Performance Analysis of Artificial Fish Swarm based Clustering for Gene Expression Data

M. Raja
Research Scholar
Department of Computer Science,
Periyar University,
Salem-636011

H. Hannah Inbarani
Assistant Professor,
Department of Computer Science,
Periyar University,
Salem-636011

M. Thangarasu
Research Scholar,
Department of Computer Science,
Periyar University,
Salem-636011

ABSTRACT

The K-Means algorithm is the widely used clustering technique. The performance ofthe K-Means algorithm depends highly on original cluster centers and converges to local minima. This paper proposes hybrid Artificial Fish Swarm Means (AFSK-Means) based clustering algorithm, by combining Particle Swarm Optimization with K-Means (PSOK) and Artificial Fish Swarm Algorithm based K-Means (AFSA). The basic idea is to search around the global solution by AFSK-Means and to increase the information exchange among genes. The effectiveness of the clustering algorithm depends on finding optimal clusters. The Clustering result shows the improved performance of hybrid clustering algorithm AFSK-Means in finding the best solution compared with the algorithms K-Means and PSOK-Means.

Keywords

Hybrid evolutionary optimization algorithm, Data clustering, K-meansclustering, Artificial Fish Swarm Algorithm, Particle swarm optimization

1. INTRODUCTION

Data clustering is unsupervised classification of patterns into groups or clusters. In clustering, the data in each cluster assign a high degree of likeness while being very dissimilar to data from other clusters. Differences are assessed according to the attributed values describing the objects. Generally, distance measures are used. Data clustering has been used in many different applications, such as data mining, machine learning, biology, and statistics (Tsai *et al*., 2004; Kao *et al.*, 2008). Modern clustering algorithms can be partitioned into two main categories: hierarchical and partition clustering. In hierarchical clustering the data are not unglued into a particular cluster in a single step. Hierarchical clustering does not need to require the number of clusters, most of which are deterministic. On the other hand, partition clustering attempts to straight partition the dataset into a set of divide clusters. The partition clustering begins with a casually chosen or user-defined clustering, and then improves the clustering according to some validity measurements (Fathian*et al*., 2007).In this paper, application on partition clustering, and in particular a public partition clustering method called K-Means clustering. Among clustering algorithms, the K-means clustering method is one of the most generally used and applied methods. The main idea of the K-Means clustering is to identify K centroids, one for both cluster. All samples in the dataset are matched with each center by means of the Euclidean distance and assigned to the nearby cluster center. The method is iterated until no sample is pending. In each stage, the center of each cluster is recalculated by using the regular vector of the items assigned to the cluster. The algorithm stops when the changes in the cluster centers from one step to the next are close to zero or smaller than a pre-specified value. Every sample is assigned to only one cluster (Mingoti and Lima, 2006). Unfortunately, the effects of the K-means are very delicate to the initial values of centers. A poor choice of centers may lead to a local optimum, which is quite inferior to theglobal optimum (Laszlo and Mukherjee, 2007). Recently, evolutionary algorithms such as the genetic algorithm (GA), Tabu search (TS), and simulated annealing (SA) have been developed to solve the clustering problem. However, most evolutionary methods such as GAs and TS are usually very slow at finding an optimal solution. Recently researchers have offered new evolutionary methods such as particle crowd algorithms to solve hard optimization problems, which not only have a better reaction but also congregate very quickly in comparison with normal evolutionary methods (Eberhart and Shi, 2001). All studies verify that the particle swarm optimization (PSO) should be taken into account as a controlling technique, which is efficient enough to handle various kinds of nonlinear optimization difficulties. Nevertheless, it may be trapped into local optima if the global best and local best positions are equal to the position of particle over a number of iterations (Niknam, 2006; Olamaei et al., 2008). To overcome this shortcoming this paper presents a dissimilar hybrid evolutionary optimization method based on PSO and AFSK-Means, called PSOK-AFSK-Means, for optimally clustering N objects into K clusters, which not only has a better response but also congregates more quickly than ordinary evolutionary algorithms.The basic idea is to search around the global solution by AFSK-Means and to increase the information exchange among particles using a mutation operator to escape local optima.

