Application of a New Algorithm, Iterative SOM, to the Detection of Gene Expressions

Marcos Lévano
Universidad Católica de Temuco
Escuela de Ingeniería Informática
Av. Manuel Montt, N 056, Casilla 15-D
Temuco - Chile
mlevano@uct.cl

Hans Nowak
Universidad Técnica Federico Santa María
Depto. Fisica
Av. España N 1680, Casilla 110-V
Valparaiso - Chile
hans.nowak@experimentos.cl

Abstract

DNA analysis by microarrays is a powerful tool that allows replication of the RNA of hundreds of thousands of genes at the same time, generating a large amount of data in multidimensional space that must be analyzed using informatics tools. Various techniques have been applied to analyze the microarrays, but they do not offer a systematic form of analysis. This paper proposes the use of the Self-Organizing Maps (SOM) in an iterative way to find patterns of expressed genes. The new method proposed (Iterative Self-Organizing Maps, ISOM) has been evaluated with up-regulated genes of the Escherichia Coli bacterium and is compared with the Self-Organizing Map (SOM) technique and a method which uses iteratively Gorban's Elastic Neural Net. In a comparative analysis of the three methods the ISOM shows the best results.

1. Introduction

Application of pattern research methods to the determination of global gene expressions in microarrays is today an important task. One of the most widely used method to determine groupings and select patterns in microarrays is the SOM (Self Organization Map) [1], [2]. The SOM defines a low dimensional manifold in the high dimensional data space of the gene expressions, in many cases a bidimensional net with nodes. Then this net will be deformed by a heuristical local optimization method and the high dimensional data is projected on the nearest nodes. Correlations between the data will be seen as data clusters on the nodes of the net which may be visualized and analized by statistical means.

The disadvantage of projecting the data only on the nearest nodes have been overcome by the Elastic Net algorithm (ENA) of Gorban [3],[4], where a low dimensional manifold is inserted into the multidimensional dataspace and then deformed by minimization of an energy functional which connects the data and the nodes by elastic forces. The projection of the data on the deformed net is not only on the nearest nodes but on the nearest points of the net. This leads to a better data and cluster distribution on the net. For the global gene expression of Escherichia coli growing on six different carbon sources [5], the ENA was used iteratively by the authors (Iterative Elastic Neural Net, IENN) [6], where the algorithm was applied multiple times to the resulting clusters. In this way a hierarchical structure of clusters were obtained and correlations between genes were found which were not present in the initial clusters. In order to find the optimal number of clusters in every iterative step, the k-Means method was used on the bidimensional data together with a quality index.

In this contribution an iterative SOM algorithm is proposed where the data is treated by a net with a large number of nodes so that the SOM applied to this net gives not too much projected data on one node but a good enough distribution of the data over the nodes. A k-Means method, adapted to data on nodes, together with quality indices is used to find an optimal number of clusters of nodes with their corresponding data. Then the SOM is applied again on every found cluster with a net of a large number of nodes and so on until the found clusters have well defined characteristics. This method is different from the hierarchical SOM [7] which repeats the SOM algorithm for the data on the nodes.

In order to compare the results of the iterative SOM with the iterative Elastic Neural Net and the SOM applied only once on a net with a small number of nodes, the Escherichia coli gene expressions on different carbon sources was analized. Receiver Operating Characteristic (ROC) [8] curves for the different cases were compared.

2. Methods

2.1. Self Organizing Maps

The Self Organizing Maps (SOM) technique is a neural net model capable to represent the topological structure of the initial data space in a discrete or continues form. It consists of a net made by a group of prototypes (weight vectors) which are associated to the neurons of the net.

The net is generated by establishing a correspondence between the input signals $x = [x_1, ..., x_n]^T, x \in \Re^n$ and the neurons which are represented by a weight vector. The input vectors, which come from a multidimensional space of the problem in question, are nonlinearely mapped and ordered on a regular array of neurons. The correspondence is obtained by a learning algorithm of competence which consist in a sequence of training steps that modify the weights of the neurons $m_i = [m_1^{(i)}, m_2^{(i)}, ..., m_n^{(i)}]^T, m_i \in \Re^n$, where i is the localization of the neuron in the net. A neuron, whose prototype is the nearest to the data x is called the Best Matching Unit (BMU) m_c and is obtained by the relation $||x - m_c|| = min_i||x - m_i||$, where $||..||$ is the distance measure.

