A Study for Comparing Various Meta-Heuristic Algorithms in Data Clustering on applying K-Means

JIGYASHA VERMA[1], RAJSHREE GULATI[2], PRAPTI HALDER[3]
[1, 2, 3] *Amity University Uttar Pradesh, Noida, UP, India*

Abstract- ***Clustering, a technique used across various fields like pattern recognition and healthcare, involves grouping data points based on their similarities. However, traditional clustering methods sometimes struggle with issues like sensitivity to initial conditions and the tendency to get stuck in suboptimal solutions. To address these challenges, researchers have turned to Nature-Inspired Algorithms (NIAs), which mimic the behaviour of natural systems to find optimal solutions efficiently. Three popular NIAs—Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), and Genetic Algorithms (GA)—have gained attention for their ability to tackle clustering problems effectively. These algorithms emulate the behaviour of ants, particles in a swarm, and genetic evolution, respectively, to search for high-quality solutions in a complex search space. In this paper, we aim to determine which NIA performs best for clustering tasks, considering factors such as convergence speed, computational efficiency, and solution quality. By conducting a comprehensive comparative study, we evaluate the performance of each algorithm and assess their effectiveness in finding optimal sets of clusters. Our findings indicate that the Genetic Algorithm (GA) stands out as the most promising approach, outperforming other NIAs in terms of convergence speed, execution time, and success rate in finding high-quality clusters. We support our conclusions with rigorous statistical tests and detailed analyses, providing strong evidence for the superiority of the GA method in clustering applications. Overall, this paper contributes to the understanding of NIAs in clustering and highlights the practical advantages of using GA for solving clustering problems effectively and efficiently.***

Indexed Terms- ***Nature Inspired Algorithms, Clustering, K-Means, ACO, PSO, GA, WOA, Accuracy, Precision.***

I. INTRODUCTION

In today's landscape, clustering finds application in a multitude of fields, spanning from process monitoring to image processing, gene expression analysis, e-learning, and healthcare. This unsupervised machine-learning technique involves categorizing data points based on their similarities [1]. Clustering methods are commonly categorized into two main types: hierarchical clustering and partition clustering [2].

While clustering proves effective in grouping similar data points into clusters, it may not be suitable for all problem types. For instance, tasks like predictive modelling, causal inference, outlier detection, and handling high-dimensional data pose challenges for traditional clustering approaches. To address these challenges, researchers have turned to nature-inspired algorithms, which draw inspiration from natural processes to develop efficient task-solving methods. Metaheuristic algorithms, which leverage natural phenomena and behaviours such as those observed in swarms, animals, and physical laws, have been explored extensively for clustering tasks. Algorithms like Particle Swarm Optimization (PSO), Multi-Objective Optimization Algorithm (MOA), Genetic Algorithm (GA), Ant Colony Optimization (ACO), Bat Algorithm (BA), Artificial Bee Colony (ABC), and Whale Optimization Algorithm (WOA) have been employed for cluster analysis. However, these algorithms often encounter issues such as population diversity, slow convergence, trade-offs, and local optima.

To mitigate these challenges, researchers have devised hybrid approaches by combining metaheuristic algorithms with other techniques or enhancing their strengths while mitigating weaknesses. For instance, hybridizing PSO with k-harmonic mean or integrating chaotic maps into PSO have been strategies to improve convergence rates and speed. Similarly, combining K-means with PSO aims to enhance K-means' performance and reduce the impact of initial centroids on clustering results.

In conclusion, clustering remains a vital technique for data analysis across diverse fields. The efficacy of clustering algorithms can be evaluated using cluster validation techniques, which encompass internal, external, and relative evaluation methods. While metaheuristic and nature-inspired algorithms

have shown promise in addressing clustering challenges, ongoing research is essential for developing new algorithms and refining existing ones to overcome inherent limitations. This paper explores the effectiveness of several metaheuristic algorithms, including Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), Whale Optimization Algorithm (WOA), and Genetic Algorithm (GA), through novel development and optimization methods [3].

