Mobile and Forensics
Vol. 6, No. 2, September 2024, pp. 52-64
ISSN: 2714-6685 (online), 2656-6257 (print)
DOI: http://dx.doi.org/10.12928/mf.v6i2.11529

Hybrid ABC–K Means for Optimal Cluster Number Determination in Unlabeled Data

[1,*]Harunur Rosyid, [2]Muhammad Modi bin Lakulu, [3]Ramlah bt. Mailok
[1]Universitas Muhammadiyah Gresik, Indonesia
[2]Universiti Pendidikan Sultan Idris, Malaysia
[3]Universiti Pendidikan Sultan Idris, Malaysia
[1,*]harun@umg.ac.id, [2]modi@meta.upsi.edu.my, [3]mramlah@meta.upsi.edu.my
[*]correspondence email

Abstract

This study presents the ABC K Means GenData algorithm, an enhancement over traditional K Means clustering that integrates the Artificial Bee Colony (ABC) optimization approach. The ABC K Means GenData algorithm addresses the issue of local optima commonly encountered in standard K Means algorithms, offering improved exploration and exploitation strategies. By utilizing the dynamic roles of employed, onlooker, and scout bees, this approach effectively navigates the clustering space for categorical data. Performance evaluations across several datasets demonstrate the algorithm's superiority. For the Zoo dataset, ABC K Means GenData achieved high Accuracy (0.8399), Precision (0.8089), and Recall (0.7286), with consistent performance compared to K Means and Fuzzy K Means. Similar results were observed for the Breast Cancer dataset, where it matched the Accuracy and Precision of K Means and surpassed Fuzzy K Means in Precision and Recall. In the Soybean dataset, the algorithm also performed excellently, showing top scores in Accuracy, Precision, Recall, and Rand Index (RI), outperforming both K Means and Fuzzy K Means.. The comprehensive results indicate that ABC K Means GenData excels in clustering categorical data, providing robust and reliable performance. Future research will explore its application to mixed data types and social media datasets, aiming to further optimize clustering techniques.
.

Keywords: ABC K Means GenData, Artificial Bee Colony (ABC) Optimization, Clustering Algorithms, Categorical Data, Performance Evaluation

INTRODUCTION

Clustering is a foundational technique in computer science, used extensively across various fields such as data mining, statistical analysis, dimensionality reduction, and vectorquantization. It involves grouping data into clusters based on similarity, with the goal of maximizing intra-cluster homogeneity and minimizing inter-cluster heterogeneity[1] This process is essential for uncovering underlying patterns and relationships within large datasets, thereby facilitating better understanding and decision-making[2]. One of the most commonly used clustering algorithms is K-Means, introduced by MacQueen in 1967[3]. This method partitions data into a predefined number of clusters by iteratively adjusting centroids to minimize the within-cluster variance[3], [4] While K-Means is effective for datasets with globular clusters and is known for its simplicity and efficiency, it faces several limitations. A primary challenge is the need to specify the number of clusters (k) beforehand, which can be difficult as the optimal value of k is often unknown and varies with the dataset's characteristics[5]. Additionally, K-Means may struggle with non-globular clusters and datasets containing outliers or varying densities[6]. To address these issues, researchers have explored various modifications and enhancements to the K-Means algorithm. One notable approach is the integration of the Artificial Bee Colony (ABC) optimization algorithm, which enhances clustering by improving the exploration and exploitation of possible solutions[7]. The ABC algorithm, inspired by the foraging behavior of honeybees, offers

Article History: *Received September 4, 2024; Revised October 1, 2024; Accepted October 1, 2024*

advantages such as better memory utilization, local search, and solution improvement mechanisms[8]. However, like K-Means, traditional ABC algorithms also face challenges related to local optima and the need for effective parameter tuning[9]. In recent years, the challenge of escaping local optima in clustering algorithms has garnered significant attention across various domains. Researchers have increasingly adopted methods such as hierarchical clustering, partition-based clustering, density-based clustering, and artificial intelligence-based clustering to mitigate this issue. Hierarchical clustering allows for a more flexible exploration of the data space by creating a tree-like structure of nested clusters. On the other hand, partition-based methods, such as K-means, which divides data into a pre-specified number of clusters, can sometimes lead to suboptimal solutions due to poor initialization of cluster centers[10] Density-based methods like DBSCAN, which are effective in discovering clusters of arbitrary shapes without requiring a predefined number of clusters, may struggle with varying densities[11]. Artificial intelligence-based techniques, including the use of evolutionary algorithms and swarm intelligence, have emerged as powerful tools to improve clustering outcomes by avoiding premature convergence to local optima[12]. Among these approaches, the Artificial Bee Colony (ABC) algorithm has shown promise in enhancing the effectiveness of clustering tasks. The ABC algorithm simulates the foraging behavior of honeybees to solve optimization problems and has been refined in recent studies to address clustering challenges. For instance, a 2021 study applied an improved ABC algorithm to clustering, demonstrating significant enhancements in cluster quality and convergence speed[13]. Among these approaches, the Artificial Bee Colony (ABC) algorithm has shown promise in enhancing the effectiveness of clustering tasks. The ABC algorithm simulates the foraging behavior of honeybees to solve optimization problems and has been refined in recent studies to address clustering challenges. For instance, a 2021 study applied an improved ABC algorithm to clustering, demonstrating significant enhancements in cluster quality and convergence speed[14]. Additionally, the integration of the ABC algorithm with other optimization techniques, such as particle swarm optimization (PSO), has been explored to further improve its robustness in complex clustering tasks. Despite these advancements, challenges remain as the algorithm may still struggle with high-dimensional data and complex function landscapes, emphasizing the need for ongoing refinement[15]

