Volume 4, Issue 9, September 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Classification of Unknown Sample by Prediciting Its Class Label by Using Hybrid K-Means PSO

Nirjharinee Parida
Computer Sc. & Engg.
Synergy Institute of Engg. and Technology
India

Narendra Kumar Rout
Computer Sc. & Engg.
Gandhi Engineering College
India

Abstract- Microarray is the collection of gene behavior under multiple conditions. Fast and high-quality clustering algorithms play an important role in helping researchers to effectively navigate, summarize, and organize the information. Recent advancement in clustering algorithms made it possible to concurrently monitor the expression levels of thousands of genes and across collection of related samples. Cluster analysis refers to partitioning a given data set into groups based on specified features so that the data points within a group are more similar to each other than the points in different groups. Many conventional clustering algorithms like K-means, FCM, hierarchical techniques are used for gene expression data clustering. But PSO based K-means gives better accuracy than these existing algorithms. In this paper, a Particle Swarm Optimization (PSO)-based K-means clustering algorithm has been proposed for clustering microarray gene expression data.

Keywords– Microarray Gene Expression data, Clustering, K-means, PSO.

I. INTRODUCTION

Microarrays may be used in a wide variety of a fields, including biotechnology, agriculture, food, cosmetics and computers This technology can simultaneously monitor and study the expression levels of thousands of genes, relationship between the genes, their functions and classifying genes or samples. The change of experimental condition, environmental change, drug, disease etc. can change the expression levels. So, gene expression profiling can help to distinguish between disease state versus healthy state, drug identification, effect of change of environmental conditions etc. The DNA microarray is used to measure the expression level of thousands of genes under different condition[3]. Microarray data can be viewed as an N * (M+1) matrix: Each of the columns represents a gene. Each of the rows represents an experimental condition (a sample, a time point, etc.). The gene expression data matrix represents m columns of genes and n rows of samples. The last column is the class label i.e. information about which sample goes to which cluster. One frequent use of this microarray technology is to determine which genes are activated and which genes are repressed when two populations of cells are compared at a given point of time in the life of the organism [10]. Total RNA can be isolated from cells or tissues under different experimental conditions and the relative amounts of transcribed RNA can be measured. A typical microarray experiment contains 102 to 104 genes and the no of samples involved in a microarray experiment is generally less than 100. One of the characteristics of gene expression data is that it is significant to cluster both genes and samples. In gene-based clustering the genes are treated as the objects while the samples are the features. But in sample-based clustering the samples are act as the objects and the genes are treated as the features. The division of gene-based clustering and sample-based clustering is based on different characteristics clustering tasks for gene expression data. In current days, only a small subset of genes take parts in any cellular procedure. In this paper, standard deviation of the genes across all the samples are calculated first, then a small set of genes are taken having high standard deviation as input to the different clustering algorithms. Some work is done on the performance of K-means, PSO and hybrid PSO clustering approaches on different data sets [1][2]. The Euclidean distance measure and International Journal of Computer Science and its Applications 233 cosine correlation measure are used as the distance metrics in these algorithms. The algorithm shows that up to 90 iterations the performance of PSO and hybrid PSO based k-means are quite similar. After 90 iterations the performance of hybrid PSO significantly improves. Particle Swarm Optimization (PSO) applies the concept of social interaction to problem solving. It uses a number of agents (particles) that constitute a swarm moving around in the search space and looking for the best solution. The K-means algorithm is the simplest clustering method and tends to converge faster than the PSO algorithm, but usually can be trapped in a local optimal area. In the general PSO algorithm, PSO can conduct a globalized searching for the optimal clustering, but requires more iteration numbers and computation than the K-means algorithm does. So, in the PSO based K-means algorithm, the ability of globalized searching of the PSO algorithm and the fast convergence of the K-means algorithm are combined and this algorithm is applied on microarray gene expression data clustering. The rest of this paper is organized as follows. Different clustering techniques are described in section II. The proposed algorithm for PSO based K-means clustering is presented in Section III and the results and experiments are reported in section IV, respectively. Section V concludes with a summary and discussion on the future scope of this work.

