Comparative Study of Data Cluster Analysis for Microarray

Lokesh Kumar Sharma[#1], Sourabh Rungta[#2]

[#]*Rungta College of Engineering and Technology, Chhattisgarh Swami Vivekanand Technical University Bhilai, Chhattisgarh, India*

***Abstract*— Microarray has been a popular method for representing biological data. Microarray technology allows biologists to monitor genome-wide patterns of gene expression in a high-throughput fashion. Clustering the biological sequences according to their components may reveal the biological functionality among the sequences. Data cluster analysis is an important task in microarray data. There is no clustering algorithm that can be universally used to solve all problems. Therefore in this paper comparative study of data cluster analysis for microarray is presented. Here the most popular cluster algorithms that can be applied for microarray data are discussed. The uncertainty of data, optimization and density estimation are considered for comparison.**

***Keywords*— Microarray Data, Data Cluster Analysis, Bioinformatics.**

I. INTRODUCTION

Biomedical research and Biotechnology has been revolutionary changing. Due to advance technology the biomedical researchers are able to collect huge amount of biomedical data. An explosive growth of biomedical data, ranging from those collected in pharmaceutical studies and cancer therapy investigations to those identified in genomics and proteomics research by discovering sequential patterns, gene functions, and protein-protein interactions. The rapid progress of biotechnology and biological data analysis methods has led to the emergence and fast growth of a promising new field: bioinformatics. A Microarray data analysis is an important and challenging task in bioinformatics. Microarrays are one of the recent discoveries in experimental molecular biology. It allows monitoring of gene expression of tens of thousands of genes in parallel. Knowledge about expression levels of all may help us in almost every field of society. Amongst those fields are diagnosing diseases or finding drugs to cure them. Analysis and handling of microarray data is becoming one of the major bottlenecks in the utilization of the technology [1] [11] [12]. The eminence of DNA microarray technology is the aptitude to be used to simultaneously monitor and study the expression levels of thousands of genes, relationship between genes, their functions and classifying genes or samples that perform in a parallel or synchronized manner during imperative biological processes. Functional genomics can be better implicit when the veiled patterns in gene expression data is elucidated, however, it is very challenging to comprehend and construe this due to the complexity of biological networks and large number of genes. The most important area of microarray bioinformatics is possibly the data clustering analysis [4].

Data clustering is an exceptional preference for initial data analysis and data mining processes. To perceive and identify appealing patterns of expression across multiple genes and experiments, reveal natural structures and compress high-dimensional array data clustering must be ascertained to allow easier management of data set. This data reduction method is a simple tool yet powerful method of organizing genes based on their interdependence behaving similarly over the different conditions in different mutants, patients or at different time points in a time series during an experiment with similar expression patterns and properties into a set of disjoint groups based on specific features so that the underlying structures can be acknowledged and explored [6] [11].

Data Cluster analysis plays an indispensable role for understanding various phenomena. It primitives exploration with little or no prior knowledge, consists of research developed across a wide variety of communities. The diversity, on one hand, equips us with many tools. On the other hand, the profusion of options causes confusion. There is no clustering algorithm that can be universally used to solve all problems. Usually, algorithms are designed with certain assumptions and favour some type of biases [14]. Therefore the systematic study of data clustering is required to identify the appropriate data cluster technique before applying this technique in biological data. This regards in this work a comparative study of various data clustering algorithms for microarray data are presented.

The rest of the paper is organized as follow. Section 2 presents the various data cluster analysis methods that can be used to classify the microarray data. The experiment and result analysis are reported in the section 3. Finally the work is concluded on section 4.

II. DATA CLUSTER TECHNIQUES FOR MICROARRAY

Data Cluster Analysis is the partition data into a certain number of group or cluster. Also researchers describe a cluster by considering the internal homogeneity and the external separation, i.e., patterns in the same cluster should be similar to each other, while patterns in different clusters should not. Both the similarity and the dissimilarity should be examinable in a clear and meaningful way. The data cluster analysis process is completed in basic four steps as feature select or extraction, cluster algorithm design or selection, cluster validation and results interpretation [14]. In this section the

data cluster techniques are presented which can be used for microarray data.