Recent advancements include the ABC-K Means GenData algorithm, which combines the strengths of both K-Means and ABC. This hybrid approach aims to overcome the limitations of traditional K-Means by incorporating ABC's optimization techniques to determine the optimal number of clusters and enhance clustering performance[16]. The ABC-K Means GenData algorithm has been tested on various datasets, showing improvements in accuracy, precision, recall, and Rand Index compared to standard K-Means and Fuzzy K-Means methods[17]

This research offers a significant contribution to the field of clustering by addressing the inherent limitations of the traditional K-Means algorithm through the integration of the Artificial Bee Colony (ABC) optimization technique. Traditional K-Means algorithms often struggle with local optima and require a predefined number of clusters, which can be challenging in complex datasets where the exact number of clusters is unknown. The modified ABC algorithm enhances both exploration and exploitation capabilities, allowing it to overcome the issue of local optima while dynamically determining the optimal number of clusters. This is achieved through the roles of employed, onlooker, and scout bees within the ABC framework, which collaboratively explore the solution space, ensuring robust convergence towards optimal clustering solutions.

Furthermore, the research highlights the practical implementation of the ABC K-Means GenData method, demonstrating its effectiveness on real-world categorical datasets. The results of experimental evaluations show that the ABC K-Means GenData algorithm excels in several performance metrics, including accuracy, precision, recall, and the Rand Index. These metrics are crucial in evaluating the quality of clustering, and the modified algorithm's ability to perform well across these criteria underscores its potential as a powerful tool in data analysis. By combining the strengths of the ABC algorithm with K-Means, the research provides a method that is not only more accurate but also more adaptable to varying data characteristics.

In addition to its practical applications, the research contributes to the broader field of optimization techniques in data analysis. The novel approach of using the ABC algorithm to optimize K-Means clustering introduces a new dimension to clustering methodologies, offering a more flexible and dynamic solution to traditional challenges. This advancement is particularly valuable in scenarios where the number of clusters is not fixed or where the dataset presents high complexity. By refining the ABC K-Means GenData approach, this research sets the stage for further developments in clustering algorithms, paving the way for more efficient and effective data analysis techniques in various fields.

METHODS

Recent advancements in clustering algorithms, particularly over the past five years, have emphasized the integration of metaheuristic approaches to address traditional clustering limitations such as sensitivity to initial parameters and the need for predefined cluster numbers. Hybrid metaheuristic approaches, such as the Improved Artificial Bee Colony (IABC) and Modified Artificial Bee Colony (MABC), have been developed to enhance exploration and exploitation capabilities, improving the accuracy of cluster formation and convergence speed. These enhanced models have shown significant promise in handling high-dimensional data and complex clustering scenarios, demonstrating superior performance over traditional methods[18]. Furthermore, recent literature highlights the importance of dynamic cluster number determination, with approaches like Adaptive K-Means and Evolutionary Clustering models incorporating swarm intelligence techniques to automatically select the number of clusters. For example, Shenghui (2020) demonstrated a dynamic population of cluster centroids that evolve over iterations, providing a flexible framework that adapts to the data structure without requiring predefined cluster numbers[19].

Additionally, recent benchmarking studies have compared traditional algorithms, such as K-Means, against more sophisticated models, including ABC-based hybrids. These studies, such as those by Sadhu et al. (2023) and Zhu et al.(2020 , reveal that hybrid models consistently outperform classical approaches in clustering accuracy, especially in datasets with overlapping clusters or varying densities, thus underscoring the value of modern hybrid techniques in practical clustering applications[20], [21]. ABC-based clustering methods have also found applications in diverse fields, including image segmentation, bioinformatics, and market segmentation, where clustering quality is crucial. For instance, Damos. (2024) applied an ABC-enhanced K-Means algorithm to segment high-resolution satellite images, achieving superior results compared to standalone K-Means and other clustering algorithms. Integrating insights from these recent studies, the ABC K-Means approach is poised to effectively address clustering challenges, offering a robust solution grounded in contemporary advances in clustering and optimization algorithms[22].

Proposed Method

Identification Phase

The Artificial Bee Colony (ABC) algorithm has emerged as a powerful tool for clustering, thanks to its unique approach of using employed, onlooker, and scout bees to optimize cluster formation dynamically. Despite its effectiveness, particularly in balancing exploration and exploitation, the algorithm's performance can be enhanced through hybrid approaches, addressing challenges such as local minima and adapting to various data complexities in several point of indicator below:

1. **ABC Algorithm and Clustering Efficiency**:
   The ABC algorithm uses three types of bees—employed, onlookers, and scouts—to explore and optimize solutions, making it suitable for clustering tasks. Its strength lies in its ability to balance exploration and exploitation during the clustering process. Recent studies have shown that the ABC algorithm can effectively optimize cluster counts by

dynamically adjusting cluster assignments, avoiding the need for a predefined cluster number [23].