II. LITERATURE SURVEY

Associated with each object is a set of G measurements which form the feature vector, X =(X1 ,...,XG). The feature vector X

Belongs to a feature space .The task is to identify groups, or clusters ,of similar objects on the basis of a set of feature vectors, X1 = x1 ,..., Xn = xn . Clustering procedures fall into two broad categories. Hierarchical methods ,either divisive or agglomerative.

These methods provide a hierarchy of clusters, from the smallest, where all objects are in one cluster, through to the largest set, where each observation is in its own cluster. Partitioning methods.

1. K means clustering : K-means (MacQueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids shoud be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

   Finally, this algorithm aims at minimizing an *objective function*, in this case a squared error function. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2, \quad (1)$$

   where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the *n* data points from their respective cluster centres.

2. Fuzyy c-means : Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method (developed by Dunn in 1973 and improved by Bezdek in 1981) is frequently used in pattern recognition. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m \left\| x_i - c_j \right\|^2 \qquad 1 \le m < \infty$$

   where *m* is any real number greater than 1, *uij* is the degree of membership of *xi* in the cluster *j*, *xi* is the *i*th of d-dimensional measured data, *cj* is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership *uij* and the cluster centers *cj* by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}}, \qquad c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m}$$

   This iteration will stop when $\max_{ij} \left\{ \left| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right| \right\} < \varepsilon$, where ε is a termination criterion between 0 and 1, whereas *k* are the iteration steps. This procedure converges to a local minimum or a saddle point of *Jm*.

3. Hierarchical clustering: hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types: Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy. In general, the merges and splits are determined in a greedy manner. The results of hierarchical clustering are usually presented in a dendrogram. In the general case, the complexity of agglomerative clustering is $\mathcal{O}(n^3)$, which makes them too slow for large data sets. Divisive clustering with an exhaustive search is $\mathcal{O}(2^n)$, which is even worse. However, for some special cases, optimal efficient agglomerative methods (of complexity $\mathcal{O}(n^2)$) are known: LINK[1] for single-linkage and CLINK[2] for complete-linkage clustering.

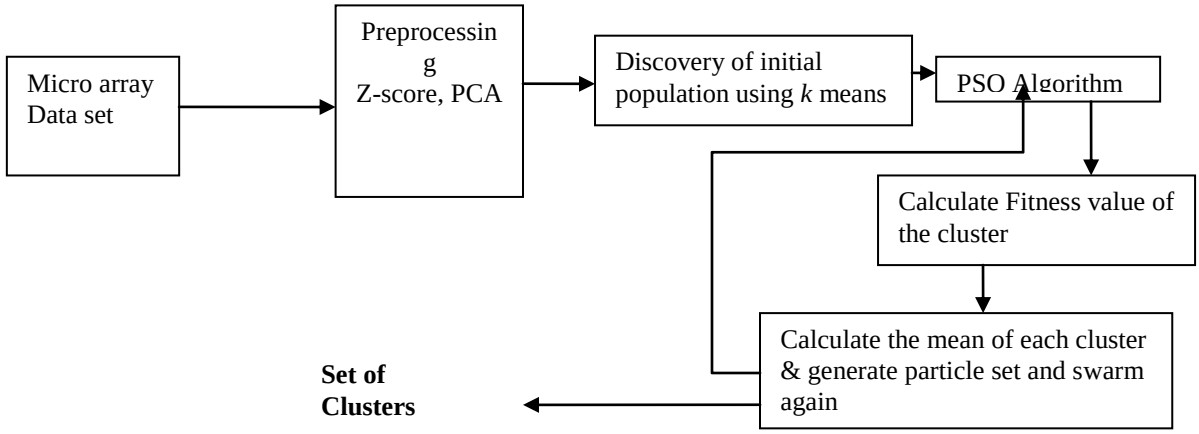
4. PSO

   In computer science, particle swarm optimization (PSO) is a computational method that optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. PSO optimizes a problem by having a population of candidate solutions, here dubbed particles, and moving these particles around in the search-space according to simple mathematical formulae over the particle's position and velocity. Each particle's movement is influenced by its local best known position and is also guided toward the best known positions in the search-space, which are updated as better positions are found by other particles. This is expected to move the swarm

