A Survey on Clustering Algorithm for Microarray Gene Expression Data

M. S. Uma
M.Phil Research scholar
Department of CS
Bharathiar University
Coimbatore, India
saranyamsuma@gmail.com

R. Porkodi
Assistant Professor
Department of CS
Bharathiar University
Coimbatore, India
Porkodi_r76buc.edu.in

Abstract— The DNA data are huge multidimensional which contains the simultaneous gene expression and it uses the microarray chip technology, also handling these data are cumbersome. Microarray technique is used to measure the expression level from tens of thousands of gene in different condition such as time series during biological process. Clustering is an unsupervised learning process which partitions the given data set into similar or dissimilar groups. The mission of this research paper is to analyze the accuracy level of the microarray data using different clustering algorithms and identify the suitable algorithm for further research process.

Keywords- *Microarray technology, Clustering techniques, Partition algorithm, Fuzzy c-means Algorithm, Hierarchical clustering, Model based clustering.*

*****

I. INTRODUCTION

Data mining is often defined as finding hidden information or extracting meaningful information from large database. The extraction of meaningful information from a large database is known as "Knowledge discovery". Clustering is the task of grouping set of object in such a way that objects in the same group called a cluster[1].A good clustering method will produce high quality clusters in which the intra class that is similarity is high. The inter class similarity is low. Clustering can be applied in many filed marketing, biology, library, insurance, city planning, www and many more. In data mining the data is mined using two learning approaches i.e. supervised learning or unsupervised clustering. Classification is supervised learning problem collection of labeled data. This model are called predictive. Data object explanatory variables and one or more dependent variables. Clustering is unsupervised learning problem so as every problem it deals with finding structure in a collection of unlabled data. This model are sometimes called descriptive model. Clustering is one of the most common untested data mining methods that explore the hidden structures embedded in a dataset. Data object dependent and explanatory variables[2]. Microarray technology measure copy number of modules in a mixture on a small slide.Thousands or millions of different kind of module can be measured [3].Thus creating large volumes of data per biological sample. The module can be DNA, RNA or protein.

Clustering gene expression data in such a group sample clustering and gene-based clustering.

Group samples:

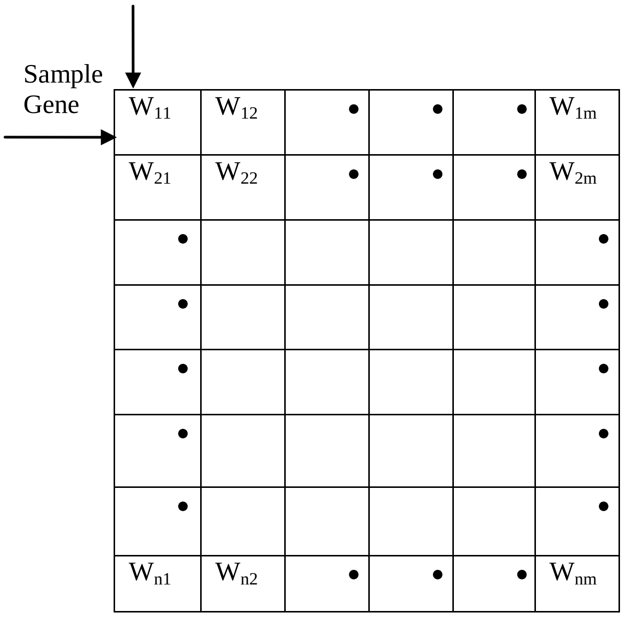
Group together tissues that are similarly affected by a disease.

Group together patients that are similarly affected by a disease.

Group genes:

Group genes: Group together gene that are similarly affected by a disease.

Group together gene that respond similarly to a experimental conditions. [4]A microarray experiment typically assesses a large number of DNA sequences (genes, cDNA clones, or expressed sequence tags [ESTs]) under multiple conditions. These conditions may be a time series during a biological process e.g., the yeast cell cycle or a collection of different tissue samples e.g., normal versus cancerous tissues. In this paper, focus on the cluster analysis of gene expression data without making a distinction among DNA sequences, which will uniformly be called "genes". Microarrays measure activity expression level of genes under varying conditions and or points in time.[5]Microarray data usually transformed into an intensity matrix. The expression patterns of genes form the matrix rows= {g_1, $g_2 \ldots g_n$}, and the expression profiles of samples represents by the matrix columns = s{$s_1, s_2, \ldots s_n$}. Each cell w_{ij} is the measured expression level of i^{th} gene in j^{th} sample. Table 1 ,A typical gene expression data represented by matrix.


