IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 4 December 2023, accepted 2 January 2024, date of publication 5 January 2024, date of current version 26 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3350442

RESEARCH ARTICLE

SA-PSO-GK++: A New Hybrid Clustering Approach for Analyzing Medical Data

AMANI ABDO[1], OMNIA ABDELKADER[2], AND LAILA ABDEL-HAMID[2]
[1]Faculty of Computing, Arab Open University, Cairo 11211, Egypt
[2]Faculty of Computers and Artificial Intelligence, Helwan University (HU), Helwan, Cairo 11795, Egypt
Corresponding author: Amani Abdo (amani.abdo@aou.edu.eg)

ABSTRACT Data clustering is an unsupervised learning task that has been extensively studied, given its wide applicability in various domains. Traditional algorithms often struggle to achieve a balance between exploration and exploitation, leading to sub-optimal solutions. This paper presents a novel hybrid algorithm named SA-PSO-GK++ that synergistically combines Particle Swarm Optimization (PSO), K-means++, Simulated Annealing (SA), and Gaussian Estimation of Distribution to tackle this issue effectively. The proposed SA-PSO-GK++ aims to overcome the drawbacks of existing methods by leveraging the strengths of each individual algorithm. The K-means++ initialization reduces the risk of poor initial centroids, while PSO aids in efficient search space exploration. GED provides a statistical model of the particle space, enabling the algorithm to generate new potential solutions that are statistically guided by the current best solutions. Additionally, the incorporation of Simulated Annealing allows the algorithm to escape local minima, thereby enhancing its global search capability. We evaluate the effectiveness of SA-PSO-GK++ using benchmark datasets from the UCI Machine Learning Repository, including the Iris, Breast cancer, Heart datasets and contraceptive method choice datasets. The proposed method outperforms conventional and some of the state-of-the-art hybrid clustering algorithms in terms of sum of euclidean distance, normalized index, and error rates. These advantages make SA-PSO-GK++ a compelling option for a wide range of clustering applications. The results offer promising avenues for future research in optimizing and applying this innovative clustering technique in diverse domains.

INDEX TERMS Swarm intelligence, particle swarm optimization (PSO), K-means, K-means++, simulated annealing, Gaussian estimation, data clustering, big data, cluster convergence, clustering metrics, local optima.

I. INTRODUCTION

Clustering remains one of the foundational techniques in data analysis [1], a bridge to understanding the intrinsic structures and relationships that pervade datasets without the guideposts of labeled examples [2]. As a key player in unsupervised learning, clustering provides a vantage point from which patterns emerge and unknown classifications become discernible [3]. The K-means algorithm, since its inception, has carved a significant niche in clustering paradigms [4]. Appreciated for its simplicity and efficiency, K-means has a wide gamut of applications, from market segmentation to image processing [5]. However, K-means is not devoid of shortcomings. It has an inherent sensitivity to initial centroid placement and, unfortunately, often finds itself ensnared in local optima, yielding sub-optimal cluster configurations [6]. The K-means++ initialization method was developed to counteract some of these limitations, ensuring a smarter initialization process that often results in faster convergence and more accurate cluster assignments [7]and provides good initial centers [8]. While traditional techniques have their merits, in the last two decades researchers have turned to swarm intelligence algorithms as solutions for clustering [9], [15], [43]. This includes approaches like PSO clustering [10], Firefly clustering [11], and Bat clustering [12] among others. Swarm intelligence, revered in the optimization domain,

The associate editor coordinating the review of this manuscript and approving it for publication was Claudio Loconsole.

takes inspiration from nature, particularly from species that naturally swarm, such as ants, fireflies, and bees [13]. The behaviors of these species, driven by motives like food search, social bonding, and obstacle avoidance, hint at inherent optimization strategies [14]. It is observed that clustering through swarm intelligence has an increased tendency to bypass local optima, making it a compelling choice for addressing clustering issues. To date, a variety of swarm intelligence techniques have been employed to tackle clustering challenges [15]. In essence, PSO is a heuristic optimization technique that harnesses collective intelligence, adjusting its search patterns based on the experience of both individual particles and the entire swarm [10], [44]. PSO's metaheuristic nature makes it a potent tool against the trappings of local optima [16], Particle Swarm Optimization (PSO) has also distinguished itself as a superior method in the realm of clustering, especially in handling complex medical datasets, when compared to other swarm intelligence algorithms like Bat algorithm in clustering context [15]. The inherent flexibility and simplicity of PSO give it an edge in exploring high-dimensional data spaces [22] commonly found in medical dataset. Moreover, studies have shown that PSO's collaborative search mechanism [24] is more adept at avoiding local optima, a common challenge in complex clustering tasks, compared to individualistic strategies employed by other swarm algorithms [54]. Despite the widespread adoption of K-means algorithm [39] in various clustering tasks, the K-means algorithm exhibits notable limitations, particularly in handling complex datasets. First and foremost, K-means is inherently sensitive to the initial choice of centroids, often leading to suboptimal clustering solutions [2] as limitation that has long plagued traditional clustering algorithms like K-means, we favored K-means++ over the conventional K-means based on its established advantages. K-means++ provides a systematic and probability-based approach to centroid initialization [28]. Furthermore, it offers quicker convergence [61], is adept at handling data clusters of diverse sizes and densities, and its superior performance is empirically validated in numerous contexts [58]. Thus, its inclusion in our methodology is a strategic choice to ensure enhanced accuracy and efficiency in clustering results.

Combining the robustness of PSO with traditional clustering has been a topic of significant research interest [19]. By amalgamating the deep search capabilities of PSO with clustering algorithms, researchers aim to harness the strengths of both worlds, yielding enhanced stability, accuracy, and performance [17], [18], [19], [20]. However, as with all methods, the fusion of PSO and K-means has its challenges, particularly in handling high-dimensional data and ensuring a balance between exploration and exploitation [21].

Among the algorithms integrated into our hybrid model is Simulated Annealing (SA), a probabilistic optimization algorithm inspired by the annealing process in metallurgy [49]. SA is renowned for its capability to escape local optima in the search space, making it an essential component for optimizing complex problems [50]. Unlike gradient descent methods, which are prone to getting stuck in local minima, SA allows for a certain probability to accept worse solutions at each step, thereby promoting exploration [51]. This is particularly beneficial when dealing with multi-modal landscapes where local optima are abundant [50]. By incorporating SA into our hybrid model, we aim to address one of the major pitfalls in clustering algorithms: the susceptibility to get trapped in locally optimal solutions. Through SA, the algorithm gains the ability to leap out of local optima, thereby increasing the likelihood of converging to a global optimum [52]. GED operates by modeling the distribution of potential solutions in the search space, providing a statistical foundation for generating new solutions.

In light of these observations, this study proposes an advanced PSO-K-means++ clustering approach SA-PSO-GK++, accentuated with Gaussian Estimation of Distribution and Simulated Annealing techniques. The aim is to maximize the exploration capabilities of PSO, ensure rapid convergence through K-means++, and ultimately deliver a more comprehensive, efficient clustering solution.

The primary objective of this research is to devise an enhanced clustering approach that addresses the pitfalls of traditional methods, particularly in handling high-dimensional data and achieving the right balance between exploration and exploitation. This innovation aims to advance clustering capabilities, avoid trapping in local optima, offering better accuracy, efficiency, and robustness, especially in medical datasets with intricate structures and relationships.

In our proposed SA-PSO-GK++ method, we've meticulously addressed several fundamental clustering challenges. The integration of K-means++ tackles the issue of initialization sensitivity, ensuring that our method starts with a robust set of initial centroids. Our adoption of Particle Swarm Optimization (PSO) and Simulated Annealing (SA) jointly work towards circumventing the pitfalls of local optima, enhancing the global search capability of our algorithm. Additionally, the Gaussian Estimation of Distribution (GED) within our hybrid mechanism adeptly handles the challenges of high dimensionality by modeling the underlying distribution of particle space, allowing for effective dimensionality reduction. The combined strength of these components assures scalability, reduced susceptibility to noise and outliers, and the capability to handle clusters of diverse shapes and sizes.

The main clustering challenges we are solving are:

- Escape from Local Minima:

The combination of simulated annealing (SA) with PSO is designed to avoid getting trapped in local optima. SA has a probabilistic mechanism that allows it to accept worse solutions temporarily, which can lead to escaping local optima. In the context of clustering, this means the ability to find better cluster centroids even if initially stuck in a less-optimal configuration.

- Scalability and Adaptability:

Using K-means++ for initialization can provide benefits in terms of scalability. The method's ability to spread out initial centroids reduces the number of iterations needed for convergence, making it better suited for larger datasets. Furthermore, the hybrid nature of the algorithm means it can potentially adapt to a variety of data shapes and distributions better than an algorithm that uses a singular approach.

The Gaussian Estimation of Distribution operates by modeling the distribution of potential solutions in the search space, providing a statistical foundation for generating new solutions [55]. This approach is particularly effective in guiding the search process towards promising regions, thereby improving the algorithm's convergence rate and robustness and is an integral component of the SA-PSO-GK++ algorithm.

The significant contributions of this paper are:

- Introducing the SA-PSO-GK++, approach, a novel amalgamation of Particle Swarm Optimization (PSO) and K-means++ with Gaussian Estimation of Distribution and Simulated Annealing techniques
- Elucidating the individual and combined strengths of these techniques in tackling the above clustering challenges and ensuring optimal cluster assignments.
- Thoroughly evaluating the proposed algorithm's performance against conventional clustering techniques, drawing insights from its efficacy, and show casing its superiority.

The paper unfolds as follows: Section II delves deep into the relevant literature, contextualizing the study within the broader realm of clustering research. Section III offers a detailed exposition on the fundamentals of PSO and K-means++, simulated annealing and Gaussian estimation of distribution followed by an intricate description of the SA-PSO-GK++, methodology. Section IV dives into the experimental design and datasets employed. Section V showcases the results, providing a comparative analysis with traditional clustering methods. The conclusion in Section VI wraps up the findings and presents avenues for future exploration in this domain.

