Quad-Tree Based Multiple Kernel Fuzzy C-Means Clustering for Gene Expression Data

E. Monica Sushil Cynthia [#1], S. Kannan[#2]

[1]Department of MCA, the American College, Madurai, Tamil Nadu, India

[2]Department of MCA, Madurai Kamaraj University, Madurai, Tamil Nadu, India

Abstract *- Minute variations in genes can have a major impact on how humans respond to disease, environmental factors such as bacteria, viruses, toxins, chemicals and drugs and other therapies.. 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. The clustering algorithms have been proven useful for identifying biologically relevant groups of genes and samples. Hence in this paper we propose a new clustering algorithm for gene expression data associated to three different types of cancer and also compare with the existing approaches to prove the novel approach proposed here, has a better performance, reliability and provide more meaningful biological significance.*

Keywords - *Clustering, Clustering Algorithms, Gene Expression analysis, Fuzzy C-Means, Hierarchical Clustering, Gene Clustering, Gene Expression data, Quad Tree, Kernel fuzzy C-means.*

I. INTRODUCTION

A person's genetic background is considered in every aspect of clinical medicine, ranging from susceptibility to diseases, pathogenesis, and clinical outcome to diversity in responses to drug treatment (pharmacogenomics). The new panoramic look at the human genome has stimulated a massive search for clinically relevant genomic information, including SNP , resulting out of substitution or deletion of one nucleotide in a DNA sequence. Individual gene expression is 99.9 percent identical, with only 0.1 percent of the gene expression data showing polymorphisms. These mutated genes are responsible for susceptibility to a particular disease. Elucidating the patterns hidden in the data is a tremendous challenge and opportunity for scientist. A first step in answering this challenge is via clustering techniques, which aim to identify sets of gene that behave similarly across the condition. In cluster analysis, we wish to partition entities into group based on given features of each entity, so that the groups are homogeneous and well separated. Each group is called *cluster*, and the partition is called *clustering*. Clustering problems arise in numerous disciplines including biology, medicine, psychology, and economics. There is a very rich literature on cluster analysis going back over three decades [1-4]. Numerous approaches were proposed to define quality criteria for solutions, stipulating the type of clustering sought, and interpreting the solution. The purpose of gene-based clustering is to group together co-expressed genes which indicate co-function and co-regulation [5]. We shall describe in this paper three of the most representative off-line clustering techniques: fuzzy C-means (FCM) clustering, mixed C-means clustering and quad tree based multiple kernel fuzzy C-means (QKFCM). The techniques are implemented and tested against a gene expression dataset for cancer. The results are presented with the comparison of different techniques and the validity is measured to find out the effectiveness of the proposed clustering approach.

II. RESEARCH MOTIVATIONS

Single nucleotide polymorphisms (SNPs) in genes predisposes humans to different diseases This makes the SNPs valuable for biomedical research and for developing pharmaceutical products or medical diagnostics. Hence there is a strong need to mine the informative genes responsible for cancer which is the leading cause of mortality in the world.

III. METHODOLOGY

The proposed algorithm QKFCM (Quad Tree based multiple kernel fuzzy C-means) is implemented against three different types of cancer dataset, and the flow is described in figure 1. A comparative analysis of QKFCM is made with the other most representative off-line clustering techniques namely Fuzzy C-means and Mixed C-means based on quality of clusters [6].

3.1 Fuzzy C-means

FCM is a method of clustering which allows one piece of data to belong to two or more clusters **(Toushmalani, 2011)**. This method [7-8] is frequently used in pattern recognition [9]. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N} \sum^{C} u^{m}_{ij} \| x_i - c_j \|^2$$

Once again, the expression X={x1,x2,…,xn} is a collection of data, where n is the number of data points and C={c1,c2,…,cc} is the set of corresponding cluster centers in the dataset X, where c is the number of clusters, μ_{ij} is the membership degree of the data x_i to the cluster center Cj.

3.2 Mixed C-Means clustering

The model proposed by Pal et al.(1997) called fuzzy-possibilistic C-means (FPCM) establishes a connections between the possibilistic and probabilistic approaches [5]. The FPCM creates both membership (in the sense of relative belonging) and typicalities, along with the usual centroids through a standard alternating optimization process.

