AN UNSUPERVISED CLUSTERING APPROACH FOR LEUKAEMIA CLASSIFICATION BASED ON DNA MICRO-ARRAYS DATA

Simone Garatti[a] Sergio Bittanti[a] Diego Liberati[b] Andrea Maffezzoli[c]

[a]Dipartimento di Elettronica e Informazione - Politecnico di Milano
Piazza Leonardo da Vinci 32, 20133 Milano, Italia. {bittanti,sgaratti}@elet.polimi.it

[b]Istituto di Elettronica ed Ingegneria dell'Informazione e delle Telecomunicazioni
Consiglio Nazionale delle Ricerche, Milano, Italia. liberati@elet.polimi.it

[c]Dipartimento di Bioingegneria - Politecnico di Milano
Piazza Leonardo da Vinci 32, 20133 Milano, Italia.

Abstract: DNA micro-arrays provide thousands of genomic expressions on the same subject. A main issue is then to find the subset of genes whose degeneration is responsible of a certain type of cancer. In this paper, starting from a paradigmatic classification problem of two kinds of Leukaemia, we discuss the use of data-mining techniques in such a context. Particular attention is devoted not only to the classification method but also to all the data analysis steps including data pre-processing and information retrieval.

Keywords: Data analysis for biomedical diagnosis, Bioinformatics, Data-mining, Clustering.

1. INTRODUCTION

Micro-arrays technology has marked a substantial improvement in making available a huge amount of data about gene expression (i.e. gene activation level) of subjects in different patho-physiological conditions. The question is then how to extract useful clinical information from such databases. In this context, two fundamental questions arise:

1. Is it possible to retrieve the (possibly unknown) patient case history from the genomic data?
2. Which are the genes responsible for the patient diversification?

Question 1 may be dealt with via automatic classification of data in homogenous groups. In this respect, a basic tool is provided by clustering procedures, which are the subject of many papers and books (see e.g. (Hand et al., [9]) for a recent volume on this topics). As is well known, one can distinguish between supervised procedures and unsupervised procedures. The former make use of a-priori additional information on the data, such as the patient case history available as a result of medical examinations. Such information along with the patient gene expression levels are used to train a classifier which should be able to distinguish among different pathologies on the basis of the gene expressions of one subject. The obtained classifier can be then used e.g. for the disease diagnosis of new patients.

On the opposite, unsupervised clustering performs the classification on the sole basis of the intrinsic characteristics of the data (gene expressions in the present work) by means of a suitable notion of distance. In this case no a-priori information on the patient case history is available and the latter should be reconstructed from the data. In this perspective, unsupervised clustering is also a tool deputed to discovery new pathologies or new forms of already known diseases.

Apart from diagnostic and disease discovering, both supervised and unsupervised clustering play a fundamental role in understanding the causes of cellular malfunctioning and tumours. As a matter of fact, it is common opinion that tumours are caused by the deregulation of some genes, i.e. such genes are over or under activated so as to produce an abnormal quantity of proteins. Needless to say, understanding in detail such deregulation mechanism would be a most valuable help for the development of therapies for tumorous diseases.

In this context, a basic issue is to understand which genes are responsible of a given pathology (question 2 above), and a classification of patients based on their genomic data is of paramount importance for determining which are such deregulated genes. However, since the number of genes to deal with is typically very large, the cluster discrimination rule returned by standard clustering procedures is based on many genes too. Therefore, a further effort is usually required in order to spot which are the (hopefully few) genes responsible of the tumour disease through a suitable information retrieval method.

In this paper, we consider a set of DNA micro-arrays data – first treated in (Golub et al., [6]) and available on Internet (http://www.broad.mit.edu/cancer/) – regarding patients suffering from two kinds of Leukaemia. Our objective is to illustrate by means of such a paradigmatic example which kinds of problems need to be overcome in order to address questions 1 and 2 above. In doing so, we will also introduce a complete data mining procedure which turned out to be effective in the considered case-study.

The methodology adopted herein is based on four steps, the third of which, devoted to clustering, is the core of the approach. In the first two, a suitable pre-processing of data is performed, as will be explained in the sequel. Finally, the retrieval of the information on the most relevant genes for patient classification is the objective of the fourth step. More in detail, our data mining procedure can be outlined as follows:

1. In order to reduce the dimensionality of the problem, a first pruning of genes not likely to be significant for the final classification is performed on the basis of the size of their inter-subject variance.
2. A principal component analysis is applied to establish a hierarchy in the remaining variables so as to point out the most representative components for clustering.
3. The unsupervised point of view is applied so as to achieve the classification without using a priori information on the patient pathology. This approach presents the advantage that it automatically highlights the (possibly unknown) patient case history. Clustering is performed by merging the classical *k-means* approach ((MacQueen, [14]) and (Hand et al., [9])) with the recently developed PDDP algorithm proposed in (Boley, [2]). According to the analysis provided in (Savaresi and Boley, [18]), cascading these two clustering algorithms results in a significant improvement of clustering performance.

4. The classification obtained at the end of step 3 is based on 178 genes as it will be seen later. This number is relatively large and one may wonder whether the classification can be based on a limited set of genes. As said before this would facilitate the comprehension of the disease mechanism. For this reason we have further elaborated the outcome of point 3 in order to shrink the number of genes to a minimum value. As we will see, our final classification is based on 8 genes only.

The leukaemia data-set has been often used as a test-bed in bioinformatics. For example, it was treated in (Golub et al., [6]), (Guyon et al., [8]) and (Liu and Iba, [12]) by resorting to supervised methods. See also (Lu and Han, [13]) for an interesting comparison between many different supervised approaches. A basic reference for an unsupervised analysis is instead (De Moor et al., [3]).

Our approach proved to be effective on the leukaemia data-set too and our final classification (which, we stress again, was obtained without any a-priori information) looks promising.

The paper is organized as follows.

