Applied Soft Computing Journal 84 (2019) 105763

Contents lists available at ScienceDirect

Applied Soft Computing Journal

journal homepage: www.elsevier.com/locate/asoc

Improving K-means clustering with enhanced Firefly Algorithms

Hailun Xie [a], Li Zhang [a,*], Chee Peng Lim [b], Yonghong Yu [c], Chengyu Liu [d], Han Liu [e], Julie Walters [a]

[a] Computational Intelligence Research Group, Department of Computer and Information Sciences, Faculty of Engineering and Environment, University of Northumbria, Newcastle, NE1 8ST, UK
[b] Institute for Intelligent Systems Research and Innovation, Deakin University, Waurn Ponds, VIC 3216, Australia
[c] College of Tongda, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu, China
[d] State Key Laboratory of Bioelectronics, School of Instrument Science and Engineering, Southeast University, Nanjing, 210096, China
[e] School of Computer Science and Informatics, Cardiff University, Cardiff, CF24 3AA, UK

ARTICLE INFO

Article history:
Received 27 December 2018
Received in revised form 15 August 2019
Accepted 3 September 2019
Available online 7 September 2019

Keywords:
Data clustering
Firefly Algorithm
Swarm intelligence algorithm
K-means clustering

ABSTRACT

In this research, we propose two variants of the Firefly Algorithm (FA), namely inward intensified exploration FA (IIEFA) and compound intensified exploration FA (CIEFA), for undertaking the obstinate problems of initialization sensitivity and local optima traps of the K-means clustering model. To enhance the capability of both exploitation and exploration, matrix-based search parameters and dispersing mechanisms are incorporated into the two proposed FA models. We first replace the attractiveness coefficient with a randomized control matrix in the IIEFA model to release the FA from the constraints of biological law, as the exploitation capability in the neighbourhood is elevated from a one-dimensional to multi-dimensional search mechanism with enhanced diversity in search scopes, scales, and directions. Besides that, we employ a dispersing mechanism in the second CIEFA model to dispatch fireflies with high similarities to new positions out of the close neighbourhood to perform global exploration. This dispersing mechanism ensures sufficient variance between fireflies in comparison to increase search efficiency. The ALL-IDB2 database, a skin lesion data set, and a total of 15 UCI data sets are employed to evaluate efficiency of the proposed FA models on clustering tasks. The minimum Redundancy Maximum Relevance (mRMR)-based feature selection method is also adopted to reduce feature dimensionality. The empirical results indicate that the proposed FA models demonstrate statistically significant superiority in both distance and performance measures for clustering tasks in comparison with conventional K-means clustering, five classical search methods, and five advanced FA variants.

1. Introduction

Clustering analysis is one of the fundamental methods of discovering and understanding underlying patterns embodied in data by partitioning data objects into several clusters according to measured or perceived intrinsic characteristics or similarity [1]. As a result of the clustering process, data samples with high similarity are grouped in the same cluster, while those with distinctions are categorized into different clusters. Clustering analysis has been widely adopted by many disciplines, such as image segmentation [2–8], text mining [9–11], bioinformatics [12,13], wireless sensor networks [14,15], and financial analysis [16]. In general, conventional clustering algorithms can be broadly categorized into two groups: partitioning and hierarchical methods. The partitioning methods divide data samples into several clusters simultaneously, where each instance can only exclusively belong to one specific cluster. On the other hand, the hierarchical methods build a hierarchy of clusters, either in an agglomerative or divisive mode. K-means (KM) clustering is one of the popular partitioning methods, and is widely used owing to its simplicity, efficiency, and ease of implementation [1].

Despite the abovementioned merits, KM clustering suffers from a number of limitations, such as initialization sensitivity [1,17], susceptibility to noise [18,19], and vulnerability to undesirable sample distributions [19]. Specifically, real-life clustering tasks pose diverse challenges to KM clustering, owing to complexity embedded in data samples, such as immense dimensionality, disturbance of noise and outliers, irregular, sparse, and imbalanced sample distributions, and clusters with overlap or narrow class margins [1]. These complexities overtly violate restrictive

* Corresponding author.
E-mail addresses: hailun.xie@northumbria.ac.uk (H. Xie), li.zhang@northumbria.ac.uk (L. Zhang), chee.lim@deakin.edu.au (C.P. Lim), yuyh@njupt.edu.cn (Y. Yu), bestlcy@sdu.edu.cn (C. Liu), liuh48@cardiff.ac.uk (H. Liu), julie.walters@northumbria.ac.uk (J. Walters).

https://doi.org/10.1016/j.asoc.2019.105763
1568-4946/

assumptions embedded in KM, i.e. spherical sample distributions and evenly sized clusters, therefore leading to limitations in interpretability for such complex data distributions [18,19]. Moreover, KM suffers from initialization sensitivity and local optima traps owing to its operating mechanism of local search around the configuration of initial centroids [1,17]. As characterized by their powerful search capability in terms of exploration and exploitation, metaheuristic search algorithms have been widely employed to assist KM to escape from local optima traps by exploring and obtaining more optimized configurations of cluster centroids. The negative impacts imposed by challenging real-life data can, therefore, be mitigated owing to more accurate cluster identification resulted from the optimized centroids. The effectiveness of such hybrid clustering models has been extensively validated by empirical studies, e.g. Tabu Search (TS) [20,21], Simulated Annealing (SA) [22], Genetic Algorithm (GA) [23], Artificial Bee Colony (ABC) [24,25], Ant Colony Optimization (ACO) [26,27], Particle Swarm Optimization (PSO) [27–29], Cuckoo Search (CS) [29,30], Firefly Algorithm (FA) [31,32], Gravitational Search Algorithm (GSA) [33,34], Black Hole Algorithm (BH) [35], and Big Bang-Big Crunch algorithm (BB-BC) [36].

