Data Analytics and Artificial Intelligence
Vol: 2(5), 2022
REST Publisher; ISBN: 978-81-948459-4-2
Website: http://restpublisher.com/book-series/data-analytics-and-artificial-intelligence/

Biological Significance of Gene Expression data Using Clustering Techniques

* **B. Sapna Devi, R. Christy Jancy Rani**
Gonzaga Arts and Science College for Women Krishnagiri, Tamil Nadu, India
*Corresponding author Email: Sapna90.it@gmail.com

Abstract. Gene clustering is the process of grouping related genes in the same cluster is at the foundation of different genomic studies that aim at analyzing the function of genes. A gene cluster is a set of two or more genes that serve to encode for the same or similar products. Gene clustering methods serve this purpose. Several advanced techniques have been proposed for data clustering and many of them have been applied to gene expression data, with partial success. The goal of gene clustering is to identify important genes and perform cluster discovery on samples. In this paper, first the concepts of Clustering Techniques are discussed briefly and the basic elements of clustering techniques are conversing for gene expression data. It also studies three of the most representative off-line clustering techniques: Rough K-Means, Fuzzy C- means clustering, and Hierarchical clustering. These techniques are implemented and tested against a Yeast gene expression Dataset, Colon Gene Expression Dataset and Breast Gene Expression Dataset. The performance of the three techniques is compared based on "goodness of clustering "evaluation measures and Rough K-Means show best performance than the other two clustering techniques for the above gene expression data..
Keywords: Clustering; Gene Expression Data; Fuzzy C-Means; styling; Rough K-Means; Hierarchical Clustering).

1. Introduction

Clustering algorithms are used extensively not only to organize and categorize data, but are also useful for data compression and model construction. By finding similarities in data, one can represent similar data with fewer symbols for example. Also if one can find groups of data, one can build a model of the problem based on those groupings. Data clustering methods have been proven to be a successful data mining technique in the analysis of gene expression data. A gene cluster is a set of two or more genes that serve to encode for the same or similar products. Because populations from a common ancestor tend to possess the same varieties of gene clusters, they are useful for tracing back recent evolutionary history. Gene expression data is usually represented by a matrix, with rows corresponding to genes and columns corresponding to conditions, experiments or time points. The content of the matrix is the expression levels of each gene under each condition. Those levels may be absolute, relative or otherwise normalized. Each column contains the results obtained from a single array in a particular condition and is called the profile of that condition. Each row vector is the expression pattern of a particular gene across all the conditions. More formal definitions will be given in the sequel. Clustering gene expression data are especially challenging computational tasks. Clustering genes groups similar genes into the same cluster based on a proximity measure. Genes in the same cluster have similar expression patterns. In gene-based clustering, the genes are treated as the objects, while the samples are the features. In sample based clustering, the samples can be partitioned into homogeneous groups where the genes are regarded as features and the samples as objects. The purpose of gene-based clustering is to group together co-expressed genes which indicate co-function and co-regulation. In this Paper, Three Clustering techniques are studied: Rough K-Means Clustering, Fuzzy C-Means Clustering and Hierarchical Clustering. The above three techniques are implemented and tested against a Gene Expression Dataset (Yeast, Colon and Breast). The results are presented with a comprehensive comparison of the different techniques and the effect of different parameters in the process. The Organization of the paper is as follows: Section 1 describes the clustering Approaches and explains the gene expression analysis: challenges and applications. Section 2 presents a review of literature on categories of gene expression data clustering. Section 3 explains the Clustering Techniques. Section 4 deals with the Experimental analysis of the proposed approach. Chapter 5 concludes the thesis.

2. Literature Survey

Clustering Techniques the goal of gene-based clustering is to group co-expressed genes together. Co-expressed genes indicate co-function and co-regulation. Gene expression data has certain special characteristics and is a challenging research problem. Here, we will first present the challenges of gene-based clustering and then review a series of gene based clustering algorithms. Partitioned Approaches K-means [3] is a typical partition-based clustering algorithm used for clustering gene expression data. It divides the data into pre-defined number of clusters in order to optimize a predefined criterion. The major advantages of it are its simplicity and speed, which allows it to run on large datasets. However, it may not yield the same result with each run of the algorithm. Often, it can be found incapable of handling outliers and is not suitable to detect clusters of arbitrary shapes. A Self Organizing Map (SOM) [6] is more robust than K-means for clustering noisy data. It requires the number of clusters and the grid layout of the neuron map as user input. Specifying the number of clusters in advance is difficult in case of gene expression data. Moreover, partitioning approaches are restricted to data of lower

