Available online at www.sciencedirect.com

ScienceDirect

International Journal of Approximate Reasoning
47 (2008) 70–84

INTERNATIONAL JOURNAL OF APPROXIMATE REASONING

www.elsevier.com/locate/ijar

Clustering and visualization approaches for human cell cycle gene expression data analysis

F. Napolitano [a], G. Raiconi [a], R. Tagliaferri [a,*], A. Ciaramella [b], A. Staiano [b], G. Miele [c,d]

[a] *Department of Mathematics and Informatics, University of Salerno, I-84084, via Ponte don Melillo, Fisciano (SA), Italy*
[b] *Department of Applied Sciences, University of Naples "Parthenope", I-80133, via A. de Gasperi 5, Napoli, Italy*
[c] *Department of Physical Sciences, University of Naples, I-80136, via Cintia 6, Napoli, Italy*
[d] *INFN, Unit of Naples, Napoli, Italy*

Received 20 April 2006; received in revised form 8 September 2006; accepted 15 March 2007
Available online 21 April 2007

Abstract

In this work a comprehensive multi-step machine learning data mining and data visualization framework is introduced. The different steps of the approach are: preprocessing, clustering, and visualization. A preprocessing based on a Robust Principal Component Analysis Neural Network for feature extraction of unevenly sampled data is used. Then a Probabilistic Principal Surfaces approach combined with an agglomerative procedure based on Fisher's and Negentropy information is applied for clustering and labeling purposes. Furthermore, a Multi-Dimensional Scaling approach for a 2-dimensional data visualization of the clustered and labeled data is used. The method, which provides a user-friendly visualization interface in both 2 and 3 dimensions, can work on noisy data with missing points, and represents an automatic procedure to get, with no a priori assumptions, the number of clusters present in the data. Analysis and identification of genes periodically expressed in a human cancer cell line (HeLa) using cDNA microarrays is carried out as test case.

Keywords: Preprocessing analysis; Data analysis; Data visualization; Microarray data

1. Introduction

Scientists have been successful in cataloguing genes through genome sequencing projects, and they can now generate large quantities of gene expression data using microarrays [19]. Proper regulation of the cell division cycle is crucial to the growth and development of all organisms; understanding this regulation is central to the study of many diseases, most notably cancer [9,10,11,21]. However, due to the sheer size of the data sets involved and to the complexity of the problems to be tackled, novel approaches to data mining and understanding, relying on artificial intelligence tools, are necessary. These tools can be divided in two main families:

* Corresponding author.
E-mail addresses: fnapolitano@unisa.it (F. Napolitano), gianni@unisa.it (G. Raiconi), rtagliaferri@unisa.it (R. Tagliaferri), angelo.ciaramella@uniparthenope.it (A. Ciaramella), antonino.staiano@uniparthenope.it (A. Staiano), miele@na.infn.it (G. Miele).

0888-613X/$ - see front matter
doi:10.1016/j.ijar.2007.03.013

tools for supervised learning, which make use of prior knowledge to group samples into different classes, and unsupervised tools, which rely only on the statistical properties of the data themselves [20,25,27,36,37,38,40]. Both approaches have been used for a variety of applications and have advantages and disadvantages: the choice of a specific tool depends on the purpose of the investigation and the structure of the data. Among various applications, we recall:

- Diagnostic: i.e. to find gene expression patterns specific to given classes mainly dealt with supervised methods [3,29].
- Clustering: aimed at grouping genes that are functionally related without attempting to model the underlying biology [18,35,34].
- Model-based approach: generation of a model that justifies the grouping of specific genes and trains the parameters of the model on the data set [6,16,41].
- Projection methods: which decompose the data set into components that have the desired properties [1,28,7]. To this class belong some methods like Principal Components Analysis (PCA), Independent Components Analysis (ICA) and Probabilistic Principal Surfaces (PPS) which are discussed in this paper.

Moreover, many of these applications can suffer from poor data visualization techniques. Regarding the clustering problems, the most commonly used clustering algorithms, such as the hierarchical clustering and k-means [17] suffer from some limitations, namely:

(i) they need an arbitrary and a priori choice of the correct number of clusters, and this greatly affects the success of the clustering procedure;
(ii) some of them, like hierarchical clustering, cannot properly handle large experimental data sets which are typically noisy and incomplete (i.e. they contain a large fraction of missing data points);
(iii) they do not feature a user-friendly data visualization, which is quite crucial for further analysis and understanding in case of large amount of data.

On the other hand, data visualization is an important mean of extracting useful information from large quantities of raw data. The human eye and brain together make a formidable pattern detection tool, but for them to work the data must be represented in a low-dimensional space, usually two or three dimensions. Even if simple relationships can seem very obscure when data is presented in tabular form, they are often very easy to see by visual inspection. Many results in experimental biology first appear in image: a photo of an organism, cells, gels, or microarray scans. As the quantity of these results grows, automatic extraction of features and meaning from experimental images becomes crucial. At the other end of the data pipeline, 2D or 3D visualizations alone are inadequate for exploring bioinformatics data. Biologists need a visual environment that facilitates the exploration of high-dimensional data depending on many parameters. In this context, research needs further work into bioinformatics visualization to develop tools that will meet the upcoming genomic and proteomic challenges. Many algorithms for data visualization have been proposed by both neural computing and statistics communities, most of which are based on a projection of data onto a two or three-dimensional visualization space. We mark that in [2] the authors introduced a multi-step approach for data clustering. In this work we take advantage of that approach and a hierarchical clustering based on both Fisher's and Negentropy information is introduced. Moreover here we introduce an advanced visualization technique and an integrated environment for clustering and 2D or 3D visualization of high-dimensional biomedical data. The approach enable the user to:

- project and visualize data on a spherical surface (which provides a useful continuous manifold that can be rotated and manipulated in several ways) or map the sphere into a 2-dimensional space;
- perform deeper studies on the data by localizing regions of interest and interacting with the data themself;
- interact with data, choose the points of interest, visualize their neighbors and similar points, print all related information etc.;
- label the data choosing the classes found by a Negentropy based dendrogram approach;
- visualize the labeled data in a bi-dimensional space using a Multi-Dimensional Scaling (MDS) approach.

In detail, in this work we propose a new approach, based on a solid mathematical formalism, able to visualize and cluster noisy gene expression data with missing data points. The method can be divided in two separated parts: the preprocessing of the data, which, as widely occurs, is specifically tailored to the characteristics of the data set under examination, and the clustering and visualization part which is of more general applicability. We stress that the preprocessing and clustering phases are absolutely independent. In this way we have the possibility to apply separated adaptive processes (non-linear PCA, PPS, hierarchical clustering). We note that this is really useful when a large data set is considered. The method was tested against the human cell cycle to identify genes periodically expressed in tumors [39]. The data set consists of gene expressions characterized during the cell division cycle in a human cancer cell line (HeLa) using cDNA microarrays.