As one of newly proposed metaheuristic search algorithms, FA possesses unique capability of automatic subdivision in comparison with other metaheuristic search algorithms. This unique property endows FA with advantages in tackling multimodal optimization problems, such as clustering analysis, which entail sub-optimal distraction and high nonlinearity [37–42]. However, the original FA model has limitations in search diversity and efficiency. As an example, with respect to search diversity, the search behaviours in FA are always constrained to a diagonal-based search in principle for any pair of fireflies in comparison. Owing to such a diagonal-based search action, instead of a region-based one, the search process tends to reduce the probability for fireflies to identify more promising search direction, leading to stagnation. On the other hand, with respect to search efficiency, the current search mechanism forces one firefly to approach the brighter ones in the neighbourhood without considering the fitness distinctiveness between them. As a result, many movements become futile and ineffective in navigating the search process to a more promising region, since there is no difference for movement towards neighbouring fireflies with large or small fitness differences to that of the current individual. Therefore, search efficiency is compromised with constrained search diversity. The limitations of KM clustering, these identified deficiencies of FA, and diverse challenges of real-life clustering tasks constitute the major motivations of this research.

This research aims to address the above drawbacks of the original FA model and resolve the initialization sensitivity and local optima traps of conventional KM clustering. Two modified FA models, namely inward intensified exploration FA (IIEFA) and compound intensified exploration FA (CIEFA), are proposed. As one of the main contributions of this research, two novel strategies are formulated to increase search diversification and efficiency. Firstly, a randomized control matrix is proposed in IIEFA to replace the attractiveness coefficient in the original FA model, in order to intensify exploitation diversity. It enables the diagonal-based search action in the original FA model to be elevated to a multi-dimensional region-based search mechanism with greater scales and directions in the search space. Secondly, besides the above strategy, the diversity of global exploration is enhanced in CIEFA by dispersing fireflies with high similarities in the early stage of the search process and relocating them in various directions and scales outside the scope between fireflies in comparison. This enables the distribution of the firefly swarm to expand to a more substantial space, therefore less likely to be trapped in local optima. The search efficiency is also improved by the guarantee of sufficient variance between fireflies in comparison, especially in the early convergence stage. The proposed FA models are incorporated into the KM clustering algorithm to enhance its clustering performance. The minimum Redundancy Maximum Relevance [43] (mRMR)-based feature selection method is adopted to reduce feature dimensionality. A total of 15 UCI data sets, a skin lesion data set, and the ALL-IDB2 database are used to evaluate the proposed models. Five clustering performance indicators, i.e. intra-cluster distances, accuracy, sensitivity, specificity, and Fscore_{M}, are used to indicate the model efficiency. The empirical results indicate that the proposed IIEFA and CIEFA models demonstrate a superior capability of dealing with both high-dimensional as well as low-dimensional clustering tasks, and outperform the KM clustering algorithm, five classical search methods, and five other FA variants statistically.

The rest of the paper is organized as follows. Section 2 introduces conventional KM clustering and FA models, modified FA variants, and the incorporation of metaheuristic algorithms with clustering models for clustering analysis. In Section 3, the proposed FA models, namely IIEFA and CIEFA, are presented comprehensively. Section 4 presents the evaluation of the proposed models and comparison with other methods. Section 5 extends the evaluation of the proposed models to high-dimensional clustering tasks with multiple classes. Section 6 further explains the distinctiveness of the two proposed IIEFA and CIEFA models. Conclusions are drawn and future research directions are presented in Section 7.

2. Related research

In this section, we firstly introduce the conventional KM clustering and FA models. Then, we review FA variants and clustering models, which incorporate metaheuristic algorithms, in the literature.

2.1. K-means clustering

The KM clustering algorithm partitions data samples into different clusters based on distance measures. It finds a partition such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized [1]. Let $O = \{O_1, O_2, \ldots, O_n\}$ be a set of n data samples to be clustered into a set of K clusters, $C = \{C_i, i = 1, \ldots, k\}$. The goal of KM clustering is to minimize the sum of the squared error over all k clusters, which is defined as follows:

$$J(C) = \sum_{i=1}^{k} \sum_{O_l \in C_i} (O_l - Z_i)^2 \tag{1}$$

where C_i, Z_i, O_l, and k represent the ith cluster, the centroid for ith cluster, data samples belonging to the ith cluster, and the total number of clusters, respectively.

In KM clustering, cluster centroids are initialized randomly. Data samples are assigned to the closest cluster, which is determined by the distances between the corresponding centroid and data samples. The centroid of each cluster is updated by calculating the mean value of all data samples within the respective cluster. Then, the process of partitioning data samples into the corresponding clusters is repeated according to the updated cluster centroids until the specified termination criteria are met. The KM clustering algorithm shows impressive performances for a wide range of applications, including computer vision [44], pattern recognition [45] and information retrieval [46]. It often serves as a pre-processing method for other complex models to provide an initial configuration.

Despite the advantages and popularity, KM clustering suffers from a number of limitations owing to its restrictive assumptions and operating mechanisms. One of the key drawbacks of KM is initialization sensitivity [1,17]. Specifically, the process of minimizing the sum of intra-cluster distances in KM is, in essence, a local search surrounding the initial centroids. As a result, the performance of KM heavily depends on the initial configuration of cluster centroids. In addition, owing to its operating mechanisms and the randomness during centroid initialization, KM is more likely to suffer from local optima traps. This drawback of KM clustering serves as one of the main motivations of this research.

