Microarray Gene Expression Data Clustering using PSO based K-means Algorithm

Lopamudra Dey
Department of Computer Science and Engineering
University of Kalyani
Kalyani-741235, Nadia, West Bengal, India
Email: lopamudra.dey1@gmail.com

Anirban Mukhopadhyay
Department of Computer Science and Engineering
University of Kalyani
Kalyani-741235, Nadia, West Bengal, India
Email: anirban@klyuniv.ac.in

***Abstract*–This paper describes the clustering analysis of microarray gene expression data. Microarray basically consists of large number of gene sequences under multiple conditions. This microarray technology has made it possible to concurrently monitor the expression levels of thousands of genes and across collection of related samples. The most important area of microarray technology is the data clustering analysis. 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*–Clustering, K-means, PSO, Microarray Gene Expression data.**

I. INTRODUCTION

The DNA microarray is a way to measure the expression level of thousands of genes at the same time in a cell mixture [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.) as shown in Figure 1.

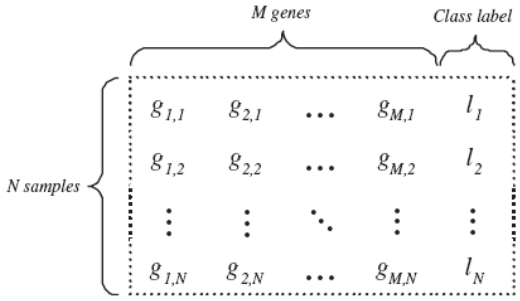

Fig 1. The gene expression data matrix represents m columns of genes and n rows of samples. The last column is the class label i.e. information about which sample goes to which cluster.

The original gene expression matrix obtained from a scanning process contains noise, missing values and systematic variations arising from the experimental procedure such as missing value estimation, data normalization etc. Genes are expressed when they are copied into mRNA or RNA. Gene structure is same in all cells in our body. 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 10^2 to 10^4 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.

Microarray is a tool for analyzing gene expression that consists of a small membrane containing samples of thousands of genes arranged in some regular pattern. 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.

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

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. CLUSTERING FUNDAMENTALS

Clustering involves dividing a set of objects into a specified number of clusters. Clustering algorithms find groups of objects in such a manner that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups. Clustering algorithms rely on a distance metric between data points. Basically it measures the dissimilarity between two data objects.

A. *Distance Metric*

A distance metric is a function that takes two data objects as inputs and has the following properties:

a) Symmetry**:** The distance should be symmetric, i.e.:

$$d(x, y) = d(y, x)$$

b) Positivity**:** The distance between any two points should be a real number greater than or equal to zero:

$$d(x, y) >= 0$$

c) Triangle inequality**:** The distance between two points *x* and *y* should be shorter than or equal to the sum of the distances from x to a third point z and from z to y:

$$d(x, y) <= d(x, z) + d(z, y)$$

B. *Different distance measures*

An important step in most clustering is to select a distance measure, which will determine how the dissimilarity of two elements is calculated. The distance between two *n*-dimensional vectors $x=(x_1,x_2,\ldots,x_n)$ and $y=(y_1,y_2,\ldots,y_n)$ according to different methods, is:

Euclidean distance

The Euclidean distance can be calculated as

$$d_E(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Manhattan distance

$$d_M(\mathbf{x}, \mathbf{y}) = |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_n - y_n| = \sum_{i=1}^{n} |x_i - y_i|$$

where $|x_i-y_i|$ represents the absolute value of the difference between x_i and y_i

C. *Clustering Techniques*

There are two major clustering techniques: Partitioning and Hierarchical. The partitioning clustering method seeks division of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset. In Hierarchical clustering a set of nested clusters organized as a hierarchical tree. The most widely used techniques in analysis of gene expression data which are applied in the early stages and proven to be useful are Hierarchical clustering, K-means clustering and Fuzzy c-means clustering.

Hierarchical Clustering

In hierarchical clustering clusters are generated by grouping genes with similar pattern of expression across a range of samples located near each other. Hierarchical clustering calculates all pairs-wise distance associations between genes and experiments to merge pairs of values that are mainly similar.

There are two types of hierarchical clustering techniques: Agglomerative (which joins clusters in a hierarchical manner) and Divisive (which splits clusters hierarchically).

The major drawbacks of hierarchical clustering are lack of robustness, vague termination criterion [9], and poor accuracy as the size of data sets increases. The time complexity of this approach is also quadratic [10].

