A hybrid unsupervised and supervised clustering applied to microarray data

Raul Măluţan, Pedro Gómez Vilda, Monica Borda

***Abstract*— Clustering methods have been often applied to large data with the main purpose of reducing the dimension, time computation and identifying clusters with similar behavior. This work presents a state-of-the-art in unsupervised clustering and cluster validation. It proposes a method for hybrid bi-clustering of microarray data combined with a supervised validation for determining the optimal amount of clusters of genes.**

***Keywords* — microarray, clustering, internal validation, external validation, gene shaving.**

I. Introduction

The microarray data processing challenges nowadays consists of how to make it more reliable, easy to use and efficient. In this task other fields of knowledge as Signal and Image Processing, Pattern Recognition, or Statistical Data Analysis [1] may help in yielding 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, etc.

Clustering techniques play an important role in analyzing high dimensional data such as microarray data. An analysis of microarray data is a search for genes that have similar, correlated patterns of expression. This indicates that some of the data might contain redundant information. For example, if a group of experiments were more closely related than it was expected, some of the redundant experiments can be ignored, or use some average of the information without loss of information.

Data analysis methods [2] can be grouped in two categories: supervised and unsupervised. In the unsupervised approach, also known as clustering, data is organized without a priori information.

This paper describes several unsupervised algorithms used mostly for microarray data, like k-means, Partitioning Around Medoids (PAM) and Expectation-Maximization (EM). These algorithms have proven to be a useful when the number of clusters is known or can be estimated. Two of them, k-means and PAM, are based on minimizing the mean squared error, while the third, EM, on mixture modeling.

The algorithms were run on several data sets [3], observing that the quality of the obtained clusters is dependent on the number of clusters specified. To assess the effectiveness of these algorithms, but also to estimate the actual number of clusters, several clustering validation method were implemented. These methods consist in calculating internal and external indices used to estimate the optimal number of clusters for each algorithm.

The microarray data has a particularity compared with other type of data. A two dimension data is in fact a gene expression matrix which usually has 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 a similarity measure between the objects, genes or experiments under comparison, has to be used. Mostly the same algorithm is used for analyzing both the genes and the experiments, *i.e.* bi-clustering of microarray data.

Considering the dimensionality of the data, large amount of genes and small number of experiments, we propose in this paper a hybrid bi-clustering, where we combine unsupervised methods with the purpose of obtaining the optimal combination of clusters.

Besides, as a supplementary validation, a supervised method was used on the same data. Supervised classification represents the issue of identifying the subset to which new observations belong, where the identity of the subset is unknown, on the basis of a training set of data containing observations whose subset is known. Therefore the classification will display a variable behavior which can be analyzed by statistics. It is required for new sample items to be placed into the respective groups based on quantitative information on one or more measurements, attributes or features and based on the training set in which previously decided groupings are already established.

Manuscript received Ortober 28, 2012, revised March 04, 2013. 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).

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).

doi: 10.11601/ijates.v2i3.21

II. Cluster Analysis

A. Unsupervised algorithms

The unsupervised clustering methods are divided in two categories: hierarchical and non-hierarchical. Hierarchical methods group the objects in an iterative way, generating a hierarchical tree structure, also known as dendogram. On the other hand, the non-hierarchical clustering, or partitioning methods, does the partitioning of a data set into a predefined number of different clusters, without a hierarchical structure. Beside, these algorithms produce an integer number of partitions, and also they optimize a certain criterion function. The partitioning methods classify the data in *k* clusters which must fulfill the following conditions:

– each group should contain at least one element;

– each object must belong to one group only.

For the microarray data the most suitable clustering methods are unsupervised ones, because we cannot observe the (real) number of clusters in the data [4]. From the unsupervised algorithms used in microarray data analysis the k-means, PAM (Partitioning Around Medoids) and EM (Expectation Maximization) are the most frequently used. The first two from the list are using the minimization of the root mean square error and the third one is using the mixing models methods.

The PAM algorithm [2] is based on the search for *k* representative objects or medoids among the objects of the dataset. These objects should represent the structure of the data. After finding a set of *k* medoids, *k* clusters are constructed by assigning each object to the nearest medoid. The goal is to find *k* representative objects which minimize the sum of the dissimilarities of the objects to their closest representative object. The algorithm first looks for a good initial set of medoids. Then it finds a local minimum for the objective function, that is, a solution such that there is no single switch of an object with a medoid that will decrease the objective.

