BI-CLUSTERING ALGORITHM FOR MICROARRAY GENE DATA BASED ON THE COMBINATION OF FCM AND LION OPTIMIZATION ALGORITHM

P. Edwin Dhas[1], Dr. B. Sankara Gomathi[2]

[1]Research Scholar, Department of Computer Science and Engineering, Anna University, Chennai, India.

[2]Professor and Head, Department of Electronics and Instrumentation Engineering, National Engineering College, Kovilpatti, India.

rb022106@gmail.com

***Abstract*--The exploitation of microarray gene data is increasing over time, owing to the technological advancement of biomedical science. The gene data reveals many important details and is necessary to analyse the data with intense care. The gene data analysis is one of the significant research areas and this work proposes an unsupervised way of gene analysis. The gene analysis can be carried out by supervised, semi-supervised or unsupervised techniques. Both the supervised and semi-supervised analysis requires the process of system training and it requires prior knowledge about the dataset. On the contrary, the unsupervised technique involves no training and the related gene data are grouped together without any prior arrangements. Hence, this work proposes an unsupervised bi-clustering algorithm for microarray gene data by combining the Fuzzy C Means (FCM) and Lion Optimization Algorithm (LOA). The performance of the proposed work is tested in terms of precision, recall, F-measure, rand index and time consumption. The experimental results prove the efficacy of the proposed clustering algorithm.**

***Keywords*--Microarray gene data, clustering, gene analysis.**

1. INTRODUCTION

Due to the advancement in medical science, the usage of gene expression data is increasing. Microarray gene data is usually exploited to detect the functionality of the genes and to recognize the processes of genes. In particular, the gene expression data is employed to differentiate between different carcinomas [1]. Yet, the major issue with respect to microarray gene data analysis is the volume of genes in the dataset. In addition to this, it is highly challenging to locate useful genes from the voluminous set. The idea is to select useful genes from the dataset, instead of processing the whole dataset.

However, the major challenge associated with this process is the accurate selection of useful or beneficial genes. Processing informative genes bring in numerous benefits to the system and on the other hand, working with the complete dataset increases the computational, memory and time complexities. Gene data analysis may involve supervised, unsupervised or semi supervised techniques. The supervised data analysis techniques work on the labels of the entities and the unsupervised data analysis techniques does not have any labels for the entities. Hence, the unsupervised techniques do not involve any training process and works on the go. Semi-supervised techniques process the entities both with and without labels.

Mostly, the microarray data analysis is carried out with the help of standard clustering algorithms such as k-Nearest Neighbour (k-NN) [2], Fuzzy C Means (FCM) [3] and the enhanced versions of these algorithms. However, these clustering algorithms face certain issues such as initial attribute selection and convergence. The algorithms can prove better performance, when the initial parameters are chosen by some other algorithm. All the stated issues can be treated with the help of bio-inspired algorithms that helps in handling the issues such as parameter selection, better convergence and so on.

The hierarchical clustering algorithms work by building hierarchy of entities and the entities are mapped to the corresponding hierarchical level [4]. The main drawback of this technique is the poor handling of overlapping clusters. This issue is addressed by the partitional clustering approach, in which all the features are given utmost importance during the process of clustering. Yet, there is a shortcoming which is the consideration of redundant and irrelevant features for performing the process of clustering. Hence, it an absolute necessary to select the optimal count of clusters along with the exclusion of redundant features. However, achieving an effective clustering operation for microarray gene data is highly complex.

Taking this challenge into consideration, this work aims to present a clustering scheme for microarray gene data based on Fuzzy C Means algorithm and the bio-inspired Lion Optimization (LO) algorithm. Hence, this work automatically

fixes the count of clusters and chooses the relevant features on the go. The remainder of this paper is organized as follows. Section 2 presents the review of literature with respect to the microarray gene data clustering. The background of the proposed approach is presented in section 3. The proposed approach is elaborated in section 4 and the performance of the proposed clustering approach is evaluated in section 5. The concluding remarks are presented in section 6.

2. REVIEW OF LITERATURE

This section studies the state-of-the-art literature with respect to microarray gene data clustering. Unsupervised algorithms are more popular, owing to the exclusion of the training process which is meant for knowledge gaining. Hence, unsupervised algorithms weed out the requirement of having prior knowledge about the dataset.

