AN EFFICIENT HYBRID APPROACH FOR DATA CLUSTERING USING DYNAMIC K-MEANS ALGORITHM AND FIREFLY ALGORITHM

Sundararajan S.[1] and Karthikeyan S.[2]
[1]Department of MCA, SNS College of Technology, Coimbatore, Tamil Nadu, India
[2]Department of Computer Applications, Karpagam University, Coimbatore, Tamil Nadu, India
E-Mail: sundar_mtp@yahoo.co.in

ABSTRACT

Clustering is an important task in data mining to group data into meaningful subsets to retrieve information from a given dataset. Clustering is also known as unsupervised learning since the data objects are pointed to a collection of clusters which can be interpreted as classes additionally. The proposed approach concentrates on the K-means algorithm for enhancing the cluster quality and for fixing the optimal number of cluster. Numerous clusters (K) are taken as input. Firefly algorithm is mainly used for solving optimization problems. The proposed approach uses dynamic K-means algorithm is used for dynamic data clustering approaches. It can be applied to both known number of clusters as well as unknown number of clusters. Hence, the user can either fix the number of clusters or they can fix the minimum number of required clusters. If the number of clusters is static, it works like K-means algorithm. If the number of clusters is dynamic, then this algorithm determines the new cluster centers by adding one to the cluster counter in each iteration until the required cluster quality is achieved. The proposed method uses Modified Firefly algorithm to determine the centroid of the user specified number of clusters. This algorithm can be extended using dynamic k-means clustering to enhance centroids and clusters. Thus the proposed Dynamic clustering method increases the cluster quality and modified firefly algorithm increases optimality for the iris and wine datasets. Experimental results proved that the proposed methodology attains maximum cluster quality within a limited time and achieves better optimality.

Keywords: clustering, dynamic K-means algorithm, modified firefly algorithm, cluster quality, optimality.

INTRODUCTION

Clustering is defined as the cataloging of objects into several groups. Clustering detaches the dataset into a subset of objects, with a goal that the data present in each subset possibly shares certain similar features frequently propinquity based on certain defined distance measure. Data clustering is a common approach for statistical data analysis, which is most generally used in various fields such as image analysis, pattern recognition and bioinformatics. (Osmar R. Zaiane)

Cluster Analysis (CA) is a probing data analysis method for managing collected data into substantial groups, clusters, or, catalogs according to the permutation, which increases the similarity of cases inside a cluster. All together, it also increases the dissimilarity between the other groups that are primarily unknown. CA makes new groupings without any fixed notion of what clusters might arise, while discriminant analysis classifies the items into previously known groups. CA does not give any details related to the existence of a particular cluster. Thus each cluster illustrates the data being collected and the class to which it belongs to. Items in each group, having similar characteristics based on certain parameters and it is dissimilar to those in other groups. (Fisher D.H., 1987)

Spatial clustering is a major constituent of spatial data mining and it is applied to reclaim a pattern from the data objects distribution in a given dataset. The resulted clusters should be illuminated to find each one's consequence in the context for which the clustering is applied. With the help of cluster analysis, the quality of the spatial clusters should be checked (Chandra *et al.*, 2011).

The simplest clustering algorithm is K-means. Since clustering algorithms are chosen in pattern recognition because of the nature of available data. K-means is one of the most widely used algorithms for clustering due to its simplicity, empirical success, ease of implementation and efficiency. (Anil K. Jain, 2009) The basic k-means algorithm creates only spherical shaped Clusters Spatial clustering algorithm. The general factors for determining particular clustering algorithm are based on the following parameters:

a) Data set size
b) Data dimensionality
c) Time complexity

Dynamic clustering problem arises while gathering data during a particular time interval. The dataset should be focused at each time interval which requires dynamic update of the clustering results. The simple method to overcome this problem is to perform clustering immediately when the data is measured at each instant of time and previous data will be neglected. The traditional clustering algorithms (McQueen, 1967; Gray R, 1984) reduce the fitness measure between the model and the data.

