Fuzzy Systems and Data Mining X
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA241403

The Impact of Firefly Algorithm (FA) Optimization on Gaussian Kernel-Based Fuzzy C-Means Clustering (GKFCM) Efficiency

Narongdech DUNGKRATOKE [a], Chantana SIMTRAKANKUL [a], Janejira LAOMALA [a] and Sayan KAENNAKHAM [b,1]
[a] *Department of Interdisciplinary Science and Internationalization, Institute of Science, Suranaree University of Technology, Nakhon Ratchasima, 30000, Thailand;*
[b] *School of Mathematics and Geoinformatics, Institute of Science, Suranaree University of Technology, Nakhon Ratchasima, 30000, Thailand;*
ORCiD ID: Narongdech Dungkratoke https://orcid.org/0009-0006-7097-9666, Chantana Simtrakankul https://orcid.org/0009-0002-0873-7384, Janejira Laomala https://orcid.org/0009-0000-9350-2541, Sayan Kaennakham https://orcid.org/0000-0001-9682-559X

Abstract. This study explores the effectiveness of the traditional Firefly algorithm (FA) in optimizing the Gaussian Kernel-based Fuzzy C-means clustering (GKFCM) algorithm by adjusting 'sigma' and 'm'. We compare GKFCM with FA optimization (With FA) to without it (Without FA) using the Calinski Harabasz (CH) index and the number of iterations. For all four datasets analyzed in this study, the findings consistently indicate that the GKFCM algorithm optimized with the Firefly algorithm (FA) performs substantially better than its non-optimized counterpart, achieving higher Calinski Harabasz (CH) scores and requiring fewer iterations across various data types. Results from two initial particle distribution styles confirm FA's robustness in refining clustering outcomes and emphasize its role in enhancing clustering quality and efficiency.

Keywords. Clustering Optimization, Firefly Algorithm, Gaussian Kernel, Calinski Harabasz Index

1. Introduction

Clustering is a fundamental technique in machine learning that organizes datasets into subsets or clusters based on shared characteristics. Among the prominent methods, the Fuzzy C-Means (FCM) Algorithm enhances traditional k-means clustering by incorporating fuzziness, allowing each data point to belong to multiple clusters with varying degrees of membership [1]. A sophisticated variant, the Kernel-based FCM, excels in managing noisy datasets, data with overlapping clusters, and non-linear, complex data distributions. Notably, the Gaussian kernel function is the most popular

[1] Corresponding Author: Sayan Kaennakham, sayan_kk@g.sut.ac.th.

choice for Kernel-based FCM, as it significantly improves clustering by mapping data into a higher-dimensional space [2,3].

Despite its advancements, the Gaussian Kernel-based FCM (Gaussian K-FCM) requires careful tuning of two critical parameters: the Gaussian shape and the fuzziness level. These parameters are highly sensitive to the specific problem, necessitating skilled and experienced users for their optimal determination [4]. This challenge has spurred researchers over the past decade to develop strategies for automatic parameter adjustment. One promising approach that has emerged is the Firefly optimization algorithm (FA) [5], which aims to autonomously optimize these parameters, enhancing the practical deployment of Gaussian K-FCM.

The Firefly Algorithm (FA) is an optimization technique inspired by the bioluminescent communication behavior of fireflies. FA utilizes the principles of attraction based on brightness, where each firefly moves towards brighter ones within the search space, thereby simulating a form of natural swarm intelligence [6]. The brightness of each firefly is determined by the value of the objective function at their location, guiding their movements towards optimal solutions. This algorithm is particularly effective for solving complex optimization problems across various domains due to its simplicity, efficiency, and the ability to escape local optima [7]. While FA has previously been applied to optimize Gaussian Kernel-based FCM [8], our study builds upon this foundation by systematically evaluating its effectiveness across multiple datasets, assessing improvements in clustering quality and efficiency.

Our work will explore the application of the Firefly Algorithm for fine-tuning the parameters of the Kernel FCM, aiming to enhance its performance and applicability in various data-driven domains.

2. Methodology

2.1. The Gaussian Kernel Fuzzy C-means (GKFCM)

The objective of GKFCM is to minimize the following function

$$J_m(U,V) = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \times \exp\left(-\left\|\mathbf{x}_i - \mathbf{v}_j\right\|^2 \Big/ \sigma^2\right) \tag{1}$$

where $U = \left[u_{ij}\right]$ represents the fuzzy membership matrix with u_{ij} denoting the degree of membership of the i-th data point $\mathbf{x}_i$ in the j-th cluster, $V = \left[\mathbf{v}_j\right]$ is the matrix of cluster centers in the transformed feature space, m is the fuzzification parameter ($m > 1$), controlling the level of cluster fuzziness. The Gaussian kernel function, with σ as the kernel width parameter, with the Euclidean distance, is used here.