In [5], the gene expression data is analysed by means of iterative signature algorithm. This work analyses the gene data by means of bi-clustering iterative signature algorithm and bimax. The performance analysis is carried out and the results proven that the bi-clustering iterative signature algorithm proves better performance. The module biomarkers of hepatocellular carcinoma are built for the gene expression data in [6]. In this work, the biomarker genes meant for the hepatocellular carcinoma is identified by means of bioinformatics and machine learning techniques. The gene expressions are described by means of network model and the genes related to hepatocellular carcinoma are detected. This detection is achieved by clustering the genes and the classification is achieved by applying Support Vector Machine (SVM).

A bi-clustering algorithm for gene expressions based on Shuffled Frog Leaping Algorithm (SFLA) is proposed in [7]. This work clusters the related genes together by clubbing the evolutionary memetic and particle swarm optimization algorithms. The performance of this algorithm is tested over the dataset yeast and the bi-clusters are obtained. However, this work is applied over a single dataset. In [8], a bi-clustering technique is proposed for microarray gene data. This work clusters the genes with the help of rules produced by the work. The bi-clusters are formed by means of k-means algorithm.

In [9], the microarray expression data analysis technique is presented. The expression levels of numerous genes, which can either be up or down regulated are considered by this work. This algorithm is based on the discrete optimization problem and is assumed that the data follows the checkerboard pattern. The clusters of this work are formed by considering the maximum likelihood approach. This algorithm is applied over three different datasets and the tumour classification is performed. A point symmetry based clustering approach is presented for microarray gene data in [10]. This approach presents a distributed parallel clustering algorithm with the help of k-NN algorithm. The performance of the work is tested upon four different datasets.

A microarray gene data clustering algorithm based on rough Fuzzy C Means (FCM) is presented in [11]. This work blends the advantages of both rough set theory and FCM algorithm. The reason for uniting the above two approaches is that the rough sets address the uncertainty and incompleteness of the cluster operation and the fuzzy technique can handle the cluster overlaps effectively. The performance of the proposed approach is tested on yeast datasets. In [12], an evolutionary clustering algorithm is proposed for microarray gene data. This algorithm encodes the complete cluster in a single chromosome and so the every gene of the chromosome represents a cluster. This algorithm does not require the number of clusters as input and the patterns of the clusters are shown.

In [13], a technique to cluster and select microarray gene data is proposed. This technique merges two different techniques such as fuzzy rough set theory and Self Organizing Map (SOM). The neighborhood neurons are defined by means of fuzzy rough sets. The so formed clusters are passed on to the decision table with respect to the classes. This work is claimed to prove better results. The optimal proximity measure for performing clustering operation over microarray gene data is presented in [14]. This work intends to find the better proximity measure for clustering the microarray gene data. This work considers about sixteen proximity measures and the better performing proximity measures are pearson, spearman and Euclidean distance. The performance of the proximity measures is tested upon 17 different datasets.

In [15], an algorithm based on subspace weighting for clustering microarray gene data is presented. This algorithm computes the gene subspace weight matrix through which the importance is gene objects is determined. This process continues until the required number of clusters is attained. The performance of this work is tested on different datasets and the results are claimed to be satisfactory. A resampling based clustering algorithm is proposed for microarray gene data in [16]. This work addresses the issues caused by noise by utilizing the mixed noise effect model for detecting the variances and the quasi Markov Chain Monte Carlo (MCMC) algorithm is employed for statistical inference. This algorithm produces better clusters.

Motivated by these existing works, this work intends to present a time conserving unsupervised microarray gene data clustering based on LO algorithm. The following section presents the background of the algorithm.

3. BACKGROUND

3.1 Fuzzy C Means (FCM) Algorithm

The FCM algorithm allots each and every entity with a representation degree of 0 to 1. Hence, the summation of representation of each entity in a cluster is 1. This problem can be seen as a minimization problem and the values are determined by means of the degree of representation. The functional value of FCM is computed by the variance between the entities and the centroid of the cluster. As this function needs to be minimized, non-linear optimization problem has to be solved and it can be attained by any meta-heuristic

algorithm. The clustering operation of FCM is performed by the following equation [17,18].

$$CL_{FCM} = \sum_{i=1}^{K} \sum_{j=1}^{C} \mu_{ij}^{f} \left|\left|a_i - c_j\right|\right|^2 \quad (1)$$

In the above equation, μ_{ij} is the fuzzy partition matrix and it ranges from 0 to 1. i is the entity, j is the centroid of the cluster and f represents the fuzziness of the clusters. The squared Euclidean distance between the entity and the centroid is computed by $\left|\left|a_i - c_j\right|\right|^2$.