II. RELATED WORK

Clustering, as a fundamental process in data analysis, has been the subject of extensive research. Traditional algorithms like K-means are popular for their simplicity and efficiency. However, they suffer from issues like convergence to local optima and sensitivity to initialization. To overcome these challenges, researchers have investigated the fusion of clustering with optimization algorithms, particularly Particle Swarm Optimization (PSO).

In 2003 Van der Merwe and Engelbrecht [10] examined the potential of Particle Swarm Optimization (PSO) for data clustering, comparing standard and hybrid PSO strategies integrated with K-means results. Their findings underscored the superiority of PSO-based techniques over traditional K-means in terms of convergence, quantization errors, and intra/inter-cluster distances. The authors signaled intentions to refine the fitness function for more precise distance metrics optimization and to handle high-dimensional datasets more robustly. Moreover, the potential to deduce optimal cluster count dynamically with PSO was also hinted at [22]. They showcased PSO's ability to efficiently locate optimal or near-optimal cluster centers, offering a robust alternative to traditional methods.

In 2005 Omran et al. [23], introduced an innovative image clustering methodology utilizing Particle Swarm Optimization (PSO). Their approach focused on optimizing quantization errors, and intra-cluster distances while enhancing inter-cluster distances. The study juxtaposed the performance of a global best (gbest) PSO and a GCPSO algorithm with traditional clustering algorithms such as K-means, FCM, KHM, H2, and GA. Notably, the PSO-driven techniques typically showcased enhanced inter- and intra-cluster distances, maintaining comparable quantization errors to the other methods.

In 2008 Ahmadyfard and Modares [17] tackled the data clustering challenge, proposing an innovative hybrid method that merges particle swarm optimization (PSO) and the K-means algorithm. By harnessing the strengths of both approaches, this new method avoids their individual limitations. The PSO, adept at thorough global search during initial stages, is employed in the PSO-K-means method's early phase. When the swarm particles are nearing the global optimum, the algorithm transitions to K-means due to its faster convergence capability. This switch is determined using a fitness function. Experiments on both real and synthetic datasets demonstrated that this hybrid method surpasses the individual performance of both K-means and PSO clustering.

In 2014 Prabha and Visalakshi [18] introduced a Min-max normalization method for value conversion of attributes into specific ranges. This study reveals that pre-clustering normalization yields higher-quality clusters. The proposed model facilitates optimal fitness function identification via normalization in unsupervised clustering. Testing on six numerical benchmarks showcased its prowess compared to existing models. This Improved PSO-based K-means clustering approach offers the potential for selecting the optimal number of clusters.

In 2014 Hüseyin Haklıand Harun Uğuz [57] introduced an innovative iteration of the Particle Swarm Optimization (PSO) algorithm, augmented with Levy flight, and aptly named LFPSO. This modification addresses critical limitations of the traditional PSO, particularly its susceptibility to premature convergence and entrapment in local minima, thereby inhibiting efficient global search capabilities.

In 2016 Jensi and Wiselin [56] Their proposed method, termed PSOLF (Particle Swarm Optimization with Levy Flight), extended the work of [57] aims to augment global search capabilities and bolster convergence efficiency. The distinctive aspect of PSOLF lies in its method of updating particle velocity using Levy flight, leading to an innovative position update mechanism.

In 2016 Wang and Sun [48] leveraged the capability of SA to escape local minima and PSO's global optimization strength to address the shortcomings of the traditional K-means approach. Their experimental results indicated that this hybrid SA-PSO algorithm exhibited improved global convergence when compared to a PSO-based K-means algorithm [Reference Number]. This work is particularly relevant as it provides valuable insights into the efficacy of incorporating SA into PSO-based clustering algorithms, laying the groundwork for further innovations such as the SA-PSO-GK++ algorithm.

In 2019 Gong et al. [15] ensured that using PSO for clustering is better than other swarm algorithms, they made an extensive comparative analysis focusing on clustering algorithms derived from swarm intelligence principles, it was observed that Particle Swarm Optimization (PSO) and Bat Algorithm outperformed others in terms of operational efficiency. This study, which evaluated Cuckoo Clustering, Firefly Clustering, PSO, and Bat Algorithm, found that Cuckoo Clustering was markedly slower, especially when compared to PSO. Furthermore, Firefly Clustering tended to lag in scenarios involving a larger number of agents.

In 2019 Gupta et al. [47] developed a hybrid PSO-GA algorithm for medical data clustering, leveraging both Particle Swarm Optimisation (PSO) and Genetic Algorithm (GA). Their approach utilized GA for global search initially, followed by PSO for local search. Tested on six medical datasets including breast cancer from the UCI machine learning repository, the hybrid PSO-GA demonstrated superiority over traditional methods like K-means, standalone PSO, and GA. The study confirmed the hybrid PSO-GA's effectiveness in clustering than regular PSO, with potential applications in solving classical mathematical problems, signifying a promising direction in medical data analysis.

In 2020 Ratanavilisagul [26] innovatively enhanced the widely adopted hybrid clustering technique, which combines Particle Swarm Optimization (PSO) and k-means (KM). Recognizing the limitations of KM and PSO-KM's tendency to trap in local optima, the paper introduces mutation operations for PSO particles to overcome these challenges.

In 2020 Gao et al. [20] extended the work of [56] and [57] and proposed a fusion of Particle Swarm Optimization (PSO) with the k-means clustering technique. What sets their approach apart is the utilization of the Gaussian Estimation of Distribution Method combined with the Lévy Flight mechanism. This combination intends to harness the exploratory power of Lévy Flight with the distribution-driven search capabilities of Gaussian Estimation. As a result, their hybrid technique demonstrated enhanced optimization performance and robustness in clustering tasks.

In 2020 Paul et al. [59] introduced an advanced data clustering methodology by integrating Particle Swarm Optimization (PSO) with the traditional K-means algorithm. This approach, known as MfPSO, leverages the global search capabilities of PSO to address the limitations typically associated with K-means, particularly its tendency to produce locally optimal solutions. Their empirical analysis, which compared the proposed MfPSO algorithm against the standard K-means and other contemporary PSO-based algorithms, demonstrated a marked improvement in clustering performance across various metrics.

In a study by Hua 2021 [27], a hybrid clustering approach was explored that merged swarm intelligence algorithms with K-means. Recognizing that hybrid algorithms, such as the hybrid particle swarm optimization clustering algorithm and others like the hybrid genetic clustering algorithm, have gained prominence in clustering domains. To address this, the researcher integrated the particle swarm algorithm with K-means++ (termed PSOK-means++), Hua introduced the empty-cluster-reassignment technique, refining PSOK-means++ into EPSOK-means++. Building on this, quantum computing theory was incorporated, culminating in the QEPSOK-means++ clustering algorithm.

In 2022 their Krishna and colleagues [60] presented a hybrid clustering algorithm that synergistically combines Particle Swarm Optimization (PSO) with the K-means algorithm. This research addresses the challenge of optimizing both global and local search strategies in clustering tasks. The proposed hybrid algorithm capitalizes on the global search capability of PSO and the local efficiency of the K-means model, aiming to enhance both accuracy and speed in clustering operations. The inertia weight, in particular, plays a pivotal role in this approach.

In 2023, Gu et al. [58] started applying PSO-K-means++ on clustering for privacy protection.they proposed an Particle Swarm Optimization (PSO) based K-means++ clustering method, which integrates multiple differential privacy protection mechanisms. By applying the K-means++ algorithm, the method achieves superior initial clustering centers, which are then refined using differential privacy techniques and incorporating Gaussian kernel functions to allocate privacy and add noise. The final optimization is carried out using the PSO algorithm.

However, as datasets have grown both in size and complexity, the need for advanced optimization techniques has become evident. Herein lies the motivation for our work by incorporating Gaussian Estimation of Distribution (GED) methods and Simulated Annealing strategy into the PSO-K-means++ framework, we aim to tackle the nuances of contemporary datasets.

Our novel approach, SA-PSO-GK++, is designed to offer an amalgamation of robust initialization, adaptive optimization, and enhanced exploration in the search space, addressing many of the challenges presented by predecessors.

III. PROBLEM DEFINITION

Data clustering is a fundamental task in machine learning, statistics, and data mining. The objective is to partition a set of data points into distinct groups, based on

some measure of similarity. However, conventional clustering algorithms like K-means suffer from limitations such as sensitivity to initial centroids and a propensity to get stuck in local optima. Various techniques have been proposed to address these issues, including Particle Swarm Optimization (PSO), Levy Flights [25], K-means, K-means++, genetic algorithm and Gaussian Estimation of Distribution.

Nevertheless, the current state-of-the-art approaches have room for improvement in terms of computational efficiency, robustness, and global optimization capabilities. This study aims to tackle these shortcomings by introducing a novel hybrid clustering algorithm, SA-PSO-GK++, that combines the strengths of Particle Swarm Optimization, K-means++, Simulated Annealing, and Gaussian Estimation of Distribution.

The proposed method aims to:

A. Improve global search capabilities by incorporating Simulated Annealing, thus avoiding local minima.
B. Enhance computational efficiency through the refined local search strategy of K-means++.
C. Integrate Gaussian Estimation of Distribution to optimize the generation of new particle positions in the PSO algorithm.

By addressing these key issues, this paper seeks to present a more robust and efficient approach to data clustering, specifically targeting high-dimensional datasets commonly encountered in the fields of bioinformatics, healthcare, and machine learning benchmarks like Iris, Breast Cancer, and Heart datasets.

IV. THE PROPOSED SA-PSO-GK++ ALGORITHM

A. PARTICLE SWARM OPTIMIZATION (PSO)

PSO is a heuristic optimization method inspired by the social behavior of bird flocking or fish schooling. It starts with a random population of particles where each particle represents a potential solution [10], [41], [42]. The position of a particle is influenced by the best-known position of itself (pBest) and the best-known position of the swarm (gBest) [29]. The choice of fitness function and the balance between exploration (global search) and exploitation (local search) are crucial for the algorithm's performance [30].

The algorithm proceeds as follows:

1) INITIALIZATION

Particles are initialized with random positions and velocities.

2) EVALUATION

Each particle's fitness is evaluated using a given objective function [44].

3) UPDATE BEST POSITIONS

Each particle's best-known position (pbest) is updated if the current position has better fitness. The global best-known position (gbest) among all particles is also updated.

