Int J Pharm Bio Sci 2015 July; 6(3): (B) 395 - 400

Research Article **Bioinformatics**

International Journal of Pharma and Bio Sciences **ISSN 0975-6299**

YEAST GENE EXPRESSION ANALYSIS USING K MEANS AND FCM

S.ANUSUYA, DR.N. USHA BHANU AND E. KASTHURI

Saveetha School of Engineering, Saveetha University, Chennai

ABSTRACT

The experiments on gene expression analysis is to analyze the thousands of genes at once to know the global picture of cell function which aids to study the regulatory gene defects like cancer and other devastating diseases, cellular responses to the environment and cell cycle variation etc. The determination of the pattern of genes help to identify the level of genetic transcription under various time periods. In this work, we apply partitional clustering algorithms such as K Means and Fuzzy C Means clustering on yeast gene expression profiles to group the similar genes. The validity of clusters is analyzed with the Davis Bouldin Index(DBI). FCM has achieved the DBI of 0.31452 for K=3 and 0.37822 for K=4 which is better than K Means clustering.

KEYWORDS: Clustering, Yeast gene expression, K Means, FCM and DBI.

*Corresponding author

S.ANUSUYA
Department of Information Technology, Saveetha School of Engineering, Saveetha University

I. INTRODUCTION

Computer assisted knowledge extraction tool or methodology provides solution to manage high throughput data. Innovative approaches on Gene Expression Analysis have been developed to group genes which are functionally related and the gene expression data are monitored concurrently for a large pool of genes. DNA Microarray technology reveals or measures expression level of tens of thousands of genes, relationship between genes and their functions simultaneously. Since, huge amount of data are produced through micro array analysis, clustering becomes an essential step and it can be used as a preprocessing step before classification in order to restrict the analysis to a specific category. K means clustering and Fuzzy C Means clustering are the partitional clustering methods widely used to cluster gene expression data where each gene assigned to one cluster.

I.1 Related Work

Fuzzy C-Means has been applied to yeast data sets to produce meaningful clusters where genes are assigned into multiple clusters according to Membership[1]. Prior Knowledge on Gene Ontology have been incorporated for initialization and finding membership. An algorithm called Dynamic MultiObjective Particle Swarm Optimization Biclustering(DMOPSOB) for mining biclusters from microarray datasets such as yeast and Human B-cells expression dataset proposed by JuwanLiu, JunwanLiu and Yiming. Authors have insisted the eminence of Particle Swarm Optimization in gene clustering and multiple fitness functions are used to reduce the shortcomings of PSO and to attain global Optimum Solutions. Semi supervised fuzzy clustering [3] has been demonstrated to gene expression data with the influence of an idea of supervised information to improve the result of unsupervised learning. Further, the algorithm has been extended to find the number of clusters automatically. The significance of algorithm is proved against some standard clustering methods in terms of sensitivity, specificity and ARI(Automated Readability Index) index. Rough Fuzzy clustering for grouping functionally similar genes from microarray data has been proposed by integrating the merits of rough sets and fuzzy sets to provide significant and relevant gene clusters [4]. The integration of K-Means and exploratory mechanism known as eXploratory K-Means is proposed to prevent the local convergence of traditional K-Means [5]. K –Means optimized using GA is proposed for automatic detection of right number of clusters. The initial seeds for clustering are adjusted using GA for quality clusters [6]. Fuzzy partition based on probabilistic data distribution is used to estimate the missing values of the dataset [7]. The method referred as Fuzzy Clustering of Large Applications based on Randomized Search (FCLARANS) is performed for gene clustering after dimensionality reduction using the same method. The biologically enriched and significant clusters are identified using this approach [8]. The application of Optimized FCM is tested on the MRI image and superiority has been proved [10]. FCM has been proposed to process high dimensional data [11].

II Clustering using FCM and K Means

II.1 Fuzzy C-Means

Fuzzy clustering groups data by permitting each sample to belong to more than one cluster. Considering a finite set X= {x_1, x_2, Xian} of N vectors in an M-dimensional space, the aim is to perform a partition on this dataset with respect to a given criterion. FCM returns cluster centers C= {c_1, c_2… c_c} and a fuzzy partition matrix. The elements in fuzzy partition matrix range from 0 to 1. FCM has been applied in different applications due to its efficacy and simplicity. The objective function of FCM is given in Eq.1.

$$\min J_{FCM} = \sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m} d_{ij}^{2} = \sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m} \| x_j - a_i \|^2 \quad (1)$$

Where,

$$a_i^{(t)} = \frac{\sum_{j=1}^{n}\left[\mu_{ij}^{(t-1)}\right]^m x_j}{\sum_{j=1}^{n}\left[\mu_{ij}^{(t-1)}\right]^m}; i = 1,2...c$$

$$\mu_{ij}^{(k)} = \left[\sum_{i=1}^{c}\left[\frac{[(x_i - a_i^{(k)})'(x_i - a_i^{(k)})]}{[(x_i - a_1^{(k)})'(x_i - a_1^{(k)})]}\right]^{1/m}\right]^{-1}$$

II.2 K Means

The K-means clustering algorithm is one of the greedy algorithms which attempts to find the best data partitioning into a given number (k) of clusters. In a K Means clustering problem, a set of n points V = {v_1, v_2, ..., v_n} and the value of k is set, and the aim is to partition P of V into k clusters, which minimizes the objective function. The steps involved in K Means clustering is given below.

