Advanced Engineering Forum *Online: 2013-06-27*
ISSN: 2234-991X, Vols. 8-9, pp 508-515
doi:10.4028/www.scientific.net/AEF.8-9.508

Combined Clustering Methods for Microarray Data Analysis

Raul Măluţan[1, a], Pedro Gómez Vilda[2] , Monica Borda[1]

[1] Communications Department, Technical University of Cluj-Napoca, 26-28 George Baritiu St., 400027 Cluj-Napoca, Romania,

[2] Departamento de Arquitectura y Tecnología de Sistemas Informáticos (DATSI), Universidad Politécnica de Madrid, Campus de Montegancedo, s/n, 28660, Boadilla del Monte, Madrid, Spain

[a]raul.malutan@com.utcluj.ro (corresponding author)

Keywords: microarray, clustering, classification, gene shaving, supervised methods.

Abstract. Data classification has an important role in analyzing high dimensional data. In this paper Gene Shaving algorithm was used for a previous supervised classification and once the cluster information was obtained, data was classified again with supervised algorithms like Support Vector Machine (SVM) and k-Nearest Neighbor (k-NN) for an optimal clustering. These algorithms have proven to be useful when the classes of the training data and the attributes of each class are well established. The algorithms were run on several data sets, observing that the quality of the obtained clusters is dependent on the number of clusters specified.

Introduction

During the last years the development of microchips capable of processing biological information was enormous. Today it is possible to retrieve enormous amounts of cellular information in a cost and time effective manner. Microarrays make it possible to get the expression levels of many thousands of genes at once which open a completely new field for researchers of many disciplines, such as computer science, statistics, biology and medicine [1]. The result of microarray processing can be an overwhelming set of files with too much amount of data which may cause problems because they will be difficult to process, separate, classify and correlate for the results to be extracted. That is why optimal algorithms for classification must be selected.

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. The classification issue is also known as supervised learning. On the other hand, the unsupervised learning is represented by clustering, where the problem is to analyze a single data-set and decide the method of dividing the observation into groups.

Cluster Analysis

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 [2], 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 [3].
6. Orthogonalize each row of X with respect to $\bar{x}_{S_{\hat{k}}}$, 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 [2]. 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.

Support Vector Machine (SVM) is a supervised learning algorithm. It is based on the concept of decision planes that define decision boundaries. A decision plane makes a separation between a set of objects having different class memberships.

A linear classifier uses a line to separate a set of objects into their respective groups. Most classification tasks, however, are not so easy to achieve, and more complex structures are adopted in order to make an optimal separation. The purpose is to classify new objects (test cases) based on examples that are available (train cases) as correctly as possible. Classification tasks based on drawing separating lines to distinguish between objects of different class memberships are known as hyperplane classifiers. Support Vector Machines are particularly suitable for this kind of tasks.

Considering the linear classifier case, one category of elements is in the lower left corner and the cases with the other category are in the upper right corner, the cases being completely separated. The SVM analysis is meant to find a dimensional hyperplane which acts as a separator between the cases based on their target categories. There are an infinite number of possible lines, but an optimal line has to be defined. [4]

Taking into consideration the non-linear classifier case, the original objects can be mapped (rearranged), using a set of mathematical functions, known as kernels. The process of rearranging the objects is known as mapping (transformation). In this new setting, the mapped objects are linearly separable, therefore, instead of constructing the complex curve, one has to find an optimal line for separation.

k-nearest neighbor algorithm (kNN) is a method for classifying objects based on closest training examples in the feature space. The classification was developed from the need to perform discriminant analysis when reliable parametric estimates of probability densities are unknown or difficult to determine.[5] The kNN classifier operates considering the idea that classification of unknown instances can be done by linking the unknown data to some known data using a distance (similarity) function. Two instances situated far apart in the instance space, defined by the appropriate distance function, are less likely to belong to the same class than two closely situated instances. The kNN classification is one of the basic and most simple classification methods and is especially recommended when there is little or no prior knowledge about the distribution of the data.

Data Analysis

Supervised Classification. The above described methods were combined in order to strongly confirm a good classification of microarray data. The data used in our study came from two different public Affymetrix databases. In the first one, the Chowdary database [6], 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 classifying them. The second database used [7] contains 24 acute lymphoblastic leukemia (ALL), 28 acute myelogenous leukemia (AML) and 20 mixed-lineage leukemia (MLL) samples.

Regarding the conclusions from [8] for the Chowdary dataset the genes must be classified into 2 groups, while the optimal number of clusters indicated by the validation indexes, in the case of the leukemia dataset is 3.

