Hybrid Clustering Using Firefly Optimization and Fuzzy C-Means Algorithm

Krishnamoorthi Murugasamy¹*, Kalamani Murugasamy²

¹Department of CSE, Bannari Amman Institute of Technology, Sathyamanaglam, India
²Department of ECE, Bannari Amman Institute of Technology, Sathyamanaglam, India

Email: krishnamoorthim@bitsathy.ac.in, kalamanim@bitsathy.ac.in

Received 29 April 2016; accepted 15 May 2016; published 19 July 2016

Abstract

Classifying the data into a meaningful group is one of the fundamental ways of understanding and learning the valuable information. High-quality clustering methods are necessary for the valuable and efficient analysis of the increasing data. The Firefly Algorithm (FA) is one of the bio-inspired algorithms and it is recently used to solve the clustering problems. In this paper, Hybrid F-Firefly algorithm is developed by combining the Fuzzy C-Means (FCM) with FA to improve the clustering accuracy with global optimum solution. The Hybrid F-Firefly algorithm is developed by incorporating FCM operator at the end of each iteration in FA algorithm. This proposed algorithm is designed to utilize the goodness of existing algorithm and to enhance the original FA algorithm by solving the shortcomings in the FCM algorithm like the trapping in local optima and sensitive to initial seed points. In this research work, the Hybrid F-Firefly algorithm is implemented and experimentally tested for various performance measures under six different benchmark datasets. From the experimental results, it is observed that the Hybrid F-Firefly algorithm significantly improves the intra-cluster distance when compared with the existing algorithms like K-means, FCM and FA algorithm.

Keywords

Clustering, Optimization, K-Means, Fuzzy C-Means, Firefly Algorithm, F-Firefly

1. Introduction

The permeation of information via the World Wide Web has generated an incessantly growing need for the improvement of techniques for discovering, accessing, and sharing knowledge from the various domains. The in-
The clustering process discovers meaningful and natural clusters in the datasets. The major obstacle in clustering is that no prior knowledge about the given dataset is available. There are many data clustering algorithms available in the literature to perform clustering. Among them most widely used categories of the clustering are partitional and hierarchical algorithms. The most popular class of partitional clustering is FCM clustering algorithm. But FCM algorithm depends on the initial seed points and converges to local optima. FCM clustering algorithm is applied to a wide variety of geostatistical data analysis problems [2]. Cannon [3] suggested an efficient implementation of FCM clustering algorithm. To enhance and tackle the shortcoming of the FCM algorithm, a unified framework for performing density-weighted FCM clustering is developed [4]. A comparative analysis of FCM and Entropy-based Fuzzy Clustering (EFC) algorithm are performed in terms of the quality of clusters and their computational time. They indicate that the major drawback of FCM algorithm is that it always converges to a local minimum or a saddle point [5]. Hence, the researchers have shown their interest in optimization algorithms to overcome the difficulties in FCM clustering technique and to improve the quality of clusters [6].

Optimization is a process of finding the feasible solutions to the problems defined. The bio-inspired optimization algorithms have attracted many researchers to solve the problems of diverse fields. The main objective of these algorithms is to efficiently determine the near optimal solution for the problem statement defined. FA is one of the recently introduced swarm intelligence techniques and it is a kind of stochastic, nature-inspired, meta-heuristic algorithm used for solving complex problems. The FA is applied for clustering and its performance comparison with the commonly used optimization methods is performed as in the other population-based algorithm, the performance of the FA depends on the population size, attractiveness factor ($\beta$), light absorption coefficient ($\gamma$) and the distance between the two firefly particles ($r$). Their performance measure indicates that the FA based clustering is an efficient, reliable, and robust method, which can be applied successfully to produce optimal cluster centers. Recently firefly algorithm has been utilized by Senthilnath [7] for solving the clustering problem and proved that FA algorithm is superior to the state-of-art algorithm in the literature. An extensive review of FA algorithm and its application in engineering practice have been widely studied and suggested ideas to develop modified or hybridized FA algorithm for solving diverse problem [8]-[13].

It is observed from the literature that combining bio-inspired algorithms with traditional data clustering techniques will produce even better results and overcome the drawbacks in partitioning clustering algorithms [14]-[17]. In recent years, many efficient hybrid clustering algorithms based optimization algorithms have been stipulated to perform clustering to get better results than those the traditional data clustering algorithms. The hybrid algorithm will overcome the problems in the traditional clustering algorithms and also converges quickly to the global optimal solution. The optimization algorithms hybridize together with traditional algorithms to enhance it to be faster and efficient [18]. Karaboga and Ozturk [19] applied ABC algorithm to fuzzy clustering of medical data [20]. Kanade and Hall [21] developed a hybrid technique by combining FCM with ACO algorithm to enhance the performance of clustering. To overcome the drawback in FCM algorithm, a hybrid algorithm based on FCM and modified Stem Cells (SC) algorithm is investigated by Taherdangkoo and Bagheri [22]. Fister [23] used quaternion for the representation of individuals in firefly algorithm to produce improved results when compared to the FA algorithm.

