A New Hybrid Approach for Data Clustering using Firefly Algorithm and K-means

Tahereh Hassanzadeh
Department of Computer Engineering and IT
Azad University, Branch of Qazvin
Qazvin, IRAN
t.hassanzadeh@qiau.ac.ir

Mohammad Reza Meybodi
Department of Computer Engineering and IT
Amirkabir University of Technology
Tehran, IRAN
mmeybodi@aut.ac.ir

Abstract— **Data clustering is a common technique for data analysis and is used in many 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 limitation 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 called K-FA. The experimental results showed the accuracy and capability of proposed algorithm to data clustering.**

Keywords-component; firefly algorithm; clustering; 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 [1], pattern recognition [2], data compression [3], machine learning [4], 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 [5], 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 k-means algorithm is hybridizing it with efficient optimization method. There is different optimization algorithm like Particle Swarm Optimization (PSO) [6], Ant Colony Optimization (ACO) [7], Artificial Fish Swarm Algorithm (AFSA) [8] and Bee Colony [9].

The Firefly algorithm was recently introduced by XIN-SHE YANG in Cambridge University [10]. 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 [11] and multimodal optimization problems [12] 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 [13], 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 Firefly Algorithm. Section 3 gives a detailed description of the k-means algorithm. The proposed algorithm is introduced in section 4. The experimental results are discussed in Section 5, and Section 6 concludes the paper.

II. FIREFLY ALGORITHM

Most of fireflies produced short and rhythmic flashes and have different flashing behavior. Fireflies use these flashes for communication and attracting the potential prey. YANG used this behavior of fireflies and introduced Firefly Algorithm in 2008 [10].

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 [10]. The pseudo code of these three rules can be shown as Fig. 1.

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;

Figure1: the pseudo code of firefly algorithm.

In the firefly algorithm there are two important issues including variation of light intensity and the formulation of the attractiveness. For simplicity, it's assumed that the attractiveness of a firefly is determined by its brightness which associated with the objective function of the optimization problem. Since a firefly's attractiveness is proportional to the light intensity seen by adjacent fireflies, we can now formulate the attractiveness of a firefly by:

$$\beta(r) = \beta_0 e^{-\gamma r^2} \qquad (1)$$

where, β_0 is the attractiveness at $r = 0$ and γ is the light absorption coefficient at the source. It should be noted that the $r_{i,j}$ which is described by equation 2, is the Cartesian distance between any two fireflies i and j at x_i and x_j, where, x_i and x_j are the spatial coordinate of the fireflies i and j, respectively.

$$r_{i,j} = \left\| x_i - x_j \right\| = \sqrt{\sum_{k=1}^{d} \left(x_{i,k} - x_{j,k} \right)^2} \qquad (2)$$

The movement of a Firefly i, which is attracted to another more attractive Firefly j is determined by:

$$X_i = x_i + \beta_0 e^{-\gamma r_{ij}^2} \left(x_j - x_i \right) + \alpha \left(rand - \frac{1}{2} \right) \qquad (3)$$

where, the second term is the attraction while the third term is randomization including randomization parameter α and the random number generator *rand* which its numbers are uniformly distributed in interval [0, 1].

For the most cases of implementations, $\beta_0 = 1$ 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 and how the FA behaves. In the most applications, it typically varies from 0.01 to 100.

III. K-MEANS CLUSTERING

The K-means algorithm groups D-dimensional data vectors into a predefined number of clusters on the basis of the Euclidean distance as the similarity criteria. Euclidean distances among data vectors are minimum for data vectors within a cluster as compared with distances to other data vectors in different clusters. Vectors of the same cluster are associated with one centroid vector, which represents the center of that cluster and is the mean of the data vectors that belong together. The standard K-means algorithm is summarized as follows:

1. Initialize k cluster center randomly.
2. Repeat
   a. For each data vector, allocate the vector to the cluster with the less Euclidean distance to centroid vector, where the distance to the centroid is determined using following equation,

$$Dis(X_p, Z_j) = \sqrt{\sum_{i=1}^{d} (X_{pi} - Z_{ji})} \qquad (4)$$

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.

b. Refine the cluster centroid vectors, using

$$Z_j = \frac{1}{n_j} \left[\sum_{\forall Xp \in Cj} X_p \right] \qquad (5)$$

where, n_j is the number of data vectors in cluster j and C_j is the subset of data vectors that form cluster j, until a stopping criterion is satisfied.

