INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

A Comparative Study on Hierarchical, K-Means and Fuzzy C-Means Clustering Algorithms and Application to Microarray Gene Expression Data

Anukampa Behera[1], Sujata Chakravarty[2]

[1]Student of M.Tech CSE, [2]HOD Deptt of CSE
Orissa Engineering College
Bhubanneswar, Odisha
[1]anukampa@citzen.net, [2]chakravartys69@gmail.com

Abstract: *The advent of DNA microarray technology has enabled biologists to monitor the expression levels (mRNA) of thousands of genes simultaneously. In this paper, three approaches to gene expression data analysis using clustering techniques have been addressed and the experiments has been done using six datasets, IRIS data, WBCD data, Iyer Serum data, Cho Yeast data, Leukaemia Golub experiment data and St. Jude Leukaemia data. The performance of various existing clustering algorithms under each of these approaches is discussed. Finally, since evaluation of the effectiveness of the clustering techniques over gene data requires validity measures and data sources for numeric data, for which the Silhouette Index is used. This paper presents a comparative analysis between Hard clustering and Soft C-Means clustering and all experimentation it is found that in majority cases soft C-Means performs better that Hard clustering algorithms.*

Keywords: *Microarray technology, gene expression data, clustering, proximity measure, clustering, cluster validation*

1. INTRODUCTION

DNA microarray technology has now made it possible to simultaneously monitor the expression levels of thousands of genes during important biological processes and across collections of related samples. [1] Elucidating the patterns hidden in gene expression data offers a tremendous opportunity for an enhanced understanding of functional genomics. However, the large number of genes and the complexity of biological networks greatly increase the challenges of comprehending and interpreting the resulting mass of data, which often consists of millions of measurements. A first step toward addressing this challenge is the use of clustering techniques, which is essential in the data mining process to reveal natural structures and identify interesting patterns in the underlying data.

Cluster analysis seeks to partition a given data set into groups based on specified features so that the data points within a group are more similar to each other than the points in different groups. A very rich literature on cluster analysis has developed over the past three decades. 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 algorithms have been proven useful for identifying biologically relevant groups of genes and samples. (1) Rest of the paper is organized as follows. In section-2 the DNA microarray technology is discussed. In section-3, a brief discussion is done on clustering gene expression data. In section-4, various clustering algorithms such as hierarchical, K-Means and Fuzzy C-Means clustering techniques are discussed. In Section-5, the whole experimental setup is discussed and results are discussed for implementation of each above said algorithms. And the paper is concluded in section-6.

2. DNA MICROARRAY TECHNOLOGY

The traditional approach to genomic research has focused on the local examination and collection of data on single genes, but using the microarray technologies now it has become possible to monitor the expression levels for tens of thousands of genes in parallel. There are two major types of microarray experiments: the cDNA microarray and oligonucleotide arrays (abbreviated oligo chip) [19]. Despite differences in the details of their experiment protocols, both types of experiments involve three common basic procedures.

A microarray experiment typically assesses a large number of DNA sequences (genes, cDNA clones, or expressed sequence tags) under multiple conditions.[4] These conditions may be a time series during a biological process or a collection of different tissue samples (e.g., normal versus cancerous tissues). In this thesis work, emphasis is given on the cluster analysis of gene expression data without making a distinction among DNA sequences, which will uniformly be called ”genes”. Similarly, it is 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 expression matrix $M = \{W_{ij} \mid 1 \le i \le n, 1 \le j \le m\}$ as shown in Figure 1.2, where the rows ($G = \{g_1 \ldots g_n\}$) form the expression patterns of genes, the columns ($S = \{S_1 \ldots S_m\}$) represent the expression profiles of samples, and each cell is the measured expression level of gene i in sample j. [2]

3. CLUSTERING GENE EXPRESSION

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

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

objects in separate classes are more dissimilar [1]. Clustering is an example of unsupervised classification. Classification refers to the method of allocating the data objects to a set of classes and Unsupervised means that this grouping (clustering) is not based on predefined training examples or existing classes while classifying the data objects.

