Microarray Gene Expression Data Analysis through a Hybrid Clustering Algorithm incorporated with Validation Techniques

Muhammad Rukunuddin Ghalib[#1], Babyna Nandeibam[†2] , D. K. Ghosh[*3],
[#†]School of Computing Science and Engineering
*Dept. of Mathematics
[1, 2] VIT University, Vellore, Tamil Nadu, India
[3]VSB Engineering College, Karur, Tamil Nadu, India
[1]ghalib.it@gmail.com
[2]babyna.nandeibam@gmail.com
[3]drdkalim@yahoo.in

Abstract— Clustering is an important technique which is used to analyze gene expression data to reveal groups of gene sharing common functional properties. With the help of the developing technology called microarray technology, clustering has now become a main technique to gene expression data analysis. Microarray technology has now made it possible to monitor the expression levels of thousands of genes during important biological processes. In this work, A Novel Hybrid Microarray Gene Expression Clustering Algorithm has been proposed which is incorporated with Hubert's Statistic Technique, Jaccard's coefficient and Dunn's Index used for cluster Validation. Its main aim is to improve the efficiency level of the quality of clusters, with optimized validation and reduce the memory requirements lower than almost all the existing clustering algorithms. It also guides in achieving high quality clusters.

Keyword- Microarray Technology, Gene Expression Analysis, Clustering Algorithm, Correlation Clustering technique, Hubert statistics, Dunn's Index Clustering, Jaccard's Coefficent.

I. INTRODUCTION

A) Microarray Technology:

Compared with the traditional approach to genomic research, which has focused on the local examination and collection of data on single genes, microarray technologies have now made it possible to study the expression levels for tens of thousands of genes in parallel. The two major types of microarray experiments are the cDNA microarray and oligonucleotide arrays. Both types of experiments comprise three common basic procedures:

Chip manufacture: A microarray is a small chip (made of chemically coated glass, nylon membrane or silicon), onto which tens of thousands of DNA molecules are attached in fixed grids. Each grid cell relates to a DNA sequence.[1]

Target preparation, labeling and hybridization: Typically, two mRNA samples (a test sample and a control sample) are reverse-transcribed into cDNA , labeled using either fluorescent dyes or radioactive isotopics, and then hybridized with the probes on the surface of the chip.[1]

The scanning process: Chips are scanned to read the signal intensity that is emitted from the labeled and hybridized targets.

B) Cluster Analysis

In clustering, the data consist of gene expression values. [2] The analytical goal is to find clusters of samples or clusters of genes such that observations within a cluster are more similar to each other than they are to observations in different clusters. Cluster analysis can be viewed as a data reduction method in that the observations in a cluster can be represented by an 'average' of the observations in that cluster. There are a large number of statistical and computational approaches available for clustering. These include hierarchical clustering [3][20][21] and k-means clustering [2][23] from the statistical literature and self-organizing maps [5] and artificial neural networks [7][14][15] from the machine learning literature.

Clustering is the process of grouping the data into classes or clusters,[1] so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters. There are several clustering techniques. They are: partitioning methods, hierarchical methods, density-based methods, grid-based methods, model-based methods, methods for high-dimensional data (such as frequent pattern–based methods), and constraint-based clustering. Clustering can also be used for outlier detection.

II. RELATED WORK

One of the most common clustering methods applied to gene expression data is Hierarchical clustering [4] which is developed as a single layered neural network. By collecting genes with similar pattern of expression, a hierarchical chain of nested clusters are generated. This pattern of expression is across a range of samples located near each other. Hierarchical clustering evaluates all pair wise distance relationships between genes and experiments to combine pairs of values that are more similar to each other for the formation of a node. Further these clusters are group together to create a higher level cluster by using inter-cluster distance and they are graphically represented by a tree called dendogram. It shows the relationship between the clusters. All clusters are joined by repeating the process of comparison between the genes or new clusters. There are two methods i.e bottom up approach (agglomerative algorithms) which joins clusters in a hierarchical way [4][25] and top down approach (diving algorithms) which splits clusters hierarchically. This method has some drawbacks including high computational intricacy, vagueness and failure due to large number of genes as data sets.

