International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 7 228 – 235

Clustering Approaches for Evaluation and Analysis on Formal Gene Expression Cancer Datasets.

Ramachandro Majji
Dept of CSE, GMRIT
Rajam, Andhra Pradesh, India.
e-mail: rama00565@gmail.com

Ravi Bramaramba
Dept of IT, GIT
GITAM University,
Visakhpatnam, AP, India.

***Abstract*:-** Enormous generation of biological data and the need of analysis of that data led to the generation of the field Bioinformatics. Data mining is the stream which is used to derive, analyze the data by exploring the hidden patterns of the biological data. Though, data mining can be used in analyzing biological data such as genomic data, proteomic data here Gene Expression (GE) Data is considered for evaluation. GE is generated from Microarrays such as DNA and oligo micro arrays. The generated data is analyzed through the clustering techniques of data mining. This study deals with an implement the basic clustering approach K-Means and various clustering approaches like Hierarchal, Som, Click and basic fuzzy based clustering approach. Eventually, the comparative study of those approaches which lead to the effective approach of cluster analysis of GE.The experimental results shows that proposed algorithm achieve a higher clustering accuracy and takes less clustering time when compared with existing algorithms.

***Keywords**: K-Means, Gene expression, Self-organizing map & Hierarchical Clustering*

*****

I. INTRODUCTION

Microarrays have emerged as a widely used technology for the monitoring of the expression levels of thousands of genes during various biological[1] processes and functions. Extracting the hidden information in this huge volume of gene[2] expression data is quite challenging, and therefore the need for computationally efficient methods to mine GE data is a thrust area for the research community.

Moreover, it is a fact that, because of the complexity of the underlying biological processes, GE data attained from DNA microarray technologies are mostly noisy and have very high dimensionality. This scenario makes mining[3] of such data very tough and challenging for prediction[4]. Several data mining techniques have been used to address the above mentioned challenge and clustering is one of the most popular tools found capable in analysing the gene [5] expression data with better accuracy. Clustering [6] Techniques (CT) identify the inherent natural structures and the interesting patterns in the dataset [7].

The purpose of clustering GE data is to reveal the natural structure inherent in the data. A good CT should depend as little as possible on prior knowledge. CT for gene expression [8] data should be capable of extracting useful information from noisy data. GE data are often highly connected and may have intersecting and embedded patterns. Therefore, algorithms for gene-based clustering [9] should be able to handle this situation effectively. Finally, biologists are not only interested in the clusters of genes, but also in the relationships among the clusters and their sub-clusters, and the relationship among the genes within a cluster.

CT provides some graphical representation of the cluster structure, which is much favoured by the biologists to approach the optimal solution, or a set of approximate solutions to a range of problems in specific areas.

II. LITERATURE REVIEW

This chapter provides overview of research carried out on clustering algorithm[10] and their application to several microarray data reported in literature.

This chapter is broadly divided into two section. First section deals with research work carried out on several microarray data and section two deals with research carried out on different clustering algorithms.

DNA microarrays are high-throughput methods for analysing complex nucleic acid samples. It makes possible to measure rapidly, efficiently and accurately the levels of expression of all genes present in a biological sample. The application of such methods in diverse experimental conditions generates lots of data. However, the main problem with these data occurs while analysing it. Derivation of meaningful biological information from raw microarray data is impeded by the complexity and vastness of the data. To overcome the problem associated with GE microarray data many statistical methods has been proposed in recent past. Some important has been explained below:

Biological data are being produced at a phenomenal rate. For example as of April 2001, the GenBank repository of nucleic acid sequences contained 1, 15, 46,000 entries and the SWISSPROT database of protein sequences contained 95,320 entries. On an average, these databases are doubling in size every 15 months.

The application of self-organizing maps, a type of mathematical cluster analysis [11] that is particularly well suited for recognizing and classification of the features in complex, multidimensional data. The approved method is the nation to exhibit and has been implemented in a

publicly stature available computer package, gene cluster, which performs the analytical calculations and provides easy data visualization.

The main types of data analysis needed to for biomedical applications include:

Clustering:: Finding new biological classes or refining existing ones.

Gene Selection:: In mining terms this is a process of attribute selection, which finds the genes most strongly related to a particular class.

Classification:: Classifying diseases or predicting outcomes based on gene expression patterns and perhaps even identifying the best treatment for given genetic signature.

Microarrays are a revolutionary new technology with great potential to provide accurate medical diagnostics help find the right treatment and cure for many diseases and provide a detailed genome-wide molecular portrait of cellular states.

DNA microarray makes it possible to quickly, efficiently and accurately measure the relative representation of each mRNA species in the total cellular mRNA population. A DNA experiment consists of measurements of the relative representation of a large number of mRNA species.