Lacking in clustering methods

In the Swarm Intelligence (SI) methods, clustering solutions are modelled after organisms, employing strategies for both intelligent intensification and diversification to navigate through the search space dynamically [4]. While these approaches offer advantages like reduced execution time compared to traditional clustering methods, they do come with certain drawbacks. For instance, Particle Swarm Optimization (PSO) and similar techniques often suffer from a lack of diversity among swarm members and mature memory elements, which can result in premature convergence and confinement to local optima. These limitations can lead to suboptimal convergence rates and the generation of low-quality solutions. Consequently, over the past two decades, numerous new SI methods and hybrid approaches have emerged aimed at addressing these constraints [4].

Contribution of the work

- During the individual testing of the algorithms, it became evident that there was room for optimization. Subsequently, we enhanced the performance and accuracy by implementing optimization functions.
- Next, we determined the optimal number of clusters using separate NIAs, applied k-means to these clusters and calculated the accuracy and standard deviation (precision).

II. LITERATURE REVIEW

Naeem et. al introduced a novel multi-objective optimization approach based on the ant colony algorithm (ACO), recommended for addressing the community detection problem in complex networks. Their findings suggest that the proposed algorithm demonstrates efficiency and effectiveness in uncovering community structures and optimizing solutions within intricate networks. Meanwhile, Marco and his team provided an overview of recent advancements in ACO. ACO algorithms prove particularly valuable when applied to "ill-structured" problems where the application of local search is ambiguous or within highly dynamic domains with limited access to local information. Their research delved into the ant colony system (ACS), a distributed algorithm applied to the traveling salesman problem (TSP). The results demonstrate that ACS surpasses other nature-inspired algorithms such as simulated annealing and evolutionary computation. They concluded their study by comparing ACS-3-opt, an augmented version of ACS with a local search procedure, to other top-performing algorithms for symmetric and asymmetric TSPs.[8]

Ahmed et al. delved into one of the most prominent paradigms of Swarm Intelligence, the Particle Swarm Optimization algorithm (PSO). Their paper analyzed existing research on PSO methods and applications published between 2017 and 2019.[10] Through a technical taxonomy of the selected content, including hybridization, improvement, and variants of PSO, as well as real-world applications across healthcare, environmental, industrial, commercial, smart city, and general aspects, they examined various technical characteristics such as accuracy, evaluation environments, and proposed case studies to evaluate the effectiveness of different PSO methods and applications.

Surbhi and colleagues developed a hybrid WOA_APSO model utilizing a convolutional neural network for categorization purposes. They conducted pre-processing and segmentation on 120 lung CT scans to differentiate nodules in tumored and untumored regions[7]. The classification algorithm's effectiveness heavily relies on the optimal feature selection process, with the convolutional neural network classification technique outperforming support vector machines and artificial neural networks, achieving an accuracy of 97.5%.

Yugal et al. investigated the effectiveness of three variations of the bat algorithm (BA-C, BA-CN, and BA-CNE) using intra-cluster distance, accuracy, and rand index parameters across twelve benchmark clustering datasets. Their simulation data indicated that the BA-CNE variation yielded superior clustering outcomes[13]. They also employed two

statistical tests to compare BA-CNE simulation results with those of other clustering algorithms, confirming BA-CNE as a reliable and efficient approach for addressing partitional clustering issues.

Raju et al. proposed a new binary gray wolf optimizer-based clustering method to enhance the energy efficiency of sensor networks, resulting in improved network performance up to 90% of its lifetime. Additionally, numerical optimization was employed to address dimensionality reduction issues, focusing on mathematical problems involving finding the best solution among a set of possible solutions. The tested functions included Rastrigin, Ackley, Rosenbrock, Sphere, and Griewank.