Firefly algorithm (FA) is a swarm intelligence optimization technique which is based on the assumption that optimization problem can be shown as a firefly which rises relatively to its quality. Therefore, each brighter firefly attracts its neighboring clusters, which makes the search space being investigated efficiently. FA can be used for nonlinear design problems (X. S. Yang, 2010) and multimodal optimization problems (X. S. Yang, 2009). FA can be used for finding global optima in two dimensional spaces.

The proposed method uses dynamic clustering of data with the help of dynamic k-means algorithm and firefly algorithm. The proposed method can be applied to the dynamic dataset for effective clustering to determine the quality of cluster data. The paper can be arranged as follows: The related works involved in the dynamic clustering, k-means algorithm and firefly algorithm is explained in the Section II. The proposed method to cluster data dynamically is summarized in the Section III. The Section IV deals with the results obtained in the proposed approach.

RELATED WORKS

Man Lung Yiu *et al*., (Man Lung *et al*., 2011) studied a top-k spatial preference query which offers a new type of ranking for spatial objects depending on the qualities of features in their neighborhood. The neighborhood of an object p is determined by the scoring function. They also presented five algorithms such as SP, GP, BB, BB* and FJ for processing top-k spatial preference queries. Based on their experimental findings, BB* is scalable to large data sets and it is the most robust algorithm with respect to various parameters.

Ahamed Shafeeq *et al*., (Ahamed Shafeeq *et al*., 2012) proposed an improved data clustering for the unknown data set. The k-means algorithm is well known for its ease and the modification is done to the k-means method for increasing the data clustering. The K-means algorithm takes several clusters as input from the user. The main disadvantage of the K-means algorithm is fixing the number of clusters. If the fixed number of cluster is very small then there is a possibility of grouping dissimilar objects. Suppose the fixed cluster is large, then the similar objects will be clustered into different groups. To overcome this problem, the optimal number of clusters is determined. The main drawback of the proposed method is it takes more computational time than the K-means for larger datasets.

The main intention of data clustering which is also known as cluster analysis is to determine the natural clustering of a points, set of patterns, or objects. Webster (Webster) presented cluster analysis as "a statistical classification technique for discovering whether the individuals of a population fall into different groups by making quantitative comparisons of multiple characteristics."

Abrantes *et al*., (Arnaldo *et al*., 1998) described a method for the segmentation of dynamic data. This work is based on an integrated framework for constrained clustering which is extended by using a motion model for the clusters. This method includes global and local evolution of the data centroids. They also proposed noise model to increase the robustness of the dynamic clustering algorithm with respect to outliers.

Hassanzadeh *et al*., (Tahereh Hassanzadeh *et al*.,) presented a new approach using firefly algorithm to cluster data. He explained how firefly algorithm can be used to find the centroid of the user specified number of clusters. This algorithm then extended to use k-means clustering to refined centroids and clusters and he named this hybrid algorithm as K-FA. K-means clustering is a common and simple approach for data clustering but this method has some limitation such as initial point sensibility and local optimal convergence. Firefly algorithm is a swarm based algorithm that use for solving optimization problems.

Christos Bouras *et al*., (Christos Bouras, 2011) proposed an unsupervised feature ranking algorithm for evaluating features using discovered bi-clusters which are local patterns that can be extracted from a data matrix. The bi-clusters can be defined as sub-matrices which are used for scoring relevant features from two aspects. They are interdependence of features and the separable of instances. Rank the features based on their accumulated scores from the total discovered bi-clusters before the classifying pattern.

METHODOLOGY

Clustering is the process of assembling the data records into significant subclasses (clusters) in a way that increases the relationship within clusters and reduces the similarity among two different clusters (Fisher D.H., 1987). Other names for clustering are unsupervised learning (machine learning) and segmentation. Clustering is used to get an overview over a given data set. A set of clusters is often enough to get insight into the data distribution within a data set. Another important use of clustering algorithms is the preprocessing for some other data mining algorithm.

