International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

Review on Performance Analysis of Gene Expression data using Fuzzy Clustering Techniques

K.Umamaheswari[1], Dr.R.P.Singh[2,] Dr.K.Vengatesan[3]

[1]*(Research Scholar, Sri Satya Sai University of Technology and Medical Sciences, India)*

[2]*(Research Guide, Sri Satya Sai University of Technology and Medical Sciences, India)*
[3]*(Associate Professor, Sanjivani College of Engineering, India)*

ABSTRACT

The analysis of gene expression data using clustering is traditional techniques and lot of research work is undergoing. The clustering is an important task in data mining, the functional analysis of gene clustering investigation can be performed by using various algorithms. This paper discussed about various fuzzy clustering methods used for gene functions, cellular process, gene regulation and subtypes of cells. The Enhanced Robust rough Fuzzy C Mean increases the probability membership of the clusters and also handles the overlapping gene clusters effectively. It is also useful in dealing with probabilistic lower approximation and possibility lower approximation. The proposed methods are used to identify the strong group of Co expressed genes and produce the best result.

Keywords: Gene Expression Data, Clustering, Micro Array, Fuzzy Clustering.

I.INTRODUCTION

Rapid retrieval of significant information from the databases has always been an important issue. Different techniques have been developed for this purpose; one of them is Data Clustering. Data clustering is methods by which clusters were made that are one way or another similar in characteristics. Clustering in computer science means unsupervised classification of data objects into different groups. It can also be referred to as partitioning of a data set into different subsets. Each data object in the subset ideally shares some common character. This section discusses various works carried out by existing researchers on data mining techniques, gene expression data, microarray, statistical methods used for measure the similarity, the advantages and limitations of existing clustering techniques, different data types and data repositories which are used for mining knowledge. The cluster is group of object one with another based on the similarity between the objects. The correlation is calculated from the micro array gene expression data to form the cluster. The performance of each work is compared with existing work.

II.RELATED WORK

The clustering is one of the major techniques used in gene expression matrix. In which clustering is categorized into three ways such as gene based clustering, in which genes are treated as object. The principle of gene-based clustering is to grouping collectively co expressed genes which point out co function and co regulation, second one is sample based clustering, in which samples are served as data objects and third one is subspace clustering, in which genes and samples treated either objects or features. DNA microarray technology has now made it possible to simultaneously monitor the expression levels of thousands of genes during important biological processes and across collections of related samples. Elucidating the patterns hidden in gene expression data offers a tremendous opportunity for an enhanced understanding of functional genomics. However, the large number of genes and the complexity of biological networks greatly increase the challenges of comprehending and interpreting the resulting mass of data, which often consists of millions of measurements. A first step toward addressing this challenge is the use of clustering techniques, which is essential in the data mining process to reveal natural structures and identify interesting patterns in the underlying data. Cluster analysis seeks to partition a given data set into groups based on specified features so that the data points within a group are more similar to each other than the points in different groups. The Figure 1 show the steps of cluster analysis are the feature selection, cluster algorithm, cluster validation, interpretation of results (Pham et al 2006, John N et al 2001).In which data set is collect from different sources, then apply preprocessing techniques to remove the noise and error data from the huge dataset, finally apply various clustering algorithms and result are report.

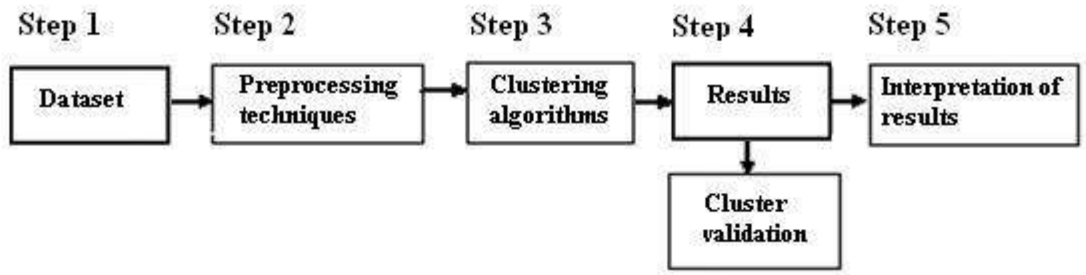

Figure 1 Framework of cluster analysis

III.MAJOR CHALLENGES OF GENE CLUSTERING

The extraction of clusters from gene expression data is one of the difficult issues, the biological field, gene-based clustering present a number of new challenges and is still an open problem, that are discussed below

❖ First the finding the structure of gene expression data and gain some initial insights regarding data distribution in clustering analysis is difficult task.

❖ Second, micro array experiments are complex procedure, which contains huge amount of noise data, in which applying clustering algorithm is also major problem.

