Gene Selection for Patient Clustering by Gaussian Mixture Model

Md. Hadiul Kabir hadi_ruo7@yahoo.com
Laboratory of Bioinformatics,
Department of Statistics
University of Rajshahi
Rajshahi-6205, Bangladesh

Md. Shakil Ahmed shakil.statru@gmail.com
Laboratory of Bioinformatics,
Department of Statistics
University of Rajshahi
Rajshahi-6205, Bangladesh

Md. Nurul Haque Mollah mollah.stat.bio@ru.ac.bd
Laboratory of Bioinformatics,
Department of Statistics
University of Rajshahi
Rajshahi-6205, Bangladesh

Abstract

Clustering is the basic composition of data analysis, which also plays a significant role in microarray analysis. Gaussian mixture model (GMM) based clustering is very popular approach for clustering. However, GMM approach is not so popular for patients/samples clustering based on gene expression data, because gene expression datasets usually contains the large number (*m*) of genes (variables) in presence of a few (*n*) samples observations, and consequently the estimates of GMM parameters are not possible for patient clustering, because there does not exists the inverse of its covariance matrix due to *m>n*. To conquer these problems, we propose to apply a few '*q*' top DE (differentially expressed) genes (i.e., *q<n/2<m*) between two or more patient classes, which are selected proportionally from all DE gene's groups. Here, the fact behind our proposal that the EE (equally expressed) genes between two or more classes have no significant contribution to the minimization of misclassification error rate (MER). For selecting few top DE genes, at first, we clustering genes (instead of patients/samples) by GMM approach. Then we detect DE and EE gene clusters (groups) by our proposed rule. Then we select *q* (few) top DE genes from different DE gene clusters by the rule of proportional to cluster size. Application of such a few '*q*' number of top DE genes overcomes the inverse problem of covariance matrix in the estimation process of GMM's parameters, and ultimately for gene expression data (patient/sample) clustering. The performance of the proposed method is investigated using both simulated and real gene expression data analysis. It is observed that the proposed method improves the performance over the traditional GMM approaches in both situations.

Keywords: Gene Expression, Patient Clustering, Gaussian Mixture Model, Inverse Problem of Covariance Matrix, Top DE genes Selection for Patient Clustering.

1. INTRODUCTION

Clustering is a useful exploratory technique for gene expression data analysis. In fact, clustering is usually performed when no information is available concerning the membership of data items to predefined classes. That's why; clustering is traditionally seen as part of unsupervised learning. However, several heuristic clustering algorithms have been proposed in microarray data analysis. Model (especially, GMM) based clustering offer a principled alternative to heuristic algorithms.

Clustering approach may understand of the functions of many genes for which information has not been previously available [1-2]. These techniques have proven to be helpful to understand gene function, gene regulation, cellular processes, and subtypes of cells. Genes with similar expression patterns i.e., co-expressed genes can be clustered together with similar cellular functions. The co-expressed genes can also be grouped in clusters based on their expression patterns [2-4]. Furthermore, the co-expressed genes in the same cluster are likely to be involved in the same cellular processes, and a strong correlation of expression patterns between those genes indicates co-regulation. Searching for common DNA sequences at the promoter regions of genes within the same cluster allows regulatory motifs specific to each gene cluster to be identified and cisregulatory elements to be proposed [4-5]. The inference of regulation through the clustering of gene expression data also gives rise to hypotheses regarding the mechanism of the transcriptional regulatory network [6-7]. And finally, clustering different samples on the basis of corresponding expression profiles may reveal sub-cell types which are hard to identify by traditional morphology-based approaches [8-9].

In practical situation, a gene expression data are typically highly-connected. And, there may be instances in which a single gene has a high correlation with two different clusters; therefore, the probabilistic feature of model-based clustering is particularly suitable for gene expression data. Mixture models are being commonly used in a wide range of application in practice. Clustering algorithms based on mixture model are being increasingly preferred over historic methods because it provides a sound statistical framework to model the cluster structure of gene expression data, and also for interpretability and validity of the results [10-14]. The basic idea of model based clustering is to model each of the subpopulations separately, and the overall population as a mixture of these subpopulations using finite mixture models [15-16]. In this approach, the observations belong to the clusters in terms of the fitted values for their posterior probabilities of the component membership of the mixture i.e., each component probability distribution to a cluster. This procedure is also known as mixture likelihood approach to clustering. There is some guidance based on statistics (e.g., AIC, BIC) associated with these approaches for solving major challenging important issues of clustering, such as, how many clusters there are, which model should be used, and how cluster should be handled?. In these clustering frameworks, several works has been done on this problem [17-23].

Most of the clustering algorithms are largely heuristically motivated, and the issue of determining the number of clusters in dataset may not rigorously solved, that’s a vital problem in most of the clustering algorithms. However, model-based clustering relies on the assumption that the data set fits a specific distribution; in particular, most of datasets usually fits a Gaussian mixture distribution, which may not be true in few cases. Modeling of the gene expression data 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 perfectly. However, gene expression data is very different from any of the data; therefore, clustering of this type of data is an important task in many application areas [24-25]. At present, a typical microarray experiment generally contains thousand to ten millions genes, and this number is expected to reach to the order of thousand millions, whereas, the number of samples involved in a microarray experiment is very small, in addition, most genes are irrelevant to microarray data clustering. From these points of views, we proposed to apply a few number of top DE genes for patients (samples) clustering. Performing top DE gene selection helps to reduce data sizes, thus improving the running time also. In fact, top DE gene selection removes a large number of irrelevant genes which improves accuracy of clustering. Focusing on dimension reduction via feature selection, model-based approaches have been also taken in studies [26-28]. However, the singular value decomposition is also important in high dimensional clustering, especially, for the bi-clustering (both genes and samples clustering) analysis [29-31]. In order to gain a better insight into the problem of clustering, systematic approaches based on global gene expression analysis have been proposed [3, 8, 32-33]. However, the feature selection algorithms are considered to be an important way of identifying crucial genes. In clustering context, several works on feature selection has been done in literature recently with some advantages and disadvantages [14, 34-42]. These methods select important genes using some objective functions. The selected genes are expected to have biological significance and should provide high accuracy. However, on many microarray datasets the performance is still limited and hence the improvements are necessitated.

