Improving K-means clustering based on firefly algorithm

Amal Mahmoud Naji Al Radhwani\textsuperscript{1} and Zakariya Yahya Algamal\textsuperscript{2}
\textsuperscript{1}Department of Statistics and Informatics, University of Mosul, Mosul, Iraq
\textsuperscript{2}Department of Statistics and Informatics, University of Mosul, Mosul, Iraq
Email\textsuperscript{1}: omeeramal@gmail.com
Email\textsuperscript{2}: zakariya.algamal@uomosul.edu.iq

Abstract

Data clustering determines a group of patterns in a dataset which are homogeneous in nature. The objective is to develop an automatic algorithm which can accurately classify an unlevelled dataset into groups. The K-means method is the most fundamental partitioned clustering concept. However, the performance of K-means method is fully depending on determining the number of clusters, K, and determining the optimal centroid for performing the clustering process. In this paper, an adaptive firefly optimization algorithm, which is a nature-inspired algorithm, is employed to improve the K-means clustering. The experimental results of clustering two real datasets show that the proposed method is able to effectively outperform other alternatives methods.

Keywords: K-means, firefly algorithm; clustering.

1. Introduction

Data clustering is the process of grouping together similar multi-dimensional data vectors into a number of clusters. Clustering algorithms have been applied to a wide range of problems, including exploratory data analysis, data mining [1], image segmentation [2] and mathematical programming [3]. Clustering techniques have been successfully used to address the scalability problem of machine learning and data mining algorithms, where prior to, and during training, training data is clustered, and samples from these clusters are selected for training thereby reducing the computational complexity of training process, and even improving generalization performance[4, 5]. Clustering algorithms can be grouped into two main classes of algorithms, namely hierarchical and partitional. The k-means clustering method [6] is one of the most commonly used partitional methods. However, the results of k-means solving the clustering problems highly depend on the initial solution and it is easy to fall into local optimal solutions. For overcoming this problem, many scholars began to solve the clustering problem using meta-heuristic algorithms. Nikham et al. have proposed an efficient hybrid evolutionary algorithm based on combining ACO and SA (simulated annealing algorithm, 1989 [7] for clustering problem [8, 9]. In 1991, A.Colorni et al. have presented an ant colony optimization (ACO) algorithm based on the behavior of ants seeking a path between their colony and a source of food. Then P.S. Shelokar and Y. Kao solved the clustering problem using the ACO algorithm [10, 11]. J. Kennedy and R.C Eberhart have proposed a particle swarm optimization (PSO) algorithm which simulates the movements of organisms in bird flock or fish school in 1995 [12]. The algorithm also has been adopted to solve this problem by M. Omran and V.D. Merwe [13, 14]. Kao et al. have presented a hybrid approach according to combination of the k-means algorithm, Nelder-Mead simplex search and PSO for clustering analysis [15]. Kevin et al. have used an evolutionary-based rough clustering algorithm for the clustering problem [16].
Although K-mean functions faster than many other clustering algorithms, it suffers from two major problems, i.e. high sensitivity to initialization and easy convergence to local optima. The major challenge is regarding the determination of the optimal centroid as most of the optimization algorithm used for determining the optimal centroids converge to the local optimum point and hence, an effective optimization algorithm is required.

In this paper, an adaptive firefly algorithm, which is a nature-inspired continuous algorithm, is proposed to improve the K-means clustering by improve the initial solution and to avoid trap into the local optimal. The superiority of the proposed method is testing using several real datasets.