K-means Clustering

K-means clustering is a method of cluster analysis which aims to partition n observations into K clusters depending on some similarity/dissimilarity metric where the value of K may or may not be known a priori. The algorithm assigns each point to the cluster whose center (also called centroid) is the nearest. The center is the average of all the points in the cluster, that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster [2]. The main

advantage of this algorithm is its simplicity and speed which allows it to run on large data sets.

The drawbacks k-means method is the lack of prior knowledge of the number of clusters in a dataset which results different results in the altering of results in successive runs since the initial clusters are selected randomly. Another problem of K-means is, it gets stuck at local optima i.e. it minimizes intra-cluster variance but does not ensure that the result has a global minimum of variance.

Fuzzy C-means Clustering

In fuzzy c- means clustering each data point has a degree of belongingness to the clusters, rather than belonging completely to just one cluster. Thus, points on the edge of a cluster may be in the cluster to a lesser degree than points in the center of cluster. With fuzzy *c*-means, the centroid of a cluster is computed as being the mean of all points, weighted by their degree of belongingness to the cluster.

The drawbacks of this approach are, it takes long computational time and it is sensitive to initial guess and noise.

D. Cluster Validity Index

Cluster validity index measures righteousness of a clustering relative to others formed by other clustering algorithms. Cluster validation is very important for clustering analysis because the outcome of clustering needs to be validated in many applications. In most clustering algorithms, the number of clusters is set as user parameter.

Silhouette Validation Method

The Silhouette validation technique calculates the silhouette width for each sample, average silhouette width for each cluster and overall average silhouette width for a total data set. Using this approach each cluster could be represented by so-called silhouette, which is based on the comparison of its tightness and separation.

To construct the silhouettes *S(i)* the subsequent formula is used:

$$S(i) = \frac{(b(i) - a(i))}{\max\{a(i), b(i)\}},$$

where *a(i)* is the average dissimilarity of *i*th object to all other objects in the same cluster; *b(i)* is the minimum of average dissimilarity *of i*th object to all objects in other cluster.

If silhouette value is close to 1, it means that sample is "well-clustered" and it was assigned to a very accurate cluster. If silhouette value is about to zero, it means that that sample could be assign to another bordering cluster as well, and the sample lies equally far away from both clusters. If silhouette value is close to –1, it means that sample is "misclassified" and is merely somewhere in between the clusters. Average silhouette value of all the samples is computed.

III. PSO BASED K-MEANS CLUSTERING ALGORITHM

Particle Swarm Optimization (PSO) [4] is a population-based, robust, stochastic optimization algorithm based on the simulation of the social behavior of birds within a flock. In the past several years, PSO has been proven to be both effective and quick to solve some optimization problems. In data clustering, it is possible to view the clustering problem as an optimization problem that locates the optimal centroids of the clusters rather than to find an optimal partition. This view offers us a chance to apply PSO optimal algorithm on the clustering solution [2]. K-means is the most popular clustering algorithm, but can only generate local optimal solution. PSO clustering algorithm performs a globalized search over entire search space. In the PSO+K-means algorithm, the ability of globalized searching of the PSO algorithm and the fast convergence of the K-means algorithm are combined [1]. The algorithm generates the highest clustering compact result in the experiments.

A. Particle Swarm Optimization (PSO)

At any particular instant, each particle has a position and a velocity. The position vector of a particle with respect to the origin of the search space represents a testing solution of the search problem. At the beginning, a population of particles is initialized with random positions denoted by vectors x_i and random velocities v_i. The population of such particles is called a "swarm" *S*. Each particle is searching for the optimum. Each particle remembers the position it was in where it had its best result so far (its personal best). The particles in the swarm co-operate. They exchange information about what they've discovered in the places they have visited.

In each time step, a particle has to move to a new position. It does this by adjusting its velocity. Velocity is updated based on information obtained in previous steps of the algorithm.