toward the best solutions. PSO is originally attributed to Kennedy, Eberhart and Shi[1][2] and was first intended for simulating social behaviour,[3] as a stylized representation of the movement of organisms in a bird flock or fish school. The algorithm was simplified and it was observed to be performing optimization. The book by Kennedy and Eberhart[4] describes many philosophical aspects of PSO and swarm intelligence. An extensive survey of PSO applications is made by Poli.[5][6]. PSO is a metaheuristic as it makes few or no assumptions about the problem being optimized and can search very large spaces of candidate solutions. A basic variant of the PSO algorithm works by having a population (called a swarm) of candidate solutions (called particles). These particles are moved around in the search-space according to a few simple formulae. The movements of the particles are guided by their own best known position in the search-space as well as the entire swarm's best known position.. A basic PSO algorithm is then:

- For each particle i = 1, ..., S do:
  - Initialize the particle's position with a uniformly distributed random vector: xi ~ U(blo, bup), where blo and bup are the lower and upper boundaries of the search-space.
  - Initialize the particle's best known position to its initial position: pi ← xi
  - If (f(pi) < f(g)) update the swarm's best known position: g ← pi
  - Initialize the particle's velocity: vi ~ U(-|bup-blo|, |bup-blo|)
- Until a termination criterion is met (e.g. number of iterations performed, or adequate fitness reached), repeat:
  - For each particle i = 1, ..., S do:
    - For each dimension d = 1, ..., n do:
      - Pick random numbers: rp, rg ~ U(0,1)
      - Update the particle's velocity: vi,d ← ω vi,d + φp rp (pi,d-xi,d) + φg rg (gd-xi,d)
    - Update the particle's position: xi ← xi + vi
    - If (f(xi) < f(pi)) do:
      - Update the particle's best known position: pi ← xi
      - If (f(pi) < f(g)) update the swarm's best known position: g ← pi
- Now g holds the best found solution.

The parameters ω, φp, and φg are selected by the practitioner and control the behaviour and efficacy of the PSO method, see below.



III. PROPOSED MODEL

This model proposes an improved clustering technique using PSO based K-means clustering which can give better accuracy than other clustering algorithms in microarray data clustering. There are two types of gene expression data clustering process (a) gene-based clustering where genes are clustered taking samples as features i.e. sample size is constant and (b) sample based clustering where samples are clustered taking genes as features.

IV. Z SCORE AND PCA

The standard score is

$$z = \frac{x-\mu}{\sigma} \qquad (5.1)$$

where:

x is a raw score to be standardized.
μ is the mean of the population.
σ is the standard deviation of the population.

The quantity z represents the distance between the raw score and the population mean in units of the standard deviation. z is negative when the raw score is below the mean, positive when above. A key point is that calculating z requires the population mean and the population standard deviation, not the sample mean or sample deviation. It requires knowing the population parameters, not the statistics of a sample drawn from the population of interest. But knowing the true standard deviation of a population is often unrealistic except in cases such as standardized testing, where the entire population is measured. In cases where it is impossible to measure every member of a population, the standard deviation may be estimated using a random sample. For example, a population of people who smoke cigarettes is not fully measured.

V. K MEANS ALGORITHM

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori [20]. The main idea is to define k centroids, one for each cluster. These centroids shoud be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycentre of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more. Finally, this algorithm aims at minimizing an *objective function*, in this case a squared error function. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{k} (\|x_i^{(j)} - c^j\|)^2 \qquad 5.2$$

Where $\|x_i^{(j)} - c_j\|$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the *n* data points from their respective cluster centres.

