ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 6, Issue 5, May 2019

Clustering Analysis of Gene Expression Profile: An Overview

[1] Arpan Kumar Das, [2] Dr.G. Sadashivappa
[1] P.G Scholar, [2] Professor
[1][2]Dept. of Telecommunication Engineering, R V College of Engineering, Bengaluru

Abstract: **Using DNA microarray technology, biologists get a large number of gene expression time series data. Clustering is a significant approach in extracting biological information from these data. This paper discusses HMM-based hierarchical clustering (HMM-HC) and Genetic clustering algorithm (GA) to analyse gene expression time series data. Some key research challenges associated with clustering analysis are also included.**

Keywords - Clustering analysis, Gene Expression, Hidden Markov Model, Hierarchical clustering, Genetic algorithm..

I. INTRODUCTION

Clustering is the process of grouping a set of objects into clusters so that objects within a cluster are similar to each other but are dissimilar to objects in other clusters [22]. Clustering has been effectively applied in a variety of engineering and scientific disciplines such as psychology, biology, medicine, computer vision, communications, and remote sensing. Cluster analysis organizes data (a set of patterns, each pattern could be a vector measurements) by abstracting underlying structure. The grouping is done such that patterns within a group (cluster) are more similar to each other than patterns belonging to different groups. Thus, organization of data using cluster analysis employs some dissimilarity measure among the set of patterns. The dissimilarity measure is defined based on the data under analysis and the purpose of the analysis.

Various types of clustering algorithms have been proposed to suit different requirements. Clustering algorithms can be broadly classified into hierarchical and partitioning algorithms based on the structure of abstraction. Hierarchical clustering algorithms construct a hierarchy of partitions, represented as a dendrogram in which each partition is nested within the partition at the next level in the hierarchy. Partitioning clustering algorithms generate a single partition, with a specified or estimated number of non-overlapping clusters, of the data in an attempt to recover natural groups present in the data. One of the important problems in partitioning clustering is to find a partition of the given data, with a specified number of clusters that minimizes the cluster variation. Unfortunately in many real life cases the number of clusters in a data set is not known a priori. Under this condition, how to automatically provide the number of clusters and find the clustering partition becomes a challenge.

In this regard, genetic algorithms are used for automatically clustering data sets [14]. Genetic algorithms (GA's) work on a coding of the parameter set over which the search has to be performed, rather than the parameters themselves [36]. These encoded parameters are called solutions or chromosomes and the objective function value at a solution is the objective function value at the corresponding parameters. GA's solve optimization problems using a population of a fixed number, called the population size, of solutions. A solution consists of a string of symbols, typically binary symbols. Genetic algorithms evolve over generations. During each generation, they produce a new population from the current population by applying genetic operator's viz., natural selection, crossover, and mutation. Each solution in the population is associated with a figure of merit (fitness value) depending on the value of the function to be optimized. The selection operator selects a solution from the current population for the next population with probability proportional to its fitness value. Crossover operates on two solution strings and results in another two strings. Typical crossover operator exchanges the segments of selected strings across a crossover point with a probability. The mutation operator toggles each position in a string with a probability, called the mutation probability. Bandyopadhyay and Maulik [26] applied the variable string length genetic algorithm with the real encoding of the coordinates of the cluster centers in the chromosome to the clustering problem. Tseng and Yang [28] proposed a genetic algorithm based approach for the clustering problem. Their method consists of two stages, nearest neighbor clustering and genetic optimization. Lin et al. [17] presented a genetic clustering algorithm based on a binary chromosome representation. This method selects the cluster centers directly from the data set. Lai [18] adopted the hierarchical genetic algorithm to solve the clustering problem. In the

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 6, Issue 5, May 2019

proposed method, the chromosome consists of two types of genes, namely, the control genes and the parametric genes.

II. CLUSTERING ANALYSIS

Cluster analysis is a class of techniques used to classify objects or cases into relatively homogeneous groups called clusters. Objects in each cluster tend to be similar to each other and dissimilar to objects in the other clusters. Cluster analysis is also called classification analysis, or numerical taxonomy. Both cluster analysis and discriminant analysis are concerned with classification. However, discriminant analysis requires prior knowledge of the cluster or group membership for each object or case included, to develop the classification rule. In contrast, in cluster analysis there is no a priori information about the group or cluster membership for any of the objects. Groups or clusters are suggested by the data, not defined a priori.

