International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Optimizing K-Means Clustering using the Artificial Firefly Algorithm

LNC. Prakash K.[1], G. Suryanarayana[2], N. Swapna[3], T. Bhaskar[4], Ajmeera Kiran[5]

Submitted: 25/04/2023 **Revised:** 26/06/2023 **Accepted:** 06/07/2023

Abstract: Data clustering is a typical data analysis approach that is utilised in a variety of domains, namely machine learning, pattern matching, and visual analytics. K-means clustering is a popular and straightforward solution to data clustering, although it has important shortcomings, including local optimum convergence and initial point sensitivities. To attend the challenge of local convergence of optimal clusters in this article a swam-based optimization technique is proposed. Firefly method is a swarm-based technique used for optimizing challenges. This research proposes a tale approach for clustering data using the firefly algorithm. It is demonstrated how the K-Means technique may be applied to locate the centroids for the known initial cluster centres. The approach was later enhanced to improve centroids and clusters using firefly optimization. This novel algorithm is known as AFA. The experimental findings demonstrated the suggested method's efficiency and capabilities for data clustering and the conclusions show that the suggested model outperform traditional K-means clustering.

Keywords: *Artificial firefly algorithm, K-Means, Optimization*

1. Introduction

Clustering is a key unsupervised categorization methodology. In clustering, the system identifies its characteristics and patterns on its own, with no input and output translation supplied. Techniques for clustering identify similarities and assumptions from different types of data sources and then group them to create definable groupings. Techniques for clustering have been used to solve a variety of issues, namely the mining of data [1], recognising patterns [2], information compression [3], and predictive modelling [4]. Clustering is used to locate homogeneous groupings of data elements in a collection of data. Each category is referred to as a cluster, and it is distinguished by the simple fact that things belonging to identical groups tend to be more comparable than those belonging to various groupings. When a few groups, K, has been determined in advance, clustering can be defined as distributing the number of n items in N dimensional space across K categories in a way entities in the same group tend to be more like each other respects than items across various groupings. This entails minimising some optimisation criteria. Clustering aids in the surface-level analysis of data that is not structured. Cluster generation is determined by various criteria such as minimum distance, charts, and data source concentration. Determining a measure of closeness connecting the items based on a measurement called the similarity measure is used to group them into groupings. It is simpler to identify closeness measurements with fewer features. Creating similarity metrics becomes more difficult as the number of attributes grows. Various kinds of segmentation procedures in data mining employ various strategies for grouping data from databases. There are various sorts of algorithms for clustering that are capable of processing various forms of distinctive data.

The one having seen most frequently about is centroid-based segmentation. It's somewhat picky about the initial settings that one provides it, but it's quick and effective. These techniques split data points depending on the presence of several centres in the data. A group of points has been allocated to each data point according to its quadratic proximity from the centre. This is probably the most popular clustering method. The basic concept underlying segmentation is to divide databases into components or to categorise items in an existing data base. A few studies [5] present a range of partitioned strategies, but the major category is clustering algorithms, which are commonly employed in data segmentation due to their outstanding performance [6, 7]. Clustering techniques seek to identify related clusters (regions) and split a database into many divisions. There have been numerous designs for grouping activities reported in the literature. For the sake of ease and straightforwardness, k-means is any of the greatest widely tapped techniques [8]. Clustering using k-means is a form of vector quantization technique derived from signal processing that seeks to segment the n data points into k groups, with every

[1] Computer Science & Engineering (DS), CVR College of Engineering, Hyderabad, India.
ORCID ID: 0000-0002-0084-1331
klnc.prakash@gmail.com
[2]Department of Information Technology, Vardhaman College of Engineering, Hyderabad, India.
ORCID ID : 0000-0002-5552-3971
surya.aits@gmaill.com
[3]Computer Science & Engineering, Vijay Rural Engineering College, Nizamabad, Telangana, India.
ORCID ID: 0000-0001-7029-2211
swapnanaralas@gmail.com
[4]Computer Science & Engineering, CMR College of Engineering and Technology, Hyderabad, Talangana, India.
ORCID ID: 0000-0002-0458-650X
bhalu7cs@gmail.com
[5]Computer Science & Engineering, MLR Institute of Technology, Dundigal, Hyderabad.
ORCID ID: 0000-0003-3826-4346
kiranphd.jntuh@gmail.com
** Corresponding Author Email: surya.aits@gmaill.com*

sample fitting to the group with the nearest mean that attends as the group's paradigm. The goal of this research ought to provide the optimized clusters using an unconventional firefly optimization which is a nature inspired optimization technique. Nature inspired optimisation (NIO) reflects the evolution of computing techniques via the actions or behaviours of various ecological creatures/plants/elements. Nature Inspired Algorithms (NIA) were developed because of these computer technologies. The majority of NIA are categorized as Evolutionary Algorithms (EA) and Swarm Intelligence (SI) related strategies (along with certain other methods that employ chemical and physical features). EA is entirely dependent on the evolutionary behaviours of ecosystems in nature. EA uses crossover and mutation procedures to optimise difficult situations. SI-based computations, on the other hand, are referred to as swarm optimisation techniques, and are designed to optimise specific issues by replicating the collective activities of living swarms. The term "swarm" refers to a group of existent living things like as the birds, fishes, and beetles such as bees, termites, and ants, among others.