2. **Local Minima Challenge**:
   Despite its strengths, the ABC algorithm can be prone to getting trapped in local minima, which affects its clustering performance, particularly in complex data sets. This limitation necessitates further refinement of the algorithm to enhance its robustness. Recent research has focused on hybrid approaches to mitigate this issue, combining ABC with other algorithms like genetic algorithms to improve convergence to the global optimum[24].
3. **Hybrid Approaches for Improved Performance**:
   Hybrid algorithms that integrate ABC with genetic algorithms have been developed to overcome the local minima challenge by introducing mechanisms like genetic selection strategies, which help maintain population diversity and guide the search process more effectively. These hybrid approaches have been found to significantly enhance clustering performance and accuracy, particularly in dynamic clustering problems where the number of clusters is unknown [25].
4. **Dynamic Adjustment of Cluster Numbers**:
   One of the main advantages of the ABC algorithm is its capability to dynamically adjust the number of clusters during the optimization process. This is especially beneficial in real-world data scenarios where the true number of clusters is not known beforehand. Studies have highlighted the ABC algorithm's potential in scenarios requiring adaptive clustering, demonstrating improved results compared to traditional methods that rely on static cluster counts[26].
5. **Applicability to Various Data Sets**:
   The ABC algorithm has been applied to a variety of data sets, showing adaptability and flexibility in clustering diverse types of data, which underscores its potential as a general-purpose clustering tool. Recent applications include its use in high-dimensional data and complex clustering scenarios, where it has outperformed conventional clustering methods by providing more accurate and reliable cluster counts [27].

Dataset

The datasets used to evaluate the performance of the proposed ABC K Means clustering algorithm include the Soybean Small dataset, which contains 47 instances with 35 attributes representing various soybean diseases; the Breast Cancer dataset, comprising 699 instances with 9 attributes used for classifying benign and malignant cases; and the Zoo dataset, consisting of 101 instances described by 16 attributes for classifying animals into 7 categories. These datasets offer diverse challenges for clustering algorithms, making them suitable benchmarks for the proposed approach as shown in table 1,

Table 1. Dataset Descriptions

Dataset	Source	Instances	Attributes	Attribute Types	Classes	Primary Use
Soybean Small	UCI Machine Learning Repository	47	35 (21 used)	Categorical (multivalued)	4 (Soybean diseases: D1-D4)	Clustering and classification of diseases

Breast Cancer	UCI Machine Learning Repository	699	9	Categorical, some missing values	2 (Benign, Malignant)	Classification of breast cancer cases
Zoo	UCI Machine Learning Repository	101	16	Boolean (except for number of legs)	7 (Animal categories)	Clustering and classification of animal types

Data Pattern Development Based on Artificial Bee Colony

The ABC K Means algorithm leverages the artificial bee colony (ABC) for clustering, involving three types of bees: employed bees, onlookers, and scouts. Each food source represents a potential solution, with the nectar amount indicating the solution's quality. The clustering task focuses on optimizing cluster centers, which are essential for determining the final clustering results.

ABC K Means Algorithm step

1. Initialization:

a. **Food Sources**: Begin by creating a population of potential solutions, known as food sources. Each food source represents a candidate set of cluster centers for the K Means algorithm.
b. **Cluster Centers**: Randomly select initial cluster centers from the dataset to populate the food sources. These initial centers serve as the starting points for clustering.
c. **Evaluation**: Evaluate each set of cluster centers using an objective function, such as the within-cluster sum of squares (WCSS). This function measures the quality of the clustering by calculating the sum of squared distances between each data point and its assigned cluster center.

2. Employed Bees Phase:

a. **Generate New Solutions**: Each employed bee generates a new set of cluster centers by making small modifications to an existing solution. This involves slightly adjusting the positions of the cluster centers.
b. **Evaluate Solutions**: Calculate the objective function (e.g., clustering error) for the newly generated set of cluster centers to assess its quality.
c. **Selection**: Compare the quality of the new solution with the old one. If the new solution has a lower clustering error (better quality), replace the old solution with this new one. If not, retain the original solution.

3. Onlooker Bees Phase:

a. **Probability-Based Selection**: Onlooker bees choose which food sources (sets of cluster centers) to explore based on their probability of being selected. This probability is determined by the quality of the solutions (better solutions have higher probabilities).
b. **Generate and Evaluate**: Onlooker bees generate new sets of cluster centers based on the selected food sources and evaluate these new solutions in the same manner as employed bees.
c. **Update Solutions**: If the new solutions produced by onlooker bees are better (i.e., have a lower clustering error), they replace the old solutions. This ensures that only high-quality solutions are retained.

4. Scout Bees Phase:

a. **Abandonment and Replacement**: If a solution does not improve over a specified number of cycles (defined by the abandonment parameter), it is abandoned. A scout bee then searches for new potential solutions to replace the abandoned one.
b. **Search New Solutions**: Scout bees explore the solution space more broadly by randomly searching for new cluster centers, which helps in discovering potentially better solutions that were not previously considered.
c. **Evaluate and Replace**: Evaluate the newly discovered solutions and replace the old, abandoned solutions if the new ones are better.

