International Journal of Scientific and Technical Advancements

ISSN: 2454-1532

A Review of Clustering Algorithms for Gene Expression Data

Ritika Manhas[1], Ayushman Koul[2], Bhawna Sharma[3], Sheetal Gandotra[4]
[1]Department of Microbiology, Panjab University, Chandigarh, India-160014
[2,3,4]Department of Computer Engineering, GCET Jammu, J&K, India-181121
[1,#]ritikamanhas05@gmail.com, [2]ayushmankoul4570@gmail.com, [3]bhawnash@gmail.com, [4]sheetalgandotra74@gmail.com

Abstract- In biological studies, the DNA microarray technology has provided us with various ways to understand the gene expression of different species and study the effect of environment on them. It is used to identify the group of genes with similar expression patterns that are extensively used to produce proteins together. It is a tedious task to analyze the huge volumes of data obtained from genome sequencing and to procure meaningful information from it. Therefore, the first major step is clustering of genes which help in identifying similar gene expression and to understand the function of each gene. In this paper we have discussed about DNA microarray technology and compared the various clustering techniques which are used in cluster analysis of gene expression data.

Keywords - DNA Microarray Technology; Clustering Algorithms; Gene Expression Data.

I. Introduction

For genomic research in the field of biology a number of traditional approaches have been used to analyse and collect data obtained to study gene expression. Analysis of gene expression data allows us to identify difference in levels of gene expression and their profile. It is of utmost importance to study gene expression data as it provides an insight into the biological processes occurring inside the cells in relation to its environment and about different genes working to produce a similar effect ultimately. A slight variation in the expression level could indicate a possible stress condition being experienced by the organism or cell and could even indicate a mutation in the genomes.

II. DNA Microarray Technology

The DNA Microarray technology is amongst the leading techniques used to study gene expression. The technique uses DNA probes of known sequences. These probes are spotted on the microarray chips in the form of spots which has led to the monitoring of thousands of genes simultaneously. The mRNA of the sample is reverse-transcribed into cDNA which is then labelled. It compares the gene expression levels of the test sample and the control sample by hybridization of cDNA with the probes. The intensity signals obtained after the hybridization reaction under different conditions like time or biological processes through scanning form the intensity matrix. The data thus obtained is called the gene expression data. The data obtained is compared to the intensity levels of the test and the control, increase in the expression levels indicates the co-expression of genes, increase and decrease in the intensity levels overtime could indicate other relations amongst the genes under study. To filter out meaningful data out of the intensity matrix various clustering algorithms are used to cluster together genes that are co-regulated, co-expressed and co-function. Clustering the data also gives us insight into the mechanism of transcription regulation. These algorithms group the gene expression data into several clusters based on the level of similarity amongst the expression levels of different genes. Those in separate clusters are more dissimilar than the ones in the same cluster. The co-expressed genes are grouped in the same clusters indicating co-regulation and co-function.

III. Clustering

Clustering is the type of unsupervised classification and it is the process of making disjoint sets called clusters, of those data objects which have higher similarity to each other and the similarity between inter cluster data objects is very minimal. Clustering is broadly used in many applications like data analysis, image processing, gene clustering, market research and pattern recognition.

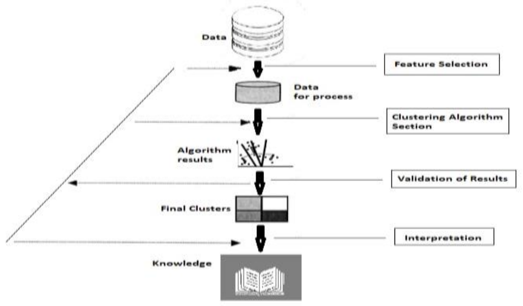

Fig. 1. Clustering procedure

.

Ritika Manhas, Ayushman Koul, Bhawna Sharma, and Sheetal Gandotra, "A Review on Clustering Algorithms for Gene Expression Data," *International Journal of Scientific and Technical Advancements*, Volume 5, Issue 1, pp. 137-140, 2019.

IV. TYPES OF CLUSTERING METHODS

In this section, we describe the different clustering methods which are used for gene based clustering whose aim is to identify group of co-expressed genes.

A. K-means Algorithm

K-means algorithm is the most popular method used for clustering. It is a typical partition based algorithm using centroid approach where a cluster is represented by a gravity centre. It is used to classify data objects into predefined number of clusters. It initially takes K known cluster centres and minimize the distance between cluster centres of given clusters and for measuring distance between data objects Euclidean distance between two data points X= ($x_1, x_2, \ldots, x_m$) and Y= (y_1, y_2,……y_m) can be calculated as follows:

$$D(X, Y) = \sqrt[2]{(x1 - y1)2 + (x2 - y2)2 + \cdots + (xm - ym)2}$$

The K-means algorithm aims to minimize the sum of squared distances between all points and cluster centre [1].This algorithm consists of following steps as explained below:

```
Algorithm: K-mean clustering algorithm [2]
Require: D = {$d_1$, $d_2$, …….., $d_n$) // set of n data points
        K = No. of desired clusters
Ensure: A set of K clusters.
Steps:
1. Arbitarily choose K data points from D as initial centroids;
2. Repeat
      Assign each point $d_i$ to the cluster which has closest centroid;
      Calculate the new mean for each cluster;
      Until convergence criteria is met.
```