The elements of the swarm interact with each other by avoiding any centralised authority and moving through the data gathered from their surroundings [9]. SI methods include ant colony [10], particle swarm optimisation [11], multi-swarm PSO [12], cuckoo search [13], and firefly technique [14], among others. When it comes to effective SI-based computations, firefly algorithm is based on populations and on the cutting edge. If our significance is determined by how we interact of exceptional agents, procedures can be categorized as attract or non-attract. firefly algorithm is an excellent instance of an fascinate-based strategy that nature has duplicated. In this article firefly procedure is treated as an optimization procedure that optimizes the clusters which are resulted from k means grouping procedure. The remaining part of this article is organized as go along. portion 2, deliberates the methodical literature revision activity counted for this study. Part 3 explains about the arrangement and operating models of firefly algorithm. Section 4, talk about the proposed methodology for cluster optimization. Section 5, discuss the empirical results of the proposed methodology. The finale of this article beside with specific probable future enhancements are conferred in Portion 6.

2. Review of literature

In this section, the sources of some algorithm-based clustering methods along with some optimizing techniques is reviewed. clustering is an effective method that positively influence the accuracy and speedup of data analysis the basic version of k means technique is discussed in [15,16]. The work [17], provides an algorithm using heuristics that is comparatively fast and has fewer preference in the amount of data found in the groups. This approach separates huge groups split two pieces sequentially, and once enough groups is obtained, the centroid is optimized once more. The final optimization enables to fine-tune the clusters' unfairness and inaccuracy. Metaheuristic algorithms for searching, which are distinguished by their robust capacity for searching in terms of discovery and utilisation, have been frequently used to help K-Means in escaping from local optimisation traps by investigating and getting more optimised arrangements of group centres. The adverse effects of difficult real-world data can thus be minimised by more accurate cluster recognition because of optimised centre points. To address the issues of optimization of k-means [18], offer the t-k-means, a durable and reliable k-means type, in addition to its rapid counterpart. According to theory, researchers construct the t-k-means and investigate its durability and stabilisation using the loss mechanism and the grouping centre expression respectively.

To address the problems of finding real groups in the traditional K-Means clustering procedure, ant colony optimisation is used in [19], The study presents two approaches for employing ants in K-Means. The first technique allows the ant to go on an unplanned stroll and select a data item. The Pick and Drop probability of that specific data item are computed. These values influence when a data item stays in a single cluster or moves to elsewhere. rather than having the ant pick up a data item at random from among we compute the pick and drop likelihood and let the ant navigate to the data item with the greatest likelihood of being transported to another cluster. In the article [20], k-means clustering is being enhanced utilizing genetic algorithm so that the challenges in finding real clusters using k-means can be overridden. The study [21], introduces an innovative Approach to address the clustering challenge. The created system separates the samples into several subgroups to maximize searched diversity and thus improve the outcome of clustering, with only one instance of the true Spiral Optimization, which iteratively spins the items throughout its self-centered core. The k-means methodology was used to enhance the value of expected methodology. In this study [22], the authors investigate which factors impair the efficiency of the k-means method along with how much of this degradation may be addressed through employing a better initialization strategy or by resuming the procedure. Our primary discovery is that whenever the groups coincide, these two approaches can considerably increase k-means. With this grouping comparison, the simple nearest point algorithm decreases the number of incorrect groups from 15% to 6% on aggregate.

A new technique for removing duplication has been suggested in [23]. This identifies all types of gathered data as either useful or unimportant in selecting appropriate information prior to it is delivered to the starting location or group head. LEACH (Low-Energy Adaptive Clustering Hierarchy) is a cluster-built routing technology that employs the establishment of clusters. The LEACH selects one of the network sensor connections, like a Cluster Head (CH), to take on the role of a new power distribution capacity. The key motivation for such spontaneous clustering was that it generated in higher overhead owing to modifications in the CH and adverts. To optimise the CH selection, Particle Swarm Optimisation (PSO) and River Formation Dynamics (RFD) are utilised. In this article a variant of firefly procedure is used to optimise the results of k-means clustering technique. Due to its habits of pursuing all lighter fireflies in the region, the initially developed FA model [24] exhibits several distinctive characteristics in the searching process, although it exhibits sluggish convergence and high computational expenses.