5. Update Best Solution:

Global Best: After completing all iterations of the algorithm, identify and update the global best solution. This solution represents the optimal or near-optimal set of cluster centers found throughout the process. It is the final clustering result that best minimizes the objective function.

The detailed approach outlined above ensures that the ABC K Means algorithm efficiently explores and exploits the solution space, adapting and improving the clustering results through each phase of the process. There are three types of artificial honeybees: employed bees, onlookers, and scouts. A food source corresponds to a possible solution of the problem to be optimised, and the nectar amount of a food source characterize the quality of the corresponding solution. In the clustering, the clustering results depend on the cluster centers. When the cluster centers are fixed, the clustering results are determined. Therefore, the clustering issue can be seen as the optimization of the cluster centers, and a set of cluster centers correspond a possible solution. For categorical data clustering, let $f_i = \{Q_1, Q_2, \ldots, Q_k\}$ denote a food source, where Q_l is the mode of cluster l. $E(f_i) = E(U, f_i)$ is the objective cost function, and where the minimize cost function aim is to divide X to k cluster.

$$E(U, f_i) = \sum_{l=1}^{k} \sum_{i=1}^{n} u_{il} d(x_i, Q_l) \tag{4.1}$$

Then, the nectar amount of a food source fi is given by:

$$Nu(f_i) = \frac{1}{E(f_i) + 1} \tag{1.2}$$

Similar to the ABC approach, the colony of artificial bees in our algorithm has two parts: the first half of the artificial bees are the employed bees, and the second half of the artificial bees are the onlookers. There exists only one employed bee for a food source, and the number of the employed bees is equal to the number of solutions in the population. Let $P_{fs} = \{f_1, f_2, \ldots, f_H\}$ denote the population of food sources, where H is the number of the food sources, and f_i is the ith food source. Then the probability of the ith food source being picked up by an onlooker is given by:

$$prob_i = \frac{Nu(f_i)}{\sum_{i=1}^{H} Nu(f_i)} \tag{1.3}$$

For deriving a candidate food source from the current one in memory, we introduce the one-step K Means procedure, called OKM, in our algorithm. The OKM procedure is essentially one iteration step in the search process of the K Means algorithm, and it is used to search the neighbor food source based on the current food source in the exploitation process performed by employed bees and onlookers.
Let f_i be the current food source, then the OKM consists of the following two steps.

1. Allocate each data object to the cluster with the nearest centroid, and then form a partition matrix U; specifically, if the ith data object belongs to the lth cluster $u_{il} = 1$; otherwise $u_{il} = 0$, where u_{il} is one element of U;
2. Calculate the new centroid on the basis of the partition matrix U, and thus form a candidate food source $f_i' = \{Q_1', Q_2', \ldots, Q_k'\}$.

For the colony of bees, an employed bee becomes a scout when its food source is exhausted. In our algorithm we adopt the parameter L, which is a predetermined number of trials to control the abandonment of a food resource. If a food source cannot be improved further through L trials, this food source is assumed to be abandoned, and the corresponding employed bee becomes a scout. Let the abandoned food source be f_i, and then the search operation of a scout finding a new food source is given by:

$$f_i' = Rand(Dom(X)) \quad (1.4)$$

where $i \in \{1, 2, \ldots, H\}$, and $Rand(Dom(X))$ is the operation of randomly selecting k data objects from the data set X. In our algorithm, the multi-source search to accelerate the convergence of the proposed algorithm.

The idea of the multi-source search is described as follows: a scout bee searches T candidate food sources at a time, and then picks up the best one as the new food source.

Having introduced the detailed calculation formula for relevant variables, the proposed ABC K Means clustering algorithm is given as follows:

1. Initialization the population of food sources $P_{fs} = \{f_1, f_2, \ldots, f_H\}$ randomly; specifically, for each food source, select k data objects randomly from the dataset X as the means of clusters; set the exploitation numbers of food sources $En_1 = 0, En_2 = 0, \ldots, En_H = 0$.
2. Evaluate the nectar amounts of the food sources $NA(f_1), NA(f_2), \ldots NA(f_H)$, according to Eq (1.2).
3. set CN(the cycles number) to 1
4. For each employed bee
   a. Generate a new food source fi' from the current food source fi by using the one-step K Means procedure OKM, and set $En_i = En_i + 1$;
   b. Evaluate the nectar amount $NA(f_i')$ for the food source f_i' according to Eq (1.2);
   c. If $NA(f_i') > NA(f_i)$, the current food source f_i is replaced by the new food source; otherwise the current food source fi is retained.
5. Evaluate the probability pro_i for each food source f_i according to Eq (1.3);
6. For each onlooker bee
   a. Pick up one food source f_i as the current food source according to the calculated probabilities;
   b. Generate a new food source f_i' from the current food source fi by using OKM, and set $En_i = En_i + 1$;
   c. Evaluate the nectar amount of f_i', that is, $NA(f_i')$;
   d. *If* $NA(f_i') > NA(f_i)$, the current food source f_i is replaced by the new food source f_i'; otherwise the current food source f_i is retained
   e. Update the probability pro_i for each food source f_i according to Eq (1.3).
7. For each food source f_i, if the exploitation number En_i is no less than L, this food source is abandoned, and the corresponding employed bee becomes a scout.
8. If there exists an abandoned food source f_i,
   a. Send the scout in the search space to find T candidate food sources $\{f_1^1, f_2^2, \ldots, f_i^T\}$ *according to* Eq (1.4)
   b. Evaluate the nectar amounts $\{NA(f_1^1), NA(f_2^2), \ldots, NA(f_i^T)\}$ of the food sources $\{f_1^1, f_2^2, \ldots, f_i^T\}$;
   c. Choose the food source with the highest nectar amount as the new food source f_i', and set $En_i = 0$;