A. Statistics associated with clustering analysis

• Agglomeration schedule: An agglomeration schedule gives information on the objects or cases being combined at each stage of a hierarchical clustering process.
• Cluster centroid. The cluster centroid is the mean values of the variables for all the cases or objects in a particular cluster.
• Cluster centers: The cluster centers are the initial starting points in non-hierarchical clustering. Clusters are built around these centers, or seeds.
• Cluster membership: Cluster membership indicates the cluster to which each object or case belongs.
• Dendrogram: A dendrogram, or tree graph, is a graphical device for displaying clustering results. Vertical lines represent clusters that are joined together. The position of the line on the scale indicates the distances at which clusters were joined. The dendrogram is read from left to right.
• Distances between cluster centers: These distances indicate how separated the individual pairs of clusters are. Clusters that are widely separated are distinct, and therefore desirable.
• Icicle diagram: An icicle diagram is a graphical display of clustering results, so called because it resembles a row of icicles hanging from the eaves of a house. The columns correspond to the objects being clustered, and the rows correspond to the number of clusters. An icicle diagram is read from bottom to top.
• Similarity/ distance coefficient matrix: A similarity/distance coefficient matrix is a lower-triangle matrix containing pairwise distances between objects or cases.

B. Conducting a clustering analysis

• Hierarchical clustering is characterized by the development of a hierarchy or tree-like structure. Hierarchical methods can be agglomerative or divisive.
• Agglomerative clustering starts with each object in a separate cluster. Clusters are formed by grouping objects into bigger and bigger clusters. This process is continued until all objects are members of a single cluster.
• Divisive clustering starts with all the objects grouped in a single cluster. Clusters are divided or split until each object is in a separate cluster.
• Agglomerative methods are commonly used in marketing research. They consist of linkage methods, error sums of squares or variance methods, and centroid methods.
• The non-hierarchical clustering methods are frequently referred to as k-means clustering. These methods include sequential threshold, parallel threshold, and optimizing partitioning.
• In the sequential threshold method, a cluster center is selected and all objects within a prespecified threshold value from the center are grouped together. Then a new cluster center or seed is selected, and the process is repeated for the unclustered points. Once an object is clustered with a seed, it is no longer considered for clustering with subsequent seeds.
• The parallel threshold method operates similarly, except that several cluster centers are selected simultaneously and objects within the threshold level are grouped with the nearest center.
• The optimizing partitioning method differs from the two threshold procedures in that objects can later be reassigned to clusters to optimize an overall criterion, such as average within cluster distance for a given number of clusters.

The steps involved in conducting a clustering analysis is given in Figure 1.

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 6, Issue 5, May 2019

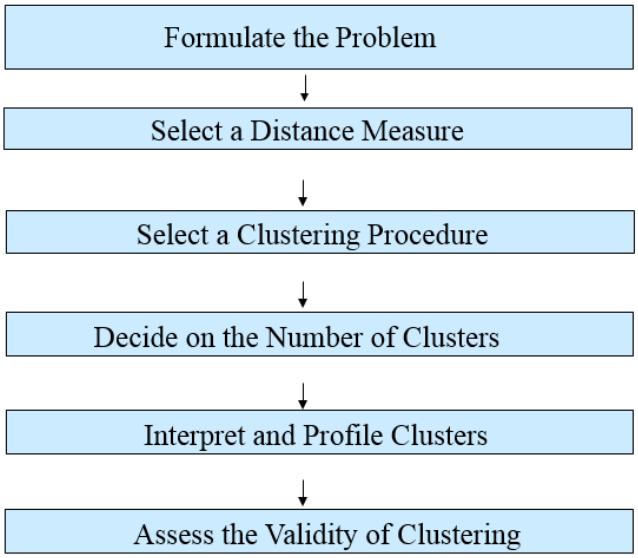

Figure 1. Flowchart for clustering analysis

III. GENE EXPRESSION PROFILE

• A 2D matrix of gene expression is assumed to have measurements as rows represent genes.
• Columns represent different experiments, time points, individuals etc. (what one can measure using one microarray)
• Individual rows or columns are referred to as profiles i.e. a row is a profile for a gene as shown in Figure 2.
• Task Definition: Given the expression profiles for a set of genes or experiments/individuals/time points (whatever columns represent)
• To do: organize profiles into clusters such that instances in the same cluster are highly similar to each other whereas instances from different clusters have low similarity to each other.