This updating of velocity and position can be described by following set of equations:

$$v_{ij}(t+1) = v_{ij}(t) + C_1 R_1 \left(p_{ij}(t) - x_{ij}(t)\right) + C_2 R_2 \left(p_{gj}(t) - x_{ij}(t)\right)$$

------ ---------- (1)

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1)$$ --------- (2)

Here, $v_{ij}(t+1)$ is the new velocity at time step (t+1)

$v_{ij}(t)$ is the old velocity at time step t

$p_{ij}(t)$ is the best position of each particle

$p_{gj}(t)$ is the Best position of swarm

$x_{ij}(t+1)$:current position of each particle

$x_{ij}(t)$:old position of each particle

R_1 and R_2 are random variables uniformly distributed within [0,1]; and C_1, C_2, are weighting factors, also called the cognitive and social parameter, respectively. In the first version of PSO, a single weight, $C = C_1 = C_2$, called acceleration constant, was used instead of the two distinct weights in equation (1).

B. Proposed Algorithm

In this algorithm the microarray gene expression data is modeled as a problem space. The datasets are very large. So standard deviation is calculated for the columns and they are sorted in descending order of standard deviations. A set of genes having high standard deviation is considered as dataset for the algorithm. The class level information is not considered initially.

Input**:**

Microarray gene expression data containing genes having high standard deviation, samples and class labels..

Output:

A set of clusters.

Algorithm:

Let, the M= $\{w_{ij} | 1<=i<=n, 1<=j<= m\}$ be the microarray gene expression data matrix and the samples of the matrix act as particles in the algorithm.

1. Initially, each particle randomly chooses k different data samples from the data matrix as initial cluster centers.
2. The particles set are encoded to generate the swarm.
3. The velocity is initialized for each particle.
4. For each particle
   (a) Assign each data vector in the data set to the closest centroid vector.
   (b) Compute fitness value based on the following equation

$$J(K) = \sum_{i=1}^{K} \sum_{x_j \in C_i} D^2(v_i, x_j),$$

   where x_j denotes the jth data point, v_i denotes the center of the ith cluster C_i, and $D(v_i; x_j)$ denotes the distance (e.g., Euclidean Distance) of x_j from v_i.

   (c) Reassign the data vector to the centroid vector according to the fitness value.
   (d)Calculate mean of each cluster and generate particle set and swarm again.
5. For each particle
   (a)Decode the Swarm set to get the particle set.
   (b) Compute the fitness value using the following equation

$$J(K) = \sum_{i=1}^{K} \sum_{x_j \in C_i} D^2(v_i, x_j),$$

   (c) The velocity and swarm position is updated using formula (1) and (2) and generate next solutions.
6. Repeat step (5) until one of the following termination conditions is satisfied.
   (a) The maximum number of iterations is exceeded or
   (b) The average change of centroid vector is less than the predefined value.
7. For each pair of datasets (particle) the clustering result is compared with the class label value of the microarray gene expression data.
   (a) If both give the same result i.e. belongs to same class or totally opposite result i.e. experimental result is different from class label value, then the value of counter1 is incremented.
   (b) Otherwise the value of counter2 is incremented.
8. The accuracy of the clustering algorithm is calculated using following formula
   Accuracy= (counter1)/ (counter1+counter2).

IV. EXPERIMENTS AND RESULTS

A. Data Sets

Here two different data sets are considered to compare the performance of K-means, FCM, Hierarchical, and PSO based K-means clustering. The data sets have been taken from www.biolab.si/supp/bi-cancer site. The number of samples varies from 20 to 40 and number of genes varies from 1500 to 4500 in microarray gene expression data.

(a) Breast Cancer (GSE349_350): The breast cancer data set (GSE349_350) includes gene expression measurements of 24 breast cancer samples. The samples were divided into two diagnostic categories based on the patient's response to no adjuvant treatment (sensitive or resistant).

(b) Lymphoma & leukemia **(**GSE1577): The lymphoma leukemia data set (GSE1577) contains gene expression measurements for 9 T-cell lymphoblastic lymphomas (T-LL) and 10 T-cell acute lymphoblastic leukemia's (T-ALL).

B. Results and Discussions

Figure 2 depicts the methodology adapted here for comparing the performance of different algorithms. The accuracy of different clustering method is measured using the formula in step (8) of proposed algorithm. Table 1 and 2 demonstrate the experimental results using K-means, FCM, Hierarchical, and PSO based K-means Clustering Algorithms. 100 iterations are considered in each case. Average values of accuracy and silhouette

index scores have been reported in Table 1 and 2, respectively.