The leukaemia data-set is described in Section 2, while the preliminary data analysis (variance analysis and principal component analysis) is given in Section 3. Then, Section 4 addresses the fundamental problem of data clustering, while Section 5 is devoted to the final “gene shrinking” step. A discussion of the obtained results is the subject of Section 6.

2. THE DATA SET DESCRIPTION

Data are taken from a public repository which has been often adopted as a reference benchmark (Golub et al., [6]) in order to test new classification techniques and compare the various methodologies each other.

The data-base is constituted by the expression of 7129 genes over 72 patients (also called subjects) suffering from Leukaemia. A small portion of such data-set is depicted in Table 1. Here, data points are collected by rows and correspond to patients. Columns, instead, denote human genes and are the descriptive variables. Each patient is determined by a sequence of 7129 real numbers each measuring the activation level (technically speaking, the *expression*) of the corresponding gene. The smaller the expression value the less the gene is activated.

Note that data are quantitative so that data points can be represented as 72 vectors in a 7129-dimensional Euclidean space. We will simply resort to the Euclidean distance in such space for measuring the genomic difference between two subjects.

In order to ease algebraic manipulations of data, the data-set can be also represented by means of a 72x7129 real matrix, which will be denoted by S^0 in the sequel. Each entry s_{ij}^0 of S^0 measures the *expression* of the j^{th} gene for the i^{th} patient.

For the Leukaemia data-set, an a-priori classification of patients is also available, as a result of medical examinations. Precisely, it is known that 47 of the 72 subjects are cases of Acute Lymphoblastic Leukaemia (ALL), while the remaining 25 are cases of Acute Myeloid Leukaemia (AML). Thus, there is a sort of label attached over each patient specifying the type of Leukaemia the patient is suffering from. Such labels could be used for *supervised* classification.

Note, however, that the data analysis has been carried out in this work as though such a-priori information were not available. Indeed, as said in the introduction, our aim was to develop an *unsupervised* procedure for knowledge discovering problems. A-priori information has been considered solely at the end of the entire data-mining process in order to test the performance of the procedure.

3. PRELIMINARY DATA-PREPROCESSING

A typical bottleneck in DNA micro-arrays experiments, making the classification problem even harder, is the difficulty to collect a high number of homogeneous subjects: not only a big matrix is involved, but such matrix has a huge number of variables (7129 genes) with a very small number of samples (72 subjects). Needless to say, finding the most significant coordinates among these 7129 variables is of paramount importance to model the data distribution and, consequently, to perform clustering (see also next Section 4).

In this work, the search of the most significant coordinates is performed in two steps:

- A preliminary variable (gene) pruning is first performed in order to eliminate those variables which are per-se of little significance.
- The most significant variable combinations are determined by resorting to the well known Principal Component Analysis (PCA).

These two points are now discussed in order.

3.1. Preliminary pruning

The first reduction of the problem dimensionality is obtained through an univariate analysis. Precisely, the empirical variance of the gene expression values over subjects is computed for each gene in order to have a first indicator of inter-subject gene expression variability. Then, the genes whose variability is below a defined threshold are rejected leading to a first pruning. The simple idea behind is that if the variability of a gene expression over the subjects is small, then that gene is similarly expressed for each patient and hence is not useful for classification purposes.

The result of the variance analysis for the 7129 genes in the Leukaemia data-set is depicted in Figure 1. As can be seen, the variance is small for thousands of genes. Having selected a suitable threshold, 6951 genes has been pruned, and attention has been focused on 178 genes only. In the sequel the remaining 72x178 data matrix will be denoted by S.

Though a gene pruning based on gene variance analysis is a bit rough way to proceed, it is often quite useful since the computational burden of subsequent steps of the data-mining procedure would be excessive without a first reduction of dimensionality (see also next Remark 3). The only delicate point in the first gene pruning procedure regards the choice of the cut-off level, which is in fact a tuning parameter of the method. To this end, note that the adopted level may be decided on the basis of biological considerations (e.g. it is known that the natural variability of gene expressions between homogenous subjects is no greater than a certain level), or technological ones (taking into account DNA micro-arrays measurement confidence) or simply following empirical considerations (e.g. selecting either a maximum of residual variables or a maximum fraction of the sum of variances to be discarded). This latter was the criterion adopted for the leukaemia data-set. The threshold was indeed selected so as to eliminate a number of genes corresponding to the 30% of the sum of variances.

Plainly, when gene pruning is applied, a piece of available information is lost, even though the threshold is chosen at best. It is, however, worth noticing that available variables in micro-arrays data (i.e. the genes) are typically highly redundant thanks to the (biological) mechanism of gene co-regulation. That is why the initial gene pruning can be considered "robust" for micro-arrays data analysis.

3.2. Principal component analysis

Principal Component Analysis ((O'Connel, [15]), (Hand et al., [9])) is a well known multivariate analysis by means of which it is possible to bring into evidence the linear combinations of variables with higher inter-subject variance, namely those combinations which are most useful for classification. More precisely, PCA returns a new set of orthogonal coordinates for the 178 dimensional data space resulting from the previous gene-pruning step. The new coordinates are ordered in such a way that the first one, the so called first principal component, denotes the direction with the greatest inter-subject variance, the second one (the second principal component) has the greatest inter-subject variance among all the directions orthogonal to the first component, and so on.

The computation of the principal components of S is made easy by the fact that, if the columns of S have zero mean, the first principal component is the eigenvector associated with the largest eigenvalue of the covariance matrix $S^T S$. Furthermore, the second principal component is the eigenvector corresponding to the second largest eigenvalue of $S^T S$, and so on (see e.g. (Hand et al., [9]) for a simple proof).