1.1 Gene Clustering

Gene expression data is usually represented by a matrix, with rows matching togenes, and columns corresponding to conditions, experiments or time points. The content of the matrix is the expression levels of each gene under each condition. Those levels may be absolute, relative or otherwise normalized. Each column contains the results obtained from a single array in a particular condition, and is called the *profile* of that condition. Each row vector is the *face pattern* of a particular gene across all the conditions. More formal definitions will begiven in the sequel.A key initial step in the analysis of gene expression data is the detection of groups of genes that exhibit similar expression patterns. This translates to the algorithmic problem of clustering. A clustering problem consists of elements and a characteristic vector for eachelement. A measure of similarity is defined between pairs of such vectors. In gene expression, elements are usually genes andthe vector of each gene is its expression pattern; similarity can be measured in various ways that are problem dependent, for example, by the correlation coefficient between vectors. The goal is to partition the elements into subsets, which are called *clusters*, so that two criteria are satisfied: *Homogeneity* -elements in the same cluster are highly similar to each other; and *separation* - elements from different clusters have low similarity to each other.

1.2 Gene Expression Pre-processing Techniques

The normal preprocessing steps include scale transformations, management of missing values, replicate handling, flat pattern filtering and pattern standardization and they are required before performing any pattern analysis. The processed data set can be sent to other patternanalysis tools.

1.3 Applications of Clustering Gene Expression Data

Clustering technique have proven to be helpful to understand gene function, gene regulation, cellular processes, and subtypes of cells. Genes with similar expression patterns (co-expressed genes) can be clustered together with similar cellular functions. This approach may further understand of the functions of many genes for which information has not been previously available [19, 20]. Furthermore, co-expressed genes in the same cluster are likely to be involved in the same cellular processes, and a strong correlation of expression patterns between those genes indicates co-regulation. Searching for common DNA sequences at the promoter regions of genes within the same cluster allows regulatory motifs specific to each gene cluster o be identified and cist-regulatory elements to be proposed [9, 19]. The inference of regulation through the clustering of gene expression data also gives rise to hypotheses regarding the mechanism of the transcriptional regulatory network [17]. Finally, clustering different samples on the basis of corresponding expression profiles may reveal sub-cell types which are hard to identify by traditional morphology-based approaches [22, 23].In the following, related work is discussed in Section 2. In Sections 3, describes the proposed work based on clustering PSOK and AFSK-Means algorithms. An experimental analysis is illustrated in Section 4. Conclusion and further implementation are discussed in Section 5.

2. RELATED WORK

2.1 K-Means Algorithm

K-means algorithm is based on identifying a preliminary amount of groups, and iteratively changing objects between clusters to the union. To set an integer K and allocate n data points.

$D_{is} \subset R_d$ To aspiration is to select a group of K centre's, consequently to minimize the potential function [1, 9, 11, 13].

$$\varphi = \sum_{x \in D_{is}} min \, ||X - C||2$$

This algorithm allocates every position to the cluster whose center is nearby. The center organizes the arithmetic mean calculation for every element one by one over all the points in the cluster [14, 21]. Suppose itr is the convergent limit, The pseudo code of K-means algorithm is presented inthe algorithm

S – Dataset

K – Number of clusters
The K-means algorithm

```
Input: A set S of examples (vectors of geneexpression
levels), a number K of clusters

Initialization: assign the examples randomly to
The K clusters
Loop:
    1.  Compute the mean for each cluster
    2.  Assign each example to the "nearest cluster"
    3.  If stop condition reached then exit loop, else
        Repeat loop
Output: A set of K clusters
```

2.2 Design of PSOK-Means

Parallel and distributed approaches are natural in swarm intelligence and they have been used intensively since the early years of this research field. Swarm systems, in fact, have often been described as intrinsically parallel computational methods. The reason for this is that many of the main computational tasks characterizing this family of heuristics are independent of each other; thus it is straightforward to perform them at the same time. This is the case, for instance, of the evaluation of the fitness of the particles in a swarm. Furthermore, by attributing a non-apodictic structure to the population, something that also finds its inspiration in nature, the operations that allow particles to update their position can also be performed independently of each other. These approaches can be useful even when there is no actual parallel or distributed implementation, thanks to the particular information Diffusion given by the more local structures of the swarms. But of course parallel and distributed approaches are at their best when the structures of the models are reflected in the actual algorithm implementations. In fact, when compared with other heuristics, swarm systems are relatively costly and slow. Both parallel and distributed implementations can boost performance and thereby allow practitioners to solve exactly or approximately, larger and more interesting problem instances thanks to the time savings afford. To future work will be oriented to the study of the computational speed and scalability of these algorithms on truly parallel architectures, like clusters of CPUs.This is the basic PSOK algorithm as introduced for instance in [13, 16, 17], where each particle is attracted by one global best position for all the swarm and one local best position. The basic PSO velocity and position-update equations for a particle are given as follows:

$$\mathbf{v(t) = w * v(t-1) + c1 * rand()} \\ \mathbf{* \left[X_{pbest}(t-1) - X(t-1)\right] + +c2} \\ \mathbf{* rand()} \\ \mathbf{* \left[X_{gbest}(t-1) - X(t-1)\right]} \quad \mathbf{(1)}$$

PSO-based K-Means (PSOK) clustering algorithm [11, 13], recognized as PSOK i.e. K-Means clustering incorporated with Parallel PSOK. An enhanced performance can be the local finest resolution found so far by the i^{th} particle, while Pg stand for the positional coordinates of the well particle found so far in the whole cluster. Once the iterations are completed, most of the particles are projected to converge to a small radius nearby the global optima of the search space. In PSO, an inhabitant of conceptual particle is initialized among random positions Xi and velocities Vi, and function, f, is calculated, using the particle's positional coordinates as input values. In n dimensional search space, Xi= (xi1, xi2, xi3, xin) and Vi= (vi1, vi2, vi3 ... vin) positions and velocities are adjusted, and the function is estimated with the new coordinates at every time step. The essential update equations for the dth dimension of the ith particle in PSO may be specified as

$$\mathbf{X_{id}^{new}} = \mathbf{X_{id}^{old}} + \mathbf{V_{id}^{new}} \qquad (2)$$

Algorithm PSOK

Step 1.At the first stage, each particle randomly chooses K different

D vectors from the Dataset as the initial cluster centroids vectors

Step 2. For each particle:
(a)Assign each vector in the Data set to the closest centroids vector.

(b) Calculate the fitness value

$$f = \frac{\sum_{i=1}^{N_c}\{\sum_{j=1}^{p^i} d(O_i, M_{ij})\}}{N_c}$$

(c) Using the velocity and particle position update equations (1) and (2) and generate thenext solutions.

Step 3.Repeat step (2) until one of the following termination conditions is satisfied.

(a)The maximum number of iterations is exceeded or
(b) The average change in centroids vectors between iterations is less than a predefined valu

$$\mathbf{V_{id}^{new}} = \mathbf{V_{id}^{old}} + \mathbf{c_1, r_1, p_{id}, x_{id}} + \mathbf{c_2, r_2, g_{id}, x_{id}}$$

The variables r1 and r2 are random positive numbers, drawn from a uniform distribution and classified by an upper limit Rmax; which is a parameter of the system. In (2), c1 and c2 are called acceleration constants whereas w is called inertia weight. Pb is the local finest solution found so far by the ith particle, while Pg correspond to the positional coordinates of the fittest particle found so far in the entire community

3. AFSA BASED CLUSTERING

3.1 Basic AFSA Design and Analysis

Fish usually stay in the place with a lot of food, so to simulate the behaviors of fish based on this characteristic to find the global optimum, which is the basic idea of the AFSO. The basic behaviors of AF are defined by equations (8),(9) as follows for maximum:

AF_Prey: This is a basic biological performance that tends to the food; generally the fish perceives the concentration of food in water to determine the movement by vision or sense and then chooses the tendency. Behavior description: Let $X_i$be the AF current state and select a state $X_j$randomly in its visual distance, *Y* is the food concentration (objective function value), the greater *Visual* is, the more easily the AF finds the global extreme value and converges.

$$X_j = X_i + \text{Visual.rand} \qquad (3)$$

If Y_i<$Y_j$in the maximum problem, it goes forward a step in this direction;

$$X_i^{(t+1)} = X_i^{(t)} + \frac{X_J - X_i^{(t)}}{\left\|X_J - X_i^{(t)}\right\|}.step.rand().(4)$$

Otherwise, select a state X$_j$ randomly again and judge whether it satisfies the forward condition. If it cannot satisfy after ***try*** number of times, it moves a step randomly. When the try number is small in AF_Prey, the AF can swim randomly, which makes it flee from the local extreme value field.