4) UPDATE VELOCITY AND POSITION

The velocity and position of each particle are updated using the equations (1), (2)

$$v_i^{(t+1)} \equiv w \cdot v_i^{(t)} + c_1 \cdot r_1 \cdot \left(pbest_i - x_i^{(t)}\right) + c_2 \cdot r_2 \cdot \left(gbest - x_i^{(t)}\right) \quad (1)$$

$$x_i^{(t+1)} = x_i^{(t)} + v_i^{(t+1)} \quad (2)$$

where w is the inertia weight, c_1 and c_2 are cognitive and social scaling factors, r_1 and r_2 are random numbers, v_i and x_i are the velocity and position of particle i.

B. FITNESS FUNCTION

The fitness function, a measure of a particle's quality, is an essential part of the PSO [30], For this work, we adopted the Sum of Squared Errors (SSE) as the fitness function [40]. The SSE calculates the sum of the squared differences between each observation and its group's mean [31] using the equation (3)

$$\text{SSE} = \sum_{i=1}^{n} \sum_{j=1}^{k} \omega_{ij} \left\| x_i - \mu_j \right\|^2 \quad (3)$$

Here, n is the total number of data points, k is the number of clusters, ω_{ij} is an indicator of the membership of data point x_i in cluster j, x_i is the i-th data point, andμ_j is the mean of cluster j. It serves to identify the variance within the clusters, with a lower SSE value indicating more tightly grouped clusters.

C. K-MEANS++

Traditional K-means clustering can sometimes produce less than optimal results due to the random initialization of cluster centroids [6]. K-means++ improves upon this by determining the initial centroids to be more optimally spaced, thus reducing the chance of random initialization adversely affecting the results. It involves the following steps:

1) INITIALIZATION

- Step 1: Choose one center uniformly at random from the data points.
- Step 2: For each data point x, compute the distance D(x) between x and the nearest center that has already been chosen.
- Step 3: Choose one new data point at random as a new center, using a weighted probability distribution where a point x is chosen with probability proportional to D(x)2.
- Step 4: Repeat Steps 2 and 3 until k centers are chosen.

2) CLUSTER ASSIGNMENT

Assign each data point x_i to the nearest centroid, denoted by C_j, using the equation (4):

$$C_j = \arg\min_{c \in C} \|x_i - c\|^2 \quad (4)$$

where C is the set of centroids.

3) CENTROID UPDATE

Update the centroids by computing the mean of the data points assigned to each cluster using equation (5):

$$C_j = \frac{1}{|C_j|} \sum_{x_i \in c_i} x_i \tag{5}$$

where $|C_j|$ is the number of data points in cluster j, and C_j is the centroid of clusterj.

4) CONVERGENCE

Repeat Steps 2 and 3 until the centroids do not change significantly.

The K-means++ initialization procedure aims to spread out the initial centroids, which can lead to improved convergence and clustering results compared to standard K-means. By initializing the centroids in a way that considers the distribution of the data points, K-means++ often results in faster convergence and can provide a more accurate final clustering [32].

D. GAUSSIAN ESTIMATION OF DISTRIBUTION (GED)

The Gaussian Estimation of Distribution Algorithm (EDA) is a probabilistic model-based optimization algorithm that models the distribution of promising solutions using Gaussian distributions. By sampling this distribution, new candidate solutions are generated, allowing the algorithm to search the solution space for optimal outcomes [33].

GED is integrated into the PSO to guide the particles toward promising regions in the search space. The new candidate solutions are generated based on the Gaussian distribution calculated from the current positions of the particles in the swarm [34]. This addition is aimed to provide a balance between exploration and exploitation in the search space. It is performed as follows:

- Compute the mean and standard deviation of the particle positions using equations (6), (7)

$$\text{Mean\_position} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{6}$$

$$\text{Std\_position} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - mean\_position)^2} \tag{7}$$

- Sample a new position from the Gaussian distribution with the computed mean and standard deviation using equation (8):

$$\text{New\_position} \sim \boldsymbol{N}(\text{mean\_position}, \text{std\_position}) \tag{8}$$

E. SIMULATED ANNEALING

Simulated Annealing is a probabilistic technique inspired by the annealing process in metallurgy. It provides a mechanism to escape local minima, making it a suitable candidate for integration into our hybrid SA-PSO-GK++ algorithm.

SA Parameters and Equations

1. Temperature (T): A parameter that controls the probability of accepting a worse solution than the current one.
2. Cooling Rate (α): The rate at which the temperature decreases.
3. Objective Function (f): The function to be minimized; in our case, this will be the same fitness function used for the PSO and K-means++ parts of the algorithm.

The Metropolis Criterion, which gives the probability P of accepting a worse solution, is defined using equation (9):

$$P = \exp(-\frac{\Delta f}{T}) \tag{9}$$

where $\Delta f = f(new\ solution) - f(current\ solution)$

The proposed hybrid algorithm, SA-PSO-GK++, introduces an avant-garde amalgamation of Particle Swarm Optimization (PSO), K-means++, Gaussian Estimation of Distribution (GED), and Simulated Annealing (SA). This unique blend aims to provide a powerful, robust, and efficient clustering method that mitigates challenges inherent to conventional clustering algorithms.

At its core, SA-PSO-GK++ fuses PSO with K-means++, forming a formidable foundation for highly efficient search space exploration aimed at optimal cluster centroid identification. K-means++ ensures the initial cluster centroids are widely distributed, minimizing the risks associated with local optima. On the other hand, PSO, inspired by swarm intelligence, fine-tunes these centroids by mimicking the social behavior of birds and fish. This dual mechanism catalyzes the algorithm's speedy convergence to a global optimum.

Further boosting this foundation is the Gaussian Estimation of Distribution (GED). Integrated into the algorithmic structure, GED statistically models the distribution of particle positions. This allows for the intelligent generation of new, promising candidate solutions based on the current swarm's best attributes. GED not only augments the robustness of the algorithm but also facilitates a more nuanced exploration of the search space, thereby increasing the likelihood of uncovering the global optimum.

Simulated Annealing (SA), the newest addition to this hybrid algorithm, brings its own set of advantages. SA introduces a probabilistic mechanism to escape local minima by accepting new positions based on a temperature-dependent probability function. As the algorithm progresses, this 'temperature' gradually decreases, allowing the algorithm to settle into a global optimum. The incorporation of SA enhances the algorithm's resilience against premature convergence and local optima, fortifying its explorative and exploitative capabilities.

The collaboration between PSO, K-means++, GED, and SA makes SA-PSO-GK++ a potent clustering algorithm, promising faster convergence and higher accuracy compared to traditional clustering methods. This amalgamation of techniques contributes a novel, effective, and versatile tool to the clustering arena, demonstrating potential applicability across various domains that require reliable and efficient clustering solutions.

PSEUDOCODE OF SA-PSO-GK++ ALGORITHM

1. Load Dataset
2. Initialize number of clusters (K), number of particles (N), max_iterations, w, c1, c2, T, alpha
3. Initialize particles with random positions and velocities
4. Initialize global best position with a random position and calculate its fitness (F_global_best)
5. For iteration = 1 to max_iterations do
   5.1 For each particle i do
      5.1.1 Calculate fitness F_i of the particle's position using K-means++ SSE
      5.1.2 If F_i < F_particle_best[i] then
         Update particle's best position
         F_particle_best[i] = F_i
      End if
   5.1.3 If F_i < F_global_best thenUpdate global best position
         F_global_best = F_i
      End if
   5.1.4 Compute cognitive and social components
      cognitive=c1∗rand()∗(particle_best_position - current_position[i])
      social=c2∗rand()∗(global_best_position -current_position[i])
   5.1.5 Update velocity[i]
      velocity[i] = w ∗ velocity[i] + cognitive + social
   5.1.6 Generate new candidate position using Gaussian Estimation
      Compute mean and std of current particle positions.
      new_position = Gaussian(mean, std)
   5.1.7 Apply Simulated Annealing
      Compute fitness F_new of new_position
      delta_F = F_new - F_i
      If delta_F < 0 or rand() < exp(-delta_F / T) then
         current_position[i] = new_position
      End if
   5.1.8 Update T = alpha * T
   5.2 End For each particle
6. End For iteration
7. Return global best position, fitness, and labels

V. EXPERIMENTAL SETUP AND ANALYSIS

This section presents the results obtained from executing the novel SA-PSO-GK++, algorithm on the three benchmark datasets. The performance of SA-PSO-GK++, is evaluated based on multiple metrics: Sum of Squared Errors (SSE), Normalized Mutual Information (NMI) and Error rate. Additionally, the results are compared against standard K-means, MinMaxK-means, K-means++, PSO–means, PSO-GA, SAPSO-Kmeans and GLPSOK to demonstrate the improvements offered by the proposed method.

A. AN EXPERIMENTAL SETUP

The algorithm was implemented with Python 3 and executed on a machine with an Intel Core i7 processor and 32GB RAM. The proposed SA-PSO-GK++ algorithm and the other comparison algorithms was run 30 times for each dataset, with each run consisting of 100 iterations. The swarm consisted of ten particles.

The cognitive and social components of PSO, responsible for individual and collective learning respectively, were both weighted 1.4. The inertia weight, controlling the influence of a particle's previous velocity, was set to 0.7.

B. EXPERIMENTAL DATA SETS

The datasets employed in this study are widely accepted benchmarks in the field of data clustering obtained from the UCI Machine Learning Repository [36]. Table 1 surmises the datasets:

1) IRIS DATASET (N=178, D=13, K=3)

Consists of 150 instances, each characterized by four features. The dataset is composed of three classes, representing three different species of Iris flowers. It is a popular dataset for its simplicity and clear class separation.

2) BREAST CANCER WISCONSIN (DIAGNOSTIC) DATA SET (n= 569, d D= 30, k = 2)

This dataset includes 569 instances with 30 real-valued features each. It consists of two classes, benign and malignant, making it a binary classification task. This dataset is favored for its real-world relevance and higher dimensionality compared to Iris.