From the literature, it is found that hybrid algorithm will improve the performance of the clustering process. Hence to overcome the shortcomings in the existing algorithms and to improve the clustering accuracy, the hybrid algorithm is proposed in this research work. The aim of the research work is to find the centroid of the clusters by minimizing the objective function, the sum of distances between the objects to their centers. For example, given N objects, the problem is to minimize the sum of squared distances between each object and allocate each object to one of the $k$ cluster centers. The research work presented here focuses on developing hybrid clustering algorithm by combining FA with FCM algorithm. The proposed hybrid algorithms form high quality clusters by utilizing the advantages of both the partitional clustering and bio-inspired algorithms.

This paper is organized as follows: Section 2 describes about the overview of Firefly Algorithm. The various
phases of proposed Hybrid F-Firefly algorithm are demonstrated in Section 3. Section 4 illustrates the experimental study and the evaluated experimental results under various performance metrics for the proposed and existing clustering algorithms are graphically shown and discussed in Section 5. Section 6 gives the conclusion of this research work.

2. Overview of Firefly Algorithm (FA)

The common approach used in this algorithm is that the fireflies will glow brighter when they attract other nearby fireflies. The attraction between the two fireflies decreases as the distance between them increases. If there are no nearby fireflies brighter than a particular firefly, then this firefly will move randomly in the search space. The flashing characteristics of the fireflies use the following three idealized rules [8]:

- All fireflies are unisex, so that one firefly is attracted to other fireflies irrespective of their sex.
- Attractiveness is proportional to the brightness that is the firefly with lesser brightness will move towards the brighter firefly. Both the attractiveness and brightness decrease as their distance increases. Firefly will move randomly if there are no fireflies brighter than the nearby fireflies.
- The brightness of the firefly shall be associated with the objective function.

There are two important features in the FA algorithm like variation in the light intensity and formulation of attractiveness. The attractiveness of the firefly depends on the brightness which in turn is associated with the objective function. FA algorithm is developed based on the rule of light intensity “I” that decreases with the increase in the square of the distance, \( r^2 \) and it is described by, \( I \propto \frac{1}{r^2} \). In the firefly based optimization problem, the brightness “I” of the firefly at a particular location “x” can be \( I(x) \propto f(x) \). However, the attractiveness “\( \beta \)” will vary with the distance between the two fireflies. The light intensity decreases with the increase in distance from its source. Light is absorbed in the media, so the attractiveness of two fireflies will vary with the degree of absorption.

In the simplest form, the light intensity \( I(r) \) varies according to the inverse square law:

\[
I(r) = \frac{I_0}{r^2}.
\]

(1)

The light intensity “I” and the absorption coefficient “\( \gamma \)” varies with the distance “\( r \)” for a given medium and it is given as:

\[
I = I_0 e^{-\gamma r}
\]

(2)

where, \( I_0 \) is the original light intensity.

The combined effect of both inverse square law and absorption can be assumed as Gaussian form and it is represented as:

\[
I = I_0 e^{-\gamma r^2}
\]

(3)

The firefly’s attractiveness is proportional to the light intensity of the adjacent fireflies. The attractiveness “\( \beta \)” of a firefly can be defined as:

\[
\beta(r) = \beta_0 e^{-\gamma r^2}
\]

(4)

where, “\( \beta_0 \)” is attractiveness at \( r = 0 \).

The Euclidean distance between the firefly “I” at \( x_i \) and firefly “J” at \( x_j \) is given by

\[
r_{ij} = \sqrt{\sum_{k=1}^{d} (x_{ik} - x_{jk})^2}
\]

(5)

The movement calculation of the firefly “I” is attracted to another more attractive (brighter) firefly “J” and it is estimated by:

\[
x_i = x_i + \beta_0 e^{-\gamma r_{ij}^2} (x_j - x_i) + \alpha \left( rand \frac{1}{2} \right)
\]

(6)

The second term represents the attraction and third term is randomization. Where, “\( \alpha \)” being randomization parameter and “\( rand \)” is a random number generator which is uniformly distributed between 0 and 1.
In the FA algorithm, the optimization of the objective function depends on the brightness and movement of the firefly. The brighter firefly will attract the nearby fireflies with low brightness. The firefly algorithm starts by initializing the population of fireflies. The brightness of the firefly determines the movement of the fireflies. In the iterative process, the intensity of the $i^{th}$ firefly is compared with intensity of $j^{th}$ firefly. Based on the difference in intensity, either $i^{th}$ firefly move towards $j^{th}$ firefly or $j^{th}$ firefly will moves towards $i^{th}$ firefly. The best solution obtained is continuously updated until certain stopping criterion is satisfied. Once the iterative process comes to an end, the best solution is determined.

