RESEARCH ARTICLE OPEN ACCESS

Lung Cancer Diagnosis Based on Microarray Data by Using ART2 Network

Fadoua Rafii [1], Badr Dine Rossi Hassani [2], M'hamed Aït Kbir [3]
LIST Laboratory [1] & [3], LABIPHABE Laboratory [2]
University of Abdelmalek Essaadi (UAE)
Tangier - Morocco

ABSTRACT

The advent of Microarray technology has given birth to gene expression profiles on a genome-wide scale. This technology is of great importance for crucial issues such as drug discovery, diagnosis and prognosis of diseases and toxicological research. The application of this advanced technology has produced attractive information; due to its capability of measuring thousands of activities of genes simultaneously. Efficient techniques are urgently needed in order to analyze the generated gene expression data. The researchers find difficulties on interpreting the large amount of Microarray data. Developing reliable computational techniques represents the major challenge in functional genomics. The present article surveys the implementation of ART2 Neural Network in order to cluster and diagnose the presence of Lung cancer based on expression profiles. To illustrate the important issues of the proposed clustering procedure, Microarray data of Lung cancer are used to perform a comparison between results obtained by the ART2 and K-means technique. The present study investigates also the effect of using reduced data by applying the PCA technique and tries to increase the accuracy of clustering.

***Keywords* :-** Microarray, Gene expression data, ART2, K-means, PCA

I. INTRODUCTION

Thanks to the advances known in Microarray technology, it is possible to study the global gene expression patterns of tens of thousands of genes in parallel [1]. Tremendous studies have been realized, such comparing the gene expressions in normal and transformed cells of human in several rumors [2] and cells under divergent conditions or environments [3]. The continuous use of Microarray technology has produced large datasets representing the molecular information. Microarrays may address many hypotheses and the results require careful preparation for further analysis. The preprocessing of the generated data is an important task of Microarray analysis. Microarray data are affected by the natural biological variability and the several procedures that are involved in the Microarray experiment. Thus, it has been focused on implementing appropriate techniques to get rid of the noisy data [4]. For further stages of analysis, the significant signal is very desirable; because the data that are not useful could mask the importance of other worthwhile biological signals. One of the major challenges of the Bioinformatics field is to develop efficient algorithms to analyze the gene expression data. The analysis of gene expression data invokes the use of machine learning techniques such as clustering. The classification of genes into clusters basing on the expression profiles represents the most important technique for investigating Microarray data [5]. The purpose of this article is to shed light on the results obtained by applying different clustering approaches for gene expression data. Clustering tissues or experiments are valuable in order to identify the samples that exist in the different disease states, to discover, or to predict different cancer types, and to evaluate the effects of new drugs and therapies [6][7][8]. Due to the motivation of trying to emulate and understand the brain aptitudes, the neural networks have been developed. The researchers find the neural nets interesting in many areas and for tremendous reasons. The applications of neural nets have shown efficient results and promise for the challenged problems such as pattern recognition. The neural nets represent a practical technique for the clustering. And these networks have specifically been utilized in applications for numerous biological systems [9].

II. MICROARRAY TECHNOLOGY

The Microarray technology provides an exceptionally powerful way to explore gene expression. By using a Microarray, it is possible to evaluate the expression levels of thousands of genes across multiple developmental stages, specific clinical conditions or time points. It is useful for understanding gene functions, biological processes and effects of medical treatments. This technology has given birth to novel directions on studying the genome.

A. Microarray birth

The utilization of Microarray technology for studying gene expression profiles in biological samples began in 1995 [10]. This technology was first proposed in the late 1980s [11]. And the DNA Microarrays were firstly described in the literature by Augenlicht and his colleagues; they spotted 4000 complementary DNA (cDNA) sequences on nitrocellulose [12] [13]. It has been found that in a Microarray, many thousands of spots are located on a rectangular grid where each spot is composed by a large number of pieces of DNA from a specific gene [14]. This technology has become one of the essential tools that many researchers and specifically biologists use to control the levels of genes in a particular organism.

B. Microarray experiment

Microarray experiments have addressed tremendous scientific tasks.