3) HEART DISEASES DATASET (n=303, d= 13, k=2)

Comprises 303 instances, each with 13 features. This dataset, like the Breast Cancer dataset, presents a binary classification task. It is chosen for its mixture of categorical and continuous features, adding to the complexity of the clustering task.

4) CONTRACEPTIVE METHOD CHOICE (CMC) DATA SET (N=1473, D=9, K=3)

The Contraceptive Method Choice dataset consists of 1473 instances, each described by 9 attributes. The attributes include the woman's age, education level, husband's education, number of children, religion, employment status, and more. The dataset has three classes, representing different contraceptive methods used (No-use, Long-term, Short-term). This dataset is selected for its relevance to public health policy and its mix of categorical and numeric features, offering a different kind of challenge for clustering algorithms.

C. COMPARISION ALGORITHMS

In the evaluation phase, we conducted a comprehensive comparative analysis between the proposed algorithm SA-PSO-GK++ and the following algorithms.

1) K-MEANS

A foundational clustering method that is A widely recognized clustering algorithm; K-means forms the foundational

TABLE 1. Properties of the UCI datasets.

Dataset	n (instances)	d (features)	k (clusters/classes)
Iris	150	4	3
Breast Cancer	569	30	2
Heart Disease	303	13	2
CMC	569	30	2

basis of the proposed SA-PSO-GK++ algorithm. It works by partitioning the dataset into K clusters, aiming to minimize the sum of intra-cluster variances. In our study, we made a comparison against K-means clustering results introduced in [46].

2) K-MEANS++

K-means++ is an enhanced initialization method for the classic K-means clustering algorithm. It aims to overcome the sensitivity to initial cluster centroids by choosing initial centroids in a smarter way, thereby improving the quality of the final clusters. The K-means++ initialization algorithm places initial centroids far apart in the data space, reducing the likelihood of suboptimal solutions. This modification often results in faster convergence and better cluster assignments compared to the standard K-means algorithm [46], [61].

3) MINMAXK- MEANS

A refined variant of K-means, MinMaxK-means assigns weights to individual clusters relative to their variances [45]. It optimizes a weighted version of the K-means objective function, where the weights are determined concurrently with cluster assignments through iterative processes. A variant of K-means that seeks to address the sensitivity of initial centroids.

4) PSO-KM

A hybrid that combines Particle Swarm Optimization with K-means. In this study, we include the PSO-KM algorithm, initially proposed by Ahmadyfard and Modares [17], as a comparison algorithm. The PSO-KM hybrid model capitalizes on the global search abilities of PSO during the initial stages of the clustering process and transitions to the K-means algorithm for faster convergence towards the end. A fitness function serves as the decision-making criterion for this transition. This duality ensures an exhaustive search in the solution space while also benefiting from quick convergence, providing a balanced and efficient clustering method. Previous studies have shown the PSO-KM algorithm's effectiveness on various real and synthetic datasets, making it a valuable benchmark for our study.

5) PSO-GA

This is a hybrid of PSO with genetic algorithm for clustering medical data was introduced in [47]. Their approach utilized GA for global search initially, followed by PSO for local search.

6) SAPSO-KMEANS

The SAPSO-KMEANS algorithm aims to overcome the limitations of K-means by enhancing it with the global search capabilities of PSO and the local search benefits of SA. The algorithm is designed to improve the likelihood of reaching the global optimum by allowing the particles to escape local minima [48].

7) GLPSOK

This is another algorithm in comparison is the hybrid PSO-K-means clustering method using Gaussian Estimation and Lévy Flight. This fusion enhances optimization performance by combining the exploratory nature of Lévy Flight with Gaussian Estimation, offering robustness over traditional PSO or K-means approaches, his work serves as a foundational basis for our current study and emphasizes the potential of merging metaheuristic techniques with traditional clustering methods [20].

These algorithms and some of the results were previously compared and evaluated on various datasets and metrics. In our study, we have extended this comparison by including SA-PSO-GK++ algorithm and by running some of these algorithms again and getting better results. This approach allows us to offer new insights and a deeper understanding of the clustering methods in comparison.

D. PARAMETER SETTINGS

To guarantee consistency and fairness across our experiments, every algorithm was executed 30 times independently on each dataset. And we standardized specific parameter settings for all the algorithms evaluated: for all algorithms grounded in PSO, we limited the maximum number of fitness evaluations to 30,000.

For algorithms like K-means, K-means++, and their variants, an iteration typically involves updating the centroids of the clusters and reassigning data points to the closest centroids. The process repeats until the centroids stabilize (i.e., there are no or minimal changes in centroid positions) or the maximum number of iterations is reached. We similarly restricted the maximum number of calculations for their respective objective functions to 30,000.

In the case of Particle Swarm Optimization (PSO) based algorithms (like PSO-KM, PSO-GA, SA-PSO-KMeans, GLPSOK, SA-PSO-GK++), an iteration involves updating the positions and velocities of particles in the search space. The max iteration limit here to 100 to ensures that the search process doesn't continue indefinitely and helps in comparing the performance of different algorithms under similar conditions.

K-means, K-means++ is an exception as it does not require any additional parameters beyond this only k (number of clusters) which will depend on dataset k=3 for iris dataset, k=2 for breast cancer, CMC and heart diseases datasets.

TABLE 2. Parameter settings for SA-PSO-GK++ and the comparison.

Algorithm	Parameter settings
Min max k-means	Pinit = 0, Pmax = 0.5, Pstep = 0.01, β = 0.3
PSO-KM	P_size=50, c1=c2=1.49, ωmax=0.9, ωmin=0.5, max_iter=100
PSO-GA	P_size=50, c1=c2=1.49, ωmax=0.9, ωmin=0.4, max_iter=100, GA_P_size=50, cr=0.8, mr=0.02, Selection=Tournament, Crossover=Uniform, Mutation=Bit Flip
SA-PSO-KMeans	P_size=50, c1=c2=1.49, ωmax=0.9, ωmin=0.5, max_iter=100, T_init= 100, r = 0.95 ,Iter_SA = 100
GLPSOK	ωmax=0.9, ωmin=0.5, c1=c2=1.49445, m=NP/2, ymax=0.7, ymin=0.3, δ=0.05, β=0.5
SA-PSO-GK++	P_size = 50 ,c1=c2=1.49, ωmax=0.9 , ωmin=0.5, T_init= 100, r = 0.95 ,Iter_SA = 100 , epsilon = 0.001,

For the other algorithms being compared, we adhered to the parameter settings as prescribed either in their original publications or their default configurations. A detailed breakdown of these settings can be found in Table 2.

Where:

P_size: Population size of the particles in PSO.

Max_FEs: Maximum number of fitness evaluations.

c_1: Personal learning coefficient.

c_2: Global learning coefficient.

w: max Inertia weight, which varies linearly from 0.9 to 0.5 over iterations.

T_init: Initial temperature for simulated annealing.

r: Cooling rate for simulated annealing.

Iter_SA: Number of iterations for simulated annealing.

epsilon: Convergence_threshold.

E. EVALUATION METRICS

In order to evaluate the effectiveness of the proposed hybrid algorithm, The metrics used in this study – Sum of Euclidean Distances (SED), Normalized Mutual Information (NMI), and error rate – are directly influenced by the addressed clustering challenges. The ability of SA-PSO-GK++ to escape local minima is reflected in the SED metric, as it measures the compactness of clusters. A lower SED indicates that the algorithm effectively found cluster centroids closer to the true centers of the data, avoiding sub-optimal solutions. NMI, as a measure of clustering quality, is enhanced by the algorithm's adaptability and scalability. This metric illustrates how well the algorithm can uncover the inherent structure of the data, regardless of the dataset's complexity or size. Lastly, the error rate is a direct indication of the algorithm's overall clustering accuracy. The integration of SA, K-means++, and Gaussian Estimation of Distribution within the SA-PSO-GK++ framework helps in reducing the error rate, showcasing the algorithm's efficiency in correctly assigning data points to their respective clusters. Thus, improvements in these metrics are a testament to the efficacy of the SA-PSO-GK++ algorithm in overcoming common clustering challenges and delivering superior performance. The detailed of the metrics are below:

1) NORMALIZED MUTUAL INFORMATION (NMI)

NMI is an information-theoretic measure that quantifies the amount of information obtained about one variable through observing the other variable [37]. In the context of clustering, it's used to measure the similarity between the true labels and the labels assigned by the clustering algorithm [38]. The NMI value ranges from 0 (no mutual information) to 1 (perfect correlation), with higher values indicating better using equation (10)

$$NMI(A, B) = \frac{2x\mathrm{I(A; B)}}{\mathrm{H\,(A)} + H(B)} \tag{10}$$

where I(A; B) is the mutual information between clusters A and B, and $H(A)$ and $H(B)$ are the entropies of A and B, respectively.

2) ERROR RATE

One of the evaluation metrics employed in this study is the Error Rate [31], This metric calculates the ratio of incorrectly clustered instances to the total number of instances. Mathematically, it can be represented as equation (11).

$$Error\ Rate = \frac{Number\ of\ Incorrect\ Clustered\ Instances}{Total\ Nuber\ of\ Instances} \tag{11}$$

3) SUM OF SQUARED ERRORS (SSE)

The SSE was used as the fitness function for the PSO. it aims to minimize the total squared deviation of each observation from its cluster mean [30]. A lower SSE value means that the data points are closer to the centroids of their respective clusters, indicating a more accurate clustering.

The SSE is calculated as the sum of the squared distances between each observation and its corresponding cluster centroid. It is formulated as equation (12):

$$SSE = \sum_{i=1}^{n} \sum_{j=1}^{k} \omega_{ij} \left\| x_i - \mu_j \right\|^2 \tag{12}$$

where:

n is the number of observations.

k is the number of clusters.

xi is the i-th observation.

cj is the centroid of the j-th cluster.

wij is 1 if observation i is in cluster j, and 0 otherwise.

F. COMPUTATIONAL COMPLEXITY ANALYSIS

As we delve into the intricacies of clustering algorithms, understanding their computational complexity becomes paramount. This is particularly vital when comparing novel hybrid algorithms against established methods, as it offers insights into their scalability and practical applicability. The

TABLE 3. Complexity calculations for SA-PSO-GK++ and the comparison algorithms.

