International Journal of Engineering & Technology, 7 (2.21) (2018) 201-205

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A performance analysis of clustering based algorithms for the microarray gene expression data

K. Yuvaraj[1*], D. Manjula[2]

[1]*Assistant Professor, Department of Computer Science, Karpagam University, Coimbatore.*
[2]*Assistant Professor, Department of Computer Science, Karpagam University, Coimbatore.*

Abstract

Current advancements in microarray technology permit simultaneous observing of the expression levels of huge number of genes over various time points. Microarrays have obtained amazing implication in the field of bioinformatics. It includes an ordered set of huge different Deoxyribonucleic Acid (DNA) sequences that can be used to measure both DNA as well as Ribonucleic Acid (RNA) dissimilarities. The Gene Expression (GE) summary aids in understanding the basic cause of gene activities, the growth of genes, determining recent disorders like cancer and as well analysing their molecular pharmacology. Clustering is a significant tool applied for analyzing such microarray gene expression data. It has developed into a greatest part of gene expression analysis. Grouping the genes having identical expression patterns is known as gene clustering. A number of clustering algorithms have been applied for the analysis of microarray gene expression data. The aim of this paper is to analyze the precision level of the microarray data by using various clustering algorithms.

Keywords: *Microarray Technology, Gene Expression Data, Clustering Algorithms.*

1. Introduction

In every living creature, the genetic data is preset in information units recognized to as genes. The complete set of genes of an organism is standard to as its genome. The conception of microarray experiments is essential to obtain, analyze, save and share the genetic information by other researchers. The microarray technology allows researchers and medical experts to measure the expressiveness of thousands of genes of a tissue model in a single experiment that also helps to determine the ailment gene. Though, the data which is generated by the experiments cannot be analyzed manually because of their huge size also high complexity. In the case, Microarrays are organized it potential to concurrently observe the expression profiles of thousands of genes under different experimental conditions [1]. Determination of co-expressed genes as well as coherent patterns is the central aim in microarray gene expression data analysis and this is a significant task in bioinformatics research. These clustering algorithms have ability to determining biologically related groups of genes and samples. The DNA microarray design is illustrated in following Fig.1.

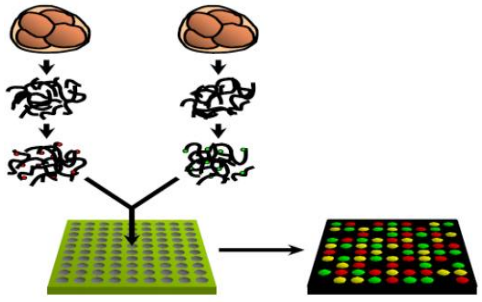

Fig.1: DNA microarray design

Genetic data are being created at an outstanding rate. Enormous quantities of such data have been produced by the global scaled human genome project. Data mining is an astounding device for separating important data from enormous databases [2]. The point of Data mining apparatus in gene expression microarray data as follows.

S.NO	Aims
1	To arrange the data in best way, this allows researchers to process existing information as well as to include new entries as they are produced.
2	To improve the tools which aid in the analysis of data
3	To use existing tools for analyze the specific systems in detail, in order to get new biological gene expression insights.

Data Mining is commonly described as Knowledge Discovery in Databases (KDD). It refers to the significant access of determining applicable, original, potentially valuable and eventually understandable patterns of data. Data mining creates the patterns in a particular representational form such as classification rules, decision trees, clustering and regression models.

Cluster analysis has developed into an imperative element of gene expression analysis.

Grouping the genes having similar patterns is known as gene clustering.

A number of clustering algorithms have been applied for the investigation of microarray gene expression data. Therefore, the

main goal of this paper is to analyze better clustering method for microarray gene expression data [3].