VI. HYBRID K MEANS WITH PSO

PSO algorithm was showed to successfully converge during the initial stages of a global search, K-Means clustering generates a specific number of disjoint, flat non-hierarchical clusters. It is well suited to generating globular clusters. The K-Means method is numerical, unsupervised, non-deterministic and iterative and quality of clusters is poor, so in this work, a hybrid algorithm combining particle swarm optimization (PSO) algorithm with k-means algorithm is proposed. We refer it as PSO-KMean algorithm. The algorithm aims to group a given set of data into a user specified number of clusters. We evaluate the performance of the proposed algorithm using three datasets. The algorithm performance is compared to K-means and PSO clustering.

VII. SIMULATION RESULTS

The simulation process is carried on the computer having Pentium processor with speed 2.6GHz and 512 MB of RAM. The matlab version used is 7.5.

1. DATASET INFORMATION

For simulation work data sets are downloaded from the 'UCI Machine Learning Repository website. It maintains 187 data sets as a service to the machine learning community. The following data sets are chosen for simulation work .These data sets don't contain any missing data. Description of each data set is given below:-

1.a. Breast Data Set.

This data set is computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. It describes characteristics of the cell nuclei present in the image. The attribute information contains 1) ID number 2) Diagnosis (M = malignant, B = benign). The Ten real-valued features are computed for each cell nucleus are a) radius (mean of distances from center to points on the perimeter) b) texture (standard deviation of gray-scale values) c) perimeter d) area e) smoothness f) compactness (perimeter^2 / area - 1.0) g) concavity (severity of concave portions of the contour) h) concave points (number of concave portions of the contour) i) symmetry j) fractal dimension ("coastline approximation" - 1)

1.b Lung Cancer Data Set

This data set is a multivariate type data set containing integer data and is used mainly for classification purpose .It contains 32 records and number of instance 56 . Attribute Information attribute 1 is the class label. All predictive attributes are nominal, taking on integer values 0-3 Missing Attribute Values: Attributes 5 and 39 (*) Class Distribution 3 classes,

1.c Leukemia Data Set

This classification model is built with probably the most famous gene expression cancer dataset (Golub et al.), containing information on gene-expression in samples from human acute myeloid (AML) and acute lymphoblastic leukemias (ALL) Leukemias are primary disorders of bone marrow. They are malignant neoplasms of hematopoietic stem cells. The total number of genes to be tested is 7129, and number of samples to be tested is 72, which are all acute leukemia patients, either acute lymphoblastic leukemia (ALL) or acute myelogenous leukemia (AML).

VIII. CONFUSION MATRIX

One of the methods to evaluate the performance of a classifier is using confusion matrix. A Confusion matrix that summarizes the number of instances predicted correctly or incorrectly by a classification model. The confusion matrix is more commonly named contingency table which is shown in the following table. For example we have two classes 1 and 2, and therefore a 2x2 confusion matrix, the matrix could be arbitrarily large. The number of correctly classified instances is the sum of diagonals in the matrix; all others are incorrectly classified. Confusion matrix is given for the better set for which better result is achieved. Better set is the set for which better accuracy is found.

Table 6.3 Breast Cancer data set

Confusion Matrix for Breast data set After k means at k=3					**Confusion Matrix for Breast data set After hybrid k means PSO at k=3**				
	Predicated					Predicted			
Actual		Class 1	Class 2	Class 3	**Actual**		Class 1	Class 2	Class 3
	Class-1	7	2	2		Class-1	10	1	0
	Class-2	5	40	6		Class-2	1	49	1
	Class-3	6	10	20		Class-3	1	3	32

Table 6.4 Lung Cancer dataset

Confusion Matrix for Lungcancer data set Using K means at k=4						**Confusion Matrix for Lungcancer data set After hybrid k means PSO at k=4**					
	Predicated						Predicated				
Actual		Class-1	Class-2	Class-3	Class-4	**Actual**		Class-1	Class-2	Class-3	Class-4
	Class-1	120	9	5	5		Class-1	131	3	1	4
	Class-2	8	4	2	3		Class-2	4	8	2	3
	Class-3	4	3	12	2		Class-3	4	3	14	0
	Class-4	7	3	1	9		Class-4	3	0	1	16