First the gene shaving method was applied and we were able to classify the Chowdary dataset in 2 clusters. The number of genes 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. 1.

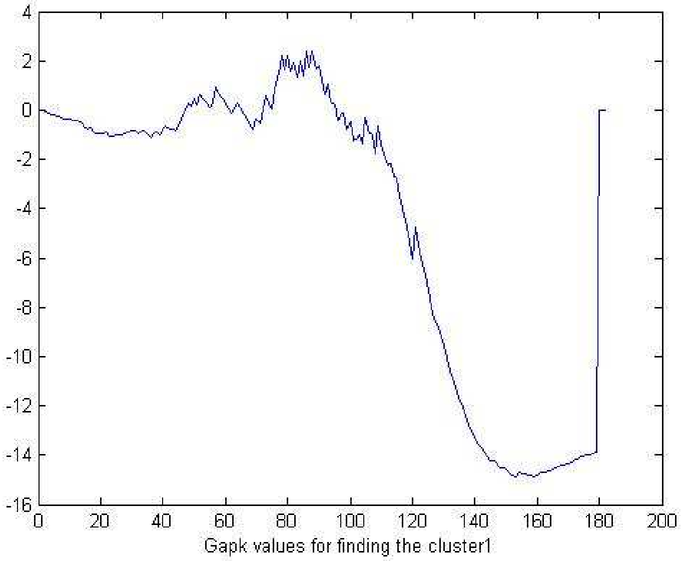

Fig. 1 *The values of the Gap(k) function for the Chowdary dataset. The maximum was obtained for a number of 89 genes.*

Once the number of elements within each cluster was determine by the gene shaving method we applied the SVM method to this database. The SVM algorithm has proved to work very well when the training data and the samples where linearly separable. This can be easily observed if applying it on the database and evaluating the figures, but the correct classification rate also. The attributes were set and the correct classification rate is given as a percentage. In Fig. 2 and Fig. 3 the classification method with a linear kernel indicates a good classification for this database.