2. Microarray gene expression data

In the rapid technological growth in the field of genomics, presence of a huge number of genes by means of no clear sequence homology with previously categorized genes and understanding genes activities is a major challenge in the future years. Micro arrays have obtained a significant role in this challenging field, since these consist of an ordered set of thousands of various DNA sequences, which can be calculated by DNA along with RNA variations [4]. Microarray technique was developed from Southern blotting. At this point, fragmented DNA is connected to a substratum and then explored with a recognized gene.

Microarrays are facilitate the quantitative learn about thousands of genes concurrently from a single sample of cells. It is exploiting in lots of applications but commonly used in expression profiling [5]. Microarray has emerged as a successful and widely used tool for resolves a wide range of problems such as the classification of disease subtypes and tumors in medical research. The below figure describe the design and implementation of the Microarray Gene Expression Data.

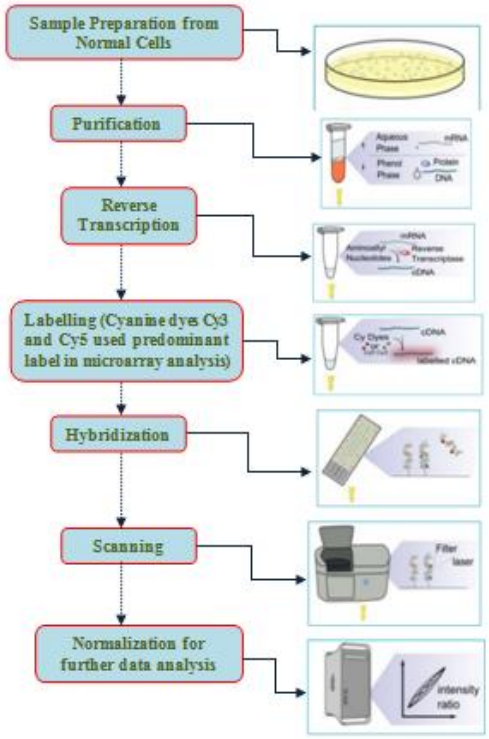

Fig.2: Implementation of microarray gene expression data

A living creature internal characteristic, function, and progression are controlled by the DNA, RNA as well as protein. Gene expression is a foremost process in the cell. Gene expression levels can vary among different cells, tissues or points in time. Therefore, the microarray innovation encourages an analyst to dissect the declaration of thousands of qualities in a solitary examination as well as provide quantitative measurements of the differential expression of these genes.

3. Clustering Techniques

In clustering techniques, it segments a huge set of data into subsets called as clusters. Every cluster is a group of data objects. They are similar to one another in the same cluster, other than dissimilar to objects in other clusters. The main objective of clustering analysis is to divide the large set of data into homogenous and different groups also reduce the complexity of data. The below figure demonstrates the clustering process.

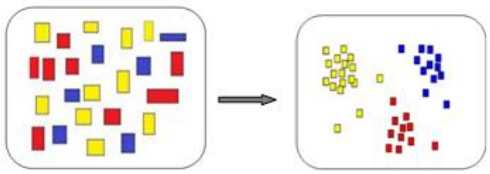

Fig.3: Clustering Process

Clustering techniques have confirmed which is supportive to identify the quality capacity, quality direction, cell forms and also subtypes of cells. Co-communicated qualities which mean qualities with same articulation designs are clustered together with the same cell capacities. In this kind of approach, it may helpful to recognize the processes of many genes for which information has not been previously obtainable [6]**.** Clustering technique is useful tool to determining structures and associating patterns in gene expression data.

Clustering algorithms can be applied to organize, model, categorize and compress data from large data set. This section consists of the different clustering algorithms used in microarray gene expression data. The algorithms are listed below,

S.NO	Clustering Algorithms	Example
1	Hierarchical Clustering	Agglomerative clustering
2	Partition Based Clustering	K-means Clustering
3	Model Based Clustering	SOM (Self-organizing map)
4	Density Based Clustering	DBSCAN (Density Based Clustering)

1. Hierarchical Clustering

In microarray analysis, clustering is a most significant method. Hierarchical clustering merge data objects into clusters and find out larger clusters from those clusters finally creating a hierarchy of clusters. It grouping the data objects into a tree of cluster. Hierarchical clustering techniques can be categorized into agglomerative and disruptive various leveled clustering. This classification in view of whether the progressive decay is created in a base up or top-down approach [7].

