Research Journal of Applied Sciences, Engineering and Technology 11(7): 692-700, 2015
DOI: 10.19026/rjaset.11.2032
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: February 26, 2015 Accepted: March 14, 2015 Published: November 05, 2015

Research Article

Hybrid Algorithm for Clustering Gene Expression Data

[1]S. Jacophine Susmi, [1]H. Khanna Nehemiah, [2]A. Kannan and [1]G. Saranya
[1]Ramanujan Computing Centre, Anna University, Chennai 600025, India
[2]Department of Information Science and Technology, Anna University, Chennai 600025, India

Abstract: Microarray gene expressions provide an insight into genomic biomarkers that aid in identifying cancerous cells and normal cells. In this study, functionally related genes are identified by partitioning gene data. Clustering is an unsupervised learning technique that partition gene data into groups based on the similarity between their expression profiles. This identifies functionally related genes. In this study, a hybrid framework is designed that uses adaptive pillar clustering algorithm and genetic algorithm. A first step towards, the proposed work is the utilization of clustering technique by adaptive pillar clustering algorithm that finds cluster centroids. The centroids and its clustering elements are calculated by average mean of pairwise inner distance. The output of adaptive pillar clustering algorithm results in number of clusters which is given as input to genetic algorithm. The microarray gene expression data set considered as input is given to adaptive pillar clustering algorithm that partitions gene data into given number of clusters so that the intra-cluster similarity is maximized and inter cluster similarity is minimized. Then for each combination of clustered gene expression, the optimum cluster is found out using genetic algorithm. The genetic algorithm initializes the population with set of clusters obtained from adaptive pillar clustering algorithm. Best chromosomes with maximum fitness are selected from the selection pool to perform genetic operations like crossover and mutation. The genetic algorithm is used to search optimum clusters based on its designed fitness function. The fitness function designed minimizes the intra cluster distance and maximizes the fitness value by tailoring a parameter that includes the weightage for diseased genes. The performance of adaptive pillar algorithm was compared with existing techniques such as k-means and pillar k-means algorithm. The clusters obtained from adaptive pillar clustering algorithm achieve a maximum cluster gain of 894.84, 812.4 and 756 for leukemia, lung and thyroid gene expression data, respectively. Further, the optimal cluster obtained by hybrid framework achieves cluster accuracy of 81.3, 80.2 and 78.2 for leukemia, lung and thyroid gene expression data respectively.

Keywords: Adaptive pillar algorithm, average pairwise inner distance, clustering, genetic algorithm, microarray gene expression data

INTRODUCTION

Gene expression profiling methods developed in the recent decades have initiated several new challenges in molecular biology research (Bandyopadhyay and Bhattacharyya, 2011). The expression profiles of genes signifies the amount of messenger RNAs (mRNAs) produced by that gene under a specific experimental condition. These expression values, over a set of time points or under different tissue samples, are frequently analyzed to study the functional coherence of genes (Bandyopadhyay and Bhattacharyya, 2011). Microarray technology monitors the expression levels of thousands of genes simultaneously and provides genetic dissection of complex diseases (Young Kim *et al*., 2005; Bose *et al*., 2013; Bidaut *et al*., 2006). The prediction of diseased genes enables physicians to understand the causes of many diseases and identifies therapeutic strategies (Yin and Chiang, 2008). However, the challenge lies in the large number of genes and the complexity of biological networks greatly increases. Data mining integrated with bioinformatics provides a way to identify key genes for predicting diseased patient and to allow the investigation of complex disease at the molecular level (Li *et al*., 2005). To address, this challenge data mining is integrated with bioinformatics that involves several tasks namely classification, clustering, prediction, affinity grouping, association rule mining and description. Classification is one of the most common data mining tasks that involve supervised learning by examining the features of a newly presented object in order to assign it to one of the predefined set of classes (Labib and Malek, 2005). Clustering is one of the unsupervised learning that does not rely on predefined classes and training while classifying the data objects (Karayiannis and Mi, 1997). Clustering partitions a given data set into groups based on specified features so that the data points

Corresponding Author: H. Khanna Nehemiah, Ramanujan Computing Centre, Anna University, Chennai 600025, India

within a group are more similar to each other than the points in different groups. Thus, clustering is distinguished from pattern recognition or the areas of statistics known as discriminant analysis and decision analysis, which seek to find rules for classifying objects from a given set of pre-classified objects. The main characteristic of gene expression data is to cluster both genes and samples. On one hand, coexpressed genes can be grouped in clusters based on their expression patterns. On the other hand, the samples can be partitioned into homogeneous groups (Park, 2004).