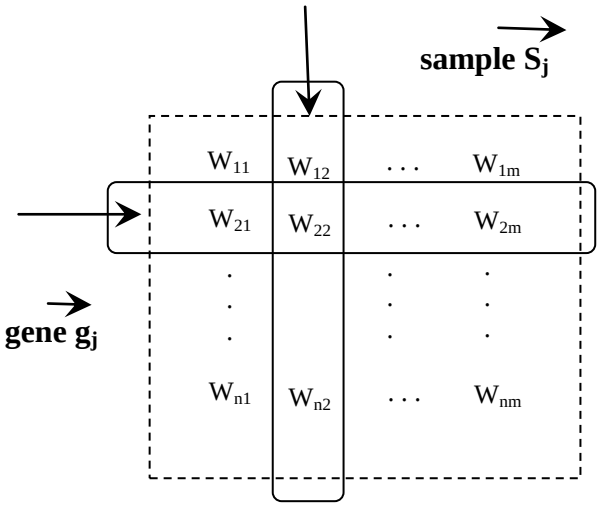

Fig 1: A gene-expression matrix

The easiest way to find genes of potential interest through several related experiments is to search for those that are consistently either up- or down regulated. To that end, a simple statistical analysis of gene-expression levels will suffice. [3] However, identifying patterns of gene expression and grouping genes into expression classes might provide much greater insight into their biological function and relevance [4]. For the analysis of gene expression data, many techniques have been used which includes statistical methods, generally referred to as cluster analysis. The years of research work in sequence analysis and related areas have shown the benefits and effectiveness of probabilistic approaches to biological data. The current DNA array data is inherently very noisy, because of experimental and biological variables that are difficult to control. However, because of the influence of experimental and biological errors inherent in the high dimensional data of DNA microarray experiments, the analysis of these data this is not a simple task.[3]

Clustering methods, which determine the natural subgroups in a data set, have some advantages over other methods, because no previous knowledge is necessary for clustering analysis [4], [5]. Clusters may be exhaustive, meaning that each object is assigned to a cluster, or non-exhaustive, meaning that some objects may be assigned to no cluster. Exclusive clusters are non-exhaustive ones to which an object is either assigned or not [2]. Objects are assigned solely to one cluster in hard clustering; whereas soft clusters, sometimes called overlapping clusters, may have common objects with non-negative value memberships [15]. Several clustering algorithms have been proposed in past few decades [8] – [11], [13] – [14].

In hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster i.e., a hard clustering algorithm allocates each pattern (or, data point) to a single cluster during its operation and in its output whereas a soft clustering method assigns degrees of membership in several clusters to each input pattern. A soft clustering can be converted to a hard clustering by assigning each pattern to the cluster with the largest measure of membership [25]. In **fuzzy clustering** (also referred to as **soft clustering**), data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters. In this chapter, study has been carried out on Hierarchical, K-means clustering (hard) and Fuzzy C-means (Soft) clustering algorithm.

4. CLUSTERING ALGORITHMS

In this section, we briefly describe three such methods, including the classic hierarchical clustering methods, k-Means and, Fuzzy C-Means.

4.1 Hierarchical Clustering

Hierarchical clustering works by iteratively joining the two closest clusters starting from singleton clusters or iteratively partitioning clusters starting with the complete set. After each joining of two clusters, the distances between all the other clusters and a new joined cluster are recalculated. The complete linkage, average linkage, and single linkage methods use maximum, average, and minimum distances between the members of two clusters respectively. [16] Since hierarchical clustering is a greedy search algorithm based on a local search, the merging decision made early in the agglomerative process are not necessarily the right ones. One possible solution to this problem is to refine a clustering produced by the agglomerative hierarchical algorithm to potentially correct the mistakes made early in the agglomerative process. Hierarchical methods are commonly used for clustering in Data Mining. A hierarchical clustering scheme produces a sequence of clustering in which each clustering is nested into the next clustering in the sequence [17].

Hierarchical clustering algorithms can be further divided into agglomerative approaches and divisive approaches based on how the hierarchical dendrogram is formed. Agglomerative algorithms (bottom-up approach) initially regard each data object as an individual cluster, and at each step, merge the closest pair of clusters until all the groups are merged into one cluster. Divisive algorithms (top-down approach) starts with one cluster containing all the data objects and, at each step split, only singleton clusters of individual

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

objects remain. For agglomerative approaches, different measures of cluster proximity, such as single link, complete link, and minimum-variance, derive various merge strategies. For divisive approaches, the essential problem is to decide how to split clusters at each step. Some are based on heuristic methods such as the deterministic annealing algorithm, while many others are based on the graph theoretical methods.[6-7]

4.2 K-Means clustering

The K-means algorithm [23] is a typical partition-based clustering method. Given a prespecified number K, the algorithm partitions the data set into K disjoint subsets which optimize the following objective function:

$$E = \sum_{i=1}^{K} \sum_{O \in C_i} |O - \mu_i|^2 \quad (1)$$

Here, O is a data object in a cluster Ci and μ_i is the centroid (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.