3.2 Lion Optimization Algorithm

LO algorithm is a bio-inspired algorithm that duplicates the behaviour of original lions. On studying the life policy of lions, it can be noticed that the lions organize themselves as resident and nomadic lions [19, 20]. The resident lions live in a group named as pride and each pride consists of about four to five female lions, their cubs and one or two male lions. Eventually, the cubs grow and they reach maturity. At this juncture, the young male lions are expelled from the pride. As these young lions do not have any company, they wander without any specific principle. However, a nomadic lion can become a part of a pride, when it can defeat the matured lion inside a pride. At this point, swap happens inside a pride and this may happen at any point of time. Hence, the stay at a pride is impermanent for any lion. The overall process of the LOA is presented as follows.

Standard deviation, Mean and Median absolute deviation is been extracted from the waveforms with the help of WT. After the extraction the most distinguished features that is much suited in the analysis will be extracted with Fuzzy C-means clustering to classify and group the distinct signals. In multi-dimensional space Fuzzy C-means algorithm groups the data points in to a specific number of cluster .The extracted feature by WT were given as a input to Fuzzy c-means clustering to determining center c_i and the membership matrix U which is done based on minimization of the objective function shown in equation (3).

LOA

Produce initial population of lions N_p
Declare the count of prides and nomadic lions
    Choose the percent of nomadic lions from N_p randomly and number of prides;
    Choose the gender rate of lions in each pride;
For each pride do
    Select a lioness randomly for hunting;
    Assign the remaining lionesses a position from the territory;
    Set the roaming percent of each male lion in the pride and the mating probability;
    Expel the weakest lion from the pride;
For each nomadic lion do
    Both the male and female nomadic lions wander randomly;
    Set the mating probability of nomadic female lion with the male lion;
    Nomadic lions attack the pride randomly;
For each pride do
    Fix the percentage of lioness that can become nomad;
Do
    Sort the nomad lions with respect to the gender;
    Choose best lionesses and distribute to fulfil the pride;
    Check the completeness of all the available prides;
    Eliminate the lions with minimal fitness value;
While (termination condition);

The activities of lionesses are quite different from the male lions and the prey hunting is usually performed by the lionesses and not lions. Each pride of lions is confined to a selective location. The lionesses encircle the targeted prey to initiate the attack. The input parameters required for this algorithm are initial population with the associated percentage of nomadic and residential lions. In the meantime, the lioness reproduces cubs and the cycle of getting inside and outside of a pride continues. The optimal solution to a problem is found by arranging the nomadic lions based on their fitness value. When the fitness value of the lions is lower, they are eliminated from consideration. Hence, the lion has to be fitter for capturing a place in the pride. Additionally, it can be considered that the fitter lions alone can have place in the pride. This process is continued until the best possible solutions are attained. Based on this background knowledge, the microarray gene clustering data is proposed as follows.

4. Proposed Microarray Gene Clustering Algorithm based on FCM and LOA

As stated above, the microarray gene data contains more intricate information. However, the data contains beneficial information and can be used for predicting the current status of a person. In order to utilize the data for achieving a goal, a powerful data analytic algorithm is required. The data analysis can be performed in two ways, which are data clustering and data classification. Data clustering exhibits more merits than data classification and the reasons are listed as follows. The overall flow of this work is depicted in figure 1.

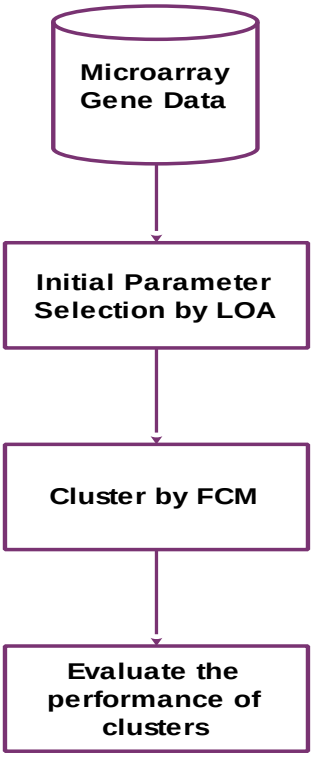

Fig.1. Overall flow of the work

Data clustering requires no knowledge in advance about the dataset and is unsupervised. Data clustering can be performed on the go, without any special preparation. On the other hand, data classification is supervised and can be accomplished only when the classifier is trained. The training process incurs extra time and having prior knowledge about the dataset is not always possible.