❖ Third, our observed study has established that gene expression data are often “extremely coupled” [37], and clusters may be greatly intersected with each other or even implanted one in another [36]. Therefore, algorithms for gene-based clustering should be able to efficiently handle these circumstances.

❖ Finally, users of microarray data may not only be concerned in the clusters of genes, but also be attracted in the association between the clusters and the affiliation between the genes within the same cluster. A clustering algorithm, which cannot only partition the data set but also provides some graphical representation of the cluster structure, would be more favored by the biologists.

One of the goals of microarray data analysis is to cluster genes or samples with similar expression profiles together, to make meaningful biological inference about the set of genes or samples. Clustering is one of the unsupervised approaches to classify data into groups of genes or samples with similar patterns that are characteristic of the group. Cluster analysis groups data objects based only on information found in the data that describes the objects and their relationships (Reynaldo Gil-García et al 2006). The goal is that the objects within a group be similar to one another and different from the objects in other groups. The greater the similarity within a group and the greater the difference between groups the better or more distinct the resulting clusters.

IV.HIERARCHICAL CLUSTERING

Hierarchical clustering is the most commonly used clustering paradigm in microarray data analysis for identifying genes with similar profiles and possibly with similar functions. The goal of hierarchical clustering is to obtain the definitive clustering that characterizes a set of patterns in the context of a given distance metric (Jinwook Seo and Shneiderman 2002, Grzegorz Boratyn et al 2006). Hierarchical clustering starts by calculating the distance matrix for all patterns in data set. The two closest patterns are merged and distance matrix is calculated again but using the new cluster instead of the two merged patterns (Yuri Stekh et al 2006).This approach produces a hierarchy of clusters by pairing genes that are most similar. The result of hierarchical clustering is usually depicted by a binary tree (or) dendrogram (Bernard Chen et al 2005, Cheng-Che Wu et al 2006). The application of hierarchical clustering are Image processing, Natural language processing, Speech and pattern recognition, Classifying DNA sequences.

V.PARTITION BASED CLUSTERING

Most partition based clustering algorithms assume a priori the number of clusters, and partition the data set accordingly. There can be many partitions of a given data set, but there will be only a few which identify the clustering in the data set (Suresh et al 2009). To arrive at a correct partition, an objective function can be formulated that measures how good a partition with respect to the data set is (Khaled Alsabti et al 1998). If a given partition minimizes the objective function, one can assume that the optimal partition has been found. Most objective function-based algorithms use cluster prototypes to facilitate the evaluation of a given partition (Fahim et al 2006) .Each prototype is assumed to be a typical representative of the group of points in that cluster. In an ideal case, each prototype will take the general shape of its cluster (Kiri Wagsta et al 2001, Amir ben-dor et al 1999). Since objective functions are typically non-linear, the optimal partition will usually have to be searched for algorithmically.

VI.FUZZY CLUSTERING

The algorithms for clustering approach under the either hard (or) crisp clustering. A hard clustering allocates each pattern to a single cluster. While crisp clustering approaches can accurately identify distinct expression patterns by grouping genes with similar expression patterns, they are unable to identify genes whose expression levels are similar to multiple distinct groups of genes (Mingrui Zhang et al 2008, Seo Young Kim and Tai Myong Choi 2005). Crisp clustering methods are likely to yield inaccurate clusters leading to incorrect conclusions when analyzing large gene expression datasets collected under different conditions (Jain et al 1999, Fu and Medico E 2003). These problems in hard clustering are avoided with fuzzy clustering approach. When we “fuzzy cluster” a data set we allow for data points to belong with varying degrees to more than one cluster.

If data objects are distributed in well-separated groups, then a crisp classification of the objects into disjoint clusters seems like an ideal approach. However, in most cases, objects in a data set cannot be partitioned into well separated clusters, and there will be certain arbitrariness in assigning an object to a particular cluster. Consider an object that lies near the boundary of two

Clusters, but is slightly closer to one of them (Valente De Oliveira and Pedrycz 2007).

VII.FUZZY C-MEAN CLUSTERING

The fuzzy c-mean algorithm uses the prototype based clustering algorithm to find the association between the data. It will assign the membership to each gene how it related to another gene, sometime a gene may be related to more than one cluster, because of gene are overlapped with one another, based on the certain boundaries the gene group is formed[19]. The fuzzy clustering methods used to identify the information regarding overlapped cluster and cellular pathways of each gene. If the membership value of the fuzzy c-means is low means, then there is a noisy environment, weakness in gene expression data. The possibilistic c-Mean algorithm produces the best coincidence clusters. The applications of fuzzy C-Mean are Image analysis, Pattern recognition, Bioinformatics, and Fraud detection

