IRAQI STATISTICIANS JOURNAL

https://isj.edu.iq/index.php/isj

ISSN: 3007-1658 (Online)

An Improved Unsupervised Feature Selection using Fuzzy C-means Clustering and Archimedes Optimization Algorithm.

Sarah Ghanim Mahmood Al- Kababchee[1], Ilham M Yacoob[2], Omar Saber Qasim[3], Zakariya Yahya Algamal[4].

[1,2] Department of Mathematics, Education College, University of AL-Hamdaniya, Mosul, Iraq.
[3] Department of Mathematics, University of Mosul, Mosul, Iraq.
[4] Department of Statistics and Informatics, University of Mosul, 41002 Mosul, Iraq; College of Engineering

ARTICLE INFO

Article history:

Received 6 January 2026
Revised 6 January 2026
Accepted 31 March 2026
Available online 3 April 2026

Keywords:

Gen selection
Feature selection
Clustering
Fuzzy C means clustering
Archimedes Optimization Algorithm

ABSTRACT

In this research, we propose a new choice of plant selection based on Archimedes Optimization Sales (BAS) in combination with a modified fuzzy c-mean algorithm. With marked patterns, BAS increases the unsecured learning of traditional unclear C-mines. Using these annotate patterns, BAS is able to model the geometry of each cluster correctly, emphasizing the properties that are most useful for discrimination against clusters. Patterns with high levels of cluster membership show small variance in the values of these functions. These symptoms have strong discriminatory properties between clusters and other groups if there is a significant distance between the prototype of a cluster and the prototype of other clusters. We used BAS to classify patterns on many real benchmarks. Our learning method was the popular K-nest neighbor (FCM). Our results showed that the functional elections made by BAS increased generalization in all data sets.

1. Introduction

The data has increased drmatically and quickly in recent decades. Gen selection process, which is the process of choosing a functional suit from the initial collection of properties, is immediately required as a result of accumulation of this data. Selecting a subset of pertinent attributes from the authentic dataset is known as function choice. Its number one goals are to growth calculation pace, lower superfluous capabilities, and beautify model accuracy. In recent years, it's been extensively researched [1].

Finding important characteristics in available data, eliminating those who are not related to the learning process, is the main goal of convenience choice [2]. Functional choices are important for two reasons: First, it increases the purity of findings; Second, it cuts the time to calculate solutions on the algorithm. Features can be divided into monitored and unprotected functional choices based on data provided for selections. The system chooses functions based on classification accuracy in the first scenario, when class labels are already provided. The second type of square label is missing, so functional elections should identify the best of properties in the absence of label information. Semi-contained learning is used when the square label is partly known and partially unknown [3, 4]. Cai & others presented Multi-Cluster Feature Selection (MCFS) introduces, which is a revolutionary insecure functional

Corresponding author E-mail address: sarahghanim@uohamdaniya.edu.iq
https://doi.org/10.62933/hcfkd736

choice approach. Inspired by the latest progression in the L1-regulated model for the most selection and spectral analysis of data, he suggested using multiple self-vests of Graf LaPlagian, defined on the Intimacy Matrix to data points, to capture the multicycle data structure. MCFS is therefore able to handle data on multiple clusters. Experimental results confirm that the new method provides much more performance for grouping and classification compared to two state-of-the-art methods, namely the Lapla Glassian score and Q, and a direct method, ie maximum variance. The performance of our proposed MCF's algorithm is very good when less than 50 functions are selected [5]. Zawbaa & others purpose of this study is to control and evaluate the capacity of hybrid Antilon-Gre Wolf-algorithm, Antale Optimization and Gray Wolf Optimization, which is for choosing important functions from datasets with relatively small number of events and a large number of characteristics. These two algorithms are combined to create a third model, and these models appear to perform better than other models used for comparison, including particle flock adaptation and genetic algorithms. This suggests that these techniques can get the right classification results on the highest selection function and are effective optimizers for large dimensional data sets with small examples. Hybrid Alo-Gwo algorithm creates a very excellent balance between the optimal solution and the discovery of the huge search area. The algorithm can prevent initial convergence and produce a variety of solutions [6]. Also, Li & others by using modern and weird networks TON-IoT data sets for classification in NIDs, he compared completely two-dimensional deficiency methods, FS and Fe. Our detailed analysis has shown that when a significant number of functions were reduced (eg 9 or 22), FE not only achieved high accuracy in the attack detection, but also demanded a short time for dimension reduction; However, for example, an increase in the number of functions (eg 33 or more), FS exceeded extraction of functions, indicating that FS showed more capacity with low functions, while Fe showed a room for improvement with a large number of functions. In addition, he found that FES continued to improve, while the effectiveness of FS fell significantly with an increased number of selected properties [7]. Cheng removing excellent functions, reducing overfeating and increasing efficiency, the study highlighted the importance of functional choices to promote machine learning model performance. We saw three primary approaches: the methods of the cover, which are accurate, but calculated expensive; Built -in methods, which add functional choices and model training; And filter methods, which are fast, but limited. We also talked about how the choice of facilities is used in different industries, including health services and banking, and how different performance measurements can be used to evaluate how well convenience election methods work [8].