dimensionality, with inherent well-separated clusters of high density. But, gene expression data sets may be high dimensional and often contain intersecting and embedded clusters. A hierarchical structure can also be built based on SOM, such as Self-Organizing Tree Algorithm (SOTA) [7]. Hierarchical Approaches Hierarchical clustering generates a hierarchy of nested clusters. These algorithms are divided into agglomerative and divisive approaches. UN weighted Pair Group Method with Arithmetic Mean (UPGMA), presented in [9], adopts an agglomerative method to graphically represent the clustered dataset. However, it is not robust in the presence of noise. In [7], the genes are split through a divisive approach, called the Deterministic-Annealing Algorithm (DAA). The Divisive Correlation Clustering Algorithm (DCCA) [2] uses Pearson's Correlation as the similarity measure. All genes in a cluster have highest average correlation with genes in that cluster. Hierarchical clustering not only groups together genes with similar expression patterns, but also provides a natural way to graphically represent the data set allowing a thorough inspection. However, a small change in the data set may greatly change the hierarchical dendrogram structure. Another drawback is its high computational complexity. D. Soft Computing Approaches Fuzzy c-means [2] and Genetic Algorithms (GA) (such as [7] and [2]) have been used effectively in clustering gene expression data. The Fuzzy c-means algorithm requires the number of clusters as an input parameter. The GA based algorithms have been found to detect biologically relevant clusters but are dependent on proper tuning of the input parameters.

3. Methodology

Clustering Gene expression data comprises three important steps: They are Preprocessing, Clustering, Pattern Analysis Preprocessing In the gene expression matrix; different genes have different ranges of intensity values. The intensity values alone may not have significant meaning, but the relative values are more intrinsic. So we first normalize the original gene intensity values into relative values. Data sometimes need to be transformed before being used. For example, attributes may be measured using different scales, such as centimeters and kilograms. In instances where the range of values differ widely from attribute to attribute, these differing attribute scales can dominate the results of the cluster analysis. It is therefore common to normalize the data so that all attributes are on the same scale.

The following are two common approaches to data normalization for each gene vector

$$m'_{i,j} = \frac{m'_{i,j} - \overline{m}_i}{\overline{m}_i}, \overline{\boldsymbol{m}_{\boldsymbol{i}}} = \frac{\sum_{j=1}^{n_s} \boldsymbol{m}_{\boldsymbol{i,j}}}{\boldsymbol{n_s}}$$

$m'_{i,j}$ Denotes the normalized value for gene vector i of sample j, $\boldsymbol{m}_{\boldsymbol{i,j}}$represents the original value for gene i of sample j, ns is the number of samples, $\overline{m_i}$is the mean of the values for gene vector i over all samples.

Clustering Algorithms, 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 a cluster until only singleton clusters of individual objects remain. For agglomerative approaches, different measures of cluster proximity, such as a 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.

1. Initialize U=[u_{ij}] matrix, $U^{(0)}$
2. calculate the probability for each gene $\mathbf{P_{j(\vec{g}_k)}P_{j(\vec{x}_k)}}=\mathbf{exp}\left(-\boldsymbol{\beta}\left|\vec{\mathbf{x}}_\mathbf{k}-\mathbf{c_j}\right|^2\right)/\sum_\mathbf{j}\mathbf{exp}\left(-\boldsymbol{\beta}\left|\vec{\mathbf{x}}_\mathbf{k}-\mathbf{c_j}\right|^2\right)$
3. cluster the gene based on the probability
4. Update cluster centroid $\mathbf{C_j}$ $\mathbf{C_j}=\sum_\mathbf{k}\vec{\mathbf{g}}_\mathbf{k}\mathbf{P_j}(\vec{\mathbf{x}}_\mathbf{k})/\sum_\mathbf{k}\mathbf{P_j}(\vec{\mathbf{x}}_\mathbf{k})$
5. If $\left\|\mathbf{U^{k+1}}-\mathbf{U^k}\right\| < \varepsilon$ then STOP; otherwise return to step 2.

Rough K-Means

The Rough K- means algorithm provides a rough set theory flavor to the conventional K-means algorithm to deal with the uncertainty involved in cluster analysis. In rough k means, each cluster has a lower approximation and an upper approximation. A lower approximation which having most similar objects that belong to the same clusters and cannot belongs to any other cluster. It is the purified and fine cluster.

