SEMARAK ILMU
PUBLISHING
202103268166(003316878-P)

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

K-Means Hybridization with Enhanced Firefly Algorithm for High-Dimension Automatic Clustering

Afroj Alam[1,*], Muhammad Kalamuddin Ahamad[1]

[1] Department of Computer Application, Integral University, Lucknow, 226026 India

ARTICLE INFO

Article history:
Received 21 July 2023
Received in revised form 17 October 2023
Accepted 29 October 2023
Available online 15 November 2023

Keywords:
K-Means clustering; Firefly algorithm; Opposition and Dimensional-Based Modified Firefly Algorithm (ODFA)

ABSTRACT

K-means clustering is not able to select the right number of clusters of data items having high-dimension. So, for determining the ideal number of clusters, we have combined PCA with the Silhouette and Elbow approaches. Additionally, we have a large number of meta-heuristic swarm intelligence algorithms which is influenced by nature and were previously used to solve the automatic data clustering problem. Firefly offers reliable and effective automatic data clustering. The Firefly algorithm automatically divides the entire population into subpopulations, which slows down the convergence and reduces the likelihood of capturing local minima in high-dimensional optimization problems. Thus, for automatic clustering, we demonstrated an improved firefly, i.e., we offered a hybridized K-means with an ODFA model. The experimental section displays the results and graphs for the Silhouette, Elbow, and Firefly algorithms.

1. Introduction

Clustering is an important, influential, and unsupervised machine-learning technique for naturally grouping objects into groups according to their similarity [1]. Objects will be segmented into different clusters based on some cluster validity index (CVI) metrics. This CVI shows the relation between cluster cohesion i.e., intra-cluster distance within a group and cluster separation i.e., the inter-cluster distance between groups. This process of partitioning the objects is called clustering. These objects do not have any external information, such as class labels; that is why clustering is an instance of unsupervised machine learning classification techniques [2]. Data clustering is very useful for partitioning the data so that we can make exploratory data analysis to find hidden patterns and valuable information from high-dimensional data [3].

Nowadays, the clustering technique is being applied in diverse fields of real-world problems, e.g., in engineering, wireless sensor networks, mobile networks, medical science, computer science, biological science, earth science, economics, and bioinformatics, yield-marketing, image analysis, web mining, spatial database analysis, insurance, statistical data analysis, fraud detection, libraries (book ordering), loan approval, community detection (e.g., LinkedIn, Facebook, Twitter, etc.), pattern

[*] *Corresponding author.*
E-mail address: alamafroj@student.iul.ac.in

https://doi.org/10.37934/araset.33.3.137153

recognition, data compression, classification of plants and animals, improving decision-making in business intelligence, etc. [4-7]. At the start clustering approach was used mainly in two domains i.e., anthropology and psychology, which were the two social science areas. Further with advancement, it's implemented also in trait theory. After that, it extended to other new study directions with major impacts like machine learning and data science. As a result, data clustering is a significant and hot topic in many domains. E.g., in artificial intelligence as well as in data mining [8,9].

Researchers have proposed several partitioning-based heuristic algorithms from the last 2-3 decades to solve the clustering problems. Among all partitioning clustering algorithms, the K-Means method appeared as the most prominent and widely used and powerful algorithm for selecting the optimum number of output clusters in a dataset. Because it is straightforward to implement and flexible, its run time complexity is less than other clustering algorithms; it is also rated under the top 10 data widely used in data mining [10]. K-means clustering is a non-deterministic method for dividing n data points into K non-overlapping clusters in an N-dimensional space. There should be not any fuzzy data points in any cluster. However, there are a lot of major challenges with K-means clustering, such as it highly relies on a predefined appropriate number of clusters K for datasets having high dimensional. This leads the algorithm to get trapped in local optima. Defining and initializing the appropriate number of clusters and their corresponding centroids the limitation of performance and their accuracy of the quality of good clusters [11]. Also, K-means clustering is NP-hard in the high dimensional dataset; there is no widely accepted theory of K-means for all types of datasets. The above drawbacks prompted data mining experts to develop new methods to address them and come up with other effective means to enhance the K-Means. To address the shortcomings mentioned earlier and tackle more sophisticated and high-dimensional data clustering challenges, the data mining researcher redefines the K-means, which is hybridized with meta-heuristic optimization methods such as swarm intelligence algorithms [12,13].

There are a lot of stochastic meta-heuristic multi-model evolutionary algorithms which is inspired by nature. These multi-model algorithms have been used to solve the challenge of data clustering, which are differential evolution (DE), Tabu search algorithm, and genetic algorithm (GA). At the same time, there are several swarm intelligence-based techniques such as symbiotic organisms search (SOS), PSO, ant colony optimization (ACO), invasive weed optimization (IWO), firefly algorithm (FA), Cuckoo algorithm (CA), ABC, teaching learning-based optimization (TLBO) etc. all these algorithms integrated with clustering to solve the clustering problems. For Clustering analysis swarm intelligence and evolutionary algorithms both act as an optimization catalyst in global search space to maximize the cohesive process inside the clusters and maximize the adhesive process for separating the clusters. The above two stochastic meta-heuristic methods are far better than traditional clustering in high-dimensional data clustering problems. These nature-inspired stochastic optimization algorithms have been widely used in different fields for solving real-life optimization problems. GA is used to solve high-dimensional clustering problems, location problems, and flow shop problems. Differential Evolutionary (DE) algorithms have been widely used to answer computer science problems and engineering optimization complex problems. The PSO is also successful in several high-profile applications in terms of high accuracy and good convergent speed [14,15]. It has been observed that using NIC not only causes automatic clustering problems to solve, but also improves the accuracy, efficiency, and robustness of the algorithms if we hybridize them with another traditional algorithm [16,17].

However, for the majority of the above stochastic meta-heuristic optimization algorithms, we have to prior specify the value of K for the number of clusters as well as the structure of data and their parameters in K-means. Unfortunately, finding the optimal value of K in advance for KM clustering is a challenging task in high-dimensional datasets. In many successful applications, the FA is another well-known popular meta-heuristic optimization technique that has been utilized to solve real-life complex application, challenges of data clustering, and scheduling problems which is not related to parallel machine execution [18]. Hence, our research idea proposed a hybridized K-means clustering with an enhanced FA method; this enhanced FA not required prior input. It will automatically calculate K i.e., number of clusters in this manner. This automatically calculated value of K by FA will help the k-means algorithm to converge to the global optima and generates a quality of clusters; this is our proposed methodology.

Further, the outline of our paper of the remaining part is as given below: In section II we have described briefly and scientifically the exhaustive literature review on the k-means clustering algorithms, meta-heuristic techniques for global optimization, and procedure for automatic clustering. Section III elaborates on the K-means clustering. Section IV elaborates on the Firefly Algorithms and its variants design concept. Section V gives a detailed description of the proposed algorithm. In Section VI, Experimental results are done, and Section VII concludes the paper.