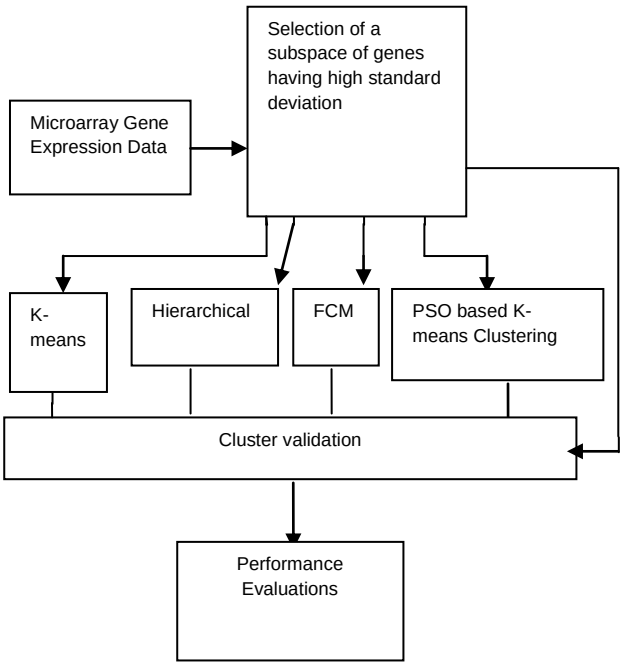

Fig 2: Performance evaluations of clustering algorithms.

Table 1. Accuracy comparison of Hierarchical, K-means, FCM and PSO based K-means clustering.

Dataset	Accuracy			
	Hierar-chical	K-means	FCM	PSO based K-means
Breast Cancer	0.5035	0.7222	0.7222	0.7863
Lympho ma & leukemia	0.5014	0.5679	0.6314	0.7900

Table 2. Silhouette Index comparison of Hierarchical, K-means, FCM and PSO based K-means clustering.

Data Set	Silhouette Index			
	Hierar-chical	K-means	FCM	PSO based K-means
Breast Cancer	0.4292	0.2839	0.2839	0.5139
Lympho ma & leukemia	0.3167	0.4176	0.3846	0.5156

The accuracy and silhouette index scores reported in Table 1 and Table 2 indicate that PSO based K-means clustering outperforms the other algorithms clearly for both the data sets considered here.

V. CONCLUSION AND FUTURE SCOPE

This paper proposes an improved clustering technique using PSO based K-means clustering which gives 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.

In this paper only sample-based clustering method is elaborated as single-objective PSO based K-means algorithm is considered. It is possible to generate multiobjective PSO based K-means clustering algorithm which can cluster both genes and samples simultaneously for gene expression data.

REFERENCES

[1] X. Cui and T. E. Potok, "Document Clustering using Particle Swarm Optimization", IEEE Swarm Intelligence Symposium, Pasadena, California, 2005.

[2] X. Cui and T. E. Potok, "Document Clustering Analysis Based on Hybrid PSO+K-means Algorithm," *Journal of Computer Sciences*, vol. Special Issue, pp. 27-33, 2005..

[3] Y. Wang, F. S. Makedon, J. C. Ford, J. Pearlman, "HykGene : a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data", Vol. 21 No.8 pp. 1530-1537, 2005.

[4] K. E. Parsopoulos, M. N. Vra, "Particle Swarm Optimization and Intelligence: Advances and Applications", Information science reference, Hershey, New York, 2010.

[5] M. R. Sierra and C. A. Coello Coello, "Multi-objective particle swarm optimizers: A survey of the state-of-the-art", International Journal of Computational Intelligence Research, Vol. 2, No. 3, pp. 287-308, 2006.

[6] R. Poli, J. Kennedy, T. Blackwell, **"**Particle Swarm Optimization", Swarm Intelligence, Vol. 1, No. 1, pp. 33-57, 2007.

[7] Y. Shi and R. C. Eberhart, "A Modified Particle Swarm Optimiser", In Proc. IEEE International Conference on Evolutionary Computation, Anchorage, Alaska, May 1998.

[8] J. Kennedy and R. C. Eberhart , "A Discrete Binary Version of the Particle Swarm Algorithm", In Proc. IEEE Conference on Systems, Man, and Cybernetics, pp. 4104–4109, 1997.

[9] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: A survey", IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 11, pp. 1370-1386, 2004.

[10] K. Premalatha, A. M. Natarajan, "A New Approach for Data Clustering Based on PSO with Local Search", Computer and Information Science, Vol. 1, No.4, November 2008.