Input: Dataset of n objects with d features, number of clusters k and values of parameters W_{lower}, W_{upper}, and epsilon.

Output: Lower approximation $U(\underline{K})$ and upper approximation $U(\overline{K})$ of k cluster

Procedure:

Step 1: Randomly assign each data object to exactly one lower approximation $U(\underline{K})$. By property 2, the data object also belongs to upper approximation $U(\overline{K})$ of the same cluster.

Step 2: Compute cluster centroids Cj

$$\text{if } U(\underline{K}) \neq \emptyset \text{ and } U(\overline{K}) - U(\underline{K}) = \emptyset$$

$$C_j = \frac{\sum_{x \in U(\underline{K})} X_j}{|U(\underline{K})|}$$

Else

$$\text{if } \underline{U}(K) \neq \emptyset \text{ and } \overline{U(K)} - \underline{U}(K) \neq \emptyset$$

$$C_j = \frac{\sum_{x \in \underline{U}(K) - \overline{U}(K)} X_j}{\left|\overline{U}(K) - \underline{U}(K)\right|}$$

Else

$$C_j = W_{lower} * \frac{\sum_{x \in \underline{U}(K)} X_j}{\left|\underline{U}(K)\right|} + W_{upper} * \frac{\sum_{x \in \underline{U}(K) - \overline{U}(K)} X_j}{\left|\overline{U}(K) - \underline{U}(K)\right|}$$

Where, j=1,2, ……, k

Step 3: Assign each object to the lower approximation $\underline{U}(K)$or upper approximation $\overline{U}(K)$ of cluster I clusters respectively. For each object vector x, let $d(x,c_j)$ be the distance between itself and the centroid of cluster cj. Let $d(x,c_j)$ be min $_{1<=j<k}$ d(x,cj). The ratio $\frac{d(x,c_i)}{d(x,c_j)}$,$1 \leq i,j \leq k$ is used to determine the membership of x as follows:If $\frac{d(x,c_i)}{d(x,c_j)} \leq$ epsilon, for any pair (i,j), the $x \in U(\overline{K_i})$ and $x \in U(\overline{K_j})$ and x will not be a part of any lower approximation. Otherwise$x \in \underline{U}(K_j)$, such that $d(x, K_j)$ is the minimum for$1 \leq i \leq k$. In addition, $x \in U(K_i)$.

Step 4: Repeat step 2 and 3 until convergence

Fuzzy C-Means Clustering

Fuzzy C-Means (FCM) [12] is a method of clustering which allows one piece of data to belong to two or more clusters. This method (developed by Dunn in 1973 and improved by Bezdek in 1981) is frequently used in pattern recognition.

1. Initialize $U=[u_{ij}]$ matrix, $U^{(0)}$
2. At k-step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$

$$\mathbf{c_j} = \frac{\sum_{i=1}^{N} \mathbf{u_{ij}^m} \cdot \mathbf{x_i}}{\sum_{i=1}^{N} \mathbf{u_{ij}^m}}$$

3. Update $U^{(k)}$, $U^{(k+1)}$

$$\mathbf{u_{ij}} = \frac{\mathbf{1}}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}}$$

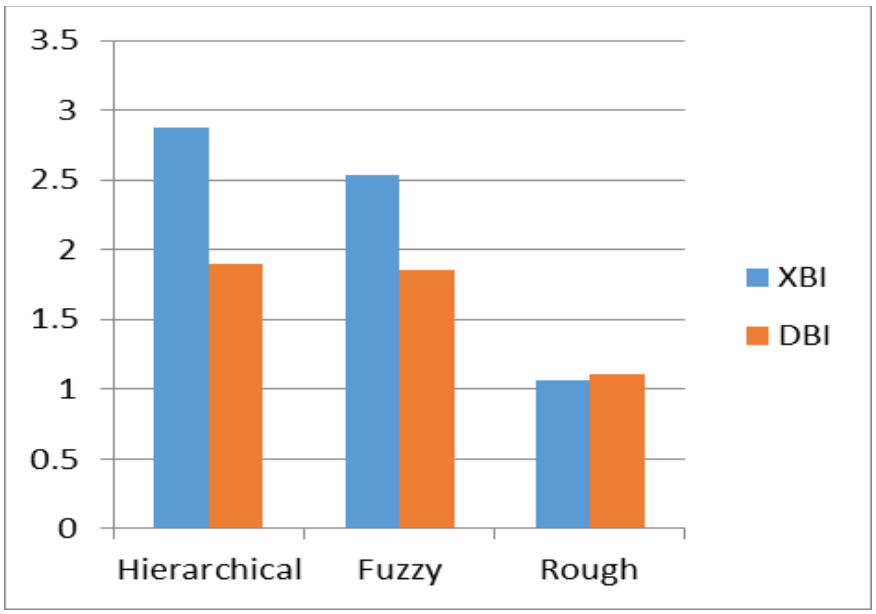
4. If $\left\|\mathbf{U^{k+1}} - \mathbf{U^{k}}\right\| < \varepsilon$ then STOP; otherwise return to step 2.