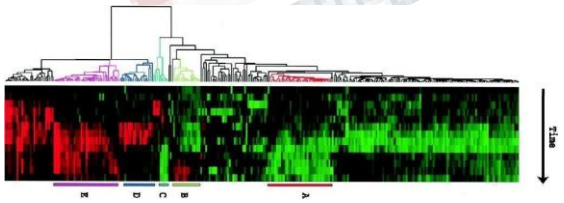

Figure 2. Hierarchical clustering of expression data

A. Distance Metrics

Distance metrics are used to measure distance (variations) between the elements (objects) of a cluster to determine its meaningfulness. This also determines the clustering algorithm that has to be used.

Distance Properties:

$\text{dist}(x_i, x_j) \geq 0$ (non-negativity)
$\text{dist}(x_i, x_j) = 0$ if and only if $x_i = x_j$ (identity)
$\text{dist}(x_i, x_j) = \text{dist}(x_j, x_i)$ (symmetry)
$\text{dist}(x_i, x_j) \leq \text{dist}(x_i, x_k) + \text{dist}(x_k, x_j)$ (triangle inequality)

Distance Metrics:

Manhattan $\quad \text{dist}(x_i, x_j) = \sum_e |x_{i,e} - x_{j,e}|$

Euclidean $\quad \text{dist}(x_i, x_j) = \sqrt{\sum_e (x_{i,e} - x_{j,e})^2}$

e ranges over the individual measurements for x_i and x_j

B. Data Processing

Gene expression data should be processed before modelling. After data normalization, gene expression data should be represented by discrete symbol in order to make HMM model. The expression-level measurements of a gene in a given situation have a roughly Gaussian distribution . With common technologies, the logarithm of the expression levels obeys normal distribution . Therefore, we can use this feature to convert data to discrete symbols using equation (1). The data normalization uses the following formula:

$$x'_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(x_{ij} - \bar{x}_i)^2}} \qquad (1)$$

Where x_{ij} is the expression data of gene i at time point j, $\bar{x}_i$; is the average expression data of gene i. The above formula makes the mean of expression data, μ, to be 0, and the standard deviation, σ, to be 1. The data normalization is to make all of the data into the same area. Furthermore, the following equation (2) is used to make the gene expression data fall into the particular interval [a, b]:

$$\tilde{x}_{ij} = \frac{(b-a)(x'_{ij} - x'_{i\min})}{x'_{i\max} - x'_{i\min}} + a \qquad (2)$$

Where $x'_{i\min}$ and $x'_{i\max}$ are the minimum and maximum values in a single gene expression time-series data sequence. 'a' and 'b' can respectively take the values of -3σ and 3σ. According to the nature of the normal distribution, it can be determined which one of the three symbols: I (increased), D (decreased) and U (unchanged), should be used to represent the data given by equation (3).

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 6, Issue 5, May 2019

$$S_{ij} = \begin{cases} U & -\sigma \le \tilde{x}_{ij} \le \sigma \\ I & \tilde{x}_{ij} > \sigma \\ D & \tilde{x}_{ij} < -\sigma \end{cases} \qquad (3)$$

Where S_{ij} is the obtained symbol representing the gene expression time series data.

IV. HMM-HC ALGORITHM

In hierarchical clustering, the data is not partitioned into a particular cluster in a single step. Instead, a series of partitions takes place, which may run from a single cluster containing all objects to 'n' clusters that each contain a single object. Hierarchical Clustering is subdivided into agglomerative methods, which proceed by a series of fusions of the n objects into groups, and divisive methods, which separate 'n' objects successively into finer groupings. Agglomerative techniques are more commonly used. Hierarchical clustering may be represented by a two-dimensional diagram known as a dendrogram, which illustrates the fusions or divisions made at each successive stage of analysis. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other. Hierarchical clustering starts by treating each observation as a separate cluster. Then, it repeatedly executes the following two steps: first, identify the two clusters that are closest together, and second, merge the two most similar clusters. This continues until all the clusters are merged together.

The choice of distance metric should be made based on theoretical concerns from the domain of study. Where there is no theoretical justification for an alternative, the Euclidean should generally be preferred, as it is usually the appropriate measure of distance in physical world.