Elucidating the patterns hidden in gene expression data offers a tremendous opportunity for an enhanced understanding of functional genomics. Many conventional clustering algorithms have been adapted or directly applied to gene expression data, and also new algorithms have recently been proposed specifically aiming at gene expression data. These clustering algorithm relevant groups of genes and samples. In this study the gene expression clustering is divide into gene based clustering and sample based clustering. It is explained that K-means a partition based clustering where no of clusters has to be mentioned earlier, in hierarchical clustering it produces a genogram, in SOM it requires the grid structure earlier before clustering

Current approaches to clustering gene expression patterns utilize hierarchical methods or methods that work for Euclidean distance metrics. A graph theoretic approach is considered, and made no assumptions on the similarity function or the number of clusters sought. The cluster structure is produced directly, without involving an intermediate tree stage.

III. PROPOSED STATEMENT

CT have proven to be helpful to understand gene function, gene regulation, cellular processes and subtypes of cells. Genes with similar expression patterns can be clustered together with similar cellular functions. Furthermore, co-expressed genes in the same cluster are likely to be involved in the same cellular processes, and a strong correlation of expression patterns between those genes indicates co-regulation. Searching for common DNA sequences at the promoter regions of genes within the same cluster allows regulatory motifs specific to each gene cluster to be identified and cis-regulatory elements to be proposed. The inference of regulation through the clustering of gene expression data also gives rise to hypotheses regarding the mechanism of the transcriptional regulatory network. Finally, clustering different samples on the basis of corresponding expression profiles may reveal sub-cell types which are hard to identify by traditional morphology-based approaches.

IV. INTRODUCTION TO MICRO ARRAY TECHNOLOGY

4.1 MEASURING MRNA LEVELS

Compared with the traditional approach to genomic research, which has focused on the local examination and collection of data on single genes, microarray technologies have now made itpossible tomonitor the expression levels for tens of thousands of genes in parallel. The two major types of microarray experiments are the cDNA microarray and oligonucleotide arrays. Despite differences in the details of their experiment protocols, both types' ofexperiments involve three common basic procedures:

4.1.1Chip manufacture:

A microarray is a small chip, onto which tens of thousands of DNA molecules are attached in fixed grids. Each grid cell relates to a DNA sequence.

The Target Preparation, Labelling & Hybridization [12]:

Typically, two mRNA samples are reverse-transcribed into cDNA, labelled using either fluorescent dyes or radioactive isotopic, and then hybridized with the probes on the surface of the chip.

The scanning process:

Chips are scanned to read the signal intensity that is emitted from the labelled and hybridized targets. Generally, both cDNA microarray and oligo chip experiments measure the expression level for each DNA sequence by the ratio of signal intensity between the test sample and the control sample. Therefore, data sets resulting from both methods share the same biological semantics. In this study, unless explicitly stated, both the cDNA microarray data[13] and the oligo chip as microarray technology and term the measurements collected via both methods as gene expression data.

4.2 PRE-PROCESSING OF GE DATA

A microarray experiment typically assesses a large number of DNA sequences under multiple conditions. These conditions may be a time series during a biological process or a collection of different tissue samples. In this, the focus on the cluster analysis[14] of GE data without making a distinction among DNA sequences, which will uniformly be

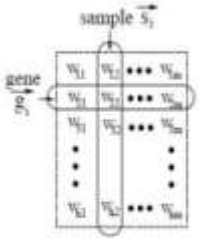
 ISSN: 2321-8169
228 – 235

called "*genes*".

Similarly and uniformly referred to all kinds of experimental conditions as "*samples*" if no confusion will be caused. A gene expression data set from a microarray experiment can be represented by a real-valued
E*xpression-matrix* [15]

$$M = \{W_{ij} \mid i \le i \le n, 1 \le j \le m\}$$

where the rows $G = \{\underset{g_1}{\rightarrow}, \underset{g_2}{\rightarrow}, \underset{g_3}{\rightarrow} \dots \dots\}$ form the expression patterns of genes, the columns $S = \{\underset{s_1}{\rightarrow}, \underset{s_2}{\rightarrow}, \underset{s_3}{\rightarrow} \dots \dots\}$ represent the expression profiles of samples, and each cell w_{ij} is the measured expression level of gene as described in the figure.



N	Number of genes
n	Number of samples
M	A gene expression matrix
W_{ij}	Each cell in gene expression Matrix
$\vec{g_i}$	A gene
$\vec{s_j}$	A sample
G, G1, G2.....	A set of genes
S, S1, S2.....	A set of samples

The original GE matrix obtained from a scanning process contains noise, missing values and systematic variations arising from the experimental procedure. Data pre-processing is indispensable before any cluster analysis can be performed. Some problems of data pre-processing have themselves become interesting research topics.