2.2. Firefly algorithm

The FA model performs the search operation according to the foraging behaviours of fireflies [47]. In FA, a swarm of fireflies is initiated randomly, and each firefly denotes one initial solution. A fitness score is calculated based on the objective function of each firefly, which is then assigned as the light intensity. According to [47], fireflies with lower light intensities are attracted to those with strong illuminations in the neighbourhood, as defined in Eq. (2).

$$x_i^{t+1} = x_i^t + \beta_0 e^{-\gamma r_{ij}^2} \left(x_j^t - x_i^t\right) + \alpha_t \varepsilon_t \tag{2}$$

where i and j denote fireflies with lower and higher light intensities, respectively, while x_i^t and x_j^t denote the current positions of fireflies i and j at the tth iteration, respectively. Parameter β_0 is the initial attractiveness while γ is the light absorption coefficient, and r_{ij} denotes the distance between fireflies i and j. In addition, α_t is a randomization coefficient, while ε_t is a vector of random numbers drawn from a Gaussian distribution or a uniform distribution.

The major advantage of FA lies in its attraction mechanism. The attractiveness-based movements enable the firefly swarm to automatically subdivide into subgroups, where each group swarms around one mode or a local optimum solution [40,47]. When the population size is sufficiently higher than the number of local optima, the subdivision ability in FA is able to find all optima simultaneously in principle, and, therefore, attain the global optima. This automatic subdivision ability enables the FA model to tackle optimization problems characterized as highly nonlinear and multimodal, which exactly match the characteristics of clustering problems evaluated in this research, namely data sets with many local optima traps and nonlinearity.

2.3. FA variants

While the original FA model demonstrates some unique properties in its search mechanism, it suffers from slow convergence and high computational complexity, owing to its behaviour of following all brighter fireflies in the neighbourhood [48]. Additionally, fireflies can fall into stagnation during the search process, as the distance between fireflies increases and the attractiveness component ($\beta_0 e^{-\gamma r_{ij}^2}$) approaches zero. Many FA variants have been proposed to overcome these problems by increasing the exploration ability and search diversification of the original FA model. The strategies employed to improve the original FA model can be generally categorized into three groups, i.e. adaptive processes of parameter tuning, population diversification, and integration of hybrid search patterns [49]. Ozsoydan and Baykasoglu [50] proposed a quantum firefly swarm model to tackle multimodal dynamic optimization problems. Four strategies were incorporated into their model: (1) multi-swarms based search; (2) two types of movements undertaken by neutral and quantum fireflies respectively in each sub-swarm; (3) simplification of firefly position updating; and (4) employment of two sub-swarm prioritizing techniques, i.e. sequential selection and roulette wheel selection. The quantum firefly swarm model was evaluated with the Moving Peaks Benchmark problem to locate and track the moving optima. The obtained results indicated that the quantum firefly swarm model was competitive and promising in comparison with 13 well-known algorithms in dynamic optimization problems, including mCPSO-with anticonvergence, mCPSO-without anticonvergence, mQSO-with anticonvergence, mQSO-without anticonvergence, SPSO, rSPSO, BSPSO, RWS, and SPSO-PD. Banerjee et al. [51] proposed a Repulsion–Propulsion FA (PropFA) model by incorporating three strategies, i.e. (1) introduction of adaptive mechanisms for both randomization coefficient α_t and light absorption coefficient γ, (2) incorporation of the global best solution as a component for swarm position update, and (3) replacement of the Euclidean distance measurement with Manhattan distance measurement. Three ratios were yielded to construct the adaptive search parameter mechanisms based on a short term memory of the last positions and light intensities of fireflies. The PropFA model was evaluated using 18 classical benchmark functions, 14 additional functions of CEC-2005, and 28 functions of CEC-2013. The results demonstrated the competitiveness of the PropFA model in finding better solutions in comparison with PSO, EDA (Estimation of Distribution Algorithms), RC-EA (Mutation Step Co-evolution), RC-Memetic (Real-Coded Memetic Algorithm), CMA-ES (Covariance Matrix Adaptation Evolution Strategy) on CEC-2005 benchmark functions, and SHADE, CoDE (DE with composite trial vector generation strategies and control parameters), Jade (Adaptive DE with optional external archive) on CEC-2013 benchmark functions. The PropFA model was also employed to estimate the spill area of a fast expanding oil spill, and the PropFA-based confinement strategy proved to be successful.

Baykasoglu and Ozsoydan [52] proposed a variant of FA, i.e., FA2, with two strategies: (1) replacing the exponential function with an inverse function of distance as the attractiveness coefficient, and (2) constructing a threshold probability for a firefly's position to be updated or otherwise. The FA2 model was tested by both static and dynamic multidimensional knapsack problems. The obtained results indicated that FA2 was more effective than GE, DA, and FA. Sadhu et al. [53] proposed a Q-learning induced FA (QFA) model. Q-learning was used to generate light absorption coefficient γ and randomization coefficient α_t with a fitness rank based rewarding and penalizing mechanism. The generated pair, $\langle\gamma, \alpha_t\rangle$, was capable of producing high-performing fireflies in each step. The QFA model was tested with fifteen benchmark functions in CEC 2015, and with a real-world path planning problem of a robotic manipulator with various obstacles. The empirical results confirmed the superiority of the QFA model in terms of solution quality and run-time complexity in comparison with other algorithms, e.g. AFA (adaptive FA), DEsPA (Differential Evolution with success-based parameter adaption), SRPSO (Self-regulating PSO), SDMS-PSO2 (Self adaptive dynamic multi-swarm PSO), SLPSO (social learning PSO), and LFABC (Levy flight Artificial Bee Colony). Zhang et al. [54] proposed a modified FA model for feature selection by incorporating three strategies, i.e. the improved attractiveness operations guided by SA-enhanced neighbouring and global optimal signals, chaotic diversified search mechanisms, and diversion of weak solutions. The modified FA model was tested with feature selection problems using 29 classification and 11 regression benchmark data sets. The experimental results indicated that the proposed FA variant outperformed 11 classical search methods in undertaking diverse feature selection tasks, i.e. PSO, GA, FA, SA, CS, TS, Differential Evolution (DE), Bat Swarm Optimization (BSO), Dragonfly Algorithm (DA), Ant-Lion Optimization (ALO), Memetic Algorithm with Local Search Chain (MA-LS), and 10 popular FA variants, i.e. FA with neighbourhood attraction (NaFA) [48], SA incorporated with FA (SFA) [55], SA incorporated with both Levy flights and FA (LSFA) [55], Opposition and Dimensional FA (ODFA) [56], FA with Logistic map as the randomization search parameter (CFA1) [57], FA with Gauss map as the attractiveness coefficient (CFA2) [58], FA with a variable step wise (VSSFA) [59], FA with a random attraction (RaFA) [60], a modified FA incorporating chaotic Tent map and global best based search operation (MCFA) [61], and a hybrid multi-objective FA (HMOFA) [62].