▪ K-means algorithm

K-Means algorithm is a clustering algorithm that classifies the input data points into multiple classes based on their inherent distance from each other. Assumes that the data features form a vector space and this algorithm tries to find natural clustering in them. The points are clustered around centroid $\mu_i \; \forall_i = 1 \ldots k$ which are obtained by minimizing the objective

$$V = \sum_{i=1}^{k} \sum_{x_j \in s_i} (x_j - \mu_i)^2 \qquad (1)$$

where there are k clusters Si, i = 1, 2,……, k and μ_i is the centroid or mean point of all the points $x_j \in S_i$

The algorithm takes a 2 dimensional image as an input. Various steps in the algorithm are as follows:

a) Compute the intensity distribution (also called the histogram) of the intensities.
b) Initialize the centroids with k random intensities.
c) Repeat the following steps until the cluster labels of the image do not change anymore.
d) Cluster the points based on distance of their intensities from the centroid intensities.

$$c^{(i)} := \arg\min_j \left\| x^{(i)} - \mu_j \right\|^2 \qquad (2)$$

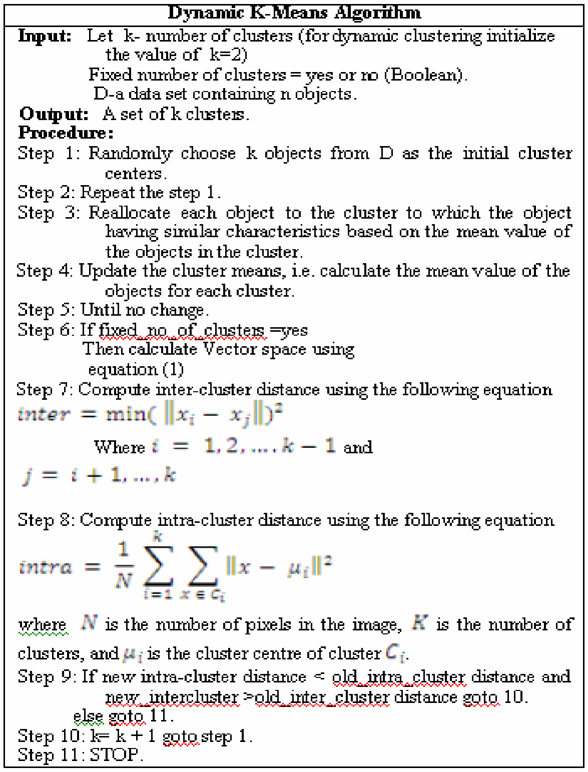
e) Compute the new centroid for each of the clusters.

$$\mu_i := \frac{\sum_{i=1}^{m} 1\{c_{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c_{(i)} = j\}} \qquad (3)$$

where k is a parameter of the algorithm (the number of clusters to be found), i iterates over the all the intensities, j iterates over all the centroids and μ_i are the centroid intensities. (Suman Tatiraju *et al.*,)

- **Dynamic K-means algorithm**

The K-means algorithm determines the predefined number of clusters. It is very much essential to find the number of clusters for unknown dataset dynamically. Initially fixing the number of clusters brings about poor quality clustering. The proposed method finds the number of clusters on the runtime which is derived from the output of the cluster quality. The proposed method can be applied to both known number of clusters and unknown number of clusters. The user can either fix the number of clusters or they can fix the minimum number of required clusters. If the number of clusters is static, it works like K-means algorithm. If the number of clusters is dynamic, then this algorithm determines the new cluster centers by adding one to the cluster counter in each iteration until the required cluster quality threshold is achieved. The dynamic k-means algorithm is summarized as follows:



```
Dynamic K-Means Algorithm
Input:  Let k- number of clusters (for dynamic clustering initialize
          the value of k=2)
        Fixed number of clusters = yes or no (Boolean).
        D-a data set containing n objects.
Output: A set of k clusters.
Procedure:
Step 1: Randomly choose k objects from D as the initial cluster
          centers.
Step 2: Repeat the step 1.
Step 3: Reallocate each object to the cluster to which the object
          having similar characteristics based on the mean value of
          the objects in the cluster.
Step 4: Update the cluster means, i.e. calculate the mean value of the
          objects for each cluster.
Step 5: Until no change.
Step 6: If fixed_no_of_clusters =yes
        Then calculate Vector space using
         equation (1)
Step 7: Compute inter-cluster distance using the following equation
$inter = \min( \|x_i - x_j\|)^2$
          Where $i = 1, 2, \ldots, k-1$ and
$j = i+1, \ldots, k$

Step 8: Compute intra-cluster distance using the following equation
$intra = \frac{1}{N} \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$
where $N$ is the number of pixels in the image, $K$ is the number of
clusters, and $\mu_i$ is the cluster centre of cluster $C_i$.
Step 9: If new intra-cluster distance < old_intra_cluster distance and
          new_intercluster >old_inter_cluster distance goto 10.
       else goto 11.
Step 10: k= k + 1 goto step 1.
Step 11: STOP.
```