AFSK-Means
Input: D dataset, K -number of clusters,
Output: K overlapping clusters of dataset
Step 1. At the first stage, each particle randomly chooses K different d vectors from the Dataset as the initial cluster centroids vectors.
Step 2. For each Fish
(a) Calculate the initial prey () value for given dataset.
(b) Assign each vector in the Data set to the closest centroid vector.
(c) Calculate the fitness value

$$X_i^{(t+1)} = X_i^{(t)} + \frac{X_j - X_i^{(t)}}{\left\|X_j - X_i^{(t)}\right\|}.step.rand().$$

(d) Using the velocity and particle position update equations (4) and (5) and generate the next solutions.
Step 3. Repeat step (2) until one of the following termination conditions is satisfied
(a)The maximum number of iterations is exceeded or if the cluster converges

$$X_i^{(t+1)} = X_i^{(t)} + Visual.rand()(5)$$

AF_Swarm: The fish will assemble in collections naturally in the moving process, which is a kind of living habits to assurance the existence of the colony and avoid dangers. Behavior description: Let X$_i$ be the AF current state, X$_c$ be the center position and n f be the number of its companions in the present neighborhood (d $_{ij}$< Visual), n is total fish number. If Y$_c$> Y$_i$ and n f n < δ, which means that the companion center has more food (higher fitness

function value) and is not very crowded, it goes forward a steep to the companion center;

$$X_i^{(t+1)} = X_i^{(t)} + \frac{X_c - X_i^{(t)}}{\left\|X_c - X_i^{(t)}\right\|}.step.rand(). \qquad (6)$$

Otherwise, executes the preying performance. The crowd factor limits the scale of swarms, and more AF only cluster at the best area, which ensures that AF will move to optimum in a wide field.

AF_Follow: In the moving process of the fish swarm, when a single fish or several ones find food, the neighborhood partners will trail and reach the food quickly. Behavior description: Let *Xi* be the AF current state, and it explores the companion $X_j$in the neighborhood (d_{ij}<Visual), which has the greatest $Y_{j.}$ If Y_j >$Y_i$and $n\ f\ _n$< δ, which means that the companion $X_j$state has higher food concentration (higher fitness function value) and the surrounding is not very crowded, it goes forward a step to the companion *X j*,

$$X_i^{(t+1)} = X_i^{(t)} + \frac{X_j - X_i^{(t)}}{\left\|X_j - X_i^{(t)}\right\|}.step.rand() \qquad (7)$$

Otherwise, executes the preying behavior.
AF_Move: Fish swim randomly in water in point, they are seeking food or companions in larger ranges. Behavior description: Chooses a state at unplanned in the vision, then it moves towards this state, in fact, it is a default behavior of AF_Prey.

$$X_i^{(t+1)} = X_i^{(t)} + Visual.rand(\ ) \qquad (8)$$

AF_Leap: Fish stop somewhere in water, every AF's behavior result will gradually be the same, the difference of objective values (food concentration, FC) become smaller within some iterations, it might fall into local extremism change the parameters randomly to the still states for leaping out present state. Behavior description: If the objective function is almost the same or difference of the objective functions are smaller than a section during the given (m−n) iterations, Chooses some fish arbitrarily in the whole fish swarm, and set parameters randomly to the selected AF. The β is a parameter or a function that can makes some fish have other abnormal actions (values), esp. is a smaller constant.

$$if(BestFC(m) - BestFC(n) < eps$$

$$X_{some}^{(t+1)} = X_{some}^{(t)} + \beta.Visual.rand(\,) \qquad (9)$$

AF_Swarm makes few fish confined in local extreme values move in the direction of a few fishes tending to global extreme value, which results in AF fleeing from the local extreme values. AF_Follow accelerates AF moving to better states, and at the same time, accelerates AF moving to the global extreme value field from the local extreme values.