To solve for U and V, the algorithm iteratively updates the membership u_{ij} and the cluster centers v_j using the following rules

$$u_{ij} = \left[\sum_{k=1}^{C} \left(\frac{K(\mathbf{x}_i, \mathbf{v}_j)}{K(\mathbf{x}_i, \mathbf{v}_k)} \right)^{\frac{1}{m-1}} \right]^{-1} \tag{2}$$

The cluster centers in GKFCM are typically not explicitly updated in the original feature space. Instead, they are indirectly influenced by the kernel distances in the calculation of memberships and by applying kernel methods.

2.2. The Firefly Algorithm (FA)

In FA, the brightness I_i of firefly i at a particular location $\mathbf{x}_i$ is determined by the value of the objective function $f(\mathbf{x}_i)$, i.e., $I_i \propto f(\mathbf{x}_i)$. For maximization problems, brightness is directly proportional to $f(\mathbf{x}_i)$. The attractiveness β of a firefly is a function of the distance r_{ij} between two fireflies i and j, given by

$$\beta(r_{ij}) = \beta_0 e^{-\gamma r_{ij}^2} \tag{3}$$

where β_0 is the attractiveness at $r = 0$ and γ is the light absorption coefficient, which controls how the attractiveness decreases with distance. The distance r_{ij} between two fireflies i and j located at $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively, is typically calculated using the Euclidean distance, simply defined as $r_{ij} = \left\| \mathbf{x}_i - \mathbf{x}_j \right\|$. A firefly i will move towards another more attractive (brighter) firefly j according to the following equation

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t + \beta(r_{ij})(\mathbf{x}_j^t - \mathbf{x}_i^t) + \alpha(\mathbf{rand} - 0.5) \tag{4}$$

where α is a randomization parameter, and **rand** is a random number generated uniformly from $[0,1]$.

3. The Proposed Algorithm

Step 1: Define an objective function $J(m, \sigma)$ based on Calinski Harabasz score.
Step 2: Initialize a population of fireflies, (m, σ), within their feasible ranges ($m \in (1,3), \sigma \in [0.01,5]$).
Step 3: For each firefly, update positions by moving towards brighter fireflies (i.e., better parameter sets), adjusting m and σ using

$$m_i^{t+1} = m_i^t + \beta(r_{ij})(m_j^t - m_i^t) + \alpha(rand - 0.5) \tag{5}$$

$$\sigma_i^{t+1} = \sigma_i^t + \beta(r_{ij})(\sigma_j^t - \sigma_i^t) + \alpha(rand - 0.5) \tag{6}$$

Step 4: Iterate until reaching 100 iterations or $\max_{All\_i}\left|r_i^{t+1}-r_i^t\right| \le 10^{-4}$.

Step 5: Choose the parameters from the position of the brightest firefly at the end of the iterations.

4. Experimental Setup

4.1. The datasets

Figure 1 illustrates four types of datasets under investigation. Types 1 and 2 exhibit nonlinear, spiral-like patterns; Type-1 has a tightly adherent distribution indicating low noisiness and high precision, while Type-2 shows similar patterns but with increased spread and noisiness, adding complexity to its modeling. Type-3 and Type-4 both feature two distinct clusters; Type-3 has compact clusters with low noisiness and minimal complexity, whereas Type-4's clusters are more dispersed, increasing the complexity and variability in data.

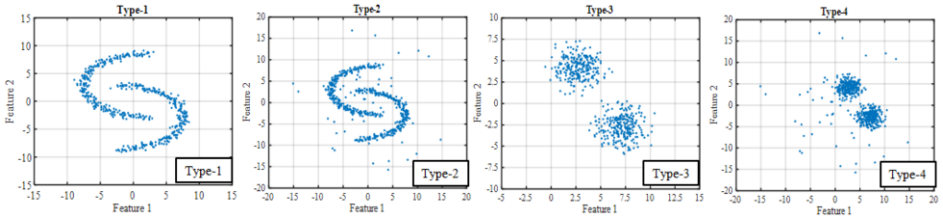

Figure 1. The four datasets: Type-1 and Type-3 each contain 500 points, while the other two contain 550 points each.

4.2. The initial FA node distribution styles

The initial node distribution in the Firefly Algorithm critically affects its performance by determining the balance between exploring the solution space and exploiting known good areas. This choice is strategic, influencing both the efficiency and effectiveness of the algorithm in response to the specific characteristics of the optimization problem. To further assess this impact, our work tested the algorithms using two styles of initial node distribution as shown in Figure 2.

