Two Way Clustering of Microarray Data Using a Hybrid Approach

Raul Măluțan, Bogdan Belean, Pedro Gómez Vilda, and Monica Borda

***Abstract*—The Microarray technique is rather powerful, as it allows to test up thousands of genes at a time, but this produces an overwhelming set of data files containing huge amounts of data, which is quite difficult to pre-process, separate, classify and correlate for interesting conclusions to be extracted. Modern machine learning, data mining and clustering techniques based on information theory, are needed to read and interpret the information contents buried in those large data sets. Independent Component Analysis method can be used to correct the data affected by corruption processes or to filter the uncorrectable one and then clustering methods can group similar genes or classify samples. In this paper a hybrid approach is used to obtain a two way unsupervised clustering for a corrected microarray data.**

***Keywords*—microarray, clustering, k-means, Expectation Maximization, external validation**

I. Introduction

DURING the last years special importance has been placed on the interpretation, classification and recognition of relationships expressed in microarray data to infer the activity of specific genes, using clustering techniques and other statistical analysis tools like blind signal separation ones [1], [2]. In most cases it was tacitly assumed that the obtention of microarray data from genetic samples is a fully reliable process *per se*, not having to take into account the large complexity of the procedures used. The importance of side fields of knowledge as Signal and Image Processing, Pattern Recognition, Statistical Data Analysis, or Automata Theory in relation with microarray data processing challenges have not completely yielded their enormous potential in solving problems as microarray image enhancement, segmentation, correction, gridding, data analysis, reliable expression estimation in relation with hybridization dynamics, etc. Others have to see with data interpretation, dimensionality reduction, cluster analysis, function prediction, etc. Summarizing, the present work uses microarray data corrected by Independent Component Analysis [3] for a hybrid clustering technique with unsupervised algorithms.

Manuscript received May 20, 2011. This work was supported by the project "Development and support of multidisciplinary postdoctoral programmes in major technical areas of national strategy of Research - Development - Innovation" 4D-POSTDOC, contract no. POSDRU/89/1.5/S/52603, project co-funded by the European Social Fund through Sectoral Operational Programme Human Resources Development 2007-2013.

R. Malutan is with the Communications Department, Technical University of Cluj-Napoca, 26-28 George Baritiu St., 400027 Cluj-Napoca, Romania, (phone: 004-0264-401564; fax: 004-264-401575; e-mail: raul.malutan@com.utcluj.ro).

B. Belean is with the Communications Department, Technical University of Cluj-Napoca, 26-28 George Baritiu St., 400027 Cluj-Napoca, Romania, (phone: 004-0264-401564; fax: 004-264-401575; e-mail: bogdan.belean@com.utcluj.ro)

P. Gómez Vilda is with Departamento de Arquitectura y Tecnología de Sistemas Informáticos (DATSI), Universidad Politécnica de Madrid, Campus de Montegancedo, s/n, 28660, Boadilla del Monte, Madrid, Spain (e-mail: pedro@pino.datsi.fi.upm.es).

M. Borda is with the Communications Department, Technical University of Cluj-Napoca, 26-28 George Baritiu St., 400027 Cluj-Napoca, Romania (e-mail: Monica.Borda@com.utcluj.ro).

II. Cluster Analysis

Microarray data is usually represented as a gene expression matrix with the rows corresponding to genes from an experiment and the columns corresponding to different experiments. If one finds that two rows are similar, it can be assumed that the genes corresponding to the rows are co-regulated and functionally related, and by comparing two columns it can found which genes are differentially expressed in each experiment. To perform a comparison between genes or experiments under comparison, a similarity measure between the objects has to be used. There are two methods by which one can study these expression matrices: supervised or unsupervised analysis. If prior knowledge is available about the results, these can be grouped, with neural networks for example, into several predefined classes. Therefore a supervised analysis can identify gene expression patterns, called features, specific to each class, but also classify new samples.

