# Performance Analysis of Firefly K-Means Clustering Algorithm Using Gene Expression Data

**M. Lakshmi**
*Department of Computer Science*
*Periyar University*
*Salem-636011, Tamilnadu*
*mlakshmics09@gmail.com*

**K. Thangavel**
*Department of Computer Science*
*Periyar University*
*Salem-636011, Tamilnadu*
*drktvelu@yahoo.com*

**P. S. Raja**
*Department of Computer Science*
*Periyar University*
*Salem-636011, Tamilnadu*
*psraja5@gmail.com*

***Abstract*- Clustering is a common technique for grouping similar objects and is used in many fields, including data mining, pattern recognition, and image analysis. K-means clustering is a bench mark algorithm for data clustering, but this method has some limitations such as the random selection of initial centroid and it results in local optimal. The initial centroid is playing vital role in K-Means clustering process. This paper adopts firefly algorithm to initialize the cluster centroids rather than selecting randomly. Firefly algorithm is an optimization algorithm which is used to find the initial centroids and then the K-Means clustering allows to refine the centroid and cluster. Through this Firefly optimization algorithm, optimal solution is optioned to find the initial cluster centroid for K-Means clustering algorithm and it applied for Asthma gene expression data. The performance of the Firefly K-Means algorithms is evaluated using the Asthma Gene expression data and compared with K-Means algorithm. The proposed Firefly K-Means algorithms out performs.**

Keywords- Gene Expression, Asthma Microarray Dataset, K-Means clustering, Firefly Optimization Algorithm, Firefly K-Means Algorithm (FIFKA)

## I. INTRODUCTION

Clustering is one of the most important unsupervised classification techniques, where a set of patterns, usually vectors in a multidimensional space, are grouped into clusters (or groups) based on some similarity metric [1]. It is often used for a difference of applications in statistical data analysis, image analysis, data mining and other fields of science and engineering. This can also be used as partitional clustering. Partitional clustering challenges to divide the data set into a set of disjoint clusters without the hierarchical structure.

The most widely used partitional clustering algorithms are the prototype-based clustering algorithms where each cluster is represented by its center. The objective function is the sum of the distance from the pattern to the center. A popular partitional clustering algorithm K-means clustering, is essentially a function of minimization technique, where the objective function is the squared error. The main drawback of K-Means algorithm is that it joins to local minima from the starting position of the search [2,3].

The Firefly algorithm was introduced by XIN SHE YANG in 2008 [4]. It is a swarm intelligence optimization technique based on the theory that solution of an optimization difficult can be shown as a firefly which glows proportionally to its quality in a considered problem location. A swarm is a group of multi-agent systems such as fireflies, in which simple agents coordinate their activities to solve the complex problem of the allocation of communication to multiple forage places in dynamic environments. Consequently, each brighter firefly attracts its group, which makes the search space being explored efficiently. This algorithm is based on the behavior of social insects (fireflies). In social insect colonies, each particular seems to have its own agenda and yet the group as a whole appears to be highly organized. Algorithms based on nature have been established to show effectiveness and efficiency to solve difficult optimization problems. In this paper, we use the Firefly algorithm to find initial optimal cluster centroids and then K-Means algorithm is performed with optimized centroid for developing them and improve clustering accuracy [5].

The rest of the paper is organized as follows: Section II describes the Firefly Algorithm and K-Means clustering algorithm. Section III gives a detailed description of the Severe Asthma microarray data taken from NCBI dataset. Section IV gives the experimental results and analysis. Section V concludes the paper with some perspectives.

## II. METHODOLOGY

### *A. Firefly Optimization Algorithm*

In this paper it is proposed to adopt firefly algorithm to initialize the centroids for K-Means clustering to analysis its performance using severe asthma Gene expression data set taken form National Centre for

Biotechnology Information (NCBI). The firefly algorithm is working based on three rules according to the glowing nature of fireflies: (i) all fireflies are unisex and each firefly is attracted towards other fireflies regardless of their sex; (ii) the attraction is proportional to their brightness.Therefore between any two flashing fireflies, the less bright one will move towards the brighter one. In the surrounding, if there is no brighter one than a particular firefly, then it has to move randomly; (iii) the brightness of a firefly is determined by the nature of objective function. Initially at the beginning of clustering algorithm, all the fireflies are randomly dispersed across the entire search space. In Firefly algorithm, Fireflies are glow-worms that glow through bioluminescence. It is a population based algorithm to find the global optima of objective functions based on swarm intelligence, investigating the foraging behavior of fireflies [4].