The paper organization is as follows: in Section 2 we introduce the preprocessing approaches to eliminate and to extract features from the data by using a Robust Principal Component Analysis Neural Network; in Section 3 we introduce the Probabilistic Principal Surfaces approach for data clustering and visualization and in Section 4 the agglomerative hierarchical clustering approach based on Fisher and Negentropy information; in Section 5 we show the results obtained to cluster, label and visualize genes periodically expressed in a human cancer cell line; finally in Section 6 some comments on the approach are introduced.

2. Preprocessing

Microarray data are very noisy, and thus preprocessing plays an important role. Preprocessing is needed to filter out noise and to deal with missing data points [30]. We used a two-step preprocessing phase: a preliminary procedure of noisy data rejection, followed by a nonlinear PCA features extraction. About the first part of the preprocessing we simple eliminate the genes that have not samples in the particular experiment that we consider. Moreover, in most cases an interpolation is needed to overcome the problem of the missing points in the data set. We stress that using a Robust PCA Neural Network (NN) technique it is possible to work with periodic unevenly sampled data without using interpolation [13,32,33].

Summarizing, in our analysis this phase is accomplished by applying an on-line Robust PCA NN for each gene that corresponds to an unevenly sampled sequence. From each of these sequences we extract the m principal components to obtain the new feature vectors. In the following subsection we detail the PCA based approach.

2.1. Robust PCA based feature extraction

The second step of preprocessing is the extraction of the principal components (eigenvectors) of the autocorrelation matrices of the genes which passed the filtering procedure. This step is used to deal with missing data points. Feature extraction process is based on a non-linear PCA method which can estimate the eigenvectors from unevenly sampled data. This approach is based on the STIMA algorithm described in [13,32,33].

We note that PCA's can be neurally realized in various ways [22,13,33,32]. In our case we used a non-linear PCA NN consisting of a one layer feedforward NN able to extract the principal components of the autocorrelation matrix of the input sources. Typically, Hebbian type learning rules are used, based on the one unit learning algorithm originally proposed by Oja and co-workers [22]. Many different versions and extensions of this basic algorithm have been proposed during the recent years [23,24,13,22]. The structure of the PCA NN can be summarized as follows: there is one input layer and one forward layer of neurons totally connected to the inputs; during the learning phase there are feedback links among neurons, that classify the network structure as either hierarchical or symmetric. After the learning phase the network becomes purely feedforward. The hierarchical case leads to the well-known GHA algorithm; in the symmetric case we have the Oja's subspace network. PCA neural algorithms can be derived from optimization problems; in this way the learning algorithms can be further classified in robust PCA algorithms and nonlinear PCA algorithms [23,24,13]. A PCA algorithm is said robust, when the objective function grows less than quadratically; examples of valid cost functions are $\ln(\cosh(t))$e$|t|$. In order to extract the principal components we use a Robust nonlinear PCA NN.

Finally, we see that the approach can be divided in the following two main steps:

- Normalization: we first calculate and subtract the average pattern to obtain a zero mean process.

- Neural computing: the fundamental learning parameters are (i) the number of output neurons m, which is equal to the estimated embedding dimension and it is the number of principal eigenvectors that we need; (ii) the number of input neurons q; (iii) the initial weight matrix $\mathbf{W}$ of $m \times q$ dimension; (iv) α, the nonlinear learning function parameter; (v) the learning rate μ and the ϵ tolerance.

This leads to the algorithm for the generic ith, m-dimensional weight matrix $\mathbf{w}_i(k)$ $(i = 1, \ldots, m)$ at time k:

$$\begin{aligned} \mathbf{w}_i(k+1) &= \mathbf{w}_i(k) + \mu_k g(y_i(k))\mathbf{e}_i(k), \\ \mathbf{e}_i(k) &= \mathbf{x}_i - \sum_{j=1}^{I(i)} y_j(k)\mathbf{w}_j(k) \end{aligned} \tag{1}$$

In the hierarchical case we have $I(i) = i$. In the symmetric case $I(i) = q$, the error vector $\mathbf{e}_i(k)$ becomes the same $\mathbf{e}_i$ for all the neurons. For more details on the algorithm see [13,33,32].

3. Probabilistic Principal Surfaces

Probabilistic Principal Surfaces (PPS) [8,2,12] are a nonlinear extension of principal components, in that each node on the PPS is the average of all data points that project near/onto it. From a theoretical standpoint, the PPS is a generalization of the Generative Topographic Mapping (GTM) [5], which can be seen as a parametric alternative to Self Organizing Maps or SOM [26]. Advantages of PPS include its parametric and flexible formulation for any geometry/topology in any dimension, guaranteed convergence (indeed the PPS training is accomplished through the Expectation–Maximization algorithm) [15]. PPS are governed by their latent topology and owing to their flexibility, a variety of PPS topologies can be created, for example as a regular grid of a $3D$ sphere. A sphere is finite, unbounded and symmetric, with all the nodes distributed on the surface, and it is therefore suitable for emulating the sparseness and peripheral property of high-D data. Furthermore, the sphere topology can be easily understood by humans and thereby used for visualizing high-D data.

PPS define a nonlinear, parametric mapping $\mathbf{y}(\mathbf{x}; \mathbf{W})$ from a Q-dimensional latent space $(\mathbf{x} \in R^Q)$ to a D-dimensional data space $(\mathbf{t} \in R^D)$, where usually $Q < D$. The (continuous and differentiable) function $\mathbf{y}(\mathbf{x}; \mathbf{W})$ maps each of the M points in the latent space (where M is the number of latent variables which is fixed *a priori*) to a corresponding point into the data space. Since the latent space is Q-dimensional, these points will be confined to a Q-dimensional manifold non-linearly embedded into the D-dimensional data space. We mark that the PPS approach builds a constrained mixture of Gaussians. Moreover, if $Q = 3$ is chosen, a spherical manifold [8,2,12] can be constructed using PPS with nodes arranged regularly on the surface of a sphere in R^3 latent space, with the latent basis functions evenly distributed on the sphere at a lower density. After a PPS model is fitted to the data, several visualization possibilities are available like the projection of the data are projected into the latent space as points onto a sphere.

Having projected the data into the latent sphere, a typical task performed by most data analyzers is the localization of the most interesting data points, for instance the ones lying far away from more dense areas (outliers), or those lying in the overlapping regions between clusters, and to investigate their characteristics by linking the data points on the sphere with their position in the original data set.

