FROM DNA MICRO-ARRAYS TO DISEASE CLASSIFICATION: AN UNSUPERVISED CLUSTERING APPROACH

Sergio Bittanti[a] Simone Garatti[a] Diego Liberati[b]

[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

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 methods 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 issues arise:

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

Point 1 is related to automatic classification of data in homogenous groups. In this connection, a basic tool is provided by clustering procedures, which are the subject of many papers and books (see (Hand et al, 2001) for a recent volume on this topics). As is well known, one can distinguish supervised procedures and unsupervised procedures. The former makes use of a-priori additional information on the data, such as the patient casuistry 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 a subject. The obtained classifier can be then used e.g. for the disease diagnosis of new subjects.

In contrast to the supervised approach, unsupervised clustering performs the classification on the sole basis of the intrinsic characteristics of the data (gene expressions) by means of a suitable notion of distance. In this case no a-priori information on the patient casuistry is available and the latter should be reconstructed from the data. In this perspective, unsupervised clustering is the 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 tumour diseases. 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 tumour diseases

In this context, a basic issue is to understand which genes are responsible of a given pathology (point 2 above), and plainly a classification of patients based on their genomic data is of paramount importance for determining which are such deregulated genes. Perhaps, it should also be said that clustering algorithms do not provide in general this kind of information directly. Indeed, clusters involve thousands of genes. Therefore, suitable information retrieval methods have to be used so as to spot which are the (hopefully few) genes responsible of the tumour disease.

In this paper, we consider a set of DNA micro-arrays data – first treated in (Golub et al 1999) and available on Internet – regarding patients suffering from two kinds of Leukaemia. Our objective is to show by means of a paradigmatic example how the two points above can be addressed through a complete data-mining approach.

The methodology we adopt herein is based on four steps, the third of which, devoted to clustering, is the core of the approach. The first two, instead, constitute a pre-processing of data. Finally, the retrieval of the information on the most relevant genes for the patient classification is the objective of the fourth step.

More in detail, the data mining procedure can be outlined as follows:

1. A first pruning of genes not likely to be significant for the final classification is performed on the basis on their small inter-subject variance, thus reducing the dimensionality of the problem
2. A principal component analysis defines a hierarchy in the remaining transformed orthogonal variables so as to point out the variables most representative for clustering.
3. The unsupervised point of view is applied so as to achieve the classification without using a priori information on the patient's pathology. This approach presents the advantage that it automatically highlights the (possibly unknown) patient casuistry. Clustering is performed by merging the classical *k-means* approach ((MacQueen 1967) and (Hand et al, 2001)) with the recently developed PDDP algorithm proposed in (Boley 1998). According to the analysis provided in (Savaresi & Boley 2004), cascading these two clustering algorithms results in a significant improvement of performances.
4. By analyzing the obtained results, the number of genes for the detection of pathologies is further reduced, so that the classification is eventually based on a minimum number of genes.

The leukaemia dataset 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. Our approach seems to present significant advantages in that the classification is eventually based on a very limited number of genes, and no a priori information is required.

1.1 Structure of the paper

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 on 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 1999) in order to test new classification techniques and compare the various methodologies each other.

The data-base is constituted by gene expression data over 72 subjects suffering from Leukaemia, relying on 7129 genes. A small portion of such data-set is depicted in Table 1.

Table 1: the Leukemia data-set

	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	…
…	…	…	…	…	…

Here, data points are collected by rows and correspond to patients (also called subjects). 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. If such a value is negative, then the gene is poorly activated with respect to a standard reference (the smaller the expression the less the gene is activated) and we will say that the gene is under-expressed. Similarly, when such value is positive, the gene is said to be over-expressed.

Note that data are quantitative and the data points can be represented as 72 vectors in a 7129-dimensions Euclidean space. A simple measure of the genomic difference between two subjects can be obtained by resorting to the Euclidean distance in such space.

In order to ease algebraic manipulations of data, the data-set can be also represented by means of a real matrix S^0 of dimension 72x7129, the entry s_{ij}^0 of which 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). Such labels attached over subjects could be used for supervised

classification. Herein, however, the data analysis has been carried out as if 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 variance of the expression values is computed for each gene across the patients, in order to have a first indicator of the relative inter-subject 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 Fig. 1.

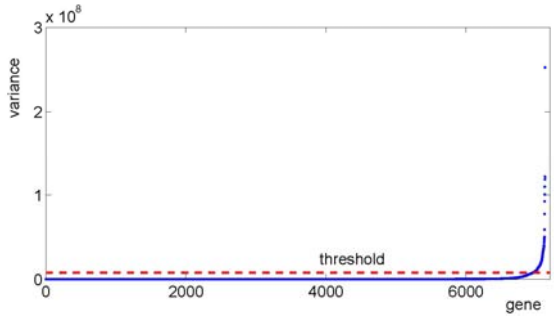


Fig. 1. inter-subject variance

As can be seen, the variance is small for thousands of genes. Having selected a suitable threshold, 6951 genes were pruned. So, attention has been focused on 178 genes only. In the sequel the remaining 72x178 data matrix will be denoted by S.