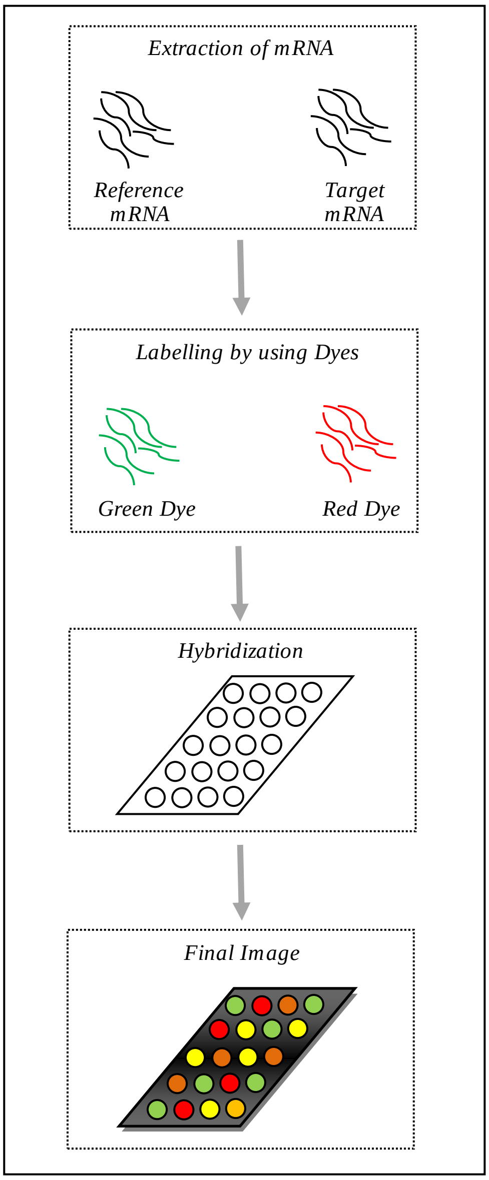

Fig. 1 Microarray experiment

The major tasks of the Microarray experiments are the identification of co-expressed genes, the recognition of sample or gene groups that have similar expression patterns, the identification of gene expression patterns that are highly differentiating by taking into consideration a set of discerned biological entities like in tumor types, and the exploration of gene activity patterns under different stress conditions as the chemical treatment.

The object of Microarray technology is to measure a cell's transcript through the abundance of mRNA molecules. This technique exploits a mighty feature of DNA duplex, which is represented by the sequence complementarity formed by the two strands. And this feature makes the hybridization feasible. The hybridization is a chemical reaction where the single-stranded of DNA or RNA molecules are combined in order to form double-stranded complexes.

III. GENE EXPRESSION DATA

A. Microarray Data Matrix

The gene expression data is described by an expression matrix which is characterized by *N x M*; where *N* is the dimension of genes and *M* represents the dimension of the samples or conditions involved on the Microarray experiment. The gene expression profile describes the expression values for one gene under many samples or experimental conditions. The array profile represents many genes across a single sample or condition. To depict the difference between the two profiles, with the equations (1) and (2) it has been defined the i^{th} gene profile of the gene expression matrix by row vector GP_i and the j^{th} array profile by column vector AP_j.

$$GP_i = (x_{i1}, x_{i2}, \ldots, x_{iM}) \qquad (1)$$

$$AP_j = (x_{1j}, x_{2j}, \ldots, x_{Nj}) \qquad (2)$$

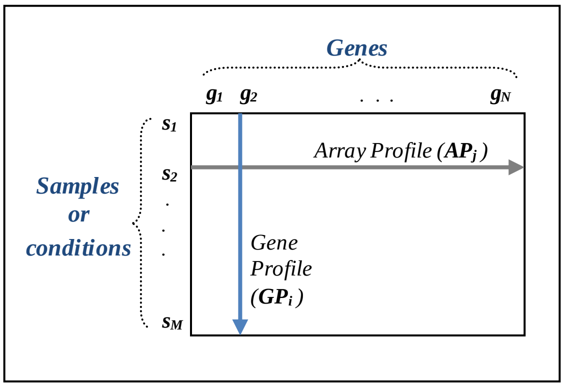

Fig. 2 Gene expression data matrix

B. Microarray Databases

Even though a Microarray experiment addresses a determined biological question, it could be followed by other experiments to evaluate supplementary factors. Thus, a Microarray experiment could be part of a huge study. For this reason, the Microarray databases have been created. These resources are able to manage the gene expression data resulted from different Microarray experiments. Some of the existent Microarray databases are:

- ArrayExpress is a public database for managing Microarray data; which provides access to Microarray data and presents an information base of gene expression profiles [15].
- caArray is a reachable open-source which manages the Microarray data. It takes into account the annotation of

Microarray data using MAGE-TAB and web-based forms [16].