During the learning proces a BMU is the winner and the net is changed in such a way that the BMU is desplaced in the direction of the input data vector x after the following adaption rule (1),

$$m_i(t+1) = m_i(t) + \alpha(t)h_{ci}(t)[x - m_i(t)], i = 1..M \quad (1)$$

where M is the number of changed prototypes. $\alpha(t) \in [0, 1]$ is the training parameter which decreases with t.

Equation (2) gives the neighbourhood function which is in most cases a Gauss function,

$$h_{ci}(t) = exp\left(-\frac{||m_c - m_i||^2}{2\sigma^2(t)}\right) \quad (2)$$

that defines the neighbourhood in which the neurons are displaced to the input data vector x. m_c and m_i denotes the coordenates of the neurons c and i on the net.

The manipulation of the functions $h_{ci}(t)$ and $\alpha(t)$ determin the speed to reach the final state of the arrangemente of the neurons on the net in which the prototypes do not change their values anymore and the net is converged. More details and other properties of the SOM can be obtained in [1].

2.2. Iterative Elastic Neural Net

The Iterative Elastic Neural Net [6], is a method which repeats the Elastic Neural Net algorithm of Gorban for every found clusters mutiple times. It may be formulated in 4 phases.

In the first phase the data of the multidimensional space is preprocessed by cleaning and normalizing and preparing the data for the algorithm.

The phase of the Elastic Neural Net (ENN) [3], creates a net of nodes and defines an energy functional,

$$U = U^{(Y)} + U^{(E)} + U^{(R)} \quad (3)$$

which will be globally minimized with respect to the position of the nodes. $U^{(Y)}$ is the interaction energy of the data with their nearest nodes. $U^{(E)}(\lambda)$ is the elastic energy between neigbouring nodes. It depends on the constant λ which controls the elasticity of the net. Finally, $U^{(R)}(\mu)$ is the deformation energy between the nodes and their neighnours and is controlled by the deformability parameter μ. Details on the formulation and minimization of the functional U are found in Gorban et.al. [3], [4].

In the phase for the pattern identification the clusters on the bidimensional net are analyzed by the k-Means method together with a quality index and an optimal number of clusters is found [6].

Finally, the phase of cluster ananlysis, for the optimal number of clusters the centroid curves are calculated and analyzed, classifying them by their separation, high and low values of the mean expression level and their up-or down-regulation respect to the different carbon sources. In this step the good clusters (expression values higher than 5.5 in a logarithmic scale and with a clear upregulation) are checked for the 345 upregulated genes found experimentally. More details of the IENN method may be consulted in [6].

3. Iterative Self Organizing Maps (ISOM)

In the ISOM method the SOM is applied to a first net with a large number of nodes (3,600) and clusters of nodes are found by a slightly modified k-Means method. The following matrix represents the number of data projected on the nearest node (hits). There are n*n (n=60) nodes (i, j) with $h(i, j)$ hits on the corresponding node.

$$X = \begin{matrix} h(1,1) & h(1,2) & \dots & h(1,n) \\ h(2,1) & h(2,2) & \dots & h(2,n) \\ \vdots & \vdots & \ddots & \vdots \\ h(n,1) & h(n,2) & \dots & h(n,n) \end{matrix}$$

One chooses now a trial number of clusters $nclust$, going from 2 to 10, for the decomposition with arbitrary cluster centers $cx(k), cy(k); k = 1, ..., nclust$ Then one proceeds in two steps:

(i) One calculates the distances between the cluster centers and the nodes in order to define which node belongs to which cluster by taking the nearest nodes with respect to the cluster center.

(ii) The calculation of the new cluster centers is now done in

the flollowing way: Let $nc(k)$ be the number of nodes with coordinates $x(i,k), y(j,k)$ and hits $h(i,j,k)$ which belong to the cluster k, then the new cluster centers $cx(k), cy(k)$ are defined by:

$$cx(k) = \sum_{(i,j)\in nc(k)} h(i,j,k) * x(i,k); \quad (4)$$

$$cy(k) = \sum_{(i,j)\in nc(k)} h(i,j,k) * y(j,k), \quad (5)$$

With this new cluster centers the steps (i) and (ii) are repeated until the cluster centers do not change anymore. In order to find the optimal number of clusters, the quality indices Davis-Bouldin [9] and I [9] are calculated for every cluster decomposition. Figure 1 shows a diagram for the ISOM method.