VIII.ROBUST ROUGH FUZZY C-MEAN ALGORITHM

It is an efficient method to handle the cluster in both possibilistic and probabilistic fuzzy sets, and upper and lower approximation of rough sets into C-Mean algorithm, while integrating both techniques, it will handle the overlapping cluster in noisy environments, also deal with vagueness, incompleteness uncertainty in cluster definition. A membership function (MF) is a arc that defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1. Fuzzy Logic consists of type 1 and Type 2 fuzzy. A type-2 fuzzy set contains the grades of membership that are themselves fuzzy. Type-1 Fuzzy Logic is unable to handle rule uncertainties. Type-2 Fuzzy Logic can handle rule uncertainties effectively and efficiently [8]. A Type-2 relationship rating can be any separation in the primary membership. For each primary membership there exists a secondary membership that defines the potential for the primary membership.

IX.ENHANCED ROBUST ROUGH FUZZY C-MEAN

It is an efficient method to handle the cluster in both possibilistic and probabilistic fuzzy sets, and upper and lower approximation of rough sets into C-Mean algorithm, while integrating both techniques, it will handle the overlapping cluster in noisy environments, also deal with vagueness, incompleteness uncertainty in cluster definition. The objective function is let Y= $\{y_1.....y_j.....y_n\}$ be a set of n objects and C= $\{c_1....c_i.....c_c\}$ be the set of centroid, where $y_j \varepsilon R^m$ and $v_j \varepsilon R^m$. Each of the clusters β_i is represented by a cluster center υ_i which follows both lower and upper approximation of the cluster β_i. The minimization function of the proposed C cluster is written as

$$J=\begin{cases} wA_1+(1-w)\beta_1 \quad if \;\; A(\beta\;)\neq\theta, B(\beta_i)\neq\theta \\ \qquad A_1 \qquad if \;\; A(\beta\;)\neq\theta, B(\beta_i)\neq\theta \\ \qquad B_1 \qquad if \;\; A(\beta\;)\neq\theta, B(\beta_i)\neq\theta \end{cases} \tag{1}$$

In which $A(\beta\;)$ is the lower approximation and $B(\beta_i)$ is the probability boundary where A_1, B_1 are represented as

$$A_1=\sum_{i=1}^{c}\sum_{x_j\varepsilon\; A(B_i)}(v_{ij})^{m_2}\left\|x_j-v_i\right\|^2+\sum_{i=1}^{c}\eta_i\sum_{x_j\varepsilon\; A(B_i)}(1-v)^{m_2} \tag{2}$$

$$B_1=\sum_{i=1}^{c}\sum_{x_j\varepsilon\; B(B_i)}(\mu_{ij})^{m_1}\left\|x_j-v_i\right\|^2 \tag{3}$$

The relative important of lower boundary is represented by the parameter w and (1-w), while 1<=m$_1$<∞ and 1<=m$_2$<∞ are the probability functions. The centroids of the cluster should be independent to the lower approximation along with memberships of the objects. The membership function between the object is represented as following equation

$$\mu_{ij}=\left[\sum_{k=1}^{c}\left(\frac{\left\|x_j-v_i\right\|^2}{\left\|x_j-v_k\right\|^2}\right)^{\frac{1}{m_1-1}}\right]^{-1} \tag{4}$$

$$v_{ij} = \left[1 + \left\{\frac{\|x_j - v_i\|^2}{\eta_i}\right\}^{\frac{1}{m_2-1}}\right]^{-1} \qquad \textbf{(5)}$$

where the scale parameter $\eta_i = K.\frac{\sum_{j=1}^{n}(v_{ij})^{m_2}\|x_j - v_i\|^2}{\sum_{j=1}^{n}(v_{ij})^{m_2}}$ which represents the size of the cluster B .

The centroids of the cluster are calculated based on the weighting average of the probabilistic boundary and possibility lower approximation. The performance of the proposed algorithm enhanced robust rough fuzzy c-mean algorithms are compared with the fuzzy c-means (FCM), Rough-Fuzzy, C-Mean(RFCM), Hard C-Mean(HCM), Cluster Identification via Connectivity Kernel (CLICK), Self Organizing Map (SOM) and robust rough fuzzy c-mean algorithm with several microarray gene expression data set.

X.CONCLUSION

In this paper discussed about various existing clustering algorithms used in micro array analysis, which is a combination of different algorithms such as rough set, c-mean algorithm and possibility, probabilistic memberships of fuzzy sets, which produced the maximum result for rough and fuzzy sets, the effects of our algorithm is compared to other algorithms. The Enhanced Fuzzy Clustering is applied to gene expression data that produce significantly better result, irrespective of the quantitative indices and microarray data sets using a biological process, molecular function and cellular components for lower approximations. It extracts the highly similar genes from lower approximation, upper approximation, any shaped gene clusters and handles efficiently overlapped gene cluster.