Furthermore, the latent variables responsibilities can be plotted on the sphere in order to obtain the data probability density function visualization. Another advantage of the $3D$ sphere representation is that, unlike $2D$ plot, it is possible to observe the distribution of the different clusters on the sphere and observe which clusters are similar to each other as the dense regions that are in close proximity. All these advanced visualization options and successful applications to other research fields (i.e., in astrophysics) have been discussed in [2].

4. Negentropy based approach

At this point we introduce the second step of the hierarchical approach that uses both Fisher's and Negentropy information to agglomerate the clusters found in the first phase (PPS clustering phase). We underline that the approach we are describing is based on that introduced in [12,2]. Such authors proposed a hierarchical agglomerative clustering where the optimal number of clusters is decided by analyzing the plateaus obtained

by varying an agglomerative threshold. We also stress that the most natural representation of a hierarchical agglomerative clustering is obtained by using a corresponding tree, called a dendrogram, which shows how the samples are grouped [17]. In this paper the optimal number of clusters is defined using the dendrogram information.

We note that the Fisher's linear discriminant is a projection method that projects high-dimensional data onto a line and performs classification in this one-dimensional space [4]. The projection maximizes the distance between the means of the two classes while minimizing the variance within each class. The Fisher criterion for two classes is given by

$$J_{\mathrm{F}}(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}} \tag{2}$$

where $\mathbf{S}_{\mathrm{B}}$ is the between-class covariance matrix and $\mathbf{S}_{\mathrm{W}}$ is the total within-class covariance matrix. From Eq. (2) differentiating with respect to $\mathbf{w}$ we find the direction where $J_{\mathrm{F}}(\mathbf{w})$ is maximized.

The definition of Negentropy J_{N} is given by

$$J_{\mathrm{N}}(\mathbf{x}) = H(\mathbf{x}_{\mathrm{Gauss}}) - H(\mathbf{x}), \tag{3}$$

where $\mathbf{x}_{\mathrm{Gauss}}$ is a Gaussian random vector of the same covariance matrix as $\mathbf{x}$ and $H(\cdot)$ is the differential entropy. Negentropy can also be interpreted as a measure of non-Gaussianity [22]. The classic method to approximate Negentropy is using higher-cumulants, through the polynomial density expansions. However, such cumulant-based methods sometimes provide a rather poor approximation of the entropy. A special approximation is obtained if one uses two functions G^1 and G^2, which are chosen so that G^1 is *odd* and G^2 is *even*. Such a system of two functions can measure the two most important features of non-Gaussian 1-D distributions. The odd function measures the asymmetry, and the even function measures the dimension of bimodality vs. peck at zero, closely related to sub- vs. supergaussianity. Then the Negentropy approximation of Eq. (3) is

$$J_{\mathrm{N}}(\mathbf{x}) \propto k_1 E\{G^1(\mathbf{x})\}^2 + k_2(E\{G^2(\mathbf{x})\} - E\{G^2(\upsilon)\})^2 \tag{4}$$

where υ is a Gaussian variable of zero mean and unit variance (i.e. standardized), the variable $\mathbf{x}$ is assumed to have also zero mean and unit variance and k_1 and k_2 are positive constants. We note that choosing the functions G^i that do not grow too fast, one obtains more robust estimators.

In this way we obtain approximations of Negentropy that give a very good compromise between the properties of the two classic non-Gaussianity measures given by kurtosis and skewness [12,2,22]. They are conceptually simple, fast to compute, yet have appealing statistical properties, especially robustness. We have to note that several methods to accomplish Independent Component Analysis are based on entropy information. At this point we remark that our aim is to agglomerate by an unsupervised method the clusters (regions) that are found by a clustering approach. The information, that we call $J_{\mathbf{NEC}}$, used to merge two clusters is based both on the Fisher's discriminant and on the Negentropy (NEC approach [12,2]):

$$J_{\mathbf{NEC}}(\mathbf{X}) = \alpha_{\mathrm{F}} J_{\mathrm{F}}(\mathbf{w}) + \alpha_{\mathrm{N}} J_{\mathrm{N}}(\mathbf{X}) \tag{5}$$

where α_{F} and α_{N} are two defined (normalizing) constants and $\mathbf{w}$ is the Fisher's direction. At this point using this information we apply the agglomerative hierarchical clustering approach and extract the dendrogram.

The NEC algorithm is described in Algorithm 1.

Algorithm 1. NEC: Agglomerative Hierarchical Clustering

Begin initialize $\hat{c} = c$, $D_i \leftarrow X_i$, $i = 1,\ldots,c$
DO $\hat{c} \leftarrow \hat{c} - 1$
find nearest clusters, say:
calculate the Fisher's direction between D_i and D_j and project the data on it
calculate the $J_{\mathbf{NEC}}$ information and merge clusters D_i and D_j with lowest $J_{\mathbf{NEC}}$ information
UNTIL $\hat{c} = 1$
return the dendogram
End