Clustering of cancer samples has become common and leads high throughput in cancer studies. Once cancer signatures are identified on a genomic level, specific drugs can be developed, improving treatment efficacy while reducing its side effects (Golub *et al*., 1999). The clustering application concerning gene expression data is found when genes that show similar expression patterns are clustered together (Jaskowiak *et al*., 2014). Several machine learning algorithms has been proposed to analyze gene expression data such as k-Nearest Neighbour (k-NN) classifiers, naive bayes classifiers, k-means clustering, decision tree algorithm and support vector machine.

In this study, a hybrid framework is proposed for gene selection approach based on adaptive pillar clustering algorithm and genetic algorithm. The hybridization results in identification of key feature genes. By grouping the genes using adaptive clustering algorithm gives similar expression values in each cluster and by selecting optimal cluster using genetic algorithm results in informative genes.

LITERATURE REVIEW

Park (2004) has presented active sampling applied to multidimensional space using Orthogonal Pillar Vectors (OPV). To remove the randomness in the sampling process, an optimal active sampling algorithm was devised for sampling the high-dimensional space using a few selected vectors. These selected vectors, called the pillar vectors, are orthogonal to each other and lie on the current estimate decision boundary. The algorithm sequentially sample class labels at each pillar vector and then updates both the estimate decision boundary and the pillar vector. The active sampling at the boundary method facilitates identifying an optimal decision boundary for pattern classification in machine learning. The result of this method is compared with the standard active learning method that uses random sampling on the decision boundary hyperplane. The validity of active sampling using pillar vectors is shown through convergence by performing real-world experiment from UCI Machine learning archive. The application selects the tic-tac-toe problem and Wisconsin breast cancer data to test the convergence. The convergence is proven in probability for active sampling method that uses the orthogonal pillar vectors tested and compared using perceptron update and support vector machine update. The orthogonal pillar vectors sampling method avoid sensitivity in sparse data distribution and skewed ratio of positive and negative labelled data compared to active and random sampling. The higher dimension data with sparse and non-uniform data, sampling that uses random selection overfits to local optima however, OPV sampling finds global optima.

Barakbah and Kiyoki (2009) have proposed a pillar algorithm for optimizing initial centroids for k-means clustering. The initial centroids are calculated by farthest accumulated distance between each data. The accumulated distance metric between each data point and their grand mean is created. The first initial centroid is selected from the data point with maximum accumulated distance metric. The next centroid is selected by comparing the data point and the previous initial centroid and then a data point with highest maximum accumulated distance is selected as new centroid. The process is repeated till all the centroids for data points are designated. The pillar algorithm also introduces detecting outliers. The performance of the pillar algorithm for optimization of k-means clustering is experimented with seven benchmark datasets taken from UCI repository Ruspini, Fossil, Iris, Thyroid, Wine, Glass, Heart and Ionosphere. The performance of pillar algorithm with k-means outperforms by its validity measurement v_w with 181.42 with Ruspini datasets when compared with other existing classical approaches k-means with Forgy approach, k-means with Mac Queen, k-means with refinement, k-means with MDC, k-means with Kaufman, k-means with random initialization.

Robideau *et al*. (2011) has presented a study on DNA barcoding with cytochrome c oxidase subunit I (COI) assessed over a significant sample of oomycete genera. In this study COI is sequenced from isolating that represents genera. A comparison to Internal Transcribed Spacer (ITS) sequences from the same isolates showed that cytochrome c oxidase subunit I (COI) identification is a practical option; complementary because it uses the mitochondrial genome instead of nuclear DNA. In some cases COI was more discriminative than ITS at the species level. Distance matrices were analysed using matrix algebra and SAS. The average intra-specific distance was calculated for each species represented by more than one strain and coded as missing data when only one strain could be obtained to avoid having a bias towards zero variation. For each pair of species, the average pair wise distance was calculated for all the possible strain comparisons. The total of the distances [TD] for each species and pairwise comparison was found by specifying the diagonal as the number of pairwise comparisons for each species and the lower triangular matrix as the total number of possible pairwise comparisons for each pair of species. A lower triangular matrix with a diagonal of 1's was created with the same number of rows and columns. The total number of pairwise distance comparisons is denoted by [ND]. The average of all the pairwise comparisons was found by

dividing total distances by number of pairwise distances, with the diagonal of the matrix giving the averages of all intraspecific comparisons and the lower matrix the averages of all interspecific comparisons.

Muda *et al*. (2009) have proposed a phylogenetic tree construction using distance-based method. The UPGMA calculates the average of possible pairwise distances to get a new distance in the clustering process. If outliers exist in the possible pairwise distances, new mean distances are calculated and the result is not robust. To overcome this problem, we implement a checking process to detect the outliers using MAD*n* criteria and the new distances using the modified one-step M-estimator (MOM). In order to evaluate the branch of the tree constructed, the bootstrap method is used and the *p*-value (bootstrap value) for both methods is compared.