Many FA versions have been suggested to address the problems by boosting the basic FA model's investigation capabilities and search heterogeneity. The tactics used to enhance the initial FA framework can be divided into three categories: adaptive methods of parameter adjustment, population diversification of operations, and hybrid searching pattern incorporation [25]. Baykasoglu and Ozsoydan [26] suggested FA2, a form of FA with two approaches: (1) substituting an exponential formula with a function that is inverse of proximity as the attraction factor and (2) creating an acceptable likelihood for a firefly's location to be altered or not. Static and dynamic multifunctional knapsack challenges were used to test the FA2 model. According to the findings, FA2 was more successful than some other optimization techniques. Firefly algorithm ((FA) and its derivatives are also commonly employed to solve multimodal optimisation challenges. FA was applied by Gandomi et al. [27] to a collection of seven diversified variable structured optimisation issues with nonlinear nature and several local optimal solutions. On these optimising tasks, research findings showed that FA was more effective than metaheuristic methods like PSO, GA, and Harmony Search (HS). In addition to FA, some Empirical investigations have widely proven the usefulness of such mixed grouping models as discussed in Tabu Search (TS) [28], Simulated Annealing (SA) [29], Genetic Algorithm (GA) [30], Artificial Bee Colony (ABC) [31], Particle Swarm Optimization (PSO) [32], Cuckoo Search (CS) [33], and Big Bang-Big Crunch procedure (BB-BC) [34]. Han et al. [35] developed BFGSA, an improved GSA framework for analysis of clustering. The standard location of the globally best resolution's seven closest neighbors was employed to help the winner avoid local optimal pitfalls. BFGSA outscored many conventional search techniques on 13 UCI data sets. Nanda and Panda [36] provide a detailed survey of metaheuristic techniques for segmentation and grouping. The study [37] examines medical forecasting technologies and resolves the difficulties using four distinct machine learning approaches. The databases such as for early detection and Binary 012 were used in this research. The precision, recall, and accuracy of K-Nearest Neighbours and Random Forest algorithms are determined using these sets of data. Since SVM outperforms KNN and Logistic Regression. the results of this study will be improved, and it is useful to investigators involved in predicting diabetes studies and development if they utilize the proposed modified phenomenon of FA. A multi-strategy firefly technique with chosen ensembles (MSEFA) has been proposed in the study [38]. The approach used three innovative search techniques with distinct properties in the method's pool. Furthermore, the chosen ensemble concept is used to build an appropriate roulette decision approach. The approach may choose appropriate methods for searching at various stages of searching while coordinating the balance of tactics to achieve superior outcomes. The parameterized adaptive modification technique is also implemented to manage the slow reduction of stride size.

3. Firefly Algorithm

In this article now it is glamorised a few of the luminous properties of fireflies to create firefly-based algorithms, for that researchers have been using the three idealised bellow given rules which can also be applicable for the new Firefly technique that is going to be proposed. 1) All fireflies seem to be same sex so a firefly would be appealed to another firefly irrespective of gender. 2) Aattractiveness is directly proportionate to intensity so the less optimistic firefly will gravitate forward concerning the more optimistic brighter. The pleasant appearance is comparative to the light intensity and decrease as the spacing between them grows. When there is no one brighter than a specific firefly, then that will start moving at random. 3) The fitness function influences or decides the brightness of the firefly.

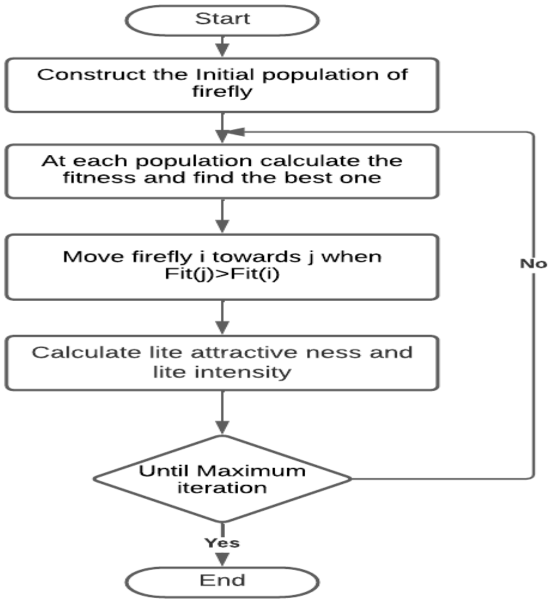

Fig 1: Basic firefly algorithm

The brightness of a maximisation problem could really be directly proportionate to the objective utility quantity. The fundamental actions of the firefly algorithm can be described as illustrated in Fig. 1 using the same three criteria.

3.1. Attractiveness:

The intensity of brightness I, of a firefly at a certain place x could be specified for the basic form in such a way that the intensity is propositional to the estimate of the objective function at x for maximization methods. Furthermore, attractiveness is subjective; it must be appraised by another firefly. As a result, it will change depending on the length from firefly i to firefly j. Further, because light strength diminishes with detachment starting from its source, and also light is immersed in the media, so it would enable the attraction to fluctuate through the amount of immersion. The intensity of the light changes with the distance r and the given medium that has a light absorption parameter ɣ. The cumulative influence of the inverse law of distance and absorption could be estimated by using Gaussian pattern shown below.

$I(r) = I_0 e^{-\gamma r^2} \ldots\ldots\ldots\ldots\ldots\ldots. Equation(1)$

Because a firefly's attractiveness is related to the brightness observed by neighbouring fireflies, it can now describe a firefly's attractiveness as

$\beta(r) = \beta_0 e^{-\gamma r^2} \ldots\ldots\ldots\ldots\ldots\ldots.. Equation\ (2)$

where β0 is the attraction at r = 0.