K-means is another clustering technique [5] [16][17]commonly used which is simple and a fast method . It is easy to implement and has small number of iterations. First, the data sets are divided into k disjoint subsets. Then the user evaluates the number of cluster (k) and calculated as an input where the computer randomly assigns each gene to one of the k clusters.

The genes inside every cluster are as close to the centre of the cluster as possible. The distance between each gene and the centre of each cluster is calculated resulting in an optimal grouping of data to clusters.

This method is effective if different values of k are attempted but it doesn't give the relationship between the clusters like hierarchical clustering, it only evaluates the number of clusters.

Some of the drawbacks of K-means clustering are lack of prior knowledge of the number of gene clusters in a gene expression data and alteration of results.

Self Organized Maps Clustering technique [6] is easy to implement and is a fast method which is scalable to large number of data sets.It is an unsupervised algorithm used to analyse vast number of data in different fields like visualization and monitoring process. It makes the visualization of complex data easier.

Validation techniques:

1. **Hubert's Statistics Technique**[9][18][19]**:**

Hubert's Statistics can be defined as follows:

Let us consider two $n \times n$ 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:

$$Y(i, j) = \begin{cases} 1 & \text{if genes } i \text{ and } j \text{ are clustered in the same} \\ & \text{cluster} \\ 0 & \text{Otherwise} \end{cases} \qquad (1)$$

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

$$\Gamma = \frac{1}{M} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\frac{X(i, j) - \overline{X}}{\sigma_X} \right) \left(\frac{Y(i, j) - \overline{Y}}{\sigma_Y} \right) \qquad (2)$$

Where M is the number of entries in the double sum and it is given by $M = n(n-1)/2$, and σ_X and σ_Y denote the sample standard deviations. The sample means of the entries of matrices X and Y is denoted by $\overline{X}$ and $\overline{Y}$.

The value of Hubert's Statistic Γ is between [-1,1]. The higher value of Γ indicates better clustering result and cluster quality.

The above formula can be simplified further to improve the performance of the approach since the computing of Γ statistic takes longer time. So by expanding the expression we get Simplistic Hubert's Γ Statistic which is denoted as follows:

$$\Gamma = \frac{M\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} X(i,j)Y(i,j) - \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} X(i,j)\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} Y(i,j)}{\sqrt{M\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} Y(i,j) - \left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} Y(i,j)\right)^2}} \quad (3)$$

2. Dunn Index[8][20][23]**:**

Dunn index identifies all sets of clusters that are dense and well separated. Let us consider any partition $U \leftrightarrow X : X_i \cup ....X_i \cup ...X_c$, where X_i represents the i^{th} cluster of the partition. The Dunn's validation index, D, is defined as eqn (4) :

$$D(U) = \min_{1\le i\le c}\left\{ \min\left\{ \frac{\delta(X_{i,}X_j)}{\max_{1\le k\le c}\{\Delta(X_k)\}} \right\} \right\} \quad (4)$$

Here $\delta(X_{i,}X_j)$ indicates the intercluster distance between clusters X_i and X_j.

$\Delta(X_k)$ indicates the intracluster distance of cluster X_k and c is the number of clusters of partition U. The main aim of this validity index is to maximize intercluster distances and minimize intracluster distances. Higher values of D represents better clusters.

3. Simple matching coefficient and Jaccard coefficient[9][12][13]**:**

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

$$G(i,j) = \begin{cases} 1 \text{ if } i \text{ and } j \text{ are clustered in the same cluster} \\ 0 \text{ otherwise} \end{cases} \quad (5)$$

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.

Table I: Confusion Matrix of Simple Matching Coefficient

Matrix P	**Matrix Q**	
	1	**0**
1	a	b
0	c	d

The simple matching coefficient is given by

$$S = \frac{a+d}{a+b+c+d} \quad (6)$$

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

The Jaccard coefficient is given by

$$S = \frac{a}{a+b+c} \quad (7)$$

Here, the negative matches 'd' are not considered

III. METHODS AND IMPLEMENTATION

The Computational microarray dataset used in our experiments are downloaded from publicly available data site (www.pnas.org) or at http://rana.stanford.edu/clustering. Data in our sample were formed on spotted DNA microarrays, for which the gene expression were studied during the diauxic shift [24][25] of budding yeast *Saccharomyces cerevisiae* [15]. We also have run through the data from Kim lab, Stanford University for our research purpose which is available in http://cmgm.stanford.edu/~kimlab/, [22].