- **Modified firefly algorithm**

Firefly Algorithm (FA) brightens firefly (local optima) in each iteration, attracts the neighboring fireflies towards itself in maximizing optimal problems. Fireflies move regardless of the global optima and it reduce the ability of the firefly algorithm to find global best in the standard Firefly algorithm.

To overcome the problems of FA and to increase the collective movement of fireflies, a modified firefly algorithm (MFA) is proposed in this paper. The global optimum in firefly's movement is used in the proposed method. Global optimum is mainly related to optimization problem and it can be a firefly which values are either maximum or minimum. Global optima will be updated in any iteration. When a firefly is compared with other fireflies instead of the one firefly being allowed to influence and to attract its neighbors, global optima (a firefly that have maximum or minimum value) in each iteration can be allowed to influence others and affects their association in the standard FA algorithm. But in the modified FA algorithm when a firefly is compared with corresponding firefly and if the corresponding firefly will be brighter, then the compared firefly will move towards corresponding firefly which is considered for global optima. Cartesian distance is used to compute the distance of fireflies to global optima which is determined using the following equation.

$$r_{i,best} = \sqrt{(x_i - x_{gbest})^2 + (y_i - y_{gbest})^2} \qquad (4)$$

The movement of the firefly can be determined by the following equation,

$$x_i = x_i + \left(\beta_0 e^{-\gamma r_{ij}^2} (x_j - x_i) + \beta_0 e^{-\gamma r_{gbest}^2} (x_{gbest} - x_i) \right) + \alpha\left(rand - \frac{1}{2}\right) \qquad (5)$$

Where $gbest$ is global optimal and x_{gbest} is the coordinate of global optima, α is the randomization parameter and rand is the random number generator in which its numbers are uniformly distributed in interval [0, 1]. β0 is the attractiveness at r = 0 and γis the light absorption coefficient at the source.

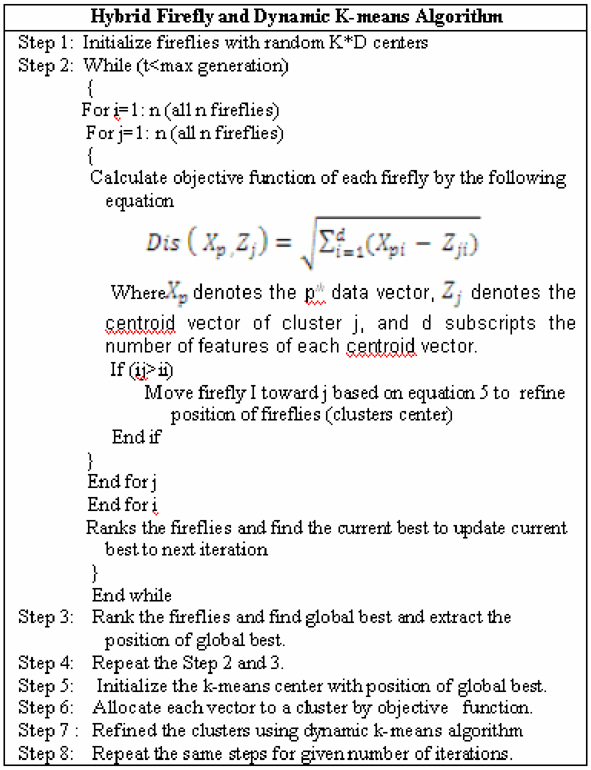
For the most cases of implementations, β0 = 1 and$\alpha \in [0, 1]$. The parameter γ characterizes the variation of the attractiveness and its value is important to determine the speed of the convergence. Mostly this value varies from 0.01 to 100.