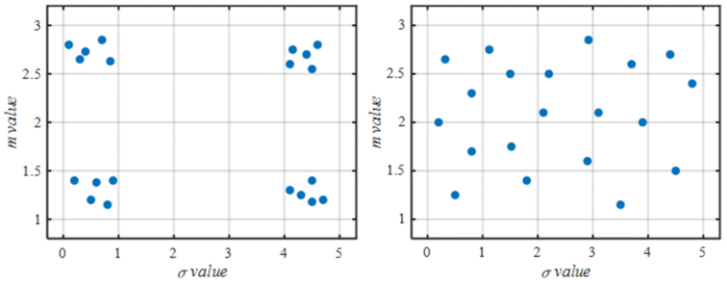

Figure 2. The initial FA node distribution styles; Left) Style-1, and Right) Style-2, containing 20 points each.

By comparing these two configurations, we aimed to understand how different initial node placements affect the algorithm's balance between exploring broadly across the solution space and concentrating on refining solutions in promising areas. Limiting the analysis to these two distinct styles also prevents excessive complexity, ensuring clear insights into the role of initial distribution without complicating the assessment with multiple configurations.

5. Main Results and General Discussions

The investigation used two numbers of fireflies; 10 and 20, each tested with two initial node distribution styles as previously shown. All simulations were performed on the same computational configuration to ensure fairness.

Table 1. Comparison of performance with and without FA, using 10 fireflies (initial distribution Style-1).

Data Type	Calinski Harabasz		No. of Iterations		(sigma, m)	
	Without FA	With FA	Without FA	With FA	Without FA	With FA
Type-1	141.1030	677.1037	42	6	(1, 2)	(5.0,2.9)
Type-2	182.1749	605.4243	102	38	(1, 2)	(5.0,1.1)
Type-3	375.3746	2751.874	57	16	(1, 2)	(4.1,2.6)
Type-4	43.6024	962.3828	97	68	(1, 2)	(4.1,2.6)

As can be seen in Table 1, the implementation of the Firefly algorithm (FA) for optimizing the Gaussian Kernel-based Fuzzy C-means clustering (GKFCM) algorithm demonstrates significant enhancements in clustering performance across various data types. The Calinski Harabasz (CH) scores, a metric assessing the ratio of between-cluster to within-cluster sums of squares, show marked improvements when optimized with FA. For example, Type-1 data sees an increase in CH score from 141.103 to 677.1037, and even more dramatically, Type-4 data's CH score escalates from 43.6024 to 962.3828. This increase signifies a more effective clustering with better defined separations between clusters when parameters are optimized by FA.

Additionally, the optimization through FA considerably reduces the number of iterations required for the GKFCM algorithm to converge, enhancing computational efficiency. Without FA, the iterations needed range from 42 to 102 across different data types, while with FA, these are reduced significantly—for instance, from 97 to 68 in Type-4 data and from 42 to 6 in Type-1 data. The optimization also influences the algorithm's parameters (σ , m), which vary with FA and are constant without it (e.g., Type-1 data with parameters changing from (1, 2) to (5, 2.9)). This tailored approach in adjusting parameters per data type contributes both to the quality of clustering and the efficiency of the algorithm, demonstrating the substantial benefits of incorporating an optimization algorithm like FA in clustering methodologies.

Table 2. Comparison of performance with and without FA, using 10 fireflies (initial distribution Style-2).

Data Type	Calinski Harabasz		No. of Iterations		(sigma, m)	
	Without FA	With FA	Without FA	With FA	Without FA	With FA
Type-1	141.1030	407.7983	42	99	(1, 2)	(3.96,2.19)
Type-2	182.1749	594.8594	102	99	(1, 2)	(4.43,2.45)
Type-3	375.3746	2751.874	57	24	(1, 2)	(4.24,2.76)
Type-4	43.6024	962.3828	97	20	(1, 2)	(4.11,2.15)

The experimental results comparing the performance of the Gaussian Kernel-based Fuzzy C-means clustering (GKFCM) algorithm with and without the optimization by the Firefly algorithm (FA) initial particle distribution style 2, as shown in Table 2, illustrate significant improvements in clustering quality and efficiency when parameters are optimized. The Calinski Harabasz (CH) index, which evaluates clustering effectiveness by measuring the ratio of between-cluster to within-cluster variance, shows substantial increases across all data types when FA is utilized. Specifically, for Type-1 data, the CH value rises from 141.103 to 407.7983, and for Type-4, it increases from 43.6024 to 962.3828. These improvements indicate that the optimized GKFCM algorithm, with tailored parameters of sigma and m, produces more distinct and well-separated clusters.