FA and its variants have also been widely used for solving multimodal optimization problems. Gandomi et al. [41] applied FA to a set of seven mixed variable structural optimization problems with nonlinearity and multiple local optima. The empirical results indicated that FA was more efficient than other metaheuristic algorithms, such as PSO, GA, and Harmony Search (HS), on these optimization tasks. Nekouie and Yaghoobi [38] proposed a hybrid method on the basis of FA for solving multimodal optimization problems. In their study, KM was used to cluster the FA population into several subpopulations. FA with a roaming technique was employed to identify multiple local optima, while SA was used to further improve the local promising solutions. A set of 15 multimodal test functions was used to evaluate the effectiveness of the hybrid model. The empirical results demonstrated its great advantages over other methods such as Niche GSA (NGSA), r2PSO (a l-best PSO with a ring topology and each member interacting with its immediate member on its right), r3PSO (a l-best PSO with a ring topology and each member interacting with its immediate members on both its left and right), r2PSO-lhc (r2PSO with no overlapping neighbourhoods), FER-PSO (Fitness Euclidean-distance Ratio based PSO), and SPSO (Speciation-based PSO). Zhang et al. [39] proposed a modified FA model for ensemble model construction for classification and regression problems. Their FA variant embedded attractiveness strategies guided by both neighbouring and global promising solutions, as well as evading mechanisms with the consideration of local and global worst experiences. Their FA variant was evaluated with standard, shifted, and composite test functions, as well as the Black-Box Optimization Benchmarking test suite and several high-dimensional UCI data sets. The experimental results indicated that their FA model outperformed several state-of-the-art FA variants and classical search methods in solving diverse complex unimodal and multimodal optimization and ensemble reduction problems. Yang [42] proposed a multi-objective FA model (MOFA) for solving optimization problems with multiple objectives and complex nonlinear constraints. Evaluated with five mathematical artificial landscapes with convex, nonconvex, discontinuous Pareto fronts, and complex Pareto sets, the empirical results indicated that MOFA outperformed seven established multi-objective algorithms, i.e. vector evaluated GA (VEGA), Non-dominated Sorting GA II (NSGA-II), multi-objective DE (MODE), DE for multi-objective optimization (DEMO), multi-objective Bees algorithms (Bees), and Strength Pareto Evolutionary Algorithm (SPEA). A comprehensive review on evolutionary algorithms for multimodal optimization is also provided in [63].

Despite the abovementioned studies, there are certain limitations in search diversification imposed by the strict obedience of biological laws in the original FA model. These limitations are rarely addressed in the existing literature. Specifically, the position updating strategy in FA in Eq. (2) is constructed according to the firefly foraging behaviours, which is employed to guide one firefly to approach another with a higher light intensity by multiplying the position difference of these two fireflies $(x_j^t - x_i^t)$ with their relative attractiveness component $(\beta_0 e^{-\gamma r_{ij}^2})$. While the inheritance of biological laws enables one firefly to approach another with a more favourable position, the dimensionality and diversity through the approaching process are severely constrained, since the movement can only happen on the diagonal direction composed by two fireflies, in accordance with the formula. As illustrated in Fig. 1, in a two-dimensional scenario, there are two fireflies, i and j. If we view both fireflies as vectors, the position difference of these two fireflies, i.e., $(x_j^t - x_i^t)$, can be represented by the dotted line denoted by Δp in Fig. 1. The calculation of attractiveness practically imposes one constant isotropic factor $(\beta_0 e^{-\gamma r_{ij}^2})$ on all dimensions of the position difference between fireflies i and j, causing the lack of variance among different dimensions. As a result, instead of exploring flexibly in the entire solution space, the fireflies can merely move along the specific diagonal trajectory between two fireflies in comparison, and the search area is shrunk drastically from a two-dimensional rectangular region enclosed with dash lines into a one-dimensional vector along the dotted line, as shown in Fig. 1. Therefore, the chances of finding the global optima are reduced, since search diversification is constrained severely owing to the limitations of the biological laws in the original FA model. In order to mitigate the limitations, matrix-based search parameters and dispersing mechanisms are proposed in this research, which are incorporated into the proposed FA models to enhance exploitation and exploration.

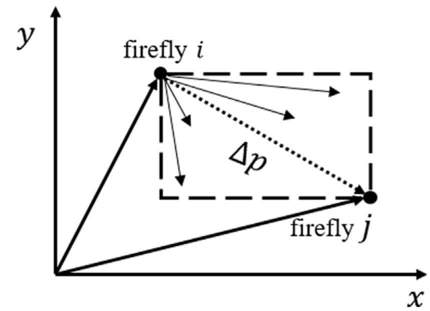


Fig. 1. The movement of fireflies in a two-dimensional search space (Δp denotes the position difference between fireflies i and j).