The detach involving every two fireflies i and j at x_i and x_j, correspondingly, is the Euclidean distance and can be calculated by using any distance measure like Euclidian, Manhattan excreta. The attraction of a firefly I to the next most interesting (brighter) firefly j is defined by.

$$x_j = x_i + \beta e^{-\gamma r_{ij}^2}(x_j - x_i) + \alpha\delta(rand - 0.5) * scale \ldots\ldots\ldots. Equation\ (3)$$

Where,

- β= attractiveness constant.
- ɣ =absorption parameter.

- α=Random ness strength.
- δ=Random deduction.
- r_{ij} = gap between fireflies i and j.
- rand= random number
- scale=modular distance involving the upper and lower bound values of a feature.

In the equation (3) the second part is partly concerned to attraction and the third part is randomization, where α is the randomization component. rand is a aimlessly generated number with a identical allocation in the range [0, 1]. Sometimes the sum of the two terms in the equation (3) may lie outside in the range of the domain values of the feature which causes to select an irrelevant centroid, to overcome from that the parameter scale is defined as modular distance between the upper and lower boundaries of the domain values of a feature. The component ɣ now distinguishes the difference of the attraction, but also its value is essential in deciding the rate of convergence and the behaviour of the firefly algorithm. In hypothesis it lies in [0, ꝏ], although in exercise, ɣ is decided by the specific section of the system to be improved, as a result, it usually ranges commencing 0.01 to 100 across most implementations. It's important to note that the distance r specified previously isn't restricted to Euclidean distances. Based on the sort of situation we are interested in; we may specify a variety of alternative distance measures in n-dimensional space.

4. Proposed Clustering Method

K-means is one of the prime unsupervised learning techniques for solving the cluster formation issue. The approach adopts a prime and uncomplicated technique for categorizing a provided records utilizing a predetermined number of groups (assuming k clusters). The key idea is to creäte k centers, one for every group. These centers must be advantageously positioned because numerous positions deliver differed conclusions. Consequently, the ideal option is to arrange them as farthest distant as reasonable. The subsequent phase is to correlate every tuple in a known data through the closest centroid. When there is no outstanding data, the first phase is accomplished, and the primary cluster age is finished. We must now again calculate k new centers as the barycentres of the groups established in the previous phase. When these k new centers have been created, we will need to rebind the identical data set elements to the nearest new center there has been created a cycle. Because of this cycle, we may note that the k centers gradually move their location until no further alterations are finished, or until the centers block moving. Ultimately, this approach seeks to minimize a fitness function known as the squared error criterion, which is specified as follows:

$$J(U) = \sum_{j=1}^{k} \sum_{i=1}^{n} \left(\|\mathrm{x_i} - \mathrm{v_j}\| \right)^2 \ldots\ldots\ldots\ldots Equation\ (4)$$

Where,

- $\|\mathrm{x_i} - \mathrm{v_j}\|$ is the detachment connecting $\mathrm{x_i}$ and $\mathrm{v_j}$.
- n is the number of data points in j^{th} cluster.
- k is the number of clusters.

The K-means method uses the Distance metric as the likeness criterion to divide D-dimensional dataset units into a pre-defined number of clusters. When contrasted to similarities towards other data points in all the other groups, distances between data points inside a cluster are the smallest. Tuples in the same cluster are linked by a single center tuple which indicates the cluster's center which is the mean of the data points in the cluster. The basic K-means method may be summed up as given bellow:

Consider X = $\{x_2, x_2, x_3 \ldots\ldots\ x_n\}$ is the data set with given tuples and C=$\{C_1, C_2, C_3, \ldots\ldots, C_k\}$ be the cluster centres.

Step 1. Pick k arbitrary cluster midpoints.

Step 2. Calculate by what means far away every data tuple is from the cluster middle.

Step 3. Assign the tuples to the group center with the smallest distance between it. and all another group midpoints and calculate $J(U)$ using equation (4).

Step 4. Using the following expression, recreate the new cluster center.

$$C_i = \left(\frac{1}{k_i}\right) \sum_{j=1}^{k_i} x_j \ldots\ldots\ldots\ldots Equation\ (5)$$

where, k_i signifies the number of tuples in i^{th} group.

Step 5. Recompute the distance between every tuple and the new clusters midpoints that are exposed and compute $J(U)$ using equation (4).

Step 6. Halt if there is no significant change in $J(U)$, otherwise continue from step 3.

4.1 Cluster optimization using firefly algorithm:

In the elementary Firefly procedure, the relocation, i.e., Eq. (3), is mostly influenced by the attraction of some another firefly; the attraction is a characteristic property of the inter-firefly distances. As a result, a firefly could be tempted to some other firefly simply because it is nearby, which could lead it to depart from the global minimum. A firefly is attracted to another firefly because their attractiveness helps to the movement. This conduct may cause a pause in the process of cooperation toward the global minimum. The concept underlying our Artificial Firefly Algorithm (AFA) is to employ local optimal intelligence such that each firefly is influenced by the attraction of a subset of fireflies rather than all of them. Depending on their local optima, this small proportion occupies the largest component of the fireflies. Therefore, a firefly is responding smartly by focusing its motion on adjacent fireflies as well as its attraction.

