialize *K-means.*

Principal Component Analysis: Rearrangement of the data matrix in new orthogonal transformed variables ordered in decreasing order of variance.

Singular Value Decomposition: Algorithm able to compute the eigen values and eigen vectors of a matrix; also used to make principal components analysis.

Unsupervised Clustering: Automatic classification of a dataset in two of more subsets on the basis of the intrinsic properties of the data without taking into account further contextual information.