Considering these points, this work proposes an unsupervised microarray gene data clustering algorithm based on FCM and LOA. FCM is an efficient clustering and the main concern of FCM is the selection of the initial points. When the initial point selection is carried out by some other algorithm, FCM can prove its best. For this sake, this work utilizes LOA which proves better performance. LOA is proven with speedy convergence and better global optimality. This is the reason for the choice of LOA and the proposed algorithm is presented as follows.

Proposed FCM-LOA Algorithm

Input : Entities, m, termination condition
Output : Clustered gene data (μ_{fsbl})
Begin
Choose initial points by LOA; $\mu_0 := LOA(entities, m)$
Do
Compute $c_j = \frac{\sum_{i=1}^{K} \mu_{ij}^{f} a_i}{\sum_{i=1}^{K} \mu_{ij}^{f}}$;
if (current solution is better than existing $\vee$ *met termination condition)*
End the process;
Else
Compute $\mu_{ij} = 1/\sum_{p=1}^{C} \left(\frac{||a_i - c_j||}{||a_i - c_p||} \right)^{\frac{2}{f-1}}$
End if;
$\mu_{fsbl} := \mu_{ij}$;
While (termination condition)
End;

The traditional FCM algorithm randomly chooses the initial points, which results in huge count of iterations for reaching the final feasible solution. This in turn consumes more resource, which is not recommendable. This shortcoming of FCM is addressed by means of LOA in which the centroids are chosen by the LOA. This work separates between two clusters and hence, the entities belong to either cluster 1 or 2. Therefore, the combination of FCM and LOA reduces the count of iterations considerably, which in turn conserves the resources and enhances the performance of the algorithm. As the optimal centroids are chosen by LOA before being passed to FCM, faster convergence is experienced. The performance of the proposed FCM-LOA algorithm is evaluated in the forthcoming section.

5. RESULTS AND DISCUSSION

The proposed FCM-LOA algorithm is simulated by exploiting MATLAB on a stand-alone computer with Intel i7 processor and 8 GB RAM. The performance of the proposed algorithm is evaluated on two prominent datasets such as Acute Lymphoblastic Leukemia – Acute Myeloid Leukemia (ALL-AML) and colon tumour [22]. The ALL-AML dataset possesses 3571 genes and 72 instances. The colon tumour gene dataset 2000 genes and the total instances of this dataset are 60. Both these dataset involves two clusters.

To start with the LOA, the initial parameters are set as follows. The initial population of 50 and the termination condition is set as 1000 iterations. The count of pride is 4 and the gender rate is fixed as 0.8. The percentage of nomadism and mating probability of lions is set as 20 and 30 percentages respectively. Totally, twenty bi-clusters are produced as result and the quality of the formed clusters are evaluated by randomly selecting the clusters.

As this work forms bi-clusters out of the whole data, each cluster must possess the genes with related expression levels. The performance of this work is evaluated by means of standard performance measures such as precision (P), recall (R), f-measure (F) and rand index [23]. The formulae for

computing the above stated performance metrics are presented as follows.

$$P(x,y) = \frac{Q_{xy}}{Q_y} \quad (2)$$

$$R(x,y) = \frac{Q_{xy}}{Q_x} \quad (3)$$

$$F(x,y) = \frac{2*R(x,y)*P(x,y)}{P(x,y)+R(x,y)} \quad (4)$$

In the above equations, $P(x,y)$ denotes the probability of an entity in cluster x to be a part of class y. Q_{xy} is the total number of entities in class y of cluster x and Q_x is the total number of entities in class x. $R(x,y)$ is the recall rate of cluster x by considering the class y. Here, Q_x is the total count of entities in class y and Q_{xy} is the total count of entities y in cluster x. The rand index compares the pairs of entities being present in the cluster. The rand index takes the value from 0 to 1. The value one indicates that the pairs are relevant to each other, which is computed by considering the actual place of an entity with the ground truth cluster. In case of perfect placement of an entity, the value is set as 1.

