ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 2, February 2013

Gene Expression Data Analysis Using Automatic Spectral MEQPSO Clustering Algorithm

Vijayalakshmi. S[1], Rajalakshmi@Manochitra. J[2], Jayanavithraa. C[3], Ramya. L[4]

Assistant Professor, Dept of IT, Christ College of Engineering and Technology, Puducherry, India[1].

B. Tech, IT, Christ College of Engineering and Technology, Puducherry, India[2,3,4]

ABSTRACT: **In the field of Genetics, thousands of gene expression levels are measured simultaneously, using Microarray Technology. In this technology, Gene Clustering approach is used to discover the similarity of biological function within the genes. In this approach, many clustering algorithms are used. In this paper a new algorithm - Spectral MEQPSO for clustering gene datasets is proposed, based on MEQPSO and Automatic clustering algorithms. MEQPSO algorithm is a promising method in gene clustering, which provide an ability of stronger global convergence towards an optimal solution. By using Spectral algorithm, cluster number can be selected automatically during the cluster process, which reduces the overall time taken to cluster the genes.**

Keywords : Gene Expression, Clustering, MEQPSO, Spectral algorithm

I. INTRODUCTION

Much research was done to generate a large amount of gene datasets, so, the clustering can be applied in molecular biology for analyzing gene expression data[1]. Using clustering algorithms, different clusters of similar expression patterns of gene dataset are assigned according to a dissimilarity measure between any two genes. The ultimate goal of the clustering process is to identify the genes with the same functions or the same regulatory mechanisms.

In clustering technology, Hierarchical[2] and K-means approaches[3] are used in the earlier process. The fundamental strategy of these clustering approaches is to imitate the evolution process of nature and evolve the solutions of clustering from one generation to the next. Then Genetic K-means algorithm[4] was used in the clustering process, which combine the robust nature of the genetic algorithm and the high performance of the K-means algorithm.

In the year 1995, a population-based random search technique, known as Particle Swarm Optimization (PSO)[5] has been applied to data clustering. It is easier to implement than the earlier approaches, since it do not undergo any complex operations such as selection, crossing and mutation. A new variant of PSO, called Quantum-behaved Particle Swarm Optimization (QPSO)[6 - 8], has been proposed to improve the global search ability of the original PSO. The iterative equation of QPSO is different from that of PSO. It was proved that this iterative equation, leads QPSO to be a global convergent than PSO, since it need no velocity vectors for particles and has only fewer parameters to adjust. The main drawback of this algorithm is, it leads to premature convergence, since the particle is guided by both global best and personal best positions.

To overcome this drawback, a new version of QPSO algorithm was introduced, known as Multi-Elitist Quantum behaved Particle Swarm Optimization (MEQPSO)[9]. In MEQPSO algorithm, the particle's search is influenced by the position, which may lie in a promising search region than that of global position. So the particles have much chance to search this region to find out the global optimal solution. As a result, MEQPSO have better overall performance than the original QPSO. The main disadvantage of this algorithm is, it cannot select the cluster number automatically during the clustering process. So, this algorithm is combined with one of the prominent automatic clustering algorithm called Spectral Clustering algorithm. By combining MEQPSO with Spectral Clustering algorithm, it provides better overall convergence to the best solution by automatically selecting the cluster number during the clustering process.

The rest of this paper is organized as follows. Section II explains about the Spectral clustering and MEQPSO algorithms. Section III provides details on how the clustering process is done automatically using the proposed Spectral MEQPSO Algorithm. Finally, the paper is concluded in Section IV.

II. DESCRIPTION OF ALGORITHMS

A. Spectral Clustering Algorithm

Many algorithms are available for automatically clustering the given datasets. In this paper, one of such automatic clustering algorithm called Spectral Clustering Algorithm is used. This algorithm is used for selecting the cluster number automatically during the clustering process. To do so, the cluster number should be automatically generated or selected according to the number of genes in the given data set. After selecting the cluster number which is to be processed, the genes in the data set then undergo matching and grouping process.

A spectral clustering algorithm is a way of grouping N number of genes (taken to be d-dimensional vectors) into

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 2, February 2013

a predefined number of clusters, *K*. The starting point is to construct an *affinity matrix A* from the data, which is an *N X N* matrix encoding the distances between the various points. The affinity matrix is then normalized to form a matrix L^{-1} by conjugating with the diagonal matrix $D^{-\frac{1}{2}}$ which has the square root values of the sum of the rows in A. This value is taken into account for the spreading of different clusters in the given gene datasets. The algorithm[10] steps are given below:

Step 1 : Given a data set consisting of *N d*-dimensional vectors which is to be partitioned into *K* clusters, and construct the affinity matrix.
Step 2 : Compute the first *K* eigenvectors of *L* (the ones with the largest eigenvalues), assemble them in an *N X K* matrix ^ *Y* and construct an another matrix named *Y* by normalizing the rows of ^Y.
Step 3 : Perform *K*-means on the matrix called *Y* (each row is treated to be a data vector).