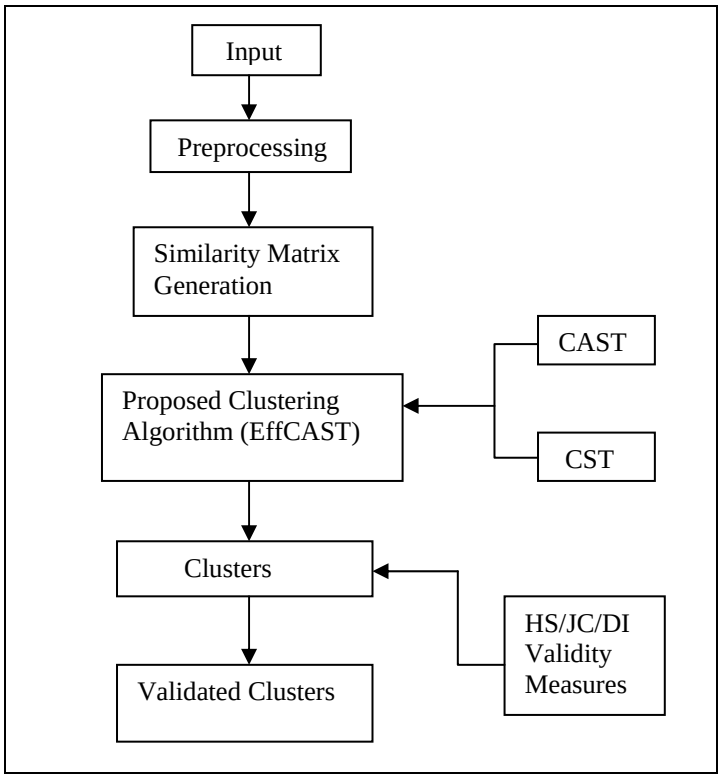


Figure 1: Illustration of the Schematic System Architecture

Figure 1 shows the detailed architecture of the proposed method. The input is a set of data generated by a correlation inclined cluster data sets generator. The input data is preprocessed by using data mining open software tool Weka or by using k-means method to remove all outliers. Then the similarity matrix is generated by using similarity measurements formula like Euclidean Distance and Correlation coefficient. The matrix stores the similarity between each pair of genes in the dataset, and the range of the degree is [0, 1]. This reduces the computation overhead occurred in some clustering algorithms that calculates the similarities dynamically. The hybrid algorithm can automatically cluster the genes according to the similarity matrix without any user input parameters.

The hybrid algorithm EffCAST is the enhanced version of CAST and CST clustering algorithms. CAST (Cluster Affinity Search Technique) [11] clustering technique takes a parameter called the affinity threshold t as an input, where $0<t<1$. The average similarity in each generated cluster in this method is higher than the threshold t. CAST generates one cluster at a time and selects the object with the most neighbors as seed for the current cluster. It groups un-clustered with high affinity i.e the average similarity between the object and the cluster is greater than t, to the current cluster and removes objects with low affinity from the current cluster. The main advantage of CAST is that it can detect outliers more effectively.

CST (Correlation Search Technique) is another clustering algorithm which aim at reducing the amount of computation to obtain a nearly optimal clustering results [9]. The idea behind CST is to integrate clustering method with validation techniques so that the genes can be clustered quickly and automatically.

The main idea behind Hybrid Clustering Algorithm EffCAST is to obtain optimal clustering results. After generating clusters from the dataset, a validation technique is apply to check the validity index of each cluster. The validation technique used is Hubert's Statistics. This technique is further expanded as Simplistic Hubert's Statistic to get better clustering results.