2. Methodology:

2.1 Clustering

The method of clustering involves arranging data in hitherto unknown groups or categories, and ensuring that the components of a cluster are similar, while in other clusters are different from these. Many domains, such as data set analysis and unprotected classification, use computer grouping broadly. If they are part of the same cluster to maximize the correlation between two objects, a data analysis method for grouping exploration is dividing data into several groups. Clustering only reveals the pattern in data without explaining its existence [2, 9]. Explain $F=[X_1, X_2, ..., X_K]$ that is a list of K clusters, where $Y_i \in R^D$, & $R=[Y_1, Y_2, ..., Y_M]$ must have a collection of M data of objects. To reduce the purpose function, every data point in R is awarded to one of the K clusters during the cluster process. SSED between every (Y_i) and (X_j) is the definition of the inner cluster variety of purpose. The following targets are function [10]:

$$E(X,Y)=\sum_{i=1}^{M}\min\{\left\|Y_i - X_j\right\|^2\},\ j=1,2,...,K \quad (1)$$

Also, for X_j is not an empty set for all values of j, the intersection between X_i and X_j equal empty set for all $i \neq j$, and the union of all $X_j = R$, such X_j is a cluster Center.

2.2 C- Means Clustering

For continuous data, C-min grouping technology separates a predetermined number of data vectors in a predetermined number of groups. [11]. Consider a partition P_c of a data set with N data patterns (each data pattern is represented by a vector $x_j \in R^m$ where $j = 1, 2, ..., N$) in C clusters, where C is an algorithm input parameter. The centroid vector $g_c \in R^m$ (where c=1, 2,…,C), represents each cluster.

Euclidian distance, a measure of inequality, is used to make clusters in K-means [Eq. (2)]. Each cycle creates a new cluster center vector for each cluster from the average of the existing data sector (until a maximum number of iterations $t_{\max kmeans}$ is reached or another stopping requirement is satisfied) (In other words, the current data pattern of the cluster is taken into account[12]. Each pattern is then assigned the cluster with the nearest Centroid in a recently generated division) [13]

$$d(x_j, g_c) = \sqrt{\sum_{k=1}^{m} (x_{jk} - g_{ck})^2} \qquad (1)$$

Where,

$$g_c = \frac{1}{N} \sum_{\forall x_l \in c} x_l \qquad (2)$$

Cluster

Squares, giv

K-Means [1

$$J(P_c) = \sum_{c=1}^{C} \sum_{\forall x \in c} d(x_j, g_c) \qquad (3)$$

2.3 Fuzzy C-means

One of the most popular methods is an unclear cluster approach[14]. This model is an unprotected approach to production and data analysis. Dun first suggested, then changed Bezdek in 1980 [15]. Fuzy clustering is more natural than stiff clustering. Objects at the limits of many categories receive a degree of membership between 0 and 1, which reflects their partial membership, rather than perfectly related to a category [16].

Algorithm [16]

1- Randomly we can generate U , where $U = [u_{ij}]$ matrix for the initial U.