Li *et al*. (2005) have proposed a gene selection method that builds hybrid between Genetic Algorithm and Support Vector Machine (GA-SVM) that utilizes two data mining tools. Genetic algorithm is used in the search engine, while support vector machine is used as the classifier. The proposed approach is hybridized to identify key feature genes. First, initial population is generated randomly with the fixed-length binary strings for N individuals. Each string represents a feature subset and the values at each position in the string are coded as either presence or absence of a particular feature. Then, fitness is calculated for each feature subset that is evaluated using linear SVM. The classification accuracy is obtained as fitness index. Genetic operators cross over and mutation is performed for selecting different feature subsets. The genetic algorithm is an iterative process in which each successive generation is produced by applying genetic operators to the members of the current generation. In this manner, iterations are repeated until good feature subsets are obtained. The classifier results in selecting optimal gene subset. This application has demonstrated to large B cell lymphoma for mining high dimensional data. The results were compared with GA-SVM and existing GA-kNN which obtains 99 and 95%, respectively.

Krishna and Murty (1999) proposed a novel hybridization of Genetic Algorithm (GA) with gradient descent algorithm used in clustering is K-means algorithm that finds a globally optimal partition of a given data into a specified number of clusters. To overcome the drawback of K-means algorithm from local minimum the improved K-means operator is applied that defines a search operator instead of crossover. A biased mutation operator is applied that is specific to clustering called distance-based-mutation. Finite Markov chain theory is also used to prove that the Genetic K-means Algorithm (GKA) converges to the global optimum. It is observed in the simulations that GKA converges to the best known optimum corresponding to the given data and also it attains search faster than some of the other evolutionary algorithms used for clustering.

Bishnu and Bhattacherjee (2012) have proposed a Quad tree based K-means algorithm (QDK) for predicting faults in program modules. Clustering is an unsupervised technique used for fault prediction in software modules. The Quad tree based initialization algorithm finds initial cluster centres for k-means algorithm. The concept of clustering gain is used as a metric to determine the quality of clusters for evaluation of the Quad Tree-based initialization algorithm as compared to other initialization techniques. The clusters obtained by Quad tree-based algorithm were found to have maximum gain values. In addition, Quad Tree based algorithm is also applied for predicting faults in program modules. The Quad tree initialization algorithm is demonstrated on four data sets namely AR3, AR4, AR5 and Iris datasets that are related to software fault prediction. The gain values for Quad tree based K-means algorithm are nearly close or equal in all the cases except AR4 data set. This indicates that cluster quality obtained by QDK is comparable with six techniques namely K-means algorithm, naive bayes, linear discriminant analysis, two stage approach and single stage approach. The overall error rates of this prediction approach are compared to other existing algorithms. The values of FPR, FNR and Error are presented for six techniques. The FPR values for QDK algorithm are improved for AR3, AR4 and AR5 data sets and FNR values are same as two stage approach and single stage approach except in the case of AR4 data sets. The overall error rates are reduced with two stage approach and single stage approach for AR3, AR4 as well as AR5 data sets.

Jacophine Susmi *et al*. (2015) have proposed a hybrid technique for classification of leukemia gene data by combining two classifiers namely, Input Discretized Neural Network (IDNN) and Genetic Algorithm-based Neural Network (GANN). The leukemia microarray gene expression data is preprocessed using probabilistic principal component analysis for dimension reduction. The dimension reduced data is subjected to two classifiers: first, an input discretized neural network and second, genetic algorithm-based neural network. In input discretized neural network, fuzzy logic is used to discretize the gene data using linguistic labels. The discretized input is used to train the neural network. The genetic algorithm-based neural network involves feature selection that selects subset of genes. The subset of genes is selected by evaluating fitness for each chromosome (solution). The subset of features with maximum fitness is used to train the neural network. The hybrid classifier designed, is experimented with the test data by subjecting it to both the trained neural networks simultaneously. The hybrid classifier employs a distance based classification that utilizes a mathematical model to predict the class type. The model utilizes the output values of IDNN and GANN with respect to the distances between the output and the median threshold, thereby predicting the class type. The performance of the hybrid classifier is compared with

existing classification techniques such as neural network classifier, input discretized neural network and genetic algorithm-based neural network. The comparative result shows that the hybrid classifier technique obtains accuracy rate of 88.23% for leukemia gene data.