From the literature, it is found that FA algorithm can outperform when compared to many other algorithms. FA algorithm expands and new variants of it started emerging to solve all kinds of optimization problems. FA is a swarm-intelligence-based algorithm, so it has similar advantages of the other swarm intelligence based algorithms. FA has two major advantages over other algorithms, namely automatic subdivision and the ability of dealing with multimodality [9].

- As the FA algorithm is designed based on the attraction and this attraction decreases with distance increases. This phenomenon leads the whole population to automatically subdivide into groups and each group can search around the local optimum. Among all these local optimum, the global solution can be found.
- This subdivision allows the fireflies to be able to find all optimum solution simultaneously if the population size is sufficiently high.

FA algorithm has some limitations such as:
- FA parameters are set fixed and they do not change over time.
- FA does not memorize any history of the best solution for each firefly and moves regardless of it, and it misses its best solution.

3. Hybrid F-Firefly Algorithm

The existing FA algorithm suffers because of the inability to memorize any history about the best solution for each firefly and also moves randomly if there are no fireflies brighter than the nearby fireflies. This problem can be overcome by the proposed Hybrid F-Firefly Algorithm in which the FCM operator is incorporated into the end of each iteration in the existing FA. The incorporation of the FCM algorithm has provides significant improvement in the performance of the firefly algorithm. The proposed Hybrid F-firefly algorithm consist of four major phases such as initialization phase, intensity calculation phase, movement calculation phase and FCM algorithm phase. The detailed description of all the phases of Hybrid F-Firefly algorithm is furnished as follows.

3.1. Initialization Phase

In the F-Firefly algorithm, the population (S) of fireflies within the search space is created. Based on the objective function, all the agents (fireflies) are randomly distributed in the search space. The position of the agents represents the possible solution (centroids) for the clustering problem. Furthermore, this phase will assign the algorithm parameters like $\beta_0$, $\gamma$, $\alpha$ and the maximum cycle number.

3.2. Intensity Calculation Phase

After the initialization phase, the intensity of each firefly is evaluated by measuring the distance between the position of the firefly and the whole data in the dataset. After calculating the distance, consider the minimum distance value among the firefly with respect to data from the dataset. Calculate the intensity value of each firefly based on the summation of minimum distance value obtained with respect to the data from the dataset. The formula for calculating the intensity is given below:

$$I(FF) = \sum_{i=1}^{n} A_i$$  \hspace{1cm} (7)

where, $i = 1,2,3,\cdots,n$ and $j = 1,2,3,\cdots,SN$. Then, FF is the Firefly and $A_i$ is the Minimum distance value for a particular firefly.

3.3. Movement Calculation Phase

The brightness of the firefly determines the movement of the fireflies in the search space. After the intensity
calculation, the fireflies are compared to find the new position. During the iterative process, the intensity of one firefly is compared with that of the other fireflies in the swarm and the difference in the brightness triggers the movement. The distance travelled depends on the attractiveness between the two fireflies. For example, consider a firefly “I” and “j”, the intensity value of the firefly “I” is compared with firefly “j”. If the intensity value of firefly “I” is more than firefly “j” then the firefly “j” is moved towards the firefly “I”. Based on the intensity value of the fireflies, the movement calculation is performed using Equation (8).

\[ x'_i = x_i + \beta_i e^{-\gamma r_{ij}} + \alpha \left( \text{rand} - \frac{1}{2} \right) \]  

(8)

where, \( x'_i \) is the new location the firefly “I”, \( x_i \) is the location of the firefly “I” that is attracted to another more brighter firefly “j” at location \( jx \) and \( \beta_i, \gamma \) and \( \alpha \) are constant values, \( \text{rand} \) is a value that varies from 0 to 1 and \( r_{ij} \) is the difference between \( x_i \) and \( x_j \).

Once the firefly “I” is moved to the new position and then update the intensity value of the firefly “I”. Similarly, the above mentioned procedure is repeated for the entire fireflies by keeping one of the fireflies as constant. The intensity value is calculated for the new position of the firefly obtained after the movement calculation is performed.