Sample
Gene

W_{11}	W_{12}	•	•	•	W_{1m}
W_{21}	W_{22}	•	•	•	W_{2m}
•					•
•					•
•					•
•					•
•					•
W_{n1}	W_{n2}	•	•	•	W_{nm}

Microarray technology is a developing technology used to study the expression of many genes at once. It involves placing thousands of gene sequences in known locations on a glass slide called a gene chip. A sample containing DNA or RNA is placed in contact with the gene chip[6] . DNA microarrays (also called Gene Chips) are devices not much larger than postage stamps. They are based printed on a glass substrate containing as many as 400,000 tiny cell each containing a microscopic spot of DNA. Each microscopic spot holds a short, synthetic, single-stranded DNA sequence from a different human gene[7].this makes it possible to carry out a very large number of genetic tests on a sample at one time. An array is an orderly arrangement of samples where matching of known and unknown DNA samples is done based on base pairing rules. An array experiment makes use of common assay systems such as micro plates or standard blotting membranes. The sample spot sizes are typically less than 200 microns in diameter usually contain thousands of spots.Thousands of spotted samples known as probes with known identity are immobilized on a solid support a microscope glass slides or silicon chips or nylon membrane[8].The spots can be DNA, cDNA, or oligonucleotides. These are used to determine complementary binding of the unknown sequences thus allowing parallel analysis for gene expression and gene discovery.

The paper organized as follows section 1, describe the literature review, section 2 describe the various clustering algorithm, section 3 describe clustering validation, section 4 describe comparison of clustering algorithm, finally the paper is concluded in section 5

II. LITERATURE REVIEW

In paper [9] A.Dharmarajan., T.Velmurugan presented about the performance of partitioning based algorithm. The algorithm was analyzed by using the selected three attributes from the total number of attribute. In this algorithm LC arff and LC.csv datasets are used. The contents of dataset are completely numeric symbols with k-means algorithm to give better result.

In paper [10] Jyrki Joutsensalo and Antti Miettinen,Tamayo,P proposed the results of SOM algorithm which includes rat CNS dataset . This algorithm used to compare with high accuracy for best result when compared with other algorithms. This paper described about the expression levels with genes during rat central nervous system development over 9 time points. Mapping to 2-D space and statistical average, are used in this technique to give better results than average method.

In paper [11] Nikhil R. Pal, Kuhu Pal, James M. Keller, and James C. Bezde presented the work with FCM algorithm and EFC algorithm of clustering. The terms of quality clusters are used and their computational time is measured. Fuzzy reasoning algorithms was developed by using the best set of clusters, thus obtained.

In paper [12] Jacob Goldberger., Tamir Tassa discussed about the hierarchical clustering with number of attributes are added automatically. The attributes are based on the priority. Human cancer data set is used in this algorithm to give best result and time complexity.

In paper [13] K.Sasirekha, P.Baby discussed the comparison between the agglomerative algorithms that can be attempted according to different factors other than those considered with different algorithms. Comparing between the results of algorithms by using hierarchical cluster either it may normalizes or non normalizes data with different results.

In paper [14] G.J.McLachlan,R.W.Been and D.Peel presented Mixture model based approach in cluster technique and to provide a sound of mathematical based model. The aim was not to provide a detailed analysis of this dataset Breast cancer data but rather to highlight the potential role and useful of mixture model based approach to microarray expression data. In this mixture model based can identify various class and subclass among tissues on based gene expression level with best result.

In paper [15] Daxin Jiang Chun Tang Aidong Zhang presented Gene expression data generated by microarray experiments offer huge potential for advances in molecular biology and functional genomics. Gene expression data can be clustered on both genes and samples.