The K-means algorithm is simple and fast. The time complexity of K-means is O(l*k*n), where l is the number of iterations and k is the number of clusters. Our empirical study has shown that the K-means algorithm typically converges in a small number of iterations. However, it also has several drawbacks as a gene-based clustering algorithm. First, the number of gene clusters in a gene expression data set is usually unknown in advance. To detect the optimal number of clusters, users usually run the algorithms repeatedly with different values of k and compare the clustering results. For a large gene expression data set which contains thousands of genes, this extensive parameter fine-tuning process may not be practical. Second, gene expression data typically contain a huge amount of noise; however, the K-means algorithm forces each gene into a cluster, which may cause the algorithm to be sensitive to noise [26].

4.3 Fuzzy C-Means

Fuzzy C-Means algorithm (FCM) [20]-[24], generalizes the Hard C-means algorithm, to allow data points to partially belonging to multiple clusters at same time. Therefore, it produces a soft partition for a given data set. In fact, it generates a constrained soft partition. To achieve this, the objective function of HCM clustering algorithm has to be extended in two ways:

1. Incorporation of degree of fuzzy membership in clusters $\{V_1,V_2,...,V_C\}$ and
2. An introduction of fuzziness parameter 'm', a weight exponent in the fuzzy membership.

The extended objective function J (MSE, Mean square error) is defined as follows:

$$Minimize(J) = \sum_{j=1}^{C} \sum_{x_i \in v_j} \left(\mu V_j(x_i)\right)^m \|x_j - v_j\|^2$$

$$for\ j \in 1,2,\dots,C\ and\ i = \{1,2\dots,n\} \quad (2)$$

Where V is a fuzzy partition of the data set X formed by clusters $\{V_1,V_2,...,V_C\}$. The parameter 'm' is a weight that determines the degree to which partial members of a cluster affect the clustering result.

Like Hard clustering algorithm, the Fuzzy C Means clustering algorithm tries to find a good partition by searching prototypes (i.e., cluster centre) vj that minimize the objective function 'J'. Unlike hard clustering, however, the Fuzzy C Means algorithm also needs to search for membership functions µVj that minimizes 'J'. To accomplish these two objectives, Fuzzy C Means Theorem has been proposed by Bezdek in 1981 [22-23].

5. EXPERIMENTAL SETUP

In this section, six datasets are used for simulating cluster formation algorithm to solve "gene expression analysis problem". Two pattern recognition data and four bioinformatics data were taken for simulation study. All the datasets on which study has been made on the benchmark bioinformatics data reported in last decade in literature. In order to identify common subtypes (cluster within clusters) in independent disease data: two different types of breast data (Golub et. al) and Leukaemia data are considered for our study on both gene/sample data as well as time series microarray data. The short description of the datasets along with their size, no. of clusters and source of the data has been given in Table 1.

Sr. no	Datasets	Dimension	No. Of clusters
1	Iris	[150x4]	3
2	WBCD	[683x9]	2
3	Iyer data/Serum data	[517x12]	11
4	Cho data (yeast data)	[386x16]	5
5	Leukaemia (Golub experiment)	[72x7129]	2
6	St. Jude Leukaemia data	[248x985]	6

Table 1: Datasets used in the experiments

These datasets are having both overlapping and non-overlapping class boundaries, where the number of features/genes ranges from 4 to 7129 and number of sample ranges from 32 to 683. The number of cluster ranges from 2 to 11. All the data have been pre-processed in such a way that all the data point which belongs to class 1 in original datasets is kept together,

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

then all the data points which belong to class 2 is kept together and so on.

5.1 Cluster Validation Metrics

Generally, cluster validity has three aspects. First, the quality of clusters can be measured in terms of homogeneity and separation on the basis of the definition of a cluster: "objects within one cluster are similar to each other, while objects in different clusters are dissimilar with each other". For this, HS Ratio has been used as cluster validation metric. The second aspect relies on a given 'ground truth' of the clusters. The 'ground truth' or 'class−labelled information' could come from domain knowledge, such as known function families of genes, or from other sources such as the clinical diagnosis of normal or cancerous tissues. Cluster validation is based on the agreement between clustering results and the "ground truth. For this, Cluster Accuracy was used as cluster validation metric. The third aspect of cluster validity focuses on the reliability of the clusters, or the likelihood that the cluster structure is not formed by chance. For this Silhouette index was used.

