# *Optimized K-means using Firefly Algorithm*

Chaithanya Chikkannaswamy
Dept of EECS
Syracuse University
cchikkan@syr.edu

Gahan Gurumurthy
Dept of EECS
Syracuse University
ggurumur@syr.edu

***Abstract*— Data clustering is a popular technique for analyzing data across various fields, including data mining, pattern recognition and image analysis. K-means clustering is a common and simple approach for data clustering but this method has some drawbacks such as local optimal convergence and initial point sensibility. Firefly algorithm is a swarm based algorithm that use for solving optimization problems. This paper presents a new approach to using firefly algorithm to cluster data. It is shown how firefly algorithm can be used to find the centroid of the user specified number of clusters. The algorithm then extended to use k-means clustering to refined centroids and clusters. This new hybrid algorithm is called K-FA. The experimental results showed the accuracy and capability of the proposed algorithm to data clustering.**

***Keywords-component; firefly algorithm; lustering; k-means; optimization.***

## I. INTRODUCTION

Clustering is a most important unsupervised classification technique. Clustering algorithms have been applied to a wide range of problems, including data mining , pattern recognition, data compression , machine learning, etc. When the number of clusters, K, is known a priori, clustering may be formulated as distribution of n objects in N dimensional space among K groups in such a way that objects in the same cluster are more similar in some aspects than the others in different clusters. This involves minimization of some optimization criterion. The K-means algorithm , starting with k random cluster centers then partitions a set of objects into k subsets. This method is a one the most popular and simple method that widely used in clustering. However, the k-means clustering has several drawbacks such as being trapped in local optima, as well as local maxima and being sensitive to initial cluster centers. One method to refined kmeans algorithm is hybridizing it with efficient optimization method. There is different optimization algorithm like Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Artificial Fish Swarm Algorithm (AFSA) and Bee Colony. The Firefly algorithm was recently introduced by XINSHE YANG in Cambridge University. This swarm intelligence optimization technique is based on the assumption that solution of an optimization problem can be shown as a firefly which glows proportionally to its quality in a considered problem setting. Consequently, each brighter firefly attracts its partners, which makes the search space being explored efficiently. Yang used the FA for nonlinear design problems and multimodal optimization problems and showed the efficiency of the FA for finding global optima in two dimensional environments. In this paper, we use the firefly algorithm to find initial optimal cluster centroid and then initial k-means algorithm with optimized centroid to refined them and improve clustering accuracy. Proposed method experimental results compared with PSO, K-means, K-PSO method on standard datasets of Iris, WDBC, Sonar, Glass and Wine. The results show that the proposed algorithm has a higher efficacy than the other algorithms. The rest of the paper is organized as follows: Section 2 describes the Problem and Data. Section 3 gives a detailed description of our Approach. Section 4 shows the Results and followed by the Conclusion and References.

## II. PROBLEM AND DATA DESCRIPTION

### *A. Problem*

Using the firefly algorithm to optimize k-means clustering. k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. The “firefly algorithm” (FFA) is a modern metaheuristic algorithm, inspired by the behavior of fireflies

### *B. Data Description*

Dataset - xclara: Bivariate Data Set with 3 Clusters
Description : An artificial data set consisting of 3000 points in 3 well-separated clusters of size 1000 each.
Format : A data frame with 3000 observations on 2 numeric variables giving the x and y coordinates of the points, respectively.

## III. Approach

Two key factors in accomplishing great K-means clustering results: one is the capacity to discover great centroid areas at the start. The other factor is the capacity to investigate worldwide optima past the local ones. For the primary factor, the underlying centroids ought to be circulated so that the bunches framed upon them will accomplish a worldwide ideal. In clustering, the worldwide ideal can be viewed as having greatest intrasimilarities and least between similitudes. Nonetheless, without knowing where the worldwide ideal is, it is computationally very hard to track down one of every enormous hunt space. K-means calculations bunch information into non-covering raised gatherings what's more, consistently join, which is its legitimacy, despite the fact that not essentially to the worldwide ideal. Similarly as with any partitional clustering calculation, they are exceptionally delicate to the underlying parameters: the number k of bunches and their centroids, separately. K-means clustering regularly prompts neighborhood optima which might be a long way from the best outcomes . By and by, to get the best clustering outcomes, the K-means calculation is frequently applied commonly to various irregular inceptions. Notwithstanding, this is done at the expense of expanded calculation and preparing time. Every individual random trial is autonomous and no assurance in finding the best outcome.