In paper [16] T. Deepika Dr. R. Porkodi discussed the wide application of microarray technologies which generates very large amounts of data. As a result, there is an increasing need for technology that can extract useful and rational, fundamental patterns of gene expression from the data. This paper reviewed clustering technology is one of the most useful and popular methods for identifying the patterns.

In paper [17] L.Boopathi1, D.Vijaybabu. presented DNA microarray technologies to monitor transcription levels with tens and thousands of genes in parallel. Gene expression data’s are generated by microarray experiments which offer tremendous potential for advances in molecular biology and functional genomics. This paper, reviewed both supervised clustering for gene expression and micro array technology which have been applied to gene expression data, and given best results.

III. CLUSTERING AGLORITHM

Clustering is one of the methods used to biological processes, particularly in the genomics level. Clearly, clustering can be used in many areas of biological data analysis. A good clustering approach may detect patterns or relationships in expression data. A clustering algorithm used for group together genes based on their expression profiles[18].In clustering, group together similar expression profiles of genes as expression data analysis are to identify the changing and unchanging levels of gene expression. A collection of data objects, the working principle of clustering is to divide the data objects into groups such that objects in the same group are similar. Objects in different groups should be dissimilar. Data belonging to one cluster are the most similar and data belonging to different clusters are the most dissimilar.

A. K-MEANS ALGORITHM

The k-Means algorithm is one of the simplest unsupervised learning algorithms that answer the well-known clustering problem . K-means is another clustering technique commonly used which is simple and a fast method. It is easy to implement and has small number of iterations[19].The K-means algorithm is a typical partition-based clustering method. given a pre-specified number K the algorithm partitions the data set into disjoint subsets which optimize the following objective function:

$$E= \sum_{i=1} \sum_{O\varepsilon Ci} |O-\mu_i|^2$$

Here O is a data object in cluster "Ci and µi is the centroid (mean of objects) of Ci Thus, the Objective function E tries to minimize the sum of the squared distances of objects from their cluster centers. The time complexity of K-means is O (i*k*n) where iis the number of interation and k is the number of clusters. however, the K-means algorithm forces each gene into a cluster, which may cause the algorithm to be sensitive to noise.

Typically the square error criterion is used, defined as,

$$E = \sum_{i=1}^{k} \sum_{p\epsilon Ci} |p-m_i|^2$$

Where E is the sum of the square error for all objects in the data set P is the point in space representing a given object Mi is the mean of cluster co For each object in each cluster, the distance from the object to its cluster center is squared and the distances are summed This criterion tries to make the resulting k clusters as compact and as separate as possible.

CLARANS was one of the first clustering algorithms that was developed specifically for use in data mining spatial data mining. CLARANS itself grew out of two clustering algorithms, PAM and CLARA , that were developed in the field of statistics.

PAM (Partitioning Around Medoids) is a "*K*-medoid" based clustering algorithm that attempts to cluster a set of *m* points into *K* clusters by performing.

CLARA (Clustering LARge Applications) is an adaptation of PAM for handling larger data sets. It works by repeatedly sampling a set of data points, calculating the medoids of the sample, and evaluating the cost of the configuration that consists of these sample-derived medoids and the entire data set.

B. *SELF ORGANIZING MAP*

Self organizing Map (SOM) is used for visualization and analysis of high-dimensional datasets. SOM facilitate presentation of high dimensional datasets into lower dimensional ones, usually 1-D, 2-D and 3-D. It is an unsupervised learning algorithm, and does not require a target vector since it learns to classify data without supervision. A SOM is formed from a grid of nodes or units to which the input data are presented. Every node is connected to the input, and there is no connection between the nodes.[20] The Self-Organizing Map (SOM) was developed by Kohonen on the basis of a single layered neural network. SOFMs were developed by observing how neurons work in the brain and in ANNs The firing of neurons impact the firing of other neurons that are near it Neurons that are far apart seem to inhibit each other Neurons seem to have specific non-overlapping tasks The term self-organizing indicates the ability of these NNs to organize the nodes into clusters based on the similarity between them Those nodes that are closer together are more similar than those that are far apart The most common example of a SOFM is the Kohonen Self organizing map It is used extensively in commercial data mining products to perform clustering There is one input layer and one special layer which produces output values that compete Multiple outputs are created and the best one is chosen This extra layer is not technically either a hidden layer or an output layer, so we refer to it here as the competitive layer Nodes in this layer are viewed as a two-dimensional grid of nodes Each input node is connected to each node in this grid[21]. Propagation occurs by sending the input value for each input node to each node in the competitive layer As with regular NNs, each arc has an associated weight and each node in the competitive layer has an activation function Thus each node in the competitive layer produces an output value, and the node with the best output wins the competition and is determined to be the output for that input An attractive feature of Kohonen nets is that the data can be fed into the multiple competitive nodes in parallel Training occurs by adjusting weights so that the best output is even better the next time this input is used "Best" is determined by computing a distance measure. A common approach is to initialize the weights on the input arcs to the competitive layer with normalized values The similarity between output nodes and input vectors is then determined by the dot product of the two vectors Given an input tuple to X=<x1,…xh> and weights on arcs input to a competitive node i as w1i,…whi, the similarity between X and i can be calculated by