Remark 1: *The requirement that the columns of* S *have zero mean can be fulfilled considering in place of* S *the unbiased matrix* $S - e \cdot w$, *where* $e = [1,1,\dots,1]^T$ *and* w, *the so called* centroid *of* S, *is a vector* $[w_1, w_2, \dots, w_p]$ *where*

$$w_k = \frac{1}{N}\sum_{i=1}^{N} s_{ik}\ ,\ k = 1,2,\dots,p$$

($[s_{i1}, s_{i2}, \dots, s_{ip}]$ *is i*th *row of* S, $i = 1,2,\dots,N$ *).*

Remark 2: *Interestingly enough, if data are projected onto the i*th *principal component, the variance of projected data is given by the i*th *largest eigenvalue, say* λ_i, *of* $S^T S$. *Correspondingly, if data are projected onto the subspace generated by the first k principal* *components, the variance of the projected data is* $\sum_{i=1}^{k} \lambda_i$ *. The squared error in terms of approximating the true data matrix using only the first k principal components is* $\sum_{i=k+1}^{178} \lambda_i / \sum_{i=1}^{178} \lambda_i$ *.*

These considerations call for the notion of singular value of a matrix, originally introduced in by Eugenio Beltrami, (Beltrami, [1]). Indeed, the eigenvalues of matrix $S^T S$ are the singular values of S squared.

To be precise, the singular value decomposition (SVD) of a matrix, say S, consists in seeing matrix S as the product of three matrices, that is $S = U\Sigma V$, where U and V are orthonormal unitary square matrices (i.e. $U'U = I$ and $V'V = I$) and Σ is diagonal with real and non negative elements on the diagonal. Such decomposition can be performed for any matrix. In our case, the dimensions of the various matrices are 72×72 for *U*, 178×178 for *V* and 72×178 for Σ. The elements on the diagonal of Σ, which can be always organized in a decreasing sequence starting from the largest element in position $(1,1)$ of Σ, are named *singular values* of S. Furthermore, note also that the columns of V corresponds to the (ordered) eigenvectors of $S^T S$.

For more details and an efficient algorithm for computing the SVD, see (Golub and Van Loan, [5]).

Remark 3: *In principle, PCA could be applied to matrix* S^0 *directly, without any preliminary variable pruning. This, however, is not a wise procedure in general because of the computational over-effort required by the high number of initial variables (7129 in our problem). The cut-off on low inter-subject gene variance is most useful in order to limit the computational burden of PCA.*

4. CLUSTERING

As far as clustering is concerned, we resort to a bisecting divisive partitioning algorithm. In brief (see e.g. (Grabmeier and Rudolph, [7]), (Jain and Dubes, [10]) and (Jain et al., [11]) for a more detailed discussion), this algorithm is first used to split the entire data-set in two clusters (bisection) so as to maximize the intra-similarity and to minimize the inter-similarity of the partition. Then, the same bisecting procedure is iteratively applied, each time dividing a single cluster among those obtained in the previous step, until a final partition of the initial data-set is reached.

The section is organized as follows. The basic algorithm for bisection is introduced in the next Section 4.1 while Section 4.2 addresses the problem of how to iterate bisections so as to obtain an optimal data-set partition. To be precise, we will tackle the following questions: Which is the best cluster to be split at each iteration? Moreover, when is it advisable to stop iterations? Some additional remarks and possible variants of the proposed method are also discussed in Section 4.3.

It is worth mentioning that the proposed approach is very general, and is not limited to the bio-informatics field. For instance it was successfully used for analyzing the data regarding a large virtual community of Internet (see (Garatti et al., [4])).

4.1 PDDP and k-means

According to the analysis developed in (Savaresi and Boley, [16]) and (Savaresi and Boley, [18]) the cluster bisection at each iteration has been performed by means of the cascade of the Principal Direction Divisive Partitioning (PDDP) algorithm and the bisecting *k-means* algorithm. For the sake of self-consistency of this paper, these two algorithms are briefly outlined in Tables 2 and 3. In both cases, the input is a $N \times p$ matrix X where data samples are the rows of the matrix, while the output is given by two matrices X_L and X_R corresponding to data bisection (i.e. the concatenation of X_L and X_R gives X).

PDDP is a recently proposed technique (Boley, [2]) which is representative of non-iterative techniques. The idea behind PDDP is that in many situations data are typically aggregated in two (possibly spherical-shaped) clusters so that the first principal component is oriented from one cluster to the other one and the data-set centroid lies between the two clusters. In such a case, the partition of the two clusters can be achieved by means of an hyper-plane orthogonal to the first principal component and passing through the centroid of the data-set.

K-means was first introduced in (MacQueen, [14]), and, probably, it is the best known and most widely used clustering technique. Hence, it is the best representative of the class of iterative centroid-based divisive algorithms. The idea behind *k*-means is as follows. If the centroids w_L and w_R of the two clusters to be find were known, the data could be partitioned grouping the points close to w_L in one cluster and those close to w_R in an another one (as it can be easily derived from the algorithm, this corresponds to separate data points through a cut orthogonal to the direction $u = (c_R - c_L)/\|c_R - c_L\|$). In practice w_L and w_R are not known in general, so that at the beginning of *k-means* the partition rule above is performed with two randomly chosen points, c_L and c_R, in place of w_L and w_R. Then the result is refined through iterations using the centroids of the clusters obtained at the previous iteration as c_L and c_R.

The main flaw of *k-means* is its initialization since different initial values of c_L and c_R may result in substantially different data partitions, (Selim and Ismail, [19]). PDDP is instead a one-shot algorithm, hence not suffering from any initialization effect. Yet, its drawback is that the assumption that data can be partitioned across the first principal component may not be fulfilled for some data configurations.

It has been proven ((Savaresi and Boley, [16]), (Savaresi et al., [17]) and (Savaresi and Boley, [18])) that the best performance (in terms of quality of partition and of computational effort) can be obtained by applying PDDP, followed by *k-means* initialized with the centroids of the clusters obtained as a result of PDDP. In such a way, the initialization problem of *k-means* is avoided, and the final bisection takes advantage of the positive features of both methods.