This paper is intended to show a basic construct of integration of a K-means clustering algorithm with bio-inspired optimization algorithms. A major difference between the integrated algorithms and original K-means is the additional exploration function which is called optimization, which progressively optimizes or improves the currently best solution with a new solution from an unexplored part of the search space

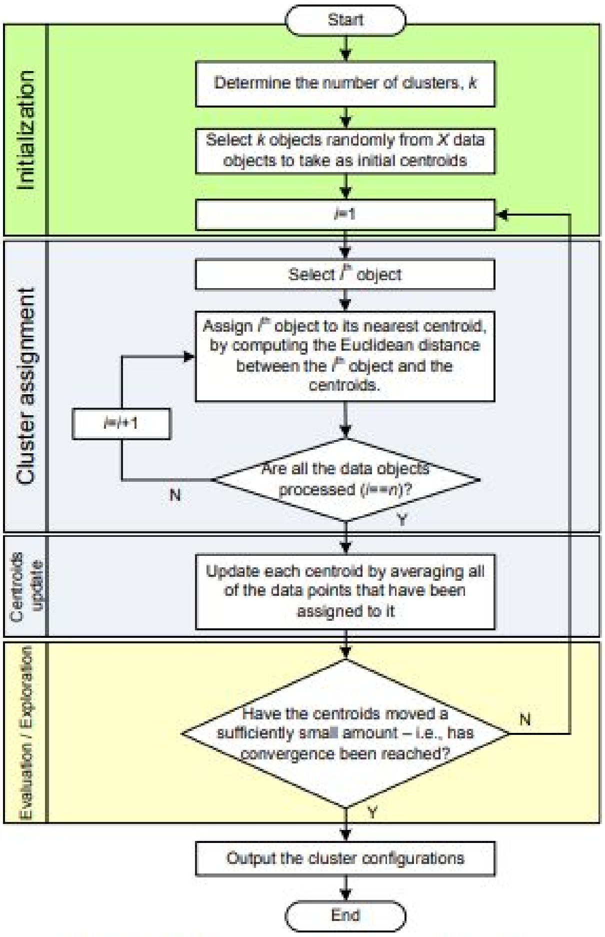
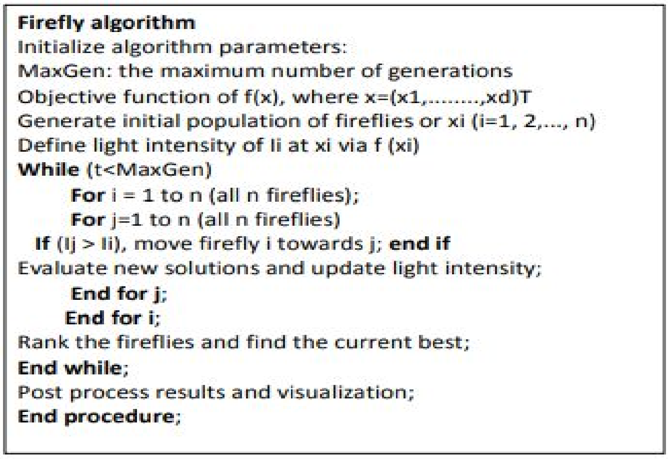