$$sim(X,i) = \sum_{j=1}^{H} x_{\mathrm{j}}\, \mathrm{w}_{\mathrm{ji}}$$

C. *FUZZY C-MEANS ALGORITHM*

Fuzzy C-means is an overlapping clustering algorithm Fuzzy c-means allows data points to be assigned into more than one cluster each data point has a degree of membership (or probability) of belonging to each cluster. Let x_i be a vector of values for data point g_i. Initialize membership $U^{(0)}$ = [u_{ij}] for data point g_i of cluster cl_j by random At the *k*-th step, compute the fuzzy centroid $C^{(k)}$ = [c_j] for *j* = 1, .., *nc*, where *nc* is the number of clusters, using[22]

$$c_j = \frac{\sum_{i=1}^{n} (u_{ij})^m x_i}{\sum_{i=1}^{n} (u_{ij})^m}$$

where *m* is the fuzzy parameter and *n* is the number of data points. Update the fuzzy membership $U^{(k)}$ = [u_{ij}], using

$$u_{ij} = \frac{\left(\frac{1}{\|x_i - c_j\|}\right)^{\frac{1}{(m-1)}}}{\sum_{j=1}^{n_c}\left(\frac{1}{\|x_i - c_j\|}\right)^{\frac{1}{(m-1)}}}$$

If $||U^{(k)} – U^{(k-1)}|| < \varepsilon$, Determine membership cutoff For each data point g_i, assign g_i to cluster cl_j if u_{ij} of $U^{(k)} > \alpha$ Allows a data point to be in multiple clusters.The representation of the behavior of genes usually are involved in multiple functions.Need to define c, the number of clusters.Need to determine membership cutoff value Clusters are sensitive to the initial assignment of centroids.

D. HIERARCHICAL CLUSTERING

Hierarchical cluster builds a cluster hierarchy in other words a tree of cluster also known as dendrogram[23].Organize elements into a tree, leaves represent genes and length of the paths between leaves represents distances between genes. Similar genes lie within same sub trees. The branches of a dendrogram not only record the formation of the clusters but also indicate the similarity between the clusters. By cutting the dendrogram at some level, we can obtain a specified number of clusters. By reordering the objects such that the branches of the corresponding dendrogram do not cross, the data set can be arranged with similar objects placed together. There two methods for hierarchical clustering.

Agglomerative: start with every element in its own cluster, and iteratively join clusters together.

Divisive: start with one cluster and iteratively divide it into cluster.

Hierarchical clustering algorithms can be further divided into agglomerative approaches and divisive approaches based on how the hierarchical dendrogram is formed. Agglomerative algorithms (bottom-up approach) start with every element in its own cluster, and iteratively join clusters together. Divisive algorithms (top-down approach) start with one cluster and iteratively divide it into smaller clusters.UPGMA (Unweighted Pair Group Method with Arithmetic Mean) and adopted a method to graphically represent the clustered data set. In this method, each cell of the gene expression matrix is colored on the basis of the measured fluorescence ratio, and the rows of the matrix are re-ordered based on the hierarchical dendrogram structure and a consistent node-ordering rule. After clustering, the original gene expression matrix is represented by a colored table a cluster image where large contiguous patches of color represent groups of genes that share similar expression patterns over multiple conditions. split the genes through a divisive approach, called the deterministic-annealing algorithm[24].First, two initial cluster centroids c_j =1,2.. Were randomly defined.The expression pattern of gene k was represented by a vector $\overrightarrow{g_k}$ and the probability of Gene k belonging to cluster j was assigned according to a two component Gaussian model.