Without any hypothesis, unsupervised approaches can discover novel biological mechanisms and reveal genetic regulatory networks in large datasets when little a priori knowledge is available. Within unsupervised learning, there are three classes of techniques: feature determination, or determining genes with interesting properties without specifically looking for a particular a priori pattern, such as principal component analysis (PCA); cluster determination, or determining groups of genes or samples with similar patterns of gene expression, such as k-means clustering and Expectation Maximization clustering; and network or graph determination representing gene–gene or gene–phenotype interactions using Boolean networks.

For the microarray data the most suitable clustering methods are unsupervised ones, because we cannot observe the (real) number of clusters in the data.

K-means [4], an unsupervised learning algorithm, has been used to form clusters of genes in gene expression data analysis. The algorithm takes the number of clusters (k) to be calculated as an input. The number of clusters is usually chosen by the user. The procedure for k-means clustering is as follows:

1. First, the user tries to estimate the number of clusters.
2. Randomly choose N points into k clusters.

3. Calculate the centroid for each cluster.

4. For each point, move it to the closest cluster.

5. Repeat steps 3 and 4 until no further points are moved to different clusters.

The Expectation-Maximization (EM) algorithm [5] is a method for finding maximum likelihood estimates of parameters in statistical models, where the model depends on unobserved latent variables. This is a general method for optimizing likelihood functions and is useful in situations where data might be missing or simpler optimization methods fail.

If one wish to estimate the parameters $\theta = \pi_1,\ldots,\pi_c,k_1,\ldots,k_c,\sigma_1,\ldots,\sigma_c$, this can be done using the maximum likelihood approach by maximization of the log-likelihood given by:

$$L(\theta|x_1,\ldots,x_n) = \sum_{i=1}^{n} ln\left[\sum_{k=1}^{c} \pi_k \phi(x_i;\mu_k,\sigma_k)\right]. \quad (1)$$

It is assumed that the components exist in a fixed proportion in the mixture, given by the π_k. Thus, it makes sense to calculate the probability that a particular point x_i belongs to one of the component densities. It is this component membership that is unknown, and why we need to use something like the EM algorithm to maximize the equation above. One can compute a posterior probability that an observation x_i belongs to component *k*. Because the posterior probability is unknown, we should compute it in an iterative manner. It is a two step process given below:

1. *E-Step*: calculate the posterior probability that the i^{th} observation belongs to the k^{th} component, given the current values of the parameters;

2. *M-Step*: update the parameter estimates.

For determining the optimal numbers of clusters for the current microarray data clustering validation methods must be applied. In general, we can apply these methods to a range of numbers of clusters in k-means or EM clustering, and determine an estimate of optimal number of clusters from the data.

Clustering validation is a technique to find a set of clusters that best fits natural partitions, *i.e.* number of clusters, without any class information. There are two types of clustering techniques [6]: *external validation,* based on previous knowledge about data and *internal validation*, based on the information intrinsic to the data alone.

Even though we can find different external and internal indexes in the literature, in this paper we choose to use three external indexes: Rand index, Jaccard coefficient, and Fowlkes and Mallows index [7].

Considering *P* the existing partition of the microarray data set, and *C* the clustering structure resulting from the use of clustering algorithms, the performance can be evaluated by comparing *C* to *P* in terms of external criteria. If x_i and x_j are a pair of samples, there are four different cases based on how x_i and x_j are placed in *C* and *P*:

1. x_i and x_j belong to the same clusters of *C* and the same category of *P*.

2. x_i and x_j belong to the same clusters of *C* but different categories of *P*.

3. x_i and x_j belong to different clusters of *C* but the same category of *P*.

4. x_i and x_j belong to different clusters of *C* and different category of *P*.