The questionnaire are beyond the scope of this survey; an examination of the problem of missing value estimation and the problem of data normalization. Furthermore, many clustering approaches apply one or more of the following pre-processing procedures:

a) Filtering out genes with expression levels which do not change significantly across samples;
b) Performing a mathematical or a logarithmic transformation function of each expression level; (or)
Standardizing each row of the gene expression matrix with a mean of zero and a variance of one.

V. CLUSTERING

In this subsection, the concept of clusters and clustering preceded with the division of the clustering tasks for gene expression data into three categories according to different clustering purposes.

5.1 CLUSTERS AND CLUSTERING

Clustering is the process of grouping data objects into a set of disjoint classes, called clusters; so that objects within a class have high similarity to each other, while objects in separate classes are more dissimilar. Clustering is an example of unsupervised classification. Classification refers to a procedure that assigns data objects to a set of classes. Thus, clustering is distinguished from pattern recognition or the areas of statistics known as discriminant analysis and decision analysis, which seek to find rules for classifying objects from a given set of pre-classified objects.

5.2 CATEGORIES OF GE DATA

Currently, a typical microarray experiment contains 10^3 to 10^4 genes, and this number is expected to reach to the order of 10^6. However, the number of samples involved in a microarray experiment is generally less than 100.

One of the characteristics of gene expression data is that it is meaningful to cluster both genes and samples. Besides, the co-expressed genes can be grouped in clusters based on their expression patterns. In such **gene-based clustering**, the genes are treated as the objects, while the samples are the features.

The samples can be partitioned into homogeneous groups. Each group may correspond to some particular macroscopic phenotype, such as clinical syndromes or cancer types such **sample-based clustering** regards the samples as the objects and the genes as the features. The distinction of gene-based clustering and sample based clustering is based on different characteristics of clustering tasks for gene expression data.
Some clustering algorithms, such as K-means and hierarchical approaches, can be used both to group genes and to partition samples.

Both the gene-based and sample-based clustering approaches search exclusive and exhaustive partitions of objects that share the same feature space.

However, current thinking in molecular biology holds that only a small subset of genes participates in any cellular process of interest and that a cellular process takes place only in a subset of the samples. This belief calls for the **subspaceclustering** to capture clusters formed by a subset of genes across a subset of samples. For subspace clustering algorithms, genes and samples are treated symmetrically, so that either genes or samples can be regarded as objects or features. Furthermore, clusters generated through such algorithms may have different feature spaces.

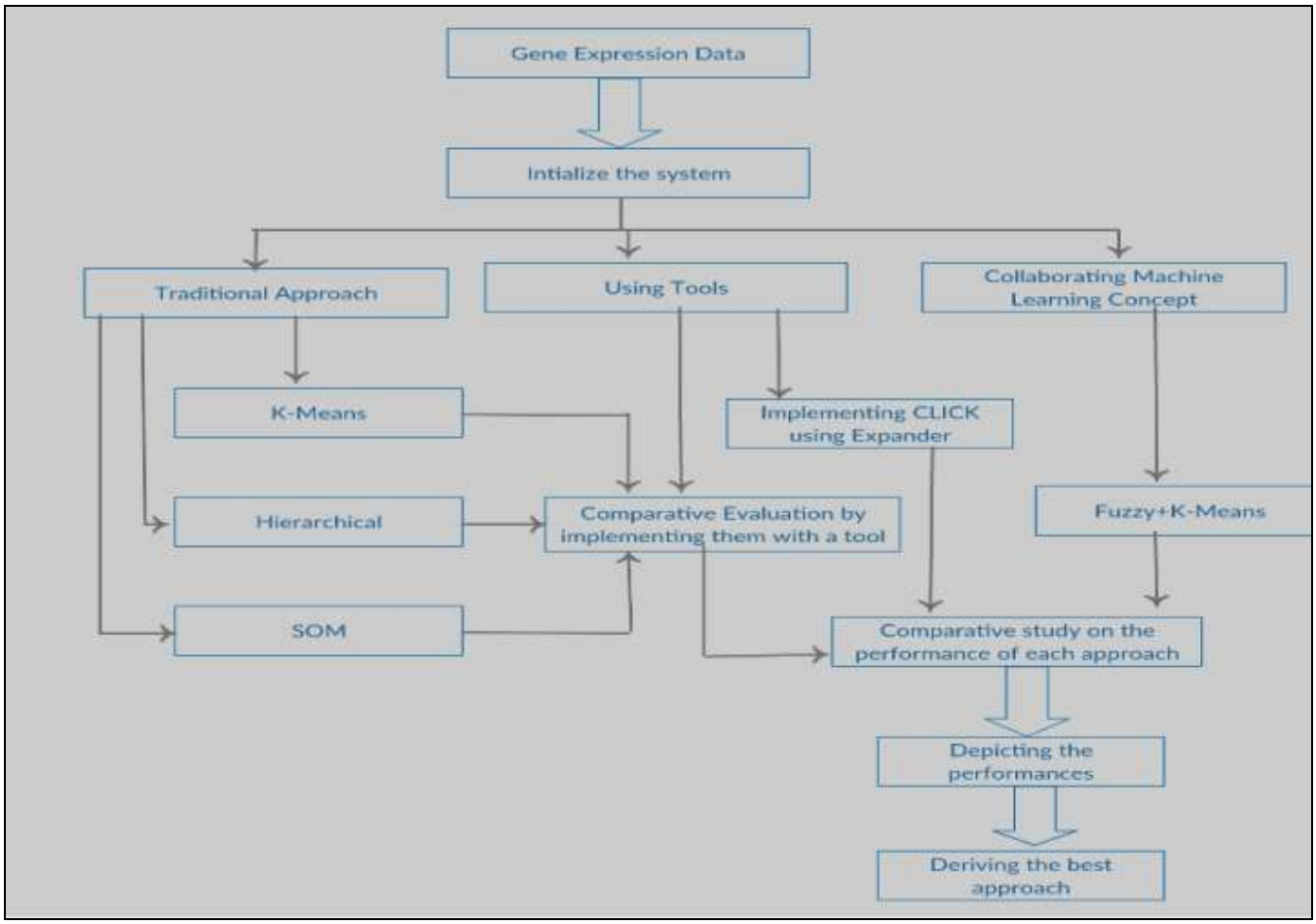
While a GE matrix can be analysed from different angles, the gene-based, sample based clustering and subspace clustering analysis[16] face very different challenges. Thus, there is a scope to adopt very different computational strategies in the three situations.