- **Proposed hybrid modified firefly and dynamic K-means algorithm**

To improve and increase the accuracy of dynamic k-means algorithm, initialize the dynamic k-means algorithm with optimal centers, which can be calculated by modified firefly algorithm. The structure of a firefly position in clustering problem space is shown below.

$$[Z_{11,} Z_{12}, \dots Z_{1d}, Z_{21,} Z_{22}, \dots, \dots Z_{2d}, \dots \dots Z_{k1,} Z_{k2}, \dots, Z_{kd}]$$

Clustering has two stages in the proposed method. At first initialize fireflies with random values as shown in the structure. Consider there are D-dimensional data and K clusters; hence each firefly has K×D dimensions. Euclidean distance which is the objective function must have minimum values. The modified firefly algorithm mechanism should be done until the given number of iteration. Then, initialize the dynamic k-means position with the best firefly. The dynamic k-means clustering refine the centers. The proposed hybrid clustering algorithm can be summarized as follows:



```
Hybrid Firefly and Dynamic K-means Algorithm
Step 1:  Initialize fireflies with random K*D centers
Step 2:  While (t<max generation)
           {
          For i=1: n (all n fireflies)
           For j=1: n (all n fireflies)
           {
            Calculate objective function of each firefly by the following
              equation
                 $Dis\left(X_p, Z_j\right) = \sqrt{\sum_{i=1}^{d}(X_{pi} - Z_{ji})}$
             Where $X_p$ denotes the $p^{th}$ data vector, $Z_j$ denotes the
              centroid vector of cluster j, and d subscripts the
              number of features of each centroid vector.
             If (ij>ii)
                  Move firefly I toward j based on equation 5 to refine
                     position of fireflies (clusters center)
             End if
           }
           End for j
           End for i
          Ranks the fireflies and find the current best to update current
             best to next iteration
           }
           End while
Step 3:    Rank the fireflies and find global best and extract the
             position of global best.
Step 4:    Repeat the Step 2 and 3.
Step 5:     Initialize the k-means center with position of global best.
Step 6:    Allocate each vector to a cluster by objective   function.
Step 7 :   Refined the clusters using dynamic k-means algorithm
Step 8:    Repeat the same steps for given number of iterations.
```

EXPERIMENTAL RESULTS

In order to evaluate the effectiveness of the proposed algorithm, the following UCI repository data set are used.

a) Iris dataset and
b) Wine dataset

The experimental results show that the proposed method is more efficient than dynamic K-means algorithm in determining the cluster quality and optimality for the unknown dataset.

Number of clusters

The number of clusters obtained in the proposed approach is shown in the Table-1.

Table-1 shows the number of clusters obtained for the Iris dataset and Wine dataset. The proposed Modified firefly algorithm with dynamic k-means algorithm provides the more number of clusters effectively.

Table-1. Comparison of the number of clusters.

Data points	**No. of clusters** (Dynamic K-means)	**No. of clusters** (Modified FA with dynamic K-means)
Iris dataset	4	6
Wine dataset	9	11

Figure-1 shows the comparison of the proposed Modified firefly algorithm and dynamic k-means with the dynamic k-means clustering algorithm. The number of clusters obtained is more in the Modified firefly algorithm with dynamic k-means algorithm.