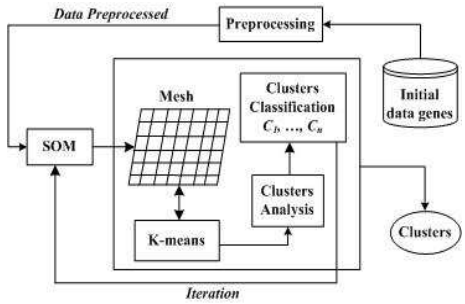

Figure 1. Flux diagram ISOM.

4. Data Collection

The data correspond to the level of gene expression of 7,312 genes obtained by the microarray technique of E.Coli [5]. These data are found in the GEO database[1] (Gene Expression Omnibus) of the National Center for Biotechnology Information [10]. The work of Liu et. al.[5] provides the 345 up-regulated genes that were tested experimentally. Each gene is described by 15 different experiments (which correspond to the dimensions for the representation of each gene) whose gene expression response is measured [5] on glucose sources. Specefically there are 5 sources of glucose, 2 sources of glycerol, 2 sources of succinate, 2 sources of alanine, 2 sources of acetate and 2 sources of proline. The definition of up-regulated genes according to [5] is given in relation to their response to the series of sources of glucose considering two factors: that its level of expression is greater then 8.5 on a log_2 scale (or 5.9 on a ln scale), and that its level of expression increases at least 3 times from the first to the last experiment on the same scale. In this evaluation a less restrictive definition was considered that includes the genes that have only an increasing activity of the level of expression with the experiments; since the definition given in [5] for up-regulated genes contain very elaborate biological information which requires a precise identification of the kind of gene to be detected.

The original data have expression level values between zero and hundreds of thousends. Such extensive scale does not offer an adequate resolution to compare expression levels; therefore a logarithmic normalization is carried out. In this case the use of the natural logarithm [11] was preferred instead of the base 2 logarithm used by Liu, because it is a more standard measure and it was used in the application of the Iterative Elastic Neural Net method [6] with that we compare our results. A limiting value of 5 in the natural logarithmic scale for the expression level was estimated by determining the threshold as the value that best separates the initial clusters. This expression level allows discarding groups of genes that have an average level lower than this value.

5. Application of ISOM and Results

The application of the new method ISOM will be done like the IENN method in four phases: data preprocessing, SOM application, pattern identification, and finally a stopping criterion and cluster selection based on the expression level and inspection of the pattern that is being sought.

In the phase of data preprocessing the set of N data to be analized is chosen, $x^j = [x_1^j, ..., x_M^j]^T, j = 1...N$, where N corresponds to the 7,312 genes of the E.coli bacterium, M to the 15 different experiments carried out on the genes and x^j is the gene expression level. The data are normalized in the form $\theta^j = ln(x^j - min(x^j) + 1)$.

In the second phase the package SOM Toolbox 2.0 for MATLAB 5 is applied to the data, using a two-dimensional net with 60*60 nodes (neurons). With such a large and more flexible net one would expect to find a good enough data distribution on the nodes and see hopefully clustering.

In the phase of pattern identification the data are analyzed by projecting them on internal coordinates for the possible formation of clusters or other patterns such as accumulation of clusters in certain regions of the net. As a typical dependence of the data in a cluster on the dimensions of the multidimensional space, the mean expression level of the data for each dimension, the clusters centroid, is calculated. For the formation of possible clusters the k-Means method is used together with the quality indices Davies-Bouldin [9]and I [9], which gives information on the best number of clusters. The centroids of each cluster are graphed and analyzed to find possible patterns.

[1] http://www.ncbi.nlm.nih.gov/projects/GEO/goes

Once the best number of clusters is obtained, the centroids' curves are used to detect and extract possible patterns with respect to the activity of the genes such as increasing, decreasing or fluctuating activity of the expression level with respect to the 15 dimensions. Clusters found with an average expression level below 5 in the natural logarithmic scale and clusters with a decreasing activity of the average expression level were discarded. For the other clusters phases 2 and 3 are repeated and the analysis of phase 4 is carried out again, repeating the process until the remaining clusters have well defined increasing activity levels.