Correspondingly, the number of pairs of samples for the four cases are denoted as *a*, *b*, *c*, and *d*, respectively. Because the total number of pairs of samples is $M = N(N-1)/2$, from a total number of N samples, we have $a+b+c+d=M$. The external indexes can then be defined as follows:

- Rand index

$$R = (a+d)/M \quad (2)$$

- Jaccard coefficient

$$J = a/(a+b+c) \quad (3)$$

- Fowlkes and Mallows index

$$FM = \sqrt{\frac{a}{a+b}\frac{a}{a+c}} \quad (4)$$

As can be seen from the definition, the larger the values of these indices, the more similar are *C* and *P*. Specifically, the values of both the Rand index and the Jaccard coefficient are in the range of [0, 1]. The major difference between these two statistics is that the Rand index emphasizes the situation that pairs of samples belong to the same group or different groups in both *C* and *P*, but the Jaccard coefficient excludes *d* in the similarity measure.

III. MICROARRAY DATA CORRECTION

Oligonucleotide Microarray technology offers the possibility of simultaneously monitoring thousands of hybridization reactions. These arrays show high potential for many medical and scientific applications as gene expression monitoring, sequence analysis, and genotyping. Nevertheless microarrays are exposed to errors during manufacturing, similar to silicon circuit electronics and the hybridization process may be contaminated by different reasons. Other sources of errors are optical noise during scanning and processing, or to interactions between molecular structures and light, dispersion among others. To reduce some of these effects replicates of experiments are used at the cost of increasing expenses. In order to detect noise contamination in microarray data, statistical tools like Independent Component Analysis (ICA) can be used. ICA allows us to better understand data in complex and noisy environments. It can separate the patterns in which we are interested from independent other effects like random sample variations or biological patterns unrelated to the subject of investigation. The technique has the potential of significantly increase the quality of the resulting data, and improve the biological validity of subsequent analysis [8].

TABLE I
NUMBER OF ICA CORRECTED UNRELIABLE PROBES

Samples	Genes	Unreliable	Very unreliable	Corrected unreliable	Corrected very unreliable
B19R	22283	15613	1088	4320	677
B19T	22283	14980	599	4249	349
C76R	22283	15076	842	4112	495
C76T	22283	14590	615	4124	378

Microarray data correction [9] was done based on the analysis of the correlation coefficient between Perfect Match and MisMatch samples within an microarray experiment. The correlation coefficient, γ is given by the following relation:

$$\gamma = 1 - cos^2 \beta \tag{5}$$

where β is the angle between the multidimensional vectors corresponding to the Perfect Match (PM) and MisMatch (MM) samples.

This parameter may be used to provide information on which probe set results were produced from normal hybridization processes, in contrast with those which may be produced by corrupted hybridization. This can help in improving the estimation reliability of microarray data prior to their use in clustering and pattern recognition. The number of unreliable gene probe sets found in a particular microarray may be quite large, thus meaning that many probe tests may have been affected by corruption processes. These probe sets were re-aligned by detecting their independent components by ICA, and re-estimating the PM-MM pairs from the independent components found.

For the microarray data correction we used the Chowdary database [10] consisting of 104 samples of breast and colon tissues on Affymetrix Human Genome U133A Array. For each sample the array contains 22238 genes and after computing the correlation coefficient, from an average number of 14603 unreliable labeled probe sets, $0.1 < \gamma < 0.5$, only an average number of 4115 probe sets showed an improving, this means a new computed $\gamma < 0.1$; and from an average number of 779 very unreliable labeled probe sets, $\gamma > 0.5$ the average number of corrected data is 474, as shown in Table I.

IV. HYBRID DATA BICLUSTERING

The corrected data is next transferred to a gene expression matrix that will be analyzed in order to extract some knowledge about the underlying biological processes. For the microarray data, one method of clustering is clustering the two-way clustering in which both the samples and the genes are grouped in the same time. This method, also known as biclustering is done usually using the same algorithm.

In our work we propose a method of hybrid biclustering by combining the previous mentioned algorithms: k-means and EM.

Before clustering the corrected data was filtered in order to reduce the dimension of the data and to eliminate the genes that do not show any interesting changes during the experiment. After filtering the genes with small variance over time and those that have very low absolute expression values a total number of 182 genes were used for clustering.