4.2 Iterating bisections

The PDDP + *k-means* algorithm can be used to bisect the original data-set in two clusters. As is obvious, in order to obtain a multi-cluster partition, one may proceed to a further bisection of one of the two clusters resulting from the first bisection. In this way, one is left with a 3-partition of the original data-set. Of course, one may further bisect one of these 3 clusters, and so on and so forth, each time dividing a single cluster among those obtained in the previous iteration.

In order to put this procedure into practice, it is necessary to take at each iteration a decision on which cluster has to be split. Moreover, one has to decide when halting the iterations as well. For these decisions, we follow the lead proposed in (Savaresi et al., [17]) as briefly

explained in the following. The interested reader is referred to (Savaresi et al., [17]) for additional technical details.
At the end of the *i*-th iteration, a set of *i*+1 clusters, say $X_1^i, X_2^i, \ldots, X_{i+1}^i$, is obtained.. For each one of these sets, a suitable performance index $\gamma\left(X_r^i\right)$ is computed as follows. Let $X_{r,L}^i$ and $X_{r,R}^i$ be the clusters which would be obtained if X_r^i were portioned through PDDP + *k-means*, and let ${X_{r,L}^i}^\perp$ and ${X_{r,R}^i}^\perp$ the projection of $X_{r,L}^i$ and $X_{r,R}^i$ over the direction orthogonal to the one along which the bisection is performed. Then, $\gamma\left(X_r^i\right)$ is defined as the ratio between the average variance of ${X_{r,L}^i}^\perp$ and ${X_{r,R}^i}^\perp$, and the half squared distance between the centroids of ${X_{r,L}^i}^\perp$ and ${X_{r,R}^i}^\perp$. Thus, $\gamma\left(X_r^i\right)$ quantifies the separation degree between $X_{r,L}^i$ and $X_{r,R}^i$, and the lower $\gamma\left(X_r^i\right)$ the better the two clusters are separated.
Thus, altogether, the meaning of γ is such that the bisection at iteration *i*+1 is performed for the cluster X_r^i for which γ takes the lowest value. Moreover, a value of $\gamma\left(X_r^i\right)$ higher than a given threshold can be taken as a halting rule of the iterative procedure.

4.3 Some additional remarks

The reader should be aware of the fact that in many cases, depending on the available data, some modifications to the basic PDDP and *k-means* algorithms presented in Section 4.1 are desirable in order to improve the quality of the data-set clustering. Here, we discuss two slight variants of PDDP and *k-means,* which, as we will see later, turned out to be useful for the Leukaemia data-set.

As far as PDDP is concerned, one should note that in the basic algorithm bisection is typically performed by an hyper-plane orthogonal to the first principal component. However, such a strategy may be not effective in certain problems. For example, consider the data-set represented in Figure 2.
As can be seen, the data give rise to two parallel clouds, and clustering them through a cut across to the first principal component would be nonsense. Rather, one should take the second principal component and perform bisection through an hyper-plane orthogonal to such component.
Plainly, in unsupervised clustering, there are no a-priori hints on which is the best principal

component across which performing bisection, and such a decision should be retrieved from the data-set itself. To this purpose, one can project data on each principal component in turn, and inspect from graphical visualizations in which case data points are better separated. Alternatively, a more rigorous approach (yet, more time consuming) consists in performing bisection according to each principal component at a time, and then deciding which component results in the best partition through an extensive computation of the index γ.

The second variant of PDDP and *k-mean*s algorithm regards *k-means*. Indeed, after that PDDP has been applied, the quality of the partition refinement obtained through *k-means* may depend on the number of variables (i.e. the number of columns) of the data matrix given as input to the algorithm. As a matter of fact, it is well known that too many variables may adversely affect the convergence of *k-means*.

For this reason, it is advisable in many cases to project data onto a moderate number of principal components (so as to reduce the number of descriptive variables) before applying *k-means*. Notice that for this problem too, an automatic procedure for the selection of a limited number of principal components can be worked out by resorting to the index γ.

5. GENE SHRINKING

As already pointed out in Section 4, after that PDDP + *k-means* is applied one time, the data turns out to be eventually bisected according to the following (linear) classification rule (w_L and w_R are the final centroids returned by the *k-means* algorithm):

$$\begin{cases} x_i \in X_L & \text{if } x_\mathrm{i} \cdot u \le K \\ x_i \in X_R & \text{if } x_\mathrm{i} \cdot u > K \end{cases} \qquad (1)$$

where $u = (w_R - w_L)/\|w_R - w_L\|$ and $K = 0{,}5 \cdot (w_R + w_L) \cdot u$.

The problem within this expression is that if the linear inequalities above are referred to the original coordinates (i.e. the genes), we obtain a classifier depending on all the 178 relevant genes, characterizing each patient. This in turn implies that the expression above is of minor interest from a biological perspective as it involves too many genes. For this reason, the classification procedure outlined in Section 4 has been complemented with a “gene shrinking” technique in order to detect which are the (hopefully few) genes actually relevant for the classification.

The approach we propose is the following one. Suppose that vector u above is written as $[u_1, u_2, \dots, u_{178}]$ where, without any loss of generality, these components are sorted by decreasing values of $|u_i|$. In a sense, $|u_i|$ measures the importance of the i^{th} variable for the classification. Then, we consider $u' = [u_1, \dots, u_{177}, 0]$ in place of u in (1) and we search for a partitioning threshold K' ($K' \neq K$ in general) such that the original data partition is preserved (i.e. the same X_L and X_R are obtained even though u' and K' are used in place of u and K). If such K' exists, a new classifier based on 177 genes only is found. This procedure can be then iterated, eliminating one by one all the less relevant components of u. The stopping rule is determined by a certain $u' = [u_1, \dots, u_{l-1}, 0, \dots, 0]$ for which no K' preserves the original data partition. This returns $u_1, \dots, u_l$ as the genes actually relevant for the patient classification.