A. Hierarchical Clustering

Hierarchical clustering builds a cluster hierarchy or a tree of cluster, it calls dendrogram. Every cluster node contains child clusters; sibling clusters; they partition the points covered by their common parent. Such an approach allows exploring data on different levels of granularity. Hierarchical clustering methods are categorized into agglomerative (bottom-up) and divisive (top-down) [7]. An agglomerative clustering approach starts with the two assumptions that each object is singleton cluster and recursively merges two or more appropriate clusters. A divisive clustering approach starts with all the objects in a same cluster and recursively splits the most appropriate cluster. The process continues until a stopping criterion (the requested number k of clusters) is achieved.

AGNES (Agglomerative Nesting) and DIANA (Divisia Analysis) are two earlier hierarchical clustering algorithms. AGNES is a bottom up algorithm and DIANA is top-up algorithm. These two algorithms are simple but selection of merge or split points are difficult. Improper merge or split point selection can be produce low quality clusters. Divisive clustering approach is expensive in computation. There are 2^{N-1} -1 possible two subset divisions for a cluster with N Objects.

B. K-Means Clustering

The K-Means algorithm is the well known clustering technique used in scientific and industrial applications. It is based on squared error criterion [14] K-mean algorithm initializes K-partition randomly and they change clusters based on their similarity between the objects and the cluster centroid C until a convergence criterion is met. The K-means algorithm is simple, relatively scalable and efficient and it can be easily implemented in solving many practical problems. The time complexity of K-mean algorithm is O (NKd) where d is number of iterations. Apart from these benefits, K-means algorithm has various loose falls. Therefore several enhancements of the K-means algorithm have been reported. K-means algorithm deals only numerical data set. Due to lack of universal method for identification, the initial number of partitions, the convergence centroids varies with different initial points. It is sensitive to noise and outlier data objects since a small number of data can influence the mean value. K-mean is the iteratively optimal procedure and it cannot guarantee convergence to a global optimum.

C. Fast Genetic k-Means Clustering

Fast Genetic k-Means Algorithm (FGKA) [16] is inspired by the Genetic K-means Algorithm (GKA) [4] but features several improvements over GKA. The experiments indicate that, while K-means algorithm might converge to a local optimum, both FGKA and GKA always converge to the global optimum eventually but FGKA runs much faster than GKA. In recent years, clustering algorithms have been effectively applied in molecular biology for gene expression data analysis. The goal of FGKA algorithm is to partition the N patterns into user-defined K groups, such that this partition minimizes the Total Within-Cluster Variation (TWCV, also called square-error in the literature), which is defined as follows.

Let X_1, X_2… X_N be the N patterns, and Xnd denotes the dth feature of pattern X_n (n=1…N). Each partitioning is represented by a string, a sequence of numbers $a_1 \ldots a_N$, where a_n takes a value from {1, 2… K} representing the cluster numbers that pattern X_n belongs to. Let G_k denote the kth cluster and Z_k denote the number of patterns in Gk. The Total Within-Cluster Variation (TWCV) is defined as

$$TWVC = \sum_{N=1}^{N} \sum_{d=1}^{D} X_{nd}^{2} - \sum_{k=1}^{K} \frac{1}{Z_k} \sum_{d=1}^{D} SF_{kd}^{2} \quad (1)$$

Where SF_{kd} is the sum of dth features of all the patterns in G_k. FGKA starts with the initializing phase, which generates the initial population P_0. The population in the next generation P_{i+1} is obtained by applying the genetic operators sequentially: the selection, the mutation, and the k-means operator on the current population P_i. The evaluation takes place until the termination condition is reached. The detail equations are given paper [16].

D. Incremental Genetic K-Means Clustering

Incremental Genetic K-Means Algorithm [17] is an extension to FGKA. IGKA outperforms FGKA when the mutation probability is small.