TABLE II
NUMBER OF CLUSTERS FOR GENE CLUSTERING BY EXTERNAL VALIDATION

Index	k-means algorithm	EM algorithm
Rand	2	3
Jaccard	2	2
Fowlkes-Mallows	2	2

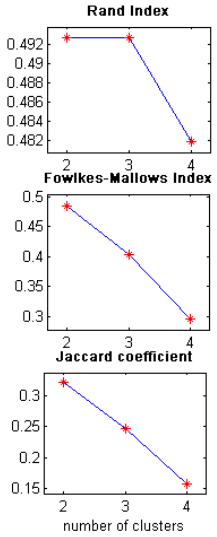

Fig. 1. Values of the external indexes determined for the k-means algorithm.

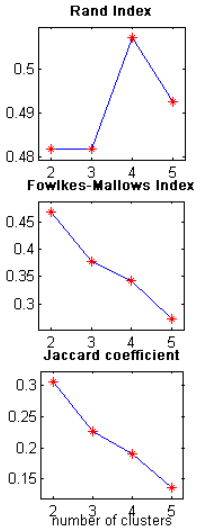

Fig. 2. Values of the external indexes determined for the EM algorithm.

After filtering we applied the cluster validation method with the external technique. For each index if the value is closer to 1 this means that the number of clusters is the one that it is expected to be. The results obtained are shown in Fig. 1 for the k-means algorithm and in Fig. 2 for the EM algorithm, and confirms that we have to classes of samples.

Once we established the optimal number of clusters for the samples we proceed in a similar manner for the genes and we obtained the results from Table II.

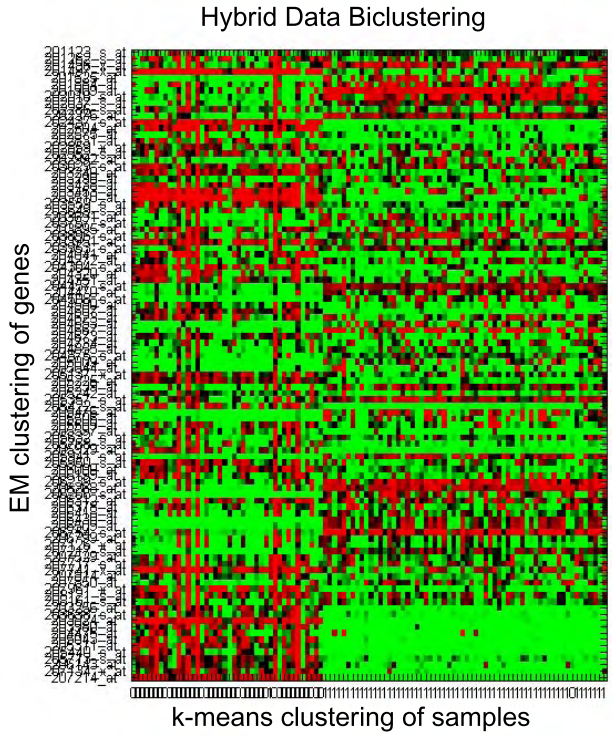

Fig. 3. Hybrid clustering of data using k-means algorithm for the samples and EM for the genes. On the ordinate axis there are the gene's labels and on the abscissa axis there are the samples labeled with 1 and 0.

