Unsupervised Mining of Genes Classifying Leukemia

Diego Liberati
Consiglio Nazionale delle Ricerche, 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 much 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 encyclopedia (Liberati, 2004), where different techniques also are referenced, and where a clustering approach to piecewise 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, 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 subsequent application 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 (2005) 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 $\widetilde{S}$ as:

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

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

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

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

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:

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