E. Fuzzy C-Means Clustering

Fuzzy clustering arises as a commonly used conceptual and algorithmic framework for data analysis and unsupervised pattern recognition. In fact, fuzzy algorithms are an extension of the classical clustering algorithms to the fuzzy domain. However, there are very few efforts, in the field of fuzzy clustering, that efficiently handle clusters of non-standard shapes. A widely used fuzzy clustering algorithm is Fuzzy C-Means (FCM). It is an extension of the K-Means algorithm for fuzzy applications [3]. FCM attempts to find i) the representative point of each cluster, which is considered to be the “center” of the cluster, and ii) the degree of membership for each object to the defined clusters. It is obvious that FCM presents the similar disadvantages with K-Means. Considering that the fuzzy clusters are represented by their centers, the degree of data membership to a cluster decreases as the points move away from the center of a cluster. Thus it mostly favours spherical clusters.

Zhang et al. [10] have developed an FCM algorithm using Pearson correlation distance as a metrics in the objective function and initialized cluster centroids with genes classified based on Gene Ontology. This algorithm was experimented with lung cancer microarray dataset and the algorithm produced more functionally significant clusters, and assigned more genes to functional groups defined in GO terms.

F. Expectation Maximization Clustering

The Expectation-Maximization (EM) algorithm is a popular iterative refinement algorithm that can be used for finding the parameter estimates. It can be viewed as an extension of the k-means paradigm, which assigns an object to the cluster with which it is most similar, based on the cluster mean. Instead of assigning each object to a dedicated cluster, EM assigns each object to a cluster according to a weight representing the probability of membership. In other words, there are no strict boundaries between clusters. Therefore, new means are computed based on weighted measures.

EM starts with an initial estimate or "guess" of the parameters of the mixture model (collectively referred to as the parameter vector). It iteratively rescores the objects against the mixture density produced by the parameter vector. The rescored objects are then used to update the parameter estimates. Each object is assigned a probability that it would possess a certain set of attribute values given that it was a member of a given cluster [2].

G. Neural Network based Clustering

Most of Artificial Neural Networks (ANN) based clustering methods use Self-Organising Maps (SOMs) or Adaptive Resonance Theory (ART) [14]. It is believed to resemble processing that occurs in the brain. Neural networks involve several layers of units that pass information from one unit to another, in an attempt to 'learn' the correct structure of clusters in a dataset. SOMs assume that the units will eventually take on the clusters' structure in space. In this form of clustering several units compete for the current object. The unit that is closest to the current object becomes the winning or active unit. The weights of the winning unit are adjusted, as well as those of its nearest neighbours, so that the units will eventually take on the structure of the clusters in space. The main disadvantage of SOMs and neural networks is the long processing time required, especially when dealing with large datasets. Kato et al. [15] proposed analysis of DNA microarray data by using self organizing maps. The rat RNA samples with DNA microarray of rat 3824 genes are used in this experiment and it has been found the result was equivalent with a usual clustering method.

H. Density based Clustering

Density based clustering methods discover cluster based on the density of points in regions. Therefore density based clustering methods are capable to produce arbitrary shapes clusters and filter out noise (outlier). Ester et al. [9] introduced density based algorithms DBSCAN and further it has generalized by using symmetric and reflexive binary predicate and introduce some non-spatial parameter "cardinality" [5]. Thus the GDBSCAN [5] algorithm can cluster point objects as well as spatially extended objects according to both, their spatial and their non-spatial attributes. Apart from this, several variants of DBSCAN algorithm have been reported in literature. The key feature of DBSCAN (Density-Based Spatial Cluster of Applications with Noise) is that for each object of a cluster the neighbourhood of a given radius ϵ has to contain at least a specified minimum number MinC of objects, i.e., the cardinality of the neighbourhood has to exceed a given threshold. Radius ϵ and minimum number MinC of objects are specified by user. Let D is a data set of objects, the distance function between the objects of D is denoted by DIST and given parameters are ϵ and MinC then DBSCAN can be specified by the following definitions. We have adopted these definitions from Ester et al. [9]

Definition 1 (Neighbourhood of an object). The ϵ-neighbourhood of an object p, denoted by $N_\epsilon(P)$ is defined by $N_\epsilon(P) = \{q \in D \mid DIST(p, q) \leq \epsilon\}$.