Algorithm	computational analysis
kmeans	O (N • K • D • MaxFs)
MinMax-means	O (N • K • D • MaxFs)
PSO-KM	O(MaxFs×[P×D+N×K×D])
PSO-GA	O(MaxFs×[P×D+GA operations])
SA-PSO-KMeans	$O(MaxIterSA\times[N\times D\times MaxFs+N2\times D+N\times K+N\times K\times D\times MaxFs+SAoperations])$
GLPSOK	O (K • D • (K + D + N) • MaxFs)
SA-PSO-GK++:	O(MaxIterSA×[N×D×MaxFs+N2×D+N×K+N×K×D×MaxFs])

following table2 presents a detailed computational complexity analysis of the proposed hybrid clustering algorithm, SA-PSO-GK++, in juxtaposition with other prevalent clustering algorithms.

Breakdown of the complexities for each:

1) **Simulated Annealing (SA)**: Given its iterative nature, the main cost comes from evaluating the objective function. If we let *MaxIterSA* denote the maximum number of SA iterations, the complexity can be approximated as $O(MaxIterSA\times cost\_of\_eval)$, where *cost_of_eval* would be the complexity of evaluating the objective function.
2) **Particle Swarm Optimization (PSO)**: Its complexity, as mentioned previously, is $O(N \times D\times MaxFs)$, where N is the number of particles, D is the dimensionality, and *MaxFs* is the maximum number of fitness evaluations.
3) **Gaussian Kernel Estimation of Distribution**: Generally, for Gaussian Kernel methods, calculating pairwise distances is the most computationally intensive step. Its complexity is $O(N2\times D)$.
4) **KMeans++**: For initialization, KMeans++ requires $O(N \times K)$, and for clustering, the standard KMeans complexity is $O(N \times K \times D\times MaxFs)$.

Considering all these parts, a rough estimation for the complexity of our hybrid algorithm would be: $O(MaxIterSA \times[N \times D\times MaxFs+N2\times D+N \times K+N \times K \times D\times MaxFs])$

VI. RESULT ANALYSIS ON DATASETS

In this section, we describe the comparative experimental results obtained by applying the proposed algorithm, SA-PSO-GK++, to four real world datasets: Iris, heart, breast cancer and CMC with three metrics NMI, SED, and error rate.

For the Iris dataset results shown in table 4, the SA-PSO-GK++ outperformed the traditional K-means and its direct upgrade, K-means++, MinMax-means and the compined algorithms PSO-KM, PSO-GA, SA-PSO-KMeans and GLPSOK in terms of the Sum of Euclidean Distances (SED) and Normalized Mutual Information (NMI) and Error Rate for the iris dataset indicating tighter clustering and better class purity alignment. Notably, the error rate was minimized which is a substantial improvement over other algorithms, demonstrating the effectiveness of the proposed method in dealing with well-separated clusters. Fig 1 summarize the results of the metrics on Iris dataset.

In the context of the Breast Cancer dataset shown in table 5, which presents a more challenging and higher-dimensional space, the SA-PSO-GK++ algorithm's performance remained robust, achieving the lowest SED among the compared algorithms but not the lowest error rate although the rate is better than other algorithms among the compared algorithms. The NMI for SA-PSO-GK++ was the highest suggesting that the incorporation of simulated annealing and Gaussian estimation within the PSO and K-means++ framework effectively captures the underlying distribution of data points.. Fig 2 summarize the results of the metrics on Breast Cancer dataset data set.

The comparative results for the Heart Disease dataset shown in table 6 exhibit a continuation of the trend observed in the previous datasets. Here, the proposed SA-PSO-GK++ algorithm achieved the lowest SED, signifying more cohesive clustering when handling medical datasets characterized by mixed feature types. The NMI for SA-PSO-GK++ was superior to some other algorithms but not the best one, suggesting an improved match between the clusters formed and the inherent data distribution error rate has low value but not the lowest and still needs enhancement in the context of heart disess. Fig 3 summarize the results of the metrics on Heart Disease data set.

The analysis on the CMC dataset shown in table 7, which often represents a multi-class clustering challenge, showed that SA-PSO-GK++ maintained a competitive edge. It registered the lowest SED and the highest NMI, which indicates they were more accurate in terms of representing the true data labels. The error rate was also among the lowest, reinforcing the algorithm's capability to maintain high performance even as the complexity of the task increases. Fig 4 summarize the results of the metrics on CMC data set

In assessing the comparative performance of the proposed SA-PSO-GK++ algorithm against the comparison algorithms, we employed the Wilcoxon signed-rank test [53], a non-parametric statistical hypothesis test. This test is particularly suitable for our analysis given its ability to manage non-normally distributed data which is common in algorithmic performance metrics. For each metric—SED, NMI, and error rate, we calculated the differences between the paired observations from 30 independent runs on the four datasets. The Wilcoxon test then assigns ranks to these absolute differences, summing ranks separately for positive and negative differences to obtain.

TABLE 4. Comparative Statistical mean (standard deviation) results of eight algorithms across three metrics on the iris dataset In 30 independent runs.

Algorithm	SED	NMI	Error rate
K-means	1.0344E +02(1.1251E+01)	6.9425E-01(9.1285E-02)	1.089E-01(1.3415E−01)
K-means++	1.0213E +02(1.1251E+01)	6.2481E-01(9.3165E-02)	1.1683E-01(0.0000E+00)
Min max k-means	9.7523E+01(5.7815E-14)	7.1667E-01(1.1292E-16)	1.1333E-01(0.0000E+00)
PSO-KM	1.0122E+02(1.4454E-14)	7.0123E-01(1.1082E-16)	1.0500E -01(4.5168E-16)
PSO-GA	9.70011E+01(5.613E-14)	6.7500E-01(1.1001E-16)	1.3800E-01(0.000E+00)
SAPSO-KMeans	9.7512E+01(5.7725E-14)	6.8500E-01(9.1211E-16)	1.350E-01(4.6168E-16)
GLPSOK	9.6655E+01(3.7413E-14)	7.6036E-01(1.1292E-16)	1.000E-02(4.5143E−16)
SA-PSO-GK++	8.4135E+01(4.2190E-14)	7.7635E-01(1.2181E-16)	0.891E-02(4.4153E−16)

TABLE 5. Comparative statistical mean (standard deviation) results of eight algorithms across three metrics on the breast cancer dataset In 30 independent runs.

Algorithm	SED	NMI	Error rate
K-means	1.5265E+05(1.1841E-10)	4.2291E-01(2.258415E-16)	7.846E-01(1.2154E-01)
K-means++	1.5065E+05(1.1001E-10)	3.8229E-01 (3.22085E-16)	15.872E-01(2.345E-16)
Min max K-mean	1.5606E+05(5.9203E- 11)	2.6215E-01(5.64600E-17)	16.872E-01(0.0000E+00)
PSO-KM	1.52414E+05(4.323E- 11)	4.23309E-01(5.3211E-17)	3.7000E-01(3.2560E-16)
PSO-GA	1.5006E+05(0.0000E+00)	3.6510E-01(0.0000E+00)	5.410E-01(0.0000E+00)
SA-PSO-KMeans	1.5042E+05(0.0000E+00)	3.6770E-01(0.0000E+00)	13.4410E-01(4.0023E-16)
GLPSOK	1.4948E+05(2.4936E+01)	4.5869E-01(2.2584E-16)	13.811E-01(4.5168E-16)
SA-PSO-GK++	1.4005E+05(2.1236E+01)	4.9910E-01(3.2151E-16)	7.0142E-01(4.234E-16)

TABLE 6. Comparative Statistical mean (standard deviation) Results of Eight Algorithms across Three Metrics on the Heart Dataset In 30 independent runs.

Algorithm	SED	NMI	Error rate
K-means	2.2223E+05(1.460E+01)	1.5139E-01 (2.481E-16)	18.13E-01(1.12E-16)
K-means++	2.2143E+05(1.312E+01)	1.6057E-01 (2.245E-16)	16.85E-01(3.21E-16)
Min max k-mean	2.30203E+05(2.56E+01)	1.4834E-01 (3.124E-16)	17.21E-01(2.12E-16)
PSO-KM	2.18711E+05(4.98E+01)	1.4512E-01 (5.146E-16)	15.22E-01(3.25E-16)
PSO-GA	2.41521E+05(2.65E+01)	1.5189E-01 (2.231E-16)	19.97E-01(0.00E+00)
SA-PSO-KMeans	2.1510E+05(1.516E+01)	1.5346E-01 (2.156E-16)	15.01E-01(2.17E-16)
GLPSOK	2.13110E+05(1.65E+01)	1.6821E-01 (3.156E-16)	15.31E-01(1.32E-16)
SA-PSO-GK++	2.10911E+05(1.34E+01)	1.5121E-01 (2.145E-16)	15.45E-01(1.89E-16)

TABLE 7. Comparative Statistical mean (standard deviation) Results of Eight Algorithms across Three Metrics on the CMC Dataset In 30 independent runs.