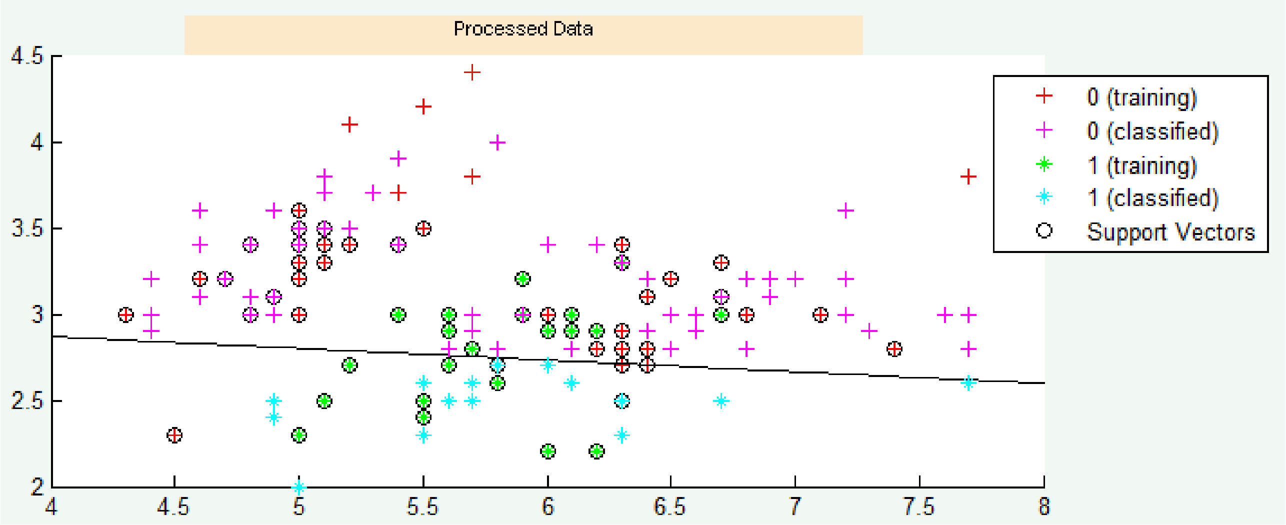

Fig. 2. *Linear kernel classification for lymph node-negative breast tumors*

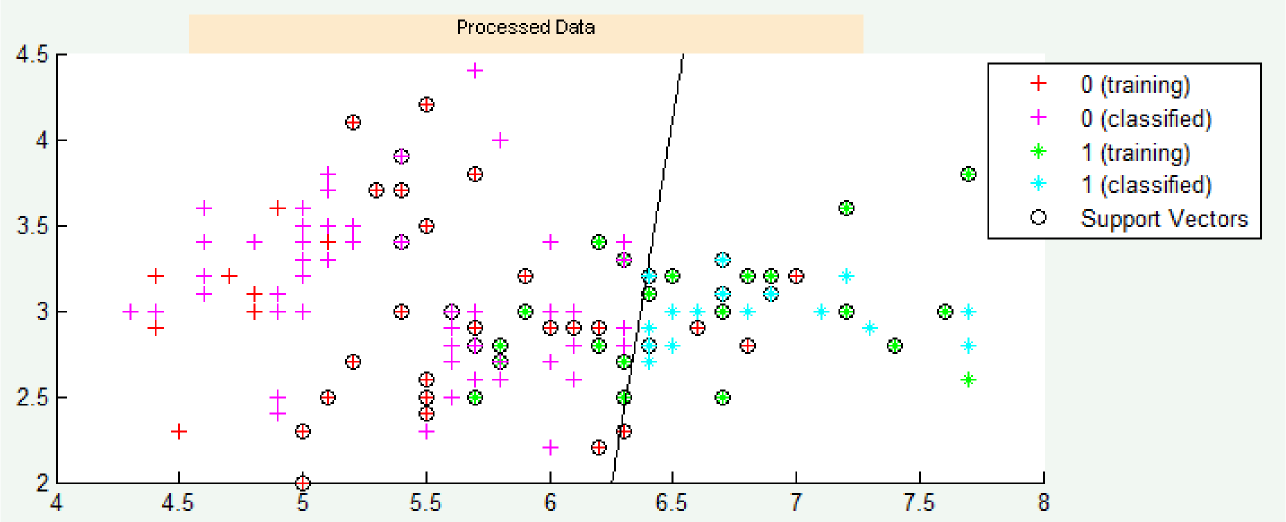

Fig. 3. *Linear kernel classification for colon tumors*

Also the high percentage rates from Table 1 and Table 2 prove that the lymph node-negative breast tumors are linearly separable from the colon tumors. Hence, in Fig. 2 and Fig. 3 there is a larger number of support vectors that are meant to maintain the hyperplane in a certain equilibrium. Usually, the less support vectors are present, the more separable the data is.

Table 1. *Sepparation coefficients for lymph node-negative breast tumors*

	Attribute 1	**Attribute 2**	**Attribute 3**	**Attribute 4**
Attribute 1	X	72	61.33	66.67
Attribute 2	72	X	73.33	64
Attribute 3	65.33	74.67	X	58.67
Attribute 4	60	77.33	60	X

Table 2. *Sepparation coefficients for colon tumors*

	Attribute 1	**Attribute 2**	**Attribute 3**	**Attribute 4**
Attribute 1	x	81.33	93.33	93.33
Attribute 2	80	x	92	94.67
Attribute 3	93.33	96	x	94.67
Attribute 4	97.33	94.67	97.33	x

For the leukemia database we computed the gap statistics 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. 4 and Fig. 5.

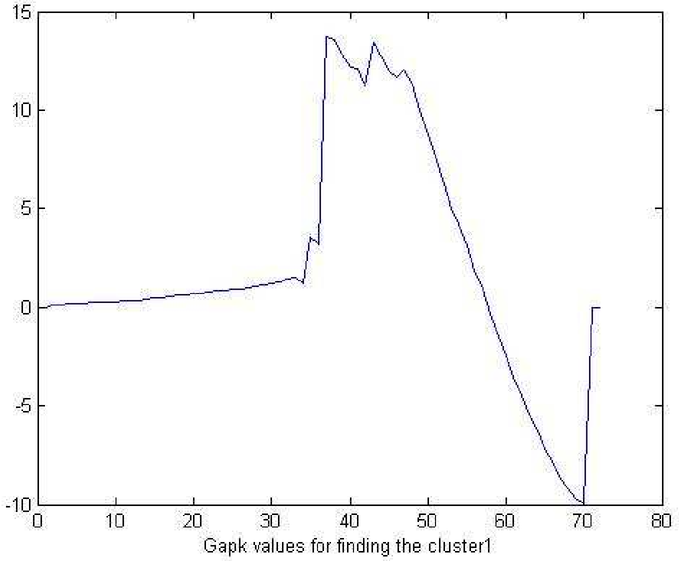

Fig. 4 *The values of the Gap(k) function for the leukemia dataset. For the first cluster the maximum was obtained for a number of 37 genes*