The cluster centroids were $p_j(\overrightarrow{g_k})\exp(-\beta|\overrightarrow{g_k} -_{c_j}|^2)/\sum_j \exp(-\beta|\overrightarrow{g_k} -c_j|^2)$ The cluster centroids were recalculated by$c_j = \sum_j \overrightarrow{g_k} p_j (\overrightarrow{g_k})/\sum_k p_j (\overrightarrow{g_k})$ an iterative process (the EM process (the EM algorithm) w as then applied to solve p_j and c_j.For β =o there was only one cluster C1 = C2.when β was increased in small steps until a threshold was reached, two distinct, converged centroids emerged. The whole data set was recursively split until each cluster contained only one gene.Hierachical clustering linkage there three type.

Single Linkage: Join clusters whose distance between closest genes is smallest.

Complete Linkage: Join clusters whose distance between furthest genes is smallest.

Average Linkage: Join clusters whose average distance is the smallest.

BIRCH[24]can typically find a good clustering with a single scan of the data and improve the quality further with a few additional scans. It is also the first clustering algorithm to handle noise effectively.

Describe some representative hierarchical clustering algorithms CURE (Clustering Using Representatives) is a clustering algorithm that uses a variety of different techniques to create an approach which can handle large datasets, outliers, and clusters with non-spherical shapes and non-uniform sizes. CURE represents each cluster by a certain number of points that are generated by selecting well-scattered points and then shrinking them toward the cluster centroid by a specified fraction. It uses a combination of random sampling and partition clustering to handle large databases.

ROCK (RObust Clustering using links) is a clustering algorithm for data with categorical and Boolean attributes. It redefines the distances between points to be the number of shared neighbors whose strength is greater than a given threshold and then uses a hierarchical clustering scheme to cluster the data.

E. MODEL BASED CLUSTERING

These model attempt to optimize the fit between the given data and some mathematically model[25]. These method find character description for each group, where each group represent a concept or class.

The goal is to estimate the parameters **(-)**$\{\boldsymbol{\theta}\mathbf{i}|\mathbf{1} < i < k\}$ and $\boldsymbol{\Gamma = \gamma_r^i \,|\leq i \leq k, 1 \leq r \leq n\}}$ That maximize the likelihood Lmin(-) $\Gamma=\sum_{i=1}^{k} \gamma_r^i$ θi) where n is the number of data object K is number of component Xr data object (i.e., a gene expression pattern),fi(Xr | θi) is the density function of x_r of component c_i with some unknown set of parameter θi model parameter and γ_r^i hidden parameters) represents the probability that belongs to c_i. usally the parameter(-) and Γ are estimated by the EM algorithm. The EM algorithm iterates between Expectation (E) steps and Maximization (M) steps. In the E step, hidden parameters are conditionally estimated from the data with the current estimated (-).In the M step, model parameters (-) are estimated so as to maximize the likelihood of complete data given the estimated hidden parameters.[26]When the EM algorithm converges, each data object is assigned to the component (cluster) with the maximum conditional probability. An important advantage of model-based approaches is that they provide an estimated probability γ_k^i that data object i will belong to cluster gene expression data are typically highly-connected there may be instances in which a single gene has a high correlation with two different clusters. Thus, the probabilistic feature of model-based clustering is particularly suitable for gene expression data.

[27]However, model-based clustering relies on the assumption that the data set fits a specific distribution The modeling of gene expression data sets, in particular, is an ongoing effort by many researchers, and, to the best of our knowledge, there is currently no well-established model to represent gene expression data. Commonly used data transformations and assessed the degree to which three gene expression data sets fit the multi-variant Gaussian model assumption. The raw values from all three data sets fit the Gaussian model poorly and there is no uniform rule to indicate which transformation would best improve this fit.