3.3 Quad Tree based Multiple Kernel Fuzzy C-means (QKFCM)

- A Quad Tree in two dimensional spaces is a 4-way branching tree that represents recursive decomposition of space using separators parallel to the coordinate axis.
- At each level a square subspace is divided into four equal size squares. This data structure was named a Quad Tree by Finkel and Bentley in 1974.
- The output of the Quad Tree algorithm presented is the set of centers. We use such centers as the initial cluster centers for the Multiple Kernel Fuzzy C-means.

Algorithm: Multiple kernel fuzzy c-means (MKFC)

Give a set of N data points $X=\{x_i\}^{N}_{i=1}$, a set of kernel functions $\{k_k\}^{M}_{k=1}$, and the desired number of cluster C, output a membership matrix $U=\{u_{ic}\}^{N,C}_{I,c=1}$ and weights $\{w_k\}^{M}_{k=1}$ for the kernels [11].

```
procedure MKFC(Data X, Number C, Kernels $\{k_k\}^{M}_{k=1}$
Initialize membership matrix $U^{(0)}$
Repeat
    $\hat{u}^{(t)}_{ic} = \frac{u^{(t)m}}{\sum^{N}_{i=1} u^{(t)}_{ic}{}^{m}}$   - calculate normalized memberships
    For (i=1..N; c=1..C; k=1..M) do
        $\alpha_{ick} \leftarrow k_k(x_i, x_i) - 2\sum_{j=1}^{N} \hat{u}^{(t)}_{ic} k_k(x_i,x_i) + \sum_{j=1}^{N} \sum_{j=1}^{N} \hat{u}_{jc}^{(t)} \hat{u}_{jc}^{(t)\,k}{}_{k} (x_j,x_{j'})$
    End for
                                                  -Calculates Coefficients
    For (k=1..M)  do
        $\beta_k \leftarrow \sum_{i=1}^{N} \sum_{c=}^{C} (u_{ic}^{(t)})^m \alpha_{ick}$
    End for
                                                  -update weights
    For (k=1..M) do
        $w^{(t)}_{k} \leftarrow \frac{\frac{1}{\beta_k}}{\frac{1}{\beta 1} + \frac{1}{\beta 2} + \ldots + \frac{1}{\beta M}}$
    End for
                                                  -calculate distance
    For (i=1..N;c=1..C) do
        $D^2_{ic} \leftarrow \sum_{k=1}^{M} \alpha_{ick} (w_k^{(t)})^2$
    End for
                                                  -Update membership
    For(i=1..N;c=1..C) do
        $u^{(t)}_{ic} \leftarrow \frac{1}{\sum^{C}_{c'=1} \left( \frac{D^2_{ic}}{D^2_{ic}} \right)^{\frac{1}{m-1}}}$
    End for
Until $\| U^{(t)} - U^{(t-1)} \| < \in$
Return $U^{(t)}$, $\{ w_k^{(t)} \}_{k=1}^{M}$
End procedure
```

To determine optimal clustering various quality indices like Ball-Hall index, C index, Xie-Beni index, etc have been proposed. We have selected the most representative Xie-Beni index and Davies-Bouldin index. Xie-Beni index has been shown to detect the correct number of clusters in several experiments [15]. The homogeneity and separability of the clusters are measured using [5, 13, 14]

The methodology adopted in this work is given in Figure 1.

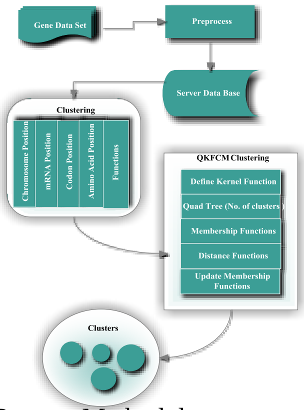

Figure 1: Process Methodology

IV. STRUCTURE OF CANCER DATASET

Cluster analysis, is an important tool in gene expression data analysis. For experimentation, we used a set of gene expression data for cancer. In Clustering gene expression data, the genes were treated as objects and the samples are treated as attributes. To evaluate the performance of the various proposed algorithms FCM, mixed C-means and QKFCM, these approaches were implemented in java. The cancer gene expression data was used for our experiments. This data is publicly available at the National Center for Biotechnology information (www.ncbi.nlm.nih.gov/SNP/). The dataset consist of three types of cancer namely leukemia, lung cancer and breast cancer.