$$RI = \frac{Total\ acceptance}{Total\ acceptance + Total\ rejections} \quad (5)$$

The total acceptance denotes the acceptance of an entity inside a cluster and the rejections indicate that the denied permission for an entity to be a part of a cluster. The experimentation is carried out by dividing the ALL-AML dataset into four datasets for easy processing. The experimental results attained by the proposed approach are presented in table 1 and 2. In the second round of performance comparison, the performance of the proposed approach is compared with the existing techniques such as FCM [17], rough fuzzy [11] and SOM [13] in terms of precision, recall, F-measure and time consumption.

Dataset ALL AML	Part 1		Part 2		Part 3		Part 4	
	Cl-1	Cl-2	Cl-3	Cl-4	Cl-5	Cl-6	Cl-7	Cl-8
Precision	1.0	0.84	0.91	1	0.66	1	0.81	0.83
Recall	0.91	1.0	1.0	0.81	1.0	0.26	0.91	0.72
F-measure	0.95	0.93	0.96	0.88	0.82	0.46	0.84	0.78
Rand-Index	0.62	0.63	0.72	0.75	0.89	0.87	0.65	0.66
Time (sec)	2.0	2.1	2.5	2.4	1.0	1.09	1.0	0.9

Table 1. Experimental results on ALL-AML dataset

Colon Dataset	Dataset 1		Dataset 2	
	Cl-1	Cl-2	Cl-3	Cl-4
Precision	0.57	1	0.61	0.54
Recall	1	0.06	0.6	0.54
F-measure	0.73	0.16	0.7	0.52
Rand-Index	0.94	0.95	0.54	0.56
Time (sec)	2.5	2.4	2.3	2.4

Table 2. Experimental results on Colon dataset

The colon dataset is divided into two parts and the performance metrics are computed for the clusters. The performance of the proposed work is proven by comparing the work with existing algorithms and the average results are presented as follows.