The k-means algorithm [5], 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 [6] 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.

B. Unsupervised clustering validation

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 [7], [8]: *external validation*, based on previous knowledge about data and *internal validation*, based on the information intrinsic to the data alone.

There are three external indexes which were used in our previous study [3], Rand index, Jaccard coefficient, and Fowlkes and Mallows index, and they are going to be used also in this paper.

In contrast to external validation, internal validation evaluates the clustering without any a priori information. The obtained values are known as internal indexes as they are computed from the data used for clustering. In this paper we evaluate for the microarray data the following internal indexes: silhouette, Calinski-Harabasz index, Krzanowski-Lai index, Hartigan index and Davies-Bouldin index.

Silhouette index calculates the silhouette width for each sample, average silhouette width for each cluster and overall average silhouette width for a total data set:

$$S(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}\,, \tag{1}$$

where *a(i)* is the average dissimilarity of *i*-object to all other objects in the same cluster; *b(i)* is the minimum of average dissimilarity of *i*-object to all objects in other cluster. The largest overall average silhouette indicates the best clustering.

Calinski-Harabasz index is defined by:

$$CH(k)=\frac{B(k)/(k-1)}{W(k)/(n-k)}\,, \tag{2}$$

where *k* denotes the number of clusters, and *B*(*k*) and *W*(*k*) denote the between and within cluster sums of squares of the partition, respectively. An optimal number of clusters is then defined as a value of *k* that maximizes *CH*(*k*).

Krzanowski-Lai index is given by the equation:

$$KL(k)=\left|\frac{DIFF(k)}{DIFF(k+1)}\right|\,, \tag{3}$$

where $DIFF(k)=(k-1)^{2/p}W(k-1)-(k)^{2/p}W(k)$ and *p* denotes the number of features in the data set. A value of *k* is optimal if it maximizes *KL*(*k*).

Davies-Bouldin index is a function of the ratio of the sum of within-cluster scatter to between-cluster separation:

$$DB=\frac{1}{n}\sum_{i=1}^{n}\max_{i\neq j}\left\{\frac{S_n(Q_i)+S_n(Q_j)}{S(Q_i,Q_j)}\right\}, \tag{4}$$

where *n* is the number of clusters, S_n is the average distance of all objects from the cluster to their cluster centre, $S(Q_i,Q_j)$ is the distance between clusters centres. Consequently, Davies-Bouldin index will have a small value for a good clustering.

For the review, the manuscript is submitted to the IJATES^2 Editorial Board only electronically in PDF format. Do not change any formatting; otherwise you undergo the risk to be directly rejected without review process.

C. Gene Shaving

The method of Gene Shaving is designed to extract coherent and typically small clusters of genes that vary as much as possible across the samples. According to [9], the algorithm consists of the following steps:

1. Start with the entire expression matrix X, each row centered to have zero mean
2. Compute the leading principal component of the rows of X
3. Shave off the proportion (typically 10 %) of the genes having smallest absolute inner-product with the leading principal component
4. Repeat steps 2 and 3 until only one gene remains
5. This produces a nested sequence of gene clusters $S_N \supset S_k \supset S_{k_1} \supset S_{k_2} \supset \cdots \supset S_1$, where S_k denotes a cluster of k genes. Estimate the optimal cluster size $\hat{k}$ using the gap statistic [10].
6. Orthogonalize each row of X with respect to $\bar{x}_{S_k}$, the average gene in $S_{\hat{k}}$
7. Repeat steps 1-5 above with the orthogonalized data, to find the second optimal cluster. This process is continued until a maximum of M clusters are found, where M is chose a priori.

When implementing this method we made some changings compared with the steps form [9]. But I have made some changes in the algorithm. First of all, we shaved off not *α %* genes, but 1 gene each time. This is because we will lose the precision of the algorithm if using *α%*. For example, supposing we remain with S_k clusters with k being 135, 122, ..., 53, 47, etc. genes, and according to the gap statistic step, the algorithm decides to have a cluster with 52 or 47 genes, which is not correct. This is the reason why we have decided to have clusters with ..., 53, 52, 51, 50, 49, 48,... genes, in this way hoping to obtain a maximum *Gap(k)* closer to ideal 50. Also when computing the gap statistics we have made some changes. If we consider all possible permutations and after that finding each *D(k)*, then, in case of 150 genes with 4 characteristics of each gene, it is required to have all possible permutations of the matrix by permuting the elements within each row. This means that each row has from 4 parameters a number of 24 possible permutations, this implies that we have another sets of 24150 matrices, which is too much (we cannot take for example 3rd permutation from gene 1 with 3rd permutation for gene 2, 3rd permutation for gene 3, and so on, because we obtain the same *D(k)*). We have tried to consider random matrices, a number of 5000, but the problem in this case is that the result is varying at every other analysis and they are also different at each simulation. Finally we have decided to use just the first input matrix and get its *D(k)* and *Gap(k)*, without any permutations.