1.1 Literature Review

The researcher Indra Kumari *et al.,* [19] in his paper proposed an improved K-means that will not generate any empty clusters in K-means. They also showed in his work that the upgraded k-means cluster has a lower running time complexity than a conventional k-means cluster. They used the concept of the backtracking method algorithm, by which the algorithm takes the help of previously calculated data for calculating the new centroids. For this algorithm, the future challenge and work is the space minimization of this algorithm.

Islam *et al.,* [20] have proposed a more efficient nature-inspired algorithm, i.e., Genclust, which combines the capacity of a genetic algorithm to conglomerate different solutions with the exploitation of a hill climber for the entire search space. The researcher has improved the traditional genetic searching approach by using the concept of hill-climbing concepts of K-means, which is faster than the traditional clustering and gives the results of higher quality clusters, and also reduces the computational resources. A researcher has also tested their algorithm on different datasets.

According to the current researcher Alam *et al.,* [21], K-means clustering is more general and easier for the prediction of disease in the field of health care. This algorithm is also prominent for the detection of fraud in the bank, detection of crime, yield management, and profitability analysis. This algorithm also applies to international super-market for predicting frequent item sets for customer attraction. We need some suitable data mining techniques like K-means, which are hybridized with metaheuristic algorithms to improve the accuracy of real-life NP-hard optimization problems.

Shi *et al.,* [22] have improved the effectiveness of data analysis in their proposed hybridized algorithm over the classical K-means algorithm, which was easily convergent to the local optima by the initial selection of cluster centroids. The proposed hybridized algorithm is a genetic algorithm and the K-means. In this paper genetic algorithm is used for cleaning and reducing the dimensionality of the datasets, then uses the K-means to optimize the selection of cluster centroids. Three different approaches they have tested are standard K-means, K-means hybridized with GA, and population-based GA with K-means clustering. The objective of all three approaches is to automatic selection of initial cluster centroids. The proposed method gives more accuracy.

The Analyst Hrosik *et al.,* proposed a hybridized approach for the division of the brain picture for recognizing diverse essential tumours. The proposed idea placed a strong emphasis on tracking various primary brain tumours: glioma, metastatic bronchogenic carcinoma, metastatic adenocarcinoma, and sarcoma. This method is analysed with standard benchmark digital images, and it improves the quality of cluster results compared to other simple K-means clustering methods [23].

The researcher has proposed an image analysis for MRI pattern recognition in brain tumour with K-means as unsupervised machine learning and nature-inspired-based PSO and FA. They improved the image division using the fitness function of Swarm-Based PSO. The accuracy of the segregation of images is improved with the help of KM and PSO. The Comparative studies have shown that the combined k-means with FA exhibited high accuracy and precision in detecting brain tumour Region-of-Interest [24].

2. Methodology

2.1 K-Means Clustering

The K-means clustering algorithm is an unsupervised machine learning technique that divides data into K clusters based on inter-cluster and intra-cluster distances. The objective of this algorithm is to minimize the intra-cluster, i.e., Euclidean distance should be minimized to get good-quality clusters. This technique assumes that a data object belongs to one of two clusters: one or none. It has been one of the most extensively utilized clustering algorithms for tackling numerous real-life situations due to its simplicity, ease, and linear time running complexity [25]. We can explore this algorithm in such a way that we have T sets of data, and we have to segregate it into K sets with non-overlapping. S={$s1_1$,$s2_2$,s_3,.....s_k},s_j≠ Φ ,j=1,2,...,k where T=U_1^k s_j; s_j ∩s_i= Φ, j,i=1............k and j≠i. During the partitioning process, we have to optimize the fitness function, i.e., Minimize intra-cluster distances between objects within the cluster while maximizing inter-cluster distances between objects between clusters [13], and this is the objective function of K-means clustering, which we have to achieve for good quality of clusters. The objective function can be defined mathematically as follows.

The given dataset T={t_1,t_2,t_3,............,tn} has a d-dimension. T is being isolated for k sets as given below,

$$P(s_k) = \sum_{s_i \in t_k} \left\| s_i - \mu_k \right\|^2 \tag{1}$$

So that the square errors sum objective function should decline across all k clusters. i.e., minimize the given equation.

$$P(t) = \sum_{k=1}^{K} \sum_{s_i \in t_k} \left\| s_i - \mu_k \right\|^2 \tag{2}$$

The main focus of automatic clustering is to identify the optimal value of k, and the distribution of data objects should be correct in all the clusters. As a result, automatic clustering, we must optimize the number of choices used to allocate R data objects to K groups.

The optimal value of K is being determined mathematically in the search space as below:

$$D(R) = \sum_{K=1}^{r} P(R, K) \tag{3}$$

Here is P is the search space. To find an optimal solution this is an NP-hard issue with K>3. The computational time of a task with a high-dimensional and huge dataset is very high. It is challenging to perform automatic clustering for such datasets without prior knowledge of the data items' characteristics. Hence creating an appropriate no of clusters and distributing the objects into their corresponding clusters is computationally exhaustive and time-consuming in traditional K-means clustering [35].

2.2 Principal Component Analysis (PCA)

Recent methodologies of KM for data having higher dimensions use deep learning methods to select attributes which are highly informative of a given data set. K-means clustering will then be applied to the reduced data set. Boutsidis *et al.,* [36] proposed an algorithm that helps to decrease the dimension to t from q by selecting a small subset of m rows from the data matrix D ε R^{qXn} where t<q. Then, on data matrix D, apply the KM approach to cluster the data objects into K groups. In consideration of variance-covariance for good clusters in a large dataset, the authors executed PCA which will include only informative attributes [37].

With the help of PCA, we can filter out irrelevant features, which reduces the training time as well as cost, and improve the performance of a given model. After PCA implementation the dataset will be passed to the K-means clustering to reduce the outlier's data from the clusters. [38].

PCA is a technique for dimension reduction that filters out uncorrelated variables from high dimensional data which do not adequately explain the original variables.

PCA is a technique for dimension reduction that filters out uncorrelated variables from high dimensional data which do not adequately explain the original variables.