2. The mathematical model of clustering
2.1. Mathematical definition of data clustering

The goal of data clustering is grouping data into a number of clusters, clustering problems can be seen in practice frequently. In this subsection, a mathematical definition of data clustering is presented. In order to explain the definition clearly, we supposed that there exists a data set \( D = \{d_1, d_2, \ldots, d_n\} \). And each individual \( d_i \) (\( 1, 2, \ldots, n \)) has many features. If the dimension is \( m \), each individual can be shown as \( d_i = (1, 1_2, \ldots, 1_m) \). Data clustering is a process which can classify the given data set \( D \) into a certain numbers of clusters \( G_1, G_2, \ldots, G_k \) (assume \( k \) clusters) based on the similarity of individuals. And \( G_1, G_2, \ldots, G_k \) should satisfy the following formulas:

1) \( G_i \neq \emptyset, \ i = 1, 2, \ldots, k \).
2) \( G_i \cap G_j = \emptyset, \ i, j = 1, 2, \ldots, k, i \neq j \).

2.2. The principle of data clustering

In the clustering process, if the given data set \( D \) should be divided into \( k \) clusters \((G_1, G_2, \ldots, G_k)\), and each cluster must have one center \( c_j \) (\( j = 1, 2, \ldots, k \)). It is supposed that \( C = (c_1, c_2, \ldots, c_k) \) are the centers of \((G_1, G_2, \ldots, G_k)\).Where \( c_j \) is the center of subset \( G_j \).

The main idea of clustering is to define \( K \) centers, one for each cluster. These centers should be placed in a crafty way, because different location will causes different result. Therefore, the better choice is to place then as far away from each other as possible. In this paper, we will use Euclidean metric as a distance metric. The expression is given as follows:

\[
 d \left( d_i, c_j \right) = \sqrt{\sum_{k=1}^{m} (d_{ik} - c_{jk})^2} \tag{1}
\]

Where \( d_i \) (\( i = 1, 2, \ldots, n \)) is an individual in the given data set \( D \), \( m \) is the number of individual features; \( c_j \) (\( j = 1, 2, \ldots, k \)) is the center of \( j \)th subset. Because individual has \( m \) features, \( c_j \) can be presented by \((c_{j,1}, c_{j,2}, \ldots, c_{j,m})\). In order to confirm which subset \( d_i \) belongs to, the distances between \( d_i \) and \( c_j \) (\( j = 1, 2, \ldots, k \)) should be calculated via \( (1) \). If the distance between \( d_i \) and \( c_{best} \) (\( best = 1, 2, \ldots, k \)) is smaller than the distances between \( d_i \) and other centers (except \( c_{best} \), we can make the decision that \( d_i \) should belongs to \( G_{best} \). For examples, if the value of \( d \left( d_i, c_j \right) < d \left( d_i, c_{j} \right), \ (j \neq 2) \), we can draw the conclusion that \( d_i \) should be distributed to \( G_1 \).
3. The proposed algorithm

Nature has been an inspiration for the introduction of many meta-heuristic algorithms. Swarm intelligence is an important tool for solving many complex problems in scientific research. Swarm intelligence algorithms have been widely studied and successfully applied to a variety of complex optimization problems. The firefly algorithm (FFA), is one of the recent novel swarm intelligence methods and the most powerful optimization algorithms, which was developed by Yang [17].

Firefly algorithm has been proved to be a good performance and the effectiveness for solving various optimization problems [18]. The firefly algorithm has been inspired by the simulation of the social behavior of fireflies on the basis of the flashing lights or the flash attractiveness. By representing the advantage of some flashing characteristics of fireflies and how fireflies interact with flashing lights, the firefly flash is a signal system which used to attract another firefly [19].

Mathematically, assume that there are \( n_f \) of fireflies in the swarm (populations size) are randomly distributed in the D-dimensional search space. During the evolutionary process, each firefly has a position vector denoted as \( \{x_{i1}, x_{i2}, \ldots, x_{id}\} \), where \( i = 1, 2, \ldots, n_f \) and \( d \in D \) is the dimensionality of the solutions. The distance between any two fireflies \( i \) and \( j \), at positions and in the search space, respectively, is the Cartesian distance which can be calculated using the following equation

\[
 r_{ij} = \|x_i - x_j\| = \sqrt{\sum_{d=1}^{D} (x_{id} - x_{jd})^2}.
\] (2)

Each firefly has its light intensity or brightness. The brightness value is used to evaluate the goodness of firefly, which is affected by the landscape of the optimization problem [20-23]. The brightness of firefly at a particular or current position \( x \) can be denoted by the objective function value as follows:

\[
 I(x_i) = f(x_i)
\] (3)

The light intensity of the firefly is directly proportional to its brightness and is related to objective values. In comparing the two fireflies, both fireflies are attracted, the firefly which has a lower light intensity is attracted toward the other firefly with the higher light intensity.