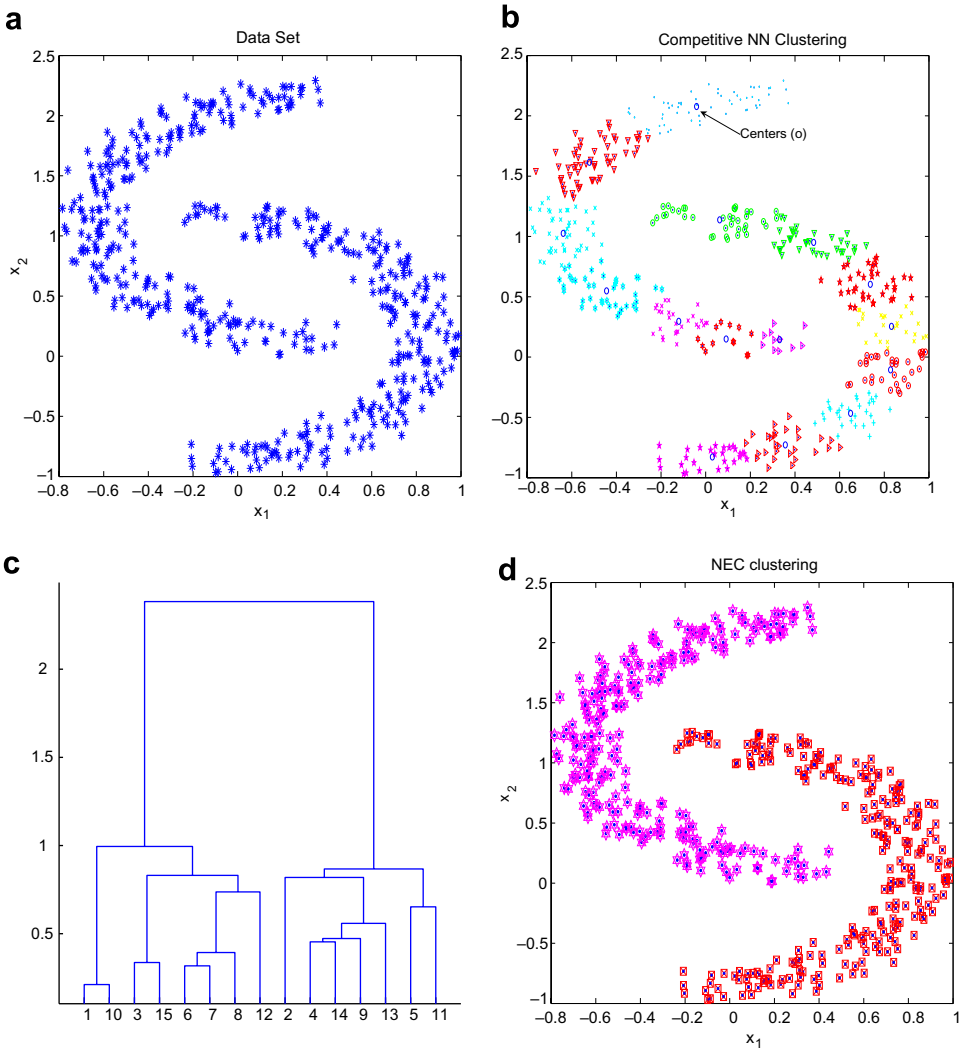

Fig. 1. NEC experimental results: (a) 2-dimensional data set; (b) comparative NN clustering; (c) NEC Dendrogram and (d) NEC clustering.

To show the agglomerative hierarchical clustering process we describe the results on a synthetic data set. We mark that the shown examples are obtained by using $\alpha_F = 0.1$, $\alpha_N = 10$. The data set that we consider is composed by two 2-dimensional classes with a complex distribution (see Fig. 1). In Fig. 1a we plot the data. In Fig. 1b we show the clusters obtained using the PPS approach and in Fig. 1c the NEC dendrogram. We note that focusing our attention on the dendrogram we clearly can define two separated regions obtaining the clusters in Fig. 1d.

5. Experimental results

To validate the proposed multi-step approach, in this section we detail its application to analyze and to visualize a data set composed by genes periodically expressed. To be more precise we focus our attention on the identification of genes periodically expressed in a human cancer cell line (HeLa) using cDNA microarrays. We stress that in [39] some results on this data set are presented. In that paper the authors used a particular Fourier based preprocessing and a hierarchical clustering method. The approach revealed coexpressed

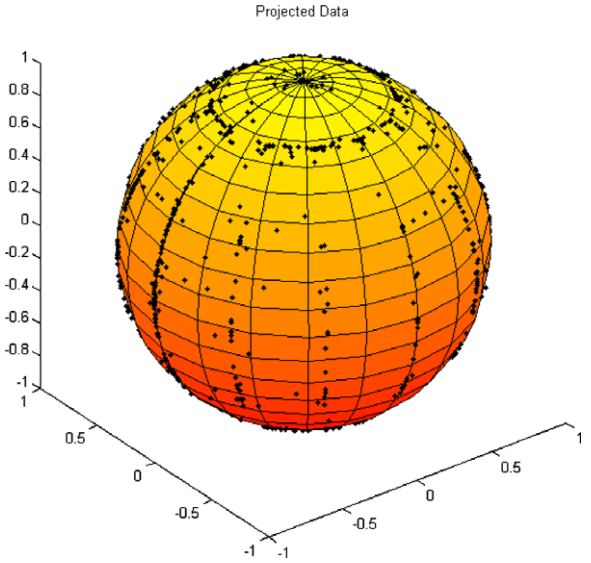

Fig. 2. PPS visualization: projection of the genes on the sphere.

groups of previously well-characterized genes involved in essential cell cycle processes such as DNA replication, chromosome segregation, and cell adhesion along with genes of uncharacterized function. In that paper transcripts of 874 genes showed periodic variation during the cell cycle.

To be more precise the data set is composed by 1134 genes where the values are obtained considering the ratio between two channels of the microarray (Cy5/Cy3). These values describe the index of the expression at a specific time. Approximatively we have one hour sampling period and the largest sequence cover about 2 days. The data set is composed by five experiments where the first 3 correspond to the same experiment but with a different overall sampling period. We also note that in the data set for some genes there are missing sequences and/or missing samples. In our analysis we use the third experiment because it covers a longer period and have less missing sequences and/or missing samples.

We remind that the first step needed to analyze a data set is the preprocessing. We wish to point out that since a Robust PCA NN method is used then we can elaborate unevenly sampled data directly. In fact in [13,32] the authors introduced and described a Robust PCA NN-based approach to analyze unevenly sampled data without using interpolation. In that paper the authors concluded that the method has better properties with respect to Fourier-based methods.

Now, in this work, using the preprocessing step, we extract for each gene the first two 10-dimensional principal components obtaining in this way a 20-dimensional feature vector for each gene. On the so obtained feature data set we apply the second step. In detail we apply the PPS approach to obtain a clustering with a high number of clusters that should be refined by using the NEC approach. In Fig. 2 we show the three-dimensional projection obtained by using the PPS approach.

Moreover, in Fig. 3 we show the 2-dimensional mapping obtained unfolding the sphere using a Robinson map projection (or orthophanic projection) [31]. At this point we apply the NEC agglomerative approach to combine the clusters obtained by PPS. We also note that the NEC approach permits to build a dendrogram tree that can be used to decide the number of clustering regions. In Fig. 4 we show the dendrogram obtained from this analysis. To compare and to validate the PPS $2D$ visualization we use a well-known methodology to map and to visualize the data in a 2-dimensional subspace: the Multi-Dimensional Scaling (MDS) which is the process of converting a set of pairwise dissimilarities for a set of points into a set of coordinates for the points. In Fig. 5 we plot the MDS projection in which an Euclidean distance is used. In Fig. 6 we show the same kind of projection obtained by using a correlation measure. At this point we stress that in data analysis/visualization a fundamental role is covered by the data labeling that can be emphasized by using different symbols or colors. In fact, now we focus our attention on the dendrogram of Fig. 4. From it we can clearly identify 4 main

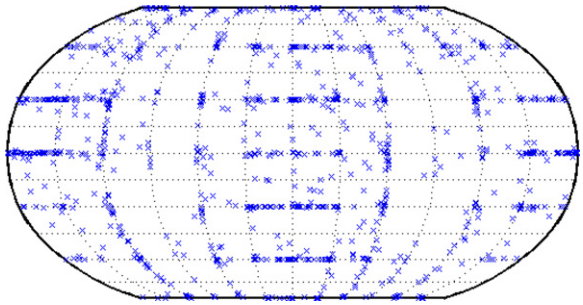

Fig. 3. 2-Dimensional Robinson map projection of the PPS 3-dimensional sphere.