Comparing to the works discussed in the literature, the work presented in this study differs in the following ways:

The novelty of the hybrid approach is the framework that combines the adaptive pillar algorithm and genetic algorithm. The output of adaptive pillar clustering algorithm is input to genetic algorithm. The novelty lies in the utilization of clustering technique by adaptive pillar clustering algorithm that finds the cluster centroid. Rather than using grand mean of data points (Barakbah and Kiyoki, 2009) this system makes use of average mean of pairwise inner distance calculation. Further, in existing work the genetic algorithm searches for optimum cluster and calculates fitness function which gives the classification accuracy (Li *et al*., 2005) whereas, in our work the fitness function is tailored to incorporate a weighted parameter for the diseased genes that minimizes the intra cluster distance and maximizes the fitness value.

METHODOLOGY

System framework: The framework of the system is illustrated in Fig. 1. The system framework makes use of two techniques:

i. Clustering
ii. Genetic algorithm

The clustering technique uses adaptive pillar clustering algorithm that gives gene expression data as input. The adaptive pillar clustering algorithm clusters the given gene expression data into number of clusters that partitions given dataset into groups based on specified features so that the data point within a group are more similar to each other than the points in different groups. Clustering algorithm adapted in the framework aims at identifying key featured genes exhibiting similar patterns of variation in expression level. The existing pillar k-means algorithm proposed by Barakbah and Kiyoki (2009) is enhanced by applying its adaptive nature by calculating average mean of pairwise inner distance that is used to partition gene data. The adaption of pillar algorithm for gene expression data finds the cluster centres. The cluster centers identified were chosen as pillars. The algorithm estimates pillar vectors proposed by Park (2004) thus avoiding randomness which results in local minima. Second, the genetic algorithm has been used as an optimisation technique for clustering algorithm. The output of adaptive pillar algorithm is given as input to genetic algorithm. GA performs search in large dimensional space and provides optimal cluster (solution) for fitness function. The effectiveness of genetic algorithm provides optimal cluster so the maximization of fitness function leads to minimization *G* as defined in Eq. (9). Further, in existing work the genetic algorithm calculates fitness function that gives the classification accuracy (Li *et al*. 2005) whereas, in this devised system the fitness function designed is tailored with a parameter that includes weight age for diseased genes that minimizes the intra cluster distance and maximize the fitness value.

Adaptive pillar clustering algorithm: The microarray gene expression data is considered as input to adaptive pillar algorithm. The adaptive pillar algorithm has been tailored to compute the average mean of pairwise inner distance value as a parameter in calculating similarity measure. The modified parameter in the distance measure is tailored in two steps: step 2 and step 3 of the algorithm. The execution step of the adaptive pillar algorithm is described below. The resultant of adaptive pillar algorithm is the cluster centres were computed and similarity grouping was done accordingly.

Algorithm:

Input: Gene expression dataset can be represented by a real valued expression matrix $X = \{G_{ij} \mid 1 \le i \le n, 1 \le j \le m\}$ where the rows $G = \{g_1, g_2, g_3, \ldots g_n\}$ represent the expression patterns of genes, the columns $S = \{s_1, s_2, s_3, \ldots s_m\}$ represent the expression profiles of individuals and G_{ij} is the measured expression level of gene i in individual j.

Step 1: Initially the algorithm initializes the cluster *C* with cluster centre *c* and *SG* is selected genes with maximum distance where, $C = \{\}$, $SG = \{\}$. The set *SG* is represented by $SG = \{SG_1, SG_2, SG_3, \ldots SG_n\}$

Step 2: Compute the sum of pairwise inner distance for each individual expression patterns for all possible pairs corresponding to each gene is computed using the mathematical model presented in Eq. (1):

$$D(i) = \sum_{j=1}^{m} \sum_{k=1}^{m} \left| G_{i,j} - G_{i,k} \right|, \forall i, i = 1 \ldots . n \quad (1)$$

where, n is the total number of genes; m is the total number of individual expression pattern corresponding to each gene; $G_{i,j}$, $G_{i,k}$ represents individual expression pattern corresponding to each gene.

Step 3: Compute the mean deviation $E(i, j)$ for each gene:

$$z(i) = \frac{D(i)}{m^2}, \forall i, where\ i = 1 \ldots . n \quad (2)$$

Step 4: Compute the deviation of each individual expression pattern from the average mean value for each gene using the mathematical model is presented in Eq. (3):