VI. METHODOLOGY

The outline of our system or simply architecture of our system is depicted below.
The major components of our system involves

 ISSN: 2321-8169
228 – 235

- Gene Expression Data,
- K-Means[17]
- Hierarchal,
- SOM [18],
- CLICK & Fuzzy based approach



As shown in the outline, the steps followed to implement are:

- First move on with the gathering of GE which is thoroughly pre-processed.
- In the second step, give the data into the system i.e.., platform
- Now the system will load the data
- After loading, run the K-means clustering algorithm on it.
- By depicting the results of it the performance of the algorithm will be noticed
- The above two steps will be performed for algorithms Hierarchal & SOM.
- Now using the Cluster 3.0 all the above executed algorithms will be executed in it.
- Later a comparative study [19] among them will be performed.
- Now an advanced clustering approach has to kept for study. For that, Expander tool is used to implement CLICK clustering algorithm on the considered data.
- To incorporate a new approach, a collaborative technique has to be implemented. For that, a machine learning[20] concept fuzzy is collaboratively used with K-means for the scope of better performance.
- Eventually all the algorithms will be kept under for a comparative study.
- Out of that study among all the above a best approach can be derived.

VII. IMPLEMENTATION PROCEDURE

The K-Means algorithm is a typical partition-based clustering method. Given a pre-specified number K, the algorithm partitions the data set into K disjoint subsets which optimize the following objective function:

$$E = \sum_{i=1}^{k} \sum_{O \in C_i} |O - \mu_i|^2$$

Here, O is a data object in cluster *Ci* and *mui*is the centroid [21] (mean of objects) of *Ci*. Thus, the objective function E tries to minimize the sum of the squared distances of objects from their cluster centres.

7.1 Algorithm:

- The K-Means algorithm accepts the "number of clusters" to group data into and the dataset to cluster the input values.
- The K-Means algorithm then creates the ¯rst k initial clusters from the data set
- The K-Means algorithm calculates the arithmetic mean of each cluster formed in the data set. The arithmetic mean is the mean of all the individual records in the cluster.
- Next K-Means assigns each record in the dataset to only one of the initial clusters. Each record is assigned to the nearest cluster using proximity measure like Euclidean distance.
- K-Means reassigns each record in the dataset to the most similar cluster and recalculates the arithmetic mean of the clusters in the dataset.
- K-Means reassigns each record in the dataset to only one of the new clusters formed.
- The preceding steps are repeated until "stable clusters" are formed and the K-Means clustering is completed

7.2 SELF ORGANIZING MAP (SOM)

The Self-Organizing Map (SOM) was developed by Kohonen in 1997, on the basis of a single layered neural network. The data objects are presented at the input, and the output neurons are organized with a simple neighbourhood structure such as a two dimensional p*q grid.

7.3 HIERARCHICAL CLUSTERING

Hierarchical clustering generates a hierarchical series of nested clusters which can be graphically represented by a tree, called *dendrogram.*

The branches of a dendrogram not only record the formation of the clusters but also indicate the similarity between the clusters.

The hierarchical clustering scheme:

Let S=*{Si, Sj}* is the input similarity matrix, where *Si,j* indicates similarity between two data objects based on Euclidean distance.