- The Stanford Microarray Database is a research tool that is offering to hundreds of researchers the capability of storing, annotating, analyzing and sharing the data produced by Microarray technology [17].
- BASE stockpiles the data generated by multiple Microarray experiments, the annotations and the biomaterial information for some techniques or data formats are accessible [18].

IV. PROBLEMATIC

The Microarray technology is offering a global view on the activity levels of a huge amount of genes simultaneously. The typical gene expression dataset contains a number of genes which is usually larger than the number of experiments. For a simple organism such as the yeast, it has six thousand genes. And for humans, it is estimated that they have nearly thirty to forty thousand genes [19]. The objective of cluster analysis is to allocate objects to the adequate clusters; such that the objects belonging to the same cluster are more similar to each other. The most challenging area of computational Biology is to study thousands of genes across diverse conditions simultaneously [3]. The analysis of gene expression data is considered as an interesting research field in the Microarray technology exploration. Data mining techniques have proven to be useful in understanding gene function, gene regulation, cellular processes and subtypes of cells [20]. They represent good solutions to properly comprehend and interpret gene expression.

V. CLUSTERING

Clustering is a data mining technique that is applied in the object of placing the elements of data into proper groups in the absence of preliminary knowledge about the group definitions. The methods of clustering are implemented in tremendous engineering applications and scientific disciplines. Furthermore, clustering methods have been also used for the analysis of biological data.

Clustering is one of the most important data mining techniques for analyzing gene expression data; in view of the fact that it is able to determine the interesting patterns and the natural structures of data. The essential step consists on identifying a set of samples that present the same expression patterns across various genes into clusters. Thereby, it reveals the existing relations amongst the samples of the Microarray experiment. A cluster of samples could be defined as a group of biologically relevant samples that are similar based on the proximity measure.

Some of the techniques used for clustering are:

- K-means: it is restricted when the clusters are not of the same sizes, densities, and when the shapes are non-globular; it shows also complications when the data contains outliers [21].
- KNN: it has restrictions where the calculation becomes more complex because of the usage of all the training samples for the classification, the performance is depending only on the training set, and there is no weight distinction between samples [22].
- SOM: its principal disadvantage is that it needs necessarily to define a priori the structure of neural networks and the number of neurons in Kohonen layer [23]. And SOM are performing remarkably worse in terms of the quantization error and in recovering the clusters structure [24].

The appearance of Artificial Neural Networks (ANNs) has gained rising popularity for the applications where the relation between dependent and independent variables is very complex or unknown. This technique can be determined as a universal algebraic function that will differentiate the signal from the existing noise directly from experimental data. The ANNs applications have made worthwhile contribution to complex relationships; which makes them highly significant for studying the biological systems. The recent applications of ANNs comprise the analysis of expression profiles and genomic and proteomic sequences [25].

VI. IMPLEMENTED TECHNIQUES

A. *Adaptive Resonance Theory (ART)*

Adaptive Resonance Theory (ART) was introduced in 1976 by Stephen Grossberg [26]. The networks created by ART are unsupervised and self-organizing [11]. Besides, ART is capable of forming stable clusters of random sequences of input patterns by learning or setting resonant states and self-organizing [11]. The adaptive resonance theory nets allow the user to assess the degree of patterns similarity on the same cluster.

1) ART derivatives:

Adaptive Resonance Theory (ART) networks represent a family of neural networks that is based on resonance. Thus, the resonant state of a neural network is when a category prototype vector suits closer to the current input vector [27].

TABLE I
THE ART FAMILY

ART version	Description
ART1	It is the version of ART networks that accepts just binary inputs [28].
ART2	It is able to support the continuous inputs [29].
ART 2A	It is a streamlined form of the ART2 version accompanied by an accelerated runtime, and qualitative results that are rarely inferior to the entire ART2 implementation [30].
ART3	It is based on ART2 by using the simulation of the neurotransmitter regulation of synaptic activity [31].
Fuzzy ART	It implements the fuzzy logic into the recognition of ART patterns in the object of improving the generalizability. The most useful feature of fuzzy ART is to use the complement coding; which is a way to incorporate the absence of pattern elements in the classifications [32].
ARTMAP	It is known as Predictive ART. And it combines lightly the two adjusted ART1 or

	ART2 units into a supervised learning structure; such the first unit takes the input data and the second unit acquires the correct output data, after that in order to make the correct classification, it tries to make the minimum possible regulation of the vigilance parameter in the first unit [32].
Fuzzy ARTMAP	It is simply ARTMAP with the application of fuzzy ART units in order to increase the efficacy [33].