Table 6.5 Leukemia data set

Confusion Matrix for Leukemia data set using k means at k=2				**Confusion Matrix for Leukemia data set After hybrid k means PSO at k=2**			
	Predicated				Predicated		
Actual		Class-1	class-2	**Actual**		Class-1	class-2
	Class-1	36	11		Class-1	44	3
	Class-2	8	17		Class-2	2	23

IX. FLU DATA SET

We have also tested our model with unknown samples of flu data from H family named as H6N2, H1N1, H5N1, H7N2, H7N3, H9N1, H3N8, H6N6, H9N2, H1N2, H5N2,H3N6,and H3N8 from Influenza primer design resources by medical college of Wisconsin. Http://www.ipdr.mcw.edu/fludb/sequenceInfo/list from year 1999 to 2010. In recent year this deadliest virus has not only affect the economics of country but also taken millions of life. The problem faced by the doctor is ignorance about the class and of virus. As there were no readymade tools available in market that predicts the class label, hence the drugs designed need to be tested in lab which is time consuming [21]. Which not only threat the life of people but also made financial loss to the state?. The above virus is also known as swine flu and bird flu. Swine flu: Influenza is a virus that infects people, birds, pigs and other animals. Swine flu, swine influenza, is a form of the virus that normally infects pigs. Occasionally, pigs transmit influenza viruses to people, mainly hog farm workers and veterinarians. Due to H1N1 virus related illness were increasing in almost all parts of the world. The World Health Organization declared the infection a global pandemic. Bird Flu: Bird Flu or Avian influenza is a disease whose causative agent is H5N1 virus, which is carried by animals especially Birds.

A. MOMENT MATRIX

A DNA sequence contains information about the A T C and G composition and their position. The composition vector, however, completely disregards the position information. Therefore, we propose a new measure, composition moment vector, which includes information about both composition and position of A T C and G in the sequence. In contrast to the composition vector it also provides functional relation with the structure content, i.e. there must not be two or more DNA sequences that would have different structure content but the same composition moment vector. The moment vector contains the same information as the composition vector plus the A T C and G position information, and intuitively should give better description of the DNA sequence. The Moment matrix M can be calculated using the equation 6.1

$$x_i^{(k)}=1/N(N-1)\ldots(N-K)\sum_{j=1}^{ki} n_{ij}^{k} \quad 6.1$$

where :

N is the length of DNA sequence and n_{ij} the j^{th} position of the i^{th} DNA and K_i is the total number of the i^{th} DNA in the sequence. [28]

Table 6.7 Moment matrix M

SLNO	A	T	C	G
1	777235.8	596916.5	478205.3	585273.8
2	908403.6	601667.1	539373	632483.8
3	866483.2	585808.6	512435.4	710256.8
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
95	122888.8	91696.87	70990.83	88535.53
96	122888.8	91696.87	70990.83	88535.53
97	122888.8	91696.87	70990.83	88535.53
98	769340.4	597374.3	477871.5	593045.3
99	894864.3	610269.6	528533	632073.1
100	843121	587678.5	498396.5	697435.6

B. NORMALIZED MATRIX

The normalized matrix is produced using z-score and PCA algorithm from the moment matrix which is shown on table 6.10.

Table 6.8 normalized matrix M_{Norm}

SLNO	A	T	C	G
1	1.91755	-0.17147	-0.02207	-0.15605
2	2.456676	-0.08597	-0.18607	0.117589
3	2.443502	0.253239	0.06277	0.008894
	…	…	…	…
	…	…	…	…
	…	…	…	…
95	-2.93415	-0.02594	-0.01752	0.026423
96	-2.93415	-0.02594	-0.01752	0.026423
97	-2.93415	-0.02594	-0.01752	0.026423
98	1.922025	-0.13981	-0.00692	-0.18563
99	2.421359	-0.11122	-0.12208	0.067359
100	2.334659	0.211898	0.103759	-0.04038