Furthermore, many popular statistical methods like Gaussian mixture based clustering could suffer from inverse operation of sample covariance matrix due to scarce samples. This problem can be resolved by regularization techniques or pseudo inversing covariance matrix. In this paper, to avoid the singularity problem of variance-covariance matrix, we are suggested to use a few number of top DE genes in model based clustering, provided that number of top DE genes must be less than the corresponding sample sizes. Our main goal here is to develop a method to solve the above problem by removing EE genes (because EEs have no contribution to enhance clustering performance) and selecting '*q*' number of top DE genes ($q < n/2 < m$) proportionally from the DE gene groups in dataset. Then, it is possible to apply popular statistical approaches (like GMM) for patient (sample) clustering in this reduced microarray dataset. In proposed algorithm, first we clustered genes (variables) instead of sample (observations) into two or more clusters (groups) by GMM approach for genes (features) clustering. Then, we categorized gene's groups into DE and EE groups using proposed objective functions (6-8), and remove EE genes group from dataset. Finally, we select few (*q*) number of top DE genes proportionally from each DE groups in order of magnitude, which provides the sufficient information for patients (samples) clustering of underlying dataset. The MER is used to judge the performance of clustering. GMM based clustering using a few top DE genes is experimented on simulated and real datasets. Besides gene selection, there are several issues related to (gene expression data) clustering that are of great concern to researchers. These issues are derived from the biological context of the problem, and the medical importance of the result. We believe that in order to have an in-depth understanding of the problem, it is necessary to study both the problem and its related issues and look at them all together. We organized this paper as follows. In section 2, we described the Gaussian mixture model based clustering and detection of the DE and EE gene groups. In section 3, we described the results of the simulated and real gene expressions datasets. Finally, we end this paper with a conclusion.

2. Gaussian Mixture Model (GMM) Based Clustering Approach

Let us assume that source data vectors originate from *c* populations $\{\Pi_1, \Pi_2, \ldots\ldots, \Pi_c\}$ and that the corresponding observe data vectors belongs to *c* different data clusters $\{D_1, D_2, \ldots, D_c\}$ in the entire data space, where *c* is assume to be unknown. In addition, we assume that the data cluster D_k occurs in the entire data space *D* due to the population Π_k, (k=1, 2 … *c*). In practice, the occurrence order of an observed data vector in the entire data space from a population is unknown. Let the data set of n vectors of observations, $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\ldots, \boldsymbol{x}_n\} \varepsilon R^m$. Here, the objective is to separate n-vectors into *c* clusters. To solve the problem, let an observed random vector x_i follows Gaussian mixture distribution [16] as,

$$p(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{c} \pi_k f(\boldsymbol{x}_i;\boldsymbol{\theta}_k) \tag{1}$$

Here, $\boldsymbol{\theta} = \{\pi_k, \boldsymbol{\theta}_k\}_{k=1}^{c}, \sum \pi_k = 1;\ f(\boldsymbol{x};\boldsymbol{\theta}_k) = (2\pi)^{p/2} |\mathbf{V}_k|^{-1/2} \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \mathbf{V}_k{}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\}$

is the probability distribution of population Π_k and π_k is the mixing proportion or prior probability of $\boldsymbol{x}_i \in \Pi_k$. Then the posterior of $\boldsymbol{x}_i \in \Pi_k$ is defined as

$$p(\Pi_k;\boldsymbol{X},\boldsymbol{\theta}) = \frac{\pi_k f(\boldsymbol{x}_i;\boldsymbol{\theta}_k)}{\sum_{k=1}^{c} \pi_k f(\boldsymbol{x}_i;\boldsymbol{\theta}_k)} \tag{2}$$

Then the observed random vector $\boldsymbol{x}_i$ is classified into population $\Pi_{k,}$ if

$$k = \underset{k' \in \{1,2,\ldots,c\}}{\text{argmax}}\; p(\Pi_{k'}; \boldsymbol{X}, \boldsymbol{\theta}) \tag{3}$$

The parameters $\boldsymbol{\theta}$ are estimated by maximizing the likelihood function of Gaussian mixture distribution (eq.1) using EM algorithm. The estimates are as follows,

$$\left.\begin{array}{l} \boldsymbol{\mu}_k{}^{(j+1)} = \dfrac{\sum_{i=1}^{n} t_{ik}{}^{(j)} \boldsymbol{x}_i}{\sum_{i=1}^{n} t_{ik}{}^{(j)}}, \boldsymbol{V}_k{}^{(j+1)} = \dfrac{\sum_{i=1}^{n} t_{ik}{}^{(j)} (\boldsymbol{x}_i - \boldsymbol{\mu}_k{}^{(j)})(\boldsymbol{x}_i - \boldsymbol{\mu}_k{}^{(j)})^T}{\sum_{i=1}^{n} t_{ik}{}^{(j)}} \\ \text{And } \pi_k{}^{(j+1)} = \dfrac{\sum_{i=1}^{n} t_{ik}{}^{(j)}}{n} \end{array}\right\} \tag{4}$$

