# Gene Expression Data Analysis using Fuzzy C-means Clustering Technique

Thomas Scaria
Research Scholar,
Periyar University,
Salem, TamilNadu

Gifty Stephen
Assistant Professor,
SSITS,Thaliparamba,
Kannur,

Juby Mathew, PhD
Assistant Professor,
Dept of MCA,
Amal Jyothi College of Engg.

## ABSTRACT
The challenging issue in microarray technique is to analyze and interpret the large volume of data. This can be achieved by clustering techniques in data mining. In hard clustering like hierarchical and k-means clustering techniques, data is divided into distinct clusters, where each data element belongs to exactly one cluster so that the outcome of the clustering may not be correct in many times. The problems addressed in hard clustering could be solved in fuzzy clustering technique. Among fuzzy based clustering, fuzzy c means (FCM) is the most suitable for microarray gene expression data. The problem associated with fuzzy c-means is the number of clusters to be generated for the given dataset needs to be specified in prior. The main objective of this proposed Possibilistic fuzzy c-means method is to determine the precise number of clusters and interpret the same efficiently. The PFCM is a good clustering algorithm to perform classification tests because it possesses capabilities to give more importance to topicalities or membership values. PFCM is a hybridization of PCM and FCM that often avoids various problems of PCM, FCM and FPCM. Based on the sample dataset 'lung' the entire research has been developed. The available research works already developed in this area are not exclusively working with cancer genes.At this juncture, using of the Modified Possibilitistic fuzzy c- means algorithm could be found matching with cancer genes in a better fashion. "Matlab" is used for the algorithm.The accuracy of the dataset may be identified with the usage of different training sets.Possibilistic fuzzy c means algorithm has provided better results while identifying the cancer gene. For evaluating the feasibility of the Possibilistic Fuzzy C-Means (PFCM) clustering approach, the researcher has carried out the experimental analysis.

## Keywords
Clustering, Microarray, Gene Expression, Fuzzy Clustering, FCM, PFCM

## 1. INTRODUCTION
An emergence of microarray technology has made it possible to monitor the expression levels of thousands of genes simultaneously. The Challenge is to effectively analyze and interpret this large volume of data. Two statistical operations commonly applied to microarray data are classification and clustering but the most significant area is clustering microarray data analysis [1][2].Since 40 years ago, clustering, which is one of the renowned data mining techniques, is being extensively studied and applied in numerous applications. In clustering, the whole data is divided into different sub-groups based on some similarity and each sub-group is a cluster. Numerous clustering algorithms have been reported in the literature for clustering the subjected data in an efficient way. They can be classified as nearest- nearest-neighbour clustering, fuzzy clustering, partitional clustering, hierarchical clustering, artificial neural network - based clustering, statistical clustering algorithms, density-based clustering algorithm, etc. Despite varieties of algorithm classes prevail, partitional clustering algorithms and hierarchical clustering algorithms grab great attention from the researchers. Generally, hierarchically clustering algorithms produce satisfactory level of clustering performance. However, these algorithms do not provide options for reallocation of entities. This may lead to poor classification at the initial stage. Moreover, majority of the hierarchical algorithms consumes high computational time and memory [7]. The problem associated with fuzzy is that the number of clusters to be generated for the given data set needs to be specified, this can be solved by the proposed method [3].

The rest of the paper is organized as follows. Section 2 describes related work. Section 3 presents the proposed methodology. Section 4 shows experimental results and evaluations. Finally, the conclusions and future work are presented in Section 5.

## 2. LITERATURE REVIEW
Clustering is a task of assigning a set of objects into groups called clusters. In general the clustering algorithms can be classified into two categories. One is hard clustering; another one is soft (fuzzy) clustering. Hard clustering, the data's are divided into distinct clusters, where each data element belongs to exactly one cluster. In soft clustering, data elements belong to more than one cluster, and associated with each element is a set of membership levels [9].

Clustering is considered an interesting approach for finding similarities in data and putting similar data into groups. Initial step in the analysis of gene expression data is the detection of groups of genes that exhibit similar expression patterns. In gene expression, elements are usually genes and the vector of each gene is its expression pattern. Patterns that are similar are allocated in the same cluster, while the patterns that differ significantly are put in different clusters. Gene expression data are usually of high dimensions and relatively small samples,which results in the main difficulty for the application of clustering algorithms[8]. Clustering the microarray matrix can be achieved in two ways:

- Genes can form a group which show similar expression across conditions,
- Samples can form a group which shows similar gene expression across all genes.

Methods of clustering can be categorized as Hard Clustering or Soft Clustering. Hard clustering requires each gene to belong to a single cluster, whereas Soft Clustering permit genes to simultaneously be members of numerous clusters. Hard Clustering tells whether agene belongs to a cluster or not. Whereas in Soft Clustering, with membership

values,every gene belongs to each cluster with a membership weight between 0 (doesn'tbelong) and 1 (belongs). Clustering algorithms which permit genes to belong to more than one cluster are more applicable to Gene expression. Gene expression data has certain special characteristics and is a challenging research problem [5].