Figure 1 depicts a simple method for the AFA approach. The new optimization model for local optimal solution is the proportion of fireflies used in the move and is calculated utilising the equation presented in (6).

$$L_i(U) = \sum_{j=1}^{n} \left(\|\mathrm{x_j} - \mathrm{v_i}\| \right)^2, \mathrm{i} = 1 \ldots \mathrm{k},\ \ldots\ldots\ldots Equation(6)$$

Where,

- $\|\mathrm{x_i} - \mathrm{v_j}\|$ is the distance between $\mathrm{x_j}$ and $\mathrm{v_i}$.
- n is the number of data points in i^{th} cluster.
- k is the number of clusters.

Algorithm: Artificial firefly algorithm (AFA) for cluster optimization.

Let the notation is a set that holds centres after k-means clustering as initial fireflies along with the intensity of each firefly.

Input: Data set, $NCS, J(U), \boldsymbol{\epsilon}$.

Output: Optimal subset of cluster centers, best value of fitness function.

Step 1. Initialize the firefly population that is resulted from k: means NCS.

Step 2. Compute the intensity $L_i(U)$ at each $x_i \in NCS$ using equation (6).

Step 3. Consider s=0.

Step 4. Compute the $J^s(U)$ using equation (4).

Step 5. Define the absorption coefficient ɣ.

Step 6. $While((J^s(U) - J^{s-1}(U)) > \epsilon)$, begin

Step 7. For each new firefly $\{C_i/C_i \in NCS\}$ begin

Step 8. For each new firefly $\{C_j/C_j \in NCS,\ i \neq j\}$ begin

Step 9. $If(L_j(U) < L_i(U))$

Step 10. $Pre_i = L_i(U)$.

Step 11. Move firefly i towards j.

Step 12. Update each firefly with position update equation using equation (3).

Step 13. Find the clusters with respect to the updated fireflies (new cluster centers).

Step 14. Find the intensities of updated fireflies (new $L_i(U)$) using equation (6).

Step 15. $If(newL_i(U) \ < Pre_i)$

Step 16. Update NCS by new firefly along with intensities found in step 14.

Step 17. End of step 15.

Step 18. End of step 9.

Step 19. End of step 8.

Step 20. End of step 7.

Step 21. Evaluate $(J^s(U)$.

Step 22. s=s+1.

Step 23. End of step 6.

Step 24. Return the final fireflies as optimal cluster centers.

By establishing all the parameters as explained previously, the basic firefly method is preserved. In the basic FA algorithm, the strength is that the position of the finest firefly has no effect on the orientation of the searching. As a result, the fireflies are not caught in a local optimum. The searching for the global minimum, on the other hand, necessitates more computer work because numerous fireflies fly around in unimportant locations. In clustering optimization problem, the global optimum can be obtained by moving towards local optimum for individual clusters, that is maximizing intra cluster similarity which can be found by using the equation (6) and the global optimum for all clusters can be obtained by minimizing inter cluster similarity which will be found by using the equation (4). The artificial firefly algorithm that is used to optimize the clusters resulted from k-means clustering algorithm is presented in above algorithm.

5. Experimental Results

Experiments are conducted out on five data sources picked out from standard data set repository UCI [39], namely Iris, WDBC, Sonar, Glass, and Wine. The underlying qualities are stated with each of them:

Iris (Iris Plant Dataset): This data set is based on the identification of Iris flowers, which contains three separate groups, each one with 50 tuples. Each tuple is identified by four characteristics. **WDBC** (Wisconsin diagnostic breast cancer): This data set contains information regarding breast cancer which was gathered at the University of Wisconsin. There are two groups with 357 and 212 tuples respectively. Every tuple in this data collection has 30 attributes. **Sonar**: This collection of data contains 208 tuples of underwater sonar waves. Sonar sounds were separated into two groups in this data collection, with 111 and 97 samples each containing 60 characteristics. **Glass** (glass recognition dataset): This collection of data contains 214 tuples of various kinds divided into six classifications. each of these categories of data having 9 properties. **Wine** (wine identification dataset): This data source has 178 entries divided into three separate groups, which include 59, 71, and 48 observations, correspondingly. Every tuple in this data gathering has 13 properties.

To enhance the end results, in the starting phase of the suggested technique, we initialize centroids with k tuples of database, that pick arbitrarily seen between tuples of given dataset. As a result, the early centers will be one of the inputs and will not be beyond the data regions. In the second stage, we used the firefly technique to compute the best cluster center and used it to optimize the k-means clustering. firefly streamlines the centers until there is a minimal variation in the centroid objects after a few cycles.

Purity is an independent assessment criteria of cluster quality based on cluster analysis. It is the percentage of the overall number of entities (data points) properly categorized in the unit range [0...1].

$$Purity = \frac{1}{N}\sum_{i=1}^{k} max_j\left|c_i \cap t_j\right| \ldots \ldots \ldots Equation\ (7)$$

Where,

- N denotes the number of entities (data points),
- k denotes the number of clusters,
- c_i is a cluster in C, and
- t_j denotes the classification with the total number for cluster c_i.