Suppose the matrix Y^T=[Y_1,Y_2,Y_3,….., Y_p] and the their matrix Σ with eigenvalues λ_1>=…… λ_p>=0 is covariance

$$X_1=a_1^T Y=a_{11}Y_1 + a_{21}Y_2 +\ldots\ldots. a_{p1}Y_p$$
$$X_2=a_2^T Y=a_{12}Y_1 + a_{22}Y_2 +\ldots\ldots. a_{p2}Y_p$$
$$X_p=a_p^T Y=a_{1p}Y_1 + a_{2p}Y_2 +\ldots\ldots. a_{pp}Y_p$$

All of these equations can be replaced by X_i, the i^{th} principal component in the given below equation.

$$Variance(X_i) = a_i^T \sum a_i = e_i^T \sum e_i; \tag{4}$$

$$Co-Variance(X_i, X_k) = a_i^T \sum a_k = e_i^T \sum e_k; \tag{5}$$

The linear combination with the greatest variance is the first principal component. In other words, it maximizes Var (Y^1) = a_1^T Σa1. It is Var(Y1) = a_1^T Σa1 can clearly be increased by multiplying any a_1 by some constant. Hence first and second principal components in Eq. (7) and Eq. (8).

$$PC1=\begin{cases} Maximize: Variance(X_1) \\ Subject\ to\ a_1^T a_1 = 1 \end{cases} \tag{6}$$

$$PC2=f(x) = \begin{cases} Maximize: Variance(X_2) \\ Subject\ to: a_2^T a_2 \\ CoVariance\ (X_1, X_2) \end{cases} \tag{7}$$

In the next stage of our algorithm, PC1 and PC2 will be used as input for our KM method. Below is the Python algorithm for PCA.

```
import the iris dataset
import PCA from sklearn
include numpy as np1
def plt_clustered_out(X, centers):
pca = PCA(2).fit_transform(np1.concatenate([X, centers], axis=0))
true_centers = np.array([np1.mean(iris_data[..], axis=(0)) for label in set(iris_labels)])
plt_clustered_out(iris_data, true_centers
```

The clustering graph of clustering of iris data after PCA implementation is in the experimental part of Figure 1.

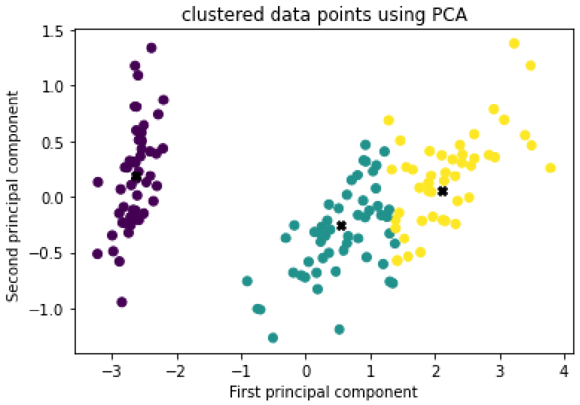

Fig. 1. Clustered data after applying PCA

2.3 Various Approaches to Identify the Value of k in Extended K-Means

Without the use of deep learning, prior to finding the optimal value of K in KM is a challenging task, hence KM will converge to the local optimal solution. Different researchers used different approaches. Some of them are explained given below:

i. *Silhouette method:*
The silhouette method is a clustering validation index tool for optimal cluster numbers. This method provides the graphical representation of intra as well as inter-cluster distances within its clusters and data points. The Individual silhouette coefficient index sc (k) value for a data point i is defined as a ratio scale data which is explained in Eq. (7) and Figure 2.

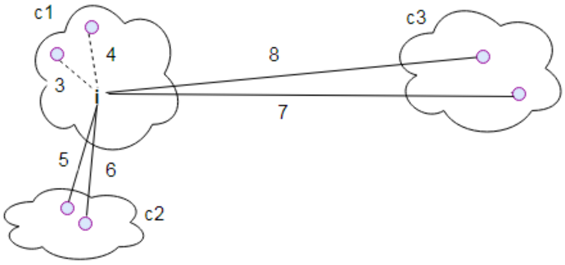

Fig. 2. Example of the Silhouette coefficient

$$sc(i) = \frac{q(k) - p(k)}{\max\{p(k), q(k)\}} \tag{8}$$

p(k) = mean value of dissimilarity of data point k to all other data points of c1.
p(k)=(3+4)/2=3.5
q(k) = minimum average dissimilarity between data point i to all other clusters c_2 and c_3
q(k)=min((8+7)/2,(5+6)/2)=5.5

The output of Eq. (3) is in between -1 and +1 i.e. -1 <= sc(k) <=+1.

+1 of s(i) indicates that data points are denser and data are correctly clustered. If its value is near -1 means bad clustering, then data point k would be more appropriate if it would be clustered in its neighbouring cluster. 0 shows an intermediate cluster. Eq. (7) Calculates the intra-inter Silhouettes value for a given data point.

Hence, larger silhouettes are well clustered as compared to small silhouette statistic values between clusters. The average value of sc(k) shows how data are tightly dense in the cluster and also it measures how the entire data has been clustered correctly by the given below equation [39].

$$\overline{SC} = \frac{1}{n}\sum_{k=1}^{n} SC(k) \tag{9}$$

Algorithm is given below for the Silhouette index

```
m = KMeans(random_state=1) //m is model
v = KElbowVisualizer(m, k=(2,10), metric='silhouette')
v.fit(X_numerics) // v is visualizer
v.show()
plt.show()
```

Experimental results for the above method are given in Figure 3. In the experimental part. Silhouette score method indicates the best options would be 5 or 6 clusters in Figure 3.

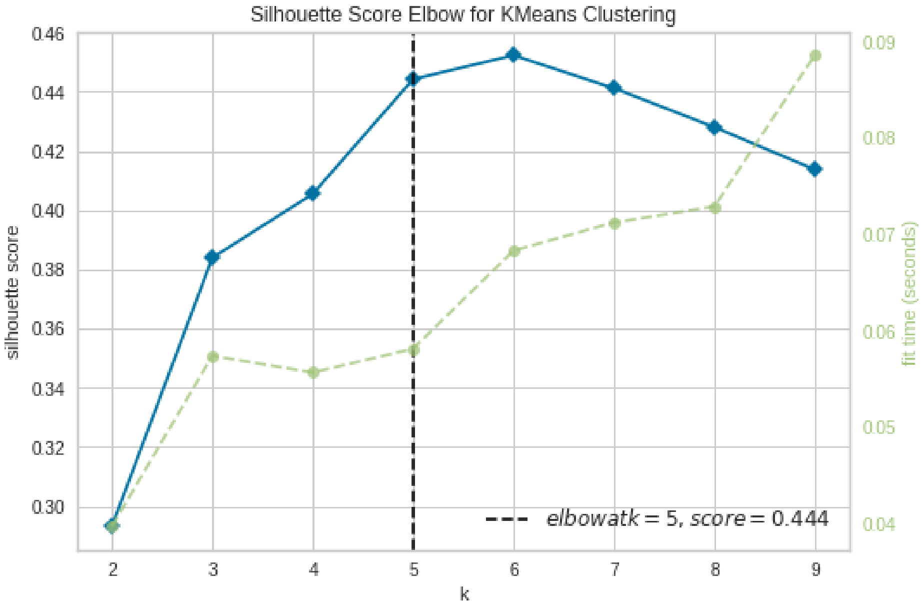

Fig. 3. Silhouette Score and Elbow value for K-means

Using Silhouette, we can also check the quality of clusters. The code is given below, and the experimental analysis results are in the experiment part in Figure 4.