K-means algorithm is fast and performs well when compared to the new clustering algorithm but has a number of limitations. First, the number of gene clusters in gene expression data set is usually unknown and to detect the optimal number of clusters, users usually run the algorithms with different values of K and for a large expression data set which contains thousands of genes, this extensive parameter fine-tuning process may not be practical [3]. Second, gene expression data typically contain a huge amount of noise; however, the K-means algorithm forces each gene into a cluster, which may cause the algorithm to be sensitive to noise [3]. The various application of K means algorithm for clustering gene expression data is also discussed in literature [4, 5, 6, 7, 8].

B. Hierarchical Clustering

Hierarchical clustering is the hierarchical decomposition of the data based on group similarities. Hierarchical clustering is extensively used in gene expression data analysis. This type of clustering is divided into two methods, agglomerative and divisive. **Agglomerative** clustering uses a bottom-up approach, in which each data point initiates its own cluster. These clusters are then joined by taking the two most similar clusters together and merging them. **Divisive** clustering uses a *top-down* approach, wherein all data points start in the same cluster then using a parametric clustering algorithm like K-Means divide the cluster into two clusters. For each cluster, we further divide it down to two clusters until we get the desired number of clusters.

Clustering Using Representatives (CURE) employs a new algorithm model which uses both centroid based and all points approach. A constant number of scattered points in a cluster act as representatives. The clusters with closest pair of representative points are the clusters that are merged together at each step. It uses random sampling and partitioning for reducing the large amount of input data set [9].

CHAMELEON algorithm works on the principle to generate a sparse graph in which nodes represent data items, and weighted edges represent similarities among the data items. It uses a graph partitioning algorithm to cluster the data items into a large number of relatively small sub-clusters and then uses an agglomerative hierarchical clustering algorithm to find the genuine clusters by repeatedly combining together these sub-clusters [10].

ROCK is also an agglomerative hierarchical clustering algorithm. It uses links to measure the similarity between a pair of data points in a cluster. Then it merges the data points of a cluster [11].

BIRCH (Balanced Iterative and Clustering using Hierarchies) is suitable for large database [12].

EISEN's method is much favored by many biologists and has become the most widely-used tool in gene expression data analysis [13].

However, the conventional agglomerative approach suffers from a lack of robustness [3], i.e., a small perturbation of the data set may greatly change the structure of the hierarchical dendrogram. Another drawback of the hierarchical approach is its high computational complexity [3].

C. Model-based clustering.

Clustering algorithms can also be developed based on probability models. In the family of model based clustering algorithms, one uses certain models for clusters and tries to optimize the fit between the data and the models. The **Expectation Maximization** (EM) algorithm [2] determines good values for its parameters iteratively. It is able to handle different shapes of cluster, and lots of iteration are required that makes this algorithm costly. The literature of the model based clustering approaches for gene expression data is discussed in [14, 15].

Self Organizing Map (SOM) by Teuvo Kohonen provides a data visualization technique which helps to understand high dimensional data by reducing the dimensions of data to a map. SOM also represents clustering concept by grouping similar data together. Therefore it can be said that SOM reduces data dimensions and displays similarities among data. Self-Organizing Map (SOM) [16] is a technique easy to implement, fast and scalable for large gene expression dataset. It is based on a single layered neural network. It is represented in a two dimensional m*n grid where data points are taken as input and output neuron. Then neurons are represented as simple neighborhood structure. A reference number is attached with each neuron, and each data point is mapped to the nearest reference vector. Each data point is act as training sample which leads the movement of reference vectors towards the deeper input space so that it will is distributed to input dataset. Clusters are identified by mapping all data points to the output neuron after the completion of training process. The Self organizing map clustering algorithm starts with the initialization of the reference vector followed by randomly selection of data points. Then nearest reference vector to the current data point is determined and finally reference vector and neighboring reference vectors are updated. SOM is very efficient method used for gene expression data clustering and it is also discussed in literature [2, 16].

TABLE I. Clustering Algorithms computational complexity

S. No.	**Clustering Algorithm Computational Complexity**		
	Clustering Algorithm	**Complexity**	**Capable of handling high dimensional data**

1	K-means	O(NKd) (time)
		O(N+K) (space)
			No

2	Hierarchial Clustering	$O(N^2)$ (time) $O(N^2)$ (space)	No
3	ROCK	$O(n^{3)}$	No
4	CHEMLEON	$O(m^2 logm)$	No
5	BIRCH	O(N) (time)	Yes
6	CURE	$O(N^2_{sample} logN_{sample})$ (time) $O(N_{sample})$ (space)	Yes
7	SOM	$O(N^2_{sample})$(time)	Yes