III. PROPOSED METHODOLOGY

The methodology involves a series of sequences which include- selection of the appropriate dataset to selection of feature and accuracy and precision computation. The steps are as shown below:

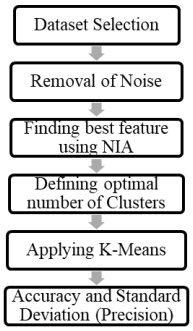

Fig 1. Steps of proposed methodology

3.1 Dataset Selection

The process of dataset selection involves choosing a subset of data from a larger collection that is crucial for a particular task or analysis. In machine learning, the effectiveness of a model heavily relies on the quality of the data it is trained on, and dataset selection plays a vital role in ensuring that the model is trained on pertinent data. This process typically entails identifying the data attributes most relevant to the task at hand, while filtering out any irrelevant or redundant data. For our research, we have opted to utilize the Iris dataset and the Seed dataset.

3.2 Removal of Noise

Feature selection, also referred to as variable or attribute selection, is the procedure of identifying a subset of relevant features from a larger set available within a dataset. In machine learning, the success of a model is greatly influenced by the quality and relevance of the features employed for training. The objective of feature selection is to pinpoint the most informative and pertinent features that are likely to enhance the model's performance, while discarding any irrelevant or redundant features that could lead to overfitting or suboptimal performance. Various techniques exist for feature selection, ranging from basic statistical measures like correlation analysis and mutual information to more complex methods such as wrapper and embedded methods, as well as dimensionality reduction techniques like principal component analysis (PCA) or t-SNE.

The features within the Iris dataset have been carefully chosen to be informative and relevant for the classification task, rendering the dataset an invaluable benchmark for testing and comparing different machine learning algorithms, as well as exploring the efficacy of various feature selection and dimensionality reduction techniques.

3.3 Finding best feature using NIA

In the feature selection process, we initially define a fitness function to calculate the cost for a specific value, fitness score, etc. This entails initializing a population of candidate solutions, assessing their fitness, selecting parents for reproduction, generating new solutions using genetic operators, evaluating the fitness of these new solutions, and selecting survivors for the subsequent generation. This iterative process continues for a predetermined number of generations until the best solution, i.e., the subset of features that maximizes the fitness function, is identified.

3.4 Defining Optimal Number of Clusters

This step involves carefully selecting the optimal data points and assigning them to distinct clusters. Hyperparameter optimization plays a crucial role here, as incorrectly optimized parameters may yield different results. Here before moving forward, we find the silhouette score to determine the which are the optimal clusters.

3.5 Applying K-means clustering

K-means clustering is a widely used unsupervised learning algorithm designed to group similar data points together in a dataset. This algorithm partitions the dataset into k clusters, with each cluster

representing a collection of data points that share similarities.

K-means clustering offers several advantages, including its simplicity, scalability, and efficiency. It finds applications in various domains such as image processing, text mining, and customer segmentation. The formula employed in K-means clustering is represented by:

$J=\sum i=1n||xi-\mu j||2$

Where J denotes the objective function, x(i) represents the ith data point, n is the total number of data points, μ (j) signifies the centroid of the jth cluster.

3.6 Accuracy and Precision outputs

We evaluated four distinct bio-inspired algorithms—Whale Optimization Algorithm (WOA), Particle Swarm Optimization (PSO), Genetic Algorithm (GA), and Ant Colony Optimization (ACO). These algorithms were utilized to select the optimal set of feature clusters from the predefined labelled datasets.

IV. RESULTS

The algorithm and results were computed on a machine equipped with 16GB of RAM, an Intel-based Kaby Lake CPU running at 3.45 GHz with 4 cores, and 150GB of disk space (SSD-v2), without a dedicated GPU. The datasets utilized in this study were Iris and Seed, comprising numerous standard data points, making them suitable for K-means clustering. Presented below is a comparative table containing the specifications for each algorithm.