4. EXPERIMENTAL ANALYSIS

4.1 Dataset

Test dataset is considered for evaluating the proposed algorithm. The test datasets are collected from UCI Repository Yeast dataset.

Yeast Dataset consists of 1484 objects characterized by 9 features:

CYT (cytosolic or cytoskeleton), NUC (nuclear), MIT (mitochondrial), ME3 (membrane protein, no N-terminal signal), ME2 (membrane protein, uncleaved signal), ME1 (membrane protein, cleaved signal), EXC (extracellular), VAC (vacuolar), POX (peroxisomal), ERL (endoplasmic reticulum lumen)

Ecoli Dataset consists of 336 objects Characterized by 8 features.

Seed Dataset Consist of 210 objects characterized 7 features.

Yeast gene expression dataset were chosen to evaluate the proposed method. The proposed method AFSK-Means based clustering was compared against different existing methods, including classic K-Means method, original PSOK. In order to have better representative number of clusters, these methods were applied to cluster the datasets into two different number of clusters, K = 200 and K = 150 clusters, respectively. A total of 10 trials for the schemes with the datasets were conducted. The experimental settings of PSOK and AFSK-Means schemes are specified in Table 1. Initial values of velocity and its maximum velocity ofPSOK and AFSK-Means are set to be the dynamic range of each dimension accordingly. Theeffects of parameters in Table 1 are discussed as follows. The w is set to 0.7 which is a typical value in between 0 and 1. The values of c1 and c2 are suggested as 2 for the general PSOK, but it requires smaller values for clustering problem and it can be referred to AFSK-Means [7]. The two values were obtained by experiment. The MSE results of the methods with K = 200 and K = 150 are listed in Table 2. It can be seen that the results of the classic K-Means clustering is the weakest for the test sets (the lower MSE value, the better result). The PSOK can have notable improvement in comparison with the classic K-Means method. Recently, AFSK-Means for gene clustering has been demonstrated to further improve the results. Among the methods, the proposed AFSA K-Means exhibits the optimized results. For evaluating convergence rate among the methods, the variations of MSE performance over time (in second) of the schemes with the test datasets are shown in Figs. 1 and 2. The experimental results with the set of K = 200 are selected to show the convergence rate. From the results depicted in the figures, it can be seen that the MSE of K-Means algorithm converges faster among all the methods due to its simplicity and efficiency, but it tends to be trapped in premature solution.

Table 1. Key experimental settings for PSOK, AFSK-Means schemes:

Input parameter	Values
Swarm Size	6
C1	0.6
C2	0.2
No.Of maximum Iteration	50

Table 2 MSE clustering results of the schemes with K = 200

S.No	**Schema**	**MSE**
1	K-Means	4.5
2	PSOK	4.0
3	AFSK-Means	3.3

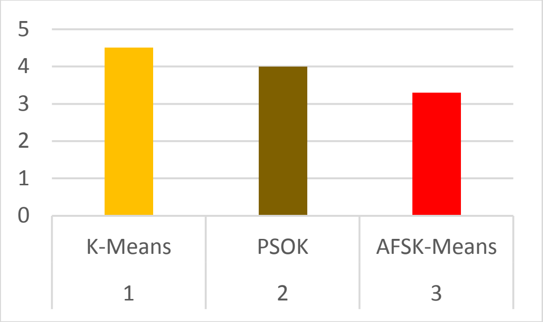

Fig. 1 MSE variance when K=200

Table 3 .MSE clustering results of the schemes with K = 150

S.NO	**Scheme**	**MSE**
1	K-Means	3.1
2	PSOK	2.9
3	AFSK-Means	2.4

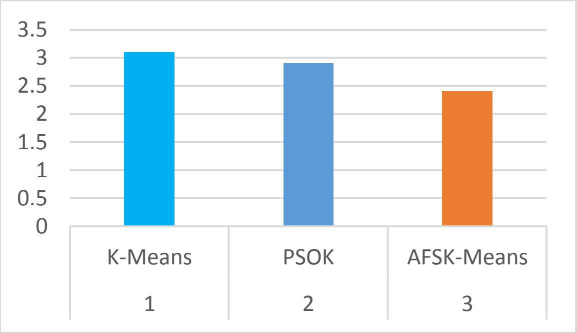

Fig. 2 MSE variance when K=150

4.2 Comparative Analysis Based On DBI

The comparative results based on DBI and DUNN validity measure for all the depicted gene expression data clustering algorithm is shown table 5.1.