Algorithm:

- ✓ Find a minimal entry s(i, j) in S, and merge clusters i and j.
- ✓ Modify S by deleting rows and columns i, j and adding a new row *i*and column *j*, with their dissimilarities defined by:
- ✓

$$s(k, i \cup j) = s(i \cup j, k) = \alpha_i s(k,i) + \alpha_j s(k,j) + \gamma |s(k,i) - s(k,j)|$$

✓If there is more than one cluster, then go to initial step.

Common variants of this scheme, obtained for appropriate choices of the parameters, are the following:

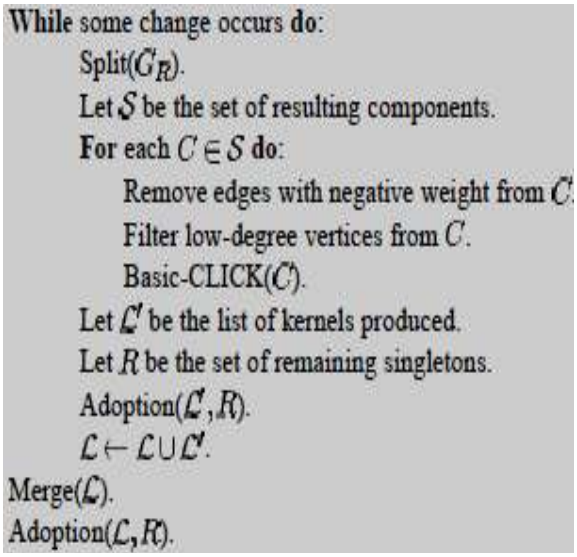
$$single linkage : s(k, i \cup j) = min = \{s(k,i), s(k,j)\}$$

$$complete linkage : s(k, i \cup j) = max\{s(k,i), s(k,j)\}$$

$$average linkage : s(k, i \cup j) = (n_i d(k,i) + n_j d(k,j))/(ni + nj),$$

Cluster Identification Via Connectivity Kernels (CLICK) ALGORITHMS

CLICK (Cluster Identification via Connectivity Kernels). This utilizes graph-theoretic and statistical techniques to identify tight groups of highly similar elements that are likely to belong to the same true cluster. Define the similarity between two sets as the average similarity between their elements. An *adoption step* repeatedly searches for a singleton *v* and a kernel *K* whose similarity is maximum among all pairs of singletons and kernels. If the value of this similarity exceeds some predefined [22] threshold, then *v* is added to *K* and removed from *R*.



- K and N are the number of clusters and genes in the data sets,
- m is a parameter which relate to `fuzziness' of resulting clusters,
- uki is the degree of membership of gene xi in cluster k,
- d2(xi; ck) is the distance from gene xi to centroid ck.

The parameters in this equation are the cluster centroid vector ck and the components of the membership vectors uki.

These unknown parameters can be optimized by La-grange method. Calculated uki shows the belonging ratio to a cluster k and centroid ck shows the representative gene expression profile of a cluster k.

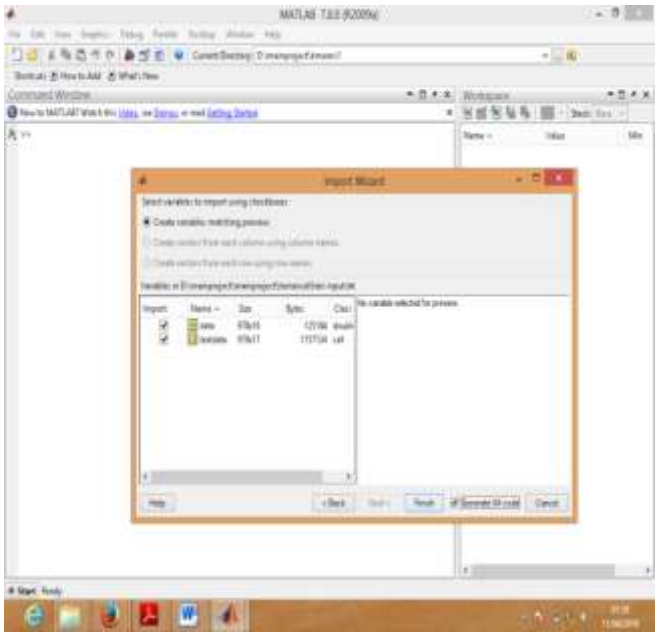
 ISSN: 2321-8169
228 – 235

In this study, a parameter m was set to 2.0 and the number of clusters was set to 5.