Silhouette index [24] is used to assess the quality of any clustering solution. This index reflects the compactness and separation of clusters. It is calculated as follows.

1. Compute $a(g_i)$, i.e., the average distance of gene i to the other genes of cluster A to which it belongs.
2. Compute $d(g_i,C_k)$ where $d(g_i,C_k)$ is the average distance of gene g_i from the genes of cluster C_k where $g_i \notin C_k$.
3. Compute $b(g_i)$, where $b(g_i) = \min\{d(g_i,C)\}$ where $C = \{C_1, C_2,..., C_m\}$ and $A \notin C$, i.e., $b(g_i)$ represents the distance of gene g_i to its closest cluster. Now compute the silhouette width of gene g_i as

$$S(g_i) = \frac{b(g_i)-a(g_i)}{max\{a(g_i),b(g_i)\}} \qquad (3)$$

4. Compute silhouette index by finding the average of S(i) over i = 1,2,...,G, where G is the total number of genes: S = average{S(gi)}.

The value of silhouette index varies from -1 to 1 with higher values indicating better clustering.[24]

We have used silhouette index to validate the cluster accuracy. Silhouette index [24] is used to assess the quality of any clustering solution. This index reflects the compactness and separation of clusters.

Table 2 represents summary of result obtained from Hierarchical and K-means clustering algorithm for six datasets. It also consists of some important relevant characteristics, such as number of classes, number of features/genes and the number of item samples. Out of six datasets maximum accuracy was reported for WBCD data while least accuracy data was reported for Iyer Serum Data.

Slno	Dataset	Dimension	# of clusters	Accuracy	
				Hierarchical	K-Means
1	Iris	[150X 4]	3	47.6821	86.7
2	WBCD	[683X 9]	2	79.0861	90.5
3	Iyer data/ Serum Data	[517X12]	11	53.1915	46.4
4	Cho data/ Yeast Data	[386X 16]	5	62.4352	61.4
5	Leukaemia (Golub experiment)	[72X 7129]	2	72.2222	77.8
6	St. Jude Leukemia data	[248X 985]	6	83.8912	93.2

Table 2: Summary of Results for Hard clustering using all six datasets

Figure – 2 and Figure – 3 represents the clustering accuracy of Hierarchical clustering and K-Means clustering algorithm using all six datasets.

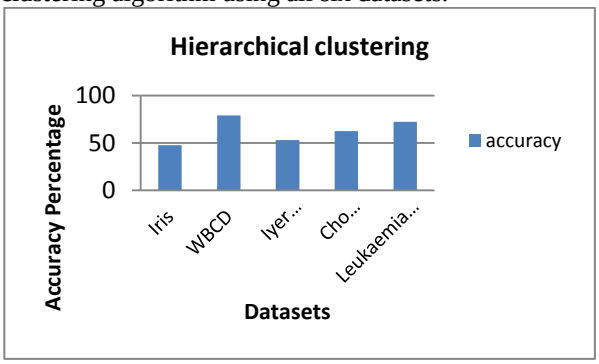

Fig – 2: clustering accuracy of Hierarchical clustering algorithm using all six datasets

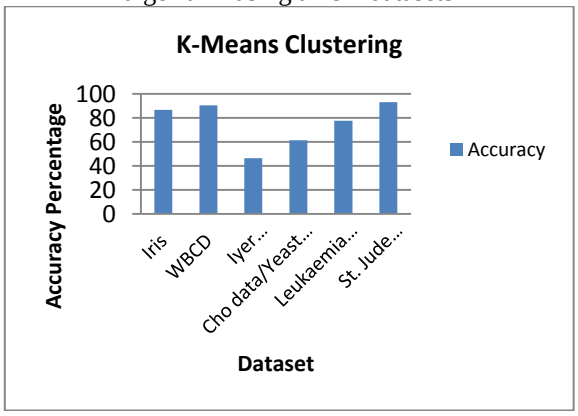

Fig – 3: clustering accuracy of K-Means clustering algorithm using all six datasets

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

Table – 3 represents summary of result of Fuzzy C-Means clustering algorithm result for six datasets. It consists some of the relevant characteristics, such as number of classes, number of features/genes and the number of item samples. These datasets are having both overlapping and non-overlapping class boundaries, where the number of features/genes ranges from 4 to 7129 and number of sample ranges from 32 to 683. The number of cluster ranges from 2 to 11.