III. Data Clustering

A. Unsupervised clustering

Three unsupervised algorithms were used to cluster microarray data. We used for our study two different datasets from public Affymetrix databases. The first set was the Chowdary database [11] the authors compared pairs of snap-frozen and RNA later preservative-suspended tissue from 62 lymph node-negative breast tumors and 42 colon tumors, with purpose of separating them. The second set [12] contains 24 acute lymphoblastic leukemia (ALL), 28 acute myelogenous leukemia (AML) and 20 mixed-lineage leukemia (MLL) samples.

For the microarray data the clustering was done by a two-way clustering or bi-clustering [13] in which both the samples and the genes are clustered in the same time using the portioning method.

Regarding the conclusions from [3] a useful classification was obtained for microarray data when EM clustered the genes and k-means the samples. In this study we will use the k-means algorithm to cluster the samples and the PAM and respectively the EM algorithm to cluster the genes. Before combining the algorithms we applied the clustering validation methods, both external and internal indexes.

Also, based on these results we combined the unsupervised method with a supervised one. So we clustered the samples by k-means algorithm and we classify the genes using the gene shaving method.

In Table I, the numbers of clusters obtained after using the internal and external indexes are indicated. For the k-means algorithm the number refers to sample partitioning, while for the PAM and EM algorithms the numbers refers to gene clustering. In the EM clustering validation only the external indexes were computed.

TABLE I
The number of clusters obtained with the clustering validation methods

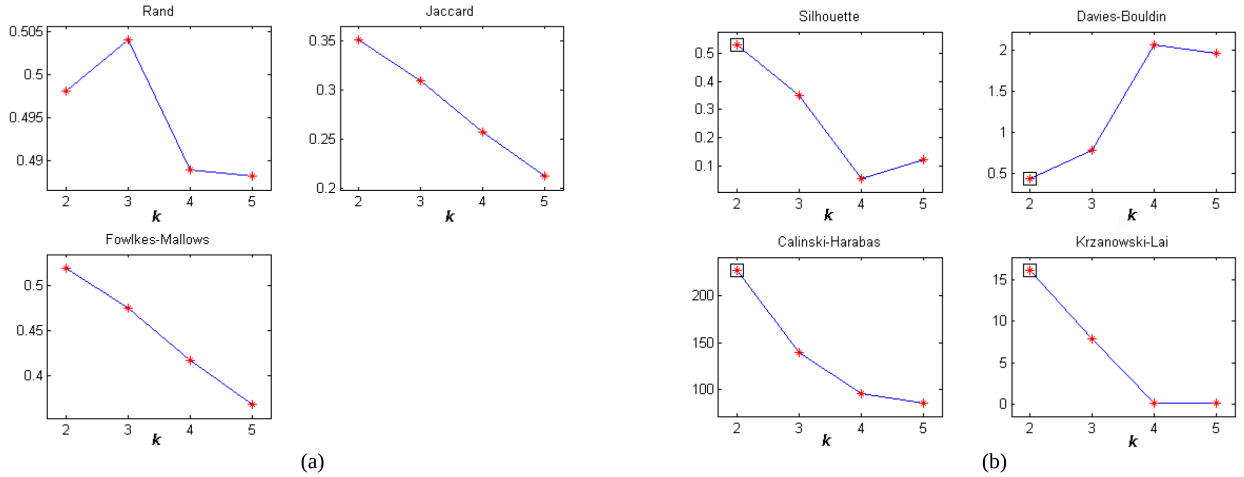
Index	Chowdary database			Leukemia database		
	k-means	PAM	EM	k-means	PAM	EM
Rand	2	3	3	3	3	4
Jaccard	2	2	2	3	3	3
Fowlkes-Mallows	2	2	2	3	3	3
Silhouette	2	2	-	2	3	-
Calinski-Harabasz	2	2	-	3	3	-
Krzanowski-Lai	2	2	-	3	3	-
Davies-Bouldin	2	2	-	2	3	-

After the validation was done we applied the combined clustering for the datasets. Thereby, for the Chowdary dataset the genes were clustered into 2 groups, with the PAM method, Fig 2.a, and then with the EM algorithm, Fig.2.b. The samples were clustered with the k-means algorithm.