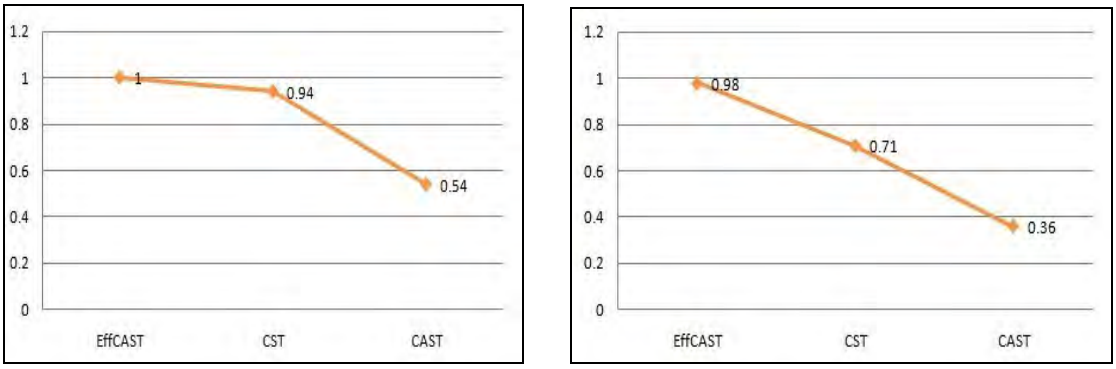
EffCAST Algorithm Detail:

Input: An *g* by *g* similarity matrix A
Pre-processing:
Similarity matrix A and AB
Collection of Closed Clusters C i.e. $K_{initial}$ = NULL
Elements not yet assigned to any cluster i.e U={1,2,……n}
q=0;
count=0;
for all iϵU such that a(i)> θ
{
q+=q(i) – θ
count++
}
g=(q/count) + θ
While (i ≠ NULL) do
Currently constructed cluster C_C= NULL
Correlation function q (.) = 0//this is the Correlation parameter
SEED:
Calculate initial Correlation function of all the genes, g_i
Pick an element i ∈ U with most neighbors.
Remove i from U
Then update the Correlation function and
Insert i into C_C
ADD: While max {q(i)|iϵU}≥ t | C_C | do Pick an element i ϵ U with maximum q(.)
Remove i from U
Add the latest constructed cluster C_C to similarity matrix Y
Add the Correlation function q(i) to XY matrix
Update the Correlation function
Insert i into C_C
REMOVE : min{q(i)|iϵ C_C } < t| C_C |do Pick an element m ∈ C_C with maximum q(.)
Remove m from C_C
Remove C_C from matrix B
Remove q(i) from matrix AB
Update the Correlation function
Insert m into i
Repeat steps ADD and REMOVE as long as there are no elements been added or removed.
C = C ∪{ C_C }
End
Done,return the collection of clusters, C.
Validate C using HS
Validate C using JC
Validate C using DI
Finally, validated C_{Val} is collected.

Figure 2: Pseudo Code of EffCAST

Cluster validity Index Computation

The quality of cluster is measured by using Hubert's Statistic [1], Dunn's index [10] and Jaccard coefficient [1]. The similarity measure is evaluated by using correlation coefficient since it is widely used in most studies on gene expression analysis. The visualized results of the clustering methods are exhibited by using an intensity image. Figure 3 shows the result for validity index measurement by using Hubert's Statistic on Dataset I and II for EffCAST, CST and CAST.



(a) (b)

Figure 3: Validation results for (a) Dataset I and (b) Dataset II.

IV. RESULTS AND DISCUSSION

1. Result on Similulated Dataset I and Dataset II

Table II and Table III show the best clustering quality of the tested methods on Dataset I and Dataset II. It also show the total execution time for both Datasets. We take out clusters whose size is greater than 50. In Dataset I Effaces has 5 clusters, CST has 4 Clusters and CAST has 2 clusters and for Dataset II EffCAST has 8 clusters,CST has 7 clusters and CAST has 4 clusters. It is observed that EffCAST outperform the other two clustering algorithms in both execution time and clustering quality in terms of validation measures.