Slno	Dataset	Dimension	# of clusters	Accuracy Fuzzy C-Means
1	Iris	[150X4]	3	90.5765
2	WBCD	[683X9]	2	92.0914
3	Iyer data/Serum Data	[517X12]	11	59.7679
4	Cho data/Yeast Data	[386X16]	5	51.5287
5	Leukaemia (Golub experiment	[72X7129]	2	77.7778
6	St. Jude Leukemia data	[248x985]	6	92.5623

Table 3: Summary of results of Fuzzy C-Means clustering applied on all six datasets

Figure – 4 represents the clustering accuracy of Fuzzy C-Means clustering algorithm using all six datasets

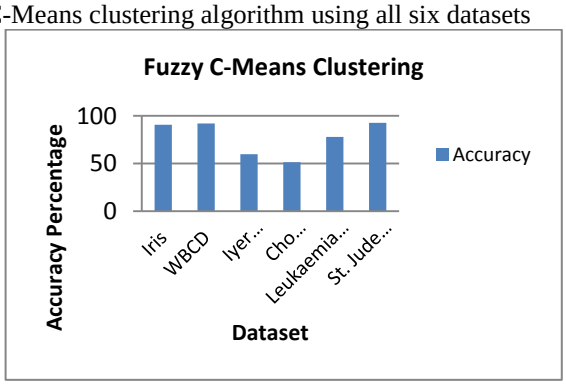

Fig – 4: clustering accuracy of Fuzzy C-Means clustering algorithm using all six datasets

5.2 A Comparative study

This section deals with comparative studies on hard clustering and Soft C-means Clustering algorithm. All the six datasets described in the previous section has been considered to compare performances of hard clustering and Soft C-means clustering algorithm. A comparative representation is presented in table – 4.

Dataset	# of clusters	Accuracy			Error		Error %		Increment
		K-Means	Hierarchical	Fuzzy C-Means	K-Means	Fuzzy C-Means	K-Means	Fuzzy C-Means	
Iris	3	86.8	47.7	90.6	17	14	11.3	9.33	2
WBCD	2	90.5	79.1	92.1	29	29	4.25	4.21	0.04
Iyer data	11	46.4	53.2	59.8	246	265	47.6	45.5	2.13
Cho data	5	61.4	62.4	51.5	151	168	39.1	43.5	-4.4
(Golub)	2	77.8	72.2	77.8	29	15	40.3	20.8	19.5
St. Jude	6	93.2	83.9	92.6	55	14	27.9	7.11	20.8

Table 4: Comparative study of Hard Clustering and Fuzzy C-Means Clustering using all ten datasets

Fig – 5 shows a comparative representation of the clustering accuracy of all the three types clustering algorithm using all six datasets

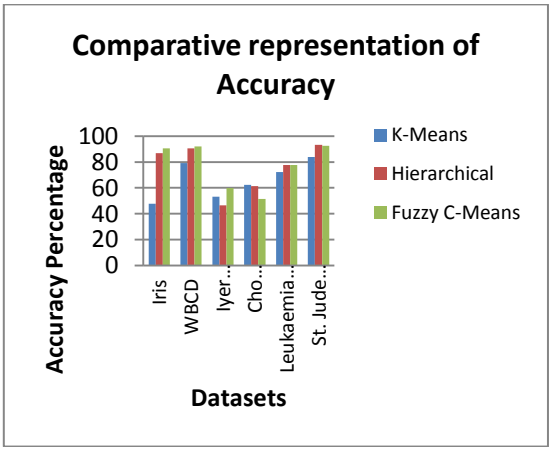

Fig – 5: A comparison chart on all three clustering

6. CONCLUSION

In this paper we have compared the hierarchical; K-means and fuzzy C-Means based clustering algorithms based on six datasets. The Fuzzy C-Means based clustering algorithm is a distinct improvement from the conventional hard clustering algorithm. Its ability to cluster independent of the data sequence provides a more stable clustering result. Computer simulation shows that Fuzzy C-means performs superior compared

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

to Hierarchical and K-Means algorithm in four out of six cases. Datasets belonging to these cases are Iris, WBCD, Iyer serum data, Leukaemia Golub experiment data. Whereas hard clustering shows superior performance in two cases. Datasets belonging to these cases are Yeast (Cho) data, St. Jude leukaemia data. As seen from the experiments, the Fuzzy C Means based clustering algorithm was able to provide the highest accuracy and generalization capabilities.