TABLE I

COMPARATIVE ANALYSIS FOR DB

	DB Index		
Dataset	**FUZZY C-Means**	**Mixed C-Means**	**QKFCM**
Leukemia	6.4	4.9	1.8
Lung Cancer	6.9	4.6	1.956
Breast Cancer	5.8	5.3	2.1

TABLE II

Comparative analysis for XB

	XB Index		
Dataset	**FUZZY C-Means**	**Mixed C-Means**	**QKFCM**
Leukemia	3.2	1.5	0.7
Lung Cancer	5.6	2.4	0.6
Breast Cancer	4.5	2.1	0.34

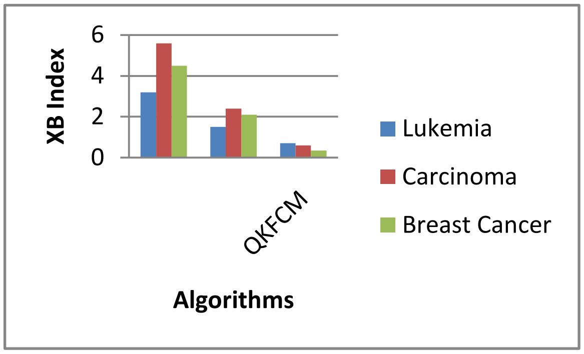

Figure 2: Comparative analysis for XB Index

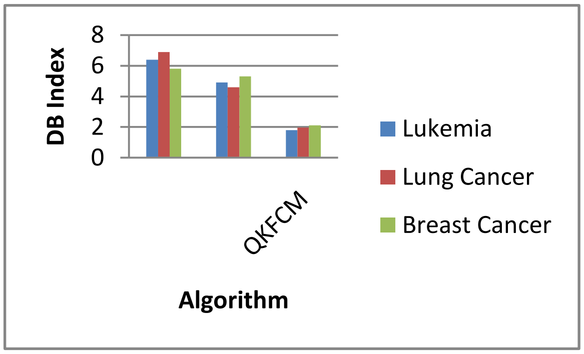

Figure 3: Comparative analysis for DB Index

V. OPTIMAL NUMBERS OF CLUSTERS

The optimal number of cluster can be determined from one of the validity measures Table 3 and Figure 5 shows the values of validity measures for various values of K. Based on the validity indices used in this paper, it can be observed that the best value of K is 2. The smaller the value of the validity index, the better the quality of clusters.

TABLE III

OPTIMAL NUMBER OF CLUSTERS

Dataset	**2K**	**4K**	**6K**	**8K**	**10K**
XB index	0.34	0.39	0.35	0.38	0.39
DB Index	2.13	2.62	2.35	2.72	2.45

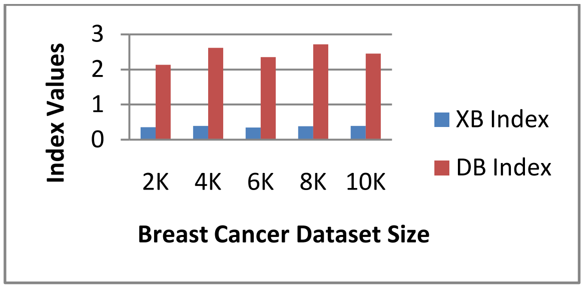

Figure 4: Optimal number of Clusters

VI. COMPUTING TIME

The computing time for the three analyzed algorithms for cancer dataset is give below in Table 4 and figure 5

TABLE IV

COMPUTATION TIME

	Computation Time		
Dataset	**FUZZY C-Means**	**Mixed C-Means**	**QKFCM**
Leukemia	31	42	25
Lung Cancer	34	44	24
Breast Cancer	36	48	22

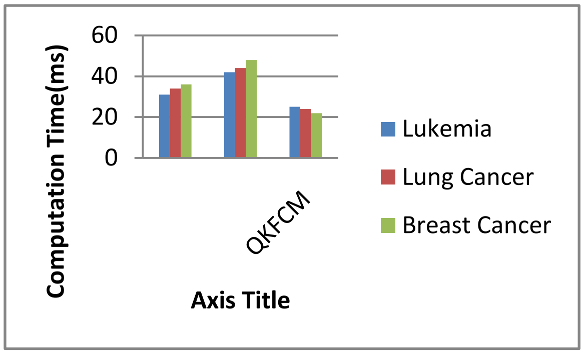

Figure 5: Computation Time

VII. INTERPRETATION OF RESULTS

Cluster analysis is an important tool in gene expression data. In clustering cancer gene expression data the genes are treated as objects and the samples are treated as attributes, the goal of the gene clustering is to identify the important gene markers. The clustering algorithms were applied to cluster cancer gene expression data and the result has given more insights to unravel the patterns hidden in the gene expression data.