After selecting a distance metric, it is necessary to determine from where distance is computed. For example, it can be computed between the two most similar parts of a cluster (single- linkage), the two least similar bits of a cluster (complete-linkage), the center of the clusters (mean or average-linkage), or some other criterion. Many linkage criteria have been developed.

As with distance metrics, the choice of linkage criteria should be made based on theoretical considerations from the domain of application. A key theoretical issue is what causes variation. Where there are no clear theoretical justifications for choice of linkage criteria, Ward's method is the sensible default. This method works out which observations to group based on reducing the sum of squared distances of each observation from the average observation in a cluster.

The HMM-HC clustering algorithm is divided into two stages. First, model gene expression time-series data, every model representing one cluster. Second, cluster models with the strategy of hierarchy. The specific steps of the algorithm are as follows:

l. For the given gene expression data sequences: {O1, O2, On}, build a HMM model λ1 for the sequence O1 at first, and then put O1 into the cluster c1, c1: {O1};

2. Calculate the probability of the sequence O2 produced by model λ1, P(O2| λ1) , if |l-P(O2| λ1)/P(O1| λ1)| < ε, put O2 into the cluster c1 , c1:{O1, O2 }, and adapt parameters of the model λ1, else build a new model λ2 for O2, and put O2 in the cluster c2,c2:{O2};

3. For the sequence Oi, assume that there have already built k models: λ1, λ2… λk and k clusters:

$$\begin{array}{ll} c_1 & \{O_{11}, O_{12}, \ldots, O_{1j_1}\} \\ c_2 & \{O_{21}, O_{22}, \ldots, O_{2j_2}\} \\ \ldots & \ldots \\ c_k & \{O_{k1}, O_{k2}, \ldots, O_{kj_k}\} \end{array}$$

Find λM which could produce the maximum probability, P (Oi | λM), and the sequence OM in the cluster cM which has the maximum probability in model λM, P (OM | λM). If |l-P (Oi | λM) / P (OM | λM) | < ε, put the sequence Oi; into cluster cM, and adjust the parameters of the model λM, else build a new model λk+1, put Oi; into new cluster ck+1, ck+1: {Oi};

4. Repeat step 3 until finding appropriate clusters for all the sequences;

5. For the HMM models obtained by the above steps, every time merge two models which have the smallest distance among all the models until only one is remained. In the merging process, first combine the clusters which the models represent, then use the members of the combined clusters to train the parameters of one of the two models. Retain the trained model and delete another.

In the algorithm, ε is a real number close to 0, and the distance between two models is defined by equation (4):

$$D_s(\lambda_x \| \lambda_y) = \frac{D(\lambda_x \| \lambda_y) + D(\lambda_y \| \lambda_x)}{2} \qquad (4)$$

Where D (λx ‖ λy) and D (λy ‖ λx) respectively represent the distance from λx to λy and from λy to λx. The time complexity of HMM-HC is smaller than existing model-based methods. When building and merging HMM models, the parameters training process of HMM-HC doesn't use all of the gene sequences to train parameters, which reduces a lot of computation.

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 6, Issue 5, May 2019

V. GENETIC CLUSTERING ALGORITHM

Various adaptations are used to enable the GA to cluster and to enhance their performance. Further the Genetic Clustering Algorithms are tested on databases, which are benchmarks for data mining applications or heuristics are added to enable the GAs to cope with a larger number of objects. Genetic algorithm for the clustering problem fall into the following areas: representation, fitness function, operators and parameter values. Figure 3 shows the dendrogram for the hierarchically related clusters.

A. Representation

Genetic representations for clustering or grouping problems are based on underlying scheme. The scheme represents the objects with gene values, and the position of these genes signifies how the objects are divided amongst the clusters. The use of simple encoding scheme causes problems of redundant codification and context insensitivity .This has led researchers to devise complicated representations and specialized operators for clustering problems. The cluster label based on n bit encoding is simple compared to parameterization of prototype location.

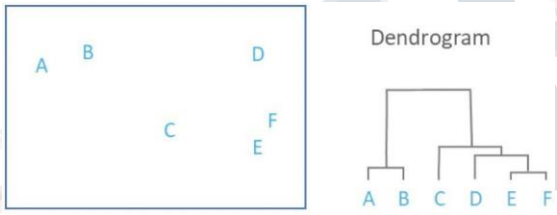

Figure 3. A dendrogram showing hierarchical relationship between clusters

In such a representation many genotype translate to a unique phenotype. The notion of cluster labels built into the representation makes little intuitive sense. Such representations have spawned off a set of pre-treatment methodologies to make the representations suitable for genetic operators.