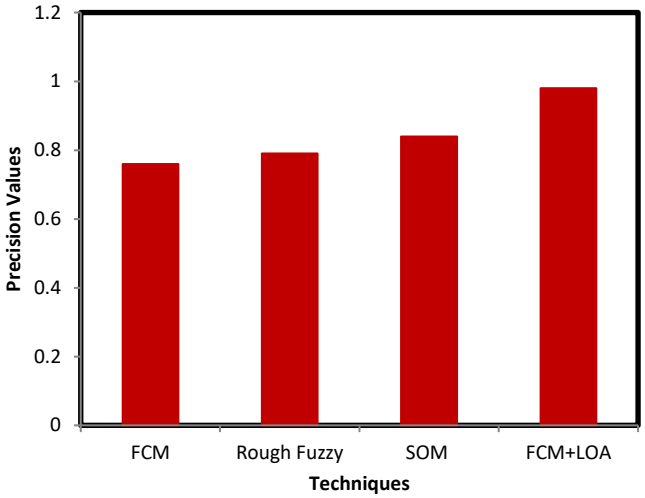

Fig.2 (a) Precision analysis

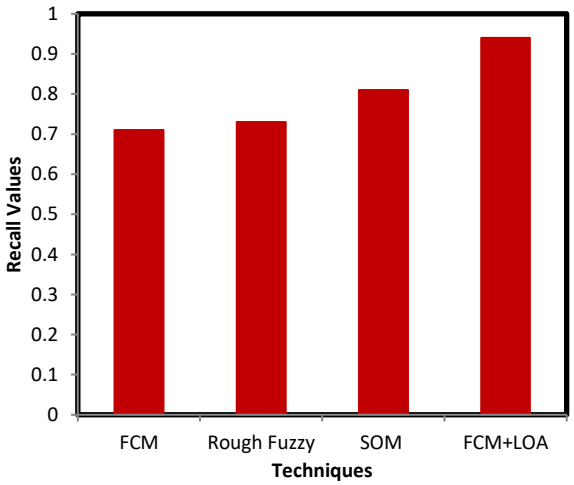

Fig.2 (b) Recall rate analysis

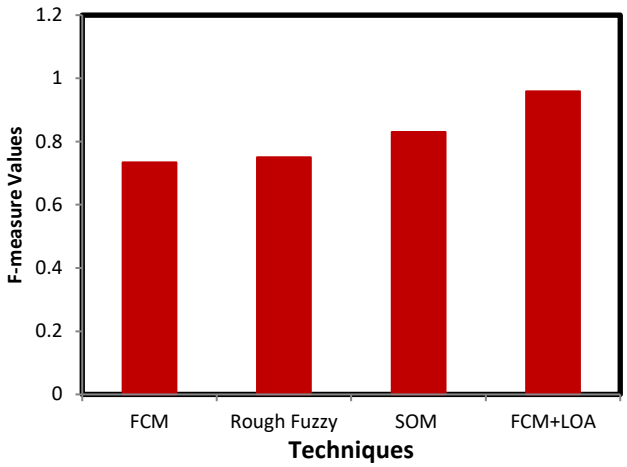

Fig.2 (c) F-measure rate analysis

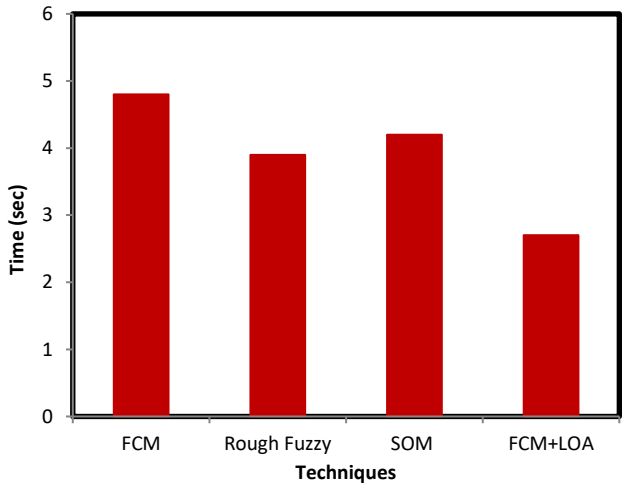

Fig.2 (d) Time consumption analysis

On analysis, it is proven that the FCM clustering algorithm works better, when the initial centroid points are chosen by metaheuristic algorithm. This work employs the metaheuristic algorithm LOA for selecting the initial centroids of the cluster. This enhances the performance of the clustering and is evident through the experimental results. The proposed approach shows greater precision, recall rates. This improves the rate of F-measure automatically and the time consumption of the proposed approach is very less because of the optimal selection of centroids by LOA. In substance, the combination of FCM and LOA results in forming better clusters of microarray gene data.

6. CONCLUSIONS

This article presents a clustering algorithm for microarray gene data, which is based on the combination of FCM and LOA. FCM performs well and the main shortcoming of this algorithm its inefficient selection of initial points. Additionally, FCM requires more number of iterations to form clusters. This drawback is addressed by employing a metaheuristic algorithm for selecting the initial points for performing clustering operation. The LOA effectively selects the initial points for FCM and the FCM forms clusters effectively. The quality of clusters is evaluated by means of precision, recall, F-measure, rand index and time consumption rates. The proposed approach is tested over two datasets such as ALL-AML and colon tumour and the performance of the proposed clustering algorithm is tested against existing techniques. From the experimental results, it is evident that the performance of the proposed work is satisfactory in terms of clustering ability. In future, this work is planned to be extended by studying the potential of metaheuristic algorithms over gene data.

REFERENCES

[1] Pal, N. R., Aguan, K., Sharma, A., & Amari, S., "Discovering biomarkers from gene expression data for predicting cancer subgroups using neural networks and relational fuzzy clustering", BMC Bioinformatics, Vol.8, No.5, pp. 1-18, 2007.

[2] D. Jiang, C. Tang, and A. Zhang, “Cluster analysis for gene expression data: A survey,” IEEE Trans. Knowl. Data Eng., Vol. 16, no. 11, pp. 1370–86, Nov. 2004.

[3] Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, 10(2-3), 191-203.

[4] Sheng W, Liu X, Fairhurst M, "A niching memetic algorithm for simultaneous clustering and feature selection", IEEE Trans Knowl Data Eng, Vol. 20, No. 7, pp. 868–879, 2008.