IV. CLUSTER VALIDATION

Clustering algorithms which partition the dataset based on different Clustering. For gene expression data, clustering results in groups of genes, groups of samples with a process. However, different clustering algorithms, or even a single clustering algorithm using different parameters, generally in different sets of clusters. Cluster validation is the process of assessing the quality and reliability of the cluster sets derived from various clustering processes Generally, cluster validity has six aspects. First, the quality of clusters can be measured in terms of hubert's, Dunn Index, Simple matching coefficient, Nmimeasure Purity and Silhouette coefficient.

A. Hubert's Statistics

Let us consider two n×n proximity matrices X (i, j) and *Y* (*i*, *j*) on the same *n* genes. X (i, j) denotes the observed distance of genes *i* and *j* and Y (i, j) is defined as follows[28]

$$\mathrm{Y(i,i)} = \begin{cases} \text{if gene i and j are cluster in the same cluster} \\ \text{0 otherwise} \end{cases}$$

The serial correlation between the matrices *X* and *Y* is represented by Hubert's Statistic ʢand it is defined as:

$$ʢ=\frac{1}{m} = \textstyle\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left[\frac{x(i,j) - \overline{x}}{\sigma x}\right] \left[\frac{y(i,j) - \bar{y}}{\sigma y}\right]$$

Where *M* is the number of entries in the double sum and it is given by M= , and σx and σy denote the sample standard deviations. The sample means of the entries of matrices *X* and *Y* is denoted by *X* and *Y* .

B. Dunn Index

Dunn index identifies all sets of clusters that are dense and well separated. Let us consider any partition *U* ↔ *X* : *Xi* ∪....*Xi* ∪...*Xc* , where *i X* represents the *i*th cluster of the partition. The Dunn's validation index, D, is defined as equation[29]

$$D(u) = \min_{1<i\leq c}\left\{\min\left\{\frac{\delta(x_i,x_j)}{max\{\Delta(x_k)\}}\right\}\right\}$$

Here δ (Xi,Xj)indicates the intercluster distance between clusters *Xi* and *Xj* . (Xk) indicates the intracluster distance of cluster *Xk* and *c* is the number of clusters of partition *U* .

C. Simple matching coefficient

Let us consider two n·n binary matrices P =[P(i, j)] and *Q* = [Q(i, j)] on the same set of n data.[30] Let the matrices *P* and *Q* indicates two distinct clustering results and the general form is defined as follows:

$$\mathrm{Y(i,i)} \begin{cases} \text{if gene i and j are cluster in the same cluster} \\ \text{0 otherwise} \end{cases}$$

Let us consider an association table of the matrices *P* and *Q* as give in the Table (1). The number of entries on which both *P* and*Q* has value 1 is indicated by 'a'. The number of entries on which *P* has value 1 and *Q* has value 0 is indicated by 'b'and so on.

Matrix of Simple Matching Coefficient

Matrix P	Matrix Q	
	1	0
1	A	B
0	C	D

The simple matching coefficient is given by

$$\mathrm{S} = \frac{a+d}{a+b+c+d}$$

that is total number of matching entries divided by total number of entries.

The Jaccard coefficient is given by

$$\mathrm{S} = \frac{\mathrm{a+d}}{\mathrm{a+b+c}}$$

Here, the negative matches ' *d* ' are not considered.

D. Nmimeasure

Is called Normalized Mutual Information (NMI). The NMI of two labeled objects can be measured as[31]

$$\mathrm{NMI(X,J)} \quad \frac{\mathrm{I(x,y)}}{\sqrt{H(X)+H}\mathrm{(Y)}}$$

Where I (X,Y) the mutual information between two random variables X and Y and H(x) denotes the entropy of X,X will be consensus clustering while Y will be the true label.

E. Purity

Purity is very similar to entropy. We calculate the purity of a set of clusters. First, we cancel the purity in each cluster. For each cluster, we have the purity $p_j = \frac{n_j}{n} \max \top n^{i}_{j}$ is the number of objects in j with class label I In other words, p_j is a fraction of the overall cluster size that the largest class of objects assigned to that cluster represents.[32] The overall purity of the clustering solution is obtained as a weighted sum of the individual cluster purities and given as:

$$\sum_{j=1}^{m} p_j = \frac{n_j}{n} p_j$$

Were n_j is the size of cluster j,m is the number of clusters and n is the total number of objects.