PSO:

Constant parameters:

Iteration for each population size:	Dimensions for each iteration:
1000	1

ACO:

Constant parameters:

Iteration for each population size:	Dimensions for each iteration:
1000	1

GA:

Constant Parameters:

GENE_LENGTH	NUM_GENERATIONS	MUTATION_RATE
6	20	0.1

WOA:

Constant Parameters:

Iteration for each population size:	Cross over size for each iteration:
1000	1

Table 4.1: Comparative table for accuracy for different NIA

Dataset	Population_size	GA	PSO	ACO	WOA
Iris	10	97.89	97.00	96.65	95.91
	35	97.65	96.79	96.65	95.90
	50	97.47	96.66	96.37	95.86
	75	97.33	96.45	96.32	95.84
	100	97.31	96.40	96.02	95.70
Seed	10	96.77	96.76	96.01	95.56
	35	96.74	96.66	95.89	95.54
	50	96.23	96.54	96.00	95.49
	75	96.22	96.45	95.78	95.39
	100	96.00	96.45	95.74	95.37

Table 4.2: Comparative table for standard deviation (precision) for different NIA

Dataset	Population_size	GA	PSO	ACO	WOA
Iris	10	0.63	0.691	0.6	0.54
	35	0.64	0.678	0.53	0.542
	50	0.624	0.599	0.682	0.48
	75	0.621	0.58	0.364	0.431
	100	0.60	0.58	0.36	0.59
Seed	10	0.32	0.384	0.30	0.29
	35	0.321	0.379	0.299	0.287
	50	0.29	0.37	0.289	0.28
	75	0.287	0.365	0.28	0.27
	100	0.279	0.36	0.28	0.27

Table 4.3: Comparative table for recall for different NIA

Dataset	Population_size	GA	PSO	ACO	WOA
Iris	10	0.94	0.65	0.78	0.85
	35	0.94	0.64	0.78	0.84

	50	0.93	0.64	0.77	0.84
	75	0.93	0.63	0.77	0.83
	100	0.93	0.60	0.77	0.83
Seed	10	0.92	0.78	0.86	0.83
	35	0.92	0.77	0.86	0.82
	50	0.92	0.77	0.86	0.82
	75	0.91	0.76	0.84	0.81
	100	0.91	0.76	0.84	0.81

CONCLUSION & FUTURE WORK

Following a thorough analysis of each algorithm and their respective outcomes, it is evident that the Genetic Algorithm (GA) surpasses the others in terms of computational complexity. Notably, it exhibits the shortest execution time, accurately determining the optimal number of clusters for each dataset. Moreover, the silhouette score exceeds 0.5, suggesting that the GA may have reached its global optimal position. These identified optimal cluster numbers (k) can be leveraged for partition clustering problems. Nonetheless, optimizing certain hyperparameters may potentially yield even more precise results within a shorter timeframe.

REFERENCES

[1] E.-G. Talbi, Metaheuristics: From Design to Implementation, John Wiley & Sons, Hoboken, NJ, USA, 2009.

[2] Boussa J. Lepagnot, and P. Siarry, "A survey on optimization metaheuristics," *Information Sciences*, vol. 237, pp. 82–117, 2013.

[3] K. S. Lee and Z. W. Geem, "A new meta-heuristic algorithm for continuous engineering optimization: harmony search theory and practice," *Computer Methods in Applied Mechanics and Engineering*, vol. 194, no. 36–38, pp. 3902–3933, 2005.

[4] Demar, J. C. (2006). Statistical comparisons of classifiers through confidence curves. Neural Computation, 18(12), 2824-2850.

[5] Neelima, S., Ramkumar, B., & Kumar, G. V. P. (2018). A hybrid algorithm based on artificial bee colony and BAT for frequent itemset mining. International Journal of Applied Engineering Research, 13(12), 10457-10465.

[6] Fister, I., Yang, X. S., & Brest, J. (2013). A hybrid bat algorithm. IEEE Transactions on Evolutionary Computation, 17(2), 476-490.

[7] Kennedy, J., & Eberhart, R. C. (1995). Particle swarm optimization. In Proceedings of the IEEE international conference on neural networks, 4(C2), 1942-1948.

[8] Srensen, K. (2015). Metaheuristics—the role of metaphor. In Metaheuristics (pp. 13-27). Springer, Cham.

[9] Garcia, S., Herrera, F., & Fernández, A. (2010). A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. Soft Computing, 14(7), 959-977.