Here, $t_{ik}{}^{(j)}$ is the posterior probabilities that can be expressed as (using Bayes' theorem),

$$t_{ik}{}^{(j)} = \frac{\pi_k{}^{(j-1)} f(\boldsymbol{x}_i; \boldsymbol{\theta}_k^{(j-1)})}{\sum_{k=1}^{c} \pi_k{}^{(j-1)} f(\boldsymbol{x}_i; \boldsymbol{\theta}_k^{(j-1)})} \tag{5}$$

Where $\boldsymbol{\mu}_k^{(j+1)}$ **,** $\boldsymbol{V}_k^{(j+1)}$ and $\pi_k^{(j+1)}$ are updated as $\boldsymbol{\mu}_k^{(j)}$, $\boldsymbol{v}_k^{(j)}$ and $\pi_k^{(j)}$ respectively. Here, $\boldsymbol{\mu}_k^{(0)}$, $\boldsymbol{V}_k^{(0)}$ and $\pi_k^{(0)}$ are the initial values for $\boldsymbol{\mu}_k$, $\boldsymbol{V}_k$ and π_k respectively at *t=0.*

However, the problem of clustering by the GMM based approach is that the number of clusters '*c*' needs to be known in advance. Several approaches to choosing the number of clusters in model-based clustering have been proposed [43-52]. One advantage of GMM approach to clustering is that it allows the use of approximate Bayes factors to find the number of clusters, and to compare models. The uses of EM algorithms to find the maximum mixture likelihood provide a more reliable approximation to twice the log Bayes factor called BIC [53], and the integrated likelihood can be simply approximated by Schwarz and Haughton [53-54]. This approximation is particularly good when some prior information is used for the parameters [55-56]. Several results suggest its appropriateness and good performance in the model-based clustering context [57-58]. Another problem arises in this approach due to the inverse problem of covariance matrix ($\boldsymbol{V}_k$). When *m>n*, the inverse of $\boldsymbol{V}_k$ does not exist. Usually, the number of genes '*m*' much larger than the number of samples (patients) in case of gene expression dataset. To overcome this problem, we are proposing to apply only few '*q'* top informative genes (instead of the whole dataset) for patient clustering by GMM approach that will be described in the following section 2.1.

2.1 Gene Selection for Patient Clustering by GMM Approach (Proposed)

Gaussian mixture model is very popular clustering approach. However, the main problem arises with such statistical based clustering approach like GMM, when sample clustering is required in presence of large number of variables (features); because, the inverse of its variance covariance matrix does not exists in this situation. In particular, patients clustering based on gene expression data by GMM approach usually suffer from this problem. However, in real situations, microarray gene expressions datasets are often contain (very) small sizes of patient (sample) in presence of large number of gene (variable). That's why; many popular statistical approaches (like GMM) are not suitable for gene expression data clustering and classification due to small sample size (SSS) problem. To solve this problem, we have suggested, first select a few number (*q*) of top DE genes that must be less than sample size (*n*) of the underlying data (i.e., q<n) by proposed feature selection technique, and then apply GMM based approach for patients clustering.

However, it is well known from several feature selection studies that most of the genes in microarray data are equally expressed (EE) that have no significant contribution to the performance of clustering i.e., to the minimization of misclassification error rate (MER). Our proposed gene selection technique is completed as; first, remove EE genes group and select a certain number '*q*' (*q*< *n/2*<*m*) of top DE genes proportionally from all DE gene groups in order of magnitude, and then apply GMM model based clustering to this reduced dataset. Now, the questions arise, how we detect EE gene group, and how we ordered DE gene groups? There are several steps involve with this approach, which are as follows:

(i) At first, perform gene clustering by GMM approach. We obtain '*c*' clusters, where one cluster will consist by the EE genes and rest of the clusters will consist by the different patterns of DE genes.

(iii) To detect EE gene cluster, we are proposing a criteria as follows.

The variance for a EE gene ($\mu_1 = \mu_2$) can be written as,

$$\sigma_{EE}^2 = \frac{n_1\sigma_1{}^2 + n_2\sigma_2{}^2}{n_1 + n_2} = \frac{1}{n}\sum_{i=1}^{2} n_i\sigma_i{}^2 \;; i = 1, 2 \tag{6}$$

Similarly, the variance for a DE gene ($\mu_1 \neq \mu_2$) can be written as,

$$\left.\begin{aligned} \sigma_{DE}^2 &= \frac{n_1(\sigma_1{}^2 + d_1{}^2) + n_2(\sigma_2{}^2 + d_2{}^2)}{n_1 + n_2} \\ \Rightarrow \sigma_{DE}^2 &> \frac{n_1\sigma_1{}^2 + n_2\sigma_2{}^2}{n_1 + n_2}; \; d_i = \mu_i - \mu, \; \mu = \frac{n_1\mu_1 + n_2\mu_2}{n_1 + n_2} \\ \Rightarrow \sigma_{DE}^2 &> \sigma_{EE}^2 \qquad\qquad ; d_i \neq 0 \end{aligned}\right\} \tag{7}$$

Therefore, if a DE gene cluster/group contain m$_1$ same pattern genes and EE cluster/group contains m_2 same pattern genes, then

$$\frac{1}{m_1}\sum_{i=1}^{m_1}\sigma_{DE}^2(i) > \frac{1}{m_2}\sum_{j=1}^{m_2}\sigma_{EE}^2(j) \tag{8}$$

Using equation (8), we can detect the EE and DE gene clusters.