Definition 2 (Direct Density Reachability). An object p is direct density reachablitiy from object q w. r. t. ϵ and MinC if $|N_\epsilon(P)| \geq MinC \wedge p \in N_\epsilon(q)$.

q is called core object when the condition $|N_\epsilon(P)| \geq MinC$ holds (Fig. 1 (a, b)).

Definition 3 (Density Reachability). An object p is density-reachable from an object q w. r. t. ϵ and MinC if there is a sequence of objects $p_1 \ldots p_n$; $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density reachable for from p_i (Fig. 1 (c)).

Definition 4 (Density Connectivity). An object p is density-connectivity to object q w. r. t. ϵ and MinC if there is an object $o \in D$ such that both p and q are density reachable from o (Fig. 1 (d)).

DBSCAN chooses an arbitrary object p. It begins by performing a region query, which finds the neighbourhood of point q. If the neighbourhood contains less than MinC objects, then object p is classified as noise. Otherwise, a cluster is created and all objects in p's neighbourhood are placed in this cluster. Then the neighbourhood of each of p's neighbours is examined to see if it can be added to the cluster. If so, the process is repeated for every point in this neighbourhood, and so on. If a cluster cannot be expanded further, DBSCAN chooses another arbitrary unclassified object and repeats the same process. This procedure is iterated until all objects in the dataset have been placed in clusters or classified as noise. Raczynski et al. [8] used this density based clustering concept to microarray data analysis. It proved that DBSCAN algorithm is better option for the performing cluster on microarray data.

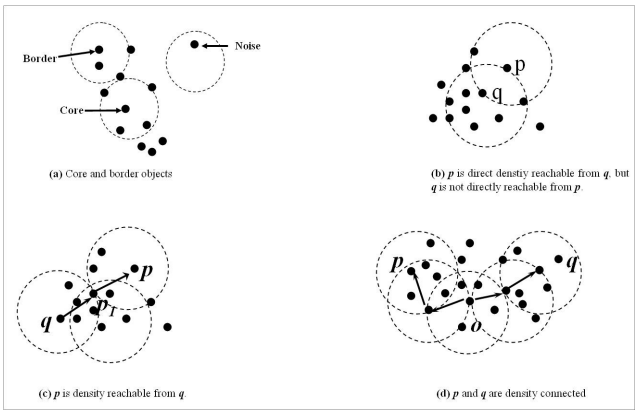

Fig. 1. Density based clustering concepts (MinC = 5).

III. EXPERIMENT AND RESULT ANALYSIS

For the comparative study, Leukemia, Lymphoma, Colon, microarray data set are used. These datasets can be downloaded from [19]. The Leukemia gene expression dataset containing expression profiles of 72 leukemia each in 7,129 genes. Pre-processed data set is used here. The pre-processing steps are presented in the paper [18]. The Lymphoma dataset contains expression measurements of 96 normal and malignant lymphocyte samples each measured using a specialized cDNA microarray, containing 40,26 genes that are preferentially expressed in lymphoid cells or which are of known immunological importance. The colon gene expression dataset containing expression values of 62 colon biopsy samples measured using high density oligonucleotide microarrays containing 2,000 genes. The above reported algorithms in section II are implemented and tested with these datasets. These data set also contain the previously classify class label. For the accuracy measure of clustering algorithm these class labels are considered and we compare with these class label and produced by algorithm. The result is shown in table 1. Overall it can be concluded that above reported cluster techniques can be applied for the microarray data set. But also the consideration of feasibility is required to consider cluster algorithm.