Agglomerative method

Agglomerative clustering method is more popular than divisive method. The proper implementation of this method is helpful for gene expression data processing in microarray techniques. This method starts based on the bottom-up approach. In this method, each objects forming a separate group and merging the objects or groups that are close to one another. This process is carried out until there is only a single cluster [8]. The following figure demonstrates the progression of agglomerative clustering,

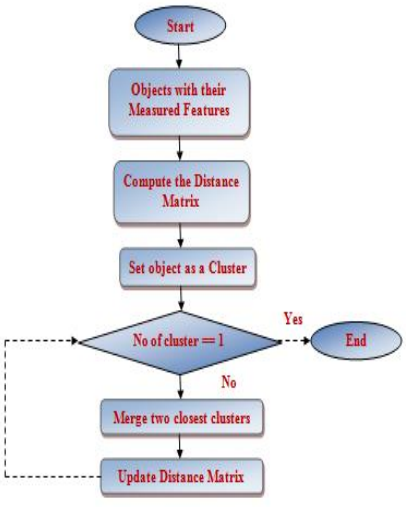

Fig.4: Flowchart of agglomerative clustering algorithm

A set N of objects {n_1, n_2... n_n}, distance function between two clusters is D and A set C of clusters {c_1, c_2...c_n}...,. The related algorithm is shown below, c_q' is the clusters obtained at each stage of hierarchical clustering,

```
For i = 1 to n
   ci = {ni}
end for
C = {c1,....,cn}
N = n + 1
while N == 1 do
        find two clusters ci and cj
        (cmin1, cmin2) = D (ci,cj) for all ci,cj
in C
        merge ci,cj into a single cluster cq
        remove cmin1 and cmin2 from C
        update the distance matrix 'D' of
new clusters
        N = N + 1
end while
```

2. Partition based clustering

Partition Based Clustering algorithm minimizes a given clustering measure by iteratively repositioning data points among clusters until a finest partition is attained. Partitioning methods consist of the K-means and K-medoids methods. The K-means calculates the distance among groups by their centroids simultaneously the K-medoids calculates the distances between groups by their center of gravity [9].

K-means clustering method

K-Means is comparatively an effective method in partition based clustering algorithm as well the most broadly used among all clustering algorithms because of its ease and efficiency. The objective of this algorithm is to identify groups in the data by the number of groups correspond to the variable K. This method functions iteratively to allocate every data object to one of K groups depends on the features which can be provided. Data objects are clustered based on feature similarity. The below figure illustrates the function of K-means clustering method.

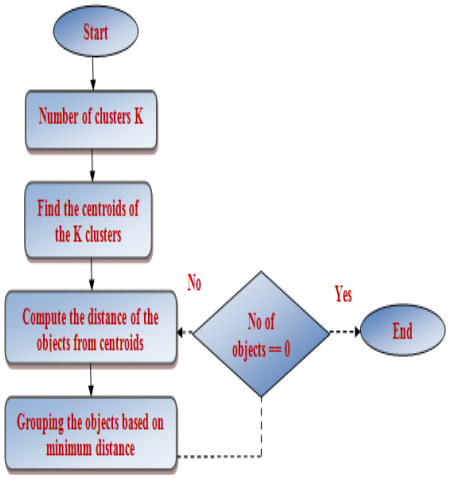

Fig.5: Flowchart of K-means clustering algorithm

K is a pre-specified number. The algorithm separates the data set into K disjoint subsets that optimize the following objective function:

$$E = \sum_{i=1}^{K} \sum_{a \in c_i} [a - \mu_i]^2$$

At this point, E is the objective function 'a' is a data object in cluster C_i 'μ_i' is the centroid (mean of objects) of C. The objective function 'E' attempts to reduce the sum of the squared distances of objects from their cluster centers [10].