Table II: Experimental Results Dataset I

Methods	Time(s)	No of Clusters	Hubert's Statistics	Dunn's Index	Jaccard Coefficient
EffCAST	8	35	1	0.96	0.94
CST	10	38	0.94	0.89	0.91
CAST	12	19	0.54	0.63	0.56

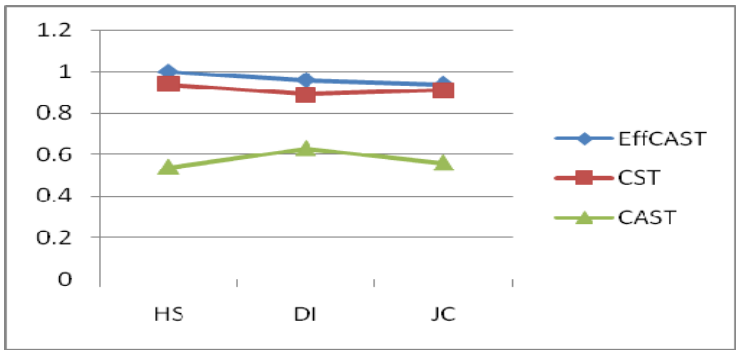


Figure 4: Line Plots of all the Average Cluster validity indices used in the experiment applied to all the three algorithms for dataset I

Table III: Experimental Results Dataset II

Methods	Time(s)	No of Clusters	Hubert's Statistics	Dunn's Index	Jaccard Coefficient
EffCAST	90	67	0.98	0.90	0.90
CST	120	74	0.71	0.66	0.70
CAST	170	27	0.36	0.31	0.34

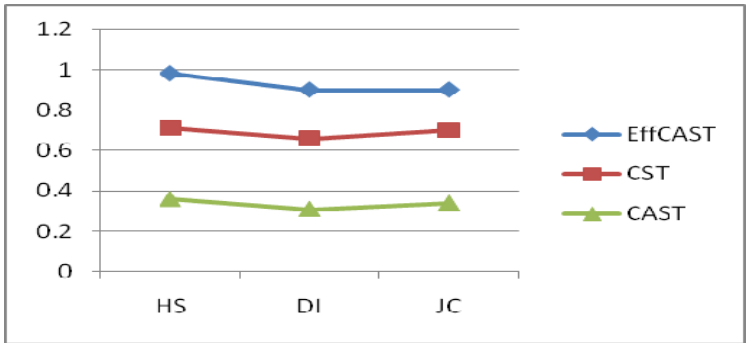
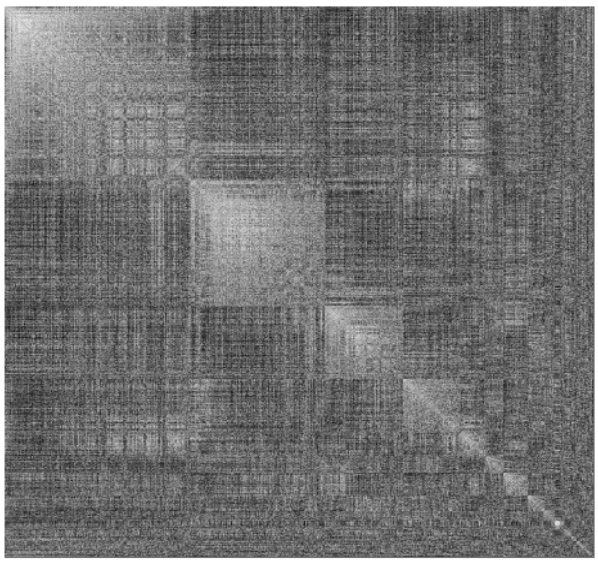


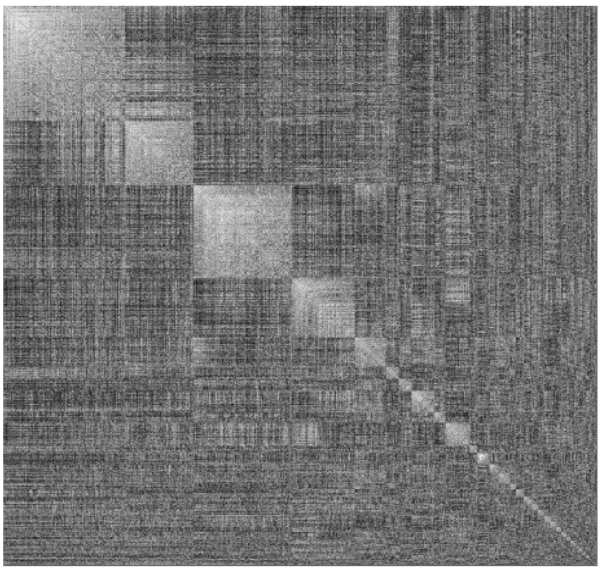
Figure 5: Line Plots of all the Cluster validity indices used in the experiment applied to all the three algorithms for the dataset II

The above result shows less difference between the number of clusters found in clustering algorithm EffCAST and CST but in terms of validity measurement EffCAST is much better than CST and CAST in both Datasets. It means that the clustering quality of EffCAST is good than other methods.