Start
Initialization
Determine the number of clusters, k
Select k objects randomly from X data objects to take as initial centroids
i=1
Cluster assignment
Select ith object
Assign ith object to its nearest centroid, by computing the Euclidean distance between the ith object and the centroids.
i=i+1
Are all the data objects processed (i==n)?
N
Y
Centroids update
Update each centroid by averaging all of the data points that have been assigned to it
Evaluation / Exploration
Have the centroids moved a sufficiently small amount – i.e., has convergence been reached?
N
Y
Output the cluster configurations
End

Figure 1: Flowchart of ordinary k-means clustering algorithm

As it can be seen in Fig. 1, the K-means algorithm can be divided in several stages, namely initialization, cluster assignment and centroids update, and exploration and evaluation. We advocate that these operation stages form the basic constructs for the extended algorithm incorporating the optimization facility. The three constructs are displayed in different color in the flowcharts for distinguishing their exclusive functions.

Firefly Algorithm: Most fireflies produce short and rhythmic flashes and have different flashing behavior. Fireflies use these flashes for communication and to attract the potential prey. YANG used this behavior of fireflies and introduced Firefly Algorithm in 2008. In Firefly algorithm, there are three idealized rules: 1) All fireflies are unisex. So, one firefly will be attracted to other fireflies regardless of their sex; 2) Attractiveness is proportional to their brightness. Thus, for any two flashing fireflies, the less brighter one will move towards the brighter one. The attractiveness is proportional to the brightness and they both decrease as their distance increases. If there is no brighter one than a particular firefly, it will move randomly; 3) The brightness of a firefly is determined by the landscape of the objective function. For a maximization problem, the brightness can simply be proportional to the value of the objective function. The pseudo code of these three rules can be shown as Fig. 2



```
Firefly algorithm
Initialize algorithm parameters:
MaxGen: the maximum number of generations
Objective function of f(x), where x=(x1,........,xd)T
Generate initial population of fireflies or xi (i=1, 2,..., n)
Define light intensity of Ii at xi via f (xi)
While (t<MaxGen)
        For i = 1 to n (all n fireflies);
        For j=1 to n (all n fireflies)
  If (Ij > Ii), move firefly i towards j; end if
Evaluate new solutions and update light intensity;
        End for j;
       End for i;
Rank the fireflies and find the current best;
End while;
Post process results and visualization;
End procedure;
```