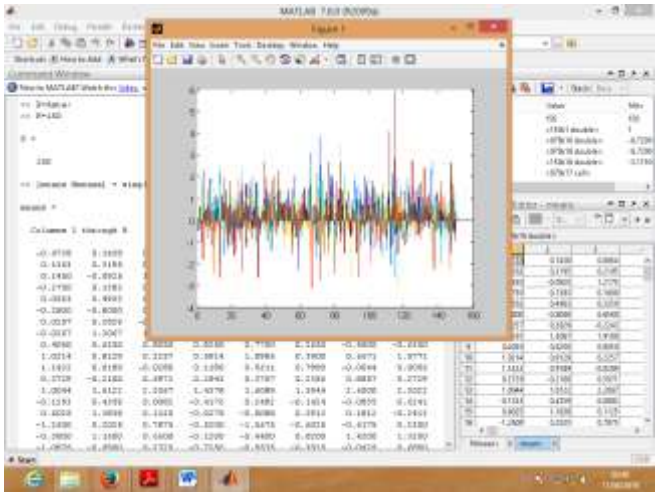
For the number of the clusters in the other clustering method, the selected are same number as that of clusters using fuzzy k-means clustering in order to comparing the clustering results.

VIII. RESULTS

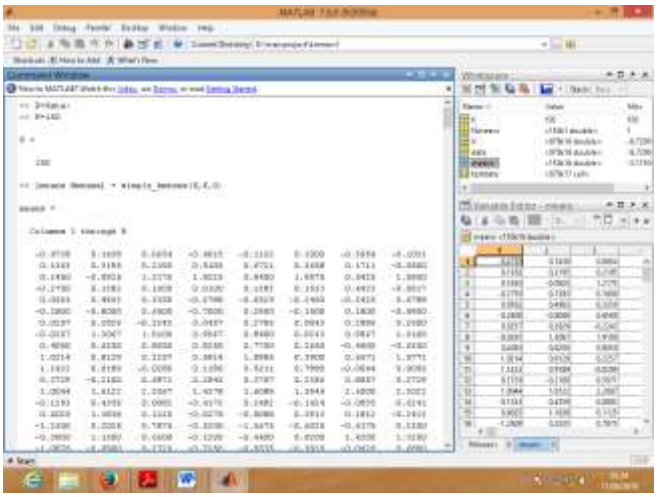
8.1 SCREEN SHORT FOR IMPORT GENE EXPRESSION DATA