d. *If NA(fi') > NA(fi)* the current food source *fi* is replaced by the new food source *fi'*; otherwise the current food source *fi* is retained.
e. CN = CN+1;
f. If CN = MCN, terminate the algorithm and output the best food source; otherwise go to step 4).

The ABC K Means clustering algorithm is a method designed to discover the optimal food source. This algorithm utilizes a bee colony to search for the best solution to the clustering problem. To begin, the algorithm takes in the size of the bee colony (N), the maximum cycle number (MCN), the number of clusters *(k)*, and L. The algorithm works in several steps. First, the population of food sources is initialized randomly and denoted as Pfs. Each food source is assigned k data objects from the dataset X as the means of clusters. The exploitation numbers of the food sources are set to 0. Next, the nectar amounts of the food sources *NA(f1), NA(f2), . . . NA(fH)* are evaluated using Eq (1.2). The cycles number CN is then set to 1. For each employed bee, a new food source fi' is generated from the current food source fi using the one-step K Means procedure OKM, and Eni is incremented.

The nectar amount NA(fi') for the new food source fi' is then evaluated using Eq (1.2). If NA(fi') is greater than NA(fi), then the current food source fi is replaced with fi'; otherwise, fi is retained. The probability proi for each food source fi is then calculated using Eq (1.3). For each onlooker bee, a food source fi is selected based on the calculated probabilities. A new food source fi' is generated from the current food source fi using OKM, and Eni is incremented. The nectar amount of fi', NA(fi'), is then evaluated. If NA(fi') is greater than NA(fi), the current food source fi is replaced with fi'; otherwise, fi is retained. The probability proi for each food source fi is then updated using Eq (1.3). For each food source fi, if the exploitation number Eni is greater than or equal to L, the food source is abandoned, and the corresponding employed bee becomes a scout. If there exists an abandoned food source fi, the scout is sent to the search space to find T candidate food sources *{f11, f22, ..., fiT}* using Eq (1.4). The nectar amounts *{NA(f11), NA(f22), ..., NA(fiT)}* of the food sources *{f11, f22, ..., fiT}* are evaluated. The food source with the highest nectar amount is selected as the new food source fi', and Eni is set to 0. If NA(fi') is greater than NA(fi), the current food source fi is replaced with fi'; otherwise, fi is retained. CN is incremented, and if CN equals MCN, the algorithm is terminated, and the best food source is outputted. The ABC K Means clustering algorithm is a systematic approach to finding the best solution to the clustering problem. It utilizes a bee colony to explore the search space, generate new potential solutions, and evaluate their quality using the nectar amount. The algorithm terminates when the maximum cycle number is reached, and the best food source is outputted as the solution to the clustering problem. For a visual representation and step-by-step implementation, refer to Figure 1, which illustrates the pseudocode and processes involved in the ABC K Means algorithm. To assess the performance of the clustering strategy, a series of experiments were conducted. Experiment 1 focused on evaluating accuracy, Experiment 2 examined precision, and Experiment 3 analyzed recall. Furthermore, Experiment 4 investigated the effectiveness of the Rand Index in evaluating the clustering findings. These experiments utilized synthetic datasets from the UCI repository.

Input:
- *Size of bee colony, N*
- *Maximum cycle number, MCN*
- *Number of clusters, k*
- *Abandonment parameter, L*

Output:
- *Best food source*

1. Initialize population of food sources Pfs = {f1, f2, ..., fH} randomly
- *For each food source, select k data objects randomly from the dataset X as initial cluster centers*
- *Set exploitation numbers En1 = 0, En2 = 0, ..., EnH = 0*

2. Evaluate nectar amounts NA(f1), NA(f2), ..., NA(fH) for each food source using objective cost function

3. Set cycle number CN = 1

4. While CN < MCN do

4.1. For each employed bee do
- *Generate a new food source fi' from the current food source fi using one-step K Means (OKM) procedure*
- *Set Eni = Eni + 1*
- *Evaluate nectar amount NA(fi') for the new food source*
- *If NA(fi') > NA(fi) then*
  - *Replace fi with fi'*
- *Else*
  - *Retain fi*