**Fuzzy Clustering**

Cluster analysis is a method of grouping data with similar characteristics into larger units of analysis. First in (Zadeh, 1965) fuzzy set theory that gave rise to the concept of partial membership, based on a membership function, fuzziness was articulated and has received increasing attention. Fuzzy clustering which produces overlapping cluster partitions has beenwidely studied and applied in various areas. In fuzzy clustering, the Fuzzy C-Means (FCM) clustering algorithm is the best known and most powerful methods used incluster analysis (Bezdek, 1981). In (Yu et. al., 2007), a general theoretical method toevaluate the performance of fuzzy clustering algorithm is proposed. The Fuzzy integrated model is accurate than rough integrated model and conventional integrated model(Banu et. al., 2011). Fuzzy clustering approach captures the uncertainty thatprevails in gene expression and becomes more suitable for tumor prediction. One ofthe important parameters in the FCM is the weighting exponent m. When m is close toone, the FCM approaches the hard C-Means algorithm. When m approaches infinity, theonly solution of the FCM will be the mass center of the data set. Therefore, theweighting exponent m plays an important role in the FCM algorithm.PFCM algorithm helps in identifying hidden pattern and providing enhanced understanding of the functional genomics in a better way.The commonly used clustering technique is K-Means clustering. But this clustering results in misclassification when large data are involved in clustering. To overcome this disadvantage, Fuzzy-Possibilistic C-Means (FPCM) algorithm can be used for clustering.FPCM combines the advantages of Possibilistic C-Means (PCM) algorithm and fuzzy logic.

Finally for subset of input data, a group of K-clusters is obtained after applying the PFCM clustering method. Likewise for each set of input data a group of K-clusters is obtained. The size of the obtained group of K-clusters is less than the size of input subset of data. It is completely based on the K value.

## 3. METHODOLOGY

The researcher is attempting to cluster a large dataset with the proposed method.The Possibilistic Fuzzy C- Means (PFCM) clustering algorithm has been applied for the set of input data which is randomly divided and with which the output ' the resultant cluster' has been obtained .The fig.1 shown below represents the architecture of the proposed clustering algorithm.

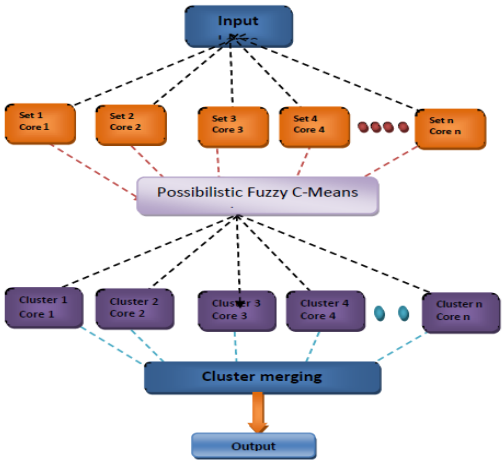


**Fig.1. Parallel architecture of the proposed algorithm**

### 3.1.Partitioning the input large dataset

Let the input be the large dataset with a size of $M \times N$ . In this processing, input large dataset using Possibilistic fuzzy c-means clustering algorithm is difficult. So dividing the input dataset randomly in to small subsets of data with equal size will make system better. So further in this proposed system the input large data set is divided into N number of subset, based on number of cores available in the system, $S = \{S_1, S_2, S_3 \ldots\ldots\ldots S_N\}$ , where N is the total number of sets with equal size. Here the each subset of data is clustered in to clusters using a standard and efficient clustering algorithm called Possibilistic Fuzzy C-Means (PFCM). Programmatically used in fork method in Java. The each single data subset S consist of a vector of d measurements, where $X = (x_1, x_2, x_3, \ldots\ldots x_d)$. The attribute of an individual data set is represented as $x_i$ and d represents the dimensionality of the vector. The Possibilistic Fuzzy C-Means (PFCM) is applied to the each subset of dataset for clustering the input dataset $n \times d$ in to k-clusters.

Possibilistic Fuzzy C-Means (PFCM) clustering method is applied to divided subset of data. The PFCM is one of the most efficient parallel clustering methods.

Let the unlabelled data set is $S = \{S_1, S_2, S_3 \ldots\ldots\ldots S_N\}$ which is further clustered in to a group of k-clusters using PFCM clustering method. This proposed PFCM is based on the minimization of the objective function given below,

$$\min_{(U,T,V)} \left\{ J_{m,\eta}(U,T,V;X) = \sum_{k=1}^{n}\sum_{i=1}^{c}\left(a u_{ik}^{m} + b t_{ik}^{n}\right) \times \|x_k - v_i\|_A^2 + \sum_{i=1}^{c}\gamma_i \sum_{k=1}^{n}(1-t_{ik})^n \right\}$$

Subject to the constraints, $\sum_{i=1}^{c} u_{ik} = 1 \quad \forall k$ , and $0 \le u_{ik}, t_{ik} \le 1$.