The light intensity of a firefly depends on the intensity \( I_0 \) of light emitted by a firefly and the distance \( r_{ij} \) between two fireflies. Light intensity \( I(r) \) can be described by a monotonically decreasing function of the \( r_{ij} \) which can be formulated as follows:

\[
 I(r) = I_0 e^{-\gamma \frac{r}{r_{ij}}},
\] (4)

where \( \gamma \) is used to control the decrease of the light intensity or brightness and can be taken as a constant.

Each firefly has its distinctive attractiveness which indicates how powerful it attracts other members in the swarm. Attractiveness, \( \beta \), is relative, which means that it must be judged by others, and therefore varies with the distance \( r_{ij} \). As mentioned earlier, the brightness decreases with the distance from the source and the light is also absorbed by the air, therefore the
attractiveness must be allowed to vary with differing degrees of absorption [24]. Thus, the main form of the attractiveness of a firefly is defined as the following equation:

$$\beta(r) = \beta_0 e^{-\gamma r^2}, \hspace{1cm} (5)$$

where $\beta(r)$ represents attractiveness function of a firefly at a distance, $r$, and $\beta_0$ denotes the initial attractiveness of a firefly at distance $r = 0$ and it can be constant. For implementation usually $\beta_0$ set to be 1 for most problems.

The fireflies will try to move to the best position. This means that the lower light intensity one will be attracted by the brighter one. The location updates for each pair of fireflies $i$ and $j$.

For each firefly $x_i$, it is compared to other all fireflies $x_j, j = 1, 2, ..., n_f$. If firefly $j$ at position $x_j$ is brighter than firefly $i$, then $x_i$ will move towards $x_j$ by the attraction. The movement is defined as:

$$x_{id}^{(i+1)} = x_{id}^{(i)} + \beta_0 e^{-\gamma r^2} \left( x_{jd}^{(i)} - x_{id}^{(i)} \right) + \alpha_t e_{id}^{(i)}, \hspace{1cm} (6)$$

where $\alpha_t$ is the randomization parameter, $\gamma$ is an absorption coefficient which controls the decrease of the light intensity, and $e_{id}^{(i)} = (\text{rand} - 0.5)$, where $\text{rand}$ is a random number from uniform distribution with $[0, 1]$.

The effect of this random movement depends on the parameter $\alpha_t$. If $\alpha_t$ is chosen to be large then the solution $x_i$ will randomly jump away from the neighborhood and explore the solution space and if it is very small then its jump will be in the neighborhood and also may become negligible compared to the movement towards brighter fireflies.

To enhance the FFA, a new mathematical form is proposed for determining the value of $\alpha_t$ as

$$\alpha_t = \frac{\alpha_t^{\text{max}}}{t^2}. \hspace{1cm} (7)$$

Equation (7) decreases quickly and the random movement of the firefly will almost vanish within small number of iterations [25].

4. The performance evaluation function of data clustering

For explaining the evaluation process explicitly. We suppose that given data set D should be divided into $k$ subset. And the dimension of individual of data set D is $m$. In order to optimize the coordinates of centers of $k$ subset, it is easily to find that the dimension of solution should be $k \times m$. The individual in the population can be described as $s = (c_1, c_2, ..., c_k)$. A great classification should minimize the sum of distances value. So, we should try to minimize the distance between individual $d_i$ and the center $c_j$ of subset it belongs to. Finally, the proposed algorithm aims at minimizing the objective function, which can be expressed as following:

$$f (D,C) = \sum_{i=1}^{n_f} \min_{k=1,2,...,K} \|d_i - c_k\|$$

$$\hspace{1cm} \left(8\right)$$
Where $D = (d_1, d_2, ..., d_n)$ is the given data set, $C = (c_1, c_2, ..., c_k)$ is the centers of subsets $(G_1, G_2, ..., G_k)$.