Fig 2. Pseudo firefly algo code

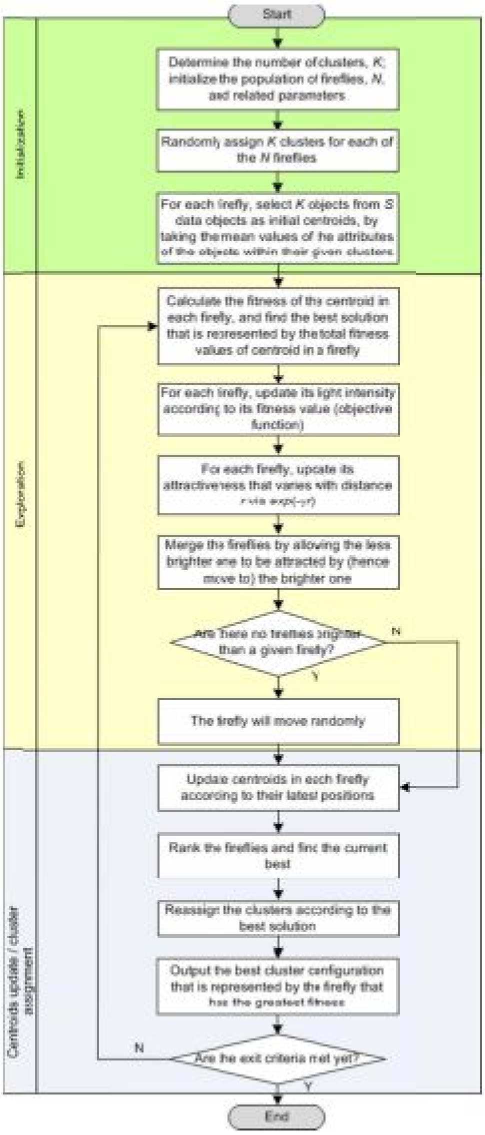

Start
Initialization
Determine the number of clusters, K; initialize the population of fireflies, N, and related parameters
Randomly assign K clusters for each of the N fireflies
For each firefly, select K objects from S data objects as initial centroids, by taking the mean values of the attributes of the objects within their given clusters
Exploration
Calculate the fitness of the centroid in each firefly, and find the best solution that is represented by the total fitness values of centroid in a firefly
For each firefly, update its light intensity according to its fitness value (objective function)
For each firefly, update its attractiveness that varies with distance r via exp(-γr)
Merge the fireflies by allowing the less brighter one to be attracted by (hence move to) the brighter one
Are there no fireflies brighter than a given firefly?
N
Y
The firefly will move randomly
Centroids update / cluster assignment
Update centroids in each firefly according to their latest positions
Rank the fireflies and find the current best
Reassign the clusters according to the best solution
Output the best cluster configuration that is represented by the firefly that has the greatest fitness
Are the exit criteria met yet?
N
Y
End

Fig 3: Firefly Algorithm workflow

## IV. Results

Here we come across how the centroid is initially chosen and how the k-means clustering with the firefly optimization results in optimized centroids. Fig 1 shows the initial data representation and Fig 2 shows the Initial K-means centroid representation and finally the Fig 3 shows the Centroid allocated after the firefly algorithm optimized the given centroids after multiple iterations.

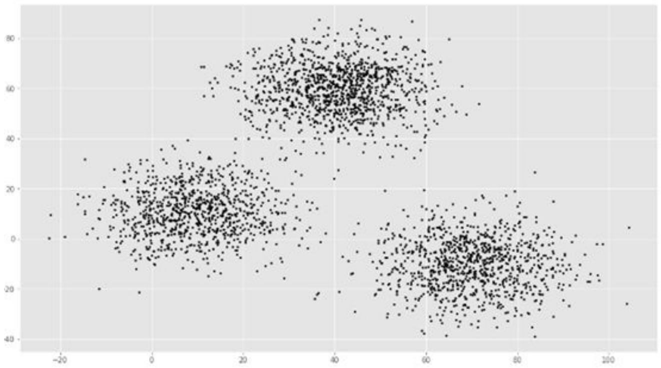


Fig 1: initial Data representation

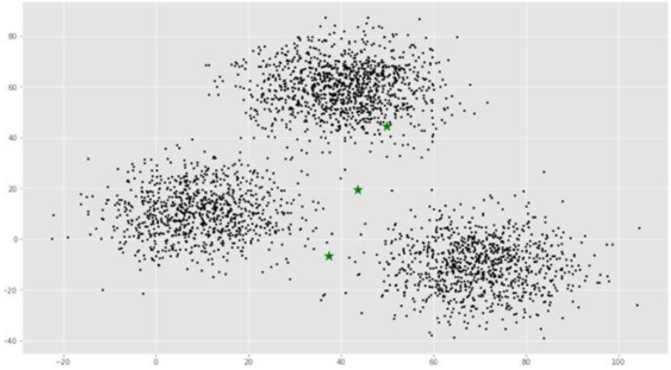


Fig 2: The initial centroid Representation

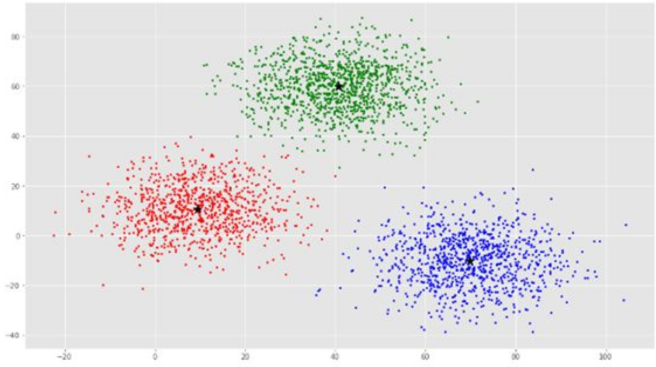


Fig 3: the final Centroids Representation

The value of the objective function is considered to vary across number of generations for which the algorithm is run and the objective function gives the indication about the quality of the clustering results.