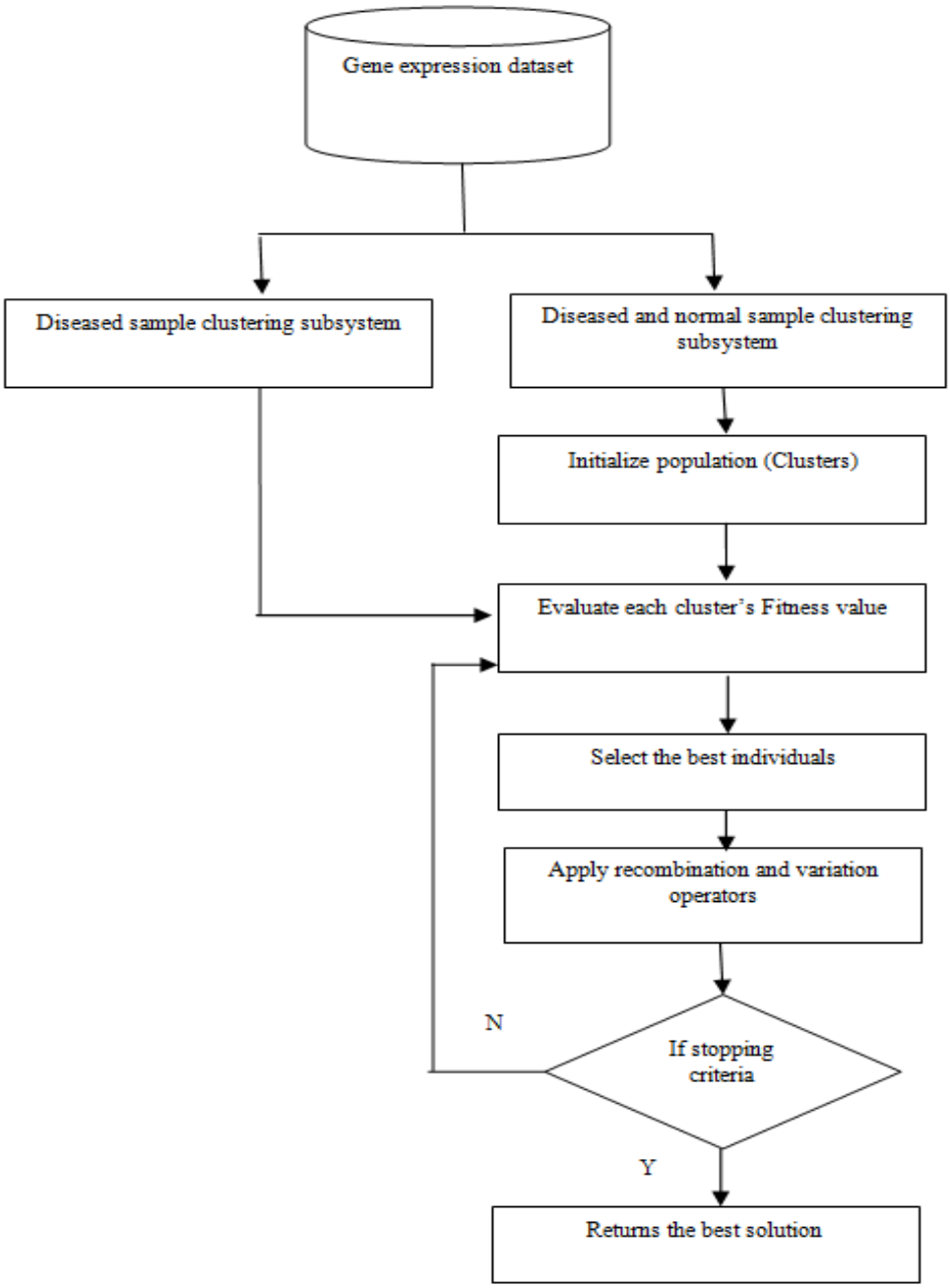

Fig. 1: System framework

$$E(i,j)=\left|G_{i,j}-z(i)\right|,\forall i,j,i=1....n,j=1....m \quad (3)$$

Step 5: Deviation of each individual expression pattern corresponding to each gene is arranged in descending order ranging from maximum to minimum and assigned it to set SG:

$$SG(i)=MAX\left\{E(i,j)\right\} \quad (4)$$

The maximum deviated value is chosen as the candidate cluster centre.

Step 6: Cluster centres are formed by satisfying the following two conditions:

Condition I: The neighbourhood boundary of the candidate cluster centre is computed using the mathematical model presented below:

$$nbis(i)=\gamma\, MAX(i),\forall i=1....p \quad (5)$$

where, p is the total number of deviated genes, γ is a probabilistic value set as 0.3. If the elements (expression patterns) are more than 30% of deviation

from the selected candidate cluster centre then those elements are not chosen as a cluster element using the mathematical model is presented in Eq. (6):

$$ND(i,j) = G_{i,j} - c_k \quad (6)$$

If condition I is satisfied, then the deviated value can be considered as a cluster centre subject to satisfying condition II.