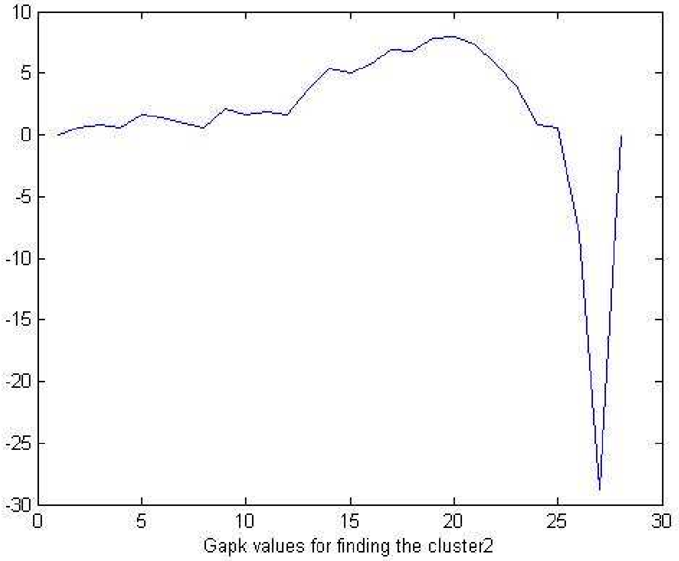

Fig. 5 *The values of the Gap(k) function for the leukemia dataset. For the second cluster the maximum was obtained for a number of 21 genes*

The *k* Nearest Neighbour algorithm has offered satisfactory results for the second database using different metrics. The training data for this method was obtained from the numerical attributes contained in the input files by using the values represented by the odd indexes in the matrices. The even values are considered to be the samples that have to be grouped.

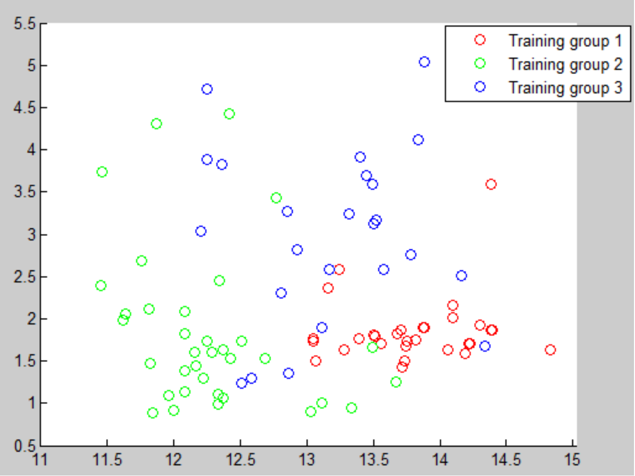

Fig. 6 *The training values for the leukemia database, grouped on classes*

The differences that occur when changing the distance option together with the number of nearest neighbors are quite noticeable. These can be illustrated in the leukemia database by applying 1-NN and 13-NN.

As it can be observed form Fig. 7 and Fig. 8 there is a difference between the Euclidean and cosine metric, both for 1-NN and 13-NN. In Fig. 6 one can easily spot the resemblances between the grouped training data and the grouped trained data, which means that the Euclidean metric has a high rate of correct classification. The differences between 1-NN and 13-NN occur because the boundaries between classes become less clear when the *k* value is increased. Even so, in this particular case, the groups still keep their initial structure.