### 3.4. FCM Algorithm Phase

The proposed Hybrid F-Firefly algorithm is the modification of the existing FA algorithm by incorporation of the FCM operator (one step of FCM algorithm) to enhance the clustering performance. The main objective of incorporating the FCM operator in this stage is that, the FCM algorithm will find the local optimal solution effectively. Based on the current intensity value, the new position of the entire population of the fireflies is calculated by applying the FCM Operator. The process of FCM operator is carried out through the update of the membership values \( u_{ij} \) and position of the firefly \( f_j \) using the Equations (9) and (10) respectively [17].

\[ u_{ij} = \frac{1}{\sum_{i=1}^{N} \left( \frac{\|x_i - f_j\|^m}{\|x_j - x_i\|^m} \right)}^{\frac{1}{m-1}}, 1 \leq i \leq N \]  

(9)

where, \( u_{ij} \) is the degree of membership of \( x_i \) in the firefly \( j \), the value of fuzziness component “\( m \)” = 2 and \( x_i \) is the data associated with the firefly under study.

\[ f_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m} \]  

(10)

where, \( f_j \) is the solution obtained after applying the FCM operator in the firefly \( j \).

The new position for each firefly is evaluated and the intensity value is updated. The fireflies are sorted based on the intensity value before moving to the next iteration. During the iterative process, the best solution obtained so far is continuously updated and all the above mentioned phases of the F-Firefly algorithm are prolonged until the stopping criteria are reached. After the iterative process comes to an end, the best solution is determined by ranking the position of the fireflies and the post process is initiated to select the final centroids. The overall process of the proposed Hybrid F-Firefly algorithm is represented in the flowchart and it is shown in Figure 1.

### 4. Experimental Study

Experiments are conducted to evaluate the performance proposed and existing algorithms. All the algorithms are implemented in JAVA language and executed on a core i5 processor, 2.1 MHZ, 4 GB RAM computer. The initial set up for the experiments is as follows: number of fireflies (\( S \)) = 20, attractiveness (\( \beta_0 \)) = 1 and light absorption coefficient (\( \gamma \)) = 1. To present the effectiveness of the proposed Hybrid F-Firefly clustering algorithm, most commonly used partitional clustering algorithms are used for comparison of the results. The two major partitional clustering algorithms such as k-means and FCM clustering algorithms are used for comparison as they are commonly used for hard and soft clustering technique in a wide range of applications. The FA algorithm is also utilized for comparison because the Hybrid F-Firefly algorithm is a modification of the original FA algorithm.
The performance of the F-Firefly algorithm is compared with the existing algorithms based on the intra-cluster distance and cluster validity metrics. Here, the internal validity measures like Beta index, Distance index, and DB index are utilized for evaluating and comparison. To perform the experiments on the Hybrid F-Firefly and existing algorithms, six benchmark datasets such as Abalone, Zoo, Iris, Wine, Liver and Thyroid are selected from the UCI machine learning repository.
5. Results and Discussion

The experimental results obtained for the proposed Hybrid F-Firefly algorithm are highlighted in this section. In order to show the effectiveness and the strength of the Hybrid F-Firefly algorithm, experiments have been conducted using intra-cluster distance and the cluster validity measures. The Hybrid F-Firefly and existing algorithms are executed for a maximum of 200 iterations and cluster size is fixed as 3. The results are tabulated for Hybrid F-Firefly, k-means, FCM and FA algorithm and comparisons made under various performance measures for six different datasets.

5.1. Intra-Cluster Distance

Lower the value of intra-cluster distance value indicates that the goodness of the clustering algorithm. The average intra-cluster distance calculated for the FCM, k-means, FA, and Proposed Hybrid F-Firefly algorithms and it is tabulated in Table 1. From the results, it is observed that the proposed Hybrid F-Firefly algorithm yields the improvements of intra-cluster distance from 0.2% to 44.4% when compared to all existing algorithms under various datasets. These results reveal that the proposed Hybrid F-Firefly algorithm outperforms when compared to all the existing algorithms used for comparison in case of abalone, zoo, thyroid and wine datasets. On all these datasets, the Hybrid F-Firefly algorithm produces minimum average intra-cluster distance values and indicates the effectiveness of the hybrid algorithm. The experimental results illustrate that, the Hybrid F-Firefly algorithm escape from trapping in local optima and discover better results in most of the cases.