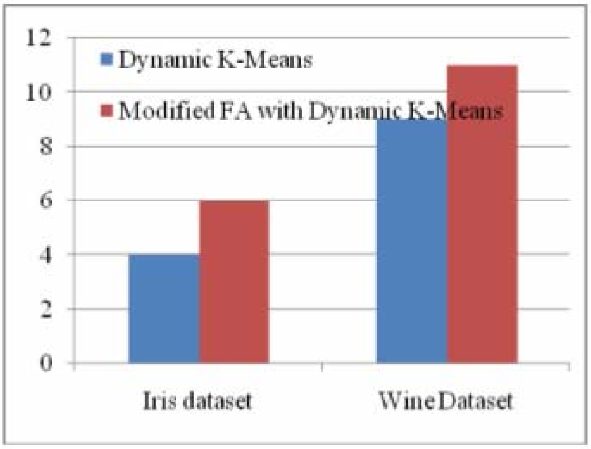


Figure-1. Comparison of the number of clusters.

Inter cluster distance

Table-2 shows the inter cluster distance values for the Iris dataset and Wine dataset. It is clear from the table that the proposed modified firefly algorithm with dynamic k-means algorithm achieves better results than the dynamic k-means algorithm.

Table-2. Comparison of the inter cluster distance values.

Data points	Dynamic K-means	Modified FA with dynamic K-means
Iris Dataset	0.06961	0.07453
Wine Dataset	0.04837	0.08077

Figure-2 reveals the comparison of the inter-cluster distance of the proposed Modified firefly algorithm with dynamic k-means with the dynamic k-means clustering algorithm. The inter cluster distance values are larger in the modified firefly algorithm with dynamic k-means algorithm.

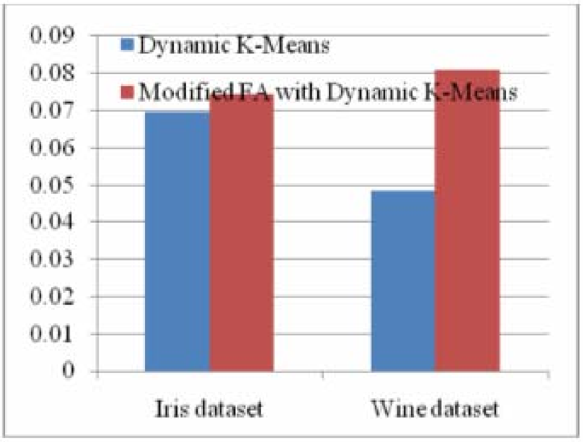


Figure-2. Comparison of the inter-cluster distance.

Intra cluster distance

Table-3 shows the intra cluster distance values for the Iris dataset and Wine dataset. It is clearly showed from the table that the proposed modified firefly algorithm with dynamic k-means algorithm has lesser intra cluster values than the dynamic k-means algorithm.

Table-3. Comparison of the intra cluster distance values.

Data points	Dynamic K-means	Modified FA with dynamic K-means
Iris dataset	0.02814	0.01686
Wine dataset	0.01376	0.01136

Figure-3 illustrates the comparison of the intra-cluster distance of the proposed modified firefly algorithm and dynamic k-means with the dynamic k-means clustering algorithm. The inter cluster distance values are smaller in the proposed modified firefly algorithm with dynamic k-means algorithm.

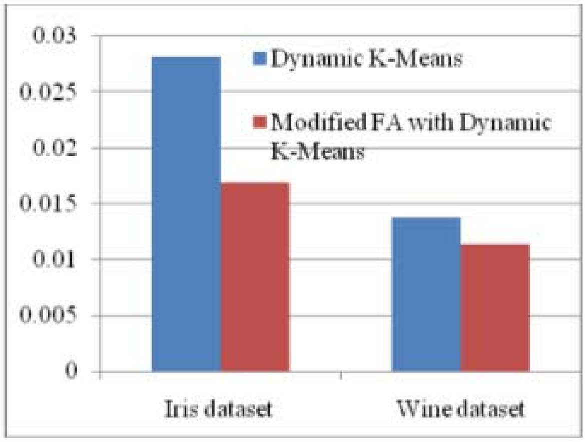


Figure-3. Comparison of the intra-cluster distance.

Execution time

Figure-4 shows the comparison of the execution time.