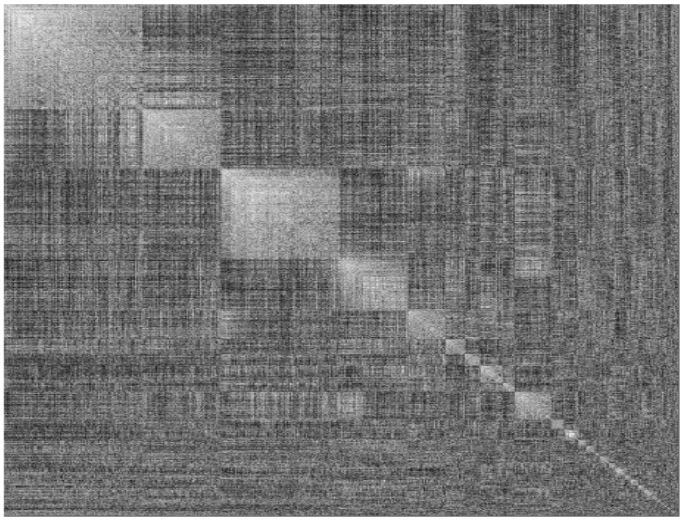
Figure 6 and 7 shows the intensity images of the original cluster structures and the clustering results for Dataset I and Dataset II.



(a)