Table 3 Comparative analysis Based on DBI

Datasets	K-Means	PSOK-Means	AFSK-Means
Ecoli dataset	5.2416	8.6217	4.5945
Seed dataset	9.0868	11.8762	4.6107
Yeast Dataset	8.1045	13.3784	7.8162

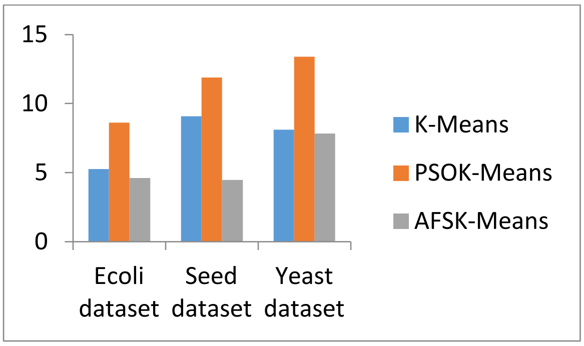

Fig 3 Comparative analysis Based on DBI

Table 4 Comparative analysis Based on DUNN

Datasets	K-Means	PSOK-Means	AFSK-Means
Ecoli dataset	5.5253	6.4512	2.9017
Seed dataset	5.6129	9.4136	1.9742
Yeast Dataset	3.6232	11.8783	1.9387

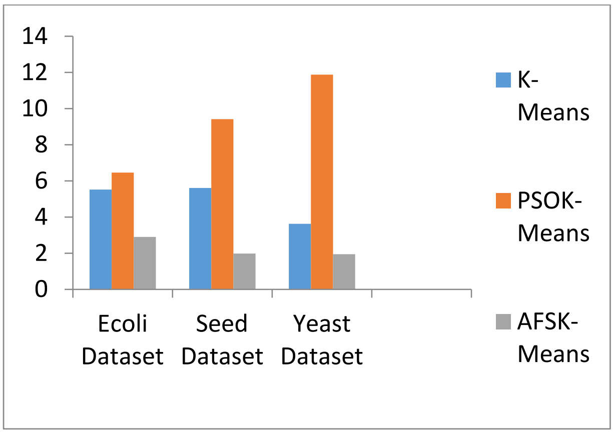

Fig 4 Comparative analysis Based on DBI

However, the three kinds of M methods can further improve the quality of the MSE results. Although they can converge at similar time, the proposed AFSK-Means can have capability of searching a better solution notably earlier. It is noted that PSOK –Means algorithm converges slowly due to three K-Means iterations conducted in its every iteration. Finally, the proposed AFSAK can provide the best solution. More details can further be observed from the results. In Figs. 1 and 2, they show that PSOK and the proposed AFSK-Means can gain greater improvement for hose high-dimensional datasets (yeast cell-cycle). Among different K values, K-Means is most dependent on the selection of initial conditions. This problem has been overcome with the integration of PSO and K-Means algorithm, a technique which is referred to as PSO-based K-Means clustering algorithm (PSOK). To proposed AFSK-Means further improves the performance in this aspect by optimizing the sequence of the clusters encoded in the particle positions.

5. CONCLUSION

To proposed a cross evolutionary algorithm based clustering algorithm, called AFSK-Means. The AFSK-Means algorithm was used to explore for the cluster centers. This algorithm minimizes the objective function of the clustering problem. When the number of clusters is known a priori, the AFSK-Means algorithm can find the cluster centers. In order to demonstrate the effectiveness of the AFSK-Means clustering algorithm in finding optimal clusters from the gene expression dataset, the dataset consist of the numbers of size ranging from 4 to 9 and the numbers of clusters ranging from 2 to 6. The simulation result of the proposed algorithm was compared with those of the *K*-means, original PSOK. The results reveal that the AFSK-Means clustering algorithm provides a performance that is significantly better than that of the *K*-means algorithm.

6. REFERENCES

[1] Brown P, Botstein D (1999) Exploring the new world of the genome with DNA microarrays. Nat Genet 21:33–37.