Algorithm	SED	NMI	Error rate
K-means	5.5431E+03(1.4594E+00)	3.2432E-02(5.7265E-04)	6.01810E-01 (2.2462E-03)
K-means++	5.54291E+03(1.4594E+00)	3.2432E-02(5.7265E-04)	6.02110E-01 (2.3442E-03)
Min max K-mean	5.5426E+03 (0.0000E+00)	3.0597E-02 (1.4115E-17)	6.08280E-01 (1.1294E-16)
PSO-KM	5.5300E+03 (2.6851E-01)	3.1781E-02 (1.2540E-03)	6.0336E-01 (1.11088E-03)
PSO-GA	5.5521E+03(1.3256E+00)	3.18120E-02(0.0000E+00)	6.0197E-01(0.0000E+00)
SA-PSO-KMeans	5.5401E+03 (9.1321E-01)	3.2825E-02 (7.0875E-18)	6.02850E-01 (1.6938E-16)
GLPSOK	5.5323E+03 (2.4391E-01)	3.1329E-02 (1.4115E-17)	6.05570E-01 (2.8230E-16)
SA-PSO-GK++	5.5300E+03 (2.1231E-01)	3.19014E-02(1.45E+00)	6.0310E-01(1.365E+00)

W+ and W− values. These sums are critical in computing the test statistic, which under the null hypothesis follows a known distribution, allowing us to derive the p-value. A low p-value (typically < 0.05) indicates a statistically significant

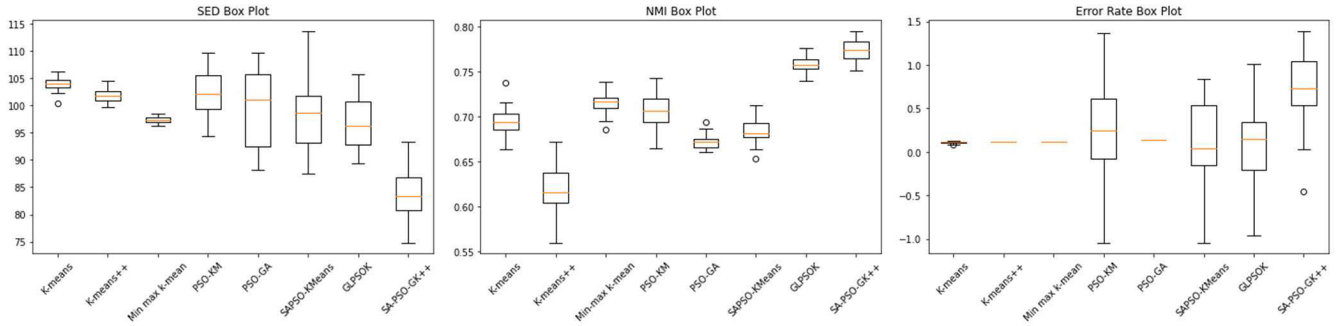

FIGURE 1. Box plots of SED, NMI, Error Rate obtained by eight algorithms on Iris Data set with 30 independent runs.

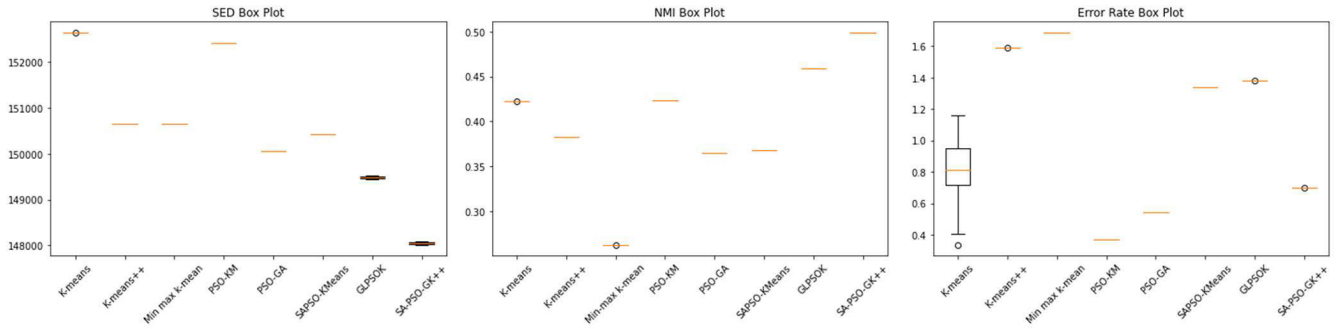

FIGURE 2. Box plots of SED, NMI, Error Rate obtained by eight algorithms on Breast cancer Data set with 30 independent runs.

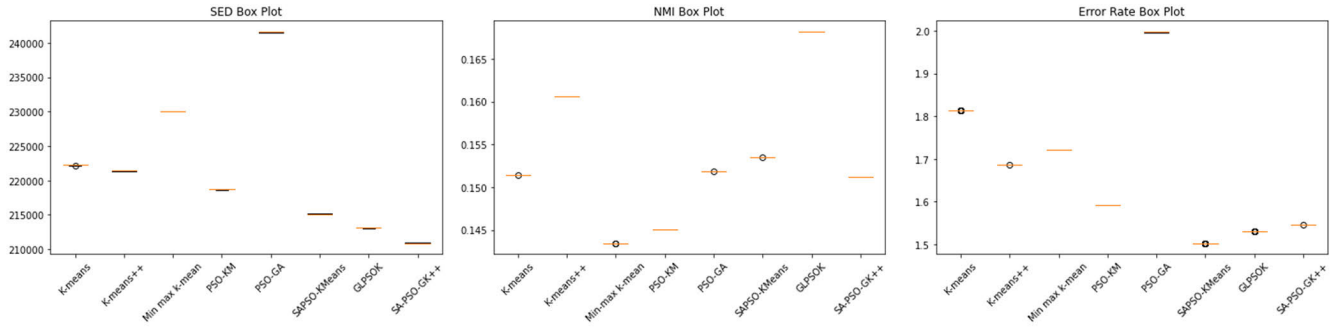

FIGURE 3. Box plots of SED, NMI, Error Rate obtained by eight algorithms on Heart diseases Data set with 30 independent runs.

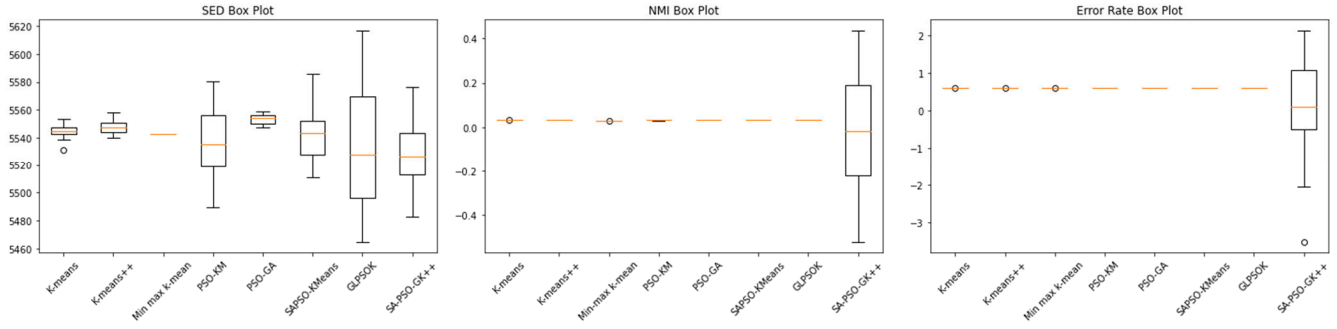

FIGURE 4. Box plots of SED, NMI, Error Rate obtained by eight algorithms on CMC Data set with 30 independent run.

difference in the performance of the algorithms, suggesting that the observed differences are not due to random variation. Conversely, a higher p-value suggests insufficient evidence to assert a significant difference in performance. Moreover,

TABLE 8. The results of the Wilcoxon signed rank test, between K-means vs. SA-PSO-GK++ with the significance level_D 0.05.

Algorithm	Dataset	Metric	W+	W-	R	p-values
K-means vs. SA-PSO-GK++	Iris	SED	461.0	4.0	4.0	1.79e-07
		NMI	120.0	345.0	120.0	8.51e-02
		Error rate	435.0	30.0	30.0	4.54e-05
	Breast Cancer	SED	465.0	0.0	0	5.96e-08
		NMI	0.0	465.0	0.0	5.96e-08
		Error rate	300	0.0	0.0	5.39e-05
	Heart Dises	SED	465.0	0.0	0.0	5.96e-08
		NMI	465.0	0.0	0.0	5.96e-08
		Error rate	8.0	457.0	8.0	4.17e-07
	CMC	SED	465.0	0.0	0.0	5.96e-08
		NMI	323.0	142.0	142.0	2.87e-01
		Error rate	323.0	142.0	142.0	2.87e-01

TABLE 9. The results of the Wilcoxon signed rank test, K-means++ vs. SA-PSO-GK++with the significance level_D 0.05.

Algorithm	Dataset	Metric	W+	W-	R	p-values
K-means++ vs. SA-PSO-GK++	Iris	SED	457.0	8.0	8.0	4.17e-07
		NMI	17.0	448.0	17.0	1.23e-05
		Error rate	465.0	0.0	0.0	5.96e-08
	Breast Cancer	SED	465.0	0.0	0.0	5.96e-08
		NMI	0.0	465.0	0.0	5.96e-08
		Error rate	465.0	0.0	0.0	5.96e-08
	Heart Dises	SED	465.0	0.0	0.0	5.96e-08
		NMI	465.0	0.0	0.0	5.96e-08
		Error rate	465.0	0.0	0.0	5.96e-08
	CMC	SED	465.0	0.0	0.0	5.96e-08
		NMI	0	465.0	0.0	6.25e-02
		Error rate	323.0	142.0	142.0	2.87e-01

TABLE 10. The results of the Wilcoxon signed rank test, Min-Max kmeans vs. SA-PSO-GK++with the significance level_D 0.05.

Algorithm	Dataset	Metric	W+	W-	R	p-values
Min-Max kmeans vs. SA-PSO-GK++	Iris	SED	465.0	0.0	0.0	5.96e-08
		NMI	0.0	465.0	0.0	5.96e-08
		Error rate	435.0	30.0	30.0	4.54e-05
	Breast Cancer	SED	465.0	0.0	0.0	5.96e-08
		NMI	1.0	464.0	1.0	5.96e-08
		Error rate	465.0	0.0	0.0	5.96e-08
	Heart Dises	SED	465.0	0.0	0.0	5.96e-08
		NMI	0.0	465.0	0.0	5.96e-08
		Error rate	465.0	0.0	0.0	5.96e-08
	CMC	SED	465.0	0.0	0.0	5.96e-08
		NMI	323.0	142.0	142.0	2.87e-01
		Error rate	323.0	142.0	142.0	2.87e-01

TABLE 11. The results of the Wilcoxon signed rank PSO-KM vs. SA-PSO-GK++test, with the significance level_D 0.05.