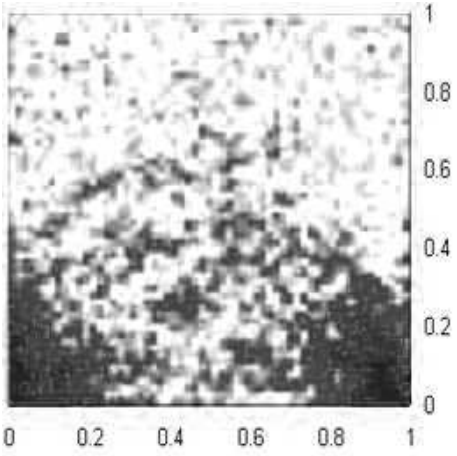

Figure 2 (a) Data distribution on the 3,600 nodes of the net after the application of the SOM on the initial data set. The darker spots indicate a higher data concentration.

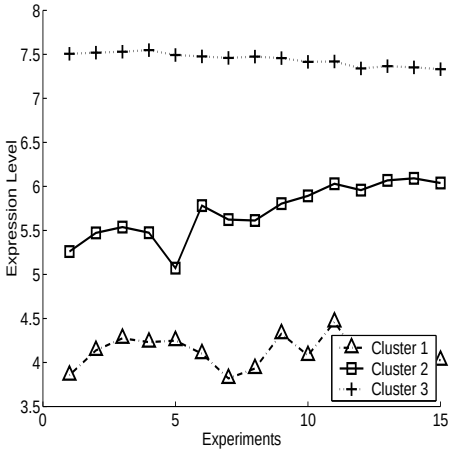

Figure 2 (b) Centroids of the 3 clusters found by the division of the initial cluster with the k-Means method as a function of the 15 experiments.

Figure 2 (a) shows the cluster distribution of the first application of the SOM whith 2 accumulations of data at the corners of net and an equal data distribution in the rest of the net. With the k-Means method and the quality index, 3 clusters were found.

In figure 2 (b) the centroids for the 3 clusters are shown and table 1 shows a list of these clusters, their number of genes, their number of up-regulated genes of the list of 345 genes found by Liu et.al. [5] and their tendency in the activity of the genes with respect to the 15 experiments.

cluster	np(k)	nup(k)	tendency
cluster 1	2,451	4	fluctuating
cluster 2	2,393	301	up
cluster 3	2,468	40	slightly down

Table 1. Division of the initial data cluster0 in 3 clusters with the number of genes $np(k)$, the number of up-regulated genes $nup(k)$ and the tendency in the activity of the genes with respect to the 15 dimensions

The centroid of cluster 1 has very low expression levels. As the interest lies in clusters with an expression level higher than 5.5, cluster 1 is not treated anymore. To the remaining two clusters the SOM is again applied with a net of 60*60 nodes. The result is given in figures 3 and 4. The k-Means method with the quality indices applied to the two clusters give for cluster 2 the 4 subclusters, cluster2-1, cluster2-2 and so on, and for cluster 3 the 3 subclusters cluster3-1, cluster3-2 and cluster3-3. Their centroids are given in figure 3 (b), 4(b) and table 2 shows their number of genes, their number of up-regulated genes and their tendency with respect to the 15 experiments.

For some of these clusters another division in subclusters with SOM was performed in order to get better results in the concentration of up-regulated genes in smaller clusters.

cluster	np(k)	nup(k)	tendency
cluster2-1	645	24	up
cluster2-2	539	7	up
cluster2-3	540	41	up
cluster2-4	669	229	up
cluster3-1	761	18	slightly up
cluster3-2	889	1	down
cluster3-3	818	21	down

Table 2. Division of the cluster2 and cluster3 in sub clusters with the number of genes $np(k)$, the number of up-regulated genes $nup(k)$ and the tendency in the activity of the genes with respect to the 15 dimensions

For the comparison of the results with the simple SOM technique and the IENN method, the results of the publication of Chacón et.al. [6] were used. With the IENN method 9 clusters with high expression level and increasing activity with the 15 experiments were found, with 1,579 genes from which 299 correspond tu up-regulated genes. This

corresponds to a concentration of 18.9% in these 9 clusters. The simple SOM technique gave for a net of 5*6 nodes 225 up-regulated genes with a total of 1,653 genes in the 9 clusters with the highest percentage of up-regulated genes which corresponds to a concentration of 13.6%.)

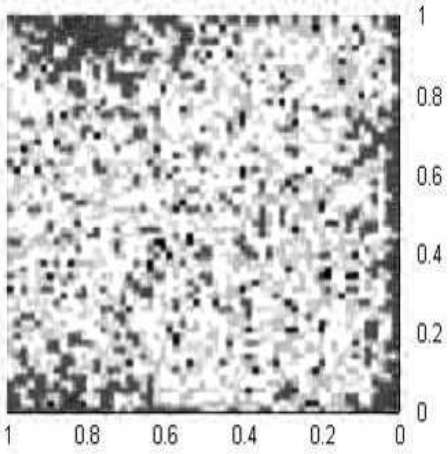

Figure 3 (a) Data distribution on the 3,600 nodes of the net after the application of the SOM on the data set of cluster 2. The darker spots indicate a higher data concentration.