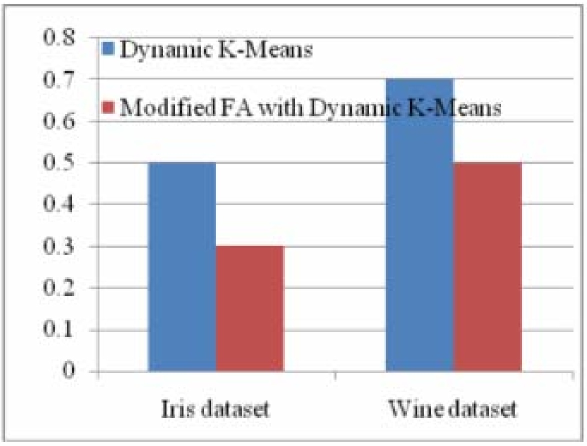


Figure-4. Comparison of the execution time.

It is clear from the experimental results that the proposed algorithm takes less time for computing the datasets than the k-means algorithm. The proposed Modified firefly algorithm with dynamic k-means algorithm provides the cluster quality effectively with lesser time.

CONCLUSIONS

Clustering is a most important unsupervised classification technique. An effective approach for data clustering using hybrid approach is proposed in the paper. The modified firefly algorithm is combined with dynamic k-means algorithm in order to increase the optimality and cluster quality. This hybrid approach outperforms the dynamic k-means algorithm and shows better results in providing cluster quality and improving optimality. The proposed determines the best global optima clusters using Modified Firefly algorithm. Hence the optimality is improved. The proposed approach finds the optimal number of clusters effectively during runtime using dynamic k-means. Thus the cluster quality is improved. The experimental result shows that the proposed find the maximum number of clusters in less time with better cluster quality and increased optimality. Thus the proposed method effectively clusters the data by using hybrid approach. As a future work, the optimality and the quality can be improved further by combining various effective clustering algorithms.

REFERENCES

Ahamed Shafeeq B.M. and Hareesha K.S. 2012. Dynamic Clustering of Data with Modified K-Means Algorithm. International Conference on Information and Computer Networks.

Anil K. Jain. 2009. Data Clustering: 50 Years Beyond K-Means. Pattern Recognition Letters.

Arnaldo J. Abrantesy and Jorge S. Marques. 1998. A Method for Dynamic Clustering of Data. British Machine Vision Conference (BMVC). doi:10.5244/C.12.16.

Chandra E. and Anuradha V. P. 2011. A Survey on Clustering Algorithms for Data in Spatial Database Management Systems. International Journal of Computer Applications. 24(9): 975-8887.

Christos Bouras and Vassilis Tsogkas. 2011. Clustering User Preferences using W-K Means. 7th International Conference on Signal Image Technology and Internet-Based Systems.

Fisher D.H. 1987. Conceptual Clustering, Learning from Examples, and Inference. Proceedings of the 4th International Workshop on Machine Learning, Irvine, CA. pp. 38-50.

Gray R. 1984. Vector Quantization. IEEE ASSP Magazine. pp. 4-24.

Man Lung Yiu, Hua Lu, Nikos Mamoulis and Michail Vaitis. 2011. Ranking Spatial Data by Quality Preferences. IEEE Transactions On Knowledge And Data Engineering. 23(3).

McQueen. 1967. Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of the 5th Berkeley Symp. Math. Stat. and Prob., Vol. 1, Univ. California Press. pp. 281-286.

Osmar R. Zaiane. Principles of Knowledge Discovery in Databases Chapter 8: Data Clustering.

Suman Tatiraju and Avi Mehta. Image Segmentation using K-Means Clustering, EM and Normalized Cuts.

Tahereh Hassanzadeh and Mohammad Reza Meybodi. A New Hybrid Approach for Data Clustering using Firefly Algorithm and K-means.

Webster. Cluster analysis. Merriam-Webster Online Dictionary, 2008 (Feb). http://www.merriam-webster-online.com.

X. S. Yang. 2009. Firefly Algorithm for Multimodal Optimization. In: Stochastic Algorithms: foundations and applications SAGA lecture notes in computer sciences. pp. 169-178.

X. S. Yang. 2010. Firefly Algorithm‹ Stochastic Test Functions and Design optimization. International Journal of bio-inspired computation.