Here $a > 0, b > 0, m > 1, \eta > 1,$ where m is any real number greater than 1, $u_{ik}$ is the degree of membership of xi in the cluster j, xi is the ith of d-dimensional measured data, $v_i$ is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center,where, $D_{ikA} = \|X_k - v_i\| A$ and $\sum_{k=1}^{n} t_{ik} = 1 \quad \forall i$

The PFCM clustering or partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership $u_{ik}$ and the cluster centers $v_i$ by,

$$u_{ik} = \left( \sum_{j=1}^{c} \left( \frac{D_{ikA}}{D_{jkA}} \right)^{2/(m-1)} \right)^{-1}, 1 \le i \le c; , 1 \le k \le n$$

$$t_{ik} = \frac{1}{1 + \left( \frac{b}{\gamma_i} D_{ikA}^2 \right)^{1/(\eta-1)}}, 1 \le i \le c; , 1 \le k \le n$$

,

$$v_i = \frac{\sum_{k=1}^{n} \left( a u_{ik}^m + b t_{ik}^\eta \right) X_k}{\sum_{k=1}^{n} \left( a u_{ik}^m + b t_{ik}^\eta \right)}, 1 \le i \le c.$$

This iteration will stop when $\max_{ik} \left\{ \left| u_{ik}^{(k+1)} - u_{ik}^{(k)} \right| \right\} < \varepsilon$

Where, $\varepsilon$ is a termination criteria between 0 and 1, whereas k are the iteration steps. This procedure converges to local minimum of $J_{m,\eta}$ .

The PFCM clustering algorithm contains various steps;

**Algorithm 1:**

Step 1: Initialize $U = [u_{ik}] matrix, U^{(0)}$

Step 2: At k step: calculate the centers vectors $C^{(k)} = [v_i] with U^{(k)}$

$$v_i = \frac{\sum_{k=1}^{n} \left( u_{ik}^m + t_{ik}^\eta \right) x_k}{\sum_{k=1}^{n} \left( u_{ik}^m + t_{ik}^\eta \right)}, 1 \le i \le c.$$

$U^{(k)}, U^{(k+1)}$

Step 3: Update

$$u_{ik} = \left( \sum_{j=1}^{c} \left( \frac{D_{ikA}}{D_{jkA}} \right)^{2/(m-1)} \right)^{-1}$$

Step 4: If $\left\| U^{(k)} - U^{(k+1)} \right\| < \varepsilon,$ then stop; otherwise return to step 2.

Finally for subset of input data, a group of K-clusters is obtained after applying the PFCM clustering method. Likewise for each set of input data a group of K-clusters is obtained. The size of the obtained group of K-clusters is less than the size of input subset of data. It is completely based on the K value.

## 4. EXPERIMENTAL RESULTS AND DISCUSSION

The experimental result of the proposed approach is furnished in this section. The experimental evaluation is conducted in order to evaluate the proposed approach. The researcher has included the microarray dataset for the application of Matlab. Each function is represented by the respective algorithm.

The proposed algorithm gives us the precise number of clusters and is illustrated in the fig.2 which depicts the finest number of clusters as components. For the implementation, the dataset evaluation, specified gene selection range and the clusters are very much relevant. Using modified Possibilistic fuzzy c-means algorithm, the dataset for evaluation is lung, the training set is 20 and number of clusters is 3 then the time taken for evaluation 0.15686. The time taken for evaluation in k-means is 0.633 times higher than fuzzy c-means. So, modified Possibilistic fuzzy c-means is a better algorithm to identify cancer gene when compared with k-means algorithm. Modified Possibilistic fuzzy c-means algorithm provides better result than fuzzy c-means algorithm.

**Table 1: Fuzzy C-Means**

| Dataset | Gene Selection Range | No.of clusters | Time (sec) | Space (mb) |
|---|---|---|---|---|
| Lung | 25 | 3 | 2.9707 | 815 |
| | 50 | 3 | 2.9311 | 812 |
| | 75 | 3 | 3.9155 | 812 |

The selection of the dataset, specified gene selection range and the number of clusters are relevant in the Modified Fuzzy C-Means implementation. The performance and the cluster are also evaluated efficiently. It is represented as PC, CE (i.e.) Performance Calculation, and Cluster Evaluation. Using fuzzy c-means algorithm, dataset taken for evaluation is lung, the training set is 25 and number of clusters is 3 then the time taken for evaluation 2.9707. The time taken for evaluation in possibilistic fuzzy c-means is 0.17268 so modified fuzzy c-means algorithm provides best result than fuzzy c-means algorithm.