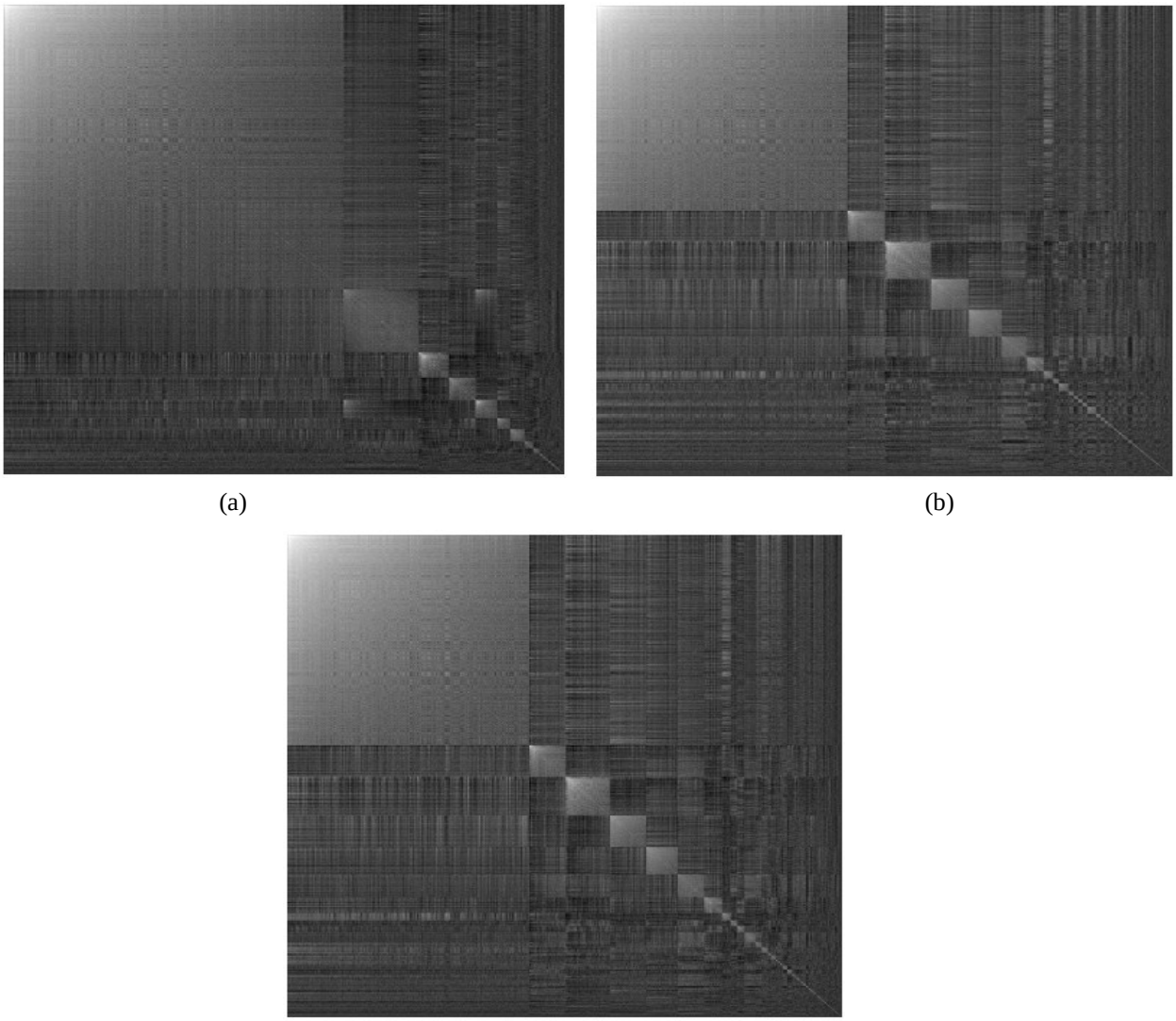


(b)



(c)

Figure 6: (a) Clustering Result of CAST (b) Clustering Result of CST (c) Clustering Result of EffCAST for Dataset I.



(a) (b)

(c)

Figure 7: (a)Clustering Result of CAST (b) Clustering Result of CST (c)Clustering Result of EffCAST for Dataset II

2. Discussion

The time consumption for all algorithms for dataset I is comparatively similar and small as given in Table 2 for the no of gene samples are less compared to dataset II. The number of clusters found by EffCAST is 35, CST is 38 and CAST is 19. In this clusters, we applied the validity measures like Hubert's statistics, Dunn's Index and Jaccard Coefficient. We found more than 90% of the clusters in EffCAST averaged to HS value ≡ 1, DI≡0.96 and JC≡0.94, which shows that EffCAST outclass other algorithms like CST and CAST separately. The detailed computational values are shown in Table II and Table III. Similarly we prove the efficiency in terms of cluster quality for the dataset II which can be seen in detail from Table III and Figure 5; the only drawback is that it is not immune to data size in terms of time constraint. Still it outclasses the CAST and CST here, but we cannot generalize it in extent.

V. CONCLUSION AND FUTURE WORK

We proposed a new microarray clustering algorithm which belongs to the family of on-the-fly clustering algorithms like CST, CAST and other algorithms. We could enhance the performance in terms of cluster quality by 90-95% in terms of cluster validity measurement, that's is almost 32 out 35 clusters found by EffCAST in Dataset I averaged the HS value to 1 and even other DI and JC values nearly to 1 which means it is good quality clusters.

In future we could consider the fuzziness, Roughness and the missing values dependence to our algorithmic design and computation by relying on various Rough-fuzzy concepts and through optimization algorithms and prediction through neural networks.

ACKNOWLEDGMENT

This work is partially supported by a grant of young scientist award from The Manipur Public Hospital and Research Institute (MPHRI/YSG/2011-RG.No.23) under the Principal Investigator of Dr. D. K. Ghosh and Co-PI of Muhammad Rukunuddin Ghalib.