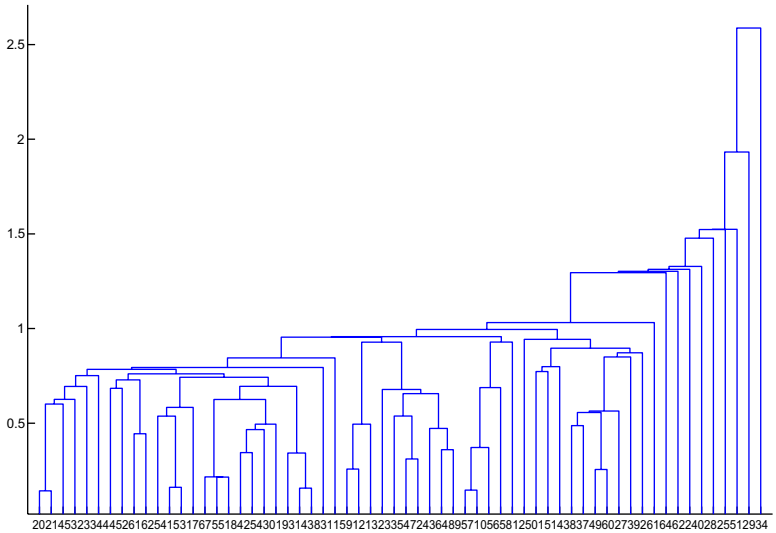

Fig. 4. NEC dendrogram.

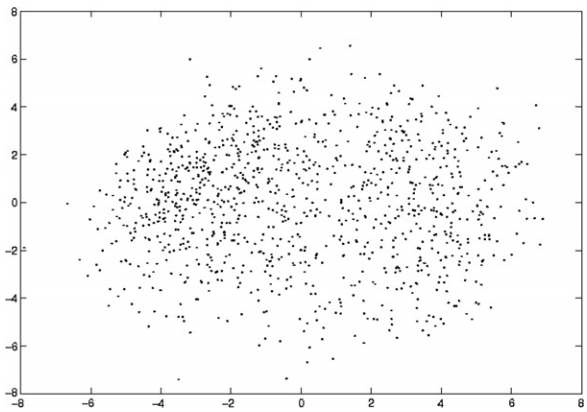

Fig. 5. MDS 2-dimensional projection by using Euclidean distance.

agglomerate regions and other 10 separated clusters. This becomes clearer when adding some graduated colors to the tree now plotted in Fig. 7. We have to underline that these are the graduated colors that we also use in the experiments described in the following so that it is simpler to understand the visualization features. We

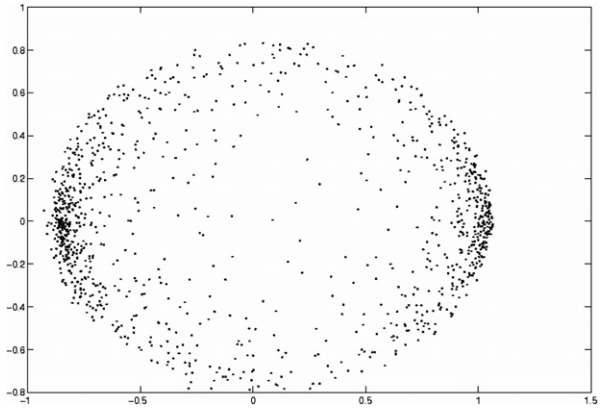

Fig. 6. MDS 2-dimensional projection by using correlation measure.

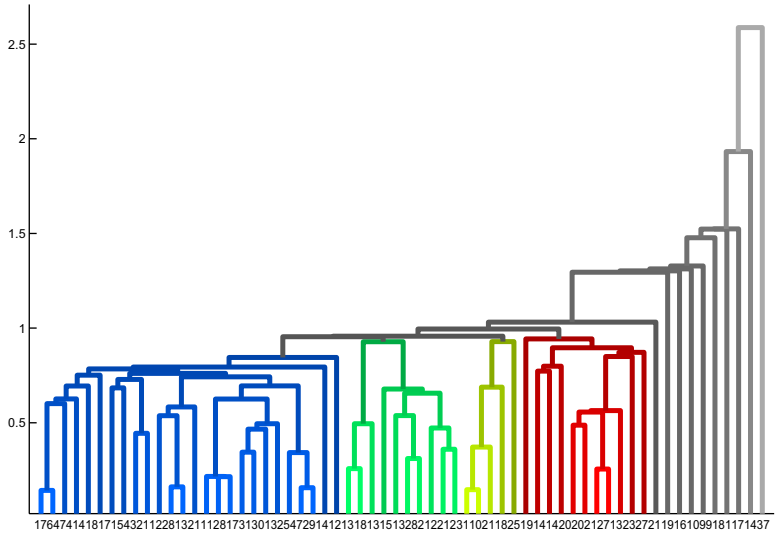

Fig. 7. NEC colored dendrogram.

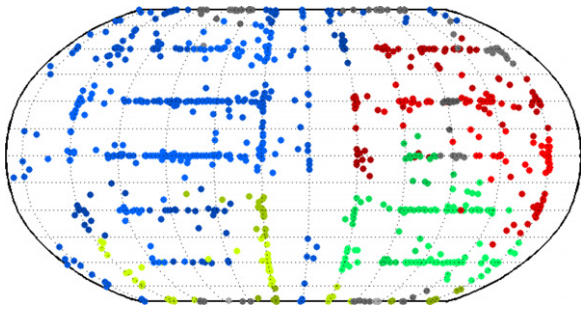

Fig. 8. PPS 2-dimensional mapping and labeling.

note that after this coloring step the data labeling is made easier. In Fig. 8 we show Fig. 3 with different labeled regions. Moreover, in Fig. 9 we plot the MDS projection by using an Euclidean distance. Now we point out that by using the multi-step approach it is simple to clusterize and to visualize high-dimensional data.

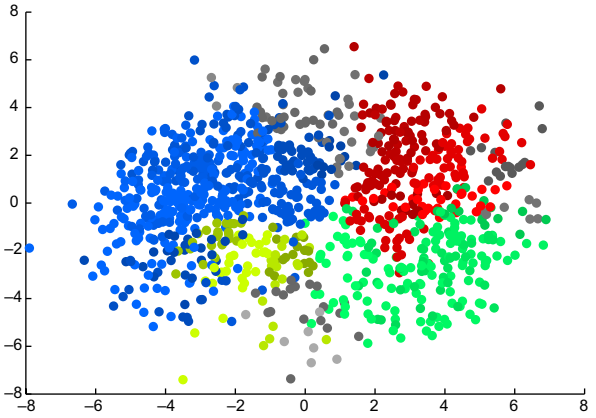

Fig. 9. MDS 2-dimensional projection and labeling.