Condition II: Minimum number of elements in a cluster is set as 0.2. If the minimum number of elements in the cluster is less than 20% then it is classified as an outlier. The mathematical model for computing minimum number of cluster element is presented below:

$$nmin = \alpha.\frac{n}{k} \quad (7)$$

where, α is a probabilistic value set as 0.2, n is number of gene expression patterns and k is the number of clusters.

Step 7: If the above two conditions are satisfied then the selected deviated value is chosen as cluster centre else the next highest deviated value is chosen as cluster centre and the above exercise is repeated.

Step 8: The next highest deviated value is chosen as the next candidate cluster center and the cluster elements are formed by repeating step 6 and then by recomputing the distance as presented in Eq. (8):

$$DM(i,j) = E(i,j) + ND(i,j) \quad (8)$$

Step 9: The process of selecting cluster centers continues till 25 cluster centers are identified. Selected cluster centre with maximum deviated value is assigned as zero for rest of the iterations.

Output: C_k clusters where $0<k<25$.

The diseased and normal sample clustering subsystem and diseased sample clustering subsystem uses the adaptive pillar clustering algorithm to cluster both genes and samples. In diseased and normal sample clustering subsystem the genes are treated as objects and samples are treated as features. On the other hand, the samples can be partitioned into homogeneous groups. Such diseased sample clustering subsystem treats samples as objects and genes as features. Within a gene expression matrix G_{ij}, there are several macroscopic phenotypes of samples related to some diseases or drug effects, such as diseased samples, normal samples, or drug treated samples. The goal of diseased sample clustering focuses to find the genes responsible for the disease. Thus, from the gene expression matrix diseased samples are taken into consideration. The phenotypes of samples can be discriminated through only a small subset of genes whose expression levels strongly correlate with the class distinction (Golub *et al*., 1999). These genes are called informative genes. The output of adaptive pillar clustering algorithm results in 25 clusters from diseased and normal sample clustering subsystem with set $C = \{c_1, c_2, c_3,\ldots c_{25}\}$ where $c_1, c_2, \ldots c_{25}$ are cluster centres and 10 clusters are formed from diseased sample clustering subsystem that results with set $D = \{d_0, d_1, d_2,\ldots d_n\}$. The resultant clusters are given as input to genetic algorithm for optimization.

Genetic algorithm: Genetic algorithm is an evolutionary search algorithm that initializes population of clusters obtained as an input from adaptive pillar algorithm. The clusters initialized are evaluated by its designed fitness function. The designed fitness computation function defined in Eq. (9) leads in maximization of the fitness function leads to minimization of G. It minimizes the intra cluster distance and maximizes the fitness value by tailoring a parameter (β) that includes the weight age for diseased genes. The approach of randomly generating a new set (generation) of solutions from an existing one, so that there is improvement in the quality of the solutions throughout generations. The optimum cluster in the population is selected by the fitness value defined in Eq. (9). The GA optimizes the search space which yields the optimum cluster that contains high influential genes. The steps of genetic algorithm in our framework are detailed below:

Algorithm:

Input: Initial population of clusters obtained from adaptive pillar algorithm.

Step 1: A population of k individuals is initialized, where k is the total number of clusters.

Step 2: The fitness of each cluster is determined as follows:

$$f(c_k) = \frac{\alpha + \beta}{G} \text{ where, } 0<i<k \quad (9)$$

$$\beta = \sum \omega(d_i) \text{ where, } \omega(d_i) = \begin{cases} 1, if\ d_i \in D \\ 0, otherwise \end{cases} \quad (10)$$

$$G = \sum_{i=1}^{q} \left\| G_i - c_i \right\| \quad (11)$$

where, $\alpha = 0.1$; $D = \{d_0, d_1, d_2,....d_n\}$; $d_i = G_i - c_i$

β represents the significance of diseased genes

D is the set of all diseased genes

q is the number of genes in cluster c_i

$G_i \in C_i$ where, c_i is the center of cluster C_i

Step 3: The individual with maximum fitness are selected and placed in the selection pool for crossover and mutation.

Step 4: The new population is generated by performing single point crossover operation with a rate of 0.8 and bit flip mutation operator with a rate of 0.01.

Step 5: The new chromosomes obtained are then placed in the selection pool.

Step 6: Repeat the process from step 2 for I_{max} number of times until individual with maximum fitness is extracted from the population pool.