2) *ART2:*

ART2 is designed in order to perform the input vectors that are characterized by continuous values. The typical architecture of ART2 is depicted in Figure 3. ART2 contains two layers F_1 and F_2. For the F_1 layer, it is composed of six types of units: *W, X, U, V, P* and *Q*. The input pattern is characterized by the dimension *n*; there are *n* units of each unit type. And there are supplemental units between:

- *W* and *X* units
- *U* and *V* units
- *P* and *Q* units

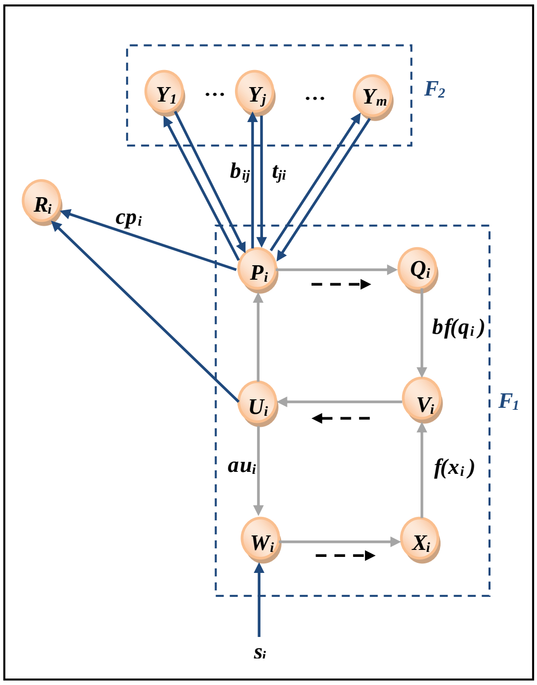

Fig. 3 ART2 Architecture

The role of the supplemental unit is to receive signals from a specific type of units, and then it computes the vector norm and sends this (inhibitory) signal to the receiver units. And each of these receives an excitatory signal from the corresponding units. Whereas, the F_2 layer consists of the Y_j units that compete in a mode of winner-take-all in order to learn each input pattern. On the other side, there is a relation between the F_1 layer and the F_2 layer, where it is illustrated by the connections between the P_i units and the Y_j units. And these connections show the weights b_{ij} and t_{ji} that multiply the signal transferred over those paths. For the winning F_2 unit, there is an activation denoted d where $0 < d < 1$.

TABLE III
ART2 PARAMETERS

Parameter	Role
n	The number of the input units
m	The defined number of cluster units
a, b	The fixed weights in the F_1 layer
c	The fixed weight that is used in testing for reset
d	The activation of winning F_2 unit
e	Small parameter used for preventing the division by zero when the vector norm is zero
θ	The parameter of noise suppression, where the sample value is $\theta = \frac{1}{\sqrt{n}}$
α	The learning rate where the small value slows the learning and ensures that the weights reach equilibrium
ρ	The vigilance parameter

The ART2 algorithm requires the following calculations for updating the F_1 activations:

$$u_i = \frac{v_i}{e + \|v\|} \qquad (3)$$

$$w_i = s_i + au_i \qquad (4)$$

$$p_i = u_i + dt_{Ji} \qquad (5)$$

$$x_i = \frac{w_i}{e + \|w\|} \qquad (6)$$

$$q_i = \frac{p_i}{e + \|p\|} \qquad (7)$$

$$v_i = f(x_i) + bf(q_i) \qquad (8)$$

The activation function is defined by:

$$f(x) = \begin{cases} x & if \quad x \geq \theta \\ 0 & if \quad x < \theta \end{cases} \qquad (9)$$

The ART2 algorithm is depicted by the following steps:

Step 1. Initialization of the parameters:
a, b, θ, c, d, e, α, ρ

Step 2. Executing the steps 3-13 for a specified number of epochs

Step 3. For each input vector s, doing the steps 4-12

Step 4. Updating the activations of F_1 unit:

$$w_i = s_i \qquad q_i = 0$$

$$u_i = 0 \qquad x_i = \frac{s_i}{e + \|s\|}$$

$$p_i = 0 \qquad v_i = f(x_i)$$

Updating the activations of F_1 unit again:

$$u_i = \frac{v_i}{e+\|v\|} \qquad w_i = s_i + au_i$$

$$p_i = u_i \qquad x_i = \frac{w_i}{e+\|w\|}$$

$$q_i = \frac{p_i}{e+\|p\|} \qquad v_i = f(x_i) + bf(q_i)$$

Step 5. Computation of signals for F_2 units:

$$y_i = \sum_i b_{ij} p_i$$

Step 6. While the reset is true, do the steps 7-8

Step 7. Finding Y_J of the F_2 unit which has the largest signal; by defining J such that $y_J \geq y_j$ with j=1, …, m

Step 8. Checking for reset:

$$u_i = \frac{v_i}{e+\|v\|} \qquad p_i = u_i + dt_{Ji}$$

$$r_i = \frac{u_i + cp_i}{e+\|u\|+c\|p\|}$$

If $\|r\| < \rho - e$, then

$$y_J = -1$$

If $\|r\| \geq \rho - e$, then

$$w_i = s_i + au_i$$

$$x_i = \frac{w_i}{e+\|w\|}$$

$$q_i = \frac{p_i}{e+\|p\|}$$

$$v_i = f(x_i) + bf(q_i)$$

Step 9. Doing the steps 10-12 for the number of iterations:

Step 10. Updating weights for the winning unit J:

$$t_{Ji} = \alpha du_i + \{1 + \alpha d(d-1)\} t_{Ji}$$

$$b_{iJ} = \alpha du_i + \{1 + \alpha d(d-1)\} b_{iJ}$$

Step 11. Updating F_1 activations:

$$u_i = \frac{v_i}{e+\|v\|}$$

$$w_i = s_i + au_i$$

$$p_i = u_i + dt_{Ji}$$

$$x_i = \frac{w_i}{e+\|w\|}$$

$$q_i = \frac{p_i}{e+\|p\|}$$

$$v_i = f(x_i) + bf(q_i)$$

Step 12. Testing the stop condition for the weight updates

Step 13. Testing the stop condition for the number of epochs

B. K-means

K-means [34] is a classic algorithm for clustering where *k* is the number of clusters. In k-means algorithm, there are the centroids that represent the clusters; which are the centers of clusters. The k-means algorithm is implemented by minimizing the sum of distances between each object and the corresponding cluster centroid.

The K-means clustering is a form of partitioning method [21]. And the function partitions the dataset containing *N* objects into *k* subsets S_j that are disjoint and contain N_j items; where they are as close to each other as possible basing on a given distance measure. In the partition, each cluster is determined by its N_j objects and by its centroid λ_j. The centroid of each cluster is represented by the point to which the sum of distances from all objects in that cluster is minimized. Consequently, the k-means algorithm could be defined as an iterative process in order to minimize the following equation:

$$E = \sum_1^k \sum_{n \in S_j} d(x_n, \lambda_j) \qquad (10)$$

- x_n represents the vector with the n^{th} object
- λ_j represents the centroid of the object in S_j
- *d* represents the distance measure.

The k-means algorithm moves objects between the clusters till E could not be reduced further. The algorithm works by choosing randomly *k* centroids. Then all the objects are allocated to the cluster such that its centroid is the nearest to them. For the new cluster centroid, it requires to be updated; in order to take into consideration the objects that were added or removed to the updated cluster. This running continues till that the objects won't further change their cluster membership.

C. PCA and SVD

PCA (Principal Component Analysis) is a technique for reducing the dimension of multivariate data. The major aims of PCA [35] consist of:

- summarizing the patterns of correlation
- reducing the number of variables
- providing the basis for the predictive models

The gene clustering has been characterized by effectiveness due to the utilization of PCA [36]. In order to reduce the Microarray data, the principal component analysis (PCA) method was employed. This technique is able to discover the variables that are correlated with one another, and the correlated subsets are combined into components. PCA gives indications on how would be the further reduction of the dimensionality, with the advantage of not losing the essential information.

Singular value decomposition (SVD) and principal component analysis (PCA) are common techniques that perform the analysis of multivariate data.