The K-means clustering algorithm will be terminates when any one of the following criteria is satisfied: when the maximum number of iterations has been exceeded, when there is little change in the centroid vectors over a number of iterations, or when there are no cluster membership changes. In the proposed hybrid algorithm of this research, the algorithm terminates when there is little change in the centroid vectors over a number of iterations.

IV. PROPOSED CLUSTERING APPROACH

In this section at first to improve standard firefly algorithm we proposed a modified firefly algorithm and then use this new algorithm to hybrid with k-mean algorithm to data cluster.

A. Modified firefly algorithm (MFA)

In the standard Firefly Algorithm (FA), in each iteration the brighter firefly (local optima) exerts its influence over other fireflies and attracts them towards itself in maximization optimization problems. In fact, in the standard FA, fireflies move regardless of the global optima and it decrease the ability of the firefly algorithm to find global best. In this paper, to eliminate weaknesses of FA and improve the collective movement of fireflies, we propose a modify firefly algorithm (MFA).

In the proposed firefly algorithm, we use global optima in firefly's movement. Global optimum is related to optimization problem and it can be a firefly that has the maximum or minimum value. And the global optima will be update in any iteration of algorithm. In the proposed algorithm, when a firefly compared with other firefly 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 affect in their movement.in the MFA, when a firefly compare with correspond firefly, if the correspond firefly be brighter, the compared firefly will move toward correspond firefly, considered by global optima.

In the MFA, we use Cartesian distance to compute the distance of fireflies to global optima same as the standard FA as it shows in equation 6,

$$r_{i,best} = \sqrt{(x_i - x_{gbest})^2 + (y_i - y_{gbest})^2} \quad (6)$$

But for movement of the firefly, equation (3) is replaced by:

$$x_i = x_i + \left(\beta_0 e^{-\gamma r_{ij}^2}(x_j - x_i) + \beta_0 e^{-\gamma r_{i,gbest}^2}(x_{gbest} - x_i)\right) + \alpha(rand - 1/2) \quad (7)$$

g_{best} is global optimal and x_{gbes} is the coordinate of global optima.

B. Proposed clustering algorithm

As it mentioned before, k-means clustering is the one of the famous and simple method for data clustering. In the k-means algorithm, at first k random cluster center defined and then each data vector will be assign to each cluster based on Euclidean distance. Each data vector compared with k center of clusters and then allocates to closer cluster and then refined the center of cluster by using equation 5. In the k-means clustering, k center of cluster initial randomly because of this the algorithm may trapped in local optima. In this paper To improve and increase the accuracy of k-means algorithm we initialize the k-means algorithm with optimal centers, which calculated by firefly algorithm.

$$\left[Z_{1,1}, Z_{1,2}, .., Z_{1,d}, Z_{2,1}, Z_{2,2}, .., Z_{2,d}, .., Z_{K,1}, Z_{K,2}, .., Z_{K,d}\right]$$

Figure2: Structure of a firefly position in clustering problem space.

In the proposed method, the clustering has two stage at first we initialize fireflies with random values. as it shows in figure 2, Since data is D-dimensional and there are K clusters, so each firefly has K×D dimension. The objective function, which must be minimum, is Euclidean distance. The mechanism of firefly algorithm must do till predefine iteration. In the second stage, the k-means will initialize with the position of best firefly. The k-means clustering refine the centers. The proposed hydride clustering algorithm can be summarized as the pseudo code show in Fig 3.

```
Initialize fireflies with random K*D centers
   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 equation 4,
      If (ij>ii)
Move firefly I toward j based on equation 7 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
Rank the fireflies and find global best and extract the position
of global best
   Repeat
Initialize the k-means center with position of global best
Allocate each vector to a cluster by equation 4,
Refined the clusters by equation 5
   Do until predefined iteration.
```

Figure3: pseudo code of proposed algorithm.