The output of genetic algorithm returns the best solution with optimal cluster. The binary encoding format is used for representing the individuals. If the j^{th} gene belongs to i^{th} cluster where $1<i<25$, $1<j<n$ then '1' is assigned to the i^{th} individual of the i^{th} allele. The fitness value is computed for each individual (clusters). The fitness value of a solution depends on the intra cluster distance. Since the objective is to minimize G, a solution with relatively small square error must have relatively high fitness value. The genetic operator's recombination and variation are done using single point crossover and bit flip mutation operator. The stopping criterion for the algorithm is until maximum fitness is reached or till a maximum generation (I_{max}) is reached. From the resultant solution the optimal cluster with maximum fitness defined in Eq. (9) is obtained as small subset of genes whose expression levels strongly correlate with the class distinction (i.e.,) diseased sample. These genes are found to be informative genes.

RESULTS AND DISCUSSION

The hybrid genetic adaptive pillar algorithm is implemented in MATLAB (version 2013a) and the results are evaluated for three microarray gene expression dataset namely leukemia taken from http://www.broadinstitute.org/cancer/software/genepattern/datasets website, lung and thyroid http:// lifesciencedb.jp/cged/website. The leukemia dataset contains expression levels of 7129 genes taken over 72 samples, lung cancer dataset contains expression levels of 2500 genes taken over 129 samples and thyroid cancer dataset contains expression levels of 2517 genes taken over 120 samples. The microarray gene expression data considered as input clusters both genes and samples using adaptive pillar clustering algorithm. The cluster from the diseased and normal sample clustering subsystem is used to initialize the individuals of genetic algorithm. The cluster from the diseased sample clustering subsystem is used for evaluating the fitness of each individual in the genetic algorithm. In the adaptive pillar algorithm of diseased and normal sample clustering subsystem the distance between the clusters is proportional to the information content of the gene, where the information content provides disease related information obtained from the cluster of genes. In diseased sample clustering subsystem the samples are clustered into 10 clusters in order to find the genes responsible for the diseases. The distance between the clusters is proportional to the severity of the disease. If the genes found in the clusters of the latter subsystem are found in the former, then that clusters of genes in the former is said to contain optimal set of genes to identify diseased samples.

Table 1: Gain values and error rate for various datasets

Data set	Techniques	Gain	Error (%)
Leukemia	K-mean	893.580	22.33
	Pillar K-mean	894.012	20.43
	Adaptive Pillar clustering algorithm	894.841	19.98
Lung Cancer	K-mean	811.5100	22.33
	Pillar K-mean	811.4220	22.33
	Adaptive Pillar clustering algorithm	812.4170	19.37
Thyroid	K-mean	753	21.33
	Pillar K-mean	754	20.33
	Adaptive Pillar clustering algorithm	756	18.37

The performance of the adaptive pillar clustering algorithm has been compared with k-means and pillar k-means algorithm and is evaluated using cluster gain for various dataset as shown in Table 1. The clustering gain defined in Eq. (12) attains a maximum value at the optimum number of clusters (Bishnu and Bhattacherjee, 2012). For the experimental setup k value has been chosen as 25 clusters for all the three datasets and to compare our clustering quality with k-means and pillar k-means algorithm. The k-values were executed with same number of clusters for all the three gene expression data which gave a maximum gain values for adaptive pillar algorithm. The gain values of adaptive pillar algorithm are comparable with k-means and pillar k-means. In fact the gain values for adaptive pillar and pillar k-means are nearly close in the case of leukemia gene data while gain values are improved in the case of lung cancer and thyroid cancer gene expression data, respectively.

The evaluation metric error rate is calculated by predicting the false positive rate and false negative rate. True negative rate indicates that the centre of cluster is lesser than the threshold which represents a non faulty mode. False positive rate indicates that the centroid point of cluster is greater than the threshold point which represents faulty mode. Similar definitions hold for false negative and true positive (Bishnu and Bhattacherjee, 2012). The values of error rates are also presented for the three gene expression datasets. The overall error rate defined in Eq. (14) is compared and is found better for adaptive pillar algorithm in all the cases:

$$Gain = \sum_{k=1}^{K} (v_k - 1)\left\| z_o - z_o^k \right\|_2^2 \quad (12)$$

$$z_o = \frac{1}{v}\sum_{i=1}^{v} o_i \quad z_o^k = \frac{1}{v_k}\sum_{i=1}^{v_k} o_i^{(k)} \quad (13)$$

Table 2: Performance of accuracy on various datasets

SI. No	Datasets	Accuracy (%)
1	Leukemia	81.3
2	Lung	80.2
3	Thyroid	78.2

where,

K is the number of clusters
v_K is the number of data points present in k^{th} cluster
z_o is the global centroid
v is the total number of data points
o is the data points
z_o^k is the centroid of the k^{th} cluster
$o_i^{(k)}$ is the data points belong to k^{th} cluster

$$Error\ rate = \frac{B + C}{A + B + C + D} \quad (14)$$

where,

A is the true negative rate
B is the false positive rate
C is the false negative rate
D is the true positive rate

Further, the performance of hybrid genetic algorithm with adaptive pillar clustering algorithm selects an optimal cluster. The performance of optimal cluster is measured by its accuracy defined in Eq. (15). The accuracy of clustering on three gene expression datasets is shown in Table 2:

$$Accuracy = \frac{a}{a + b} \times 100\% \quad (15)$$

where,

a is the number of informative genes in same cluster
b is the number of non informative genes in same cluster

The informative genes can be obtained from the value of (β) defined in Eq. (10).