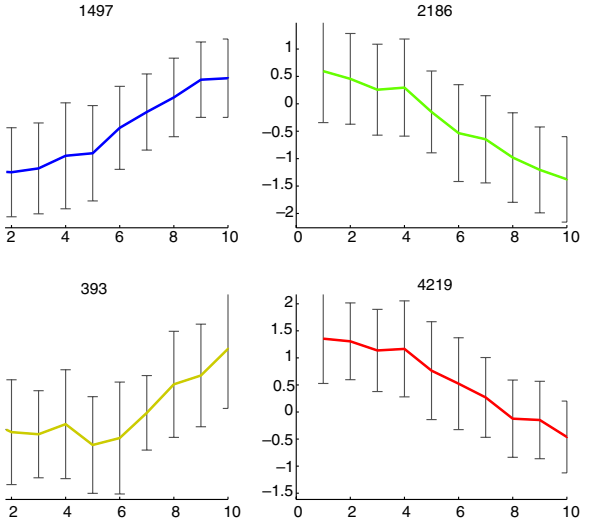

Fig. 10. The archetype behavior with corresponding standard deviation for the particularly significant clusters found in our analysis.

Moreover, after the clusterization, we can label the data to obtain useful information also in a $2D$ subspace. To show the performance of the approach on the considered data set we plot in Fig. 10 the error bars of the main 4 clusters. We mark that in the figure the colors assigned in the previous step are conserved. In detail we plot the mean and the variance of the first component extract by Robust PCA NN. We clearly confirm that the clusters have a different trend. In fact, we can deduct that we have mainly two different classes both composed by two subclasses. Moreover, in Fig. 11 we plot the error bars of the outliers clusters (gray). In this case we can deduct that their trends are different from those of the previous 4 clusters. We also stress that to validate these results a biological validation is needed. To be clearer and to show the differences between the proposed approach and the standard hierarchical agglomerative algorithm, which is the most used in literature, in Fig. 12 we show the dendrogram obtained by interpolating data and by using an Euclidean distance and in Fig. 13 by using the correlation. It is clear that in these cases it is extremely difficult to define the main clusters and consequently to label the data. Moreover in Figs. 14 and 15 we show the MDS for both the cases on the same interpolated data.

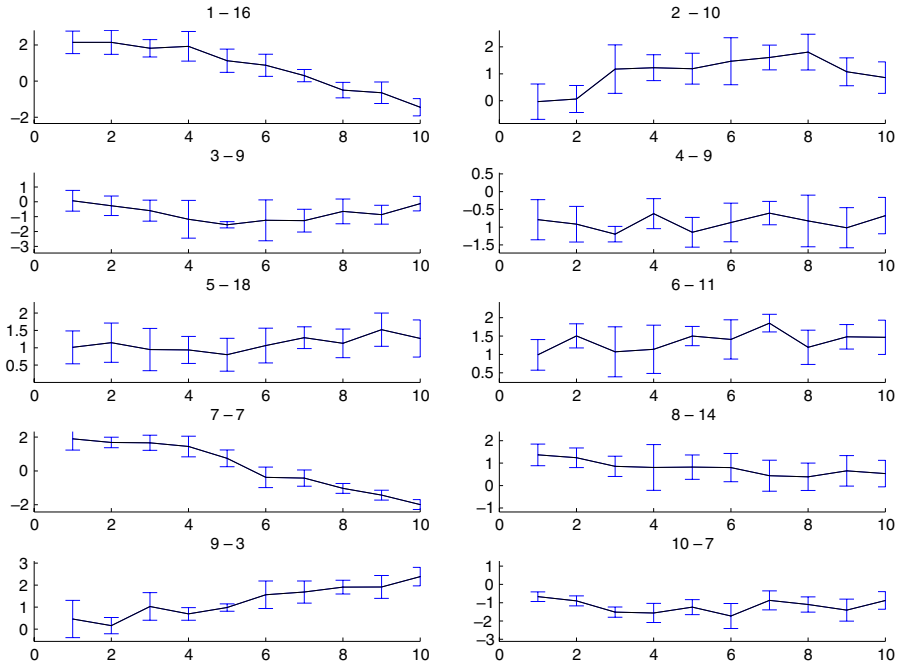

Fig. 11. The archetype behavior with corresponding standard deviation for the 10 outliers clusters found in our analysis.

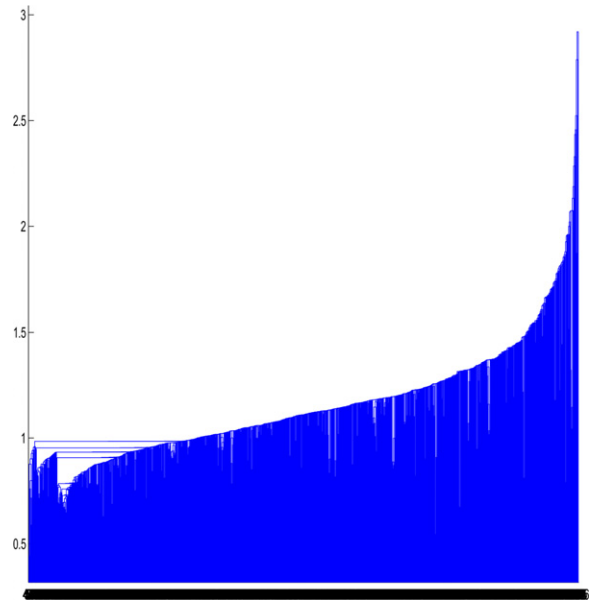

Fig. 12. Hierarchical agglomerative clustering obtained by using an Euclidean distance.

6. Discussion and conclusions

In this work we presented a new and complete machine learning data mining framework for data analysis and data visualization. The overall process is composed by a preprocessing and a clustering/visualization phases. The preprocessing consists in a filtering procedure and a Robust PCA NN for feature extraction. The second phase is based on a PPS combined with an agglomerative approach based on Fisher and

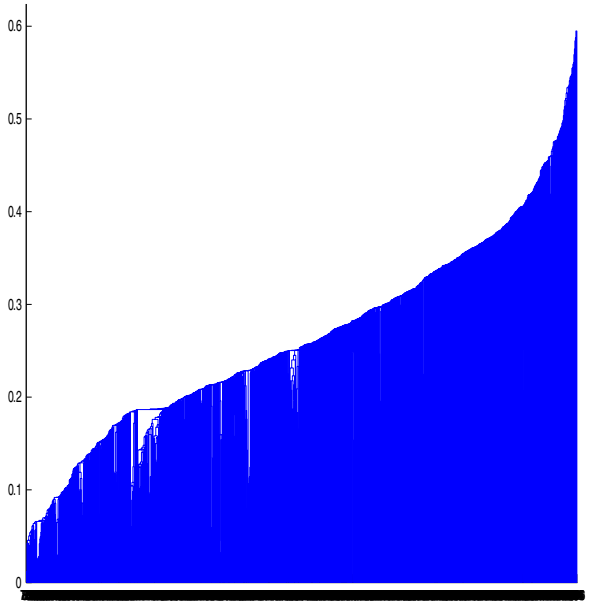

Fig. 13. Hierarchical agglomerative clustering obtained by using a correlation measure.