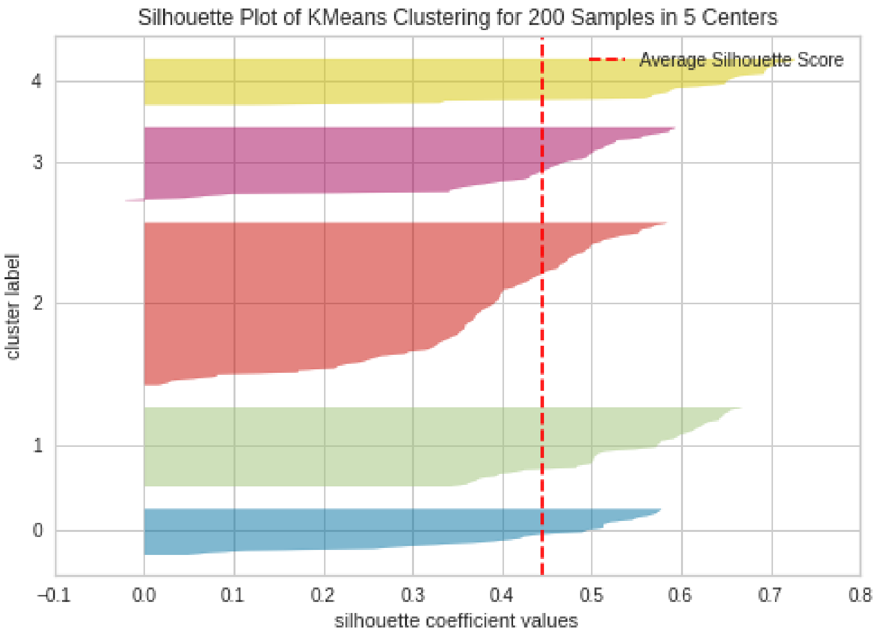

Fig. 4. Quality of Cluster using Silhouette method

```
from yellowbrick.cluster import SilhouetteVisualizer
m = KMeans(nc=5, rs=0)
v= SV(model, colors='greenbrick')
v.fit(X_numerics)
visualizer.show()
plt.show()
```

here m is model, v is visualizer, SV is Silhouette Visualizer

ii. *Elbow method*

It is a graphical and old approach for finding the optimal number of K in the K-means algorithm. It uses the concept of sum of squares error (SSE) as an objective function for clustering validity as well as cluster quality. SSE is defined as follows.

$$SSE = \sum_{k=1}^{K} \sum_{x_i \in s_k} \left\| X_i - C_k \right\|_2^2 \tag{10}$$

with K= number of clusters, Sk is the kth cluster, xi is the element of kth cluster, Ck is the centroid of the cluster Sk, || · || is the Euclidean distance between two data patterns.

We have to plot the graph of the line chart between SSE and their corresponding cluster value K. K starts with 2 and it will increase by 1 in each step. If the graph of the line chart shows a drastic decrease in SSE i.e., like an arm, then the “elbow” on the arm is a value to indicate the appropriate number of cluster k in K-means clustering [40].

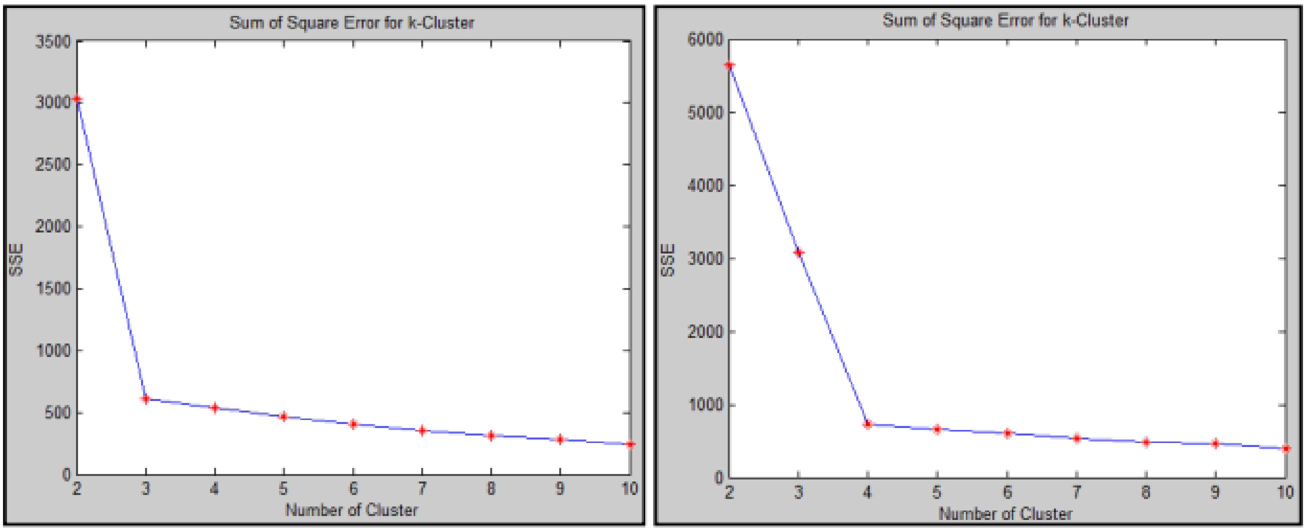

Fig. 5. Appropriate number of clusters in the graph by the relationship between SSE and Number of clusters [41]

For optimal clusters, the Algorithm of the Elbow method is given.

```
from yellowbrick.cluster import KElbowVisualizer
m = KMeans(random_state=1)
v = KElbowVisualizer(m, k=(2,10))
v.fit(X_numerics)
v.show()
plt.show()
```

The Experimental results for the above method are shown in Figure 6.

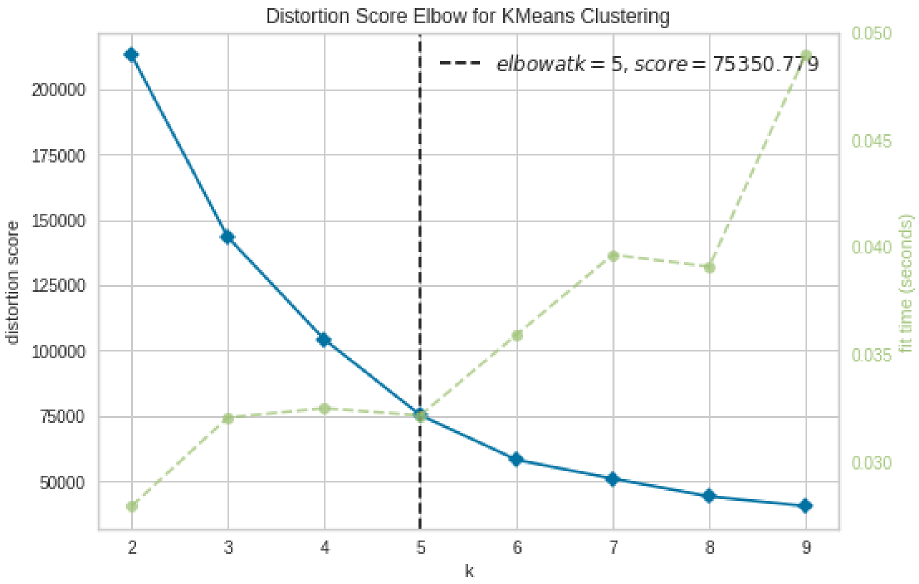

Fig. 6. Elbow method for optimal cluster

Python code is given below for clustering the data after getting an optimal number of clusters:

```
m = KMeans(n_clusters=6, random_state=0)
v = SV(model, colors='greenbrick')
```

```
v.fit(X_numerics)
v.show()
plt.show()
```

The Experimental results for the above method are shown in Figure 7.