(iv) We select top q_i DE genes proportionally from each i^{th} DE gene groups in order of their variances i.e., for i^{th} DE gene's group, first we ranking all DE genes of i^{th} DE gene's group with respect their variances, and then we select first few q_i ('q_i' is chosen according to proportional allocation rule in sampling design) DE genes from i^{th} DE group. For example, for two DE gene groups, we select q_1 top DE genes from one DE gene group, and similarly, select q_2 top DE genes from another DE gene's group; so, we have q_1+q_2=*q* (*q*<*n/2*<*m*) top DE genes in total from both DE gene's groups, the total number of DE genes (selected from all DE genes groups) must be less than the number of samples. Finally, these *q* top DE genes (reduced gene expression data) are used to estimate GMM and ultimately for required patients clustering.

3. SIMULATED AND REAL GENE EXPRESSION DATA ANALYSIS

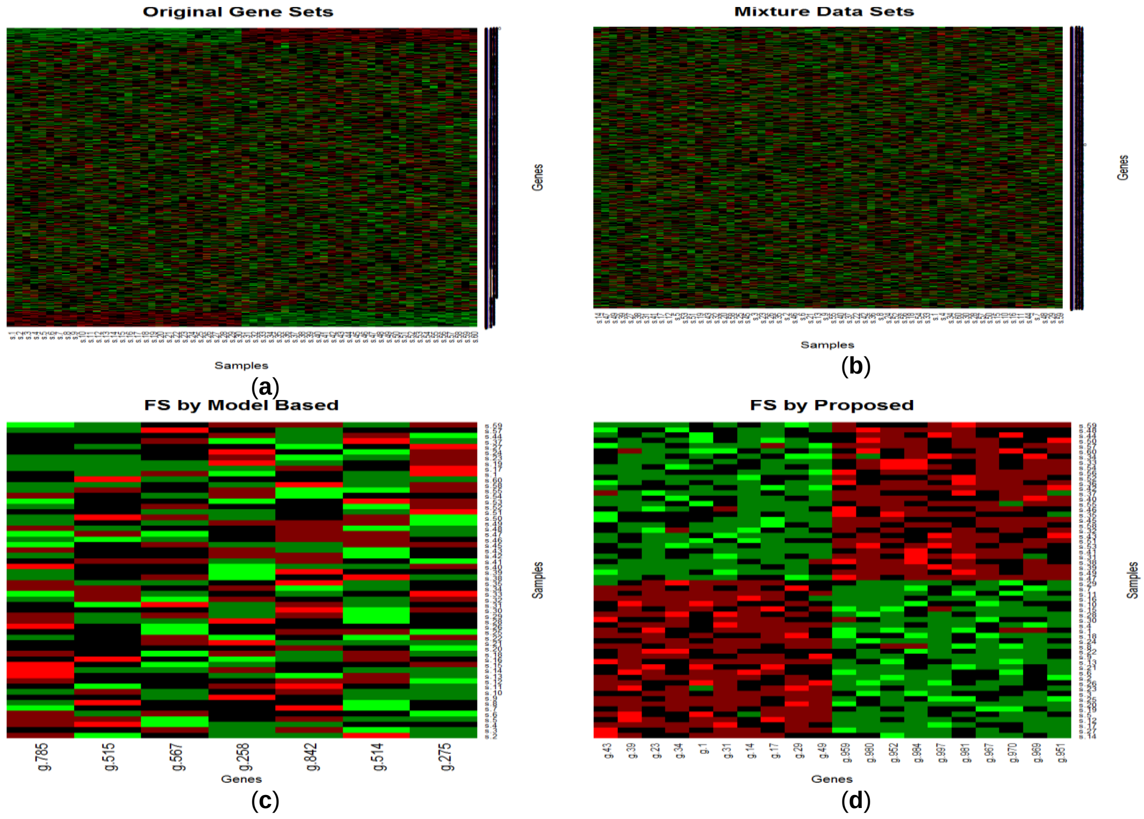
To investigate the performance of the proposed feature selection technique in GMM approach for patients clustering. Here, we have used only a few number (*q*<*n*) of top DE genes, which selected by our proposed feature selection approach, for patient clustering by GMM, and evaluated error rate in both simulated and real gene expression microarray datasets.

3.1 Simulated Gene Expression Data Analysis

We generated three set of simulated data using the data generating model as described in figure-1, where the row represents the gene groups like *A, B, C* and column sample group like P_1, P_2. For randomization, we add the Gaussian noise in the data set with parameters *d*=2 and σ^2=1, where the sample size is *n* = 20 and number of genes m_1=5 and for group *A* genes denoted by $\{A_1, A_2,\ldots,A_5\}$ ε *A*; group *B*, m_2= *90* genes denoted by $\{B_1,B_2,\ldots,B_{90}\}$ ε *B*, and finally group *C* m_3=*5* genes denoted by $\{C_1,C_2,\ldots,C_5\}$ ε *C,* these groups are generated from the normal distribution. That is, $n=n_1+n_2$=10+10=20 and $m=(m_1+m_2+m_3)$=100 are the total number of samples and genes respectively; and, the first group *A* is the *DE* (up-regulated) genes group that contains 5 genes, 2nd group *B* is the *EE* gene group that contains 90 genes, and finally the 3rd group is also DE genes (down-regulated) that contains 5 genes.