2- After that in k-step: we find centers vector $C^{(k)} = [c_1] with U^{(k)}$

$$C_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}$$

3- Improve $U^{(k)}, U^{(k+1)}$.

4- $d_{ij} = \sqrt{\sum_{i=1}^{n} (x_i - c_i)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}}$$

5- If the difference between $U^{(k+1)}$ and U^k less than ε So stop; or go back to step 2.

6- Here m is a constant belong to Real Number

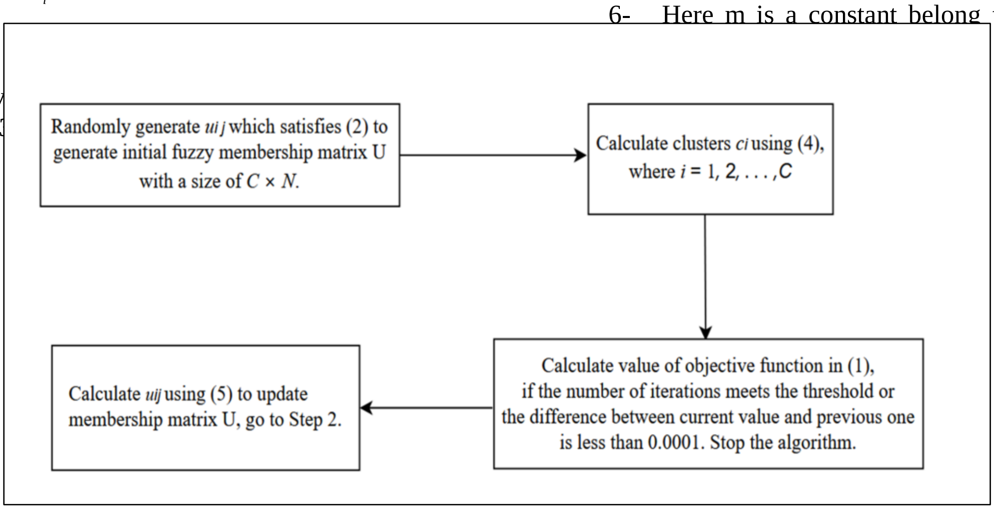

Figure 1. show the steps of fuzzy C-means.

2.4 Archimedes Optimization Algorithm (AOA)

AOA is an algorithm based on the population. Individual population consists of organisms submerged [17]. Like other traditional population -based algorithms, AOA begins to search with an early collection of organisms (permanent solutions) of random size, density and acceleration [18]. In this phase, each element is also placed in liquid in a random place. The agreement is made in cases of repetition until the need for expiration is not satisfied after evaluating the suitability of the initial package [19]. AOA adjusts the size and density of each object on each frequency. The acceleration of the object is updated in response to a collision with other surrounding objects. The new location of the object is determined by its modern density, size and acceleration [20]. Now we explain procedure of Archimedes Optimization Algorithm [18]:

Start population objects using random locations, density and versions using AOA (population size N, maximum iteration $t_{\max}, C_1, C_2, C_3, C_4$) after evaluating the original population, choose the population with the highest fitness value. Set t=1 as a repetition disk.

When $t < t_{\max}$ do

For any object i find

Update the volume and density of each object.

Update the transfer together with density matters T, F and D.

If T $TF < 0.5 then \rightarrow$ Exploration phase

Adjust the acceleration, adjust the position and adjust the acceleration.

Else $\rightarrow$ Exploration steps

Improve acceleration and normalize acceleration

Improve direction flag F and position

End if

End for

After evaluating every select the object with the highest fitness value, and set t=t+1.

2.5 Binary Archimedes optimization algorithm:

Archimede's optimization algorithm can handle intricate adaptation issues with many locally OS "optimal solutions", as it holds the collection of solutions and detects a large area for BGS "best global solution". This property makes it possible to use the binary version of the algorithm on the large metric dimensions[21].

In the constant form of AOA, institutions can move through the search space by using the position vector in constant real domains. Constant AOA variables can be converted to binary values using an S-shaped transmission function to convert a binary. It is necessary to turn between 0 and 1 to move positions to the discreet binary search site. In the this phase, the following equation is used[22]:

$$Obinaray_{ij} = \begin{cases} 1 & rand() < 0.5 \\ 0 & else \end{cases} \quad (4)$$

Rand: random number $\in [0,1]$.