8.2 SCREEN SHOT FOR FINALLY PLOTTING MEANS AND NMEANS

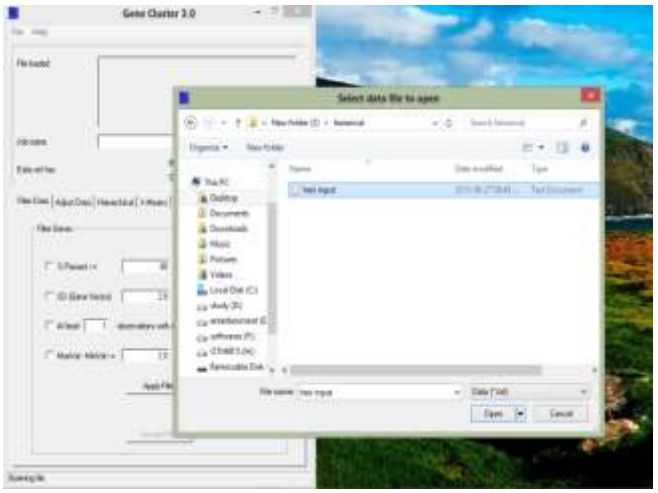


8.3 SCREEN SHOT FOR K-MEANS RESULT

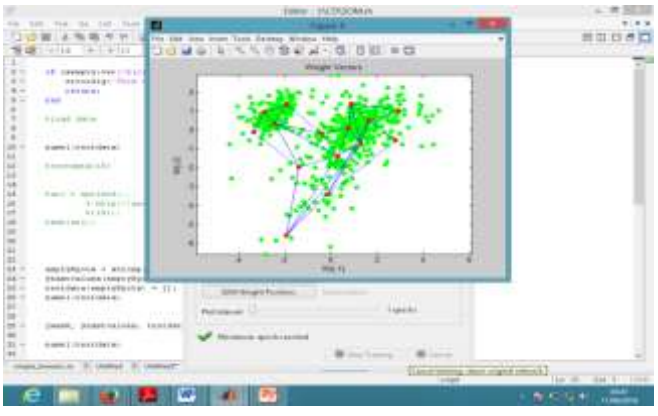


SELF ORGANIZING MAP RESULT

8.4 GIVING THE GENE EXPRESSION DATA TO THE CLUSTER 3.0

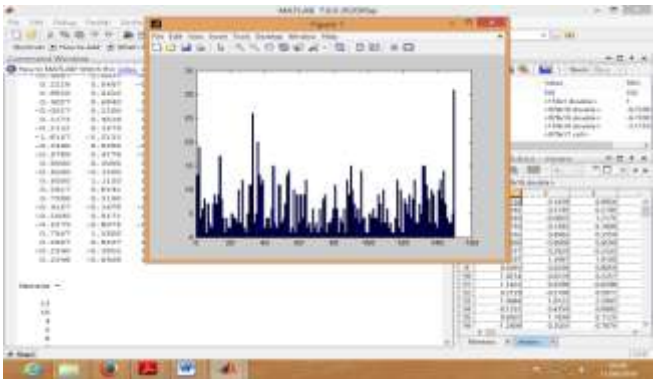


8.5 ANALYSIS ON GENE EXPRESSION DATA TO THE CLUSTER 3.0

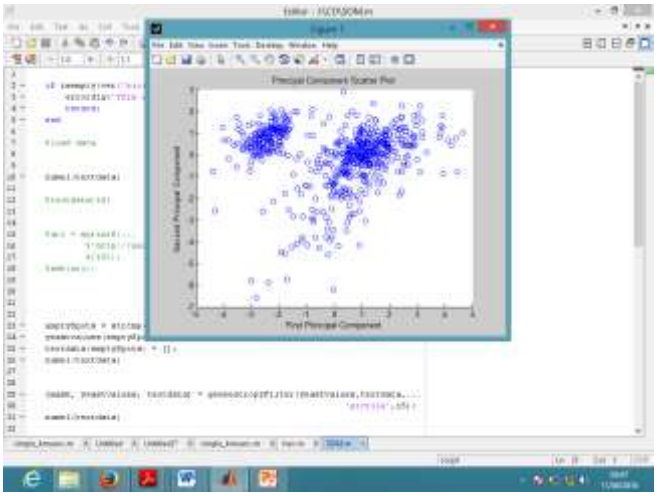


8.6 SELF ORGANIZING MAP RESULT

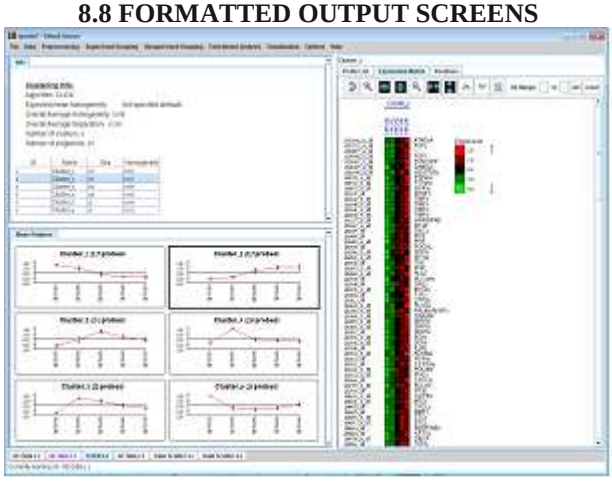
 ISSN: 2321-8169
228 – 235



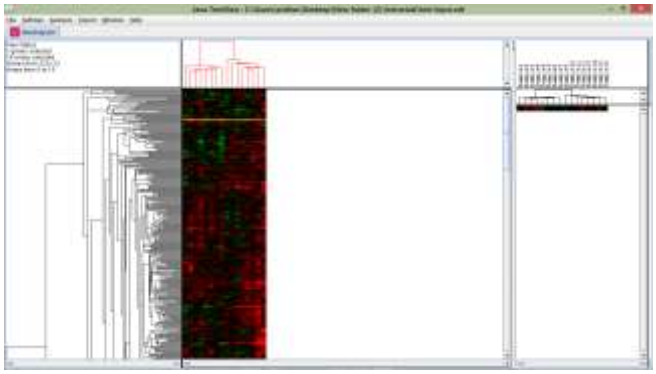
8.7 SCREEN SHOT FOR SELF ORGANIZING MAP RESULT



8.8 FORMATTED OUTPUT SCREENS



8.9 DIPICTION OF THE TREE



IX. CONCLUSION

In conclusion, cluster analysis requires experience and knowledge about the behaviour of clustering algorithms, and can benefit from a priori knowledge about the data and underlying biological processes. When a priori knowledge about the data is not available or insufficient, it may be desirable to try different algorithms to explore the data and get meaningful clustering results through comparisons. However for the considered data set CLICK gives it best with less complexity. Because of the agglorative approach of Hierarchal Clustering too gave its best. Fuzzy is the adaptive approach as it allows one gene to cluster among two or three clusters.

Given the variety of available clustering algorithms, one of the problems faced by biologists is the selection of the algorithm most appropriate to a given gene expression data set. A gene expression data set typically contains thousands of genes. However, biologists oftenhave different requirements on cluster granularity for different subsets of genes. For some purpose, biologists may be particularly interested in some specific subsets of genes and prefer small and tight clusters. While for other genes, people may only need a coarse overview of the data structure. However, most of the existing clustering algorithms only provide a crisp set of clusters and may not be flexible to different requirements for cluster granularity on a single data set.