Sample Group Gene Group	P_1	P_2
A	$d + N(0, \sigma^2)$	$-d + N(0, \sigma^2)$
B	$d + N(0, \sigma^2)$	$d + N(0, \sigma^2)$
C	$-d + N(0, \sigma^2)$	$d + N(0, \sigma^2)$

FIGURE 1: Gene Expression Data Model.

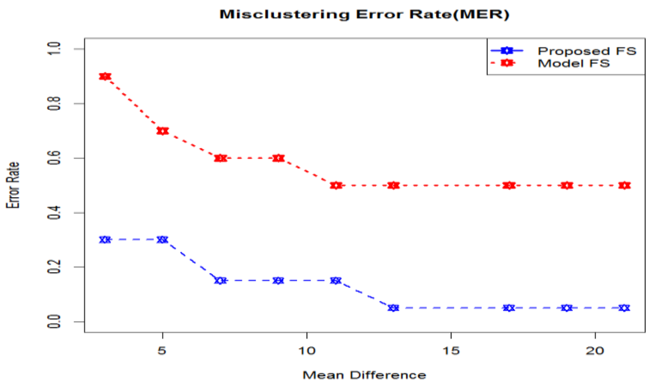


(a)

(b)

(c)

(d)



(**e**)

FIGURE 2: (a) Simulated full dataset, (b) Randomly allocated dataset, (c) Clustering results by GMM based clustering, (e) Clustering results by proposed FS approach (f) MER of the proposed FS and model based FS approaches.

Feature selection by model based and proposed approaches				MER for model based and proposed FS approaches		
Gene group	True DE	Top DE		Mean Difference	MER	
		Model FS	Proposed FS		Model FS	Proposed FS
1st Group	g.1 g.2 g.3 g.4 g.5	g.32 g.40	g.1 g.2	0.0	0.9	0.3
				0.2	0.7	0.3
				0.6	0.6	0.2
				1.0	0.6	0.2
2nd group	g.96 g.97 g.98 g.99 g.100	g.98	g.96 g.98	1.5	0.5	0.2
				2.0	0.5	0.2
				3.0	0.5	0.2
				4.0	0.5	0.2

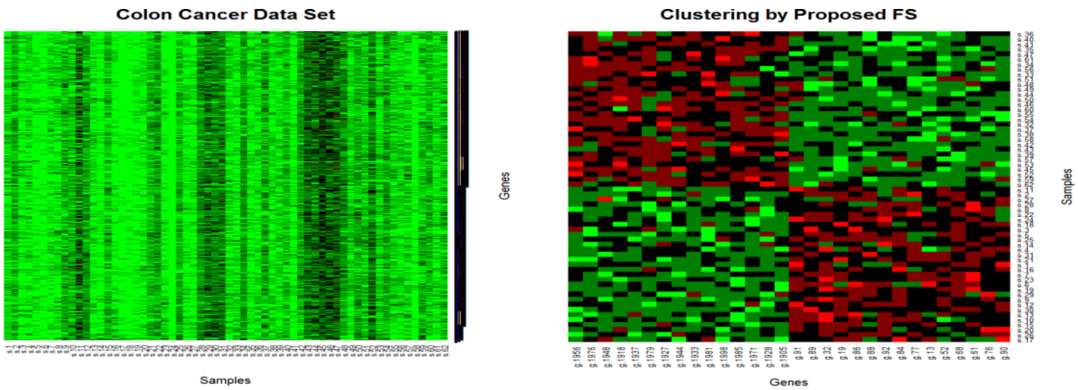
TABLE 1: Top DE genes and MER for the proposed and model based approaches.

However, In this simulated microarray data (describe in figure 1), the number of genes *m* (100) is higher than the number of samples *n* (20); therefore, patients clustering based on this microarray data using GMM based approach was face the inverse problem of the variance-covariance matrix, In this situation feature selection is most important for especially, statistical theory based clustering algorithms like GMM. To avoid this difficulty, first we clustering (grouping) genes instead of samples (patients) by GMM approach, and observed that this method have recovered the three gene clusters (groups) as the original pattern of the simulated data (Figure 1). Now, we have to detect, which group contains EE genes out of these three genes groups (clusters) by the proposed FS techniques, and remove the EE gene's group from underlying datasets, then we finally select few (q) top *DE* genes from the reaming two DE's groups in order of magnitude for required patients clustering by GMM approach.