In the FA, physical entities (agents or fireflies) are randomly distributed in the search space. Agents are thought of as fireflies that carry a luminescence quality, called Lucifer in, that emit light proportional to this value. The objective function is related to the sum of all training set instances of Euclidean distance in an n-dimensional space, as given in. The two phases of the firefly algorithm are as follows: (i) Variation of light intensity: Light intensity is related to objective values. So for a maximization /minimization problem a firefly with high/low intensity will attract another firefly with high/low intensity. Assume that there exists a swarm of n agents (fireflies) and represents a solution for a firefly i, whereas f ($x_i$) denotes its fitness value. Here the brightness i of a firefly is selected to reflect its current position of its fitness value f ($x_i$) [5].

$$l_i = \mathbf{f}(x_i),\ \mathbf{1 \le i \le n}. \quad (1)$$

ii. Movement toward attractive firefly: A firefly attractiveness is proportional to the light intensity seen by adjacent fireflies. Each firefly has its distinctive attractiveness β which implies how strong it attracts other members of the swarm. However, the attractiveness β is relative it will vary with the distance between two fireflies $x_i$ and $x_j$ at locations and respectively, is given as is a graphic representation of the distribution of data.

$$r_{ij} = \|x_i - x_j\|. \quad (2)$$

The attractiveness function β(r) of the firefly is determined by

$$\beta(r) = \beta_0 e^{-\gamma r^2} \quad (3)$$

Where $\beta_0$ is the attractiveness at r = 0 and γ is the light absorption coefficient. The movement of a firefly i at location $x_i$ attracted to another more attractive (brighter) firefly at $x_j$ location is determined by [6].

$$x_i(t+1) = x_i(t) + \beta_0 e^{-\gamma r^2}(x_j - x_i) b_i \quad (4)$$

Pseudo-Code: Description of Firefly Algorithm

```
Input:
Create an initial population of fireflies n within
d-dimensional search space (x_ik) , i = 1, 2, . . ., n and k = 1, 2, …, d
Evaluate the fitness of the population (x_ik) which is directly proportional to light intensity (l_ik)
Algorithm's parameter— β0, γ
Output:
Obtained minimum location: X_min
 begin
 repeat
        for i = 1 to n
             for j =1 to n
                  if ((l_j<l_i)
                     Move firefly i toward j in
                     d-dimension using (4)
                end if
               Attractiveness varies with distance r via
                exp [−r^2 ]
              Evaluate new solutions and update light intensity using (1)
          end for j
      end for i
     Rank the fireflies and find the current best
end
```

### B. *K-Means Algorithm for clustering*

K–Means clustering is a method of cluster analysis which aims at portioning of n observations into K clusters [2]. Each of the observation belongs to a cluster with the minimum distance between cluster centre and the observation point. It is done iteratively so that the observation point is at least distance from the center of cluster. The mean distance between the cluster center and observation is minimized during this iteration process[3].

The K-Means algorithm clusters d-dimensional data vectors into a predefined number of clusters on the basis of the Euclidean distance as the comparison criteria. Euclidean distances among data vectors are smallest amount for data vectors within a group as compared with distances to other vectors in various clusters. Vectors of the similar cluster are connected with one Centroid vector, which shows the middle of that group and is the mean of the data vectors that belong jointly [8].

K-Means algorithm is described hereunder.
Input:
d= $[d_1, d_2, d_3, \ldots . d_n]$ // set of n data items
K = number of desired clusters
Output:
A set of K clusters.
Steps:
i. Select K data items from d as initial centroids;
ii. Repeat the process of selecting the items
iii. Assign the each item di to the cluster which has the nearest
a. and the suitable centroids;
iv. Calculate the new mean value for each cluster.
v. Repeat the process until the criteria is satisfied.

## III. DATASET

Gene expression data is usually represented by a matrix, with rows corresponding to genes and columns corresponding to conditions, experiments or time points. The content of the matrix is the expression levels of each gene under each condition. The Severe Asthma: Bronchial Epithelial Cells Dataset is downloaded from National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO) website and used for this research work. The Severe Asthma microarray data contains about 41,000 genes with their corresponding Severe Asthma samples about 108. The Human Genome Database is a scientific database of the molecular biology and genetics of the Homosapiens [7],[13]. Table I shows the number of the genes and samples.