Plainly, the only delicate point in the above outlined 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 possibly 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), technological knowledge (taking into account DNA micro-arrays measurement confidence) or simply following empirical considerations (e.g. imposing either a maximum of residual variables or a maximum fraction of variance to be discarded).

3.2. Principal component analysis

Principal Component Analysis ((O'Connel 1974), (Hand et al, 2001)) 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, 2001) 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 $e = [1,1,\ldots,1]^T$ and w, the so called* centroid *of S, is a vector $[w_1, w_2, \ldots, w_p]$ where*

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

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

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 $\sum_{i=k+1}^{178}\lambda_i / \sum_{i=1}^{178}\lambda_i$.

Since the eigenvalues of S^TS are the (square of the) singular values of S, PCA can be performed through the singular value decomposition (SVD) of S. That is, write $S = U\Sigma V$, where Σ is a diagonal 72×178 matrix whose non null elements are the singular values of S, and U and V are orthonormal unitary square matrices of dimension 72×72 and 178×178, respectively; V is also the matrix of eigenvectors of S^TS. As it is well known, there are many efficient algorithms for computing the SVD (see e.g. Golub and Van Loan (1996)).

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.*

4. CLUSTERING

As far as clustering is concerned, we resort to a bisecting divisive partitioning algorithm. In brief (see e.g. (Jain et al (1999)) and (MacQueen (1967)) for a more detailed discussion), these algorithms are first used to split the entire data-set in two clusters 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.

The decision on which cluster has to be split at each iteration as well as on when halting the iterations has been guided by the criterion suggested in (Savaresi et al (2002)). This criterion is based on the computation of a certain performance index for a given data matrix (which can be both the initial data set or one of the clusters during the iterations). This index is an indication of the "separation degree" of the two clusters which would be obtained after performing the data matrix bisection. Through an intensive computation of this index, the most convenient splitting at a given iteration can be determined as the one maximizing the performance of the bisection. A performance improvement lower than a given level is taken also as a stopping rule for the algorithm.

According to the analysis developed in (Savaresi and Boley (2001) and (Savaresi and Boley (2004)) 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 5 and 6. In both cases, the input is a $N\times p$ matrix X where data samples are the rows of the matrix, and outputs are two matrices X_L and X_R.

Table 2: 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:

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

Table 3: 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:

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

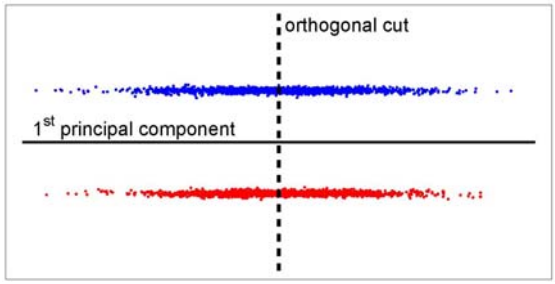
Step 3. Compute the centroids of X_L and X_R, w_L and w_R.

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.

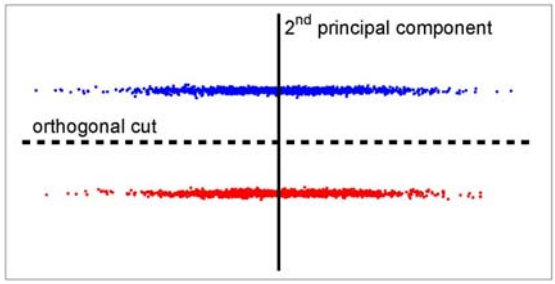
PDDP is a recently proposed technique (Boley (1998)) which is representative of non-iterative techniques. The idea behind PDDP is that data are typically aggregated in two clusters along a principal direction, separated by the centroid. Therefore, the partition of the two clusters can be achieved by means of a cut orthogonal to the considered principal component and passing through the centroid of the data set. It is worth noting that, normally, the bisection is performed by considering the first principal component. However, this has to be considered in the light of available data, and possibly one can skip the first principal component and start with the second one. As an example, consider the data set represented in Fig. 2. As can be seen, the data give rise to two parallel clouds and clustering them orthogonally to the first principal component would be nonsense. Rather, one should take the second component. Note that this is exactly what happens in the Leukaemia data, as we will see later.

K-means was first introduced in (MacQueen (1967)), 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 is as follows. If the centroids c_L and c_R of the two clusters to be find were known, the data could be partitioned grouping the points close to c_L in one cluster and those close to c_R in an another one. Since c_L and c_R are not known in general, the procedure is initially performed with two randomly chosen points, and

then the result is refined through iterations using the centroids of the clusters obtained at the previous iteration as c_L and c_R.