5. Real data results

To test our proposed algorithm, two real datasets are used. The comparison is conducted between our proposed algorithm, PFFA, the original FFA, and the standard K-means algorithm. The parameter configurations for our proposed method are presented as follows: "The number of fireflies is $n_f = 50$, $\beta_0 = 1$, $\gamma = 0.2$, $\alpha = 0.1$, $\alpha_{\text{max}} = 2$, and the maximum number of iterations is $t_{\text{max}} = 150$.

The two datasets are explained as: Iris data ($N = 150, d = 4, k = 3$), this data set with 150 random samples of flowers from the iris species setosa, versicolor, and virginica. From each species there are 50 observations for sepal length, sepal width, petal length, and petal width in cm. This data set was used by Fisher (1936) in this initiation of the linear-discriminant-function technique [26, 27]. Wisconsin breast cancer ( $N = 683, d = 9, k = 2$), which consists of 683 objects characterized by nine features: clump thickness, cell size uniformity, cell shape uniformity, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses. There are two categories in the data: malignant (444 objects) and benign (239 objects) [26, 27].

The comparison of algorithm for data set Iris and Wisconsin breast cancer is listed in Tables 1 and 2, respectively. Table 1 shows that the best value, worst value, mean value and standard deviation of our proposed algorithm, PFFA, are all better than the original FFA algorithm. Convergence curves of two algorithms for data set Iris are shown in Figure 1. The curves in Figure 1 show that PFFA has faster convergence speed. And the convergence curve of PFFA is smoother.

Related to Wisconsin breast cancer dataset, the comparison of algorithm is listed in Table 2. And the convergence curves of algorithms are shown in Figure 2. The best value, worst value, mean value and the standard deviation of PFFA are the best comparing to FFA. Convergence curves for dataset Wisconsin breast cancer shown in Figure 3 clearly show that PFFA has a faster convergence speed”.

| Table 1: Results obtained by the algorithms for 30 different runs on Iris dataset |
|-----------------|-----------------|-----------------|
| PFFA            | FFA             | K-means         |
| Best            | 94.711          | 96.842          | 97.192          |
| Worst           | 94.711          | 96.842          | 122.367         |
| Mean            | 94.711          | 96.842          | 100.963         |
| STD             | $8.6 \times 10^{-12}$ | $9.1 \times 10^{-12}$ | $6.378$         |

| Table 2: Results obtained by the algorithms for 30 different runs on Wisconsin breast cancer dataset |
|-----------------|-----------------|-----------------|
| PFFA            | FFA             | K-means         |
| Best            | 2917.651        | 2936.947        | 2977.973        |
| Worst           | 2917.651        | 2936.947        | 2989.551        |
| Mean            | 2917.651        | 2936.947        | 2983.144        |
| STD             | $7.3 \times 10^{-12}$ | $8.6 \times 10^{-12}$ | $5.023$         |
6. Conclusion
This paper proposed an adaptive procedure for improving the firefly algorithm for clustering the data. The PFFA algorithm computes the optimal centroid for performing the data clustering that is based on the minimum fitness function. The clustered data are the optimally clustered data and it provide all the valuable information for the decision-making process. Performance analysis carried out using the two datasets prove that the PFFA outperforms the existing FFA and K-means by attaining a minimum value of the objective function.