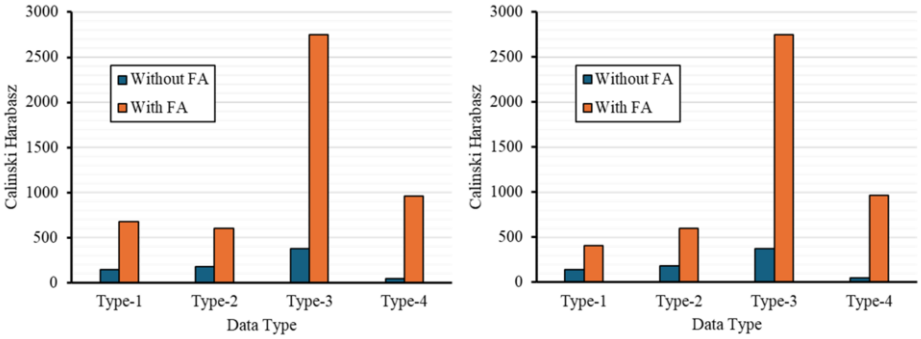

Figure 3. Calinski Harabasz values produced by both cases, using 20 fireflies, with; (left) Initial particle distribution style 1 and (right) Initial particle distribution style 2.

Contrary to the expected trend, the number of iterations needed for the algorithm to converge increased in some cases when optimized by FA. For example, Type-1 data required an increase from 42 to 99 iterations, and similar trends are observed in Type-2 and Type-3 data. This suggests that while FA enhances the clustering quality significantly, it may require more iterations to fine-tune the parameters to achieve the optimal clustering solution. The optimized parameters (sigma, m) show considerable variation across data types, demonstrating the adaptive capability of FA to tailor these parameters to specific dataset characteristics, further contributing to the enhanced performance of the GKFCM algorithm. For the case of using 20 fireflies, it was found that the overall results were similar to those discussed so far; therefore, they are not detailed here due to space limitations. Nevertheless, some results are illustrated in Figure 3.

6. Conclusion

The experiment comparing two initial particle distribution styles of the Firefly algorithm (FA) for optimizing the Gaussian Kernel-based Fuzzy C-means clustering (GKFCM) algorithm demonstrates significant benefits. FA optimization markedly improves Calinski Harabasz (CH) values, indicating better cluster separations. For instance, Type-4 data saw a CH increase from 43.6024 to 962.3828 with FA. This improvement is coupled with a reduction in iterations required for convergence, showcasing FA's computational efficiency. For Type-1 data, iterations dropped from 42 to 6. These results confirm that optimizing 'sigma' and '*m*' via FA enhances clustering quality and

accelerates convergence. Future work will extend FA-optimized GKFCM across diverse datasets and explore other optimizers, such as PSO, for optimal parameter tuning across varied applications.

Acknowledgments

This work was supported by (i) Suranaree University of Technology (SUT, http://www.sut.ac.th), (ii) Thailand Science Research Innovation (TSRI, https://www.tsri.or.th), and (iii) National Science, Research and Innovation Fund (NSRF) (NRIIS number 195619). The grant recipient is S. Kaennakham who would like to express his sincere gratitude for their support.

References

[1] Suganya R, Shanthi R. Fuzzy c-means algorithm-a review. International Journal of Scientific and Research Publications. 2012 Nov;2(11):1.
[2] Simões EC, de Carvalho FD. Gaussian kernel fuzzy c-means with width parameter computation and regularization. Pattern Recognition. 2023 Nov 1;143:109749.
[3] Chinta S, Tripathy BK, Rajulu KG. Kernelized intuitionistic fuzzy C-means algorithms fused with firefly algorithm for image segmentation. In2017 international conference on microelectronic devices, circuits and systems. 2017 Aug 10 (pp. 1-6). IEEE.
[4] Ding Y, Fu X. Kernel-based fuzzy c-means clustering algorithm based on genetic algorithm. Neurocomputing. 2016 May 5;188:233-8.
[5] Yang XS, He X. Firefly algorithm: recent advances and applications. International journal of swarm intelligence. 2013 Jan 1;1(1):36-50.
[6] Yang XS. Firefly algorithms for multimodal optimization. International symposium on stochastic algorithms 2009 Oct 26 (pp. 169-178). Berlin, Heidelberg: Springer Berlin Heidelberg.
[7] Zare M, Ghasemi M, Zahedi A, Golalipour K, Mohammadi SK, Mirjalili S, Abualigah L. A global best-guided firefly algorithm for engineering problems. Journal of Bionic Engineering. 2023 Sep;20(5):2359-88.
[8] Thomas E, Kumar SN. Fuzzy C Means Clustering Coupled with Firefly Optimization Algorithm for the Segmentation of Neurodisorder Magnetic Resonance Images. Procedia Computer Science. 2024 Jan 1;235:1577-89.