3. Model based clustering

In model based clustering, the data can be produced by a model as well as attempt to restore the unique model from the data. The model that recuperates from the data at that point characterizes clusters and a task of reports to clusters. Self Organizing Maps (SOM) is the by and large utilized simulated neural system based unsupervised learning technique that is identified with K-implies.

Self Organizing Maps (SOM)

SOM is a unique instrument in the looking at phase of data mining. This is a neural system master of mapping high dimensional data onto a low dimensional lattice. Those particularly comparable data components are mapped together as nearly as could be expected under the circumstances. The Euclidean separation was normally connected to contrast each hub weight vector and a data test. Minimization of the trouble of the data space was an alluring property while managing substantial data sets that is every now and again the instance of genomic data. This method can be efficiently utilized to visualize as well as explore the properties of the data [11]. The following figure shows the process of self organizing maps.

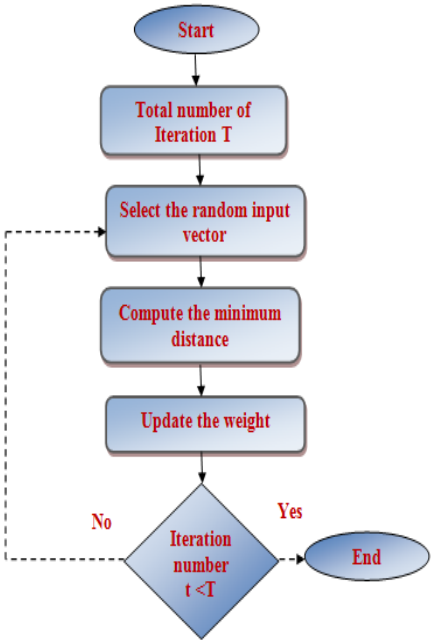

Fig.6: Flowchart of K-means clustering algorithm

The steps of the SOM algorithm can be summarised as follows,

- *Initialization*: Select random values used for the initial weight vectors w_j
- *Sampling*: Build a model training input vector x from the input space.
- *Matching*: Identify the winning neuron I(x), which has weight vector neighbouring to the input vector, i.e. the least value of $d_j(\mathbf{x}) = \sum_{i=1}^{D} (x_i - w_{ji})^2$
- *Updating*: Apply the weight update equation $\Delta w_{ji} = \eta(t)\, T_{j,I(\mathbf{x})}(t)\, (x_i - w_{ji})$ here $T_{j,I(\mathbf{x})}(t)$ is a Gaussian neighbourhood and η (t) is the learning rate.
- *Continuation* – continue returning to step 2 until the feature map stops changing [12].

4. Density based clustering

Density based clustering algorithms are created to determine arbitrary shaped clusters. In this method, a cluster is considered as a region in that the density of data objects exceeded a threshold. Density based clustering method can be classified into Density-

based Spatial Clustering of Applications with Noise (DBSCAN) and Shared Nearest Neighbour (SSN) [13].

DBSCAN

The DBSCAN algorithm based on a density based view of clusters. Clusters are determined by appearing at the density of points. Regions with a highest density of points represent the existence of clusters, while those with a least density of points point out clusters of noise or else clusters of outliers. This algorithm is mainly suitable for large datasets with noise and it is capable to find clusters with different sizes and shapes [14]. This algorithms includes the following steps,

- Randomly select a point P.
- Recover every points density-reachable from P wrt Eps and MinPts
- If P is a core point, a cluster is created.
- If P is a border, no points are density reachable form P and DBSCAN visits the subsequent point of the database.
- Keep on the process until every of the points have been processed.