Continuous values are mapped for binary values using transfer functions, the Sigmoid (s) feature it is utilized in the following way:

$$S = \frac{1}{1+e^{-10x^d}} \quad (5)$$

Such that S function is sent out and the x^d dimension is the point of a continuous value of d. A binary value is produced using the equation below:

$$Obinaray_{ij} = \begin{cases} 1 & rand() < S \\ 0 & otherwise \end{cases} \quad (6)$$

The technique looks for the optimal solution by treating the problems as an optimization problem. Consequently, a one-dimensional vector can be used to represent each thing $Obinaray = \{O_{i1}, O_{i2}, ..., O_{ij}\}$, $Obinaray_{ij}$ is a binary vector if the j-th element of the vector has of 1, it means that vertex j belongs to B. if every $v \in V(G)$ has a different representation $v(B/v)$, then B is a predominant solving set a value of the transfer function of the S-shaped is calculated to provide binary status vector value. In the BAOA process, a top from V of B is added to a unit when unsuitable as a dominant solution set. This change is made to the device a dominated solution set.

We can write any solution in the algorithm as a series of 0 or 1 values, where "1" refers to a choice of main solution kit (where the respective value is "1") and "0" reflect the non-choice of the main solution sets (where the related value is "0") [22].

2.6 The Purposed method

The proposed method aims to improve the performance of the Fuzzy C-Means (FCM) clustering algorithm by optimizing both the number of clusters (C) and the initial cluster centers using the Archimedes Optimization Algorithm (AOA).

The main idea is based on the fact that the performance of FCM is highly sensitive to the choice of the number of clusters and the initialization of cluster centers. Therefore, optimizing these values can significantly enhance clustering accuracy and stability.

In the proposed approach, both the number of clusters C and the membership function parameters are encoded in a binary string format, and the AOA is applied as a population-based optimization algorithm to select the best combination of parameters that yields the highest clustering performance.

Detailed Steps of the Proposed Method

Step 1: Initialization of Parameters

First, the following parameters are initialized:

Number of clusters C

Population size N=5 (number of agents or candidate solutions per iteration)

Maximum number of iterations T=10

Dataset D

AOA parameters: initial density, volume, and acceleration

Step 2: Generation of Initial Population

A population of agents is created, where each agent represents a potential solution consisting of a set of cluster centers in the feature space.

The cluster centers for each agent are initialized randomly within the data boundaries.

Each agent is assigned initial values for density, volume, and acceleration according to the AOA equations.

Step 3: Fitness Evaluation Using Fuzzy C-Means

For each agent in the population:

$$Fit = \alpha * \text{CE} + (1 - \alpha) * \text{NF} \quad , \quad \alpha \in [0,1]$$

CE :Classification_Error

NF : Normalized_Features

The Fuzzy C-Means algorithm is executed using the cluster centers represented by the agent as the initial centers.

After running FCM, the objective function is computed to evaluate clustering quality.

For Fuzzy C-Means, the objective is to minimize the intra-cluster distances, weighted by membership values.

For Hard C-Means, the objective is to minimize the sum of squared distances between data points and their corresponding cluster centers.

The resulting error value is used as the fitness score for that agent. Lower error values indicate better clustering performance (i.e., higher fitness).

Step 4: Updating Density and Volume

The density and volume of each agent are updated using the physics-inspired equations of the Archimedes Optimization Algorithm, which simulate buoyancy and the motion of objects in fluid environments.

Step 5: Calculation of the Transfer Function (TF)

The Transfer Function (TF) is computed to control the balance between:

Exploration: global search for new regions in the early stages.

Exploitation: fine-tuning around promising regions in the later stages.

The TF value gradually increases with each iteration to smoothly transition from exploration to exploitation.

Step 6: Updating Acceleration and Positions

The acceleration and position of each agent are updated using the AOA formulas.

The new position corresponds to an updated set of cluster centers in the feature space.

These updates allow agents to move toward better solutions based on the simulated physical forces.