4.2. Evaluate the probability proi for each food source fi

4.3. For each onlooker bee do
- *Select a food source fi based on calculated probabilities proi*
- *Generate a new food source fi' from fi using OKM*
- *Set Eni = Eni + 1*
- *Evaluate nectar amount NA(fi') for fi'*
- *If NA(fi') > NA(fi) then*
  - *Replace fi with fi'*
- *Else*
  - *Retain fi*
- *Update the probability proi*

4.4. For each food source fi do
- *If Eni >= L then*
  - *Abandon fi and convert the corresponding employed bee to a scout*
  - *Scout searches for T candidate food sources {f11, f22, ..., fiT}*
  - *Evaluate nectar amounts {NA(f11), NA(f22), ..., NA(fiT)}*
  - *Select the best candidate food source fi' from the candidates*
  - *Set Eni = 0*
  - *If NA(fi') > NA(fi) then*
    - *Replace fi with fi'*
  - *Else*
    - *Retain fi*

4.5. Increment cycle number CN

5. Output the best food source as the final clustering solution

Fig. 1. Pseudocode of ABC K Means Clustering Algorithm

Result and Analysis

There some attention about choosing the units such as

In this research, we adopt Yang's accuracy measure and the Rand Index to assess the obtained clustering results. In Yang's method, the definitions of accuracy (AC), precision (PR), and recall (RE) are given as follows:

$$AC = \frac{\sum_{i=1}^{k} a_i}{n} \tag{1.5}$$

$$PR = \frac{\sum_{i=1}^{k} \frac{a_i}{a_i + b_i}}{k} \tag{1.6}$$

$$RE = \frac{\sum_{i=1}^{k} \frac{a_i}{a_i + c_i}}{k}, \tag{1.7}$$

where a_i is the number of data objects that are correctly allocated to class C_i, b_i is the number of data objects that are incorrectly allocated to class C_i, c_i is the number of data objects that are incorrectly denied from class C_i, k is the total number of classes contained in a dataset, and n is the total number of data objects in a dataset. In the above measures, the AC has the same meaning as the clustering accuracy r. Given a dataset $X = \{x_1, x_2, \ldots, x_n\}$ as well as two partitions of this dataset: $Y = \{y_1; y_2; \ldots; y_t\}$ and $Y' = \{y_1'; y_2'; \ldots; y_t'\}$, the Rand Index (RI) is given by:

$$RI = \frac{\sum_{i,j} a_{ij}}{\binom{n}{2}}, \tag{1.8}$$

Where

$$a_{ij} \begin{cases} 1, & if\ there\ exist\ t\ and\ t'such\ that\ both\ x_i\ and\ x_j\ are\ in\ both\ y_i\ and\ y'_{ti}, \\ 1, & if\ there\ exist\ t\ and\ t'such\ that\ x_i\ is\ in\ both\ y_t\ and\ y'_{ti}\ while\ neither\ y_t\ or\ y'_{ti}, \\ 0, & otherwise \end{cases}$$

To evaluate clustering methods on categorical datasets, the true clustering and algorithmic clustering are combined before calculating the Rand Index (RI). Metrics such as purity, entropy, and RI compare clustering results with known class labels or ground truth. Additionally, external validation indices like F-measure, precision, recall, and accuracy measure the alignment between clustering results and true labels, assessing how well the method groups similar instances and separates dissimilar ones. Considering the specific characteristics of the categorical datasets is also crucial for accurate performance evaluation.

Table 2. Result of Proposed Algorithm Perform

Dataset	Algorithm	Best (Avg of Metrics)	Average (Avg of Metrics)	Std Dev (Avg of Metrics)
Zoo	ABC K Means GenData	0.8135	0.8027	0.0115
Zoo	K Means	0.8223	0.7196	0.0789
Zoo	Fuzzy K Means	0.7847	0.6654	0.0935
Breast Cancer	ABC K Means GenData	0.8135	0.7827	0.0115
Breast Cancer	K Means	0.8047	0.7040	0.0789

Breast Cancer	Fuzzy K Means	0.7797	0.6407	0.0935
Soybean	ABC K Means GenData	0.8135	0.7827	0.0115
Soybean	K Means	0.8223	0.7196	0.0789