Considering an *m x n* matrix denoted *X* and contains real valued data. The equation of the SVD of *X* is:

$$X = USV^{T} \qquad (11)$$

- U is a matrix of $m \times n$, and its columns are called the left singular vectors.
- S is a diagonal matrix of $n \times n$. Its elements are called the singular values and are nonzero on the diagonal.
- V_T is a matrix of $n \times n$. Its elements form an orthonormal basis.
- The SVD of X generates two orthonormal bases:
- The first base is defined by the singular vectors $\{U_k\}$
- The second base is determined by the singular vectors $\{V_k\}$

VII. CLUSTERING MICROARRAY DATA

A. The Microarray dataset

The selected dataset concerns Lung cancer of human gene expression. This dataset has been published by Gordon et al. [37], and was used for classifying two types of lung cancer; which are malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA). From this dataset, it has been formed another one to apply the clustering techniques:

- 62 tissue samples where 31 correspond to MPM and 31 to ADCA;
- Each sample is described by 12533 genes.

B. Results

The samples of the two clusters MPM and ADCA have been presented to the ART2 algorithm as input patterns with definition of the number of clusters. The advantage of the ART is that in the execution step of the algorithm, it's remarkable that the first input pattern of the dataset is allocated to the first created node. After that, when the following input patterns are not similar to the previous ones, another node is added.

To assess the performance of the used algorithm, the accuracy was calculated:

$$Acc = \left({}^{N_c}\!/\!{}_{N_t}\right) * 100 \qquad (12)$$

- Acc is the accuracy representing the performance measure of the clusters
- N_c is the number of instances that are clustered correctly
- N_t is the total number of instances

In order to compare the effect of reducing a large dataset into a smaller set, the PCA technique was applied. The object is to get new components without losing the important information; where each component is representing a linear combination of the original ones, and it is called a linear projection. After the execution of PCA algorithm, the variation of the cumulative variance explained with the number of principle components has been traced; which is depicted on the Figure 4. The traced variation shows that the value of cumulative variance is more than 94% when the number of the principle components is 13. Basing on the results represented on the Figure 4, thirteen principle components have been retained for representing the genes of the Microarray dataset.

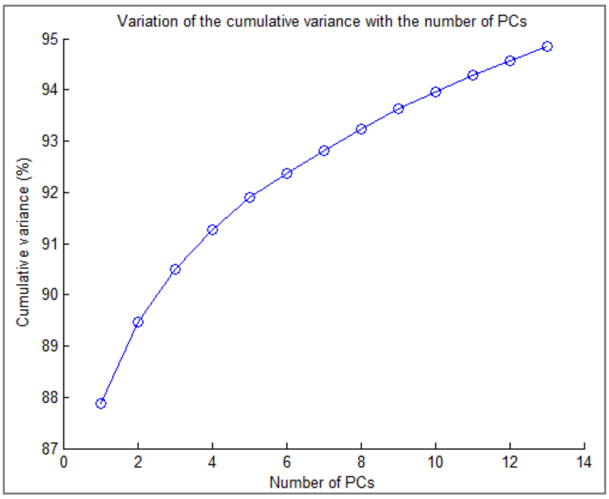

Fig. 4. The percent of cumulative variance contained by each principal component. The 13 first principal components contain more than 94% of the variance

The resulted reduced variable set was used as input to ART2 algorithm. In the object of comparing the effect of the reducing Microarray data, both the original dataset and the reduced one were presented to ART2 algorithm, and the accuracies have been calculated. For the ART2 parameters, the same values were chosen for all the tests:

- a=0.9; b=0.9;
- c=0.2; d=0.99; e=0.9;
- α=0.2;
- ρ=0.9;
- m=2;
- $\theta = \frac{1}{\sqrt{n}}$

The obtained values are depicted on the following table:

TABLE IIIII
ACCURACY OF ART2 AND K-MEANS

Number of iterations	Dataset dimension	ART2 Accuracy (%)	K-means Accuracy (%)
1	12533	80.64	61.29
5	12533	87.10	70.97
1	13	87.10	67.74
5	13	93.55	66.13

The results show the efficiency of the ART2 algorithm against the K-means one. In fact, the ART2 algorithm provides high accuracy even with few numbers of iterations. Indeed, in the first test, where raw data is used, the Microarray dataset was presented to the two algorithms with the number of iterations equal to 1, large difference between the two results is noticeable. On the second test, where reduced data are used, the difference between the two techniques is also significant.