6. RESULTS AND DISCUSSION

The procedure outlined above was applied to the Leukaemia data-set. The application of the first phase of clustering based on PDDP algorithm with the aid of the γ index (see Section 4.3) led to a bisection by an hyper-plane orthogonal to the *second* principal component. Then, the subsequent application of *k-means*, again worked out by means of the γ index, was limited to the principal components from the 2nd to the 11th.

In principle, further bisection trials should have be performed after the first one. However, the use of the γ index (see Section 4.2) highlighted that there was no significant improvement proceeding with further bisections.

The final result of the clustering procedure led to the partition of the original set of 72 patients into two clusters consisting of 23 and 49 subjects (Figure 3).

Recall that our partition is obtained in a fully unsupervised and blind way, namely without exploiting the a priori information on the pathology of the patients (ALL or AML) which, we recall, was available as a result of medical examinations.

Although such additional information was not used, the adopted data-mining procedure was able to detect a (genomic) difference between two kinds of patients. I.e. if labels had been not known, it would have provide evidence of the existence of two type of Leukaemia. This, of course, would have been of paramount importance for starting further medical investigations and for developing different treatments for patients.

In the case of Leukaemia we were already aware of the existence of two kinds of Leukaemia, so that the result is not surprising, though it confirms the validity of the approach. Furthermore, available labels can be used to a-posteriori evaluate the quality of the obtained partition. This can be simply done by verifying how ALL and AML patients are distributed among the two clusters returned by our method.

Interestingly enough, all the 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. The misclassification percentage is 2/72=3%. Thus, altogether, not only the unsupervised procedure recognizes that there is an intrinsic difference between two groups of patients, but also the obtained partition corresponds with great accuracy to the actual partition in ALL/AML patients.

Remark 3: *Note that in our procedure all data points are used to detect the best data partition. There is no distinction between training and validation samples as it is common for other methods. The reason for this relies on the unsupervised nature of our approach. Indeed, a separation between training and validation samples is required in supervised procedure (where labels are known and are used to train the classifier) to avoid data over-fitting. In unsupervised approach (where the classification is retrieved from genomic data only based on their relative distance) over-fitting is not an issue as there are no labels which the classifier can over-fit. In a sense, in unsupervised methods the separation between training and validation data points is already achieved separating genomic data (from which the classifier is trained) from labels (which in fact can be a-posteriori used to test the quality of the results).*

A further observation which is worth doing is that the final gene-shrinking step leads to a very small number of significant genes, precisely to the 8 genes listed in Table 4.

The final classifier, we are left with after that the gene-shrinking step is performed, is given by $u' = [-0.235, -0.195, -0.228, -0.207, -0.159, -0.197, -0.172, -0.164]$ as weighting vector, and $K' = -1.1$ as partitioning threshold.

Our results are comparable with those obtained through different approaches. E.g. in (Golub et al., [6]), by using a supervised tool and thus splitting the 72 patients in 38 training samples and 34 testing samples, a correct classification was obtained for 29 (about 85%) of the 34 test subjects.

Perhaps it is worth noticing that the classification in (Golub et al., [6]) was based on genes in general different from those in Table 4. Only 3 genes (namely, CST3 Cystatin C, Azurocidin and Interleukin-8 precursor) appear in both the approaches. A possible interpretation is that these three genes within the intersection of the two subsets are probably really determinant, whereas the complementing genes identified by the procedure proposed in the present paper and those in the subset of Golub and co-workers serve only to refine the partition and are somehow equivalent.

In order to highlight the relative importance of the 8 genes in Leukaemia classification, their expression across 72 patients is shown in Figure 4, where the 47 ALL patients are grouped at the left side and the 25 AML patients are grouped at the right side of each subfigure. The two sets are separated by a vertical line in order to visually highlight the two groups.

As it can be appreciated even by a visual inspection of Figure 4, the CST3 Cystatin C (M27892_at) gene is the most influential one for classification, since its expression in the leftmost 47 ALL patients is uniformly low in comparison to the other class of patients. In fact, such gene is the one maximally projecting on the principal component used for clustering. However, it is worth noting that a classification based on this gene alone would not be effective.

7. CONCLUSION

In this paper, we have faced the problem of discriminating two kinds of Leukaemia on the basis of a suitable data-mining procedure for micro-arrays genetic data. The unsupervised nature of the presented approach enables the classification without any knowledge on the pathology of patients. The results of the data analysis show that the discrimination can be effectively performed by means of only 8 genes of the original 7129 genes available on the micro-array.

ACKNOWLEDGEMENTS

Paper supported by the MIUR National Research Project “Innovative Techniques for Identification and Adaptive Control of Industrial Systems” and partially by CNR-IEIIT.

REFERENCES

[1] Beltrami, E. (1873). “Sulle funzioni bilineari”. Giornale di matematiche, Vol. XI, 98-106.

[2] Boley, D.L. (1998). “Principal Direction Divisive Partitioning”. Data Mining and Knowledge Discovery, 2(4), 325-344.

[3] De Moor, B., K. Marchal, J. Mathys and Y. Moreau (2003). “Bioinformatics: Organism from Venus, Technology from Jupiter, Algorithms from Mars”. European Journal of Control 9 (2-3).

[4] Garatti, S., S. Savaresi, S. Bittanti (2004). “On the relationships between user profiles and navigation sessions in virtual communities: a data-Mining approach”. Intelligent Data Analysis 8(6): 576-600.

[5] Golub, G.H., C.F. van Loan (1996). “Matrix Computations (3rd edition)”. The Johns Hopkins University Press.

[6] Golub, T.R., D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander (1999). “Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression”. Science 286:531-537 .

[7] Grabmeier, J., A. Rudolph (2002). “Techniques of cluster algorithms in data mining”. Data Mining and Knowledge Discovery, 6(4), 303-360.

[8] Guyon, I., J. Weston, S. Barnhill, and V. Vapnik (2002). “Gene selection for cancer classification using support vector machines”. Machine Learning, 46:389-422 .