Algorithm	Dataset	Metric	W+	W-	R	p-values
PSO-KM vs. SA-PSO-GK++	Iris	SED	465.0	0.0	0.0	5.96e-08
		NMI	0.0	465.0	0.0	5.96e-08
		Error rate	465.0	0	0.0	5.96e-08
	Breast Cancer	SED	465.0	0.0	0.0	5.96e-08
		NMI	0.0	465.0	0.0	5.96e-08
		Error rate	465.0	0.0	0.0	5.96e-08
	Heart Dises	SED	465.0	0.0	0.0	5.96e-08
		NMI	0	465.0	0.0	5.96e-08
		Error rate	0	465.0	0.0	5.96e-08
	CMC	SED	401.0	64.0	64.0	6.73e-03
		NMI	323.0	142.0	142.0	2.87e-01
		Error rate	323.0	142.0	142.0	2.87e-01

to facilitate the interpretation of these tests, we also provide the z-value—a standardized measure of the test statistic. The z-value indicates how many standard deviations the observed statistic is from the expected value under the null hypothesis,

TABLE 12. The results of the Wilcoxon signed rank test, PSO-GA vs. SA-PSO-GK++with the significance level_D 0.05.

Algorithm	Dataset	Metric	W+	W-	R	p-values
PSO-GA vs. SA-PSO-GK++	Iris	SED	465.0	0.0	0.0	5.96e-08
		NMI	0.0	465.0	0.0	5.96e-08
		Error rate	465.0	0.0	0.0	5.96e-08
	Breast Cancer	SED	465.0	0.0	0.0	5.96e-08
		NMI	0.0	465.0	0.0	5.96e-08
		Error rate	465.0	0.0	0.0	5.96e-08
	Heart Dises	SED	465.0	0.0	0.0	5.96e-08
		NMI	465.0	0.0	0.0	5.96e-08
		Error rate	465.0	0.0	0.0	5.96e-08
	CMC	SED	323.0	142.0	142.0	2.87e-01
		NMI	323.0	142.0	142.0	2.87e-01
		Error rate	0.0	465.0	0.0	2.87e-01

TABLE 13. The results of the Wilcoxon signed rank test, SA-PSO-KMeans vs. SA-PSO-GK++ with the significance level_D 0.05.

Algorithm	Dataset	Metric	W+	W-	R	p-values
SA-PSO-KMeans vs. SA-PSO-GK++	Iris	SED	465.0	0.0	0.0	5.96e-08
		NMI	0.0	465.0	0.0	5.96e-08
		Error rate	465.0	0.0	0.0	5.96e-08
	Breast Cancer	SED	465.0	0.0	0.0	5.96e-08
		NMI	0.0	465.0	0.0	5.96e-08
		Error rate	465.0	0.0	0.0	5.96e-08
	Heart Dises	SED	465.0	0.0	0.0	5.96e-08
		NMI	465.0	0.0	0.0	5.96e-08
		Error rate	0.0	465.0	0.0	5.96e-08
	CMC	SED	465.0	0.0	0.0	5.96e-08
		NMI	323.0	142.0	142.0	2.87e-01
		Error rate	323.0	142.0	142.0	2.87e-01

TABLE 14. The results of the Wilcoxon signed rank test, GLPSOK vs. SA-PSO-GK++ with the significance level_D 0.05.

Algorithm	Dataset	Metric	W+	W-	R	p-values
GLPSOK vs. SA-PSO-GK++	Iris	SED	465.0	0.0	0.0	5.96e-08
		NMI	0.0	465.0	0.0	5.96e-08
		Error rate	465.0	0.0	0.0	5.96e-08
	Breast Cancer	SED	465.0	0.0	0.0	5.96e-08
		NMI	0.0	465.0	0.0	5.96e-08
		Error rate	465.0	0.0	0.0	5.96e-08
	Heart Dises	SED	465.0	0.0	0.0	5.96e-08
		NMI	465.0	0.0	0.0	5.96e-08
		Error rate	0.0	465.0	0.0	5.96e-08
	CMC	SED	465.0	0.0	0.0	5.96e-08
		NMI	465.0	0.0	0.0	5.96e-08
		Error rate	323.0	142.0	142.0	2.87e-01

which, alongside the p-value, provides a robust understanding of the statistical significance of the performance differences observed. Wilcoxon signed rank test results are shown in Table 8 to table 14.

The consistency in the performance of SA-PSO-GK++ across various datasets, underlines the algorithm's robustness and versatility. It demonstrates that the integration of SA and GED within the PSO and K-means++ framework is not only theoretically sound but also practically viable across different data complexities. These results underscore the potential of SA-PSO-GK++ in a wide range of clustering scenarios, including those with multiple classes and mixed feature types commonly encountered in the medical field. Across both datasets, it also showcased superior performance metrics, highlighting the algorithm's adaptability and scalability. The use of simulated annealing helped in avoiding local optima, a common pitfall for traditional algorithms like K-means, while Gaussian estimation provided a probabilistic approach to centroid initialization and particle updates in PSO, enhancing the global search capabilities.

VII. CONCLUSION AND FUTURE WORK

A. CONCLUSION

In this study, we have proposed a novel hybrid clustering algorithm, SA-PSO-GK++, that synergistically combines Particle Swarm Optimization (PSO), K-means++, Gaussian Estimation of Distribution (GED), and Simulated Annealing (SA). Our hybrid model demonstrates the capability of escaping local minima thanks to the Simulated Annealing component, while the PSO and K-means++ components

ensure rapid and efficient global search. The Gaussian Estimation of Distribution further refines the solution by modeling the underlying distribution of the particle space, leading to statistically guided updates.

Experimental results on benchmark datasets like Iris, Breast cancer, Heart and CMC show that our hybrid algorithm outperforms traditional clustering algorithms in terms of NMI, SSE, and error rate. The algorithm is particularly effective in scenarios where initial clustering centroids are unknown or poorly defined, thereby manifesting its robustness.

B. FUTURE WORK

In light of the findings and contributions of this study, several avenues for future research emerge. One of the most immediate next steps is to apply the SA-PSO-GK++ algorithm to a wider array of datasets, including those with higher dimensions and different types of data distributions, to better evaluate its generalizability. Additionally, it would be valuable to explore the integration of other optimization techniques or meta-heuristics with our proposed model to investigate whether they could further enhance the performance. The computational complexity of SA-PSO-GK++ could be another focus, aiming to make the algorithm more scalable for large datasets. A comparative study involving more evaluation metrics could also be beneficial to provide a more comprehensive performance assessment. Furthermore, the algorithm's applicability in real-world scenarios, such as image segmentation, text mining, and bioinformatics, warrants investigation, clustering for privacy protection. These future explorations are anticipated to further validate and extend the utility of the proposed SA-PSO-GK++ algorithm in the field of data clustering.

REFERENCES

[1] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Upper Saddle River, NJ, USA: Prentice-Hall, 1988.

[2] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv. (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[3] P. K. Bharne, V. S. Gulhane, and S. K. Yewale, "Data clustering algorithms based on swarm intelligence," in *Proc. 3rd Int. Conf. Electron. Comput. Technol.*, vol. 4, Kanyakumari, India, Apr. 2011, pp. 407–411, doi: 10.1109/ICECTECH.2011.5941931.

[4] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning With Sparsity: The Lasso and Generalizations*. Boca Raton, FL, USA: CRC Press, 2015.

[5] M. E. Celebi, H. A. Kingravi, and P. A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm," *Exp. Syst. Appl.*, vol. 40, no. 1, pp. 200–210, Jan. 2013.

[6] K. Singh, D. Malik, and N. Sharma, "Evolving limitations in K-means algorithm in data mining and their removal," *Int. J. Comput. Eng. Manag.*, vol. 12, no. 1, pp. 105–109, 2011.

[7] D. Arthur and S. Vassilvitskii, "K-means++ the advantages of careful seeding," in *Proc. 11th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2007, pp. 1027–1035.

[8] Y. Xu, W. Qu, Z. Li, G. Min, K. Li, and Z. Liu, "Efficient k-means++ approximation with MapReduce," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 12, pp. 3135–3144, Dec. 2014.

[9] A. Abraham, S. Das, and S. Roy, "Swarm intelligence algorithms for data clustering," in *Soft Computing for Knowledge Scovery and Data Mining*. Boston, MA, USA: Springer, 2008. 279-313.

[10] D. W. van der Merwe and A. P. Engelbrecht, "Data clustering using particle swarm optimization," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, vol. 1, Dec. 2003, pp. 215–220, doi: 10.1109/CEC.2003.1299577.

[11] T. Hassanzadeh and M. R. Meybodi, "A new hybrid approach for data clustering using firefly algorithm and K-means," in *Proc. 16th CSI Int. Symp. Artif. Intell. Signal Process. (AISP)*, Shiraz, Iran, May 2012, pp. 007–011, doi: 10.1109/AISP.2012.6313708.

[12] X. S. Yang and X. He, "Bat algorithm: Literature review and applications," *Int. J. Bio-Inspired Comput.*, vol. 5, no. 3, pp. 141–149, 2013.

[13] A. Chakraborty and A. K. Kar, "Swarm intelligence: A review of algorithms," in *Nature-Inspired Computing and Optimization: Theory and Applications*. 2017, pp. 475–494.

[14] R. C. Eberhart, Y. Shi, and J. Kennedy, *Swarm Intelligence*. Amsterdam, The Netherlands: Elsevier, 2001.

[15] X. Gong, L. Liu, S. Fong, Q. Xu, T. Wen, and Z. Liu, "Comparative research of swarm intelligence clustering algorithms for analyzing medical data," *IEEE Access*, vol. 7, pp. 137560–137569, 2019.

[16] S. M. A. Salehizadeh, P. Yadmellat, and M. B. Menhaj, "Local optima avoidable particle swarm optimization," in *Proc. IEEE Swarm Intell. Symp.*, Mar. 2009, pp. 16–21.

[17] A. Ahmadyfard and H. Modares, "Combining PSO and k-means to enhance data clustering," in *Proc. Int. Symp. Telecommun.*, Tehran, Iran, Aug. 2008, pp. 688–691, doi: 10.1109/ISTEL.2008.4651388.

[18] K. A. Prabha and N. K. Visalakshi, "Improved particle swarm optimization based K-means clustering," in *Proc. Int. Conf. Intell. Comput. Appl.*, Coimbatore, India, Mar. 2014, pp. 59–63, doi: 10.1109/ICICA.2014.21.