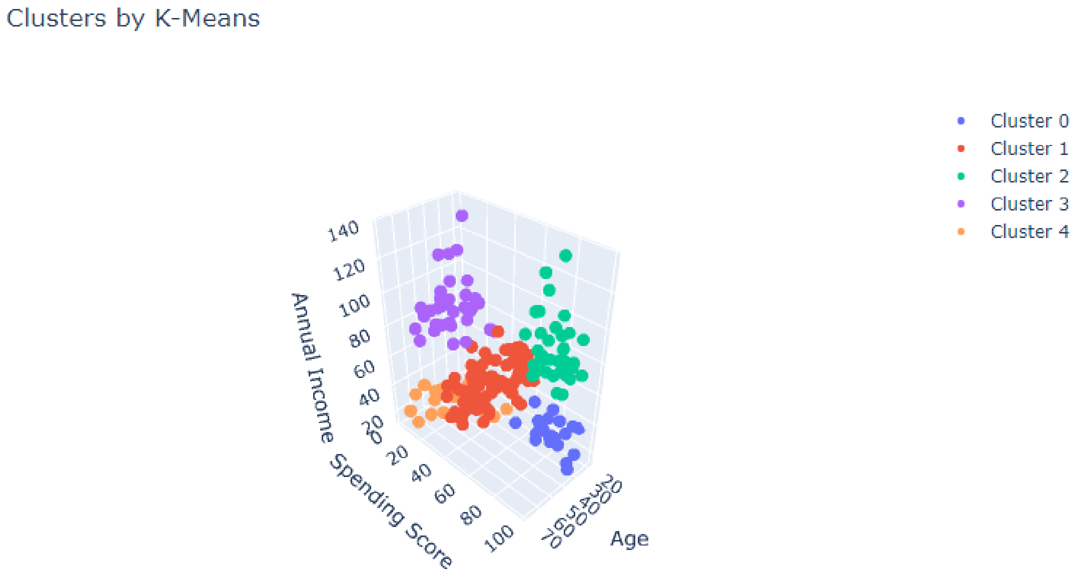

Fig. 7. Data clustered in 6 clusters using K-means

2.4 Firefly Algorithm (FA)

Xin-She Yang gave the concept of FAin, as a meta-heuristic optimization-based technique. The FA is based on the attraction behaviour of tropical fireflies and the flashing patterns of their idealized behaviour. Since 2010, the FA algorithm has been implemented in different real applications for solving different optimization tasks [42]. The attractiveness of fireflies and variation in light intensity is the important factor of fireflies [43].

The Firefly algorithm borrowed 3 ideas from the firefly behaviour for a making mathematical model of the algorithms.

i. Because all fireflies are unisex, their attractiveness is not determined by their gender.
ii. All Fireflies attract each other on the basis based on their fitness function of the brightness of the light. Fireflies having the worst fitness function move towards better fitness function fireflies.
iii. A firefly's brightness and the landscape of the objective function are directly connected.

There are two main tasks we have to define for designing a standard FA, i.e., formulation of light intensity and variation of attractiveness.

As we know that the intensity of light is inversely proportional to the distance from its source, so it can be approximated as follows. i.e.,

$$I = I_0 e^{-\gamma r^2} \tag{11}$$

where i_0 denotes the original light intensity at r=0 at the specified source, and γ is the constant coefficient of light absorption.

Now, we can define the attractiveness of FA according to the intensity of light proportion which is to be viewed by fireflies adjacent as given below. i.e.,

$$\beta = \beta_0 e^{-\lambda r^2} \tag{12}$$

where β 0 denotes light's attraction at r = 0, i.e., its greatest attractiveness. The following is a description of the Cartesian distance rij between firefly i and firefly j:

$$r_{ij} = \left\| x_i - x_j \right\|_2 = \sqrt{\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2} \tag{13}$$

where d is the number of dimensions [44]. The movement of firefly i towards firefly j by their respective brightness is as follows:

$$x_i^{t+1} = x_i^t + \beta e^{\gamma . r_{ij}^2} (x_j^t - x_i^t) + \alpha^t \varepsilon_i^t \tag{14}$$

New solutions of a firefly i depend upon the previous location x_i ones, which is represented by the following equation.

The second part of the above equation is the attraction, whereas, in the third term, α is the randomization parameter.

Methods of FA used for clustering are given below with output:

```
firefly = FireflyAlgorithm(d=d_iris, n=n, range_min=range_min, range_max=range_max,alpha=1.0, beta_max=1.0, gamma=0.5
```

Output:

```
firefly {'d': 12, 'n': 150, 'range_min': -5.0, 'range_max': 5.0, 'alpha': 1.0, 'beta_max': 1.0, 'gamma': 0.5}
```

Firefly-Algorithm on xclara dataset for automatic clustering

```
data_frame = pd.read_csv('../input/testverim/xclara.csv',index_col=False)
cols = [1,2]
data = data_frame[data_frame.columns[cols]]
len_data_points = data.shape[0]
print("Number of data points : {}".format(len_data_points)
print("v1 max : {}, v1 min : {}, v2 max : {}, v2 min : {}".format(v1_max,v1_min,v2_max,v2_min))
```

Graph of clusters using Firefly on **xclara** data is presented in Figure 8.