Thus, the size of the search space the genetic algorithm has to search is much larger than the original space of solutions. This augmented space may reduce the efficiency of the genetic algorithm. In addition, the redundant encoding also causes the undesirable effect of casting context dependent information out of context under the standard crossover, i.e., equal parents can originate different offspring.

B. Fitness Function

Objective functions used for traditional clustering algorithms can act as fitness functions for Genetic Clustering algorithms. However if the optimal clustering corresponds to the minimal objective function value, one will need to transform the objective function value, since GAs work to maximize their fitness values. In addition fitness values in a GA need to be positive if one is using fitness proportional selection.

C. Genetic Operators

The operators pass genetic information between subsequent generations of the population. As a result, operators need to be matched with or designed for the representation, so that the offspring are valid and inherit characteristics from their parents. Operators used for genetic clustering or grouping includes some of the selection, crossover and mutation methods.

1) Selection

Chromosomes are selected for reproduction based on their relative fitness. Thus the representation is not a factor when choosing an appropriate selection operator, but the fitness function is. If all fitness values are positive, and the maximum fitness value corresponds to the optimal clustering, then fitness proportional selection may be appropriate. Otherwise, a ranking selection method may be used.

In proposed Genetic Clustering Algorithm, the genotypes corresponding to each generation are selected, which does not admit negative objective function values. For this reason, a constant equal to one is summed up to the objective function before the selection procedure takes place. The highest fitness genotype is always copied into the succeeding generation.

2) Crossover

The crossover operator is designed to transfer genetic material from one generation to the next. The major concerns with this operator are validity and context insensitivity. It may be necessary to check that offspring produced by a certain operator are valid and reject any invalid chromosomes. The proposed Genetic Clustering Algorithm crossover operator combines clustering solutions coming from different genotypes. It works in the following way. First, two genotypes (G1 and G2) are selected. Then, assuming that G1 represents k1 clusters, the Genetic Clustering Algorithm randomly chooses $c \in \{1, 2, K1\}$ clusters to copy into G2. The unchanged clusters of G2 are maintained and the changed ones have their instances allocated to the corresponding nearest clusters (according to their centroids). In this way, an offspring G3 is obtained. The same procedure is employed to get an offspring G4, but now considering that the changed clusters of G2 are copied

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 6, Issue 5, May 2019

into G1.Note that, although the crossover operator produces offspring usually formed by a number of clusters that is neither smaller nor larger than the number of clusters of their parents, this operator is able to increase or decrease the number of clusters.

3) Mutation

Mutation introduces new genetic material into the population. In a clustering context this corresponds to moving an object from one cluster to another. Two operators for mutation are used in the Genetic Clustering Algorithm. The first operator works only on genotypes that encode more than two clusters. It eliminates a randomly chosen cluster, placing its instances into the nearest remaining clusters (according to their centroids). The second operator divides a randomly selected cluster into two new ones. The first cluster is formed by the instances closer to the original centroid, whereas the other cluster is formed by those instances closer to the farthest instance from the centroid.

VI. RESEARCH CHALLENGES

The data required for conducting a clustering analysis are divided into the two parts namely sparse data and dense data. For sparse data, the accuracy is low for the available clustering methods. For dense data, the clustering accuracy is high. Hence, a subspace is used for clustering high dimensional datasets to improve the accuracy. In order to form effective clusters in complex data types, higher order statistical methods like linear moment (summarize the shape of probability distribution) and Pareto index (specifying the Pareto distribution) are used. Clustering technique also needs to focus on how to reduce the time complexity without compromising cluster quality and optimality. On similar lines several new investigations may be carried out in future. Some of the key research challenges that are faced during clustering analysis are:

1) For large real data sets, the relational distribution of data is difficult which helps to understand where the class labels are.

2) For a large number of fragmental clusters, scalability remains an important issue.

3) For huge biological data sets, high cost of computation remains a major problem for the researchers.

4) Cluster Validation remains to be a fundamental problem in unsupervised clustering algorithms.

5) Functions that enable users to efficiently search within sub-networks of interest is not yet added to the existing algorithms.