TABLE I. DATASET

| S.no | Name | No of genes | Samples |
|---|---|---|---|
| 1 | Severe Asthma | 41,0000 | 108 |

### A. *Dunn's Validity Index*

The Dunn Index (DI) introduced by J. C. Dunn in 1974, is a metric for evaluating clustering algorithms. This is part of a group of validity indices including the Davies–Bouldin index or Silhouette index, in that it is an internal evaluation scheme, where the result is based on the clustered data itself. Dunn index: One of the most cited indices is proposed by The Dunn index identifies clusters which are well separated and compact. The goal is therefore to maximize the inter-cluster distance while minimizing the intra-cluster distance. The Dunn index for K clusters is defined and represented in equation (5). If the Dunn index is large, then well separated clusters exist. Therefore, the maximum is observed for K equal to the most probable number of clusters in the data set [12].

The Dunn index definition is given by equation.

$$DU_k = \min_{i=1,\ldots,k} \left\{ \min_{j=1+1,\ldots,k} \left( \frac{diss(c_i, c_j)}{\max_{m=1,\ldots,k} diam(c_m)} \right) \right\} \quad (5)$$

diss ($c_i$, $c_j$) is distance between clusters $c_i$ and $c_j$, where dist ($c_i$, $c_j$) = $\min_{x_i \in c_i, x_j \in c_j}$ d ($x_i$, $x_j$),

d ($x_i$, $x_j$) is distance between data points $x_i \in c_i$and $x_j \in c_j$,diam ($c_l$) is diameter of cluster $c_l$ where

$$diam(c_i) = \max_{x_{I1}, x_{I1} \in c_i} d(x_{I1}, x_{I2}) \quad (6)$$

*B. Davies Bouldin Index*

The Davies–Bouldin index (DBI) introduced by David L. Davies and Donald W. Bouldin in 1979 is a metric for evaluating clustering algorithms. Similar to the Dunn index, Davies-Bouldin index identifies clusters which are far from each other and compact. Davies-Bouldin index (DB) is defined according to Equation.

$$DB_k = \frac{1}{k}\sum_{i=1}^{k} \max_{j=1,\ldots,k, i \neq j} \left\{ \frac{diam(c_i) + diam(c_j)}{\|c_i - c_j\|} \right\} \quad (7)$$

In this case, the diameter of a cluster is defined as:

$$diam(c_1)\left(\frac{1}{n_i}\sum_{x \in c_1} \|x - z_1\|^2\right)^{1\backslash 2} \quad (8)$$

where $n_i$ is the number of points and $z_i$ is the centroid of cluster $c_i$. Since the objective is to obtain clusters with minimum intra-cluster distances, small values or DB are interesting. Therefore, this index is minimized when looking for the best number of clusters [8,9].

*C. Silhouette Validity Index*

Silhouette refers to a method of interpretation and validation of consistency within cluster of data. The technique provides a brief graphical representation of how well each object lies within its cluster. It was first described by Peter J. Rousseeuw in 1986. Silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and badly matched to neighboring clusters. The silhouette statistic is another well-known way of estimating the number of groups in a data set. The Silhouette Index (SI) computes for each point a width depending on its membership in any cluster. This silhouette width is then an average over all observations. This is represented is (9).

$$SI_k = \frac{1}{n}\sum_{i=1}^{n} \frac{(b_i - a_i)}{max(a_i, b_i)} \quad (9)$$

Where $a_i$ is average dissimilarity of $i^{th}$data point to all other points in the same cluster; $b_i$ is minimum of average dissimilarity of $i^{th}$ data point to all data points in other cluster. Value of SI close to 1 indicates that the data point is assigned to a very appropriate cluster. If is close to zero, it means that that data pint could be assign to another closest cluster as well because it is equidistant from both the clusters. If is close to –1, it means that data is misclassified and lies somewhere in between the clusters. The overall average silhouette width for the entire data set is the average for all data points in the whole dataset. The largest overall average silhouette indicates the best clustering. Therefore, the number of cluster with maximum overall average silhouette width is taken as the optimal number of the clusters [10].

*D. Average Correlation Value*

It is used to evaluate the homogeneity of a cluster. High ACV indicates the best similarities among the genes [11]. Matrix A = $(A_{ij})$ has the ACV which is represented in the following equation

$$ACV = \left\{ max \sum_{i=1}^{i=m} \sum_{j=1}^{m} \frac{|c\_row_{ij}|}{m^2 - m}, \sum_{p=1}^{n} \frac{|c\_col_{pq}|}{n^2 - n}^{-n} \right\} \quad (10)$$

Where C_$row_{ij}$– is the correlation coefficient between rows i and j,

C_$col_{pq}$ - is the correlation coefficient between columns p and q, ACV approaching 1 denote a significant cluster. Such technique may be found in.