REFERENCES

[1] Daxin Jiang, Chun Tang, Aidong Zhang, (2004)"Cluster Analysis for Gene Expression Data: A Survey",IEEE Transactions On Knowledge And Data Engineering, Vol 16,No 11,.
[2] William Shannon,Robert Culverhouse &Jill Duncan,(2003) "Analyzing microarray data using cluster analysis",© Ashley Publications Ltd ISSN 1462-2416.
[3] Grzegorz M Boratyn, Member IEEE, Susmita Datta, and Somnath Datta,(2006) "Biologically Supervised Hierarchical Clustering Algorithms for Gene Expression Data", Proceedings of the 28th IEEE EMBS Annual International Conference New York City, USA, Aug 30-Sept 3.
[4] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown and David Botstein,(1998) "Cluster analysis and display of genome-wide expression patterns", Proc. Natl. Acad. Sci. USAVol. 95, pp. 14863–14868.
[5] Erfaneh Naghieh and Yonghong Peng,(2009) "Microarray Gene Expression Data Mining: Clustering Analysis Review".
[6] Petri Toronena, Mikko Kolehmainenb, Garry Wonga, Eero Castren, (1999) "Analysis of gene expression data using self-organizing maps", FEBS Letters 451. pages 142-146
[7] Maria Halkidi,Yannis Batistakis,Michalis Vazirgiannis,(2001) " On Clustering Validation Techniques", Journal of Intelligent Information Systems, 17:2/3, pages 107–145.
[8] 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
[9] Vincent S. Tseng and Ching-Pin Kao, (2005) "Efficiently Mining Gene Expression Data via a Novel Parameterless Clustering Method", IEEE/ACM Transactions On Computational Biology And Bioinformatics,Vol.2, No.4,pages 355-365.
[10] Hui-Huang Hsu,(2006) "Advanced data mining technologies in bioinformatics", IGI Global pub, USA
[11] D.R. Westhead, J.H. Parish and R.M. Twyman, 2003, 'Bioinformatics', Viva Books Pvt. Ltd. New Delhi.
[12] Zoe Lacroix and Terence Critchlow, (2003), 'Bioinformatics managing scientific data', Morgan Kauffman publishers, Statistical Meetings of the American Statistical Association (Biometrics Section).
[13] Brazma, Alvis and Vilo, Jaak. ,(2000) 'Minireview: Gene expression data analysis'. Federation of European Biochemical societies, Vol. 480, no. 1, pp17–24.
[14] David R. Bickel. , (2001), 'Robust Cluster Analysis of DNA Microarray Data: An Application of Nonparametric Correlation Dissimilarity', Proceedings of the Annual Meeting of the American Statistical Association, pp.1-6.
[15] Eisen, Michael B., Spellman, Paul T.,Brown, Patrick O.and Botstein David., (1998), 'Cluster analysis and display of genome-wide expression patterns', Proc. Natl. Acad. Sci. USA, Vol. 95, no. 25, pp.14863–14868.
[16] Yeung, K.Y., Haynor, D.R. and Ruzzo, W.L., (2001) 'Validating Clustering for Gene Expression Data'. Bioinformatics, Vol.17, no. 4, pp.309–318
[17] Francis D. Gibbons and Frederick P. Roth , (2002), 'Judging the quality of gene expression-based clustering methods using gene annotation', Genome Research, Vol.12, no. 10, pp. 1574-1581.
[18] Juntao Wang and Xiaolong Su , (2011), 'An improved k-means clustering algorithm', IEEE Third International Conference on Communication Software and Networks (ICCSN), pp. 44-46.
[19] Shi Na, Liu Xumin, Guan Yong, (2010), 'Research on k-means clustering algorithm', Third International Symposium on Intelligent Information Technology and Security Informatics (IITSI), pp. 63-67
[20] Ankur Mazumdar, Muhammad Rukunuddin Ghalib, (2011), 'Qualitative and Quantitative metrics based analysis of Gene Expression Data Clustering Algorithms',Intl. J. of Computer Information Systems, Vol II, Issue IV, Pages 44-48.
[21] Muhammad Rukunuddin Ghalib, Ritwika Ghosh, Priti Saswal, Udisha Pande,(2013), "Microarray Gene Expression Analysis using Enhanced k-means Clustering Algorithm", Intl. J of Advances in Engineering and Technology, Vol V, Issue II, Pages 373-380.
[22] S.Kim, Department of Developmental Biology, Stanford university, http://cmgm.stanford.edu/~kimlab/
[23] Yeung, K.Y., Haynor, D.R. and Ruzzo, W.L., (2001) "Validating Clustering for Gene Expression Data. Bioinformatics, Vol.17 (4):pp309–318.
[24] Shalon, D., Smith, S. J. & Brown, P. O. (1996) Genome Res. 6, Pages 639–645.
[25] DeRisi, J. L., Iyer, V. R. & Brown, P. O. (1997) Science 278, Pages 680–686.