1. Select K number of initial centroids randomly from data points {v_1, v_2,….v_n}.
2. Assign a data point v_i to the cluster c_j where $0 \le j \le k$

$$K_{obj} = \min \sum_{i=1}^{K}\sum_{x \in v_i} \left\| x - c_j \right\|^2 \quad (2)$$

Where μ_i is the mean of points in v_i.

3. Compute new centroids as $c_j = \frac{1}{|c_j|}\sum_{x_i \in c_j} x_i$ and repeat step 2.

4.If less or no change of the centroids, then stop, otherwise continue to the step 3

III Cluster Validity

In this work, the cluster validity is tested with DBI index[9]. The formula for DBI is given in Eq.3

$$DBI = \frac{1}{n}\sum_{i=1}^{n} max\left[\frac{P_i + P_j}{Q_{i,j}}\right] \quad (3)$$

Where n is the number of clusters, P_i and P_j are the intra distances of cluster i and j and $Q_{i,j}$ is the inter cluster distance between cluster i and j.

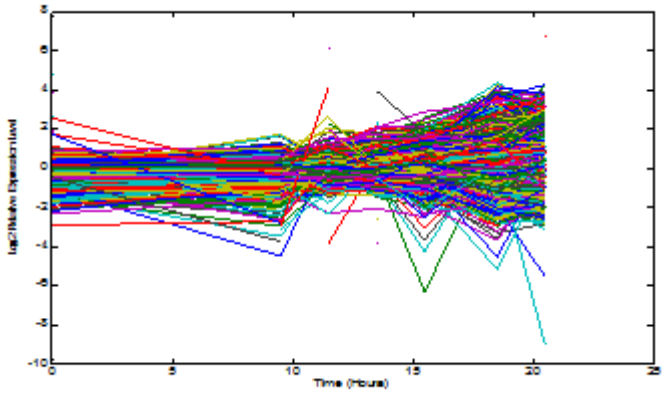
$$R_{i,j} = \frac{P_i + P_j}{Q_{i,j}}$$

The properties of Ri,j are Ri,j≥0 and Ri,j = Rji.

IV EXPERIMENTAL RESULTS

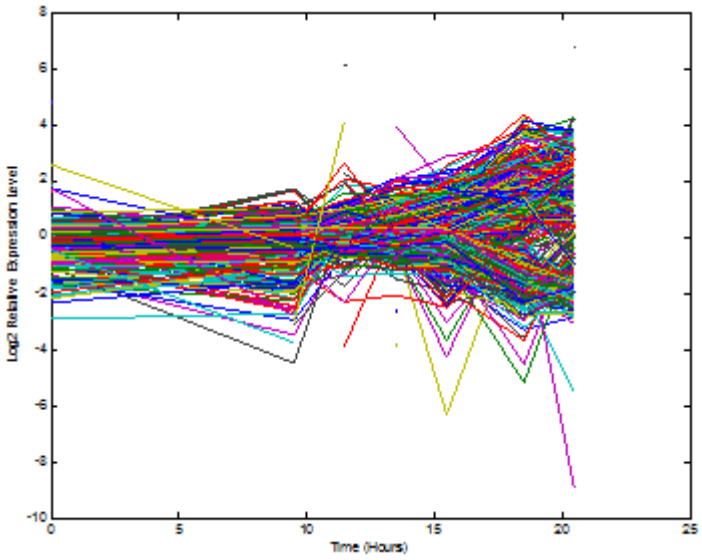
Yeast gene expression dataset consist of gene expression level values for seven time steps and it consists of 6400 genes before filtration. The expression levels of all genes with is shown in Figure 1.

Figure 1
Expression levels of yeast gene.



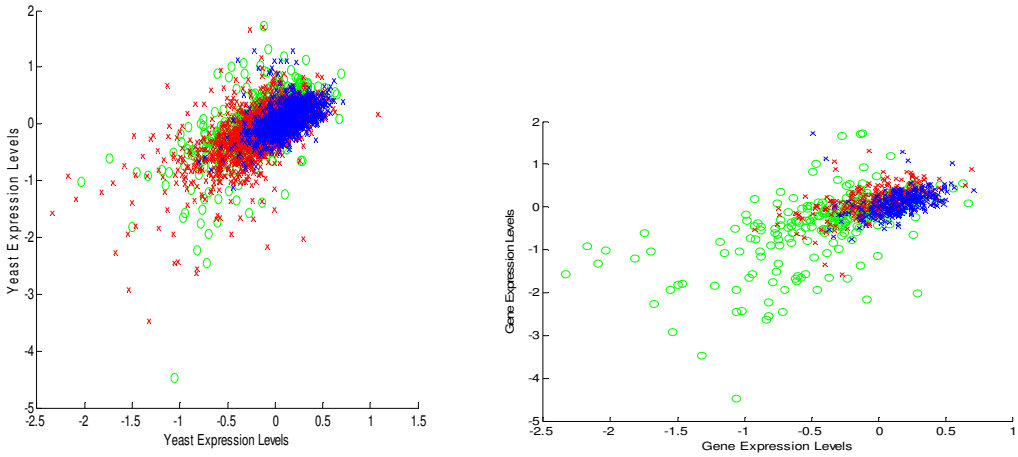
The expression levels of gene profile after applying low variation filter is shown in Figure 2.