To ensure a fair evaluation of clustering algorithms, it's essential to establish consistent boundaries and considerations for the datasets used. This includes ensuring that datasets like Zoo, Breast Cancer, and Soybean have comparable sizes and feature types, and applying uniform preprocessing and normalization techniques. Parameters for each algorithm should be optimized and initialized consistently, and performance metrics (Accuracy, Precision, Recall, Rand Index) should be used uniformly. Additionally, multiple runs with different initializations should be conducted for algorithms sensitive to initial conditions, and statistical significance tests should be performed to ensure observed differences in performance are meaningful rather than due to chance. These measures help prevent biases and ensure a balanced comparison of the algorithms. Based on table 2. shown The performance analysis of three clustering algorithms—ABC K Means GenData, K Means, and Fuzzy K Means—was conducted across Zoo, Breast Cancer, and Soybean datasets using metrics like Accuracy, Precision, Recall, and Rand Index. ABC K Means GenData consistently demonstrated superior clustering performance with higher average scores and lower standard deviations compared to the other algorithms, indicating its effectiveness in exploring and exploiting the solution space robustly. This superior performance can be attributed to the integration of the Artificial Bee Colony (ABC) optimization into the K Means algorithm, which enhances global search capabilities and avoids local optima, resulting in more consistent clustering outcomes. In contrast, the traditional K Means algorithm showed competitive best scores, particularly for the Zoo and Soybean datasets, but its performance was characterized by higher variability in average and standard deviation values. This suggests that while K Means can reach optimal solutions, its results are less stable due to its susceptibility to local minima and dependence on initial cluster center selection. Fuzzy K Means, which allows data points to belong to multiple clusters with varying degrees of membership, demonstrated lower overall performance with the highest variability. The flexibility of Fuzzy K Means introduces uncertainty, resulting in less precise cluster definitions compared to ABC K Means GenData and K Means. The analysis highlights that the hybrid ABC K Means GenData approach effectively addresses common limitations of traditional K Means by combining ABC's global optimization capabilities with K Means' partitioning strengths. This leads to higher accuracy and more reliable clustering outcomes, making it a strong candidate for complex data environments where consistency and precision are critical. On the other hand, traditional K Means remains a viable choice for applications prioritizing simplicity and computational efficiency, although it may require enhancements like multiple runs or advanced initialization techniques to improve robustness. Fuzzy K Means can be suitable for applications needing flexible cluster boundaries, though its practical use may be constrained by lower precision and higher variability in results. This study underscores the potential of hybrid optimization techniques like ABC K Means GenData in achieving stable and precise clustering results across diverse datasets.

Conclusion

In conclusion, this study introduces the ABC K Means GenData algorithm, which enhances traditional K Means clustering by integrating the Artificial Bee Colony (ABC) optimization approach. This advanced algorithm effectively addresses the issue of local optima in standard K Means by employing dynamic roles of employed, onlooker, and scout bees to improve exploration and exploitation in clustering. Performance evaluations across various datasets underscore its superiority. For the Zoo dataset, ABC K Means GenData achieved high scores in Accuracy (0.8399), Precision (0.8089), and Recall (0.7286), demonstrating consistent performance compared to K Means and Fuzzy K Means. The Breast Cancer dataset reflected similar results, with ABC K Means GenData matching K Means in Accuracy and Precision, and outperforming Fuzzy K Means in Precision and Recall. In the Soybean dataset, the algorithm excelled with top

scores in Accuracy, Precision, Recall, and Rand Index (RI), surpassing both K Means and Fuzzy K Means. These results highlight ABC K Means GenData's robustness and reliability in clustering categorical data. Future research will focus on applying this algorithm to mixed data types and social media datasets to further enhance clustering techniques.

BIBLIOGRAPHY

[1] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," 2000.
[2] I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD '01. New York, NY, USA: Association for Computing Machinery, 2001, pp. 269–274. doi: 10.1145/502512.502550.
[3] J. Macqueen, "SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS," vol. 233, no. 233, pp. 281–297.
[4] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans Inf Theory*, vol. 28, no. 2, pp. 129–137, 1982, doi: 10.1109/TIT.1982.1056489.
[5] X. Wu and V. Kumar, *The Top Ten Algorithms in Data Mining*. 2009.
[6] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," 1979.
[7] D. Karaboğa, "AN IDEA BASED ON HONEY BEE SWARM FOR NUMERICAL OPTIMIZATION," 2005. [Online]. Available: https://api.semanticscholar.org/CorpusID:8215393
[8] D. Karaboga and B. Basturk, "On the performance of artificial bee colony (ABC) algorithm," *Appl Soft Comput*, vol. 8, no. 1, pp. 687–697, 2008, doi: https://doi.org/10.1016/j.asoc.2007.05.007.
[9] J. Redha and J. Redha Mutar, "A Review of Clustering Algorithms," *International Journal of Computer Science and Mobile Applications*, vol. 10, pp. 44–50, 2022, doi: 10.5281/zenodo.7243829.
[10] S. Naeem, A. Ali, S. Anam, and M. M. Ahmed, "An Unsupervised Machine Learning Algorithms: Comprehensive Review," *International Journal of Computing and Digital Systems*, vol. 13, no. 1, pp. 911–921, 2023, doi: 10.12785/ijcds/130172.
[11] Y. Chen *et al.*, "Fast density peak clustering for large scale data based on kNN," *Knowl Based Syst*, vol. 187, p. 104824, 2020, doi: https://doi.org/10.1016/j.knosys.2019.06.032.
[12] T. A. Khan and S. H. Ling, "A novel hybrid gravitational search particle swarm optimization algorithm," *Eng Appl Artif Intell*, vol. 102, p. 104263, 2021, doi: https://doi.org/10.1016/j.engappai.2021.104263.
[13] X. Pan, Y. Wang, Y. Lu, and N. Sun, "Improved artificial bee colony algorithm based on two-dimensional queue structure for complex optimization problems," *Alexandria Engineering Journal*, vol. 86, pp. 669–679, 2024, doi: https://doi.org/10.1016/j.aej.2023.12.011.
[14] Z. Zhang, J. Lan, and Z. Zhang, "K-means clustering algorithm based on bee colony strategy," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Sep. 2021. doi: 10.1088/1742-6596/2031/1/012058.
[15] I. Arfiani, H. Yuliansyah, and M. D. Suratin, "Implementasi Bee Colony Optimization Pada Pemilihan Centroid (Klaster Pusat) Dalam Algoritma K-Means," *Building of Informatics, Technology and Science (BITS)*, vol. 3, no. 4, pp. 756–763, Mar. 2022, doi: 10.47065/bits.v3i4.1446.
[16] B. Zhou, B. Lu, and S. Saeidlou, "A Hybrid Clustering Method Based on the Several Diverse Basic Clustering and Meta-Clustering Aggregation Technique," *Cybern Syst*, vol. 55, no. 1, pp. 203–229, 2024, doi: 10.1080/01969722.2022.2110682.
[17] S. Ghosh and S. K. Dubey, "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms," 2013. [Online]. Available: www.ijacsa.thesai.org