[5] K. Vengatesan ; Sanjeevikumar Padmanaban ; R. P. Singh ; T. Nadana Ravishankar ; Mahajan Sagar Bhaskar ; M. Ramkumar, "Performance analysis of gene expression data using biclustering iterative signature algorithm", International Conference on Intelligent Computing, Instrumentation and Control Technologies, 6-7 July, Kannur, Kerala, 2017.

[6] Chen Shen; Zhi-Ping, "Identifying module biomarkers of hepatocellular carcinoma from gene expression data", Chinese Automation Congress, 20-22 Oct, Jinan, China, 2017.

[7] Priyojit Das ; Sujay Saha, "A novel SFLA based method for gene expression biclustering", Third International Conference on Research in Computational Intelligence and Communication Networks, 3-5 Nov, Kolkata, India, 2017.

[8] M P Shruthi, "Biclustering on gene expression data", International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies, 16-18 Feb, Chennai, India, 2017.

[9] Desmond J. Higham; Gabriela Kalna; J. Keith Vass, "Spectral analysis of two-signed microarray expression data", A Journal of the IMA Mathematical Medicine and Biology, Vol.24, No.2, pp.131-148, 2007.

[10] Anasua Sarkar; Ujjwal Maulik, "Rough Based Symmetrical Clustering for Gene Expression Profile Analysis", IEEE Transactions on NanoBioscience, Vol.14, No.4, pp.360-367, 2015.

[11] Pradipta Maji; Sushmita Paul, "Rough-Fuzzy Clustering for Grouping Functionally Similar Genes from Microarray Data", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol.10, No.2, pp. 286-299, 2013.

[12] P.C.H. Ma ; K.C.C. Chan ; Xin Yao ; D.K.Y. Chiu, "An evolutionary clustering algorithm for gene expression microarray data analysis", IEEE Transactions on Evolutionary Computation, Vol.10, No.3, pp. 296-314, 2006.

[13] Shubhra Sankar Ray; Avatharam Ganivada; Sankar K. Pal, "A Granular Self-Organizing Map for Clustering and Gene Selection in Microarray Data", IEEE Transactions on Neural Networks and Learning Systems, Vol.27, No.9, pp.1890-1906, 2016.

[14] Pablo A. Jaskowiak; Ricardo J.G.B. Campello; Ivan G. Costa,"Proximity Measures for Clustering Gene Expression Microarray Data: A Validation Methodology and a Comparative Analysis", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol.10, No.4, pp.845-857, 2013.

[15] Xiaojun Chen ; Joshua Zhexue Huang ; Qingyao Wu ; Min Yang, "Subspace Weighting Co-Clustering of Gene Expression Data", IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp.1-1, 2017. DoI. 10.1109/TCBB.2017.2705686

[16] Han Li ; Chun Li ; Jie Hu ; Xiaodan Fan, "A Resampling Based Clustering Algorithm for Replicated Gene Expression Data", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol.12, No.6, pp. 1295-1303, 2015.

[17] Bezdek,J.C., Ehrlich, R., Full, W., 1984. FCM: The fuzzy c-means clustering algorithm. Comput. Geosci. 10(2–3), 191–203.

[18] Zhou, K., Fu, C., Yang, S., 2014. Fuzziness parameter selection in fuzzy c-means: The perspective of cluster validation. Sci. China Inf. Sci. 57(11), 1–8.

[19] Mccomb, K, et al. Female lions can identify potentially infanticidal males from their roars. Proc. R. Soc. Lond. Ser B: Biol. Sci. 1993; 252 (1333)59–64.

[20] Schaller GB. The Serengeti lion: study of predator–prey relations. Wildlife behavior and ecology series. Chicago, Illinois, USA: University of Chicago Press; 1972.

[21] Cheng, & Church, GM 2000, 'Biclustering of expression data': proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, pp. 93-103.

[22] http://csse.szu.edu.cn/staff/zhuzx/Datasets.html

[23] Rajesh.E and Srinivasan Alavandar, 'ANN based data mining system for early detection and classification of ECG signals using wavelet', International Journal of Printing, Packaging and Allied Sciences, vol. 04, no. 5, pp. 3567-3580, 2016.

[24] Eva Freyhult, Mattias Landfors, Jenny Önskog, Torgeir R. Hvidsten, Patrik Rydén. "Challenges in microarray class discovery: a comprehensive examination of normalization, gene selection and clustering", BMC Bioinformatics, Vol.11, pp.503, 2010.