## IV. Experimental Results And Analysis

Table I shows the computational results of Davies-Bouldin index, Dunn Index, Silhouette Index and Average Correlation Value for the proposed FFKM algorithm and the traditional K-Means algorithm.

The parameter value is FireFly K-Means algorithm using the best value of β= 0. 9424.

Table II. Computational Results

| Cluster Methods | No. of Clusters | Davies-Bouldin index | Dunn Index | Silhouette Index | ACV |
|---|---|---|---|---|---|
| K-Means | 2 | 0.9967 | 0.0099 | 0.3762 | 0.7431 |
| | 4 | 1.4248 | 0.0123 | 0.1885 | 0.8010 |
| | 6 | 1.3119 | 0.00802 | 0.1354 | 0.7576 |
| FFKM | 2 | 0.7967 | 0.0886 | 0.661 | 0.9424 |
| | 4 | 0.9761 | 0.0684 | 0.5885 | 0.9314 |
| | 6 | 1.015 | 0.0309 | 0.5231 | 0.9071 |

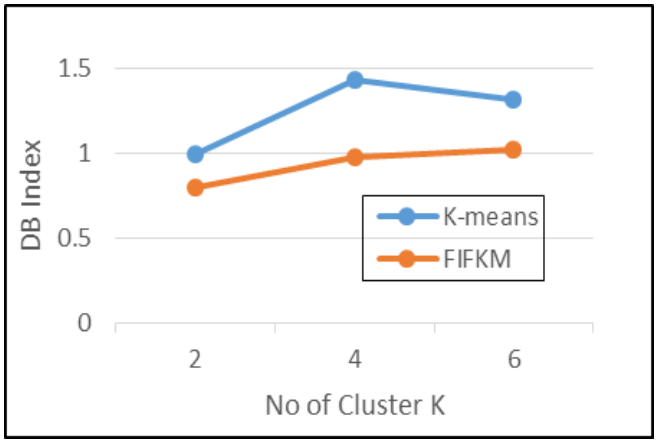


Figure 1. DB Index vs No of cluster

Figure 1 shows the graph plot of Davies Bouldin (DB) validity index values as a function of the number of clusters K. This index attempts to minimize the average distance between consequently cluster and the one most similar to it. From the graph it is clear that firefly K-Means clustering provides the minimum values for this index, thus outperforming the other techniques.

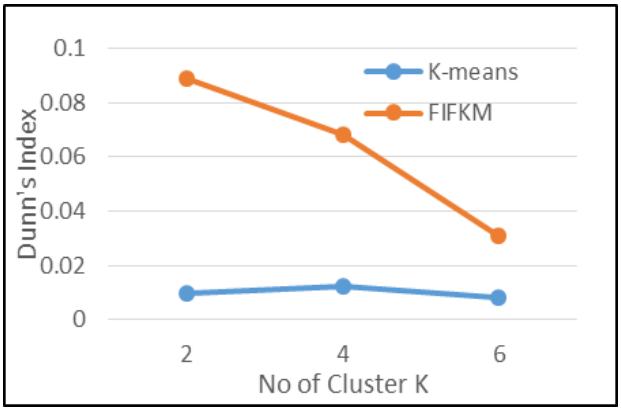


Figure 2. Dunn's Index vs No of cluster

Figure 2 shows the graph plot of Dunn's validity index values as a function of the number of clusters K. The main objective of the measure is to maximize the inter-cluster distances and minimize the intra-cluster distances. From the graph it is clear that we are success best performance using that firefly K-Means clustering in terms of the Dunn's validity Index.