F. Silhouette coefficient

The popular method of silhouette coefficient combines both cohesion and separation.

The value of the silhouette coefficient can very between -1 and 1.A negative value is undesirable because this correspond case in which a_i the average distance the point in the cluster in greater then b_i the minimum average distance to point in another cluster.

Cohesion a(x): average distance of x to all other vectors in the same cluster.

Separation b(x): average distance of x to the vectors in other clusters. Find the minimum among the clusters.

silhouette s(x):[33]

$$s(x) = \frac{b(x) - a(x)}{\max\{\ a(x), b(x)\}}$$

s(*x*) = [-1, +1]: -1=bad, 0=indifferent, 1=good Silhouette coefficient (SC):

$$SC = \frac{1}{N} \sum_{i=1}^{N} s(x)$$

V. COMPARSION OF CLUSTERING ALGORITHM USED FOR MICRO ARRAY GENE EXPRESSION DATA

This paper also presents the comparative study on the above mention clustering algorithms based on their accuracy and demerits. The outcome of study shows that the clustering algorithm gives better accuracy for k-means [9] and then other algorithm in clustering.

VI. CONCLUSION

Clustering algorithms are useful for identifying biologically relevant groups of genes and sample clustering techniques are essential in the data mining process to reveal natural structure and identifying pattern in the data sets. From the above context identified that, for microarray clustering techniques k-means algorithm is used widely. The quality of clustered data measured using different validation parameters as Hubert's statistics, Dunn's Index, Simple matching coefficient, Nmimeasure and Purity and Silhouette coefficient the similarity measures are used extensively in most of recent studies on gene expression data. The future research direction towards the gene ontology (GO) terms which comes under microarray gene expression data and k-means clustering algorithm is hybridized with other algorithms to achieve quality and accuracy.

Table 1 Comparison of clustering algorithm

S.NO	AUTHOR NAME	ALGORITHM	DATASET	OUTCOME	DEMERTICS
1	A.Dharmarajan.,T. Velmurugan [9]	Partition method K-means	LC.arff	Time complexity is high.	Difficult to compare a quality of cluster.
2	Jyrki Joutsensalo and Antti Miettinen, Tamayo, P.[10]	SOM	Rat CNS	High Accuracy.	Computational process is high.
3	Nikhil R. Pal, Kuhu Pal, James M. Keller, James C. Bezdek[11]	Fuzzy c-means	Blood cancer data	High times complexity with best result	Iteration process is expensive
4	Jacob Goldberger., Tamir Tassa [12]	Hierachical clustering Divisive	Human cancer data	Accurate Result.	Times process is slow.
5	K.Sasirekha,P.Baby[13]	Agglomerative	Yeast cell cycle	High degree accuracy.	Generate different Results in computational Process.
6	McLachlan, G.J., Bean R.W. and Peel D. [14]	Model based Clustering	Breast cancer data	Iteration Process is high.	Hard to estimate the number of clusters.

REFERENCE

[1] Mann AK, Kaur n. Survey paper on clustering techniques. Ijsetr. 2013 apr; 2 (4):803–6.

[2] Kaufman, L. and Rousseeuw, P.J. Finding Groups in Data: an Introduction to Cluster Analysis. JohnWiley and Sons, 1990

[3] Brazma, Alvis and Vilo, Jaak. Mini review: Gene expression data analysis. Federation of European Biochemical societies, 480:17–24, June 2000.36

[4] Siedow, J. N. Meeting report: Making sense of microarrays. Genome Biology,2(2):reports 4003.1–4003.2, 2001.

[5] Derisi, J.l. , Iyer, V.R., and brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science, pages 680–686, 1997.

[6] Smet, Frank De, Mathys, Janick, Marchal, Kathleen, Thijs, Gert, Moor, Bart De and Moreau, Yves. Adaptive quality-based clustering of gene expression profiles. Bioinformatics, 18:735–746, 2002

[7] R.Sharan, R.Elkon, R.Shamir, Cluster Analysis and its Applications to Gene Expression Data.