Step 7: Fitness Re-evaluation

After updating the positions, the FCM algorithm is re-executed for each agent using its new cluster centers. The fitness function is then recalculated as in Step 3 to assess the quality of the updated solution.

Step 8: Updating the Best Solution

The fitness values of all agents are compared.

If any agent produces a better fitness score (i.e., lower clustering error), the global best solution is updated to that agent's configuration.

Step 9: Iteration Until Convergence

Steps (4)–(8) are repeated until one of the following conditions is met:

The maximum number of iterations T is reached, or

The improvement in the best fitness value becomes negligible (convergence achieved).

Step 10: Output

After convergence, the algorithm outputs:

The best set of cluster centers obtained through the AOA optimization process.

The final clustering results obtained by applying FCM initialized with these optimized centers.

The proposed hybrid AOA–FCM algorithm combines the global exploration ability of the Archimedes Optimization Algorithm with the local refinement capability of the Fuzzy C-Means algorithm.

This integration leads to:

- Improved clustering accuracy
- Reduced sensitivity to initial cluster center selection
- More effective determination of the optimal number of clusters

The proposed method uses a population-based search approach with the BAS-FCM operators to simultaneously optimize the number of clusters and feature selection for fuzzy C-means clustering. The fitness evaluation is based on the fuzzy C-means clustering technique, which evaluates the clustering solution's efficacy. The BAS-FCM operators are used to generate new solutions by modifying and combining the population's preexisting solutions. The method terminates when a stopping condition is met, such as a maximum number of iterations or convergence of the optimal fitness value.

Figure1 is an illustration of the solution representation, which shows how the C and feature selection are encoded in a binary string format. The first part of the string represents the C value, and the second part represents the feature selection, where a 1 indicates that the corresponding feature is selected and a 0 indicates that it is not selected.

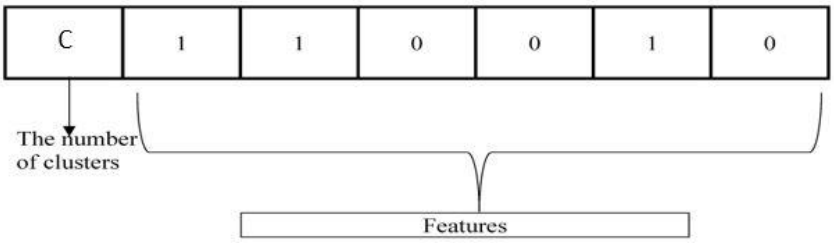

Figure (1): Representation of feature selection in the proposed algorithm.

3. Results and Discussion

Because it is important to determine the original values suitable for acceleration of the method, unclear grouping algorithm is sensitive to early values. To determine the first settings of the algorithm, including the cluster and the number of proposed centers, we used Archimedes optimization algorithm in this study. After choosing the number of centers and groups, unclear clustering was used, and the membership function was used to determine the degree of membership in each element.

We used the BAS-FCM algorithm proposed on five different publicly available data sets to assess its efficiency. To determine the ideal number of clusters, we also evaluated the performance against a traditional unclear approach using BAS. We were able to find out if the proposed algorithm does better than the traditional method of this comparison.

A brief summary of the dataset used in tests, which comes from UCI Machine Learning Repository, is given in Table 1. Five data sets in the real world with different dimensions, observational calculations and clusters were used in tests. The table gives a brief description of each dataset to give a general idea of its properties.

Table 1: The datasets used.

Dataset	N	Dim
Data1 (Cmc)	1473	9
Data2 (Glass)	214	9
Data3 (Parkinsons)	195	23
Data4 (Seeds_dataset)	210	7
Data5 (Wine)	178	13

Table 2: The Silhouette value results for BAS-FCM and FCM clustering accuracy.