B. MEQPSO Algorithm

In the MEQPSO algorithm, by using a parameter β called the growth rate is calculated to find the degree of evolution for each particle. The value of β is increased when the fitness value of the particle of the tth iteration is better than that of (t-1)th iteration of the same particle.
In this algorithm, two best positions are used. They are pbest and gbest. The pbest (personal best) is the value of each particle which track its coordinates within the problem space that are associated with the best solution (fitness) which it has achieved so far. And, best is the global best value of particle, which takes all the populations which are present in the problem space as its topological neighbors.
On each iteration, the best position of every particle is updated. The pbest position which has a better fitness value, than that of gbest position which are obtained before are taken into a candidate area. The updating of gbest position is based on the selection probability Pc. Before updating, the random number is generated. If the random number is greater than Pc and the candidate area is not empty, the gbest position is replaced by pbest position with the highest growth rate β, selected from the candidate area. If not, the gbest position is considered to be the best fitness value of a particle in a present population. The algorithm is terminated, when the limit on the number of iterations is reached. The pseudo code of MEQPSO algorithm[9] is given below:

Input:
Population size M;
Maximum number of iteration MAX_ITER;Selection ProbabilityPc;
Contraction Expansion Coefficient α;
Output: Global best solution gbest;
{Initialize the particles, the pbest, the best positions and the growth rate β of each particle to zero;}

For t = 1 : MAX_ITER
Calculate the mean best position
For i = 1 : m
Update the ith particle;
If (fitness value of the ith particle in the tth iteration > that of the ith particle in the (t-1)th iteration)
β(t) = β(t-1)+1;
End
Update the ith pbest;
If (fitness of the ith pbest > that of the gbest)
Choose the ith pbest ot put into candidate area;
End
End
Calculate β of every candidate, and record the candidate with the maximal value of β (βmax);
If (rand > pc and candidate area is not empty)
Update gbest to become the candidate with β max;
Else
Update the gbest to become the particle having highest fitness value;
End
End

Output : global best solution;

By using above algorithm, the process of matching and grouping is carried out and then the genes in the data set undergo an updation process if needed.

III. AUTOMATIC CLUSTERING USING THE PROPOSED SPECTRAL MEQPSO ALGORITHM

In this paper, the automatic clustering of gene dataset using a Spectral MEQPSO algorithm, is done based on the following activities:

Automatic selection of cluster number, Matching and Grouping of the dataset, and finally updating of gene dataset.

Initially, before the above mentioned activities to be carried out, the input gene data set to undergo the clustering process should be given. It contains all the required description about the genes.

The input for clustering gene dataset is given artificially or provided by experimental results. An example of input for gene dataset[9] is shown below:

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 2, February 2013

TABLE 1
DESCRIPTION OF GENE DATASET

Data set	Number of data points	Data dimensions	Number of clusters
AD_15	110	10	5
GAL	223	80	9
Yeast cell	1663	77	8
Rat CNS	134	9	4

With the input of gene data set, the following operations are performed.

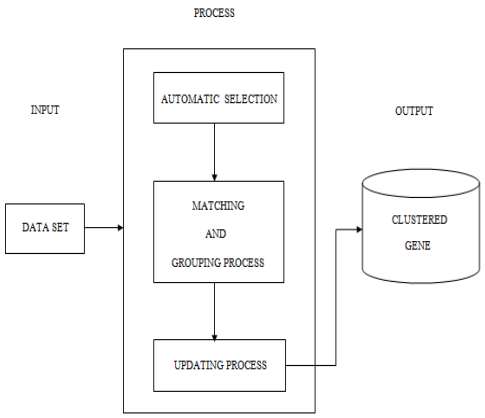

Fig 1. Proposed Architecture

A. Automatic Selection

The cluster number should be selected or generated automatically, according to the number of genes in the given dataset. For this purpose, Spectral algorithm is used. After selecting the cluster number, genes in the dataset undergo matching and grouping operations.

B. Matching and Grouping Process

These processes are performed to find the similarities between genes in the given dataset, in order to produce different clusters or groups of genes which have the same properties. In this process, a random gene is generated from the given dataset. By keeping this gene as reference, the matching process is done, by comparing it with other genes which are present in the dataset. Then, the genes which have a same properties are grouped to form a cluster. These matching and grouping processes are implemented by MEQPSO Algorithm. After processing these operations, the data set should be updated.

C. Updating Gene Dataset

After clustering all the genes from the given dataset, the genes are stored in a separate database. If any new set of genes is to be clustered, they iteratively undergo the matching and grouping process for comparing with the previous clusters of gene, until all the genes are grouped. If all the genes are grouped based on their similarities, they produce the clustered output of genes for a given data set.

The proposed Spectral MEQPSO Algorithm is given below:

Initialize the particles, pbest and gbest positions
Initialize growth rate to be 0
Create a set of data points
Measure of the similarity between points
Eigenvector and values are constructed in the matrix
Treating each row as point, start the iteration process
For every iteration
Calculate the mean best position
For every particle
Update the position value. Then calculate the growth rate for each particle.
Present iteration of the particles is selected
TWCV of particle is calculated.
Then update gbest and pbest positions by considering all the particles.
End
End
Output : optimal solution(gbest)

The output clusters of gene are stored in the database. Thus, the clustering of gene is done automatically by implementing the above proposed algorithm.