5.2. Beta Index

The experimental results of the beta index obtained by Hybrid F-Firefly and existing algorithms using six different datasets are shown graphically in Figure 2. The higher the beta index values, the better the clustering algorithm. The beta index values are significantly improved in all the datasets for the Hybrid F-Firefly algorithm when compared to all the existing algorithms. The percentage of improvement in beta index for Hybrid F-Firefly algorithm when compared to firefly algorithm are made and it is observed that proposed algorithm improves the beta index value from 20% to 40% when compared to firefly algorithm for all the datasets.

5.3. Davies-Bouldin (DB) Index

The performance of the Hybrid F-Firefly and exiting algorithms are evaluated using Davies-Bouldin index with the six benchmark datasets are shown in Figure 3. It is well known that the minimum DB-Index value indicates the effectiveness of the clustering algorithm. From the graphical results, it is observed that the Hybrid F-Firefly algorithm produces minimum DB-Index values when compared to that of the k-means, FCM and FA algorithms. The percentage of improvement from the experimental results conducted for the proposed and FA algorithm using the six benchmark datasets are made and proposed algorithm improves DB index value from 20% to 23% when compared to firefly algorithm under all six datasets.

5.4. Distance Index

The distance index is calculated as the ratio of average intra-cluster distance and average inter-cluster distance.

Table 1. Comparison of average intra-cluster distance of existing and proposed algorithms.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>FCM</th>
<th>k-means</th>
<th>Firefly</th>
<th>F-Firefly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abalone</td>
<td>16,765.43</td>
<td>16,103.12</td>
<td>15,718.34</td>
<td>15,619.5</td>
</tr>
<tr>
<td>Zoo</td>
<td>17,603.54</td>
<td>17,116.29</td>
<td>16,380.46</td>
<td>16,334.7</td>
</tr>
<tr>
<td>Liver</td>
<td>7581.02</td>
<td>7309.69</td>
<td>6520.27</td>
<td>6508.23</td>
</tr>
<tr>
<td>Iris</td>
<td>128.44</td>
<td>126.5</td>
<td>76.9</td>
<td>71.39</td>
</tr>
<tr>
<td>Thyroid</td>
<td>11,356.37</td>
<td>11,062.62</td>
<td>10,088.94</td>
<td>9943.71</td>
</tr>
<tr>
<td>Wine</td>
<td>16,885.46</td>
<td>16,530.33</td>
<td>16,231.6</td>
<td>16,201.4</td>
</tr>
</tbody>
</table>
It is well-known that the minimum distance index value illustrates better performance of the clustering algorithm. To evaluate the impact of the Hybrid F-Firefly algorithm, the comparisons are performed with the existing algorithms like k-means, FCM, FA algorithm. In Figure 4, the distance index values obtained for various algorithms are shown graphically. When considering all the six datasets, the Hybrid F-Firefly algorithm achieves minimum distance index value in most of the cases and produces better results than the existing algorithms. For the Hybrid F-Firefly algorithm, the intra-cluster distance reduces while inter-cluster distance increases, so the distance index value is minimized. There is 26% - 33% of improvement in distance index value for Hybrid F-Firefly algorithm when compared to Firefly algorithm.
6. Conclusion

In this research work, a Hybrid F-Firefly algorithm is developed by incorporating FCM operator at the end of each iteration in FA algorithm. Various performance measures are evaluated for the proposed and existing algorithms under six different datasets. From the results, it is observed that the average intra-cluster distance computed using the six different datasets is minimum and yields the percentage improvements of 0.2% - 44.4% for the Hybrid F-Firefly algorithm when compared with the FCM, k-means, and FA algorithms. In addition, the proposed Hybrid F-Firefly algorithm provides 20% - 40% improvement of beta index value, 20% - 23% improvements of DB index value, and 26% - 33% improvement of distance index value when compared to the FA algorithm. These improvements in the performance of the Hybrid F-Firefly clustering algorithm are due to local search capability of FCM algorithm and so the global search property of FA algorithm is combined in the Hybrid F-Firefly algorithm. From the experimental results, it is observed that the performance of proposed Hybrid F-Firefly algorithm is better when compared with the original FA algorithms.

References


Evolutionary Computation, 13, 34-46. http://dx.doi.org/10.1016/j.swevo.2013.06.001


Submit or recommend next manuscript to SCIRP and we will provide best service for you:

Accepting pre-submission inquiries through Email, Facebook, Linkedin, Twitter, etc
A wide selection of journals (inclusive of 9 subjects, more than 200 journals)
Providing a 24-hour high-quality service
User-friendly online submission system
Fair and swift peer-review system
Efficient typesetting and proofreading procedure
Display of the result of downloads and visits, as well as the number of cited articles
Maximum dissemination of your research work
Submit your manuscript at: http://papersubmission.scirp.org/