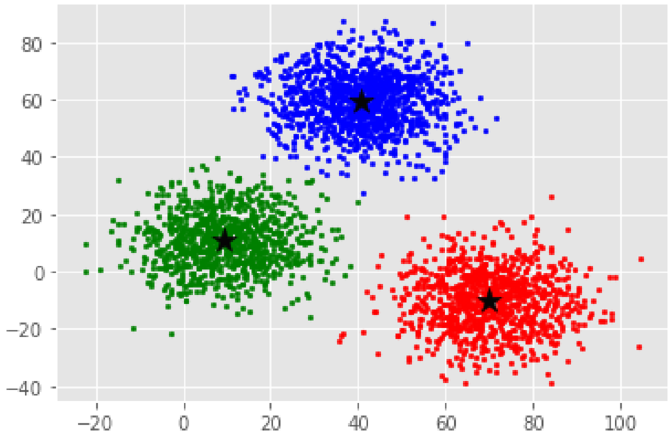

Fig. 8. Data is clustered in 3 clusters using Firefly algorithm

2.4.1 Opposition and dimensional based FA

Standard FA trap to local optima in high-dimensional data. A meta-heuristic is an excellent algorithm if it effectively balances the intensification and diversification components of the algorithms in the entire search space for getting better performance. Our proposed Opposition-based FA balance the intensification and diversification functions in every generation and fine the global best solution in each dimension, which leads to optimization and getting the global optima.

```
Algorithm:  Opposition and dimensional-based FA
Whole Firefly population X= {$X_1$, $X_2$, $X_3$,……$X_n$} is divided into  two groups i.e. $X_i$ and  ~$X_i$, where ~$X_i$ is the opposite population of $X_i$ and m<=X<=n
Initialize the population $X_i$={$X_{i1}$,$X_{i2}$,$X_{i3}$,…. $X_{iD}$} and  ~$X_{ij}$=$m_j$+$n_j$- $X_{ij}$  and 1<=i<=N and 1<=j<=D
evaluate the fitness function of each firefly by ,
f(Xi)=f($X_{i1}$,$X_{i2}$,$X_{i3}$,…. $X_{iD}$)
Intensity of light of each firefly is equal to f($x_i$)
 do
  Loop i=1 to D(dimensions)
       Loop j=1 to N(Firefly)
                 Y:=$G_{best}$
                 Y:=X($_{j,i}$)
             End loop
End loop
  Loop i=1 to N
By applying the supplied formula, you may update the global best in the entire population:
```

$$r_{ij} = \left\| x_i - x_j \right\|_2 = \sqrt{\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2}$$

attractiveness of firefly is determining by

$$\beta = \beta_0 e^{-\lambda r^2}$$

```
Update the fireflies with global best

$x_i^{new} = x_i^{old} + \beta_0 e^{-\gamma r_{ij}^2}(x_i^{old} - G_{bestpos}) + \alpha(rand - \frac{1}{2})$

  End loop
while(t < MaxGenerationFA)
End.
```

2.4.2 Proposed clustering approach based on Opposition Modified Firefly Algorithm (ODFA) model

In this novel clustering model, there is less probability of centroid initialization sensitivity problem and trapping to local optima problem in comparison to the standard K-means clustering algorithm. For good quality clusters in K-means, we have to consider some important points: The selection of feasible cluster centroids from the given dataset at the start and the ability to find global convergence rather than local convergence. As we know, K-means clustering is highly sensitive to the selection of cluster centroids, with parameter k for several clusters at the start. To get good quality clusters, we have to apply K-means clustering many times with random initialization of cluster centroids, which leads to bad convergence results. Our proposed method integrated the K-means with Opposition-based FA because FA used more factors than other nature-inspired optimization algorithms. During the optimization phase, these factors in FA algorithms include particle position, distance, light intensity, and velocity. We have provided more factors in FA that provide better results compared to another optimization algorithm. We divided the clustering task into two steps in our suggested technique; in the 1st step, we employed the firefly algorithm to discover good centroid locations. The objective function in the early steps is to keep the Euclidean distance to a minimum. The number of cores available in the system is the starting point for the method, which starts with the ideal number of clusters. Our Opposition-based FA finds the G_{best} solution in every dimension till predefine iteration. In the second step, the initialization of cluster centroids in K-means will be the G_{best} of all the fireflies.

3. Experimental Results

Experiments were carried out on Iris data sets, the xclara dataset, a customer dataset chosen from the standard data set, UCI, and the Kaggle data camp. We carried out this experiment in Python using the Google Colab online tool. We used principal component analysis to assess the quality of clusters, as shown in Figure 1. We also used the Silhouette score and the elbow method in our paper to determine the optimal number of clusters in K-means. Figure 5 and Figure 6 show the results. We have also plotted the Silhouette graph in Figure 7 with their coefficient values and their cluster labels for checking the quality of clusters and their Python code is mentioned in the above section III part.

We also wrote Python code for K-means with Silhouette Visualizer() to produce high-quality clusters. The silhouette graph provided us with the value 6 for K as an input. Figure 8 depicts a Clustered Graph of K-means. We also used the Xclara dataset to implement the Fire Fly algorithm for automatic clustering in Python programming. Section IV of this paper contains the code. The results of the experiments are shown at the end of this section.

```
Output of Firefly Algorithm:
Number of data points : 3000
v1 max : 104.3766, v1 min : -22.49599, v2 max : 87.3137, v2 min : -38.7955
```

Firefly Matrix:

```
[[[77. 6.]    [7. 178.]          [94. 180.]        [[195. 132.][5. 173.]    [156. 55.]]
[[114. 151.]            [150. 61.]        [197. 67.]]
[[69. 132.] [22. 35.]              [187. 66.]]
[[191. 54.]   [133. 42.]       [125. 43.]]]
[0. 605.97229931 582.9588196 355.98673507 530.20899625]
Best firefly matrix:
[[87.63257421 82.49212689]
 [-6.37814739 72.15306479]
 [-0.848309   79.40971558]]
Initial Centroid Values:
[[87.63257421 82.49212689]
 [-6.37814739 72.15306479]
 [-0.848309   79.40971558]]
[[69.92418447 -10.11964119]
 [9.4780459   10.686052  ]
 [ 40.68362784  59.71589274]]
```

4. Conclusions

Traditional K-means clustering has a lot of challenges, such as being highly reliant on the predefined appropriate number of clusters K. This algorithm also sticks to local optima and it is NP-hard in high-dimensional data. So, we have implemented the Silhouette and the Elbow methods with PCA to solve the problem of predefining the appropriate number of clusters K in K-means. We have also hybridized the traditional K-means with a nature-inspired meta-heuristic Firefly optimization algorithm to solve the problem of automatic clustering. The FA divides the entire population into sub-modules automatically by attracting the firefly's flash intensity. This is the unique feature of FA as compared to other swarm-based intelligent algorithms. Due to these unique features, we have explored FA and its variants in our paper. Our proposed opposition-based FA has a higher convergence speed than traditional FA. It also picks up the best from all dimensions. The result of our Silhouette and the Elbow method with PCA and our proposed approach gives the best quality cluster solution in terms of intra-cluster distance and processing CPU time as compared to traditional K-means clustering.

Acknowledgement

This work is cited as IU/R&D/2023-MCN 0002071 in the Integral University manuscript database.

References

[1] Gonbadi, Arman Mohammadi, Seyed Hasan Tabatabaei, and Emmanuel John M. Carranza. "Supervised geochemical anomaly detection by pattern recognition." *Journal of Geochemical Exploration* 157 (2015): 81-91. https://doi.org/10.1016/j.gexplo.2015.06.001

[2] Ezugwu, Absalom E. "Nature-inspired metaheuristic techniques for automatic clustering: a survey and performance study." *SN Applied Sciences* 2 (2020): 1-57. https://doi.org/10.1007/s42452-020-2073-0

[3] Ezugwu, Absalom El-Shamir, Moyinoluwa B. Agbaje, Nahla Aljojo, Rosanne Els, Haruna Chiroma, and Mohamed Abd Elaziz. "A comparative performance study of hybrid firefly algorithms for automatic data clustering." *IEEE Access* 8 (2020): 121089-121118. https://doi.org/10.1109/ACCESS.2020.3006173

[4] Anderberg, Michael R. *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks*. Vol. 19. Academic press, 2014.