2.4. Clustering models integrated with metaheuristic algorithms

A number of metaheuristic search algorithms have been employed to overcome the problems of initialization sensitivity and local optima traps of classical clustering algorithms. Karaboga and Ozturk [24] proposed an ABC-based clustering method by incorporating the original ABC model with KM clustering. The ABC-based clustering method was evaluated using 13 UCI data sets. The obtained results demonstrated the competitiveness of the combination of ABC with KM clustering in managing clustering tasks in comparison with those of PSO and nine classification techniques (e.g. Bayes Net, MultiLayer Perceptron Artificial Neural Network (MLP), Radial Basis Function Artificial Neural Network (RBF), Naïve Bayes Tree (NBTree), and Bagging). Shelokar et al. [26] incorporated the original ACO model with KM clustering. Two simulated and three UCI data sets were used to evaluate the performance of the proposed ACO-based clustering method. The ACO-based clustering method showed advantages in comparison with SA, GA, and TS in terms of quality of solution, average number of function evaluations, and processing time. Chen and Ye [28] proposed a PSO-based clustering method (PSO-clustering) and evaluated its performance on four artificial data sets. The obtained results indicated a better performance of PSO-clustering over those of KM and Fuzzy C-Means clustering algorithms. Senthilnath et al. [31] employed FA for clustering analysis. The performance of the FA-based clustering method was tested with 13 UCI data sets. The FA model demonstrated superiority in terms of clustering error rates and computational efficiency over ABC, PSO, and nine other traditional classification methods (e.g. Bayes Net, MLP, and RBF).

Hatamlou et al. [33] formulated a hybrid clustering method, namely GSA-KM, by combining GSA and KM clustering. The GSA-KM method was tested with five UCI data sets. It demonstrated advantages in terms of quality of solutions and convergence speed in comparison with seven well-known algorithms, i.e. KM clustering, GA, SA, ACO, PSO, GSA, and Honey Bee Mating Optimization (HBMO). Hatamlou [35] then produced a Black Hole (BH) Algorithm to enhance the KM clustering performance. The BH-based clustering method was tested with six UCI data sets. It demonstrated a better performance in comparison with those of KM clustering, GSA, and PSO. Moreover, Hatamlou et al. [36] also applied the Big Bang-Big Crunch algorithm (BB-BC) to clustering analysis. The BB-BC results outperformed those of KM clustering, GA, and PSO with several UCI data sets.

A number of modified metaheuristic search algorithms are available to further improve the performance of the original metaheuristic algorithm-based clustering methods. Das et al. [25] proposed a modified Bee Colony Optimization (MBCO) model by adopting both fairness and cloning concepts. The introduction of a fairness concept allowed bees with low probabilities to have a chance to be selected for enhancing search diversity. The employed cloning concept enabled the global best solution to be kept in the next iteration to accelerate convergence. Two hybrid clustering methods, namely MKCLUST and KMCLUST, were subsequently constructed based on MBCO. Additionally, a probability based selection method was introduced to allocate the remaining unassigned data samples to clusters. The MBCO method was evaluated with seven UCI data sets. It outperformed some existing algorithms, e.g. ACO, PSO, and KM clustering, while the proposed hybrid MKCLUST and KMCLUST models, on average, outperformed some existing hybrid methods, e.g. K-PSO (combination of PSO and KM), K-HS (combination of Harmony Search and KM), and IBCOCLUST (improved BCO clustering algorithm). In Niknam and Amiri [27], a hybrid evolutionary clustering model, namely FAPSO-ACO-K, was proposed by combining three traditional algorithms, i.e. FAPSO (fuzzy adaptive PSO), ACO, and KM. The proposed model was tested with four artificial and six UCI data sets. FAPSO-ACO-K was able to resolve the problem of initialization sensitivity in KM clustering. It outperformed other algorithms, such as PSO, ACO, SA, PSO-SA (combination of PSO and SA), ACO-SA (combination of ACO and SA), PSO-ACO (combination of PSO and ACO), GA, and TS. Boushaki et al. [30] constructed a quantum chaotic Cuckoo Search (QCCS) algorithm using chaotic maps and nonhomogeneous update based on the quantum theory to increase global exploration. The QCCS model was tested with six UCI data sets. QCCS outperformed eight well-known methods, including GQCS (genetic quantum CS), HCSDE (hybrid CS and DE), KICS (hybrid KM and improved CS), CS, QPSO (quantum PSO), KCPSO (hybrid KM chaotic PSO), GA, and DE, for solving clustering problems. In Zhou and Li [32], two FA variants, namely the probabilistic firefly KM (PFK) and the greedy probabilistic firefly KM (GPFK), were proposed for data clustering. The PFK model employed a cluster channel array to store the probability of each data object belonging to each cluster in the encoding system. Instead of moving towards all brighter fireflies as in PFK, the GPFK algorithm adopted a greedy search strategy, in which each firefly only moved towards the brightest firefly in the swarm. The PFK and GPFK models outperformed KM clustering and FA based on the evaluation of four UCI data sets. Hassanzadeh and Meybodi [64] proposed a modified FA model (MFA) for clustering analysis. The MFA model not only employed neighbouring brighter fireflies but also the global best solution to provide guidance for the search process. The MFA model was evaluated with five UCI data sets. It outperformed three other clustering methods, including KM, PSO, and KPSO. Han et al. [34] proposed a modified GSA model for clustering analysis, namely BFGSA. The mean position of the seven nearest neighbours of the global best solution was used to enable the leader to escape from the local optima traps. Based on 13 UCI data sets, BFGSA outperformed nine classical search methods, including GSA, PSO, ABC, FA, KM, NM-PSO (fusion of Nelder–Mead simplex and PSO), K-PSO (fusion of KM and PSO), K-NM-PSO (fusion of KM, Nelder–Mead simplex and PSO), and CPSO (Chaotic PSO) [34]. A comprehensive survey on metaheuristic algorithms for partitioning clustering can be found in Nanda and Panda [65].