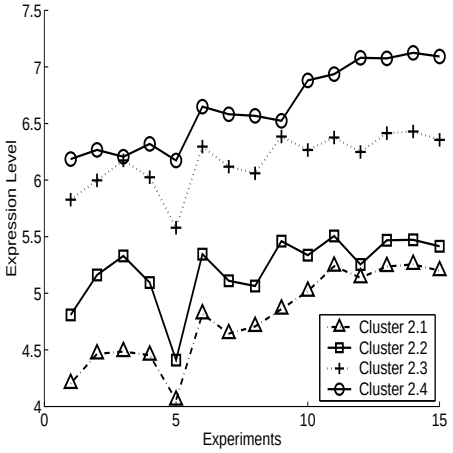

Figure 3 (b) Centroids of the 4 clusters found by the division of cluster 2 with the k-Means method as a function of the 15 experiments.

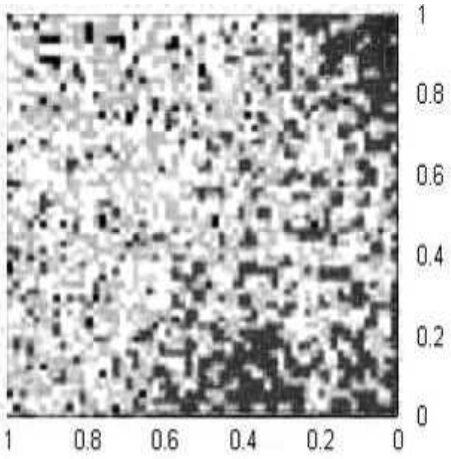

Figure 4 (a) Data distribution on the 3,600 nodes of the net after the application of the SOM on the data set of cluster 3. The darker spots indicate a higher data concentration.

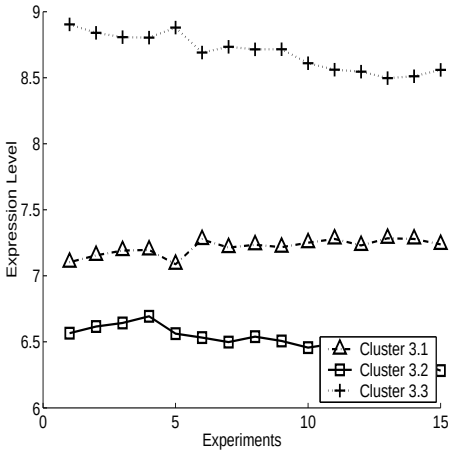

Figure 4 (b) Centroids of the 3 clusters found by the division of cluster 3 with the k-Means method as a function of the 15 experiments.

6. Discussion and Conclusion

In order to see a good distribution of the data on the nodes of the net with the SOM, it is necessary to use a large number of nodes. With a good distribution of the data on the nodes it is possible to visualize and define possibles clusters of nodes. The selection of a large number of nodes produces additionally a more adjustable net. With more efficient computers it is now possible to use large numbers of nodes. We used a net with 3,600 nodes which seems to be enough to see a good data distribution (see fig.2 (a)) of the initial data.

The large number of nodes leads not only to a clear formation and visualization of clusters, but these clusters can then be analized again for further structures, applying SOM again to every cluster. The hierarchical application of SOM shows a better resolution than the simple SOM.

From figures 3(b) and 4(b) one notes a clear separation of the centroids of nearly all 7 clusters. Some of the clusters have similar expression values but a different tendency in the activity. All subclusters of cluster 2 show a clear increasing tendency over the 15 experiments in their centroides but cluster2-1 and cluster2-2 have low expression values. They contain only a few up-regulated genes. Cluster2-4 with the clearest increasing tendency and the highest expression values contains 229 up-regulated genes. That is 2/3 of all up-regulated genes.

The three subclusters of cluster 3 have high expression values, but cluster3-2 and cluster3-3 show a decreasing tendency over 15 dimensions and cluster3-1 a slight increasing tendency, all with only few up-regulated genes.