V. EXPERIMENTAL RESULTS

Experiments have been performed on five data sets including Iris, WDBC, Sonar, Glass and Wine that were selected from standard data set UCI [14]. Which the characteristics of each of them are described in the following:

Iris (fisher's iris plants database): This data set is according to the Iris flowers recognition that has three different classes

and each class consists of 50 samples. Every sample has four attributes.

WDBC (Wisconsin diagnostic breast cancer) : this data set is about breast cancer that is collected at the university of Wisconsin. That has two different classes including 357 and 212 samples. In this data set, each sample has 30 features.

Sonar: this data set is about sonar signals of submarine that totally has 208 samples. In this data set, Sonar signals divided in tow classes including 111 and 97 samples with 60 features.

Glass (glass identification database): this data set is about several types of glass that has totally 214 samples in 6 classes. These classes are about building_windows_float_processed, vehicle_windows_float_processe, containers‹, ableware, building_windows_non_float_processed and headlamps and each data has 9 attributes.

Wine (wine recognition data): This data set is regarding to drinks recognition that totally has 178 samples classified into three different classes including 59, 71 and 48 samples, respectively. In this data set, each sample has 13 attributes.

In the first step of proposed method, to improve the final results we initial fireflies with k vector of dataset, which select randomly between the vectors of data set. Because of this the initial points of centers for fireflies will be among the data and doesn't be outside of the data areas. In the proposed algorithm, the population size set as 150 with 50 inner iteration. Also, we set $\gamma = 1$, $\alpha = 0.7$ and $\beta_0 = 1$.

In the second step we initialized the k-means clustering with optimal cluster center, that calculated by firefly algorithm. K-means refine the centers till the little change in the centroid vectors over a number of iterations happen. We set this value 0.1.

In PSO c1 and c2 values are considered 2 and inertia weight on each attempt is obtained by W=rand/2 + 0.5 [15]. Experiments were repeated 30 times.

For evaluation of obtained results we use intra-cluster distance, calculated by equation 4 and clustering error, calculated by equation 8.

$$Err = \left(\frac{1}{N} \sum_{i=1}^{N} (\text{if } (\text{Class}(i) = \text{Cluster}(i)) \text{ then 0 else 1}) \right) \times 100 \quad (8)$$

where, N is total number of samples. Class (i), is the class of data vector of i, and cluster (i), is the number of the cluster of the data vector. With this equation we can calculate the clustering error.

The best, mean and standard division of Intra-cluster distance for PSO, K-means, K-PSO and the proposed algorithm of KFA for different data set including Iris, WDBC, Sonar, Glass and Wine are shown in tables 1 to 5. As it shown in tables, the proposed method could decrease the intra-cluster distance of each cluster in all cases and its cause the proper initialization of the k-means algorithm.

Table 1: COMPARISON OF INTRA-CLUSTER DISTANCE BETWEEN DIFFERENT METHODS FOR IRIS DATA SET.

Algorithm	Best	Mean	Std.Dev
k-means	97.32	102.57	11.34
PSO	97.10	102.26	5.81
KPSO	96.78	99.61	7.21
KFA	**96.13**	103.87	**2.45**

Table 2: COMPARISON OF INTRA-CLUSTER DISTANCE BETWEEN DIFFERENT METHODS FOR WDBC DATA SET.

Algorithm	Best	Mean	Std.Dev
k-means	152647.25	179794.25	55222.17
PSO	149537.73	49830.87	364.73
KPSO	149480.93	149594.05	198.31
KFA	**149450.33**	**149590.21**	**197.31**

Table 3: COMPARISON OF INTRA-CLUSTER DISTANCE BETWEEN DIFFERENT METHODS FOR SONAR DATA SET.