VII. CONCLUSIONS

Data mining is the process of retrieving a data from extracting useful knowledge, to achieve the effective utilization of data resources. In this research the number of contributions that are brought in an effective way to accomplish the predicted results. Clustering is the unsupervised classification of observations, data points or feature vectors into groups. The clustering problem has been discussed in many contexts and by the investigators in many disciplines; this shows its widespread interest and usefulness as one of the steps in exploratory data analysis. One such is the field of genetics where it is used for the extraction of significant patterns from disease data warehouses for the efficient guess of heart diseases, cancer and tumor. Based on the computation of substantial weightage, the numerous patterns having value greater than predefined threshold are chosen for valuable prediction of cancer, heart disease and tumor.

An HMM-based hierarchical clustering (HMM-HC) method is used to analyse gene expression time series data. Compared to conventional methods, the HMM-HC method can take advantage of the special feature that in a gene expression profile the time point data is correlated with others. Moreover, the hierarchical strategy increases the efficiency of the method so that it is suitable to cluster high through put gene expression time series data. The HMM-HC method can not only produce high quality clusters, but find out the appropriate cluster number. Additionally, the Genetic algorithm is very powerful in providing optimization for large subsets of the search space.

REFERENCES

[1] Dongmin Seo, Yunsoo Choi, Min-Ho Lee, Seok-Jung Yu, "Development of biological network crawling, clustering and visualization system", International Conference on Electrical Engineering/Electronics, Computer, Telecom and IT,pp.726-728,2017

[2] Amutha Priya and R. Lawrence, "Algorithm for clustering analysis of gene expression data using MapReduce framework", International Conference on Data Storage and Data Engineering, pp.1-4, 2016

[3] Zhinwen Yu, Hanato Chen, Jiming,Wong, Han and Li, "Adaptive Fuzzy consensus clustering framework for cluster analysis of cancer data", IEEE Transactions on Computational Biology and Bioinformatics,pp.1-14,2013

[4] Liming Wang and Xiaodong Wang, "A Non-parametric Bayesian clustering for Gene expression data", IEEE Statistical Signal Processing Workshop, pp.556-559, 2012