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



(a) 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.

The main flaw of K-means is its initialization since different starting centroids may lead to different results in general. PDDP is instead a one-shot algorithm, which provides a unique solution. Yet, its drawback is that the assumption on data separability along one of the principal components may not be fulfilled for some data configurations.

It has been proven ((Savaresi and Boley (2001), (Savaresi et al (2002)) and (Savaresi and Boley (2004))) 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.

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 2004)).

5. GENE SHRINKING

Going through PDDP and K-means, it can be seen that the final data bisection is performed according to the following (linear) classification rule (c_L and c_R are the final centroids returned by the K-means algorithm):

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

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

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 imply 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 $[u_1, u_2, \ldots, u_{178}]$ 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 $u' = [u_1, \ldots, u_{177}, 0]$ in place of u in (1) and search for a partitioning threshold K' ($K' \neq K$ in general) such that the original data partition is preserved. If such K' exist, a new classifier based on 177 genes only is obtained. This procedure can be iterated, eliminating one by one all the less relevant component of u. The stopping rule is determined by a certain $u' = [u_1, \ldots, u_{l-1}, 0, \ldots, 0]$ for which no K' preserves the original data partition. This returns $u_1, \ldots, u_l$ as the genes actually relevant for the patient classification.

6. RESULTS AND DISCUSSION

By applying the above methodology to the Leukaemia database, the set of 72 subjects has been subdivided into two clusters represented in Fig. 3 (only three principal components are shown here), containing 23 and 49 patients, respectively. It is worth noting that the data bisection is basically performed orthogonally to the second principal component, for the reasons explained in Section 4.

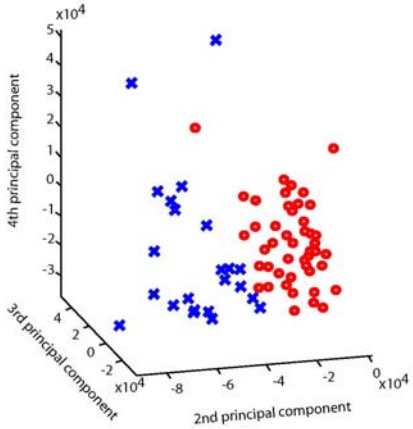


Fig. 3. PDDP+K-means data partition ("x"=first cluster, "o"=second cluster) .

Recall that our portioning has been obtained in a fully automatic and blind way, namely without exploiting a priori information on the pathology of the patients (ALL or AML).

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. Thus the misclassification percentage is 2/72=3%.
In addition, it should be pointed out that the final gene-shrinking step leads to a very small number of significant genes, precisely to the 7 genes listed in Table 4.
Note that in (Golub et al (1999)) the attention is focused on a supervised approach. By 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. Interestingly enough, the intersection between the set of discriminating genes found in that paper and our set of 7 genes is non empty: Cystatin C, Azurocidin and Interleukin-8 precursor appear in both sets. A possible interpretation is that the three genes within the intersection of the two subset are probably really determinant, whereas the complementing 4 genes identified by the procedure proposed in the present paper better discriminate than the complementing 43 in the subset of Golub and coworkers.

Table 4: The 7 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. Glutathione peroxidase 1 Y00433_at
6. INTERLEUKIN-8 PRECURSOR Y00787_s_at
7. VIM Vimentin Z19554_s_at

7. CONCLUSION

In this paper, we have faced the problem of discriminating two kinds of Leukaemia on the basis of data-mining on micro-arrays genetic data. 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 results of the data analysis show that the discrimination can be effectively performed by means of 7 genes only 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. The authors would also like to thank Andrea Maffezzoli who was in charge of the computation in fulfilment of his master thesis at the Politecnico di Milano.

REFERENCES

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

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):237-278.

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): 579-600.

Golub, G.H., C.F. van Loan (1996). “Matrix Computations”. The Johns Hopkins University Press.

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 .

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

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

MacQueen, J. (1967): “Some methods for classification and analysis of multivariate observations”. In Proceeedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, LM LE Cam & J Neyman (eds.), Berkeley, pp. 291-297.

O'Connel, M.J. (1974), “Search Program for Significant Variables”. Computer Physics Communication 8: 49-55.

Savaresi, S.M., D.L. Boley (2001). “On the performance of bisecting K-means and PDDP”. 1st SIAM Conference on Data Mining, Chicago, IL, USA, pp.1-14.

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

Savaresi, S.M., D.L. Boley (2004). “A Comparative Analysis on the Bisecting K-Means and the PDDP Clustering Algorithms”. Intelligent Data Analysis 8(4):345-362.

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.