[18] Q. Tan, H. Wu, B. Hu, and X. Liu, "An improved Artificial Bee Colony algorithm for clustering," in *GECCO 2014 - Companion Publication of the 2014 Genetic and Evolutionary Computation Conference*, Association for Computing Machinery, 2014, pp. 19–20. doi: 10.1145/2598394.2598464.

[19] W. Shenghui and L. Hanbing, "Adaptive K-valued K-means clustering algorithm," in *2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, 2020, pp. 1442–1445. doi: 10.1109/ICMCCE51767.2020.00316.

[20] T. Sadhu, S. Chowdhury, S. Mondal, J. Roy, J. Chakrabarty, and S. K. Lahiri, "A COMPARATIVE STUDY OF METAHEURISTICS ALGORITHMS BASED ON THEIR PERFORMANCE OF COMPLEX BENCHMARK PROBLEMS," *Decision Making: Applications in Management and Engineering*, vol. 6, no. 1, pp. 341–364, Apr. 2023, doi: 10.31181/dmame0306102022r.

[21] S. Zhu, L. Xu, and E. D. Goodman, "Evolutionary multi-objective automatic clustering enhanced with quality metrics and ensemble strategy," *Knowl Based Syst*, vol. 188, p. 105018, 2020, doi: https://doi.org/10.1016/j.knosys.2019.105018.

[22] M. A. Damos *et al.*, "Enhancing the K-Means Algorithm through a Genetic Algorithm Based on Survey and Social Media Tourism Objectives for Tourism Path Recommendations," *ISPRS Int J Geoinf*, vol. 13, no. 2, Feb. 2024, doi: 10.3390/ijgi13020040.

[23] I. Arfiani, H. Yuliansyah, and M. D. Suratin, "Implementasi Bee Colony Optimization Pada Pemilihan Centroid (Klaster Pusat) Dalam Algoritma K-Means," *Building of Informatics, Technology and Science (BITS)*, vol. 3, no. 4, pp. 756–763, Mar. 2022, doi: 10.47065/bits.v3i4.1446.

[24] N. Kaur and S. Aggarwal, "Comparative Analysis of Hybrid K-Mean Algorithms on Data Clustering," 2017. [Online]. Available: www.ijcat.com384

[25] S. Liu and Y. Zou, "An improved hybrid clustering algorithm based on particle swarm optimization and K-means," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, Mar. 2020. doi: 10.1088/1757-899X/750/1/012152.

[26] X. Pan, Y. Wang, Y. Lu, and N. Sun, "Improved artificial bee colony algorithm based on two-dimensional queue structure for complex optimization problems," *Alexandria Engineering Journal*, vol. 86, pp. 669–679, 2024, doi: https://doi.org/10.1016/j.aej.2023.12.011.

[27] M. Zhao, X. Song, and S. Xing, "Improved Artificial Bee Colony Algorithm with Adaptive Parameter for Numerical Optimization," *Applied Artificial Intelligence*, vol. 36, no. 1, 2022, doi: 10.1080/08839514.2021.2008147.

AUTHORS BIBLIOGRAPHY

Harunur Rosyid Harunur Rosyid was Born in Gresik on April 16th, 1975, and completed a Bachelor of Informatics Engineering at Universitas Islam Indonesia (UII) Yogyakarta in 2000. In 2012, he completed a Master of Informatics Engineering at Institut Teknologi Sepuluh Nopember (ITS) Surabaya with a concentration in Software Engineering. In 2023, he completed his Ph.D of Computer Science, field of Artificial Intelligence study at Universiti Pendidikan Sultan Idris Malaysia. He is an Informatics Engineering lecturer and Dean of the Faculty of Engineering at Universitas Muhammadiyah Gresik. His expertise is in the areas of software engineering: AI. His Id Scopus: 57263461000. He can be contacted at email: harun@umg.ac.id.

Muhammad Modi bin Lakulu A lecturer from the FACULTY OF COMPUTING AND META-TECHNOLOGY , Universiti Pendidikan Sultan Idris (UPSI), Tanjung Malim, Perak, Malaysia. Specializes in educational technology, Data Science, Artificial Intelligent

Madya Dr. Ramlah bt. Mailok A lecturer from the FACULTY OF COMPUTING AND META-TECHNOLOGY , Universiti Pendidikan Sultan Idris (UPSI), Tanjung Malim, Perak, Malaysia.