Besides, it is remarkable that ART2 technique has given more accurate results after the application of PCA technique. Thus, the ART2 inputs need to be cautiously selected; the application of PCA technique has offered an efficient supervised selection of ART2 inputs.

VIII. CONCLUSION

The cancer clustering based on Microarray data has resulted attractive information. Microarray experiments produce huge amount of data. And the majority of Microarray data matrixes are characterized by a very large number of genes. For this reason, the clustering represents a very challenging task. The present work develops a procedure by taking advantages of the adaptive resonance theory neural network. The followed method used also the PCA technique for reducing the Microarray dataset dimensionality. It combines between PCA and SVD to explore better the multivariate data. The right decision of the numbers of PCs has been taken. And the resulted set was used as input for ART2 algorithm. The strategy used has shed light on the importance of ART2 compared to K-means technique. The efficiency of ART2 technique is also illustrated by the measurements of accuracy clustering utilizing the lung cancer gene expression dataset. For this study, the coherence survey of ART2 technique was used for a predefined number of clusters. In the next work, we are looking forward for developing a strategy without a priori determination of the clusters number.

REFERENCES

[1] P. O. Brown and D. Botstein, “Exploring the new world of the genome with DNA microarrays”, *Nature genetics*, vol. 21, pp. 33–37, 1999.

[2] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, “Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays”, *Proceedings of the National Academy of Sciences*, vol. 96, no. 12, pp. 6745–6750, 1999.

[3] M. G. Schueler, A. W. Higgins, M. K. Rudd, K. Gustashaw, and H. F. Willard, “Genomic and genetic definition of a functional human centromere”, *Science*, vol. 294, no. 5540, pp. 109–115, 2001.

[4] Fadoua Rafii, M. Aït Kbir and B. D. Rossi, “Data Preprocessing and Reducing for Microarray Data Exploration and Analysis”, *International Journal of Computer Applications*, vol. 132, no. 16, 2015.

[5] K. Hakamada, M. Okamoto, and T. Hanai, “Novel technique for preprocessing high dimensional time-course data from DNA microarray: mathematical model-based clustering”, *Bioinformatics*, vol. 22, no. 7, pp. 843–848, Apr. 2006.

[6] A. Alizadeh et al., “Distinct types of diffuse large B-cell Lymphoma identified by gene expression profiling”, Nature, vol. 403, pp. 503–511, 2000.

[7] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander, “Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring”, *Science*, vol. 286, pp. 531–537, 1999.

[8] U. Scherf, D. Ross, M. Waltham, L. Smith, J. Lee, L. Tanabe, K. Kohn, W. Reinhold, T. Myers, D. Andrews, D. Scudiero, M. Eisen, E. Sausville, Y. Pommier,D. Botstein, P. Brown, and J.Weinstein, “A gene expression database for the molecular pharmacology of cancer”, *Nature Genetics*, vol. 24, no. 3, pp. 236–244, 2000.

[9] J. S. Almeida, “Predictive non-linear modeling of complex data by artificial neural networks”, *Current Opinion in Biotechnology*, vol. 13, no. 1, pp. 72–76, Feb. 2002.

[10] Lobenhofer EK, et al, “Progress in the application of DNA microarrays”, *Environ Health Perspect*, vol. 109, no. 9, pp. 881-891, 2001.

[11] Jizhong Zhou and Dorothea K. Thompson, “Microaarray technology and applications in environmental microbiology”, *Advances in Agronomy*, Volume 82, 2004.

[12] L. Augenlicht, M. Wahrman, H. Halsey, L. Anderson, J. Taylor and M. Lipkin, “Expression of cloned sequences in biopsies of human colonic tissue and in colonic–carcinoma cells induced to differentiate invitro”, *Cancer Res*, vol. 47, pp. 6017–6021, 1987.

[13] L. Augenlicht, J. Taylor. L. Anderson and M. Lipkin, “Patterns of gene expression that characterize the colonic mucosa in patients at genetic risk for colonic cancer”, *Proc. Nat. Acad. Sci. 88*, pp. 3286–3289, 1991.

[14] C. Naidu and Y. Suneetha, “Review Article: Current Knowledge on Microarray Technology - An Overview”, *Tropical Journal of Pharmaceutical Research*, vol. 11, no. 1, Mar. 2012.