[9] Hand, D., H. Mannila, P. Smyth (2001). “Principles of Data-Mining”. The MIT press, Cambridge, Massachusetts, USA.

[10] Jain, A., R.C. Dubes (1988). “Algorithms for clustering data”. Prentice-Hall.

[11] Jain, A.K, M.N. Murty, P.J. Flynn (1999). “Data Clustering: a Review”. ACM Computing Surveys, Vol.31, n.3, pp.264-323.

[12] Liu, J., and H. Iba (2001). “Selecting informative genes with parallel genetic algorithms in tissue classification”. Genome Informatics, 12:14-23.

[13] Lu, L., and J. Han (2003). “Cancer classification using gene expression data”. Information Systems, 28(4):243-268.

[14] MacQueen, J. (1967). “Some methods for classification and analysis of multivariate observations”. In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, LM LE Cam & J Neyman (eds.), Univeristy of California Press, Berkeley, pp. 291-297.

[15] O'Connel, M.J. (1974). “Search Program for Significant Variables”. Computational Physic Communications, 8:49.

[16] Savaresi, S.M., D.L. Boley (2001). “On the performance of bisecting *k-means* and PDDP”. In Proceedings of the 1st SIAM Conference on Data Mining, Chicago, IL, USA, paper n.5, pp.1-14.

[17] Savaresi, S.M., D.L. Boley, S. Bittanti, G. Gazzaniga (2002). “Cluster selection in divisive clustering algorithms”. In Proceedings of the 2nd SIAM International Conference on Data Mining, Arlington, VI, USA, pp.299-314.

[18] Savaresi, S.M., D.L. Boley (2004). “A Comparative Analysis on the Bisecting *k-means* and the PDDP Clustering Algorithms”. International Journal on Intelligent Data Analysis 8(4): 345-362.

[19] Selim, S.Z., M.A. Ismail (1984). “*K-means*-type algorithms: a generalized convergence theorem and characterization of local optimality”. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.6, n.1, pp.81-86.

	BioB 5_at	**BioB 5_st**	**CreX 5_st**	**DapX M_at**	…
Patient 1	-214	206	-118	311	…
Patient 2	-139	74	-141	134	…
Patient 3	-76	-215	84	378	…
Patient 4	-135	31	107	268	…
Patient 5	-106	252	1	118	…
Patient 6	-138	193	-1	154	…
…	…	…	…	…	…

Table 1: the Leukemia data-set

PDDP clustering algorithm

Compute the centroid w of X and compute the unbiased matrix $\tilde{X} = X - ew$, $e = [1,1,\ldots,1]^T$.
Compute v, the first principal component of $\tilde{X}$.
Divide $X = [x_1, x_2, \ldots, x_N]^T$ into two subclusters X_L and X_R, according to the following rule:

$$\begin{cases} x_i \in X_L \text{ if } v^T(x_i - w) \leq 0 \\ x_i \in X_R \text{ if } v^T(x_i - w) > 0 \end{cases}$$

Table 2: PDDP clustering algorithm.

Bisecting *k-means* algorithm

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

Step 2. Divide $X = [x_1, x_2, \ldots, x_N]^T$ into two sub-clusters X_L and X_R, according to the following rule:

$$\begin{cases} x_i \in X_L & \text{if } \|x_i - c_L\| \le \|x_i - c_R\| \\ x_i \in X_R & \text{if } \|x_i - c_L\| > \|x_i - c_R\| \end{cases}$$

Step 3. Compute the centroids of X_L and X_R, w_L and w_R.

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

Table 3: Bisecting *k-means* algorithm.

The 8 genes able to discriminate between AML and ALL

1. FTL Ferritin, light polypeptide M11147_at
2. MPO Myeloperoxidase M19507_at
3. CST3 Cystatin C (amyloid angiopaty and cerebral hemorrage) 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
8. GB DEF Cystic fibrosis antigen mRNA M26311_s_at