Pattern Analysis (Validity Measures) The validity indices [5] are used for measuring "goodness" of a clustering result comparing to other ones which were created by other clustering algorithms, or by the same algorithms but using different parameter values. If the number of clusters is not known prior to commencing an algorithm, the clustering validity index may also be used to find the optimal number of clusters [2]. Xie-Beni Validity Index is discussed in [9] which measure the compactness and separation of clusters.

4. Experimental Analysis

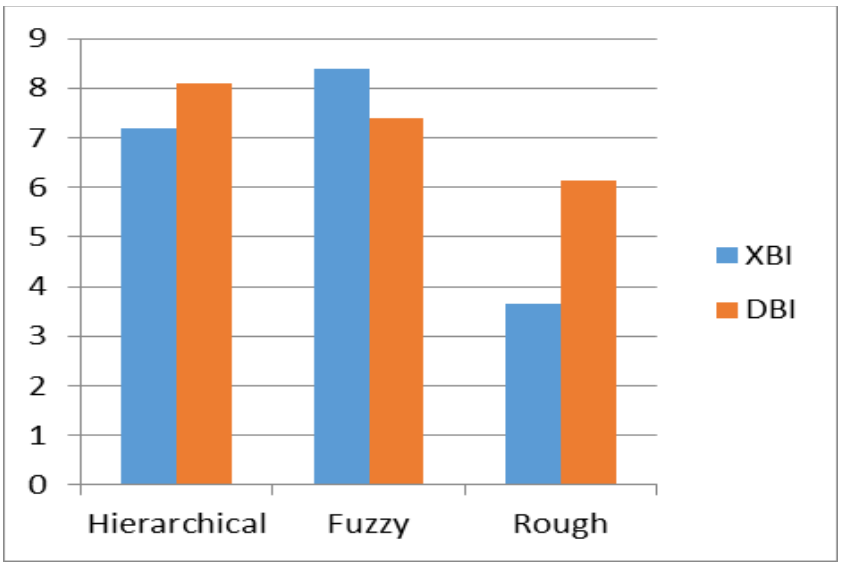
Cluster analysis, is an important tool in gene expression data analysis. For experimentation, we used a set of gene expression data that contains a series of gene expression measurements of the transcript (mRNA) levels of gene expression. In clustering gene expression data, the genes were treated as objects and the samples are treated as attributes. To evaluate the performance of the various clustering Rough K-Means, Fuzzy C-Means and Hierarchical Clustering these approaches were implemented in MATLAB. This dataset is publicly available in UCI Repository. The XBI and DB index were used as validation measures for comparative analysis. The dataset is explained below. Comparative analysis: Clustering approaches for Yeast Gene Expression data The comparative results based on XBI and DBI validity measure for all the depicted gene expression data clustering algorithms is shown in Table 4.1. It can be observed from the Table 4.1 and Figure 4.1 that Rough K-Means exhibits best performance than Fuzzy- C-Means and Hierarchical Clustering since XBI and DBI value of Rough K-Means is lower than that of Fuzzy- C-Means and Hierarchical Clustering.

Clustering Algorithms	**Yeast Gene Expression Data**	
	XBI	**DBI**
hierarchical clustering	2.8765	1.8956
fuzzy- c-means	2.5400	1.8510
rough k-means	1.0668	1.1075