Datasets	BAS-FCM	FCM
Data1 (Cmc)	0. 9699	0. 6413
Data2 (Glass)	0. 8615	0. 6667
Data3 (Parkinsons)	0. 7997	0. 5399
Data4 (Seeds dataset)	0. 8056	0. 7032
Data5 (Wine)	0. 6796	0. 4780

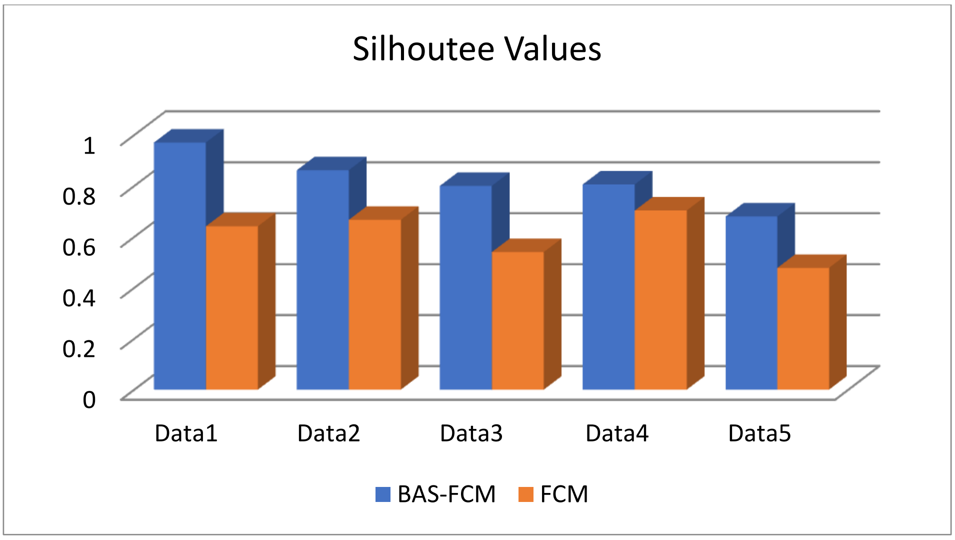

Figure (3): Comparison between BAS-FCM and FCM in Silhouette value results.

Table 3 shows that the BAS-FCM algorithm yields the best feature selection across all datasets. This implies that BAS-FCM is capable create groups they near the expected clusters on average. Additionally, BAS-FCM outshines FCM in terms of computational complexity.

Table 3: demonstrates a feature selection comparison between the BAS-FCM and FCM algorithms.

Dataset	BAS-FCM	FCM
Data1	4	9
Data2	5	9
Data3	6	23
Data4	2	7
Data5	4	13

4. Conclusions

this study introduces a new method called BAS-FCM to improve Fuzzy C-MINE's clustering. It is combined with the Archimedes algorithm to determine the ideal number of clusters and functional choices. User on five datasets in tables (3–4) suggests that the BAS-FCM algorithm provides more effective and effective solutions for cluster difficulties. According to the results, the proposed method works better than the current algorithm in terms of average convenience choice and silhouette value. Applications for proposed approaches are found in several domains, including data analysis and practical adjustment problems. Overall, by offering a new technique to improve Fuzzy C-Mine's cluster performance using an optimal method of discovery, this study links the knowledge of knowledge of grouping.

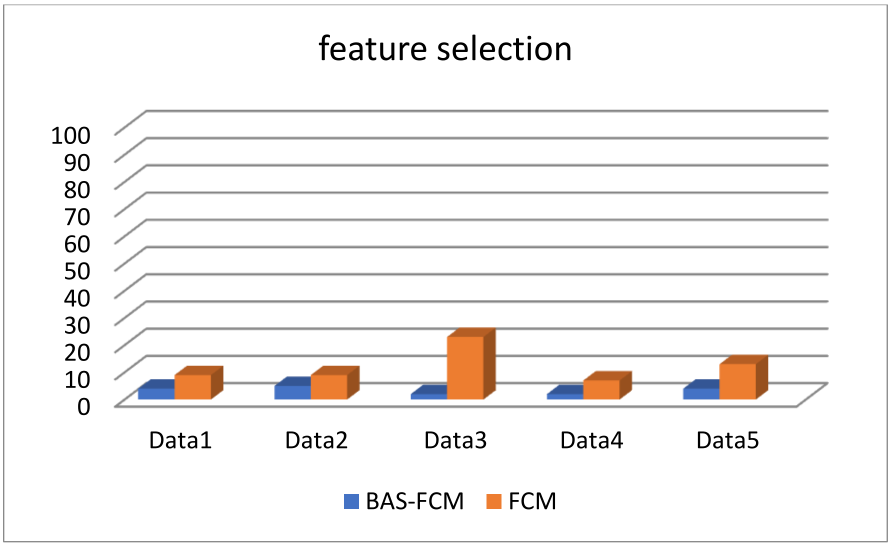

Figure (4): Comparison between BAS-FCM and FCM in average feature selection.