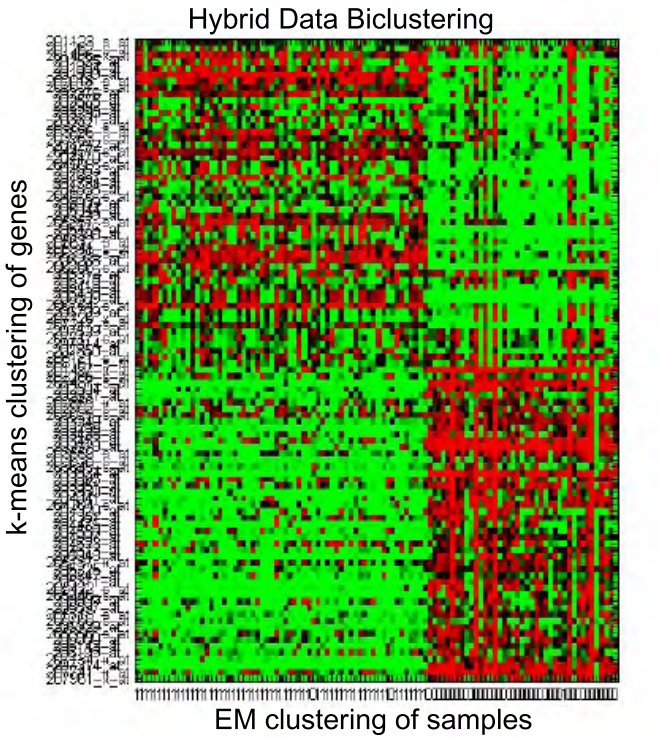

Fig. 4. Hybrid clustering of data using EM algorithm for the samples and k-means algorithm for the genes. On the ordinate axis there are the gene's labels and on the abscissa axis there are the samples labeled with 1 and 0

Considering the results from the external validations methods we applied the hybrid bilcustering on the data using first a k-means clustering of the samples and an EM clustering for the genes. We were able to group almost all the samples which belong a class to its class and we obtained to homogenous groups of genes. The results for this approach can be seen in Fig. 3.

We apply the same method but reversing the algorithms and this time the accuracy of the results was higher when the samples were clustered and we obtained more compact groups of genes as it can be seen in Fig. 4.

V. Conclusions

The analysis of the microarray data was done using to different methods with different purposes. Firstly we used Independent Component Analysis as a technique powerful enough to specifically correct deviations produced by unknown factors by extracting them and using their trace to be removed from the observations, and then we combined known clustering techniques for a two-way clustering of the data in order to classify the samples into their classes and to obtain groups of genes with similar behavior. The clustering methods were validated by some external indexes, which indicate a two class clustering. Still, there is left an internal validation of the clusters obtained. Regarding the clustering algorithm, a useful classification was obtained when EM clustered the genes and k-means the samples. Other supervised and unsupervised methods are planned to be used in a hybrid approach to microarray data.

References

[1] S. González, L. Guerra, V. Robles, J. M. Peña, F. Famili, "CliDaPa: A new approach to combining clinical data with DNA microarrays", *Intelligent Data Analysis Journal, Issue: Knowledge Discovery in Bioinformatics*, vol. 14(2), pp. 207–223, Jan. 2010

[2] K. Y. Yip, L. Cheung, D. W. Cheung, L. Jing, M. K. Ng, "A semi-supervised approach to projected clustering with applications to microarray data", *International Journal of Data Mining and Bioinformatics*, vol. 3(3), pp. 229-259, 2009

[3] A. Hyvarinen, J. Karhunen,, E. Oja, *Independent Component Analysis*, John Willey & Sons, 2001

[4] A. K. Jai, M. N. Murty, P. J. Flynn, "Data Clustering : A Review", *ACM Computing Surveys*, vol. 31(3), September 1999

[5] T. K. Moon, "The Expectation Maximization Algorithm", *IEEE Signal Processing Magazine*, November 1996

[6] M. Halkidi, Y. Batistakis, M. Vazirgiannis, "Cluster Validity methods: Part I", *SIGMOD Record*, vol. 31(2), pp. 40-45, 2002

[7] A. Jain, R. Dubes, *Algorithms for clustering data*, Prentice Hall, Englewood Cliffs, NJ, 1988

[8] S.I. Lee, S. Batzoglou, "Application of independent component analysis to microarrays", *Genome Biology*, vol. 4, pp. R76.1 - R76.21, 2003

[9] R. Malutan, P. Gómez, M. Borda, "Independent component analysis algorithms for microarray data analysis", *Intelligent Data Analysis Journal, Issue: Knowledge Discovery in Bioinformatics*, vol. 14(2), pp. 193–206, Jan. 2010

[10] D. Chowdary, J. Lathrop, J. Skelton, K. Curtin *et al.*, "Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative", *J Mol Diagn*, vol. 8(1), pp. 31-39, Feb 2006