[15] A. Brazma, M. Kapushesky, H. Parkinson, U. Sarkans, and M. Shojatalab, “[20] Data Storage and Analysis in ArrayExpress”, *in Methods in Enzymology*, vol. 411, Elsevier, pp. 370–386, 2006.

[16] https://array.nci.nih.gov/caarray/home.action

[17] Janos Demeter, Catherine Beauheim, Jeremy Gollub, Tina Hernandez-Boussard, Heng Jin, Donald Maier, John C. Matese, Michael Nitzberg, Farrell Wymore, Zachariah K. Zachariah, Patrick O. Brown, Gavin Sherlock and Catherine A. Ball, “The Standford Microarray Database:

implementation of new analysis tools and open source release of software", pp. D766-D770, 2006.

[18] Jason Comander, Griffin M. Weber, Michael A. Gimbrone, Jr, et al., "Argus-A New Database System for Web-Based Analysis of Multiple Microarray Data Sets", pp. 1603-1610, 2001.

[19] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, and others, "Initial sequencing and analysis of the human genome", *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[20] R. Das, Coherent Gene Expression Pattern Finding Using Clustering Approaches, Ph.D. dissertation, Dept. Of Computer Science and Engineering, Tezpur University, 2010.

[21] X. Amatriain, A. Jaimes, N. Oliver, and J. M. Pujol, "Data Mining Methods for Recommender Systems", in Recommender Systems Handbook, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Boston, MA: Springer US, pp. 39–71, 2011.

[22] N. Suguna and K. Thanushkodi, "An improved k-nearest neighbor classification using genetic algorithm", *International Journal of Computer Science Issues*, vol. 7, no. 2, pp. 18–20, 2010.

[23] I. Mokriš and R. Forgáč, "Decreasing the feature space dimension by Kohonen self-organizing maps", *in 2nd Slovakian–Hungarian Joint Symposium on Applied Machine Intelligence*, 2004.

[24] Flexer, "Limitations of self-organizing maps for vector quantization and multidimensional scaling", *Advances in neural information processing systems*, pp. 445–451, 1997.

[25] J. S. Almeida, "Predictive non-linear modeling of complex data by artificial neural networks", *Current Opinion in Biotechnology*, vol. 13, no. 1, pp. 72–76, Feb. 2002.

[26] Stephen Grossberg, "Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors", *Biological Cybernetics*, vol. 23, pp.121-134, 1976.

[27] K. Sarvesh, R. P. Singh, A. Mishra, and K. Hemant, "Computation of Neural Network using C# with Respect to Bioinformatics", *International Journal of Scientific and Research Publications*, vol. 3, Issue 9, September 2013.

[28] http://techlab.bu.edu/files/resources/articles_cns/carpenter_grossberg_SRT_2003.pdf

[29] S. Grossberg, "Competitive learning: From interactive activation to adaptive resonance", Cognitive science, vol. 11, no. 1, pp. 23–63, 1987.

[30] G.A. Carpenter & S. Grossberg, "ART 2: Self-organization of stable category recognition codes for analog input patterns", *Applied Optics*, vol. 26, no. 23, pp. 4919–4930, 1987.

[31] G.A. Carpenter, S. Grossberg & D.B. Rosen, "ART 2-A: An adaptive resonance algorithm for rapid category learning and recognition", Neural Networks, vol. 4, pp. 493–504, 1991

[32] G.A. Carpenter & S. Grossberg, "ART 3: Hierarchical search using chemical transmitters in self-organizing pattern recognition architectures", Neural Networks, vol.3, pp. 129–152, 1990.

[33] http://gemi.mpl.ird.fr/PDF/Lek.EM.1999.pdf

[34] J. MacQueen, "Some methods for classification and analysis of multivariate observations", *In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281-297, 1965.

[35] Tabachnick B.G., Fidel L.S., "Using Multivariate Statistics 3rd Edition", *Harper Collins College Publisher*, pp. 635-708, 1996.

[36] Yeung K.Y., Ruzzo W.L., "Principal component analysis for clustering gene expression data", *Bioinformatics*, vol. 17, pp. 763-774, 2001.

[37] Gordon GJ, Jensen RV, Hsiao LL, Gullans SR, Blumenstock JE, Ramaswamy S, Richards, WG, Sugarbaker DJ, Bueno R., "Translation of Microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma", *Cancer Res*, 62: pp. 4963–4967, 2002.