Figure 2
Expression levels of yeast gene after low variance filter.



The genevarfilter function to filter out genes with small variance over time.

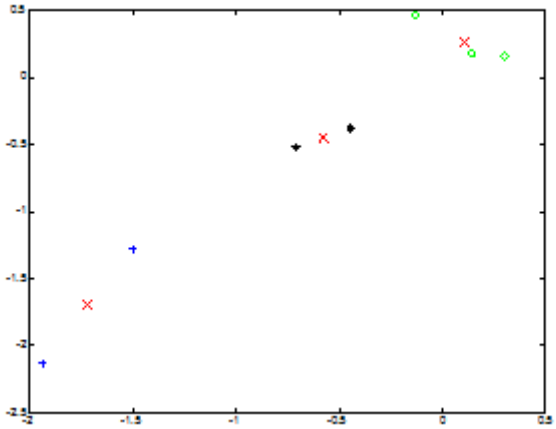
Figure 3.a Fuzzy C means clustering of yeast gene expression for K=3, 3.b Fuzzy C means clustering of yeast gene expression for K=4.



3. a **3.b**

Clustering results of yeast gene expression profile for K=3 and for K=4 is shown in Figure 3.

Figure 4
The centroids obtained using K Means for K=3.



The centres obtained in K means shown in figure 4 and the obtained DBI on FCM and K Means is shown in Table 1.

Table 1
DBI value obtained on different number of clusters

Method	Number of clusters (K)	DBI value
FCM	3	0.31452
FCM	4	0.37822
K Means	3	0.39821
K Means	4	0.39132

V CONCLUSION

Data analysis of gene expression profile analysis has become an area of intense research. In this work, we use the clustering methods FCM and K Means to cluster yeast gene expression data. Generally, the gene database is embedded with noise and many

preprocessing techniques have been used to remove or reduce the noise. In this work, the deviating genes, the genes with low variance are filtered first, then the clustering has been applied with K=3 and K=4. The validity of clusters are tested with DBI index and found that FCM produces better results of 0.31452 for K=3 and 0.37822 for K=4.

REFERENCES

1. Luis Tari., Chitta Baral., Seungchan Kim. Fuzzy c-means clustering with prior biological knowledge. Journal of Biomedical Informatics, 42(1): 74–81, (2009)
2. Junwan Liu., Junwan Liu.,Yiming Chen. Dynamic Biclustering of Microarray Data with MOPSO. IEEE International Conference on Granular Computing, 978-0-7695-4161-7/10, DOI 10.1109/GrC.2010.44, (2010)
3. Ioannis A., Maraziotis. A semi-supervised fuzzy clustering algorithm applied to gene expression data. Pattern Recognition, 45(1): 637–648, (2012)
4. Pradipta Maji., Sushmita Paul. Rough-Fuzzy Clustering for Grouping Functionally Similar Genes from Microarray Data. IEEE/ACM Transactions On Computational Biology And Bioinformatics, 10(2):286-299, (2012)
5. Yau King Lam., Peter W.M. Tsang. eXploratory K-Means: A new simple and efficient algorithm for gene clustering. Applied Soft Computing, 12(3): 1149-1157, (2012)
6. Md Anisur Rahman., Md Zahidul Islam. A hybrid clustering technique combining a novel genetic algorithm with K-Means**.** Knowledge-Based Systems, 71(1):345-365, (2014)
7. Thanh Le., Altman T., Gardiner K.J. Probability-based Imputation Method for Fuzzy Cluster Analysis of Gene Expression Microarray Data. Ninth International Conference on Information Technology: New Generations (ITNG), 42 – 47, (2012)
8. Sampreeti Ghosh., Sushmita Mitra., Rana Dattagupta. Fuzzy clustering Applied Soft Computing. Journal of Applied Soft Computing, 16(1): 102-111, (2014)
9. Davies D.L., Bouldin D.W. A cluster separation measure. IEEE Transaction on Pattern Analysis and Machine Intelligence, PAMI-1, 224–227, (1979)
10. Anusuya Venkatesan., Latha Parthiban., K.Arul. ROI Detection and Segmentation Of Medical Images Using Optimized Thresholding And Clustering. Int J Pharm Bio Sci, 4(3): 1235 – 1245, (2013)
11. Xianen Qiu.,Yanyi Qiu., Guocan Feng ., Peixing Li. A sparse fuzzy c-means algorithm based on sparse clustering framework, Neurocomputing, 157(1): 290–295, (2015)