One notes in table 2 that 2 clusters with 1,209 genes contain 270 up-regulated genes. That is a concentration of 22.3% of up-regulated genes in these 2 clusters. In a further division of the subclusters 2 and 3 one finds for the 5 clusters with the highest concentration of up-regulated genes 267 up-regulated genes with a total of 768 genes, that is a concentration of 35%. This is a good result and it will be compared to the application of a simple SOM or of the IENN.

Since in this application to the genes of E.Coli one accounts on the 345 upregulated genes [5] identified in the laboratory, it is possible to carry out an evaluation considering the three methods (ISOM, IENN and SOM) as classifiers. Moreover, if the number of clusters which contain the up-regulated genes is considered a classification parameter, it is possible to make an analysis by means of Receiver Operating Charactersitic (ROC), varying this number according the concentration of upregulated genes. The ROC analysis will be given for the three cases. One calculates the confusion matrix for different numbers of clusters. The clusters are ordered after the concentration of up-regulated genes. The ROC analysis for the simple SOM and the IENN method are given in [6]. Figure 5 shows the 3 ROC curves.

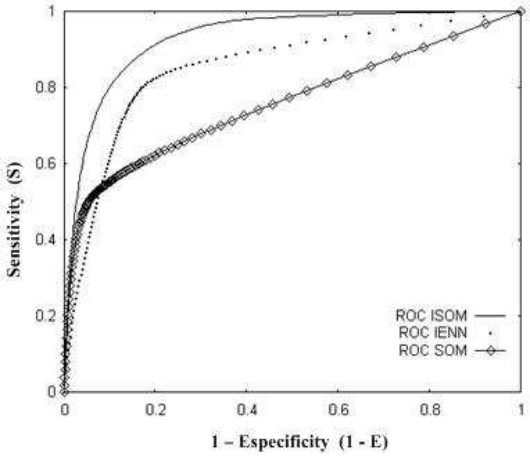

Figure 5 Curves ROC for ISOM, IENN and SOM.

When considering an overall analysis of the classifier using the expression level as a parameter, it is important to consider the area under the ROC curve. One notes clearly from these classification curves that the area beneath the ISOM curve gives the highest value and a best optimum decision level.

The results of the application to the discovery of up-regulated genes of E.Coli show a clear advantage of the proposel over the traditional use of the SOM method or the new IENN method. We chose to carry out a comparison with these two methods where the SOM is frequently used in the field of bioinformatics and the IENN recently established in this field, but it is also necessary to evaluate other alternatives which consider the robustnes of the methods. One of the recent methods, consensus clustering [12], uses new resampling techniques which should give information about the stability of the found clusters and confidence that they represent real structure.

References

[1] Kohonen T. Self-organizing maps, Berlin: Springer-Verlag (2001)

[2] Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander E and Golub T. Interpreting patterns of expression with self-organizing maps: Methods and application to hematopoietic differentiation. Genetics **96**, (1999) 2907–12.

[3] Gorban A, and Zinovyev A. Method of elastic maps and its applications in data visualization and data modeling. International Journal of Computing Anticipatory Systems, CHAOS. **12**, (2001) 353–69.

[4] Gorban A and Zinovyev A. Elastic principal graphs and manifolds and their practical applications, Computing **75**, (2005) Springer-Verlag 359–379.

[5] Liu M, Durfee T, Cabrera T, Zhao K, Jin D, and Blattner F Global transcriptional programs reveal a carbon source foraging strategy by E. Coli. J Biol Chem **280**, (2005) 15921-7.

[6] Chacón M, Lévano M, Allende H, Nowak H. Detection of gene expressions in microarrays by applying iteratively elastic neural net. ICANNGA 07, LNCS **4432**, (2007) 355–363.

[7] Lampinen J and Oja E. Clustering properties of hierachical self-organizing maps, Journal of Mathematical Imaging and Vision, Vol. 2, (1992) 261–272.

[8] Provost F and Fawcett T. Robust classification for imprecise environments. Machine Learning, Vol. **42**, No3, (2001) 203–231.

[9] Maulik U, Bandyodpadhyay S Performance evaluation of some clustering algorithms and validity indices, IEEE PAMI **24**, (2002) 1650–4.

[10] Edgar R, Domrachev M, and Lash A. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research, Vol. **30**, No1, (2002) 207–210.

[11] Quackenbush J. Microarrays data normalization and transformation, Nature Genetic, Maryland USA, 2001.

[12] Monti S, Tamayo P, Mesirov and Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Springer Netherlands, Vol. **52**, (2003) 91–118.