However, feature selection for GMM based clustering (especially, in case of gene expression data) using such a contributed R package *clustvarsel* (developed by Scrucca and Raftery, 2014) is very much time consuming and computationally intensive[59], which also gives misleading results for selecting most informative (top DE) genes indeed. At this moment, our proposed feature selection approach was reducing time and computational cost, which also selects most informative genes appropriately that ultimately minimize the misclassification error rate (MER) for patient clustering. However, the artificial gene expression data contain 5+5=10 DE genes (out of 100 genes), in which 5 DE genes comes from P_1 pattern, and others 5 DE genes comes from P_2 pattern. Model based FS (feature selection) method selects g.32 and g.40 genes for 1st DE gene's group, but these genes (g.32, g.40) was considered as equally expressed (EE) genes; whereas, our proposed FS approach select g.1 and g.2 genes for 1st DE gene's group, and these genes was considered as DE genes for this group indeed. And, for the 2nd DE gene's group, model based FS approach select g.98 gene as the most top DE gene, and our proposed FS approach selects g.96 and g.98 genes as top DE genes; these genes selecting from both approaches have considered as top DE genes for the 2nd DE gene's group really. As a result, these features, g.32, g.40 and g.98 (selected by model based FS method) can't recover the original sample group (Figure 2.c); but our proposed selecting features (g.1, g.2, g.96 and g.98) fully recover the original sample group (Figure 2.d). We have also checked the validity for determine the existing '*number of cluster*' in dataset by BIC plot, and observed the maximum BIC showed against 2 clusters (groups). Misclassification error rates (MER) of several mean differenced point shows that our proposed FS approach is better than Gaussian mixture model based FS approach (Figure 2.e); here, the minimum error rates for the model based and our proposed FS approaches are 0.50 and 0.20 respectively. That is, the proposed FS technique shows minimum error rate than model based FS method. Another problem of GMM model based FS is the computation time i.e., the benchmarking time is very high than our proposed FS approach. Hence, our proposed FS approach is out performing than model based FS method especially, for patient clustering in case of microarray gene expression datasets.

3.2 Example of Real Gene Expression Data Analysis

The colon cancer microarray gene expression dataset is one of the most popular dataset in the modeling of microarray gene expression data, in this dataset there are 62 samples and 2000 genes. In this study we used this dataset for samples clustering based on the limited number (*q<62*) of top DE genes, which are selected by our proposed FS approach, for patients clustering by GMM approach. The original pattern of colon cancer dataset has two sample groups (Fig 3.a). However, the main limitation of GMM based clustering approach is the inverse problem of variance-covariance matrix for higher number of genes than the samples. In this case our proposed feature selection approach for GMM based clustering reduced the dimension of given data due to avoid the singularity problem; here, it should be note that the number of genes must be less than the number of samples in this reduced dataset. Figure 3(b) shows that our proposed approach in this colon cancer data recover the original sample group; whereas, model based FS is very difficult to apply for this colon cancer dataset; because, this approach works as similarly as an embedded or greedy search techniques, and very much computationally intensive. Therefore, this method hasn't suitable for this type of high dimensional datasets.



(a) (b)

FGURE 3: (a) The original pattern of colon cancer dataset, (b) Clustering result by proposed FS approach.

4. CONCLUSION

Patient clustering using gene expression data is one of the major research areas in the medical field. Accurate clustering has great value to diagnosis patient and drug discovery. However, most studies of patients clustering are mainly clinical based and have limited diagnostic ability. In this paper, we propose GMM approach based on few selected top DE genes for patients clustering. First, we clustering genes instead of samples by GMM approach and get several genes' clusters (groups). Now, we propose a very simple rule (based on variance property) to detect the EE gene's cluster (group), and remove EE group from the given dataset, and then we select small number of top DE genes proportionally from each DE gene's group in order of magnitude. i.e., for each DE gene group, we ranking it's all DE genes (according to their variances) and select first top q_i informative genes accordingly i.e., for two DE gene's, we have (q_1+q_2=q) top DE genes in total, ultimately, these selected (q) top DE genes are used for required patient clustering by GMM approach. However, GMM estimation for patient clustering, in case of high dimensional but small sample sizes gene expression datasets, controlled by applying these few 'q' top DE genes; here, these 'q' top DE is used as an alternative to entire dataset, which also provide minimum misclassification error rate (MER). Therefore, to avoid the singularity problem of GMM, we propose to apply only few top DE genes for patients clustering. We also investigate the performance of our proposed technique on both simulated and real gene expression datasets. Finally, we conclude that our propose gene selection by Gaussian mixture model has a promising result for patient (sample) clustering.

In this paper, we have mainly focused on the development of feature selection algorithm by GMM especially, for patient clustering into two levels (binary classes) only; in future, we would like to extend our proposed algorithms with multi-conditional (more than two levels) patient (sample) clustering from the same platforms, and a detail comparative evaluation also be discussed. Furthermore, we will study others biological datasets (e.g., protein expression, tissue microarray data etc) in similar manners, in addition, to compare or correlate the changes in gene expression profiles with changes in proteomic or tissue profiles.

5. ACKNOWLEDGEMENTS

This work was supported by the sub-project (CP-3603, W2, R3) of Higher Education Quality Enhancement Project (HEQEP) of UGC, Bangladesh.

6. REFERENCES

[1] Michael A. Beer and Saeed Tavazoie. "Predicting Gene Expression from Sequence" *Cell*, Vol. 117, pp. 185–198, April 2004.

[2] Michael B. Eisen, Paul T.Spellman, Patrick O. Brown, and David Botstein. "Cluster analysis and display of ge-nome-wide expression patterns" *Proc. Natl. Acad. Sci. USA, vol* 95, pp. 14863–14868, Dec 1998.

[3] Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. & Yakhini, Z. "Tissue classification with gene expression profiles." *Journal of Computational Biology, vol 7*,pp. 559–584, 2000.

[4] Ben-Dor, A., Shamir, R. & Yakhini, Z. "Clustering gene expression patterns", *Journal of Computational Biology,* vol. 6 (3-4), pp. 281–97,1999.

[5] Brazma,A., Robinson,A., Cameron,G. and Ashburner,M. "One-stop shop for microarray data." *Nature,* vol. 403, pp. 699-700, 2000.

[6] D'haeseleer P. "How does gene expression clustering work?" Nat Biotechnol, vol. 23(12), pp.1499-1501, 2005.