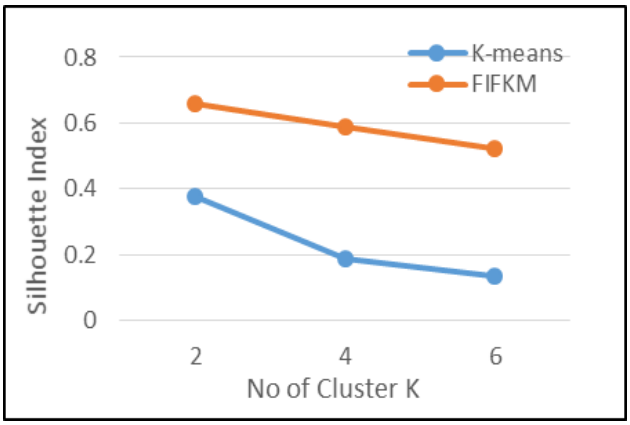


Figure 3. Silhouette Index vs No of cluster

Figure 3 shows the graph plot of Silhouette validity index values as a function of the number of clusters K. Silhouette values range between -1 and 1. A value close to 1 indicates that the data point is assigned to a very perfect cluster. A value is close to zero means that data pint could be assign to another closest cluster as well because it is significant from both the clusters. If the value is close to –1, it means that data is misclassified and lies anywhere in between the clusters. Our results show that all the techniques generate the values between 0 and 1.

## V. CONCLUSION

In this paper Firefly algorithm has been used to intialise the centroids for K-Means to improve the performance of the K-Means algorithm in the domain gene expression data. identify centroid values of traditional K- Mean's clustering algorithm using Firefly optimization method. In the traditional K-Means clustering algorithm, the centroid values are randomly selected. By using the proposed FIFKM algorithm the best values for the initial cluster centroids can be obtained. The traditional K-means clustering algorithm is compared with firefly K-means clustering algorithm for severe asthma dataset taken from NCBI. The firefly K-Means clustering algorithm gives the best result when compared by the traditional K-Means clustering algorithm in relations of Davies-Bouldin index, Dunn Index, Silhouette Index and Average Correlation Value.

## ACKNOWLEDGEMENT

The authors immensely acknowledge the UGC, New Delhi for partial financial assistance under UGC-SAP (DRS) Grant No. F.3-50/2011.

## REFERENCES

[1] E. Naghieh and Y. Peng, "Microarray Gene Expression Data Mining: Clustering Analysis Review", Techniques, 2009.

[2] E N Sathishkumar K Thangavel D Arul Pon Daniel "Effective Clustering Algorithm for Gas Sensor Array Drift Dataset" International Journal of Computational Intelligence and Informatics, Vol. 3: No. 3, October - December 2013.

[3] Jyoti Yadav#1 , Monika Sharma*2" A Review of K-mean Algorithm" International Journal of Engineering Trends and Technology (IJETT) – Volume 4 Issue 7- July 2013.

[4] Adil Hashmi, Nishant Goel, Shruti Goel, Divya Gupta, "Firefly Algorithm for Unconstrained Optimization"e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 11, Issue 1, PP 75-78, May. - Jun. 2013.

[5] X. S. Yang, Firefly Algorithm: Stochastic Test Functions and Design Optimisation, Int. J. Bio- Inspired Computation, Vol. 2, No. 2, pp.78–84, 2010.

[6] Saibal K. Pal, Dr., C.S Rai, Dr. , Prof., Amrit Pal Singh, Asst. Prof." Comparative Study of Firefly Algorithm and Particle Swarm Optimization for Noisy NonLinear Optimization Problems" I.J. Intelligent Systems and Applications, 2012, 10, 50-57 Published Online in MECS (http://www.mecs-press.org/), September 2012.

[7] http://www.ncbi.nlm.nih.gov/gquery/?term=gse43696.

[8] Sandro Saitta, Benny Raphael, and Ian F.C. Smith, "A Bounded Index for Cluster Validity", Springer Conference, 2007.

[9] Juan Carlos Rojas Thomas" New Version of Davies-Bouldin Index for Clustering Validation Based on Cylindrical Distance"

[10] Peter J. ROUSSEEUW" Silhouettes: a graphical aid to the interpretation and validation of cluster analysis" Journal of Computational and Applied Mathematics 20 (1987) 53-65 Received 13 June 1986 Revised 27 November 1986.

[11] R.Rathipriya , K.Thangavel , J.Bagyamani, "Evolutionary Biclustering of Clickstream Data" , IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011.

[12] Hitesh Kumari Yadav (Student), Sunil Dhankar (Reader" Dynamic parallel K-Means Algorithm Based On Dunn's Index Method" International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume – 5 Issue -03 March, 2016.

[13] Porkodi Rajendran and Deepika Thangavel, "Clustering of Microarray Data to Identify Enriched Go Terms of Genes in Severe Asthma Dataset using Gene Enrichment Analysis" Indian Journal of Science and Technology, Vol 9(8), DOI: 10.17485/ijst/2016/v9i8/86068, February 2016.