TABLE I
SUMMERY OF RESULT

Algorithm	**Incorrectly classify**		
	Leukemia	**Colon**	**Lymphoma**
K-means	3	10	5
EM	2	15	20
FGKA	3	5	3
IGKA	3	5	3
FCM	Not Applicable		
DBSCAN	Producing single cluster and noises		
SOM	5	15	6

IV. CONCLUSIONS

We have focused on presenting an overview of l clustering of microarray data. Microarray is a revolutionary technology. As shown above it includes many stages until a microarray is prepared and further stages until it can be analyzed. The performance of every clustering algorithm may vary significantly with diverse data sets, and there is no absolute finest algorithm among the clustering algorithms.

REFERENCES

[1] A. L. Tarca, R. Romero, and S. Draghici, "Analysis of microarray experiments of gene expression profiling",American Journal of Obstetrics and Gynecology (2006) 195, pp. 373–88.

[2] C. Escudero et al., "Classification of Gene Expression Profiles: Comparison of k-means and expectation maximization algorithms", IEEE Computer Society, 2008, pp. 831-836.

[3] D. Dembele and P. Kastner, "Fuzzy C-means method for clustering microarray data", Bioinformatics, Vol. 19, Issue 8, 2003, pp. 973-980.

[4] E. Naghieh and Y. Peng, "Microarray Gene Expression Data Mining: Clustering Analysis Review", Techniques, 2009.

[5] J. Sander, M. Ester, H. P. Kriegel and X. Xu, "Density- Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications", Journal of Data Mining and Knowledge Discovery, Kluwer Academic Publishers vol. 2, 1998 pp. 169-194.

[6] K. Krishna and M. Murty, "Genetic K-Means Algorithm", IEEE Transactions on Systems Man. and Cybernetics vol. 29, NO. 3, 1999, pp. 433-439.

[7] L. Kaufman and P. J. Rousseeuw, "Finding Group in Data: an Introduction to Cluster Analysis", John Wiley and Sons, 1990.

[8] L. Raczynski, J. Wozniak, T. Rubel and K. Zaremba,"Application of Density Based Clustering to Microarray Data Analysis", Int. Journal of Electronics and Telecommunications, 2010, Vol. 56, No. 3, pp. 281-286.

[9] M. Ester, H. P. Kriegel, J. Sander and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", In: Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD'96), Portland, 2006, AAAI Press 291-316.

[10] M. Zhang et al., "A fuzzy C-means algorithm using a correlation metrics and gene ontology", IEEE 19th Int. Conf. on Pattern Recognition, 2008. pp. 1-4.

[11] P. Valarmathie, T. Ravichandran, K. Dinakaran, "Survey of Clustering Algorithms for Microarray Gene Expression Data", European Journal of Scientific Research, Vol. 69, No. 1, 2012, pp. 5-20.

[12] R. D. Bin and D. Risso,"Clustering via nonparametric density estimation:an application to microarray data", BMC Bioinformatics, 2011 pp. 102-105.

[13] R. Suzuki and H. Shimodaira, "An application of multiscale bootstrap resampling to hierarchical clustering of microarray data: How accurate are these clusters?" proc. by the Fifteenth Int. Conference on Genome Informatics (GIW 2004). 2004. p. P034.

[14] R. Xu and D. Wunsh, "Survey of Clustering Algorithms", IEEE Transactions on Neural Networks, Vol. 16, No. 3, May 2005, pp. 645-678.

[15] T. Kato, K. Fujimura, H. Tokutaka, "Analysis of DNA Microarray Data by Using Self-Organizing Maps", Genome Informatics 14, 2003, pp. 328-329.

[16] Y. Lu, S. Lu, F. Fotouhi, Y. Deng, and S. Brown, "FGKA: A Fast Genetic K-means Clustering Algorithm", ACM 1-58113-812-, 2004.

[17] Y. Lu, S. Lu, F. Fotouhi, Y. Deng, and S. Brown, "Incremental genetic K-means algorithm and its application in gene expression data analysis", BMC Bioinformatics 5:172, 2004.

[18] K. Deb and A. R. Reddy, "Classification of Two and Multi Class Cancer Data Reliably using Multi Objective Evolutionary Algorithms", IIT Kanpur, KanGAL Report Number 2003006.

[19] http://www.iitk.ac.in/kangal/bioinfo.shtml valid on 21 May 2012.