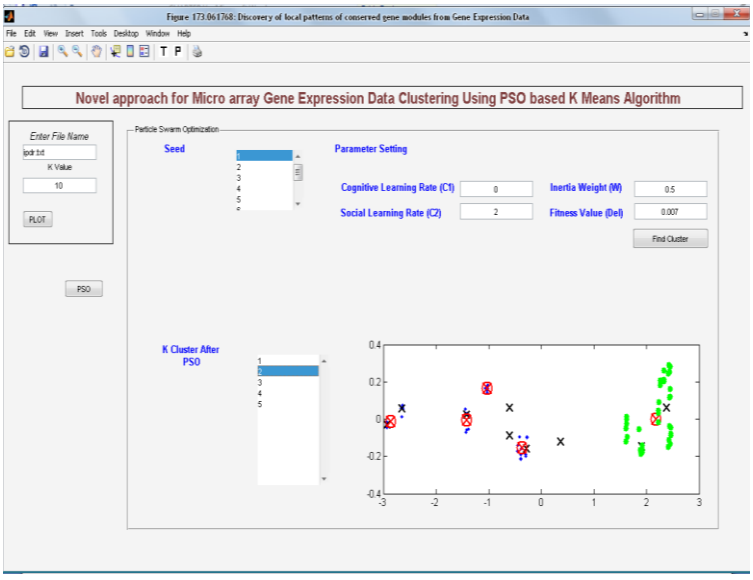


Figure 6.1. K *M*eans clustering of H series virus data set

C. PHYLOGENETIC TREE

A phylogenetic tree or evolutionary tree is a branching diagram or tree showing the inferred evolutionary relationships among various biological species or other entities based upon similarities and differences in their physical and/or genetic characteristics. The taxa joined together in the tree are implied to have descended from a common ancestor. In a rooted phylogenetic tree, each node with descendants represents the inferred most recent common ancestor of the descendants , and the edge lengths in some trees may be interpreted as time estimates. Each node is called a taxonomic unit. Internal nodes are generally called hypothetical taxonomic units as they cannot be directly observed. Trees are useful in fields of biology such as bioinformatics, systematics and comparative phylogenetics.

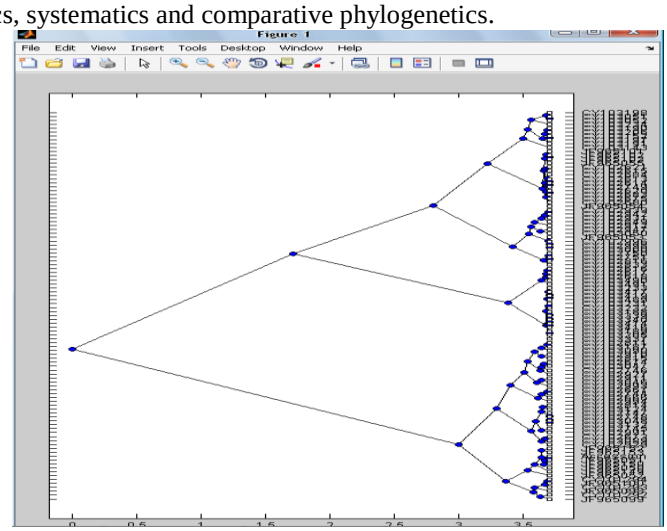


Figure 6.3 Phelogenetic tree for PSO-Kmeans clustering

D. RESULT ANALYSIS OF FLU VIRUS

Table 6.9 H series virus with cluster label

Number of cluster Identified	**Row Index of H series data set of Table 5.1**	**Total Number of Virus**
Cluster 1	2 3 4 5 6 24 27 38 39 48 54 55 62 63 68 69 75 76 80 85 87 89 90 92 93 99 100	27
Cluster 2	15 16 17 20 21 23 26 28 29 31 32 33 34 35 41 42 43 44 50 51 59 60 70 72 94	25
Cluster 3	10 11 12 13 18 25 30 49 56 57 71 77 78 79 91	15
Cluster 4	14 19 22 36 37 40 45 52 58 61 73 81 82 83 84 88 95 96 97	19

X. CONCLUSION

PSO algorithm was showed to successfully converge during the initial stages of a global search which is very faster The k-means algorithm can't determine appropriate clusters depending upon the users and the quality of clusters is very poor .so in this work, a hybrid algorithm combining particle swarm optimization (PSO) algorithm with k-means algorithm is proposed we refer to it as PSO-Kmeans algorithm. The algorithm aims to group a given set of data into a user specified number of clusters. We evaluate the performance of the proposed algorithm using 3 data set. The algorithm performance is compared to K-means and PSO clustering.