[7] D'haeseleer, P., Wen, X., Fuhrman, S. and Somogyi, R. "Linear modeling of mRNA expression levels during CNS development and injury." *Pacific Symposium on Biocomputing* vol. 4, pp. 41-52, 1999.

[8] Alizadeh, A.A., et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling." *Nature*, vol. 403(6769), pp. 503-511, 2000.

[9] Golub, T.R., et al., “Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring.” *Science, vol.* 286(5439), pp. 531-537, 1999.

[10] Fraley, C. and Raftery, A.E. “Model-Based Clustering, Discriminant Analysis, and Density Estimation.”*Journal of the American Statistical Association*, vol. 97, pp. 611-631, 2002.

[11] Fraley, C. and Raftery, A.E. (1998). “How many clusters? Which clustering methods? Answers via model-based cluster analysis.” *Computer Journal*, vol. 41, pp. 578-588, 1998.

[12] Yeung KY, et al. “Model-based clustering and data transformations for gene expression data.” *Bioinformatics,* vol. 17(10), pp. 977-87, 2001.

[13] Ghosh, D. and Chinnaiyan, A. M. “Mixture modelling of gene expression data from microarray experiments.” *Bioinformatics*, vol. 18, pp. 275–286, 2002.

[14] McLachlan, G. J. Bean R. W. and Peel D. “A mixture model-based approach to the clustering of microarray expression data” *Bioinformatics*, vol. 18 (3), pp. 413–422, 2002.

[15] Wolfe, J. H. “Object cluster analysis of social areas.” *Master’s thesis, University of California, Berkele, .*1963.

[16] McLachlan, G., and Peel, D. (2000) “Finite mixture models” *New York,* John Wiley & Sons.

[17] Law, M. H., Jain, A. K., and Figueiredo, M. A. T. “Feature Selection in Mixture-Based Clustering,” in Proceedings of Conference of Neural Infor-mation Processing Systems, Vancouver, 2002.

[18] Vaithyanathan S. and Dom B. “Generalized Model Selection for Un-supervised Learning in High Dimensions,” in Proceedings of Neural Infor-mation Processing Systems ,eds. Solla S. A., Leen T. K. and Muller K.R., Cambridge, MA: MIT Press, pp. 970–976, 1999.

[19] Liu J. S., Zhang J. L., Palumbo M. J. and Lawrence, C. E. “Bayesian clustering With Variable and Transformation Selections,” in Bayesian Statistics, Vol.7, eds. Bernardo J.M., Bayarri M.J, Dawid A.P., Berger J.O., Heckerman D., Smith A. F. M. and West M., Oxford University Press, pp. 249–275, 2003.

[20] Ding C., He X., Zha H. and Simon H. D. “Adaptive Dimension Reduction for Clustering High-Dimensional Data,” in Proceedings of the IEEE International Conference on Data Mining, Maebashi, Japan, pp. 147–154, 2002.

[21] Chakrabarti, K., and Mehrotra, S. “Local Dimensionality Reduction: A New Approach to Indexing High-Dimensional Spaces.” *The VLDB Journal*, pp. 89–100, 2000.

[22] Mitra, P., Murthy, C. A., and Pal, S. K. “Unsupervised Feature Selection Using FeatureSimilarity.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 301–312, 2002.

[23] Talavera L. “Dependency-Based Feature Selection for Clustering Symbolic Data.” *Intelligent Data Analysis,* vol. 4, pp. 19–28, 2000.

[24] Datta S. and Datta, S. “Comparisons and validation of statistical clustering techniques for microarray gene expression data.” *Bioinformatics,* vol. 19, pp. 459-466, 2003.

[25] George Loewenstein, Scott Rick and Jonathan D. Cohen “Neuroeconomics” *Annu. Rev. Psychol.*, vol. 59, pp. 647–72, 2008, Article’s doi: 10.1146/annurev.psych.59.103006.093710.

[26] Pan W., Shen X., Jiang A. and Hebbel R.P. "Semi-Supervised Learning via Penalized Mixture Model with Application to Microarray Sample Classification." *Bioinformatics*, vol. 22, pp. 2381-2387, 2006.

[27] Pan W. "Incorporating gene functions as priors in model-based clustering of microarray gene expression data." *Bioinformatics*, vol. 22, pp. 795-801, 2006.

[28] Wang S. and Zhu J. "Variable selection for model-based high-dimensional clustering and its application to microarray data." *Biometrics*, vol. 64, pp. 440-448, 2008.

[29] Liu G., Loraine A.E., Shigeta R., Cline M., Cheng J., Valmeekam V., Sun S., Kulp D. and Siani-Rose,M.A. "NetAffx: affymetrix probesets and annotations." *Nucleic Acids Res*., vol. 31, pp. 82–86, 2003.

[30] Wall M., Rechtsteiner A. and Rocha L. "Singular Value Decomposition and Principal Component Analysis." In Berrar D., Dubitzky W. and Granzow M. (eds.), "A Practical Approach to Microarray Data Analysis." Springer US, pp. 91–109, 2003.

[31]Wall M. E., Dyck P. A. and Brettin T. S. "SVDMAN—singular Value decomposition analysis of microarray data." *Bioinformatics*, vol. 17(6), pp. 566–568, 2001.

[32] Alon U., Barka ,N., Notterman D.A., Gish K., Mack S.Y.D. and Levine J. "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays." Proc. Natl. Acad. Sci. USA, vol. 96, pp. 6745–6750, 1999.