Algorithm	Best	Mean	Std.Dev
k-means	234.77	235.06	0.15
PSO	271.83	276.68	3.79
KPSO	234.65	234.92	0.22
KFA	**229.35**	**231.36**	2.94

Table4: COMPARISON OF INTRA-CLUSTER DISTANCE BETWEEN DIFFERENT METHODS FOR GLASS DATA SET.

Algorithm	Best	Mean	Std.Dev
k-means	213.42	241.03	25.32
PSO	230.54	258.02	12.24
KPSO	212.03	233.28	14.05
KFA	**210.51**	**221.87**	**10.5**

Table 5: COMPARISON OF INTRA-CLUSTER DISTANCE BETWEEN DIFFERENT METHODS FOR WINE DATA SET.

Algorithm	Best	Mean	Std.Dev
k-means	16555.68	17662.73	1878.07
PSO	16307.16	16320.67	9.53
KPSO	16298.92	16307.58	7.23
KFA	**16284.01**	**16327.53**	10.10

Table 6: COMPARISON OF CLUSTERING ERROR BETWEEN DIFFERENT METHODS FOR ALL DATA SET.

Data Set	K-means	PSO	KPSO	KFA
Iris	16.05±10.10	10.64±4.50	12.58±7.67	**7.33±1.10**
WDBC	19.12±9.22	13.18±1.80e-15	13.18±1.81e-15	**12.42**±1.02
Sonar	44.95±0.97	46.60±0.42	44.98±0.84	**35.57**±5.55
Glass	48.30±3.14	48.72±1.34	47.80±1.98	**45.54**±3.37
Wine	34.38±6.08	28.74±0.39	28.59±0.47	**28.15**±0.95

One of the most important characteristic of a clustering method is the ability of it in decreasing clustering error. An appropriate clustering method, beside of intra-cluster error should be able to reduce the clustering error and assign the data vector to accurate cluster. In the table 6 we have shown the clustering error of PSO, KPSO, K-means and proposed method on the Iris, WDBC, Sonar, Glass and Wine data sets. As it shown, the obtained results proved the accuracy and efficiency of proposed method. This method could decrease the clustering error in all cases in compared of other methods.

VI. CONCLUSION

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 5 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 than PSO, KPSO and K-means.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] Kaufman, L., & Rousseeuw, P. J. (1990). Finding groups in data: An introduction to cluster analysis. New York: John Wiley & Sons.

[6] J. Kennedy, R. C. Eberhart, "Particle Swarm Optimization". In: IEEE International Conference on Neural network، pp. 1942—194, 1995.

[7] M. Darigo, M. Birattari, T.Stutzle, "Ant Colony Optimization". In: IEEE Computational Intelligent Magazine، Vol. 1، pp. 28--39, 2006.

[8] L. X. Li, Z. J. Shao, J. X. Qian, "An Optimizing Method Based on Autonomous Animate: Fish Swarm Algorithm." In: Proceeding of System Engineering Theory and Practice، Vol. 11، pp. 32—38, 2002.

[9] D.T. Pham, A. Ghanbarzadeh, E. Koc, S. Otri, S. Rahim, M. Zaidi, " The Bees Algorithm - A Novel Tool for Complex Optimisation Problems". Proceedings of IPROMS 2006 Conference, pp. 454-461, 2006.

[10] X. S. Yang, "Nature-Inspired etaheuristic Algorithms". Luniver Press, 2008.

[11] X. S. Yang, "Firefly algorithm، stochastic Test Functions and Design ptimization".Int. J. bio-inspired computation .2010.

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

[13] D. W. van der Merwe and A. P. Engelbrecht, "Data Clustering Using Particle Swarm Optimization", in the 2003 Congress on Evolutionary Computation, Vol. 1, pp. 215-220, December 2003.

[14] http://archive.ics.uci.edu/ml/

[15] Y. T. Kao, E. Zahara and I. W. Kao, "A Hibridized Approach to Data Clustering", in Elsevier Journal on Expert System with Applications, pp. 1754-1762, 2008.