V. CONCLUSION

In this review, we discussed three types of traditional methods of clustering i.e. K means, Hierarchical Clustering and Model Based Clustering. A good clustering algorithm should be able to generate arbitrary shapes of clusters, able to handle large volumes of input data, should not be affected by the order of input data, able to handle noise in input data and should be able to produce desired and correct results with higher accuracy in less time. In the field of biology clustering algorithms have made it possible to identify those genes which perform the same function and also helped to study the effect of environment on genes and predict the diseases before they actually show symptoms. Examples of such implementations of clustering algorithms include the use of hierarchical clustering on DNA microarray data by Alizadeh et al, in 2000, which led to the discovery of three distinct subtypes of the diffuse large B-cell lymphoma (DLBCL) [17]. The K-Means algorithm was found to be efficient for clustering the lung cancer dataset with Attribute Relation File Format (ARFF) [18] and K-means algorithm was implemented on images from the Mammography Image Analysis Society (MIAS) to determine the stage of malignant breast cancer [19]. Traditional clustering algorithms despite of proving useful still have scope of improvement to curtail the significant drawbacks that they pose for gene expression data analysis. To tackle these issues new algorithms have been implemented recently to improve the drawbacks of traditional approaches in order to get better accuracy. The Enhanced Automatic Generation of Merge Factor for ISODATA (EAGMFI) Clustering Microarray Data based on K-Means and AGMFI clustering algorithms was implemented in [1] to overcome random selection of initial seed point of desired clusters. Similarly as discussed methods like BIRCH and CURE based on hierarchical approaches perform better when applied to large databases whereas model based approaches are costly compared to these as they require lot of iterations.

REFERENCES

[1] T.Chandrasekhar , K.Thangavel and E.Elayaraja, “Performance Analysis of Enhanced Clustering Algorithm for Gene Expression Data,” *International Journal of Computer Science Issues*, Vol. 8, Issue 6, No 3, 2011

[2] Dempster AP, Laird NM, Rubin DB, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the royal statistical society. Series B (methodological).*pages 1-38, 1977.

[3] Daxin Jiang, Chun Tang, Aidong zhang, “Cluster analysis for gene expression data: a survey*”, IEEE Transactions on knowledge and data engineering*, vol. 16, issue 11,pages 1370-1386, 2004

[4] J. H. Do and D. -K. Choi, "Clustering Approaches to Identifying Gene Expression Patterns from DNA Microarray Data," *Molecular Cells*, vol. 25, no. 2, 2007.

[5] Do JH, Choi D. “Clustering approaches to identifying gene expression patterns from DNA microarray data”, *Molecules and cells*, pages 242-279, 2008

[6] Thalamuthu A, Mukhopadhyay I, Zheng X, Tseng GC, “Evaluation and comparison of gene clustering methods in microarray analysis”, *Bioinformatics*, pages 2405-12, 2006

[7] Costa IG, de Carvalho FD, de Souto MC, “Comparative analysis of clustering methods for gene expression time course data”, *Genetics and Molecular Biology*, pages 623-631, 2004

[8] Borg A, Lavesson N, Boeva V, “Comparison of Clustering Approaches for Gene Expression Data”, *In-SCAI*, pages 55-64, 2013

[9] Guha S, Rastogi R, Shim K, “CURE: an efficient clustering algorithm for large databases”, *In ACM SIGMOD*, vol. 27, pages 73-84,1998

[10] Karypis G, Han EH, Kumar V, “Chameleon: Hierarchical clustering using dynamic modeling. Computer”, pages 68-75,1999

[11] Guha S, Rastogi R, Shim K. “ROCK: A robust clustering algorithm for categorical attributes” *in Data Engineering, Proceedings, 15th International Conference*, pp. 512-521, 1999

[12] Zhang T, Ramakrishnan R, Livny M., “BIRCH: an efficient data clustering method for very large data-bases*” in ACM Sigmod* , vol. 25, No. 2, pp. 103-114, 1996

[13] Eisen, Michael B., Spellman, Paul T., Brown, Patrick O. and Botstein, David, “Cluster analysis and display of genome-wide expression patterns”. *Proc. Natl. Acad. Sci. USA*, pages 14863–14868, 1998.

[14] Datta S, Datta S., “Comparisons and validation of statistical clustering techniques for microarray gene expression data”, *Bioinformatics*, pages 459-66, 2003

[15] Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR., “Interpreting patterns of gene expression with

self-organizing maps: methods and application to hematopoietic differentiation”, *Proceedings of the National Academy of Sciences*, pages 2907-12, 1999

[16] Tomida S, Hanai T, Honda H, Kobayashi T., “Analysis of expression profile using fuzzy adaptive resonance theory”, *Bioinformatics*, pages 1073-83, 2002

[17] Alizadeh AA, Eisen MB, Davis RE, et al. “Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling”,*Nature*., pages 503-11, 2000

[18] Dharmarajan A, Velmurugan T, “Lung cancer data analysis by k-means and farthest first clustering algorithms”, *Indian J Sci Techno*,. 2015

[19] Karmilasari SW, Hermita M, Agustiyani NP, Hanum Y, Lussiana ETP, “Sample K-Means Clustering Method for Determining the Stage of Breast Cancer Malignancy Based on Cancer Size on Mammogram Image Basis”, *IJACSA) Int J Adv Comput Sci Appl*. , pages 86-90, 2014