4. Proposed modified clustering technique

This paper proposes a Modified Clustering Algorithm. The major objective of calculation is to set two straightforward data structures to keep up the names of cluster and in addition the separation of each the data articles to the nearby cluster amid the each cycle, which can be utilized as a part of next emphasis. It measure the separation among the present data protest and the new cluster focus, if the deliberate separation is lesser than or equivalent to the separation to the past focus, the data question remains in its cluster, which was allocated to in past emphasis. Along these lines, No compelling reason to gauge the separation from introduce data protest the past k-1 clustering focuses, this keep up the getting to time to the k-1 cluster focuses. Else, it should quantify the separation from the present data question each k cluster focuses likewise distinguish the neighboring cluster focus. In this technique, this point is doled out to the closest cluster focus and after that independently record the separation to its inside [15]. In every cycle, a few data focuses still remain in the first cluster, it implies that a few sections of the data focuses won't be estimated, sparing an aggregate time of figuring the separation, it improving the effectiveness of the calculation.

The process of the Modified Clustering algorithm is illustrated below. In this algorithm, input is the quantity of wanted clusters 'K', Dataset 'D' containing 'n' data questions as D= {d1, d2… dn}, di is an arrangement of characteristics on one data point { a1,a2,… am }. At long last, the yield of this calculation is set of K clusters.

Stage 1: The different sub tests {s1, s2. . . sj} are drawn from the first dataset.

Stage 2: Executes the stage 3 for m=1 to n.

Stage 3: Use the joined approach for sub test.

Stage 4: At each set, get the inside point as the underlying centroid.

Stage 5: Measure the separation among each datum point to the whole introductory centroids.

Stage 6: Find out the neighboring centroid and additionally allot to

Stage 7: Select littlest estimation of least separation from cluster focus criteria.

Step8: Apply new calculation again to the dataset D for K clusters.-

Step9: Merge two neighboring clusters into single cluster.

Step10: Recomputed the new cluster community for the combined cluster still the quantity of clusters decreases into k.

V. Results and dialog

This paper chooses quality articulation dataset storehouse of machine learning databases to explore the effectiveness of the Modified Clustering Algorithm (MCA) and the standard calculations like Agglomerative clustering, K-implies Clustering, SOM and DBSCAN [16]. The tests have been done to analyze the execution of the proposed calculation. The clustering calculations are connected to datasets on WEKA data mining instrument. In these trials, time involved for each clustering calculation and their exactness is figured [17]. The accompanying table gives depiction of the datasets.

Table 1: Data set size

Dataset	No of attributes	No of records
Gene Expression Data	17	5713

Table 2: Comparative analysis clustering algorithms

Factors	Agglomerative	K-means	SOM	DBSCAN	MCA
Execution Time(ms)	1430	1279	307	985	1011
Error Rate	0.7385	0.7465	0.7158	0.6995	0.2527
No of Clusters	6	6	6	6	4

The above table demonstrates the constraints of the standard Agglomerative, K-means, SOM and DBSCAN clustering calculation. The computational trouble of the standard calculation is high in light of the fact that reassign the data focuses at number of times in each emphasis. It lessens the proficiency of standard clustering calculation. In this paper, the proposed calculation introduces a simple and powerful approach for doling out data focuses to clusters. From the experimental results, modified clustering algorithm can develop the execution time and SOM algorithm improves the work efficiency on large dataset.