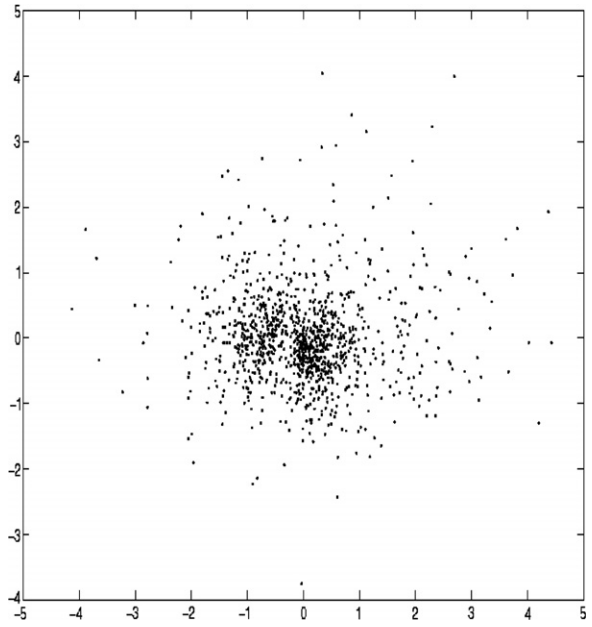

Fig. 14. MDS 2-dimensional projection obtained by using an Euclidean distance.

Negentropy information aimed at clustering and visualization. The approach has been applied to a microarray data set and more precisely to the analysis and identification of genes periodically expressed in a human cancer cell line (HeLa) using cDNA microarrays. Genes which pass the filtering procedure undergo a feature extraction process, based on a Robust PCA method, which allows us to obtain a matrix of eigenvectors from unevenly sampled data. The results of the above preprocessing are then analyzed by a PPS algorithm which has proven to be a very powerful and efficient model in several data mining activities, and in particular for high-D data visualization and clustering. Spherical PPS, which consist of a spherical latent manifold lying

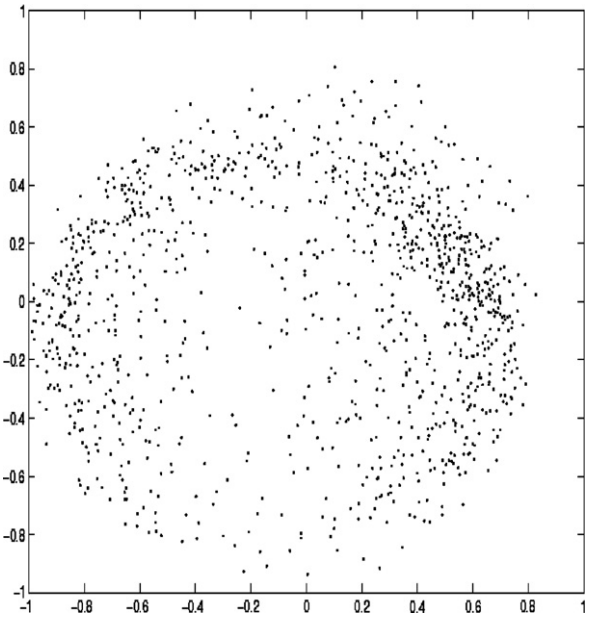

Fig. 15. MDS 2-dimensional projection obtained by using a correlation measure.

in a three-dimensional latent space, better deal with high-D data. The sphere, in fact, is able to capture the sparsity and periphery of data in large input spaces which are due to the curse of dimensionality [4]. The regions found by PPS are then agglomerated by NEC in a completely unsupervised manner obtaining in a simple way a dendrogram. In this way we can label the data in few regions that we decide. We stress that the $3D$ sphere is an innovative way of looking at gene expression data as compared to hierarchical clustering that displays a $2D$ plot. Furthermore, a mapping of the sphere in $2D$ can be obtained. Using a MDS approach, moreover, we also map our data into a bi-dimensional space keeping in mind the labeling obtained from the dendrogram. In the case of the analyzed data set, our visualizations capture the information on clusters of genes that do share similar behavior.

We have also to remark that other neural-based methods have already been investigated with success on microarray data, for example, in [35,34], Self Organizing Map (SOM) have been used for both visualization and clustering purposes. However, SOM, while nice and understandable and appealing in a number of application fields, are less flexible in the sense that gives no way to directly interact with data and exhibits strong border effects on $2D$ neurons grid since its manifold is bounded.

As shown in the paper by de Lichtenberg et al. [14] the jungle of available clustering methods often leads to contradictory results, and often, it is difficult to choose a specific benchmark.

In this paper we use a data set of genes periodically expressed (see [39]) to show the performance of the proposed method in both clusterization and visualization. We however note that in [39] the authors clustered the genes using the hierarchical bottom-up algorithm which groups the genes that have similar relative concentration profiles during cell life cycle and they use a Fourier based preprocessing. In [13,33,32] the authors experimentally demonstrated that for unevenly sampled data a Robust PCA NN permits us to obtain a better performance than the Fourier based approaches. For these reasons the authors in the next future will focus their attention to use a frequency based preprocessing by using a MUSIC frequency estimator and a Robust PCA NN.

Moreover, we have that using Fisher and Negentropy information the clustering sequence is represented by a clear hierarchical tree which makes simple the identification of the clusters. In fact, in our case we identified 4 main clusters and other 10 different clusters that helped us to label the data. From this labeling we map the multidimensional data set in two dimensions using the MDS approach or the PPS mapping in a 2-dimensional space. We also underline how our multi-step approach permits us to add graduated colors that help us to

understand and to improve the data analysis process. In the next future the authors will focus their attention to compare and to validate the found clusters from a biological point of view.