[8] Smet, Frank De, Mathys, Janick, Marchal, Kathleen, Thijs, Gert, Moor, Bart De and Moreau, Yves. Adaptive quality-based clustering of gene expression profiles. Bioinformatics, 18:735–746, 2002.

[9] In A.Dharmarajan.,T.Velmurugan k-means algorithm.

[10] Jyrki Joutsensalo and Antti Miettinen,Tamayo, P. And others, interpreting patterns of gene expression with self organizing Maps, pnas 96, p.2907--2912, 1999

[11] Nikhil R. Pal, Kuhu Pal, James M. Keller, and James C. Bezdek fuzzy c-means algorithm Springer, Berlin, Heidelberg, 1995.

[12] Jacob Goldberger., Tamir Tassa Hierarchical clustering algorithm 2nd edution springer-verlage 1998.

[13] K.Sasirekha, P.Baby.,Agglomerative algorithm 2nd edution springer-verlage 1998.

[14] McLachlan, G.J., Bean R.W. and Peel D. A mixture model-based approach to the clustering of microarray expression data. Bioinformatics, 18:413–422, 2002.

[15] Daxin Jiang Chun Tang Aidong Zhang et.al application of data mining in bioinformatics book.

[16] T. Deepika*, Dr. R. Porkodi Cluster Analysis and micro array technology bioinformatics.

[17] L.Boopathi1, D.Vijaybabu Cluster Analysis and its Applications to Gene Expression Data

[18] Khalid Raza application of data mining in bioinformatics book.

[19] Hartigan j.a clustering algorithm wiley,new york Hartigan j.a and Wong m.a(1979) a k-means clustering.

[20] Teuvo Kohonen. *Self-Organizing Maps*. Springer, Berlin, Heidelberg, 1995.

[21] Tamayo, P. And others, interpreting patterns of gene expression with self organizing Maps, pnas 96, p.2907--2912, 1999.

[22] J.C. Bezdek, J. Hathaway, M.J. Sabin, and W. T. Tucker,"Convergence Theory for Fuzzy C-Means: Counterexamples and Repairs", IEEE Trans, 17, pp. 873-877, Sept./Oct. 1987.

[23] Rose, K. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. Proc. IEEE, 96:2210–2239, 1998.

[24] Zhang, T., Ramakrishnman, R., and Linvy, M. (1996). BIRCH: An Efficient Method for Very Large Databases. *ACM SIGMOD*, Montreal, Canada.

[25] Fraley C. and Raftery A.E. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. The Computer Journal, 41(8):578–588, 1998.

[26] McLachlan, G.J., Bean R.W. and Peel D. A mixture model-based approach to the clustering of microarray expression data. Bioinformatics, 18:413–422, 2002.

[27] Ghosh, D. and Chinnaiyan, A.M. Mixture modelling of gene expression data from microarray experiments. Bioinformatics, 18:275–286, 2002.

[28] N. Bolshakova and F. Azuaje, (2003)"Cluster validation techniques for genome expression data", journal signal processing, special issue, genomic signal processing, vol 83, issue 4, pp 825-833.

[29] Yeung, K.Y., Haynor, D.R. and Ruzzo, W.L., (2001) "Validating Clustering for Gene Expression Data. Bioinformatics, Vol.17(4):pp309–318.

[30] Ankerst, Mihael, Breunig, Markus M., Kriegel, Hans-Peter, Sander, J_rg. OPTICS: Ordering Points To Identify the Clustering Structure. Sigmod, pages 49–60, 1999.

[31] Rendon L. Eréndira, Garcia Rene, Abundez Itzel, Gutierrez Citlallih, et. al. Niva:A Robust Cluster Validity. 2 th. WSEAS Int.Conf. Scientific Computation and Soft Computing, Crete, Greece, 2002, pp. 209-213.

[32] Legány C., Juhász S.and A. Babos. Cluster Validity Measurement Techniques Proceeding of the 5th.WSEAS Int.Conf. on Artificial, Knowledge Engineering and Data bases, Madrid, Spain, February 15 -17,2006, pp. 388-393.

[33] Kovács F. and R. Ivancsy. Cluster Validity Measurement for arbitrary Shaped clustering Proceeding of the 5th. WSEAS Int.Conf. on Artificial, Knowledge Engineering and Data bases.