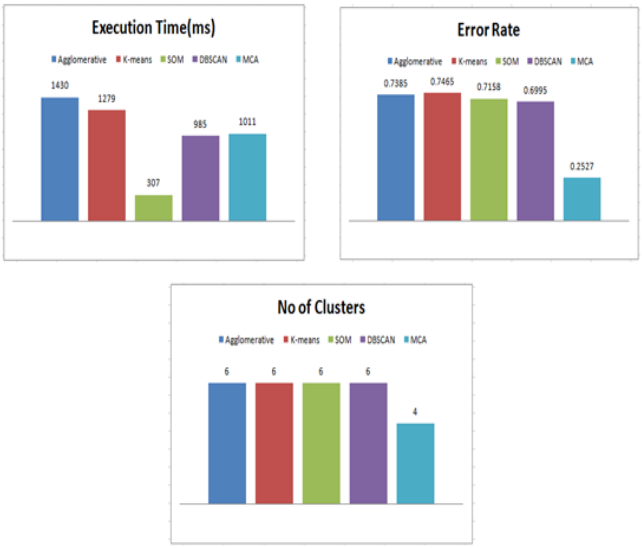

Fig.7: Performance comparison of clustering based algorithms

5. Conclusion

Managing huge data is major issues in the field of data analysis. In such analysis, clustering techniques play an important role to handle with the large-scale data like microarray gene expression dataset. Microarrays are prepared it feasible to concurrently observe the expression profiles of thousands of genes under various experimental conditions. There are many clustering techniques are available to handle but the selection of an appropriate method for a particular dataset is difficult to predict. Agglomerative, K-means, SOM, DBSCAN and proposed modified clustering algorithms are applied to the gene expression data. The experimental results shows, modified clustering algorithm have to increase the execution time and SOM algorithm can develops the work efficiency on gene expression dataset.

Reference

[1] Sherlock G, "Analysis of large-scale Gene Expression Data", *Curr. Opin. Immunol.*, Vol.12, (2000), pp. 201–205,.

[2] Segal E, Friedman N, Kaminski N, Regev A & Koller D, "From signatures to models: understanding cancer using microarrays", *Nature Genetics*, Vol.37, (2005), pp.38-45,.

[3] Mann AK & Kaur N, "Survey paper on clustering techniques", *Ijsetr*, Vol.2, No. 4, (2013), pp.803–806.

[4] Lipschultz RJ, Fodor SPA, Gingeras TR & Lockhar DJ, 'High density synthetic oligonucleotide arrays', *Suppl Nat. Genet.*, Vol.21, (1999), pp.20-24.

[5] Bowtell DDL, 'Options available from start to finish- for obtaining expression data by microarray', *Nature*, Vol.21, (1999), pp.25-32.

[6] Tavazoie S, Hughes D, Campbell MJ, Cho RJ & Church GM, 'Systematic determination of genetic network architecture', *Nature Genet*, (1999), pp.281–285.

[7] Yogita R & Harish R, 'A Study of Hierarchical Clustering Algorithm', *International Journal of Information and Computation Technology*, Vol.3, No.11, (2013), pp.1225-1232.

[8] Kaufman L & Rousseeuw PJ, *Finding Groups in Data*, Wiley, (1990).

[9] Zahra Z, Amirhossein H & Ali MN, "Computational methodologies for analyzing, modeling and controlling gene regulatory networks", *Biomedical Engineering and Computational Biology*, Vol.2, (2010), pp.47–62.

[10] Dey L & Mukhopadhyay A, 'Microarray Gene Expression Data Clustering using PSO based K-means Algorithm', *Proceedings of the International Conference Advanced Computing, Communication and Networks*, (2011), pp.587-591.

[11] Kohonen T, 'The self-organizing map', *Proc. IEEE*, Vol.78, No.9, (1990), pp.1464–1480.

[12] Vesanto J & Alhoniemi E, 'Clustering of the Self Organizing Map', *IEEE Transactions on Neural Networks*, Vol.11, (2000), pp.586–600.

[13] Ester M, Kriegel HP, Sander J & Xu X, 'A density-based algorithm for discovering clusters in large spatial databases with noise', *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, (1996), pp.226–231.

[14] Adriano M, Maribel, Y & Sofia C, "Density based clustering algorithms", *DBSCAN and SNN*, (2005).

[15] Huang Z, "A fast clustering algorithm to cluster very large categorical data sets in data mining", *Proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*. Tucson, (1997), pp.146-151.

[16] Hinneburg A & Keim D, "An efficient approach to clustering in large multimedia databases with noise", *American Association for Artificial Intelligence*, (1998), pp.58-65.

[17] Zhang T, Ramakrishnan R & Livny M, "BIRCH An efficient data clustering method for very large databases", *SIGMOD International Conference on Management of Data*, (1996), pp.103-114.