[5] Friedman, Menahem, and Abraham Kandel. *Introduction to pattern recognition: statistical, structural, neural, and fuzzy logic approaches*. Vol. 32. World scientific, 1999. https://doi.org/10.1142/3641

[6] Roberts, Stephen J. "Parametric and non-parametric unsupervised cluster analysis." *Pattern Recognition* 30, no. 2 (1997): 261-272. https://doi.org/10.1016/S0031-3203(96)00079-9

[7] Gan, Guojun, Chaoqun Ma, and Jianhong Wu. *Data clustering: theory, algorithms, and applications*. Society for Industrial and Applied Mathematics, 2020.

[8] Zhou, Yongquan, Haizhou Wu, Qifang Luo, and Mohamed Abdel-Baset. "Automatic data clustering using nature-inspired symbiotic organism search algorithm." *Knowledge-Based Systems* 163 (2019): 546-557. https://doi.org/10.1016/j.knosys.2018.09.013

[9] Cattell, Raymond B. "The description of personality: Basic traits resolved into clusters." *The journal of abnormal and social psychology* 38, no. 4 (1943): 476. https://doi.org/10.1037/h0054116

[10] Wong, Andrew KC, and Gary CL Li. "Simultaneous pattern and data clustering for pattern cluster analysis." *IEEE Transactions on Knowledge and Data Engineering* 20, no. 7 (2008): 911-923. https://doi.org/10.1109/TKDE.2008.38

[11] José-García, Adán, and Wilfrido Gómez-Flores. "Automatic clustering using nature-inspired metaheuristics: A survey." *Applied Soft Computing* 41 (2016): 192-213. https://doi.org/10.1016/j.asoc.2015.12.001

[12] Ikotun, Abiodun M., Mubarak S. Almutari, and Absalom E. Ezugwu. "K-means-based nature-inspired metaheuristic algorithms for automatic data clustering problems: Recent advances and future directions." *Applied Sciences* 11, no. 23 (2021): 11246. https://doi.org/10.3390/app112311246

[13] Alam, Afroj, Sahar Qazi, Naiyar Iqbal, and Khalid Raza. "Fog, edge and pervasive computing in intelligent internet of things driven applications in healthcare: Challenges, limitations and future use." *Fog, edge, and pervasive computing in intelligent IoT driven applications* (2020): 1-26. https://doi.org/10.1002/9781119670087.ch1

[14] Liu, Yongguo, Xindong Wu, and Yidong Shen. "Automatic clustering using genetic algorithms." *Applied mathematics and computation* 218, no. 4 (2011): 1267-1279. https://doi.org/10.1016/j.amc.2011.06.007

[15] Zabihi, Farzaneh, and Babak Nasiri. "A novel history-driven artificial bee colony algorithm for data clustering." *Applied Soft Computing* 71 (2018): 226-241. https://doi.org/10.1016/j.asoc.2018.06.013

[16] Das, Swagatam, Ajith Abraham, and Amit Konar. "Automatic clustering using an improved differential evolution algorithm." *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans* 38, no. 1 (2007): 218-237. https://doi.org/10.1109/TSMCA.2007.909595

[17] Yildiz, Ali R. "A new hybrid differential evolution algorithm for the selection of optimal machining parameters in milling operations." *Applied Soft Computing* 13, no. 3 (2013): 1561-1566. https://doi.org/10.1016/j.asoc.2011.12.016

[18] Ezugwu, Absalom E. "Nature-inspired metaheuristic techniques for automatic clustering: a survey and performance study." *SN Applied Sciences* 2 (2020): 1-57. https://doi.org/10.1007/s42452-020-2073-0

[19] Indrakumari, R., T. Poongodi, and Soumya Ranjan Jena. "Heart disease prediction using exploratory data analysis." *Procedia Computer Science* 173 (2020): 130-139. https://doi.org/10.1016/j.procs.2020.06.017

[20] Izakian, Zahedeh, Mohammad Saadi Mesgari, and Ajith Abraham. "Automated clustering of trajectory data using a particle swarm optimization." *Computers, Environment and Urban Systems* 55 (2016): 55-65. https://doi.org/10.1016/j.compenvurbsys.2015.10.009

[21] Alam, Afroj, Mohd Muqeem, and Sultan Ahmad. "Comprehensive review on Clustering Techniques and its application on High Dimensional Data." *International Journal of Computer Science & Network Security* 21, no. 6 (2021): 237-244.

[22] Sunarti, S., Irawan Dwi Wahyono, Hari Putranto, Djoko Saryono, Herri Akhmad Bukhori, Tiksno Widyatmoko, Mohd Shafie Rosli, Nurbiha A. Shukor, and Noor Dayana Abdul Halim. "Optimization Silhouette Coefficient on Genetic Algorithm for Clustering in Auto Correction German Learning." In *2022 International Conference on Electrical Engineering, Computer and Information Technology (ICEECIT)*, pp. 39-43. IEEE, 2022. https://doi.org/10.1109/ICEECIT55908.2022.10030619

[23] Hrosik, R. Capor, Eva Tuba, Edin Dolicanin, Raka Jovanovic, and Milan Tuba. "Brain image segmentation based on firefly algorithm combined with k-means clustering." *Stud. Inform. Control* 28, no. 2 (2019): 167-176. https://doi.org/10.24846/v28i2y201905

[24] Kapoor, Anjali, and Rekha Agarwal. "Enhanced Brain Tumour MRI Segmentation using K-means with machine learning based PSO and Firefly Algorithm." *EAI Endorsed Transactions on Pervasive Health and Technology* 7, no. 26 (2021): e2-e2.