References

[1] O. Alter, P.O. Brown, D. Botstein, Singular value decomposition for genome-wide expression data processing and modeling, Proc. Natl. Acad. Sci. USA 97 (2000) 10101–10106.
[2] R. Amato, A. Ciaramella, N. Deniskina, et al., A multi-step approach to time series analysis and gene expression clustering, Bioinformatics 22 (5) (2006) 589–596.
[3] T. Ando, M. Suguro, H. Hanai, M. Seto, Fuzzy neural network applied to gene expression profiling for producing the prognosis of diffuse large B-cell lymphoma, Cancer Res. 93 (2002) 1207–1212.
[4] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995.
[5] C.M. Bishop, M. Svensen, C.K.I. Williams, GTM: the generative topographic mapping, Neural Computation, 10(1) (1998) 215–234.
[6] H.J. Bussermaker, H. Li, E.D. Siggia, Regulatory element detection using correlation with expression, Nat. Genet. 27 (2001) 167–174.
[7] J.H. Chang, S.W. Chi, B.T. Zhang, Gene expression pattern analysis via latent variable models coupled with topographic clustering, Genomics Inform. 1 (2003) 32–39.
[8] K. Chang, J. Ghosh, A unified model for probabilistic principal surfaces, IEEE Trans. Pattern Anal. Mach. Intell. 23 (1) (2001).
[9] Y. Chen, E.R. Dougherty, M.L. Bittner, Ratio-based decision and the quantitative analysis of cDNA microarray images, J. Biomed. Opt. (1997) 364–374.
[10] R.J. Cho, M.J. Cambell, E.A. Winzeler, L. Steinmerz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D. Lockhard, R.W. Davis, A genom-wide transcriptional analysis of the mitotic cells, Mol. Cell 2 (1) (1998) 65–73.
[11] K.R. Christie et al., Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms, Nucleic Acids Res. (2004) 32, Database issue: D311-4.
[12] A. Ciaramella, G. Longo, A. Staiano, R. Tagliaferri, NEC: a hierarchical agglomerative clustering based on Fisher and Negentropy information, Lect. Notes Comp. Sci. 3931 (2006) 49–56.
[13] A. Ciaramella, C. Bongardo, H.D. Aller, M.F. Aller, G. De Zotti, A. Lähteenmaki, G. Longo, L. Milano, R. Tagliaferri, H. Teräsranta, M. Tornikoski, S. Urpo, A multifrequency analysis of radio variability of blazars, Astron. Astrophys. J. 419 (2004) 485–500.
[14] U. de Lichtenberg, L.J. Jensen, A. Fausbøll, T.S. Jensen, P. Bork, S. Brunak, Comparison of computational methods for the identification of cell cycle-regulated genes, Bioinformatics 21 (2005) 1164–1171.
[15] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum-likelihood from incomplete data via the EM algorithm, J. Roy. Stat. Soc. 39 (1) (1977).
[16] D. di Bernardo, M.J. Thompson, T.S. Gardner, S.E. Chobot, E.L. Eastwood, A.P. Wojtovich, S.J. Elliott, S.E. Schaus, J.J. Collins, Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks, Nat. Biotechnol. 23 (3) (2005) 377–383.
[17] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, second ed., John Wiley & Sons Inc., 2001.
[18] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Clustering analysis and display of genome-wide expression patterns, PNAS USA 95 (1998) 14863–14868.
[19] O. Ermolaeva et al., Data management and analysis for gene expression arrays, Nat. Genet. 20 (1998) 19–23.
[20] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer-Verlag, New York, 2001.
[21] T.R. Hughes et al., Functional discovery via a compendium of expression profiles, Cell 102 (2000) 109–126.
[22] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, John Wiley & Sons, 2001.
[23] J. Karhunen, J. Joutsensalo, Representation and separation of signals using non-linear PCA type learning, Neural Networks 7 (1994) 113–127.
[24] J. Karhunen, J. Joutsensalo, Generalizations of principal component analysys, optimization problems and neural networks, Neural Networks 8 (1995) 549–563.
[25] M.K. Kerr, G.A. Churchill, Statistical design and the analysis of gene expression microarray data, Genet. Res. 77 (2001) 123–128.
[26] T. Kohonen, Self-Organizing Maps, Springer-Verlag, Berlin, 1995.
[27] W. Liebermeister, Linear modes of gene expression determined by independent component analysis, Bioinformatics 18 (2002) 51–60.
[28] J. Misra, W. Schmitt, D. Hwang, L. Hsiao, S. Gullans, G. Stephanopoulos, Interactive exploration of microarray gene expression patterns in a reduced dimensional space, Genome Res. 12 (2002) 1112–1120.
[29] M. Mukherjee, P. Tamago, J.P. Mesirov, D. Slorim, A. Verni, T. Poggio, Support vector machine classification of microarray data, Technical Report, Cambridge: MIT, N.182, 1999.
[30] E. Purdom, S.P. Holmes, Error distribution for gene expression data, Stat. Appl. Genet. Mol. Biol. 4 (1) (2005) 16.
[31] A. Robinson, A new map projection: its development and characteristics, International Yearbook of Cartography 14 (1974) 145–155.
[32] R. Tagliaferri, A. Ciaramella, L. Milano, F. Barone, G. Longo, Spectral analysis of stellar light curves by means of neural networks, Astron. Astrophys. Suppl. Ser. 137 (1999) 391–405.
[33] R. Tagliaferri, N. Pelosi, A. Ciaramella, G. Longo, L. Milano, F. Barone, Soft computing methodologies for spectral analysis in cyclostratigraphy, Comp. Geosci. 27 (2001) 535–548.
[34] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, T.R. Golub, Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, Proc. Natl. Acad. Sci. USA 96 (1999) 2907–2912.

[35] P. Törönen, M. Kolehmainen, G. Wong, E. Castrén, Analysis of gene expression data using self-organizing maps, FEBS Lett. 451 (1999) 142–146.
[36] J.P. Townsend, D.L. Hartl, Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple treatments or samples, Genome Biol. 3 (2002), research0071.1-0071.16.
[37] J.P. Townsend, Resolution of large and small differences in gene expression using models for the Bayesian analysis of gene expression levels and spotted DNA microarrays, BMC Bioinform. 5 (2004) 54.
[38] G.C. Tseng, M.K. Oh, L. Rohlin, et al., Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects, Nucleic Acids Res. 29 (2001) 2549–2557.
[39] M.L. Whitfield, G. Sherlock, A.J. Saldanha, J.I. Murray, C.A. Ball, K.E. Alexander, J.C. Matese, C.M. Perou, M.M. Hurt, P.O. Brown, D. Botstein, Identification of genes periodically expressed in the human cell cycle and their expression in tumors, Mol. Biol. Cell 13 (2002) 1977–2000.
[40] R.D. Wolfinger, G. Gibson, E. Wolfinger, L. Bennett, H. Hamadeh, et al., Assessing gene significance from cDNA microarray expression data via mixed models, J. Comput. Biol. 8 (2001) 625–637.
[41] K.Y. Yeung, C. Fraley, A. Murua, A.E. Raftery, W.L. Ruzzo, Model based clustering and data transformations for gene expression data, Bioinformatics 17 (2001) 977–987.