In this work a hybrid PSO-K-means algorithm has been proposed which combines the steps of dimensionality reduction through PCA, a novel initialization approach of cluster centers and the steps of assigning data points to appropriate clusters. Using the proposed algorithm a given data set was partitioned in to k clusters in such a way that the sum of the total clustering errors for all clusters was reduced as much as possible while inter distances between clusters are maintained to be as large as possible.

The experimental results show that the proposed algorithm provides better efficiency and accuracy comparison to original k-means algorithm with reduced time. Though the proposed method gave better quality results in all cases, over random initialization methods, still there is a limitation associated with this, i.e. the number of clusters (k) is required to be given as input. Again the method to find the initial centroids may not be reliable for vary large dataset. Evolving some statistical methods to compute the value of k, depending on the data distribution is suggested for future research. Methods for refining the computation of initial centroids are worth investigating.

REFERENCES

[1] Madan Babu,M., Luscombe, N., Aravind, L., Gerstein, M., Teichmann, S.A., Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol*. (2004)

[2] Sara C. Madeira and Arlindo L. Oliveira: Biclustering Algorithms for Biological Data Analysis: Survey***,** *Inesc-Id Technical Report* 1/2004, (2004)

[3] Haixun Wang, Wei Wang , Jiong Yang, Philip S. Yu : Clustering by pattern similarity in large data sets, *International Conference on Management of Data, Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data*, Pages: 394 - 405 , (2007)

[4] Avi Silberschatz , Henry F. Korth and S. Sudarshan, Database System Concepts, Fifth Edition, (2005)

[5] UCI Repository for Machine Learning Data bases retrieved from the *http://www.ics.uci.edu*

[6] Pei, J., Zhang, X., Cho, M., et al.,"'MaPle: A Fast Algorithm for Maximal Pattern based Clustering"', *IEEE International Conference on Data Mining* , (2003)

[7] Lizhuang Zhao and Mohammed J. Zaki, " MicroCluster: Efficient Deterministic Biclustering of Microarray Data", *IEEE Intelligent Systems*, pp. 40-49, (2005)

[8] Jiun-Rung Chen and Ye-ln, " A Condition Enumeration Tree method for mining bilcusters from DNA microarray data sets", *Biosystem*, pp. 44-59.(2009)

[9] Kerr MK, Churchill GA , "Experimental design for gene expression microarrays", *Biostatistics*, (2001) 2:183-201

[10] L.J.P., Postma E.O. and Herik H.J. Van Den , "Dimensionality Reduction: A Comparative Review", *Tech. Rrep*. University of Maastricht,2007

[11] Zou, Trevor Hastiey, Robert Tibshirani,"Sparse Principal Component Analysis", *Journal of Computational and Graphical Statics*, 15(2), 2004, pp. 265-286.

[12] Pavei , Slobodan Ribari and Benjamin Grad , "Comparison of PCA -, MDF -, and RDLDA - based Feature Extraction Approaches for Hand-based Personal Recognition", *International Conference on Computer Systems and Technologies*, (2007).

[13] Wen-Hui Yang, Dao-Qing Dai, Hong Yan, "Finding Correlated Biclusters from Gene Expression Data," *IEEE Transactions on Knowledge and Data Engineering*, (2010).

[14] Kin-On Cheng, Ngai-Fong Law, Wan-Chi Siu, and Alan Wee-Chung Liew, "Identification of coherent patterns in geneexpression data using an efficient biclustering algorithm and parallel coordinate visualization", *BMC Bioinformatics* (2008); 9: 210.