[33] Hamadeh H. K. et al. "An overview of toxic-genomics." *Curr. Issues Mol. Biol*., vol. 4, pp. 45–56, 2002.

[34]Dy J. G. and Brodley C. E. "Feature Subset Selection and Order Identification for Unsupervised Learning." in Proceedings of the Seventeenth International Conference on Machine Learning, San Francisco, CA, pp. 247–254, 2000.

[35] Lazzeroni L. and Owen A. "Plaid Models for Gene Expression Data," *Statistica Sinica*, vol. 12, pp. 61–86, 2002.

[36] McCallum A., Nigam K. and Ungar L. "Efficient Clustering of High-Dimensional Data Sets With Application to Reference Matching." in Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 169–178, 2000.

[37] Brusco M. J. and Cradit J. D. "A Variable Selection Heuristic for k-Means Clustering." *Psychometrika*, vol. 66, pp. 249–270, 2001.

[38] Devaney M., and Ram A. "Efficient Feature Selection in Conceptual Clustering," in Machine Learning: Proceedings of the Fourteenth International Conference, *Nashville, TN*, pp. 92–97, 1997.

[39] Friedman J. H. and Meulman J. J. "Clustering Objects on Subsets of Attributes." *Journal of the Royal Statistical Society,* Ser.B, vol. 66, pp. 1–25, 2004.

[40] Gnanadesikan R., Kettenring J. R. and Tsao, S. L. "Weighting and Selection of Variables for Cluster Analysis." *Journal of Classification*, vol. 12, pp. 113–136, 1995.

[41] Desarbo W. S., Carroll J. D., Clarck L. A. and Green P. E. "Synthesized Clustering: A Method for Amalgamating Clustering Bases with Differential Weighting of Variables." *Psychometrika*, vol. 49, pp. 57–78, 1984.

[42] Kabir M. D. and Mollah M. N. H. "Outlier Modification and Gene Selection for Binary Cancer Classification using Gaussian Linear Bayes Classifier." *International Journal of Biometrics and Bioinformatics (IJBB)*, vol. 9(2), 2015.

[43] McLachlan G. J. and Basford K.E. “Mixture Models: Inference and Applications to Clustering.” *New York, Marcel Dekker, pp.* xi + 259 , 1988.

[44] Banfield J.D. and Raftery A.E. “Model-based Gaussian and non-Gaussian clustering.” *Biometrics,* vol. 49*,* pp. 803-821, 1993.

[45] Cheeseman P. and Stutz J. “Bayesian classification (auto-class): theory and results.” In: Fayyad,U.et al. (ed.), *Advances in Knowledge Discovery and Data Mining*, *AAAI Press, Menlo Park, CA,* pp. 61–83, 1995.

[46] Chickering D. M., Heckerman D. and Meek C. “A Bayesian approach to learning Bayesian networks with local structure.” *UAI*, pp. 80–89, 1997.

[47] Bozdogan H.“Mixture-model cluster analysis using a new informational complexity and model selection criteria.” In Bozdogan,H. (ed.), *Multivariate Statistical Modeling*, vol. 2, 1994, Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach, Kluwer Academic Publishers, Netherlands, Dordrecht, pp.69–113.

[48] Celeux G. and Soromenho G. "An entropy criterion for assessing the number of clusters in a mixture model". *Journal of Classification, vol.* 13, pp. 195-212, 1996.

[49] Biernacki C., Celeux G. and Govaert G. "An improvement of the NEC criterion for assessing the number of components arising from a mixture." *Pattern Recognition Letters*, vol. 20, pp. 267-272, 1999.

[50] Biernacki C., Celeux G. and Govaert G. “Assessing a mixture model for clustering with the integrated completed likelihood.” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 22(7), pp. 719–725, 2000.

[51] Biernacki C. and Govaert G. “Choosing models in model-based clustering and discriminant analysis.” *J. Stat. Comput. Simul*., vol. 64, pp. 49–71, 1999.

[52] Bensmail H., Golek J., Moody M. M., Semmes O. J. and Haoudi A. “A novel approach for clustering proteomics data using Bayesian fast Fourier transform.” *Bioinformatics, v*ol. 21(10), pp. 2210–2224, 2005, doi:10.1093/bioinformatics/bti383.

[53] Schwarz G. “Estimating the Dimension of a Model.” *The Annals of Statistics,* vol. 6(2), pp. 461-464, 1978.

[54] Haughton D. “On the choice of a model to fit data from an exponential family.” *Annals of Statistics*, vol. 16(1), pp. 342–355, 1988.

[55] Kass R. E. and Wasserman L. “The Selection of Prior Distribution by Formal Rules.” *Journal of the American Statistical Association (JASA)*, vol. 91, No. 435, pp. 1343-1370, Sep 1996.

[56] Raftery A. E. “Bayesian model selection in social research.” *Journal of the American Statistical Association (JASA)*, vol. 90, No. 430, pp. 773-795, Jun 1995.

[57] Leroux B. G. “Maximum-likelihood estimation for hidden Markov models.” *Stochastic Processes and their Applications*, vol. 40, pp. 127-143, 1992.

[58] Keribin C. “Consistent estimation of the order of mixture models.” *The Indian Journal of Statistics.* Series A 62(1), pp. 49–66, 2000.

[59] Scrucca L. and Raftery A. E. “clustvarsel: A Package Implementing Variable Selection for Model-based Clustering in R.” Technical Report no. 629, Department of Statistics, University of Washington. Also arXiv:1411.0606, 2014.