3. Methodology

We construct the hybrid clustering models on the basis of FA owing to its unique property of automatic subdivision and its advantages in tackling multimodal optimization problems [40,47]. However, the identified limitations pertaining to search diversity and search efficiency in the original FA model may impose certain constraints on identification of optimal centroids in clustering analysis. Therefore, in this research, we propose two modified FA models, namely IIEFA and CIEFA, to overcome limitations of the original FA model and mitigate the problems of initialization sensitivity and local optima traps of KM clustering. The proposed models intensify the diversification of exploration both in the neighbourhood and global search space, and lift the constraints of the biological laws in the original FA model. We introduce the proposed models in detail in the following sub-sections.

3.1. The proposed inward intensified exploration FA (IIEFA) model

The aim of IIEFA is to expand the one-dimensional search in the original FA model to a multi-dimensional scale by replacing the attractiveness term $\beta_0 e^{-\gamma r_{ij}^2}$ with a random matrix μ, as illustrated in Eq. (3).

$$x_i^{t+1} = x_i^t + \mu \times \left(x_j^t - x_i^t\right) + \alpha_t \varepsilon_t \tag{3}$$

$$\alpha_{t+1} = \alpha_t \times \theta \tag{4}$$

where μ denotes a control matrix where each element is drawn from [0, 1] randomly, while α_t denotes an adaptive randomization step based on a geometric annealing schedule, with θ as an adaptive coefficient. According to [47], θ is recommended to have a value in the range of 0.95 to 0.99. We set θ to 0.97 in this study, in accordance with the recommendation in [47] and several trial-and-error results in our experiments. This adaptive randomization step enables the search process to start with a larger random step to increase global exploration and fine-tune the solution vectors in subsequent iterations with a smaller search parameter.

By multiplying the control matrix, μ, each dimension of the position difference $(x_j^t - x_i^t)$ between two fireflies is assigned with a unique random number in [0,1], therefore being shrunk disproportionately with various magnitudes. Subsequently, the resulting solutions after this operation can be any vectors originated from the current firefly solution, randomly distributed in the rectangular region, as compared with those along the dotted diagonal line as in original FA model, as illustrated in Fig. 1. The random control matrix operation possesses two-fold advantages. Firstly, the search directions in the neighbourhood are not constrained to the diagonal line, but become more diversified. Secondly, the movement scales become more diverse owing to the impact of various magnitudes on each dimension. Fig. 1 provides an example of the possible directions and scales in the neighbourhood search, indicated by vectors with arrows within the rectangular region. Therefore, IIEFA possesses a better search capability by extending exploration of fireflies from a one-dimensional diagonal direction to a multi-dimensional space in the neighbourhood. In other words, exploration of the swarm increases along with the firefly congregation process. This first proposed FA variant is hereby characterized as an inward intensified exploration FA model. The pseudo-code of IIEFA is presented in Algorithm 1.

3.2. The proposed compound intensified exploration FA (CIEFA) model

In the original FA model, after being initiated, the whole firefly swarm tends to congregate continuously until convergence at one point. As such, the search process can be deemed as an inward contracting process, no matter how early the search stage is, or how close or similar two neighbouring fireflies are. Consequently, the approaching movement between fireflies with similar light intensities (i.e. fitness scores) at an early stage is more likely to result in waste of the resource, since the fitness score of the current firefly is very unlikely to be drastically improved under this circumstance by following the neighbouring slightly better solution, but with a high probability of being trapped in local optima. Therefore, we propose the second FA variant, i.e. a compound intensified exploration FA (CIEFA) model, by integrating both inward and outward search mechanisms to overcome this limitation inherent in the original FA model. This new CIEFA model is produced based on the first IIEFA model. Specifically, CIEFA combines the inward exploration strategy embedded in IIEFA with a newly proposed dispersing mechanism based on dissimilarity measures to increase diversification. Eq. (5) defines the proposed dissimilarity measure $M_{dissimilarity}$ between two fireflies.

$$M_{dissimilarity} = \left(I_j^t - I_i^t\right) / \left(I_g^t - I_i^t\right) \tag{5}$$

where I_i^t and I_j^t represent the fitness scores of fireflies i and j, respectively, in the tth iteration, while g represents the current global best solution, and I_g^t denotes its fitness score in the tth iteration.

As illustrated in Eq. (5), we employ $M_{dissimilarity}$ to distinguish fireflies with weak or strong light intensity differences to that of the current firefly, whereby the neighbouring solutions, with $M_{dissimilarity} < 0.5$, are labelled as 'ineffective individuals', whereas those with distinctive variance in light intensities, i.e. $M_{dissimilarity} > 0.5$, are labelled as 'effective individuals', through the position updating process. Eqs. (6)–(7) define the outward search operation for the 'ineffective individuals', with $M_{dissimilarity} < 0.5$. This new outward search operation enables firefly i to not only perform local exploitation of firefly j, but also force firefly i to jump out of the space between i and j so as to explore an outer space. It expands search exploration of the weaker firefly i to accelerate convergence. On the contrary, when $M_{dissimilarity} > 0.5$, the inward intensified exploration formula in IIEFA is used to dispatch firefly i using 'effective individuals'.

$$x_i^{t+1} = x_j^t + \varphi .* \tau .* \left(x_j^t - x_i^t\right) + \alpha_t \varepsilon_t \tag{6}$$

$$\tau = (1 - t/T_{total}) * (1 + \mu) \tag{7}$$

In Eq. (6), τ denotes a step control matrix for this new outward operation, while φ represents a direction control matrix with each element being drawn randomly from -1 and 1. The step control matrix, τ, for the outward search operation is further defined in Eq. (7), where t represents the current iteration number while T_{total} is the maximum number of iterations. Parameter μ denotes the control matrix that consists of random numbers in [0, 1], as defined earlier in IIEFA, with the same feature dimension as that of the firefly swarm.