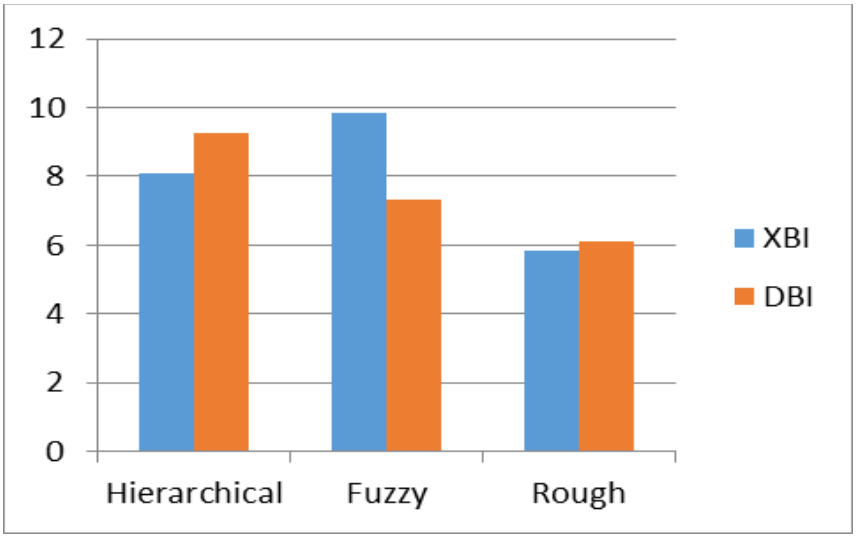
Comparative analysis: Clustering approaches for Colon Gene Expression data the comparative results based on XBI validity measure for all the depicted gene expression data clustering algorithms are shown Table 4.2. It can be observed from the Table 4.2 and Figure 4.2 that Rough K-Means exhibits best performance than Fuzzy- C-Means and Hierarchical Clustering since XBI value of Rough K-Means is lower than that of Fuzzy- C-Means and Hierarchical Clustering.

Clustering algorithms	Colon Gene Expression Data	
	XBI	DBI
Hierarchical clustering	7.1833	8.1102
Fuzzy- c-means	8.3783	7.3850
Rough k-means	3.6373	6.1305



Comparative analysis: Clustering approaches for Leukemia Gene Expression data the comparative results based on XBI validity measure for all the depicted gene expression data clustering algorithms are shown Table 4.3. It can be observed from the Table 4.3 and Figure 4.3 that Rough K-Means exhibits best performance than Fuzzy- C-Means and Hierarchical Clustering since XBI value of Rough K-Means is lower than that of Fuzzy- C-Means and Hierarchical Clustering.

Clustering Algorithms	Leukemia Gene Expression Data	
	XBI	DBI
hierarchical clustering	8.0760	9.2702
fuzzy- c-means	9.8398	7.3081
rough k-means	5.8322	6.1026



5. Conclusion

In this dissertation three clustering techniques such as Hierarchical, Rough K-Means, and Fuzzy C Means studied. To validate the performance of these clustering techniques, Xie-Beni index and Davis Boulin index measures of use in this study. From the empirical study, it has been observed that rough k means clustering techniques perform good, other two techniques (Hierarchical clustering, Fuzzy C Means clustering).

References

[1]. Ao. S.I, Kevin Y.P, Michael Ng, David Cheung, Fong .P, Ian Melhado and Sham. C., "CLUSTAG: hierarchical clustering and graph methods for selecting SNPs", Oxford University press, 21(5), 1735-1736, 2005.
[2]. Ben-Dor A., Shamir R. and Yakhini Z. "Clustering gene expression patterns". Journal of Computational Biology, vo1.6 (3/4), pp. 281–297, June 1999.
[3]. Daxin Jiang, Chun Tang, Aidong Zhang, "Cluster Analysis for Gene Expression Data: A Survey," IEEE Transactions on Knowledge and Data Engineering, vol. 16, No.11, November 2004.
[4]. D. Dembele and P. Kastner, "Fuzzy C-Means Method For Clustering Microarray Data", BioInformatics, Vol. 19, No.8, PP 973-980, 2003
[5]. Du. Z, Lin. F, "A Novel Parallelization Approach For Hierarchical Clustering". Parallel Computing, 31, 523-527,2005
[6]. Luo F, Tang K, Khan L., "Hierarchical Clustering of Gene Expression Data". Proceedings of the Third IEEE Symposium on BioInformatics and BioEngineering , 2003.
[7]. Pawlak. Z. "Rough Sets", International Journal of Computer and Information Sciences, 341-356,1982.
[8]. Pawan Lingras, Chad West. "Interval set Clustering of Web users with Rough k-Means", submitted to the Journal of Intelligent Information System in 2002.
[9]. [Yeung K.Y, Haynor D.R, Ruzzo W.L. "Validating Clustering For Gene Expression Data". Bioinformatics. 2001.