[10] Osman, I. H., & Laporte, G. (1996). Metaheuristics: Theory and applications. Kluwer Academic Publishers, Boston, MA.

[11] Battiti, R., & Tecchiolli, G. (1994). The reactive tabu search. ORSA Journal on Computing, 6(2), 126-140.

[12] Bansal, J. C., & Farswan, M. (2015). A novel migration mechanism for biogeography-based optimization. Soft Computing, 19(4), 891-918.

[13] Vanita, & Kusum, D. (2015). A comprehensive review of biogeography-based optimization (BBO). Artificial Intelligence Review, 44(1), 1-59; Verma, M., & Kesswani, N. (2015). A review of bio-inspired migration optimization techniques.

[14] Ma, H., & Simon, D. (2011b). Blended biogeography-based optimization for constrained optimization problems. Applied Mathematics and Computation, 217(17), 7228-7242.

[15] Shahraki, M. J., & Zahiri, E. (2017). Inclined plane optimization: A novel metaheuristic approach. Applied Soft Computing, 55, 1-13.

[16] Li, X., Zhang, W., & Liu, L. (2020). A hybrid PSO and simulated annealing algorithm for feature selection. Applied Intelligence, 50(2), 554-570.

[17] Liu, Z., Zhang, X., & Zhang, H. (2021). A PSO method with adaptive neighborhood selection for multi-objective optimization. Knowledge-Based Systems, 223, 107037.

[18] Jiang, K., Luo, H., & Liu, Y. (2022). Partial swarm optimization with adaptive communication for job shop scheduling. Applied Intelligence, 52(10), 6081-6104.

[19] Ching-Yi Chen and Fun Ye. (2012). Particle swarm optimization algorithm and its application to clustering analysis 4219-7865.

[20] Laith Mohammad Abualigah, Ahamad Tajudin Khader, and Essam Said Hanandeh (2018). A New Feature Selection Method to Improve the Document Clustering Using Particle Swarm Optimization Algorithm, 7865.

[21] Laith Mohammad Abualigah, Ahamad Tajudin Khader, and Essam Said Hanandeh (2018). A New Feature Selection Method to Improve the Document Clustering Using Particle Swarm Optimization Algorithm, 7865.

[22] Qingjian Ni, Qianqian Pan, Huimin Du, Cen Cao, and Yuqing Zhai (2015). A Novel Cluster Head Selection Algorithm Based on Fuzzy Clustering and Particle Swarm Optimization, 4205.

[23] C. Mageshkumar, S. Karthik & V. P. Arunachalam (2018). Hybrid metaheuristic algorithm for improving the efficiency of data clustering, 5665.

[24] S De, S Dey, S Bhattacharyya (2020). Recent advances in hybrid metaheuristics for data clustering, 8985.

[25] AM Ikotun, MS Almutari, AE Ezugwu (2021). K-means-based nature-inspired metaheuristic algorithms for automatic data clustering problems: Recent advances and future directions, 865-990.

[26] R.J. Kuo, T.C. Lin, F.E. Zulvia, C.Y. Tsai (2018). A hybrid metaheuristic and kernel intuitionistic fuzzy c-means algorithm for cluster analysis, 685-778.

[27] RK Yadav, RP Mahapatra (2022). Hybrid metaheuristic algorithm for optimal cluster head selection in wireless sensor network.

[28] P Hosseinioun, M Kheirabadi, SRK Tabbakh (2020). A new energy-aware tasks scheduling approach in fog computing using hybrid meta-heuristic algorithm, 4985.

[29] Moyinoluwa B. Agbaje; Absalom E. Ezugwu; Rosanne Els (2019). Automatic Data Clustering Using Hybrid Firefly Particle Swarm Optimization Algorithm, 556-698.

[30] Saeed Saeedvand, Hadi S. Aghdasi & Jacky Baltes (2019). Robust multi-objective multi-humanoid robots task allocation based on novel hybrid metaheuristic algorithm.