[25] Alam, Afroj, Ismail Rashid, and Khalid Raza. "Application, functionality, and security issues of data mining techniques in healthcare informatics." In *Translational bioinformatics in healthcare and medicine*, pp. 149-156. Academic Press, 2021. https://doi.org/10.1016/B978-0-323-89824-9.00012-4

[26] Draa, Amer, Zeyneb Benayad, and Fatima Zahra Djenna. "An opposition-based firefly algorithm for medical image contrast enhancement." *International Journal of Information and Communication Technology* 7, no. 4-5 (2015): 385-405. https://doi.org/10.1504/IJICT.2015.070299

[27] Horng, Ming-Huwi, Yun-Xiang Lee, Ming-Chi Lee, and Ren-Jean Liou. "Firefly metaheuristic algorithm for training the radial basis function network for data classification and disease diagnosis." *Theory and new applications of swarm intelligence* 4, no. 7 (2012): 115-132. https://doi.org/10.5772/39084

[28] Reddy, Gadekallu Thippa, and Neelu Khare. "Hybrid firefly-bat optimized fuzzy artificial neural network based classifier for diabetes diagnosis." *International Journal of Intelligent Engineering & Systems* 10, no. 4 (2017). https://doi.org/10.22266/ijies2017.0831.03

[29] Zhang, Jian, Bo Gao, Haiting Chai, Zhiqiang Ma, and Guifu Yang. "Identification of DNA-binding proteins using multi-features fusion and binary firefly optimization algorithm." *BMC bioinformatics* 17, no. 1 (2016): 1-12. https://doi.org/10.1186/s12859-016-1201-8

[30] Chinta, Sai Srujan, Abhay Jain, and B. K. Tripathy. "Image segmentation using hybridized firefly algorithm and intuitionistic fuzzy C-Means." In *Proceedings of First International Conference on Smart System, Innovations and Computing: SSIC 2017, Jaipur, India*, pp. 651-659. Springer Singapore, 2018. https://doi.org/10.1007/978-981-10-5828-8_62

[31] Zhang, Li, Kamlesh Mistry, Siew Chin Neoh, and Chee Peng Lim. "Intelligent facial emotion recognition using moth-firefly optimization." *Knowledge-Based Systems* 111 (2016): 248-267. https://doi.org/10.1016/j.knosys.2016.08.018

[32] Ghosh, Partha, Kalyani Mali, and Sitansu Kumar Das. "Chaotic firefly algorithm-based fuzzy C-means algorithm for segmentation of brain tissues in magnetic resonance images." *Journal of Visual Communication and Image Representation* 54 (2018): 63-79. https://doi.org/10.1016/j.jvcir.2018.04.007

[33] Nayak, Janmenjoy, Matrupallab Nanda, Kamlesh Nayak, Bighnaraj Naik, and Himansu Sekhar Behera. "An improved firefly fuzzy c-means (FAFCM) algorithm for clustering real world data sets." In *Advanced Computing, Networking and Informatics-Volume 1: Advanced Computing and Informatics Proceedings of the Second International Conference on Advanced Computing, Networking and Informatics (ICACNI-2014)*, pp. 339-348. Springer International Publishing, 2014. https://doi.org/10.1007/978-3-319-07353-8_40

[34] Bhattacharyya, Saugat, Abhronil Sengupta, Tathagatha Chakraborti, Amit Konar, and D. N. Tibarewala. "Automatic feature selection of motor imagery EEG signals using differential evolution and learning automata." *Medical & biological engineering & computing* 52 (2014): 131-139. https://doi.org/10.1007/s11517-013-1123-9

[35] Cowgill, Marcus Charles, Robert J. Harvey, and Layne T. Watson. "A genetic algorithm approach to cluster analysis." *Computers & Mathematics with Applications* 37, no. 7 (1999): 99-108. https://doi.org/10.1016/S0898-1221(99)00090-5

[36] Boutsidis, Christos, Anastasios Zouzias, Michael W. Mahoney, and Petros Drineas. "Randomized dimensionality reduction for $ k $-means clustering." *IEEE Transactions on Information Theory* 61, no. 2 (2014): 1045-1062. https://doi.org/10.1109/TIT.2014.2375327

[37] de Almeida, Fabrício Alves, Ana Carolina Oliveira Santos, Anderson Paulo de Paiva, Guilherme Ferreira Gomes, and José Henrique de Freitas Gomes. "Multivariate Taguchi loss function optimization based on principal components analysis and normal boundary intersection." *Engineering with Computers* (2022): 1-17.

[38] Zhu, Changsheng, Christian Uwa Idemudia, and Wenfang Feng. "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques." *Informatics in Medicine Unlocked* 17 (2019): 100179. https://doi.org/10.1016/j.imu.2019.100179

[39] Zhou, Hong Bo, and Jun Tao Gao. "Automatic method for determining cluster number based on silhouette coefficient." *Advanced materials research* 951 (2014): 227-230. https://doi.org/10.4028/www.scientific.net/AMR.951.227

[40] Syakur, M. A., B. K. Khotimah, E. M. S. Rochman, and Budi Dwi Satoto. "Integration k-means clustering method and elbow method for identification of the best customer profile cluster." In *IOP conference series: materials science and engineering*, vol. 336, p. 012017. IOP Publishing, 2018. https://doi.org/10.1088/1757-899X/336/1/012017

[41] Thinsungnoena, Tippaya, Nuntawut Kaoungkub, Pongsakorn Durongdumronchaib, Kittisak Kerdprasopb, and Nittaya Kerdprasopb. "The clustering validity with silhouette and sum of squared errors." *learning* 3, no. 7 (2015). https://doi.org/10.12792/iciae2015.012

[42] Zhang, Lina, Liqiang Liu, Xin-She Yang, and Yuntao Dai. "A novel hybrid firefly algorithm for global optimization." *PloS one* 11, no. 9 (2016): e0163230. https://doi.org/10.1371/journal.pone.0163230

[43] Ikotun, Abiodun M., Mubarak S. Almutari, and Absalom E. Ezugwu. "K-means-based nature-inspired metaheuristic algorithms for automatic data clustering problems: Recent advances and future directions." *Applied Sciences* 11, no. 23 (2021): 11246. https://doi.org/10.3390/app112311246

[44] Li, Guocheng, Pei Liu, Chengyi Le, and Benda Zhou. "A novel hybrid meta-heuristic algorithm based on the cross-entropy method and firefly algorithm for global optimization." *Entropy* 21, no. 5 (2019): 494. https://doi.org/10.3390/e21050494

[45] Alam, Afroj. "Optimization of K-means clustering using Artificial Bee Colony Algorithm on Big Data."

[46] Dey, Nilanjan, Jyotismita Chaki, Luminița Moraru, Simon Fong, and Xin-She Yang. "Firefly algorithm and its variants in digital image processing: A comprehensive review." *Applications of Firefly Algorithm and Its Variants: Case Studies and New Developments* (2020): 1-28. https://doi.org/10.1007/978-981-15-0306-1_1

[47] Banerjee, Abhijit, Dipendranath Ghosh, and Suvrojit Das. "Modified firefly algorithm for area estimation and tracking of fast expanding oil spills." *Applied Soft Computing* 73 (2018): 829-847. https://doi.org/10.1016/j.asoc.2018.09.024

[48] Nekouie, Nadia, and Mahdi Yaghoobi. "A new method in multimodal optimization based on firefly algorithm." *Artificial Intelligence Review* 46 (2016): 267-287. https://doi.org/10.1007/s10462-016-9463-0

[49] Tian, Jinlan, Lin Zhu, Suqin Zhang, and Lu Liu. "Improvement and parallelism of k-means clustering algorithm." *Tsinghua Science and Technology* 10, no. 3 (2005): 277-281. https://doi.org/10.1016/S1007-0214(05)70069-9