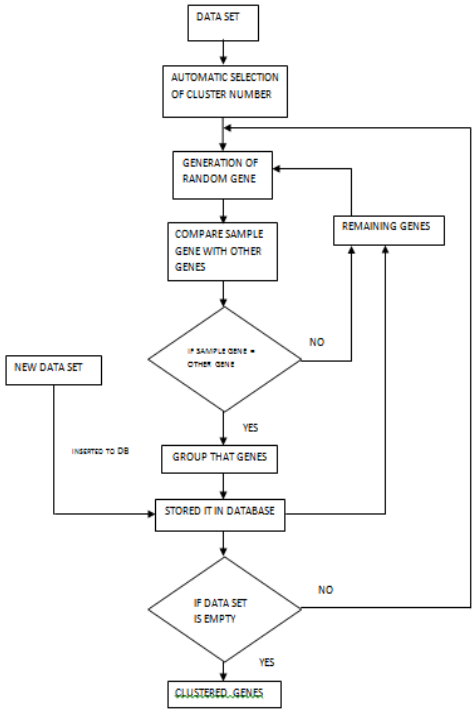

Fig 2. Flow Diagram of Proposed work

IV. CONCLUSION

This paper presented the Spectral MEQPSO algorithm for automatic clustering of the given gene data set. The proposed algorithm is the fusion of Spectral Clustering and MEQPSO Algorithm. Thus, the proposed algorithm provides a better global convergence towards the optimal

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 2, February 2013

solution and it reduces the overall time taken to cluster the genes in a dataset.

Acknowledgment

S. Vijayalakshmi is currently working as Assistant Professor in Information Technology at Christ College of Engineering and Technology, Puducherry, India. She received her MCA from University of Madras in the year 1998. After her graduation, she worked as a Lecturer in Computer Science at Theivanai Ammal Women's College for two years. Subsequently she joined as a Lecturer in Computer Science at Saradha Gangadharan College for five years. She completed her M.Tech., degree in Distributed Computing Systems at Pondicherry Engineering College in the year 2009. Later, she joined as a Lecturer in Information Technology, Christ College of Engineering and Technology, Puducherry, in the year 2009. She has promoted as Assistant Professor in the year 2010. She has published more than nine papers in reputed International Journals and Conferences. Her areas of specialization include Swarm intelligence, Agent Technology and Software Engineering. She is a Life Member of CSI and ISTE.

REFERENCES

[1] Shamir, R., Sharan, R., "Approaches to clustering gene expression data". In: Jiang, T., Smith, T.Y., Xu, Zhang, M.Q. (Eds.), Current Topics in Computational Biology, MIT press, 2001.

[2] Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., "Cluster analysis and display of genome-wide expression patterns. "Proc. Natl Acad. Sci. USA 95 (14), 863–14868, 1998.

[3] Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., Church, G.M., "Systematic determination of genetic network architecture." Nat. Genet. 22, 281–285, 1999.

[4] Krishna, K., Murty, M., "Genetic K-means algorithm". IEEE Trans.Syst. Man Cybern.—B: Cybern. 29, 433–439, 1999.

[5] Kennedy,J., Eberhart, R., "Particle Swarm Optimization". In: Proceedings of the IEEE International Conference On Neural Network, pp. 1942–1948, 1995.

[6] Sun, J., Feng, B., Xu, W.-B.,. "Particle swarm optimization with particles having quantum behavior." In: Proceedings of Congress on Evolutionary Computation, June 2004, pp. 325–331, 2004.

[7] Sun, J., Xu, W.-B., Feng, B., "A global search strategy of quantum-behaved particle swarm optimization". In: Proceedings of IEEE Conference on Cyber- netics and Intelligent Systems, December 2004, pp. 111–116, 2004.

[8] Sun, J., Xu, W.-B., Feng, B., "Adaptive parameter control for quantum-behaved particle swarm optimization on an individual level." In: Proceedings of IEEE International Conference on Systems, Man and Cybernetics, October 2005, pp. 3049–3054, 2005.

[9] Sun, J., Chen, W., Fang, W., Wun, X., Xu, W., "Gene Expression data analysis with the clustering method based on an improved quantum-behaved Particle Swarm Optimization. "SciVerse ScienceDirect, Engineering Applications of Artificial Intelligence 25 (2012) 376–391, 2012.

[10] Andrew Y. Ng., Michael I. Jordan., Yair Weiss," On Spectral Clustering: Analysis and an Algorithm", 2001.

[11] Angeline, P.J., "Using selection to improve particle swarm optimization". In: Proceedings of the 1998 IEEE International Conference on Evolutionary Computation, pp. 84–89, 1998.

[12] Bandyopadhyay, S., Mukhopadhyay, A., Maulik, U., "An improved algorithm for clustering gene expression data". Bioinformatics 23, 2859–2865, 2007.

[13] Ben-Dor, A., Yakini, Z., "Clustering gene expression patterns". In: Proceedings of the Third Annual International Conference on Computational Molecular Biology, DECOMB99. Lyon, France, 1999.