The next assessment we perform is to compute the Ranking Index (RI), which quantifies the proportion of accurate decisions. The F-measure was used to estimate the validity of the clustering methods. In knowledge discovery, the F-measure is a composite of accuracy and recall values. Each cluster acquired must be regarded as the recognized by a class, but each pre-classified class may be regarded as a desired group of tuples for those results. We consider every cluster as if it was the consequence of a question, and every class as if it was the pertinent set of tuples of that class. To find the Precision we retrieve the standard class with the greatest set of entities allocated for each cluster. The largest number of points for each group is then added together and divided by the total number of clustered items. To find the recall value we retrieve the group with the greatest number of items allocated for each standard class. The highest number of items for each standard class is then added together and divided by the total number of clustered and uncluttered instances Finally the harmonic mean of precision and recall is the F1-score.

PSO algorithms have been shown in several studies to surpass genetic algorithms [40] and other traditional algorithms in tackling a variety of optimization issues. This is probably due to current best approximations' transmitting capabilities allows for greater and faster approach to optimality. Shilane et al. [41] provide a detailed approach for measuring the statistical effectiveness of optimization computation. For several typical test features, we will now contrast the Firefly Algorithm with PSO and K-Means clustering algorithm. Tables 1 to 5 demonstrate the purity and F-measure for PSO, K-means and planned procedure of AFA for various data sets such as Iris, WDBC, Sonar, Glass, and Wine.

Table 1: Comparison of performance metrics among dissimilar procedures for iris data set.

Algorithm	Purity	F-Measure
K-Means	0.83	0.76
PSO	0.84	0.78
AFA	0.85	0.80

Table 2: Comparison of performance metrics among different methods for WDBC data set.

Algorithm	Purity	F-Measure
K-Means	0.81	0.74
PSO	0.823	0.751
AFA	0.828	0.76

Table 3: Comparison of performance metrics among different methods for Glass data set.

Algorithm	Purity	F-Measure
K-Means	0.79	0.76
PSO	0.80	0.768
AFA	0.82	0.79

Table 4: Comparison of performance metrics among different methods for Sonar data set.

Algorithm	Purity	F-Measure
K-Means	0.83	0.75
PSO	0.838	0.759
AFA	0.845	0.764

Table 5: Comparison of performance metrics among different methods for Wine data set.

Algorithm	Purity	F-Measure
K-Means	0.83	0.79
PSO	0.836	0.81
AFA	0.843	0.82

It can observe that the AFA is significantly more well-organized in terms of locating global finest solution, with substantially greater performance. On a contemporary computer, each functional execution is almost speedy.

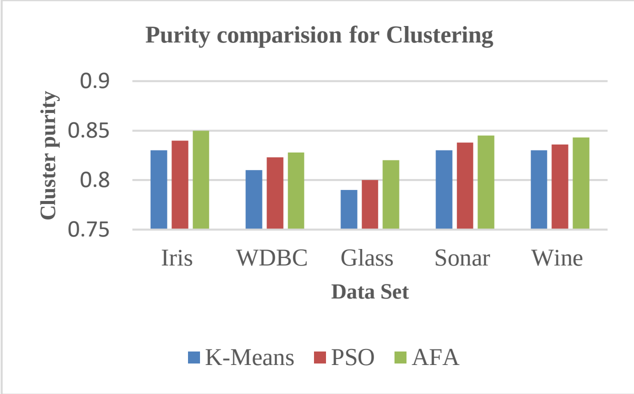

Fig 2: Comparison of Purity among Different methods for specified data sets.

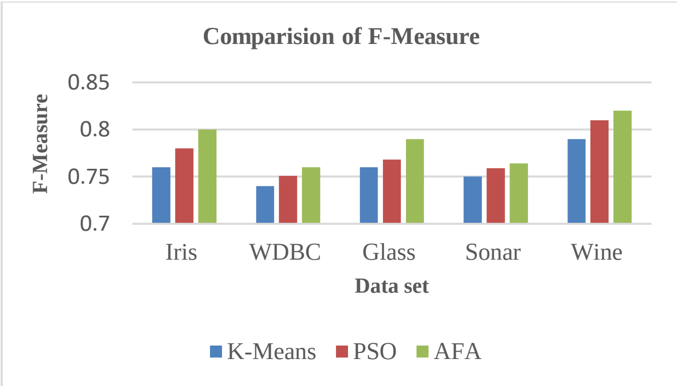

Fig 3: Comparison of F-Measure among Different methods for specified data sets.

The effectiveness of a clustering approach to reduce clustering error is one of its most essential characteristics. Apart from intra-cluster error, a good clustering algorithm must be talented to decrease clustering error and allocate the data to the correct group. On the Iris, WDBC, Sonar, Glass, and Wine data sets, it is displayed the gathering error of PSO, K-means, and the suggested technique in the following figure 2. The acquired results demonstrated the correctness and effectiveness of the suggested approach, as illustrated. When compared to previous approaches, this strategy might reduce clustering error in all circumstances.