Clustering of cancer gene expression data also reveals whether the mutated gene falls in the coding or promoter region. It also reveals whether there is a change in the amino acid which will code for a different protein or there is no change. This kind of clustering helps the clinicians and researchers to predict the outcome with standard drug therapy and dietary protocols.

FCM, mixed C-means and QKFCM Clustering algorithm were used in the comparative analysis of cancer gene expression data. Among this clustering algorithms QKFCM produce better cluster results which is obtained by Quad Tree Based algorithm and the computing time was greatly enhanced as in figure 5 and table 4. Recently gene therapy is employed for cancer patients. However these studies are in their infancy. But the improved technology employed here shows reasonable promise as this clustering unravels the structural and functional patterns in cancer genes. Therefore future treatment decisions based on the clustering analysis would help the clinicians in a better and efficient way.

IX.CONCLUSION

A significant step in the analysis of gene expression data is the detection of groups of genes that manifest similar expression patterns. Clustering is a classical exploratory technique of discovering similar expression patterns and functional modules. However, gene expression data are usually of high dimensions, which results in the main difficulty for the application of the clustering algorithms. The proposed QKFCM algorithms have several advantages over other algorithms that have been exploited for clustering gene expression data. The proposed clusters are unrelated and cluster boundaries are determined by the algorithm without human intervention. Moreover, the number of clusters is determined by the Quad function, instead of being a constant given as input parameter to the program. The proposed algorithm also has a better computing time compared to the other algorithms analyzed. A convenient user interface with several visualization tools was developed. We intend to develop the program further and turn it into a robust working tool for gene expression analysis which would help the clinicians in designing and standardizing diagnostic tools as well as drug therapy and dietary protocols.

REFERENCES

[1]. Duda , R. O. & Hart, P. E.(1973), Pattern classification and scene analysis, Wiley, New York.

[2]. Everitt, B.S. Cluster Analysis. 1993. Third Edition. (New York and Toronto: Halsted Press, of John Wiley & Sons Inc.).

[3]. M Telgarsky, A Vattani Hartigan's Method: k-means Clustering without Voronoi.

[4]. Mirkin CA, Letsinger RL, Mucic RC, Storhoff JJ. A DNA-based method for rationally assembling nanoparticles into macroscopic materials.

[5]. Jiang, D., Tang, C. and Zhang, A.(2004) 'Cluster analysis for gene expression data: a survey', IEEE Transaction on Knowledge and Data Engineering, Vol. 16, No. 11, pp.1370-1386.

[6]. Selva Kumar, S. and Hannah Inbarani, H. (2013) 'Analysis of mixed C-means clustering approach for brain tumor gene expression data', Int, J. Data Analysis Techniques and Strategies, Vol. 5, No. 2, pp.214-228.

[7]. J. C. Dunn A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters.

[8]. Dezdek, J. C., 1981, Pattern Recognition with fuzzy objective function algorithms, Plenum press, Newyork, NY.

[9]. Tomida S, et al. (2002) Analysis of expression profile using fuzzy adaptive resonance theory. *Bioinformatics* 18(8):1073-83

[10]. Bishnu PS, Bhattacherjee V (2012) Software fault prediction using quad tree-based K-means clustering algorithm. IEEE Trans Knowl Data Eng 24(6):1146–1150

[11]. H.-C. Huang, Y.-Y. Chuang, and C.-S. Chen, "Multiple kernel fuzzy clustering," Fuzzy Systems, IEEE Transactions on, vol. 20, no. 1, pp. 120 –134, feb. 2012.

[12]. X.L. Xie and G. Beni. A validity measure for fuzzy clustering. IEEE transactions on Pattern Analysis and Machine Intelligence, 13(4):841-846, 1991.

[13]. D. L. Davies and D. W. Bouldin. A cluster separation measures IEEE Transactions and Pattern Analysis and Machine Intelligence, PAMI-1, no. 2:222-227,1979.

[14]. Sitansu Mohanty, Kaberi Das, Debahuti Mishra, Ruchi Ranjan"Cluster Validity Indices for Gene Expression Data"International Journal of Computer Trends and Technology (IJCTT),V4(5):1465-1470 May 2013.ISSN 2231-2803.

[15]. D.Vanisri, Dr.C.Loganathan An Efficient Fuzzy Possibilistic C-Means with Penalized and Compensated Constraints