The obtained values were compared with the given values from the microarray databases and a similarity between these values was observed.

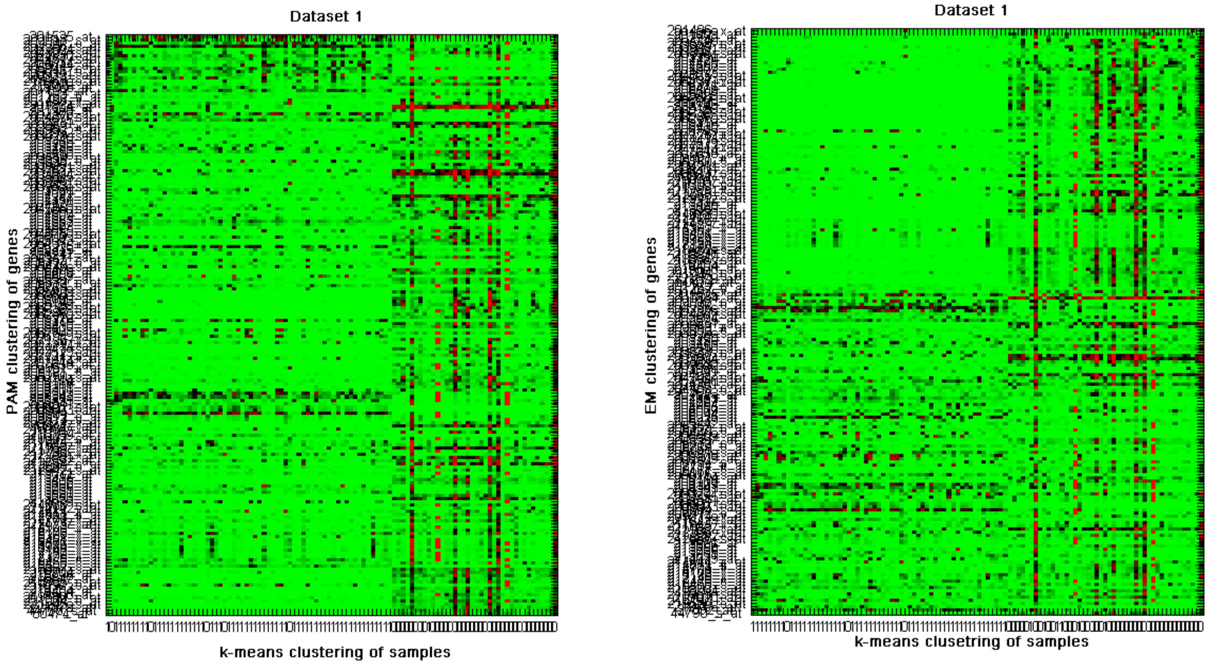
Fig. 1.a and 1.b shows all the computed indexes for the Chowdary database, internal and external, in the case of PAM algorithm. For the external indexes the highest values obtained gave the optimal number of clusters. In the case of internal indexes the optimal value was marked by a square in Fig. 1.b.

According with the optimal number of clusters indicated by the validation indexes, in the case of the leukemia dataset the clustering was done in 3 clusters. Thus the samples were group into three sets by the k-means algorithm, while the genes formed also three groups once with the PAM method