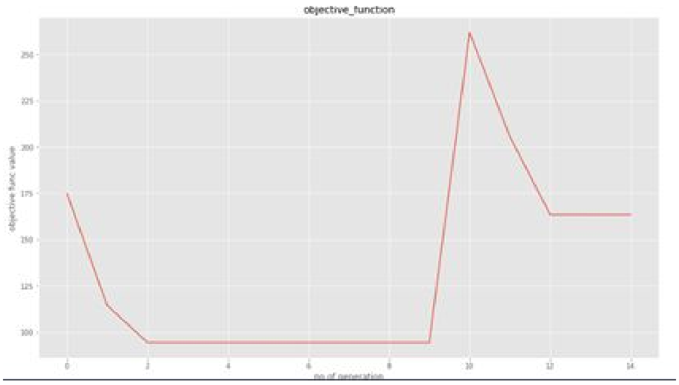


Fig 4: Objective function value variation

## V. DISCUSSION

In this paper, a new hybridizes method based on firefly algorithm and k-means clustering method proposed to cluster data. In the proposed method, at first we used firefly algorithm to find optimal cluster centers and then initialized the k-means algorithm with this centers to refine the centers. This method applies to xclara dataset. Experimental results for optimizing fitness function related to intra-cluster distance showed that the proposed obtained results that are relatively stable in different performance. Generally, experimental results showed that the proposed algorithm had better efficiency K-means.

## VI. REFERENCES

[1]Hassanzadeh, Tahereh & Meybodi, Mohammad. (2012). A New Hybrid Approach for Data Clustering using Firefly Algorithm and K-means. AISP 2012 - 16th CSI International Symposium on Artificial Intelligence and Signal Processing. 10.1109/AISP.2012.6313708. .

[2] Tang, Rui & Fong, Simon & Yang, Xin-She & Deb, Suash. (2012). Integrating nature-inspired optimization algorithms to K-means clustering. 10.1109/ICDIM.2012.6360145.

[3] Yang, Xin-She. (2020). The Firefly Algorithm: An Introduction.

[4] C. Pizzuti and D. Talia, ‘‘P-AutoClass: scalable parallel clustering for mining large data sets’’, in

IEEE transaction on Knowledge and data engineering, Vol. 15, pp. 629-641, May 2003.

[5] K. C. Wong and G. C. L. Li, “Simultaneous Pattern and Data Clustering for Pattern Cluster Analysis’’, in IEEE Transaction on Knowledge and Data Engineering, Vol. 20, pp. 911-923, Los Angeles, USA, June 2008.

[6] J. Marr, ‘‘Comparison Of Several Clustering Algorithms for Data Rate Compression of LPC Parameters’’, in IEEE International Conference on Acoustics Speech, and Signal Processing, Vol. 6, pp. 964-966, January 2003.

[7] X. L. Yang, Q. Song and W. B. Zhang, ‘‘Kernel-based Deterministic Annealing Algorithm For Data Clustering’’, in IEEE Proceedings on Vision, Image and Signal Processing, Vol. 153, pp. 557-568, March 2007.

[8] X. S. Yang, “Nature-Inspired etaheuristic Algorithms”. Luniver Press, 2008.

[9] X. S. Yang, “Firefly algorithm‹ stochastic Test Functions and Design optimization”.Int. J. bio-inspired computation .2010.

[10] X. S. Yang, “Firefly algorithm for multimodal optimization.”In:StochasticAlgorithms: foundations and applications ‹SAGA ‹lecture notes in computer sciences, pp. 169-178, 2009