CONCLUSION

This study uses a hybrid framework for selection of informative genes. The framework utilizes adaptive pillar clustering algorithm and genetic algorithm tested with three microarray gene expression datasets namely leukemia, lung and thyroid data. The performance of the hybrid framework was compared with k-means and pillar k-means algorithm. The hybrid framework has effectively selected the optimal cluster with improved clustering accuracy. Further, this study can be extended to temporal and time-series data.

REFERENCES

Bandyopadhyay, S. and M. Bhattacharyya, 2011. A biologically inspired measure for coexpression analysis. IEEE ACM T. Comput. Bi., 8(4): 929-942.

Barakbah, A.R. and Y. Kiyoki, 2009. A pillar algorithm for K-means optimization by distance maximization for initial centroid designation. Proceeding of IEEE Symposium on Computational Intelligence and Data Mining (CIDM'09), pp: 61-68.

Bidaut, G., F.J. Manion, C. Garcia and M.F. Ochs, 2006. Wave Read: Automatic measurement of relative gene expression levels from microarrays using wavelet analysis. J. Biomed. Inform., 39(4): 379-388.

Bishnu, P.S. and V. Bhattacherjee, 2012. Software fault prediction using quad tree-based K-means clustering algorithm. IEEE T. Knowl. Data En., 24(6): 1146-1150.

Bose, S., C. Das, T. Gangopadhyay and S. Chattopadhyay, 2013. A modified local least squares-based missing value estimation method in microarray gene expression data. Proceeding of IEEE 2nd International Conference on Advanced Computing, Networking and Security (ADCONS), pp: 18-23.

Golub, T.R., D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov and H. Coller, 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, 286(5439): 531-537.

Jacophine Susmi, S., H. Khanna Nehemiah, A. Kannan and J. Jabez Christopher 2015. A hybrid classifier for leukemia gene expression data. Res. J. Appl. Sci. Eng. Technol., 10(2): 197-205.

Jaskowiak, P.A., R.J. Campello and I.G. Costa , 2014. On the selection of appropriate distances for gene expression data clustering. BMC Bioinformatics, 15(S-2): S2.

Karayiannis, N.B. and G.W. Mi, 1997. Growing radial basis neural networks: Merging supervised and unsupervised learning with network growth techniques. IEEE T. Neural Networ., 8(6): 1492-1506.

Krishna, K. and M.N. Murty, 1999. Genetic K-means algorithm. IEEE T. Syst. Man Cy. B, 29(3): 433-439.

Labib, N.M. and M.N. Malek, 2005. Data mining for cancer management in Egypt case study: Childhood acute lymphoblastic Leukemia. World Acad. Sci. Eng. Technol., 8(61): 309-314.

Li, L., W. Jiang, X. Li, K.L. Moser, Z. Guo, L. Du, Q. Wang, E.J. Topol, Q. Wang and S. Rao, 2005. A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. Genomics, 85(1):16-23.

Muda, N., A.R. Othman, N. Najimudin and Z.A.M. Hussein, 2009. The phylogenetic tree of RNA polymerase constructed using MOM method. Proceeding of IEEE International Conference of Soft Computing and Pattern Recognition (SOCPAR'09), pp: 484-489.

Park, J.M., 2004. Convergence and application of online active sampling using orthogonal pillar vectors. IEEE T. Pattern Anal., 26(9): 1197-1207.

Robideau, G.P., A.W. De Cock, M.D. Coffey, H. Voglmayr, H. Brouwer and K. Bala, 2011. DNA barcoding of oomycetes with cytochrome c oxidase subunit I and internal transcribed spacer. Mol. Ecol. Resour., 11(6): 1002-1011.

Yin, Z.X. and J.H. Chiang, 2008. Novel algorithm for coexpression detection in time-varying microarray datasets. IEEE ACM T. Comput. Bi., 5(1): 120-135.

Young Kim, S., J. Won Lee and J. Sung Bae, 2005. Iterative clustering algorithm for analyzing temporal patterns of gene expression. World Acad. Sci. Eng. Technol., 4(3): 8-11.