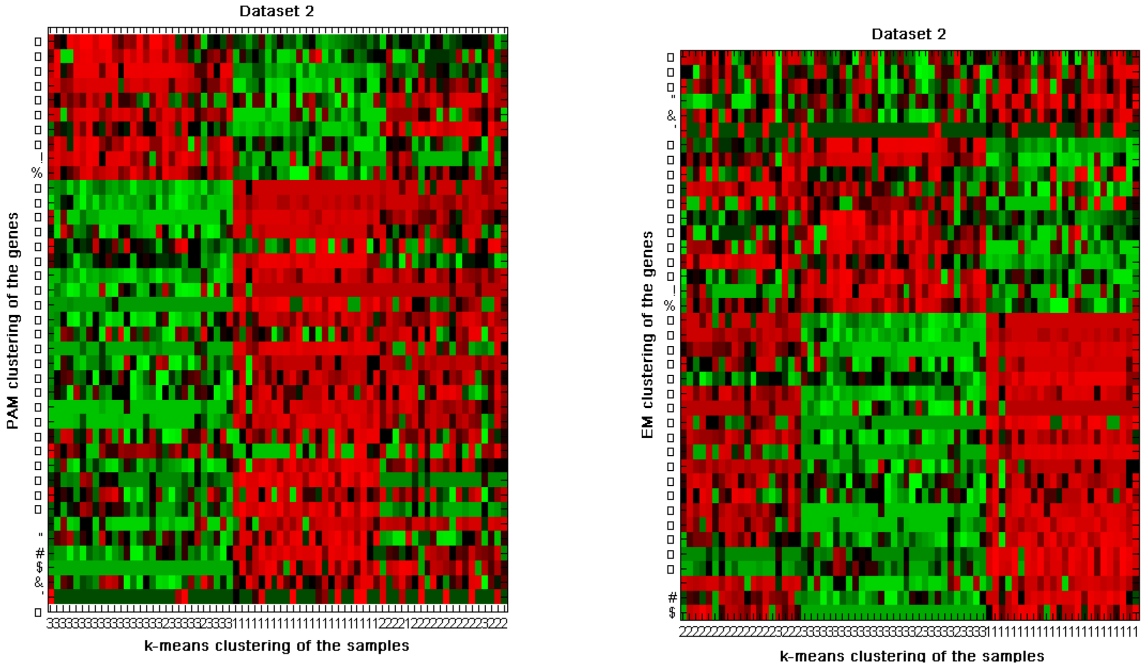
(a) (b)

Fig. 1. Values of the (a) external indexes, (b) internal indexes determined for PAM clustering of genes for Chowdary database



(a) (b)

Fig. 2. Microarray data clustering by the unsupervised algorithms: (a) PAM clustering, (b) EM clustering of the genes for the Chowdary dataset.



(a) (b)

Fig. 3.a Microarray data clustering by the unsupervised algorithms: (a) PAM clustering, (b) EM clustering of the genes for the leukemia dataset.

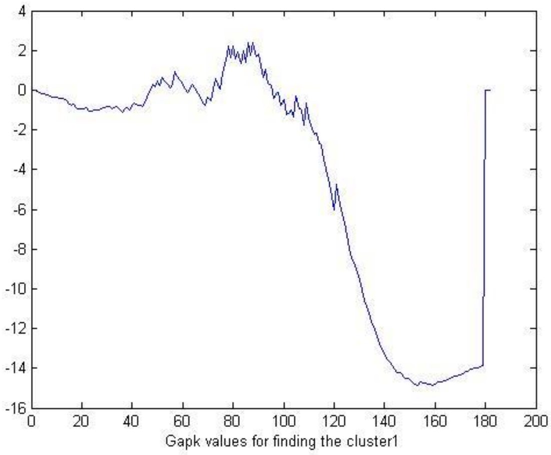


Fig. 4 The values of the Gap(k) function for the Chowdary dataset. The maximum was obtained for a number of 89 genes.

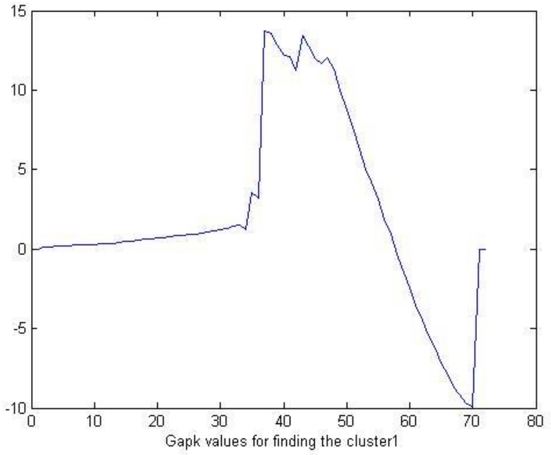


Fig. 5 The values of the Gap(k) function for the leukemia dataset. For the first cluster the maximum was obtained for a number of 37 genes