REFERENCES

[1.] P. Maji and C. Das, Protein Functional Sites Prediction Using Modified Bio-Basis Function and Quantitative Indices, *IEEE Transactions on NanoBioscience*, 9(4), pp. 250--257, December 2010.

[2.] P. Maji and S. K. Pal, Feature Selection Using *f*-Information Measures in Fuzzy Approximation Spaces, *IEEE Transactions on Knowledge and Data Engineering*, 22(6), pp. 854--867, June 2010.

[3.] P. Maji, *f*-Information Measures for Efficient Selection of Discriminative Genes from Microarray Data, *IEEE Transactions on Biomedical Engineering*, 56(4), pp. 1063--1069, April 2009.

[4.] P. Maji and S. K. Pal, Rough Set Based Generalized Fuzzy *C*-Means Algorithm and Quantitative Indices, *IEEE Transactions on System, Man and Cybernetics, Part B, Cybernetics*, 37(6), pp. 1529--1540, December 2007.

[5.] P. Maji, M. K. Kundu, and B. Chanda, Second Order Fuzzy Measure and Weighted Co-Occurrence Matrix for Segmentation of Brain MR Images, *Fundamenta Informaticae*, 88(1-2), pp. 161--176, 2008.

[6.] H.Causton,J.Quackenbush, and Brazama, Microarray Gene Expression Data Analysis, A Beginners guide, Wiley-Blackwell,2003.

[7.] E. Domany, Cluster Analysis of gene Expression data", J. Statistical Physics,vol.110,nos.3-6, pp.1117-1139,2003

[8.] D.Jiang, C.Tang, and A.Zhang, "Cluster Analysis of gene Expression Data: A survey,IEEE Trans. Kwnowledge anda data Eng,vol.16,no.11,pp.1370-1386,Nov.2004.

[9.] A.Brazma and J.Vilo,"Minireview:Gene Expression Data Analysis",Federation of European Biochemical Societies letters,vol.480,no.1,pp.17-24,2000.

[10.] L.Heyer, S.Kruglyak,and S.Yooseph,"Exploring Expression Data:Identification and analysis of Coexpressed genes", Genome Research,vol.0,no.11,pp. 1106-1115,1999.

[11.] P.Tamayo, D.Slonim,J.Mesirov,Q.Zhu,S.Kitareewan,E.Dmitrovsky,E.S.Lander,and T.R.Golub, "Interpreting Patterns of Gene Expression with Self Organizing Maps: Methods and Application to Hematopoietic Differentiation," Proc Nat'l Academy of science USA,vol.96,no.6,pp.2007-2912,199.

[12.] E.Hartuv and R.Shamir, "A Clustering Algorithm based on Graph Expression Patterns",J. Computational Biology,vol.6,nos ¾,pp.281-297,1999.

[13.] R.Sharmir, and R.Sharan,"Click :A Clustering Algorithm for Gene Expression Analysis",Proc. Eight Int'lConf. Intelligent System for Molecular Biology,2000.

[14.] D.Ghosh and A.M.Chinnaiyan,"Mixture Modelling of gene Expression Data from Microarray Experiments",Bioinformatics,vol.18,no.2,pp.275-286,2002.

[15.] D.Jiang,J.Pei and A.Zhang, "DHC:A Density Based Hierarchical Clustering Methods for Time Series Gene Expression Data," Proc,IEEE Third Int'l Symp, Bioinformatics and BioEng,pp.393-400,2003,

[16.] P.J.Woolf and Y.Wang,"Fuzzy logic approach to Analyzing Gene Expression Data",Physiological Genomics,vol.3,pp.9-15,2000.

[17.] R.Krishnapuram and J.M.Keller,"A Possibilistic Approach to Clustering ",IEEE Trans, Fuzzy System, Vol. 1, no.2, pp.98-110,1993.

[18.] N.Belacel,M.Cuperlovic-Culf,M.Laflamme,and R.Quellette,"Fuzzy J-Means and VNS methods for clustering Gene from Microarray Data," Bioinformatics, Vol. 20,no.11,pp.1670-1701,2004.

[19.] Vengatesan K., and S. Selvarajan: Improved T-Cluster based scheme for combination gene scale expression data. International Conference on Radar, Communication and Computing (ICRCC), pp. 131-136. IEEE (2012).

[20.] Kalaivanan M., and K. Vengatesan.: Recommendation system based on statistical analysis of ranking from user. International Conferenceon Information Communication and Embedded Systems (ICICES), pp.479-484, IEEE, (2013).

[21.] K. Vengatesan, S. Selvarajan: The performance Analysis of Microarray Data using Occurrence Clustering. International Journal of Mathematical Science and Engineering, Vol.3 (2) ,pp 69-75 (2014).