Several existing approaches, such as hierarchical clustering, SOM, and some of the major CT algorithms can graphically represent the cluster structure. However, these algorithms may not be able to adapt to different userrequirements on cluster granularity for different subsets of the data. In the study, the functions of some genes have been studied in the literature, which can provide guidance to the clustering. Furthermore, some groups of the experimental conditions are known to be strongly correlated, and the *differences* among the cluster structures under these different groups may be of particular interest. If a clustering algorithm could integrate such partial knowledge as some clustering constraints when carrying out the clustering task, we can expect the clustering results would be more biologically meaningful. In this way, clustering could cease to be a “pure” unsupervised process and become an interactive exploration of the data set.

References

[1] **M Ramachandra & R Bramamrambha**, Implementation of classification rule discovery on biological datsets using Ant Colony Optimization, IJETM, 3(4): 1-12, 2016.

[2] **Ben-Dor A., Shamir R. and Yakhini Z.** Clustering gene expression patterns. Journal of ComputationalBiology, 6(3/4):281–297, 1999.

[3] **Arun K Pujari,** Data mining Techniques, University Press, Hyderabad, 2002.

[4] **Prasad, Vadamodula, T. Srinivasa Rao, and PVGD Prasad Reddy.** "Improvised prophecy using regularization method of machine learning algorithms on medical data." *Personalized Medicine Universe* 5 (2016): 32-40.

[5] **Daxin Jiang, Chun Tang and Aidong Zhang,** Cluster Analysis for Gene expression Data: A Survey", IEEE Transactions on Knowledge and Data Engineering, Vol. 16,No.11, November 2004, pp:1370-1386.

[6] **M Ramachandra & R Bramamrambha**, Comparing clustering techniques for gene expression data analysis, Research India Publications, 10(11): 10397-10399, 2015.

[7] **Vadamodula, Prasad, et al**. "Scrutiny of Data Sets Through Procedural Algorithms for Categorization." *Data Engineering and Intelligent Computing*. Springer, Singapore, 2018. 437-444.

[8] **Herrero J., Valencia A. and Dopazo J,** A hierarchical unsupervised growing neural network for clustering gene expression patterns", Bioinformatics, Vol:17, 2001, pp:126-136.

[9] **Sharan R, Elkon R, and Shamir R,** Cluster Analysis and its Applications to GeneExpression Data", Ernest Schering workshop on Bioinformatics and Genome Analysis,Springer Verlog, 2002.

[10] **Shamir R. and Sharan R. Click**: A clustering algorithm for gene expression analysis. In Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB '00). AAAIPress., 2000.

[11] **Prasad, V., T. Srinivasa Rao, and M. Purnachandrarao**. "Proportional analysis of non linear trained datasets on identified test datasets." *International Conference on recent trends and research issues in computer science & engineering*. Vol. 1. No. 1. 2015.

[12] **Prasad, V., T. Srinivasa Rao, and B. Sai Ram**. "Information clustering based upon rough sets." *Int. J. Sci. Eng. Technol. Res.(IJSETR)* 3 (2014): 8330-8333.

[13] **Prasad, V., R. Siva Kumar, and M. Mamtha**. "Plug in generator to produce variant outputs for unique data." *Int J Res Eng Sci* 2.4 (2014): 1-7.

[14] **Prasad, V., T. Srinivas Rao, and M. Surendra Prasad Babu.** "Offline analysis & optimistic approach on livestock expert advisory system." *Artificial Intelligent Systems and Machine Learning* 5.12 (2013): 488.

[15] **Prasad, V., and T. Srinivasa Rao**. "Permissible thyroid datasets assessment through kernel PC Algorithm and Vapnik Chervonenkis theory for categorization and classification." Data-Enabled Discovery and Applications, 1.1 (2017): 1-11

[16] **Eisen, Michael B., Spellman, Paul T., Brown, Patrick O. and Botstein, David .** Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA, 95(25):14863–14868, December1998.

[17] **Prasad,Vadamodula.TamadaSrinivasaRao**.Implement ation of Regularization Method Ridge Regression on Specific Medical Datasets."." *International Journal of Research in Computer Applications & Information Technology* 3 (2015): 25-33.

[18] **Kohonen T.** Self-Organization and Associative Memory.Spring-Verlag, Berlin, 1984.

[19] **Prasad, V., et al.** "Comparative study of medical datasets IETD and UCITD using statistical methods." (2015).

[20] **Prasad, V., T. Srinivasa Rao, and M. Surendra Prasad Babu**. "Thyroid disease diagnosis via hybrid architecture composing rough data sets theory and machine learning algorithms." *Soft Computing* 20.3 (2016): 1179-1189.

[21] **M Ramachandra & R Bramamrambha**, Particle swarm optimization for initial centroid selection in portioned clustering, IJAREEIE, 4(2): 175-184, 2015.

[22] **Prasad, Vadamodula, Thamada Srinivasa Rao, and Ankit Kumar Surana.** "Standard cog exploration on medicinal data." *International Journal of Computer Applications* 119.10 (2015).