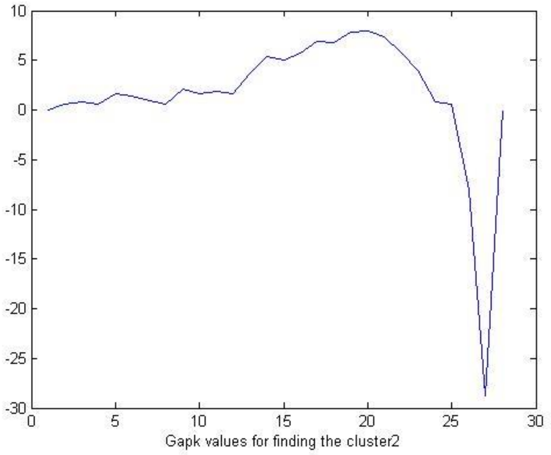


Fig. 6 The values of the Gap(k) function for the leukemia dataset. For the second cluster the maximum was obtained for a number of 21 genes

and then with the EM algorithm, as it can be seen in Fig. 3.a and Fig. 3.b.

For the gene shaving algorithm we used the same number of clusters as the one given by the validation methods. In the case of Chowdary dataset the genes were classified in 2 sets, while in the case of leukemia dataset we applied gene shaving with 3 clusters.

Once applying the gene shaving method we were able to classify the Chowdary dataset in 2 clusters. The number of gene in the first cluster was determined by using the gap statistic method. The value of 89 genes was given by the maximum of the graph from Fig. 4.

In a similar manner we computed the gap statistics for the leukemia dataset and we were able to classify the data into 3 clusters: the first one with 37 genes, the second one with 21 genes and the third one with the remaining 14 genes from a total of 72 genes. The values of the Gap function for this dataset are shown in Fig 5 and 6.

IV. Conclusion

In this paper some unsupervised and supervised clustering methods were applied to microarray dataset in order to obtain a bi-clustering of the data with different algorithms. The selection of the PAM algorithm for clustering the genes, despite of the k-means method was done because of the robustness of the first one. The EM algorithm showed some disadvantages when working with large datasets that is way the data was previously filtered.

Gene shaving, a supervised method was also combined with k-means in order to classify the genes according with the number of clusters given by the validation methods.

For cluster validation we used both internal and external methods by computing some indexes which gave us the optimal number of clusters for each clustering method. As future work we take into consideration the combination of these methods with other data mining algorithms like Independent Component Analysis, which can be used for large datasets.

References

[1] S. González, L. Guerra , V. Robles, JM. Peña, F. Famili, “CliDaPa: A new approach to combining clinical data with DNA microarrays” *Intelligent Data Analysis Journal*, vol 14(2), pp. 207 – 223, 2010

[2] J. Han, M. Kamber, *Data mining: Concepts and techniques*. Morgan Kaufmann, 2000

[3] R. Malutan, B. Belean, P.G. Vilda, M. Borda, “Two way clustering of microarray data using a hybrid approach” in *Proc. of 34th Int. Conf. on Telecommunications and Signal Processing*, Budapest, 2011, pp. 417 - 420

[4] G. McLachlan, K-A. Do, C. Ambroise, *Analyzing Microarray Gene Expression Data*. Wiley-Interscience, 2004

[5] M.B. Zoubi, A. Hudaib, A. Huneiti, B. Hammo, “New efficient strategy to accelerate k-means clustering algorithm” *American Journal of Applied Sciences*, vol 5(9), pp. 1247 – 1250, 2008

[6] G. McLachlan, T. Krishnan, *The EM Algorithm and Extensions*. John Willey & Sons 2008

[7] N. Bolshakova, F. Azuaje, “Cluster validation techniques for genome expression data” *Signal Processing* vol 83, pp. 825 – 833, 2002

[8] K. Wang, B. Wang, L. Peng, “CVAP: Validation for Cluster Analyses” *Data Science Journal* vol 8, pp. 88 – 93, 2009

[9] T. Hastie *et. al.*, “Gene shaving as a method for identifying distinct sets of genes with similar expression patterns”, *Genome Biology*, vol. I(2), research 0003, pp. 1-21, 2000

[10] W. L. Martinez, A. R. Martinez, *Exploratory Data Analysis with MATLAB*, CRC Press LLC, 2005

[11] D. Chowdary *et al*, “Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative” *J Mol Diagn* vol 8(1), pp. 31 – 39, 2006

[12] S. Armstrong *et al*, “MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia” *Nature Genetics* vol 30, pp. 41 – 47, 2001

[13] S. C. Madeira, A. L. Oliveira, “Biclustering algorithms for biological data analysis: A survey” *IEEE/ACM Transactions on Computational Biology and Bioinformatics* vol 1 (1), pp. 24-45, 2004