[5] Damodar Edla,Seshaiah Machavarapu and Prasanta Jana, "An Improved MST based clustering for Biological data",International Conference on Data Science and Engineering", pp.42-45,2012
[6] Ricardo Campello, Davoud Moulavi and Jorg Sander, "A simpler and more accurate framework for clustering and visualization of biological data", IEEE Transactions on Computational Biology and Bioinformatics,Vol.9,No.6,pp.1850-1852,Nov/Dec.2012
[7] Xiao Zhang, Aichen Li, You Zhang and Yongpeng Xiao, "Validity of Cluster technique for Genome expression data", Chinese Control and Decision Conference, pp.3737-3741, 2012
[8] Suvendu Kanungo, Gagadhar Sahoo and Manoj Gore, "A Co-clustering technique for Gene expression data using Bi-partite Graph approach", IEEE Statistical Signal Processing Workshop, pp.1-5, 2010
[9] Guoqing Zhao and Wei Dang, "An HMM based hierarchical clustering method for gene expression time series data", IEEE Statistical Signal Processing Workshop, pp.219-222, 2010
[10] Gunjan Gupta, Alexander Liu and Joydeep Ghosh, "Automated hierarchical density shaving:A robust automated clustering and visualization framework for large biological data sets",IEEE Transactions on Computational Biology and Bioinformatics, Vol.7, No.2, April/May, 2010
[11] R. Xu, S. Damelin, B. Nadler, D. C. Wunsch II, "Clustering of highdimensional gene expression data with feature filtering methods and diffusion maps," Artificial Intelligence in Medicine, vol. 48, pp. 91-98, 2010.
[12] A. Mukhopadhyay, U. Maulik, 'Towards improving fuzzy clustering using support vector machine: Application to gene expression data," Pattern Recognition, vol. 42, pp. 2744-2763, 2009.
[13] Hae-Sang Park and,Chi-Hyuck Jun,"A simple and fast algorithm for k-medoids clustering," Expert System with applications,3336- 3341,2009.
[14] Amit Banerjee and Sushil J.Louis," A Recursive Clustering Methodology using a genetic algorithm", IEEE Trans., 2007.
[15] L. Wang, F. Chu, W. Xie, "Accurate cancer classification using expressions of very few genes," IEEEIACM Transactions on Computational Biology and Bioinformatics, vol. 4 , issue I, pp. 40-53, 2007.
[16] C.M. Yang, B.K. Wan, X.F. Gao, "Internal validation technology research of the gene clustering result," Progress in Natural Science, vol. 17, no. 9, pp. 1181-1188,2007.
[17] H. J. Lin, F. W. Yang and Y. T. Kao", An efficient GA-based clustering technique, "Tamkang Journal of Science and Engineering, vol. 8, no. 2,pp. 113-122, 2005.
[18] C. C. Lai,"A novel clustering approach using hierarchical genetic algorithms," Intelligent Automation and Soft Computing, vol. 11, no. 3,pp. 143-153, 2005.
[19] A. Schliep, I.G. Costa, C. Steinhoff, A. Schonhuth, "Analyzing gene expression time-courses," IEEEIACM Transactions on Computational Biology and Bioinformatics, vol. 2, no. 3, pp. 179-193,2005.
[20] X. Sun, Z.H. Lu, J.M. Xie, Foundation of bioinformatics, G.X. Chen, C.M. Zhao, Ed. Beijing, China: Tsinghua University Press, 2005.
[21] E. R. Hruschka, R. J. G. B. Campello, and L. N. de Castro,"Improving the efficiency of a clustering genetic algorithm, In Advances in Artificial Intelligence," IBERAMIA 2004, volume 3315 of LNCS, pages 861–870, 2004.
[22] Jiawei Han and M.kamber", Data mining: Concepts and Techniques,"Morgan Kaufmann,2004.
[23] E.R. Hruschka, N.F.F. Ebecken,"A genetic algorithm for cluster analysis",Intell. Data Anal. 7 (1) 15–25., 2003.
[24] Z. Shi, G. Joydeep, "A unified framework for model-based clustering," Journal of Machine Learning Research, vol. 4, no 6, pp. 1001--I037, 2003.
[25] N. Bolshakova, F. Azuaje, "Cluster validation techniques for genome expression data," Signal Processing, vol. 83, pp. 825-833,2003.
[26] S. Bandyopadhyay and U. Maulik,"An evolutionary technique based on K-means algorithm for optimal clustering in RN," Information Sciences,vol. 146, no.1-4, pp. 221-237,2002.
[27] Y. Xu, V. Olman, D. Xu, "Clustering gene expression data using a graph-theoretic approach, an application of minimum spanning trees," Bioinformatics, vol. 18, no. 4, pp. 536-545,2002.
[28] L. Y. Tseng and S. B. Yang,"A genetic approach to the automatic clustering algorithm," Pattern Recognition, vol. 34, no. 2, pp. 415- 424,2001
[29] B.S. Everitt, S. Landau, M. Leese,"Cluster Analysis," Arnold Publishers, London, 2001.
[30] P. Baldi and S. Brunak, Bioinformatics: The machine learning approach, 2nd ed., T. Dietterich, Ed. London, England: The MIT Press, 2001.
[31] KY. Yeung, D.R. Haynor, W.L. Ruzzo, "Validating clustering for gene expression data," Bioinformatics, vol. 17, no. 4, pp. 309-318, 2001.
[32] V. R. Iyer, M. B. Eisen, D.T. Ross, G. Schuler, T. Moore, J.C. F. Lee, J.M. Trent, L. M. Staudt, HJ. James, M.S. Boguski, D. Lashkari, D. Shalon, D. Botstein, P.O. Brown, "The transcriptional program in the response of human fibroblasts to serum," Science, vol. 283, pp. 83-87, 1999.

[33] RJ. Cho, MJ. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, DJ. Lockhart, R.W. Davis, "A genome-wide transcriptional analysis of the mitotic cell cycle," Mol. Cell., vol. 2, no. I, pp. 65-73, July 1998.
[34] G. P. Babu and M. N. Murty,A near-optimal initial seed selection in K-means algorithm using a genetic algorithm," Pattern Recognit. Lett.,vol. 14, pp. 763–769, 1993.
[35] L. Kaufman, P. J. Rousseeuw,"Finding Groups in Data—An Introduction to Cluster Analysis," Wiley Series in Probability and Mathematical Statistics, 1990.
[36] Goldberg, D.E, "Genetic Algorithms in Search, Optimization and Machine Learning, "Addison-Wesley, 1989
[37] G.L. Liu," Introduction to Combinatorial Mathematics," McGraw Hill, New York, 1968.