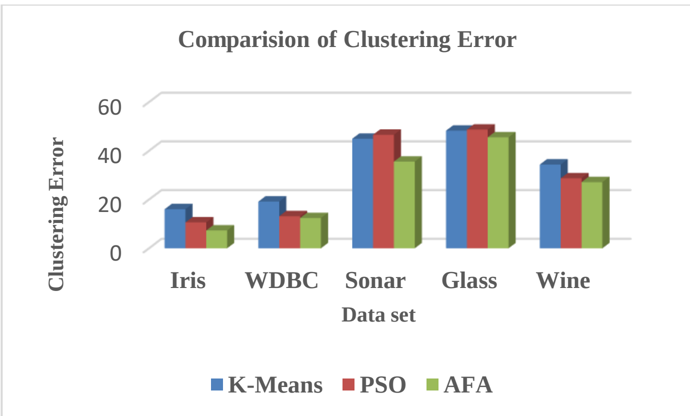

Fig 4: Assessment of Clustering Error among Dissimilar approaches for specified data sets.

6. Conclusion

In this study to cluster the data, a new hybrid method built on the firefly procedure and the k-means clustering method is suggested. In the suggested technique, we utilized the K-means method to locate ideal cluster centers, then used the firefly algorithm to enhance the centers using these centers as input. The developed novel firefly method is examined its commonalities and dissimilarities with particle swarm optimization and k-Means procedure in this research. These methods were then developed and compared, our computational results for determining the global optima of different testing functions show that particle swarm beats classic methods such as the K-Means method, while the innovative firefly approach exceeds both PSO and K-Means in terms of effectiveness and performance level.

The suggested method produced results that are reasonably robust in diverse performance, according to empirical results for optimizing fitness function linked to intra-cluster separation. In summary, experimental findings revealed that the projected algorithm outperformed PSO and K-means in terms of effectiveness. The Firefly Procedure may be extended more to address multi objective optimization issues as a reasonably simple addition. Furthermore, the use of firefly techniques in conjunction with several other techniques may provide an attractive topic for future study.

7. 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] J. Grabmeier, A. Rudolph, “Techniques of cluster algorithms in data mining”, Data Mining and Knowledge Discovery, vol. 6(4), 2002, pp. 303–360. doi: 10.1023/A: 1016308404627.

[6] H. Yu, Z.G. Liu, G.Y. Wang, “Automatically determining the number of clusters using decision-theoretic rough set”, in: J. Yao et al. (Eds.), Proceedings of the 6th International Conference on Rough Sets and Knowledge Technology, RSKT’11, LNCS, vol. 6954, 2011, pp. 504–513, https://doi.org/10.1016/j.ijar.2013.03.018

[7] Narayana G. Surya, and D. Vasumathi. “Clustering for high dimensional categorical data based on text similarity.” In Proceedings othe 2nd International Conference on Communication and Information Processing (ICCIP-2016) pp. 17–21. 2016.

[8] LNC. Prakash K, Dr. K. Anuradha “Optimal Feature Selection for multi valued Attributes using Transaction Weights as Utility Scale” proceedings of second international conference on computational intelligence and informatics (ICCII-2017). https://doi.org/10.1007/978-981-10-822 8-3.

[9] Rajasekhar A, et al. Computing with the collective intelligence of honey bees–a survey. Swarm Evolution Comput. 2017;32:25–48.

[10] Dorigo M, Stützle T. The ant colony optimization metaheuristic: algorithms, applications, and advances. Handbook of metaheuristics. Boston: Springer; 2003. p. 250–85.

[11] Kennedy J, Eberhart RC. Particle swarm optimization proceeding IEEE international conference of neural network IV. In: IEEE Service centre, Piscataway. 1995.

[12] Zhao S-Z, Ponnuthurai NS, Swagatam D. Dynamic multi-swarm particle swarm optimizer with sub-regional harmony search. In: IEEE Congress on Evolutionary Computation. IEEE, 2010.

[13] Yang X-S, Deb S. Cuckoo search: recent advances and applications. Neural Comput Appl. 2014;24(1):169–74.

[14] Yang X-S. Firefy algorithm, stochastic test functions and design optimisation. Int J Bio-inspired Comput. 2010;2(2):78–84.

[15] MacQueen, J. (1967) Some methods for classification and analysis of multivariate observations, Proceedings of the Fifth Berkeley Symposium On Mathematical Statistics and Probabilities, 1, 281-296.

[16] Morissette L, Chartier S. The k-means clustering technique: General considerations and implementation in Mathematica. Tutorials in Quantitative Methods for Psychology. 2013 Feb;9(1):15-24.

[17] A. Ito, "Successive Binary Partition K-means Method for Clustering with Less Cluster Size Bias," 2022 7th International Conference on Signal and Image Processing (ICSIP), Suzhou, China, 2022, pp. 772-776, doi: 10.1109/ICSIP55141.2022.9886452.

[18] Y. Li, Y. Zhang, Q. Tang, W. Huang, Y. Jiang and S. -T. Xia, "t-k-means: A ROBUST AND STABLE k-means VARIANT," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 3120-3124, doi: 10.1109/ICASSP39728.2021.9414687.

[19] T. N. Reddy and K. P. Supreethi, "Optimization of K-means algorithm: Ant colony optimization," 2017 International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2017, pp. 530-535, doi: 10.1109/ICCMC.2017.8282522.