References

1. Abdollahi, M., et al., Imperialist competitive algorithm for solving systems of nonlinear equations. 2013. **65**(12): p. 1894-1908.
2. K, S., Feature Selection using K-Means Clustering for Data Mining. Conference: International Symposia in Advance Computing (ISAC'2K11), 2011.
3. Sutar, S., Feature selection algorithm using fast clustering and correlation measure. 2015. **2**(7): p. 236-241.
4. Mahmood, S.G., R.S. Karyakos, and I.M.J.E.-E.J.o.E.T. Yacoob, Hybrid gene selection method based on mutual information technique and dragonfly optimization algorithm. 2021. **3**(3): p. 111.
5. Cai, D., C. Zhang, and X. He. Unsupervised feature selection for multi-cluster data. in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. 2010.
6. Zawbaa, H.M., et al., Large-dimensionality small-instance set feature selection: A hybrid bio-inspired heuristic approach. 2018. **42**: p. 29-42.
7. Li, J., et al., Optimizing IoT intrusion detection system: feature selection versus feature extraction in machine learning. 2024. **11**(1): p. 36.
8. Cheng, X.J.I.i.C., Signals and Systems, A Comprehensive Study of Feature Selection Techniques in Machine Learning Models. 2024. **1**(1): p. 10.70088.
9. Al-kababchee, S.G.M., Z.Y. Algamal, and O.S.J.J.o.I.S. Qasim, Enhancement of K-means clustering in big data based on equilibrium optimizer algorithm. 2023. **32**(1): p. 20220230.
10. Krishnasamy, G., A.J. Kulkarni, and R. Paramesran, A hybrid approach for data clustering based on modified cohort intelligence and K-means. Expert Systems with Applications, 2014. **41**(13): p. 6009-6016.
11. MacQueen, J. Some methods for classification and analysis of multivariate observations. in Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. 1967. Oakland, CA, USA.
12. Arora, J., K. Khatter, and M.J.S.E.P.o.C. Tushir, Fuzzy c-means clustering strategies: A review of distance measures. 2019: p. 153-162.
13. Pacifico, L.D.S. and T.B. Ludermir, An evaluation of k-means as a local search operator in hybrid memetic group search optimization for data clustering. Natural Computing, 2020.
14. Krasnov, D., et al., Fuzzy c-means clustering: A review of applications in breast cancer detection. 2023. **25**(7): p. 1021.
15. Xu, J., et al. Robust and sparse fuzzy k-means clustering. in IJCAI. 2016.
16. Suganya, R., R.J.I.J.o.S. Shanthi, and R. Publications, Fuzzy c-means algorithm-a review. 2012. **2**(11): p. 1.
17. Nurmuhammed, M., et al., Modified Archimedes optimization algorithm for global optimization problems: A comparative study. 2024. **36**(14): p. 8007-8038.
18. Ge, Q., et al., Support Vector Machine to Predict the Pile Settlement using Novel Optimization Algorithm. 2023: p. 1-15.
19. Allogmani, A.S., et al., Enhanced cervical precancerous lesions detection and classification using Archimedes Optimization Algorithm with transfer learning. 2024. **14**(1): p. 12076.
20. Hashim, F.A., et al., Archimedes optimization algorithm: a new metaheuristic algorithm for solving optimization problems. 2021. **51**: p. 1531-1551.
21. Fang, L., Y. Yao, and X.J.E.S.w.A. Liang, New binary archimedes optimization algorithm and its application. 2023. **230**: p. 120639.

22. Mohamed, B., et al., Binary Archimedes Optimization Algorithm for Computing Dominant Metric Dimension Problem. 2023. **38**(1).