[19] H. A. Atabay, M. J. Sheikhzadeh, and M. Torshizi, "A clustering algorithm based on integration of K-means and PSO," in *Proc. 1st Conf. Swarm Intell. Evol. Comput. (CSIEC)*, Bam, Iran, Mar. 2016, pp. 59–63, doi: 10.1109/CSIEC.2016.7482110.

[20] H. Gao, Y. Li, P. Kabalyants, H. Xu, and R. Martínez-Béjar, "A novel hybrid PSO-K-means clustering algorithm using Gaussian estimation of distribution method and Lévy flight," *IEEE Access*, vol. 8, pp. 122848–122863, 2020, doi: 10.1109/ACCESS.2020.3007498.

[21] Y.-K. Lam, P. W. M. Tsang, and C.-S. Leung, "PSO-based K-means clustering with enhanced cluster matching for gene expression data," *Neural Comput. Appl.*, vol. 22, nos. 7–8, pp. 1349–1355, Jun. 2013.

[22] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. IEEE ICNN*, vol. 4. Nov./Dec. 1995, pp. 1942–1948, doi: 10.1109/ICNN.1995.488968.

[23] M. Omran, A. P. Engelbrecht, and A. Salman, "Particle swarm optimization method for image clustering," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 19, no. 3, pp. 297–321, May 2005.

[24] S. Janson and M. Middendorf, "A hierarchical particle swarm optimizer and its adaptive variant," *IEEE Trans. Syst., Man, Cybern., B*, vol. 35, no. 6, pp. 1272–1282, Dec. 2005, doi: 10.1109/TSMCB.2005.850530.

[25] X. -S. Yang and S. Deb, "Cuckoo search via Lévy flights," in *Proc. World Congr. Nature Biologically Inspired Comput. (NaBIC)*, Coimbatore, India, 2009, pp. 210–214, doi: 10.1109/NABIC.2009.5393690.

[26] C. Ratanavilisagul, "A novel modified particle swarm optimization algorithm with mutation for data clustering problem," in *Proc. 5th Int. Conf. Comput. Intell. Appl. (ICCIA)*, Beijing, China, Jun. 2020, pp. 55–59.

[27] C. Hua, "A quantum-inspired particle swarm optimization K-means++ clustering algorithm," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Orlando, FL, USA, Dec. 2021, pp. 1–6, doi: 10.1109/SSCI50451.2021.9659549.

[28] H. Li and J. Wang, "Collaborative annealing power k-means++ clustering," *Knowl.-Based Syst.*, vol. 255, Nov. 2022, Art. no. 109593.

[29] A. Carlisle and G. Dozier, "An off-the-shelf PSO," in *Proc. Workshop Particle Swarm Optim.*, vol. 1, 2001, pp. 1–6.

[30] T. Peram, K. Veeramachaneni, and C. K. Mohan, "Fitness-distance-ratio based particle swarm optimization," in *Proc. IEEE Swarm Intell. Symp.*, Apr. 2003, pp. 174–181.

[31] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ, USA: Wiley, 2009.

[32] S. Daoudi, C. M. A. Zouaoui, M. C. El-Mezouar, and N. Taleb, "Parallelization of the k-means++ clustering algorithm," *Ingénierie des systèmes d Inf.*, vol. 26, no. 1, pp. 59–66, Feb. 2021.

[33] C. F. Gauss, *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*. vol. 7. 1877.

[34] E. Bengoetxea and P. Larrañaga, "EDA-PSO: A hybrid paradigm combining estimation of distribution algorithms and particle swarm optimization," in *Swarm Intelligence*. Berlin, Germany: Springer, 2010.

[35] F. Kayaalp and P. Erdogmus, "Benchmarking the clustering performances of evolutionary algorithms: A case study on varying data size," *IRBM*, vol. 41, no. 5, pp. 267–275, Oct. 2020.
[36] D. Dua. *UCI Machine Learning Repository*. Accessed: Mar. 20, 2019. [Online]. Available: http://archive.ics.uci.edu/ml
[37] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, vol. 106, no. 4, pp. 620–630, May 1957.
[38] Z.. X. Fern and C. E. Brodley, "Random projection for high dimensional data clustering: A cluster ensemble approach," in *Proc. 20th Int. Conf. Mach. Learn.*, 2003, pp. 186–193.
[39] K. A. Nazeer and M. P. Sebastian, "Improving the accuracy and efficiency of the k-means clustering algorithm," in *Proc. World Congr. Eng.*, vol. 1. London, U.K.: Association of Engineers London, 2009, pp. 1–3.
[40] T. Thinsungnoena, N. Kaoungkub, P. Durongdumronchaib, K. Kerdprasopb, and N. Kerdprasopb, "The clustering validity with silhouette and sum of squared errors," *Learning*, vol. 3, no. 7, 2015.
[41] A. A. A. Esmin, R. A. Coelho, and S. Matwin, "A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data," *Artif. Intell. Rev.*, vol. 44, no. 1, pp. 23–45, Jun. 2015.
[42] M. W. Rand, "Objective criteria for the evaluation of clustering methods," *J. Amer. Stat. Assoc.*, vol. 66, no. 336, pp. 846–850, 1971.
[43] B. H. Nguyen, B. Xue, and M. Zhang, "A survey on swarm intelligence approaches to feature selection in data mining," *Swarm Evol. Comput.*, vol. 54, May 2020, Art. no. 100663.
[44] M. Jain, V. Saihjpal, N. Singh, and S. B. Singh, "An overview of variants and advancements of PSO algorithm," *Appl. Sci.*, vol. 12, no. 17, p. 8392, Aug. 2022.
[45] G. Tzortzis and A. Likas, "The MinMax k-means clustering algorithm," *Pattern Recognit.*, vol. 47, no. 7, pp. 2505–2516, Jul. 2014.
[46] R. Xu and D. Wunsch II, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.
[47] I. K. Gupta, V. Yadav, and S. Kumar, "Medical data clustering based on particle swarm optimisation and genetic algorithm," *Int. J. Adv. Intell. Paradigms*, vol. 14, no. 3, pp. 345–358, 2019.
[48] X. Wang and Q. Sun, "The study of K-means based on hybrid SA-PSO algorithm," in *Proc. 9th Int. Symp. Comput. Intell. Design (ISCID)*, vol. 2, Hangzhou, China, Dec. 2016, pp. 211–214, doi: 10.1109/ISCID.2016.2057.
[49] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.
[50] S. N. Sivanandam, *Genetic Algorithms*. Berlin, Germany: Springer, 2008.
[51] S. Bandyopadhyay, S. Saha, U. Maulik, and K. Deb, "A simulated annealing-based multiobjective optimization algorithm: AMOSA," *IEEE Trans. Evol. Comput.*, vol. 12, no. 3, pp. 269–283, Jun. 2008.
[52] B. Suman and P. Kumar, "A survey of simulated annealing as a tool for single and multiobjective optimization," *J. Oper. Res. Soc.*, vol. 57, no. 10, pp. 1143–1160, Oct. 2006.
[53] F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in Statistics: Methodology and Distribution*. New York, NY, USA: Springer, 1992. 196-202.
[54] R. Poli, J. Kennedy, T. Blackwell, and A. Freitas, "Particle swarms: The second decade," *J. Artif. Evol. Appl.*, vol. 2008, pp. 1–3, May 2008.
[55] X. Wang, H. Zhao, T. Han, Z. Wei, Y. Liang, and Y. Li, "A Gaussian estimation of distribution algorithm with random walk strategies and its application in optimal missile guidance handover for multi-UCAV in over-the-horizon air combat," *IEEE Access*, vol. 7, pp. 43298–43317, 2019.
[56] R. Jensi and G. W. Jiji, "An enhanced particle swarm optimization with Lévy flight for global optimization," *Appl. Soft Comput.*, vol. 43, pp. 248–261, Jun. 2016.
[57] H. Hakli and H. Uguz, "A novel particle swarm optimization algorithm with Lévy flight," *Appl. Soft Comput.*, vol. 23, pp. 333–345, Oct. 2014.
[58] Y. Gu, Y. Zhang, and H. Zhang, "Particle swarm K-means++ clustering method based on multiple differential privacy protection mechanism," in *Proc. IEEE 3rd Int. Conf. Inf. Technol., Big Data Artif. Intell. (ICIBA)*, vol. 3, May 2023, pp. 769–773.
[59] S. Paul, S. De, and S. Dey, "A novel approach of data clustering using an improved particle swarm optimization based K-means clustering algorithm," in *Proc. IEEE Int. Conf. Electron., Comput. Commun. Technol. (CONECCT)*, Jul. 2020, pp. 1–6.
[60] E. R. Krishna, N. Devarakonda, M. Y. H. Al-Shamri, and D. Revathi, "A novel hybrid clustering analysis based on combination of K-means and PSO algorithm," in *Data Intelligence and Cognitive Informatics*. Singapore: Springer, 2022. 139-150.
[61] A. Deshpande, P. Kacham, and R. Pratap, "Robust k-means++," in *Proc. Conf. Uncertainty Artif. Intell.*, 2020, pp. 799–808.

AMANI ABDO was born in 1980. She received the bachelor's degree in computers and information, the master's degree in information systems, and the Ph.D. degree in bioinformatics from the Department of Information Systems, Faculty of Computers and Information, Helwan University, in 2000, 2004, and 2010, respectively. She has supervised many graduation projects in the field of data mining, big data, and artificial intelligence. She also supervised many master's and doctoral dissertations in the fields of machine learning, software engineering, information systems, medical informatics, and bioinformatics.

OMNIA ABDELKADER was born in 1990. She received the bachelor's degree in communication and computer engineering from Helwan University, Egypt, where she is currently pursuing the master's degree with the Software Engineering Program. She is also with Udacity as a Senior Programming Instructor.

LAILA ABDEL-HAMID received the bachelor's, M.Sc., and Ph.D. degrees in information systems from Helwan University, in 2005, 2011, and 2018, respectively. She is a Lecturer with the Faculty of Computer and Artificial Intelligence, Helwan University. Her research focuses on data streaming, data mining, sentiment analysis, and software engineering.