The step control matrix, τ, is employed to regulate the extent of outward exploration in each dimension and the balance between exploration and exploitation through the whole search

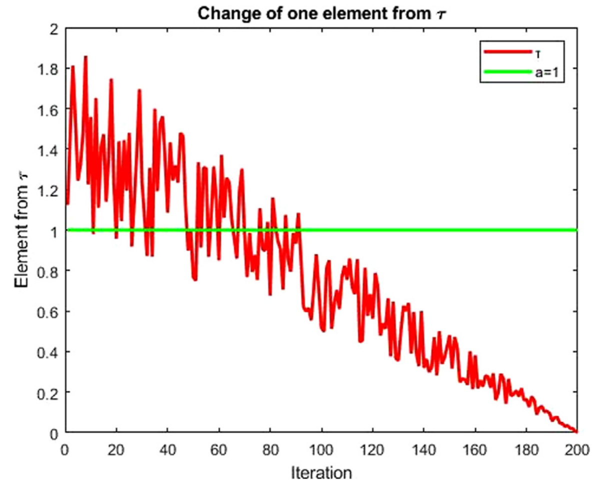


Fig. 2. An example of the change of one element from the step control matrix, τ, through iterations.

process. Owing to the randomness introduced by the control matrix, μ, in IIEFA, as defined in Eq. (3), the elements in τ possess different values from each other, but all follow the same trend of variation as the iteration number builds up. As an example, the change of one element from τ against the iteration number is illustrated in Fig. 2. This example element in τ decreases from 2 to 0, governing the exploration scale on each dimension as the count of iterations builds up. The exploration operation is conducted outwardly when the element in τ is greater than 1, otherwise the exploration operation is performed inwardly.

Based on the variance of the element in Fig. 2, it is observed that the whole search process of 'ineffective individuals' with low fitness dissimilarities ($M_{dissimilarity} < 0.5$) goes through three stages as the iteration builds up. In the first stage, the outward exploration action dominates the first 50 (out of 200) iterations approximately, where the 'ineffective individuals' are dispersed to explore a greater unexploited search domain. In the second stage, both inward and outward explorations reside in the 50th–90th iterations, in order to balance between exploitation and exploration. In the third stage, the inward exploration operation replaces the outward exploration movement, and takes control once the number of iterations exceeds 90, as the whole swarm gradually congregates and converges altogether. It should be noted that the iteration numbers used for the division of three search modes fluctuate slightly around the thresholds given in the illustrated example in Fig. 2, since the randomness of μ affects the magnitude of elements in τ delicately. Nevertheless, the general adaptive patterns coherently apply to the whole search process with respect to all dimensions in fireflies. Moreover, each element (either -1 or 1) in φ controls the direction of the movement along each corresponding dimension, which enables fireflies to fully explore and exploit the search space.

The whole search process of 'ineffective individuals' with low dissimilarity levels ($M_{dissimilarity} < 0.5$) is depicted in Fig. 3. With the assistance of three different position updating operations (indicated in three colours) in Fig. 3, not only the search diversity in direction and scope among fireflies with high similarities is improved significantly and local stagnation is mitigated effectively. The search efficiency is also enhanced because of the guarantee of heterogeneity between fireflies in movement. On the other hand, the movement of 'effective individuals' with distinctive position variance follows the same strategy in IIEFA, as illustrated in Eq. (3). In short, CIEFA enhances diversity of exploration one step further, and inherits all merits by combining both inward and outward intensified exploration mechanisms.

Algorithm 1 – The pseudo-code of the proposed IIEFA model

1. **Start**
2. Initialize a population of m fireflies
3. Initialize randomization parameter α_t and set experimental parameters
4. Define the objective function/light intensity $I = f(x)$
5. Calculate light intensity for each firefly
6. **while** (t< Max iteration) or (other converging criteria not being met)
7.   **for** $i <= m$
8.   **for** $j <= m$
9.   **if** $I_i < I_j$
10.   Generate a control matrix μ
11.   Update the position of firefly i by moving towards firefly j using **Eq. (3)**
12.   **end if**
13.   Check the new position not to exceed the range of problem variables
14.   **end for**
15.   **end for**
16.   Update the randomization step α_t using **Eq. (4)**
17. **end while**
18. Export the global best position P_g, and global best fitness value I_g
19. **End**

Moreover, according to the empirical results, the proportion of calling the dispersing search mechanism in CIEFA for 'ineffective individuals' among the total number of position updating varies slightly, and is dependent on the parameter settings (e.g. the maximum number of iterations and the size of the firefly population) as well as the problems at hand (e.g. the employed data sets). Taking the Sonar data set as an example, the proportion of running the dispersing mechanism varies between 40% and 52% for each trial with a population of 50 fireflies and a maximum number of 200 iterations. The average proportion of calling the dispersing mechanism in CIEFA over a series of 30 trials is 47.18% under the same setting. The pseudo-code of CIEFA is provided in Algorithm 2.

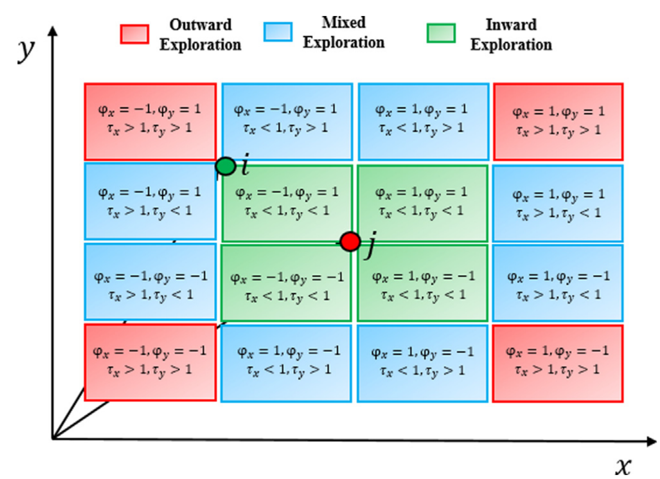


Fig. 3. Distribution of the updated positions of firefly i through iterations in the CIEFA model in a two-dimensional search space when $M_{dissimilarity} < 0.5$.