Table 4: The 8 genes able to discriminate between AML and ALL

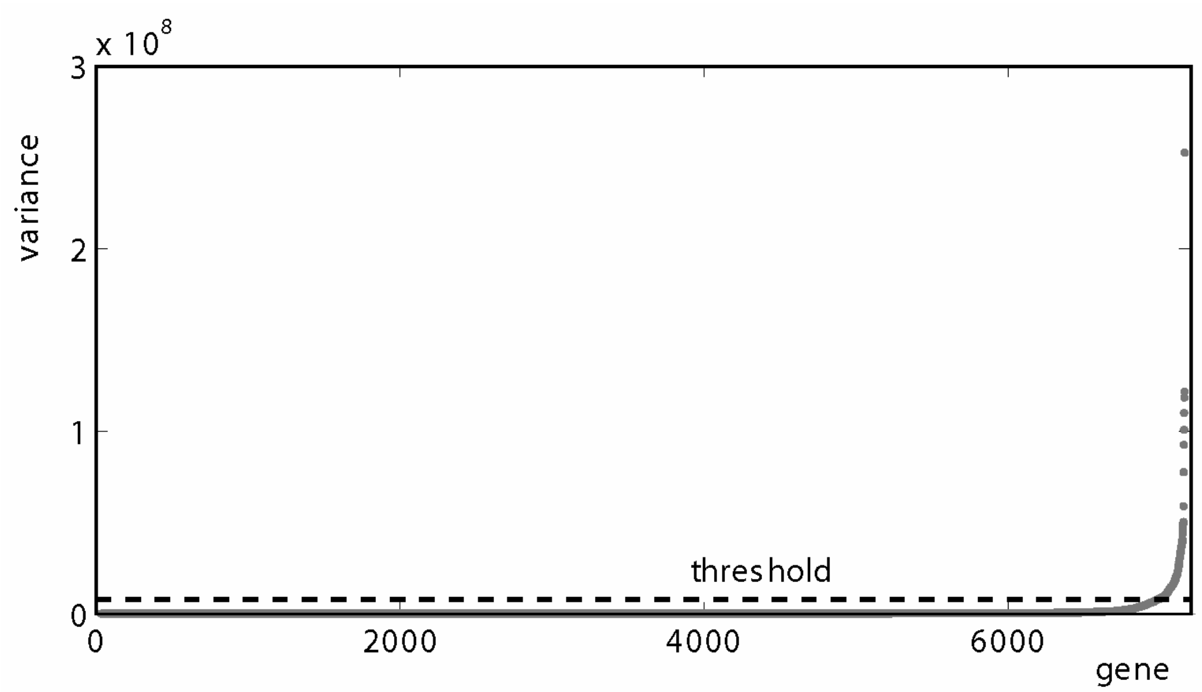
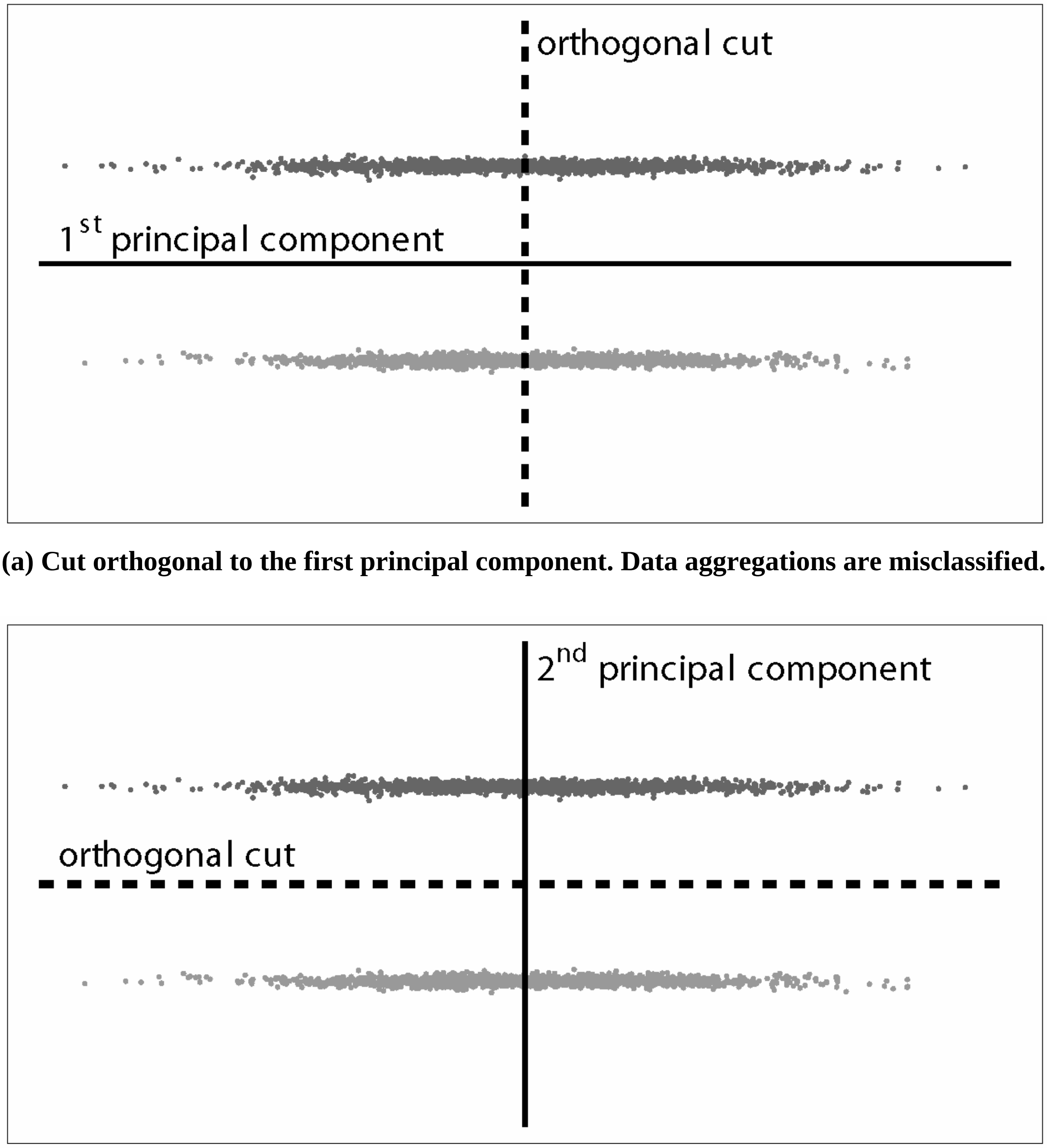


Fig. 1: Inter-subject variance



(a) Cut orthogonal to the first principal component. Data aggregations are misclassified.

(b) Cut orthogonal to the second principal component. Data are correctly clustered.

Fig. 2: A data configuration where bisection should be performed orthogonally to the second principal component.