[20] S. Kapil, M. Chawla and M. D. Ansari, "On K-means data clustering algorithm with genetic algorithm," 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), Waknaghat, India, 2016, pp. 202-206, doi: 10.1109/PDGC.2016.7913145.

[21] Prakash K., L.N.C., Surya Narayana, G., Ansari, M.D., Gunjan, V.K. (2022). Optimization of K-Means Clustering with Modified Spiral Phenomena. In: Kumar, A., Mozar, S. (eds) ICCCE 2021. Lecture Notes in Electrical Engineering, vol 828. Springer, Singapore. https://doi.org/10.1007/978-981-16-7985-8_126.

[22] Pasi Fränti, Sami Sieranoja, “How much can k-means be improved by using better initialization and repeats?”, Pattern Recognition, Volume 93, 2019, Pages 95-112, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2019.04.014.

[23] Rani, R. U. ., Rao, P. S. ., Lavanaya, K. ., Satyanarayana, N. ., Lallitha, S. ., & Prasad J, P. . (2023). Optimization of Energy-Efficient Cluster Head Selection Algorithm for Internet of Things in Wireless Sensor Networks. *International Journal on Recent and Innovation Trends in Computing and Communication*, *11*(4), 238–248. https://doi.org/10.17762/ijritcc.v11i4.6445.

[24] H. Wang, W. Wang, X. Zhou, H. Sun, J. Zhao, X. Yu, Z. Cui, Firefly algorithm with neighbourhood attraction, Inform. Sci. 382–383 (2017) 374–387.

[25] I. Fister, I. Fister Jr, X.S. Yang, J. Brest, A comprehensive review of firefly algorithms, Swarm Evol. Comput. 13 (2013) 34–46.

[26] A. Baykasoğlu, F.B. Ozsoydan, An improved firefly algorithm for solving dynamic multidimensional knapsack problems, Expert Syst. Appl. 41 (8) (2014) 3712–3725.

[27] A.H. Gandomi, X.-S. Yang, A.H. Alavi, Mixed variable structural optimization using firefly algorithm, Comput. Struct. 89 (23) (2011) 2325–2336.

[28] C.S. Sung, H.W. Jin, A tabu-search-based heuristic for clustering, Pattern Recognit. 33 (5) (2000) 849–858.

[29] S.Z. Selim, K. Alsultan, A simulated annealing algorithm for the clustering problem, Pattern Recognit. 24 (10) (1991) 1003–1008.

[30] U. Maulik, S. Bandyopadhyay, Genetic algorithm-based clustering technique, Pattern Recognit. 33 (9) (2000) 1455–1465.

[31] P. Das, D.K. Das, S. Dey, A modified bee colony optimization (MBCO) and its hybridization with k-means for an application to data clustering, Appl. Soft Comput. 70 (2018) 590–603.
[32] 27 T. Niknam, B. Amiri, An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis, Appl. Soft Comput. 10 (1) (2010) 183–197.
[33] 30 S.I. Boushaki, N. Kamel, O. Bendjeghaba, A new quantum chaotic cuckoo search algorithm for data clustering, Expert Syst. Appl. 96 (2018) 358–372.
[34] 36 A. Hatamlou, S. Abdullah, M. Hatamlou, Data clustering using big bang–big crunch algorithm, in: Innovative Computing Technology, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.
[35] X.H. Han, L. Quan, X.Y. Xiong, M. Almeter, J. Xiang, Y. Lan, A novel data clustering algorithm based on modified gravitational search algorithm, Eng. Appl. Artif. Intell. 61 (2017) 1–7.
[36] S.J. Nanda, G. Panda, A survey on nature inspired metaheuristic algorithms for partitional clustering, Swarm Evol. Comput. 16 (2014) 1–18.
[37] L. H.N., A. S. Reddy and K. Naidu, "Analysis of Diabetic Prediction Using Machine Learning Algorithms on BRFSS Dataset," 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2023, pp. 1024-1028, doi: 10.1109/ICOEI56765.2023.10125804.
[38] Hu Peng, Wenhui Xiao, Yupeng Han, Aiwen Jiang, Zhenzhen Xu, Mengmeng Li, Zhijian Wu, Multi-strategy firefly algorithm with selective ensemble for complex engineering optimization problems, Applied Soft Computing, Volume 120, 2022, 108634, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2022.108634.
[39] http://archive.ics.uci.edu/ml/
[40] Goldberg D. E., Genetic Algorithms in Search, Optimisation and Machine Learning, Reading, Mass.: Addison Wesley (1989).
[41] Shilane D., Martikainen J., Dudoit S., Ovaska S. J., A general framework for statistical performance comparison of evolutionary computation algorithms, Information Sciences: an Int. Journal, 178, 2870-2879 (2008).
[42] Yang, L. ., & Daimin, G. . (2023). Children's Perspective on Digital Picure Book: A Brief Analysis . International Journal on Recent and Innovation Trends in Computing and Communication, 11(3), 166–177. https://doi.org/10.17762/ijritcc.v11i3.6336
[43] Thomas, C., Wright, S., Hernandez, M., Flores, A., & García, M. Enhancing Student Engagement in Engineering Education with Machine Learning. Kuwait Journal of Machine Learning, 1(2). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/123