[2] Brazma A, Robinson A, Cameron G, Ashburner M (2000) Onestop shop for microarray data. Nature 403:699–700

[3] Asyali MH et al (2006) Gene expression profile classification: a review. Curr Bioinform 1:55–73

[4] Dopazo J (2006) Functional interpretation of microarray experiments. OMICS 10:3

[5] Kerr G, Ruskin HJ, Crane M, Doolan P (2008) Techniques for clustering gene expression data. Comput Biol Med 38:283–293

[6] Hartigan JA, Wong MA (1979) A K-Means clustering algorithm. Appl Stat 28:126–130

[7] Du Z et al (2008) PK-Means: a new algorithm for gene clustering. Comput Biol Chem 32(4):243–247

[8] Sun J et al (2012) Gene expression data analysis with the clustering method based on an improved quantum behaved particle swarm optimization. Eng Appl Artif I

[9] Shi Y, Eberhart RC (1998) A modified particle swarm optimizer. In: Proceedings of the IEEE international conference on evolutionary computation, IEEE Press, Piscataway, NJ, pp 69–73

[10] Lam YK, Tsang PWM, Leung CS (2011) Improved gene clustering based on particle swarm optimization, K-Means, and cluster matching. In: ICONIP 2011, part , LNCS, Springer, Heidelberg, vol. 7062, pp 654–661

[11] Alizadeh AA et al (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403:503–511

[12] Spellman PT et al (1998) Comprehensive identification of cell cycle-regulated genes of the yeast. Saccharomyces

cerevisiae by microarray hybridization. Mol Biol Cell 9:3273– 3297.

[13] Chu S et al (1998) The transcriptional program of sporulation in budding yeast. Science 282:699–705.

[14] Troyanskaya O et al (2001) Missing value estimation methods for DNA microarrays. Bioinformatics 17:520–525

[15] Lockhart, D. et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat. Biotechnol, 14:1675–1680, 1996.

[16] Schena,M., D. Shalon, R. Davis, and P.Brown. Quantitative monitoring of gene expression patterns with a compolementatry DNA microarray. Science, 270:467–470, 1995.

[17] Tefferi, A., Bolander, E., Ansell, M., Webern, D. and Spelsberg C. Primer on Medical Genomics Part III: Microarray Experiments and Data Analysis. Mayo Clin Proc., 77:927– 940, 2002.

[18] D'haeseleer, P., Wen, X., Fuhrman, S., Somogyi, R. Mining the Gene Expression Matrix: Inferring Gene Relationships fromLarge Scale Gene Expression Data. Information Processing in Cells and Tissues, pages 203–212, 1998.

[19] Tavazoie, S., Hughes, D., Campbell, M.J., Cho, R.J. and Church, G.M. Systematic determination of genetic network architecture. Nature Genet, pages.Essen, Michael B., Spellman, Paul T., Brown, Patrick O. and Botstein, David . Cluster analysis and display of genome-wide expression patterns. Proc. Natl. [21] Brahma, Alvis and Vilo, Jaak. Minireview: Gene expression data analysis. Federation of European Biochemical societies, 480:17–24, June 2000

[20] Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J. Jr, Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Staudt, L.M. et al. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. Nature, Vol.403:503–511, February 2000.

[21] Golub T.R., Slonim D.K., Tamayo P., Huard C., Gassenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield D.D., and Lander E.S. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, Vol. 286(15):531–537, October 1999.

[22] M Thangarasu, H Hannah Inbarani, "Analysis of K-Means with Multi View Point Similarity and Cosine Similarity Measures for Clustering the Document", International Journal of Applied Engineering Research, Vol. 10, pp. 6672-6675, 2015.

[23] M Thangarasu, R Manavalan, "Design and Development of Stemmer for Tamil Language: Cluster Analysis", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, pp. 813-818

[24] M Thangarasu, R Manavalan, "Stemmers for Tamil Language: Performance Analysis", International Journal of Computer Science & Engineering Technology, Vol. 4, pp. 902-908

[25] M Thangarasu, R Manavalan, "Tree-Based Mining with Sentiment Analysis for Discovering Patterns of Human Interaction in Meetings Tamil Document", International Journal of Computational Intelligence and Informatics, Vol.3, pp. 151-159