References

[1] Daxin Jiang Chun Tang Aidong Zhang "Cluster Analysis for Gene Expression Data – A Survey," Knowledge and Data Engineering, IEEE Transactions on (Volume:16 , Issue: 11

[2] Sajid Nagi, Dhruba K. Bhattacharyya, Jugal K. Kalita "Gene Expression Data Clustering Analysis: A Survey," Emerging Trends and Applications in Computer Science (NCETACS), 2011 2nd National Conference

[3] Ortiz-Gama, S. ; Tec de Monterrey, Morelos, Mexico ; Sucar, L.E. ; Rodriguez, A.F. "Clustering gene expression data: an experimental analysis", Computer Science, 2004. ENC 2004. Proceedings of the Fifth Mexican International Conference on 20-24 Sept. 2004, Pages :168 – 175, Print ISBN:0-7695-2160-6,IEEE

[4] Quackenbush, John. "Computational Analysis of Microarray Data." - Nature Reviews Genetics. Vol. 2, June 2001. P. 418-427.

[5] Yin, L. ; Dept. of Comput. Sci. & Eng., Connecticut Univ., Storrs, CT ; Chun-Hsi Huang, "Clustering of Gene Expression Data: Performance and Similarity Analysis" - Computer and Computational Sciences, 2006. IMSCCS '06. First International Multi-Symposiums on (Volume:1) 20-24 June 2006Print ISBN:0-7695-2581-4, IEEE

[6] Everitt, B., "Cluster analysis", Halstead, New York. 1980

[7] Hartigan, J., "Clustering algorithms", Wiley, New York.1973

[8] Ping Guo ; Sch. of Comput. Sci., Chongqing Univ., Chongqing, China ; Xiao-yan Deng, "Gene Expression Data Cluster Analysis", Information Engineering, 2009. ICIE '09. WASE International Conference on (Volume:1), 10-11 July 2009, Page(s):99 – 102, Print ISBN: 978-0-7695-3679-8, IEEE

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

[10] Joaquín Dopazo and José María Carazo, "Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree." J. Mol. Evol., 44: 226-233. 1997 Kohonen, T. "Self-Organizing Maps", Springer, Berlin. 1995.

[11] Sneath and Sokal. "Hierarchical Clustering", 1973

[12] Tamayo, P. Dmitrovsky, E., "Interpreting patterns of gene expression with self-organizing maps: Methods and applications to hematopoietic differentiation", Proc. Nat. Acad. Sci 96, 2907-2912, 1999

[13] Herrero, J. & Dopazo, J." Combining hierarchical clustering and self-organizing maps for exploratory analysis of gene expression patterns". Journal of Proteome Research, 1(5):467-470. 2002

[14] Harun Pirim, Burak Eks-io˘glu, Andy D. Perkins, Cetin Yuceer "Clustering of high throughput gene expression data", Computers & operations research 12/2012;39(12):3046-3061.DOI: 10.1016/j.cor.2012.03.008

[15] Alvis Brazma, Jaak Vilo "Minireview Gene expression data analysis", edited by Gianni Cesareni, FEBS 23893

[16] Manpreet kaur, Usvir Kaur "Comparison Between K-Mean and Hierarchical Algorithm Using Query Redirection" - International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, July 2013

[17] E.J. Moler, M.L. Chow, and I.S. Mian, "Analysis of Molecular Profile Data Using Generative and Discriminative Methods." Physiological Genomics, vol. 4, no. 2, pp. 109-126, 2000.

[18] D. Lockhart et al., "Expression Monitoring by Hybridization to High-Density Oligonucleotide Arrays," Nature Biotechnology, vol. 14, pp. 1675-1680, 1996.

[19] F.D. Smet, J. Mathys, K. Marchal, G. Thijs, M. Moor, D. Bart, and Y. Moreau, "Adaptive Quality-Based Clustering of Gene Expression Profiles," Bioinformatics, vol. 18, pp. 735-746, 2002.

[20] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," Journal of Cybernetics, vol. 3, pp. 32–57, 1973.

[21] J. C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algoritms." New York: Plenum Press, 1981.

[22] J. C. Bezdek, "A convergence theorem for the fuzzy isodata clustering algorithms," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 2, no. 1, pp. 1–8, 1980.

[23] P. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis". Journal of Computational and Applied Mathematics, vol. 20, pp. 153–165, 1987

[24] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A Review", ser. 3. ACM Computing Surveys, September 1999, vol. 31.

[25] K. Krishna and M. Narasimha Murty, "Genetic K-Means Algorithm", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 29, NO. 3, JUNE 1.