References
Uncategorized References
[1] Ng R T and Han J 1994 Efficient and effective clustering methods for spatial data mining. In: Proceedings of VLDB, pp 144-55
[2] Bhanu B and Peng J 2000 Adaptive integrated image segmentation and object recognition IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 30 427-41
[3] Rao M 1971 Cluster analysis and mathematical programming Journal of the American statistical association 66 622-6
[4] Fisher D H 1987 Knowledge acquisition via incremental conceptual clustering Machine learning 2 139-72
[5] Tsai C-F, Hsu Y-F, Lin C-Y and Lin W-Y 2009 Intrusion detection by machine learning: A review expert systems with applications 36 11994-2000
[6] Žalik K R 2008 An efficient k’-means clustering algorithm Pattern Recognition Letters 29 1385-91
[7] Aarts E and Korst J 1988 Simulated annealing and Boltzmann machines
[8] Niknam T, Olamaie J and Amiri B 2008 A hybrid evolutionary algorithm based on ACO and SA for cluster analysis Journal of Applied Science 8 2695-702
[9] Niknam T, Firouzi B B and Nayeripour M 2008 An efficient hybrid evolutionary algorithm for cluster analysis. In: World Applied Sciences Journal: Citeseer
[10] Shelokar P, Jayaraman V K and Kulkarni B D 2004 An ant colony approach for clustering Analytica Chimica Acta 509 187-95
[11] Kao Y and Cheng K 2006 An ACO-based clustering algorithm. In: International Workshop on Ant Colony Optimization and Swarm Intelligence: Springer) pp 340-7
[12] Kennedy J and Eberhart R 1995 Particle swarm optimization. In: Proceedings of ICNN’95-International Conference on Neural Networks: IEEE) pp 1942-8
[13] Van der Merwe D and Engelbrecht A P 2003 Data clustering using particle swarm optimization. In: The 2003 Congress on Evolutionary Computation, 2003. CEC’03.: IEEE) pp 215-20
[14] Omran M, Engelbrecht A P and Salman A 2005 Particle swarm optimization method for image clustering International Journal of Pattern Recognition and Artificial Intelligence 19 297-321
[15] Kao Y-T, Zahara E and Kao I-W 2008 A hybridized approach to data clustering Expert Systems with Applications 34 1754-62
[16] Voges K E and Pope N K L 2012 Rough clustering using an evolutionary algorithm. In: 2012 45th Hawaii International Conference on System Sciences: IEEE) pp 1138-45
[17] Yang X-S 2013 Multiobjective firefly algorithm for continuous optimization Engineering with Computers 29 175-84
[18] Fister I, Fister I, Yang X-S and Brest J 2013 A comprehensive review of firefly algorithms Swarm and Evolutionary Computation 13 34-46
[19] Yelghi A and Köse C 2018 A modified firefly algorithm for global minimum optimization Appl. Soft. Comput. 62 29-44
[20] Algamal Z 2019 Variable Selection in Count Data Regression Model based on Firefly Algorithm Statistics, Optimization & Information Computing 7
[21] Al-Thanoon N A, Qasim O S and Algamal Z Y 2018 Tuning parameter estimation in SCAD-support vector machine using firefly algorithm with application in gene selection and cancer classification Comput Biol Med 103 262-8
[22] Al-Thanoon N A, Qasim O S and Algamal Z Y 2019 A new hybrid firefly algorithm and particle swarm optimization for tuning parameter estimation in penalized support vector machine with application in chemometrics Chemometrics and Intelligent Laboratory Systems 184 142-52
[23] Kahya M A, Althamir S A and Algamal Z Y 2020 Improving firefly algorithm-based logistic regression for feature selection Journal of Interdisciplinary Mathematics 22 1577-81
[24] Karthikeyan S, Asokan P and Nickolas S 2014 A hybrid discrete firefly algorithm for multi-objective flexible job shop scheduling problem with limited resource constraints *The International Journal of Advanced Manufacturing Technology* **72** 1567-79

[25] Qasim O S, Al-Thanoon N A and Algamal Z Y 2020 Feature selection based on chaotic binary black hole algorithm for data classification *Chemometrics and Intelligent Laboratory Systems* **204**

[26] Niknam T and Amiri B 2010 An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis *Applied soft computing* **10** 183-97

[27] Blake C 1998 UCI repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html](http://www.ics.uci.edu/~mlearn/MLRepository.html)