**Table 2.Possibilistic Fuzzy C MEANS**

| Dataset | Gene Selection Range | No.of clusters | Time (sec) | Space (mb) |
|---|---|---|---|---|
| Lung | 25 | 3 | 0.17268 | 807 |
| | | 5 | 0.16916 | 808 |
| | 50 | 4 | 0.17164 | 783 |
| | | 6 | 0.16094 | 778 |
| | 75 | 3 | 0.16198 | 782 |
| | | 5 | 0.1829 | 782 |

**Accuracy:**

The proposed algorithm as well as the other algorithms was applied to the reference test set using all of the ten random realizations of each training set size. Using the five different training sets of a given size, five accuracy values (in percent) were obtained. The overall accuracy was then computed by taking the mean of the resulting ten accuracy values. Table reports the overall accuracies and standard deviations on the reference test data corresponding to the training patterns of different sizes.[11] The results are given below:

**Table 3: Accuracy**

| Parameter | K-Means | Fuzzy C-Means | Possibilistic Fuzzy C-Means |
|---|---|---|---|
| Time | 0.889368 | 0.338008 | 0.159944 |
| Memory | 779.5789 | 749.3448 | 784.4118 |
| Accuracy | 91.52533 | 93.36081 | 94.95383 |

## 5. CONCLUSION

In order to gain a deep insight into the cancer classification problem, it is necessary to take a closer look at the problem, the proposed solutions and the related issues all together. In this paper, the algorithms such as k-means, fuzzy c-means are used to find the cancer affected genes in the sample dataset. The sample dataset that has been taken for research work is lung. The specified algorithms are not well functioned, in a cancer genes. So, the modified Possibilistic fuzzy c-means algorithm is proposed to grasp the cancer genes. While comparing to the algorithms like k-means and fuzzy c-means, modified PFCM is better. The time criteria are also comparatively high. The modified fuzzy c-means attains the merits of time concern and correct gene identification algorithm. The performances are also evaluated effectively by using the result of the specified algorithms.

The proposed approach is designed to address mainly for the difficulty to cluster large data bases. The proposed approach used a PFCM algorithm to handle the large data set. Our proposed method is compared with the performance of the existing k-means and fuzzy c-means clustering algorithm. The performance analysis and experimental result showed that our proposed method provide better result. Also the experimental analysis showed that the proposed approach obtained upper head over existing method in terms of accuracy, memory used and time. The highest accuracy achieved by the proposed approach is 94.95%.

## 6. REFERENCES

[1] Michel B Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein,Cluster analysis and display of genome-wide expression patterns", Proc, Natl.Acad. Sci. USA, Vol. 95, pp, 14863- 14868, December 1998.

[2] Mathew, Juby, and R. Vijayakumar. "Scalableparallel clustering approach for large data usinggenetic possibilistic fuzzy c-means algorithm", 2014 IEEEInternational Conference on Computational Intelligence and Computing Research,2014.

[3] RM Suresh, K Dinakaran, P Valarmathie, "Model based modified k-means clustering for microarray data", International Conference on Information Management and Engineering, Vol.13, pp 271-273, 2009, IEEE.

[4] AnirbanMukhopadhyay, UjjwalMaulik and Sanghamitrabandyopadhyay, "Efficient two stage fuzzy clustering of microarray gene expression data", International Conference on Information Technology (ICIT'06), 2006 IEEE.

[5] Seo Young Kim, Tai MyongChoi,"Fuzzy types clustering for microarray data", PWASET Volume 4 February 2005 ISSN 1307-6884

[6] A. L. Tarca, R. Romero, and S. Draghici, "Analysis of microarrayexperiments of gene expression profiling",American Journal of Obstetrics and Gynecology (2006) 195, pp. 373–88.

[7] C. Escudero et al., "Classification of Gene Expression Profiles: Comparison of k-means and expectation maximization algorithms", IEEE Computer Society, 2008, pp. 831-836.

[8] D. Dembele and P. Kastner, "Fuzzy C-means method for clustering microarray data", Bioinformatics, Vol. 19, Issue 8, 2003, pp. 973-980.

[9] E. Naghieh and Y. Peng, "Microarray Gene Expression Data Mining: Clustering Analysis Review", Techniques, 2009.

[10] P valarmathie,MV Srinath,T.Ravichandran and K Dinakaran "Hybrid Fuzzy C-Means clustering Technique for Gene Expression Data",Vol 1,Issue 1,International Journal of research and reviews in Applied Sciences,ISSn2076-734X.

[11] Mathew,Juby,and R Vijayakumar."Scalable parallel clustering approach for large data using parallel K means and firefly algorithms".International Conference on High Performance Computing and Applications(ICHPA),2014