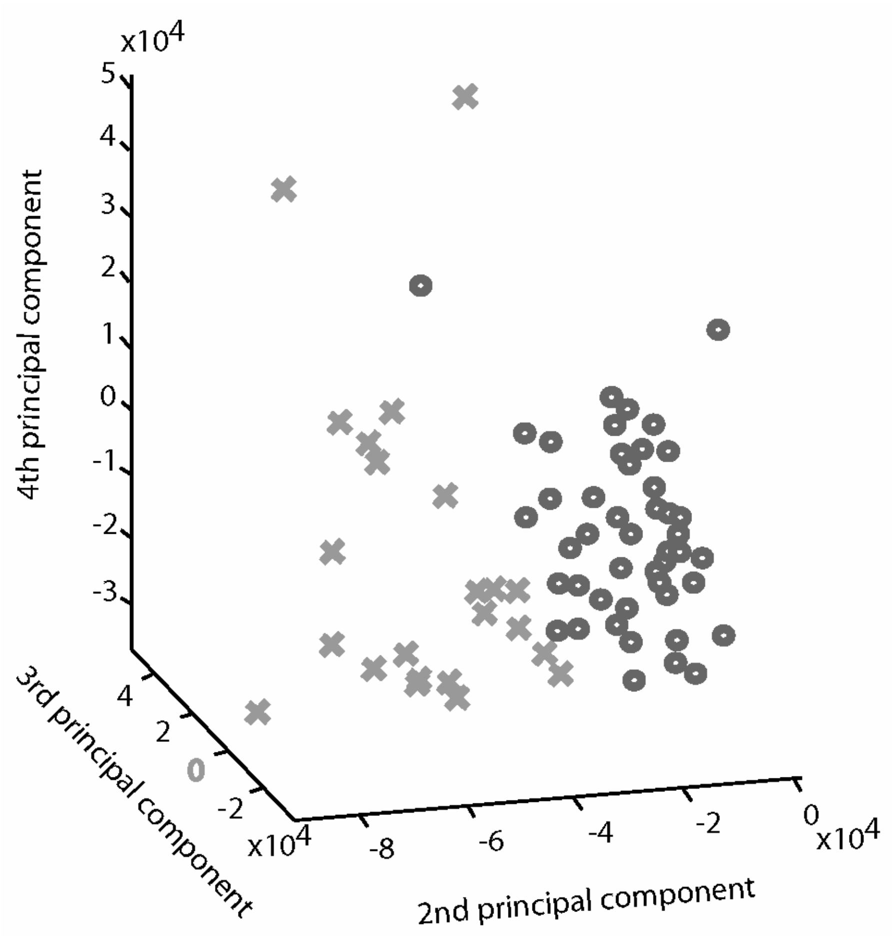


Fig. 3: PDDP + *k-means* data partition ("x"=first cluster, "o"=second cluster) .

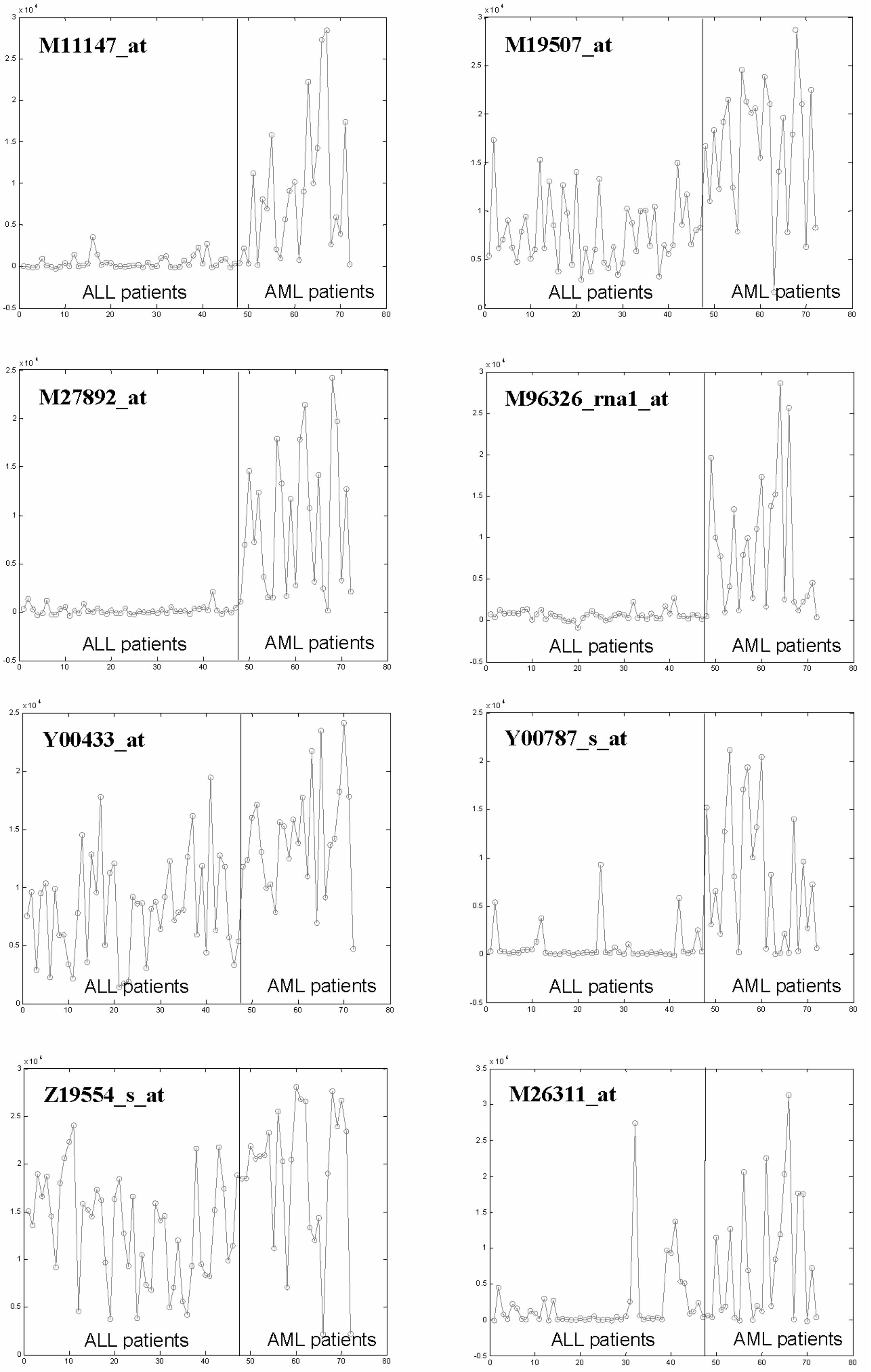


Fig. 4: expression values for the genes classifying Leukaemia patients