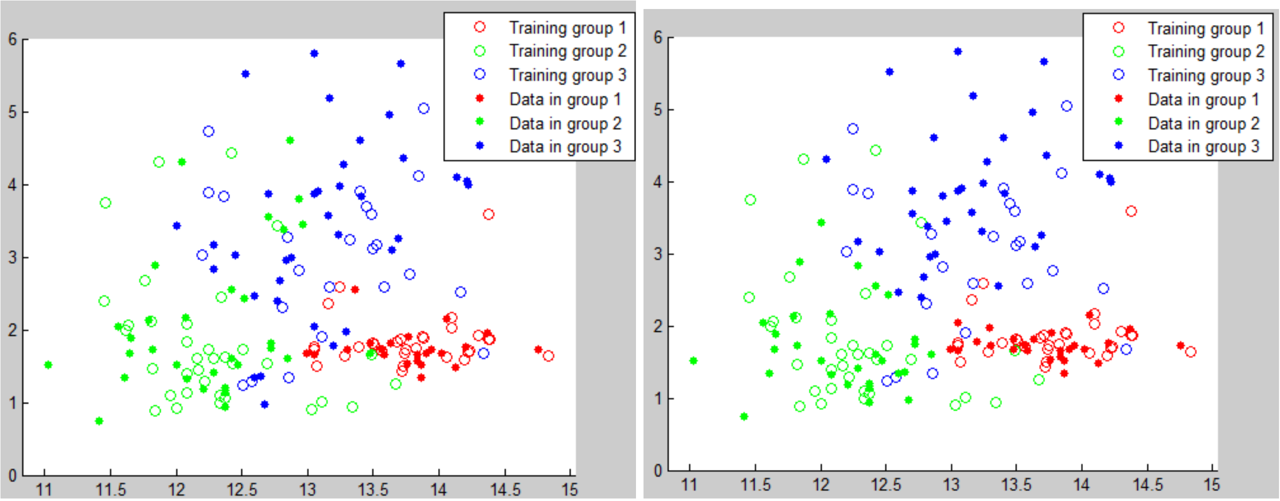

Fig. 7 *Comparison between the 1-NN (on the left) and 13-NN (on the right) using the Euclidean distance*

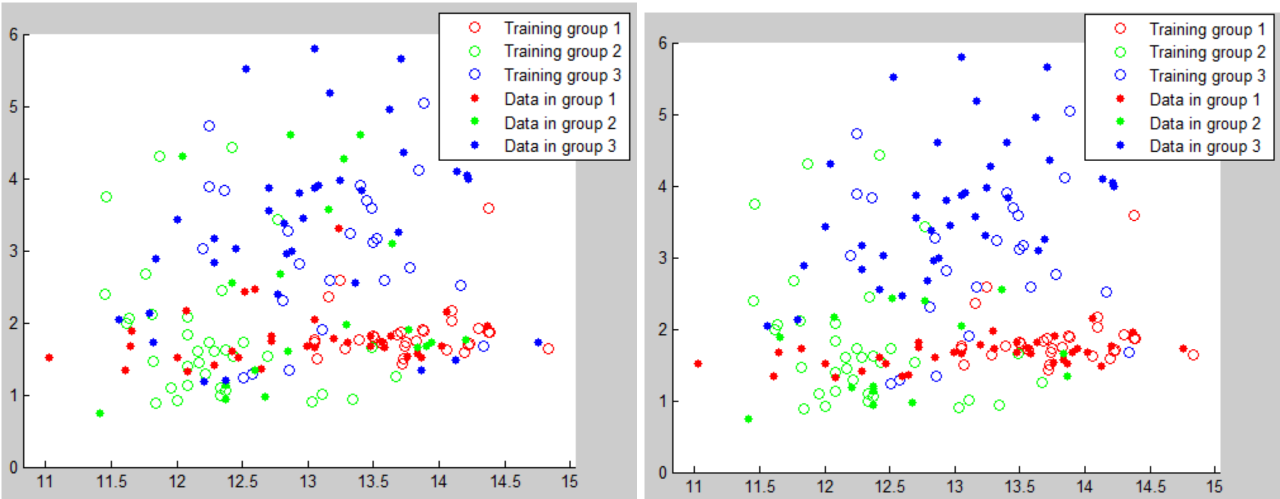

Fig. 8 *Comparison between the 1-NN (on the left) and 13-NN (on the right) using the Cosine distance*

Conclusions

In this paper some supervised clustering methods were applied to microarray dataset in order to efficiently classify predefined microarray data.

Validation methods applied to same database in early research confirmed the number of clusters for both databases. Later, gene shaving supervised method classified the data according with the number of clusters.

Each database responded in a better way to different supervised method, therefore were applied different methods. The SVM algorithm returns excellent results if the classes are linearly separable. On the other hand, the *k*-NN algorithm offers very good results on the second database when Euclidean metric was set, but seems not to be as accurate for the Cosine metric.

These algorithms are considered to be efficient due to their high correct classification rate. Because of their reliability, the next step towards improving these algorithms is to make them faster and more efficient in real-life situations. This is obtainable by developing a series of semi-supervised algorithms that do not need all the labels for classification.

Acknowledgment

This paper 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.

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, 14(2) (2010) 207 – 223.

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

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

[4] P. Sherrod, DTREG Predictive Modeling Software manual, (2001) 289-295

[5] G. Shakhnarovich, T. Darrell, P. Indyk (eds.), Nearest-Neighbor Methods in Learning and Vision: Theory and Practice (Neural Information Processing series), The MIT Press, 2006

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

[7] S. Armstrong *et al,* MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia, Nature Genetics, 30 (2001) 41 – 47

[8] 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, (2011) 417 - 420