3.3. *The proposed clustering approach based on the IIEFA and CIEFA models*

The proposed IIEFA and CIEFA algorithms are subsequently employed to construct two novel clustering models to undertake initialization sensitivity and local optima traps of the original KM clustering algorithm. The flowchart and pseudo-code of the proposed clustering method are presented in Fig. 4 and Algorithm 3, respectively.

In order to improve search efficiency and increase convergence, a seed solution for cluster centroids is generated firstly by the original KM clustering algorithm, and is used to replace the first firefly in the swarm. The similarities among data samples are measured by the Euclidean distance during the partitioning process. Quality of the centroid solution represented by each firefly is evaluated based on the sum of intra-cluster distance measures. The search process and movement patterns of the swarm are governed and regulated by the proposed IIEFA and CIEFA models. Benefited from the enhanced diversity of the search scopes, scales, and directions in IIEFA and CIEFA, a cluster centroid solution with a better quality is identified through the intensified

Algorithm 2 – The pseudo-code of the proposed CIEFA model

```
Start
Initialize a swarm of m fireflies
Initialize randomization parameter α_t and set experiment parameters
Define the objective function/light intensity I = f(x)
Calculate light intensity for each firefly
while (t< Max iteration) or (other converging criteria not being met)
    for i <= m
        for j <= m
            if I_i < I_j
                Calculate M_dissimilarity using Eq. (5)
                Generate a random matrix μ
                if M_dissimilarity < 0.5
                    Calculate control matrix τ using Eq. (7)
                    Generate direction matrix φ
                    Update position of firefly i by moving towards j using Eq. (6)
                else M_dissimilarity ≥ 0.5
                    Update position of firefly i by moving towards j using Eq. (3)
                end if
            Check the new position not to exceed the range of problem variables
            end if
        end for
    end for
    Update α_t using Eq. (4)
end while
Export the global best position P_g, and the global best fitness value I_g
End
```

neighbouring and global search processes, and the possibility of being trapped in local optima is significantly reduced.

Moreover, as mentioned earlier, nearly all the hybrid KM-based clustering models partition data samples into the corresponding clusters based on the Euclidean distance, and quality of clustering centroids is improved by minimizing the sum of intra-cluster distance measures. Therefore, irrelevant and redundant features contained in the data samples can negatively impact the distance-based clustering measures, since the distance measures under such circumstances are not able to represent the compactness of the clusters accurately. Owing to the high dimensionality of some of the data sets evaluated in this study, e.g. 80 for ALL, 72 for Ozone, and 60 for Sonar, and the implementation of feature selection on the these data sets as validated in previous studies [54,66], we employ mRMR [43] to conduct feature dimensionality reduction and improve clustering performance by eliminating redundant and irrelevant features. A comprehensive evaluation of the proposed clustering method is presented in the next section.

4. Evaluation and discussion

To investigate the clustering performance in an objective and comprehensive manner, the proposed FA models are evaluated and compared with not only FA related methods, but also several other classical metaheuristic search methods. In view of their novelties and contributions to the development of a variety of metaheuristic algorithms, GA and ACO are two most successful metaheuristic search methods [67]. As such, we evaluate and compare the proposed IIEFA and CIEFA models against GA [68], ACO [69], and four other classical methods i.e. KM clustering, FA [47], Dragonfly (DA) [70], and Sine Cosine Algorithm (SCA) [71], as well as five FA variants i.e. CFA1 [57], CFA2 [58], NaFA [48], VSSFA [59], and MFA [64]. Each optimization model

Algorithm 3 – The pseudo-code of the proposed clustering method

1. **Start**
2. Import data sets and set initial parameters
3. Initialize a firefly swarm S as a series of possible cluster centroids
4. Run KM on the data set and generate the initial cluster centroids C_o as a seed solution
5. Replace the first firefly in the swarm S with C_o
6. **while** (t< Max iteration) or (other termination criteria not being met)
7. Use each firefly as the centroids to cluster the data set based on Euclidean distance
8. Evaluate fitness value/light intensity of each firefly using the sum of intra-cluster distance measure f as defined in **Eq. (8)** in Section 4.3
9. Update firefly positions using the proposed IIEFA/CIEFA models
10. **end while**
11. Export the global best position P_g, and the global best fitness value I_g
12. **End**

is integrated with KM clustering for performance comparison. A total of ten data sets characterized with a wide range of dimensionalities are first evaluated with five performance indicators, namely sum of intra-cluster distances (i.e. fitness scores), average accuracy [72], average sensitivity, average specificity and macro-average F-score (Fscore_M) [72]. To ensure a fair comparison, we employ the same number of function evaluations (i.e. population size × the maximum number of iterations) as the stopping criterion for all the search methods. The population size and the maximum number of iterations are set to 50 and 200, respectively, in our experiments. We also employ 30 independent runs in each experiment, in order to mitigate the influence of fluctuation of the results.

4.1. Parameter settings

The parameter settings of search methods employed in our study are the same as those reported in the respective original studies. As such, the following initial parameters are applied to both the original FA model and FA variants, in accordance with the empirical study in [37], i.e. initial attractiveness=1.0, absorption coefficient=1.0, and randomization parameter=0.2, while the proposed IIEFA and CIEFA models employ randomized search parameters as indicated in Section 3. The details of parameter settings of each search method are listed in Table 1.

4.2. Data sets

Clustering performance is significantly influenced by characteristics of data samples, such as data distribution, noise, and dimensionality. Therefore, the following data sets with various characteristics from different domains are used to investigate efficiency of the proposed models. Specifically, we employ the ALL-IDB2 database [73], denoted as ALL (Acute Lymphoblastic Leukaemia), and nine data sets from the UCI machine learning repository [74], namely Sonar, Ozone, Wisconsin breast cancer diagnostic data set (Wbc1), Wisconsin breast cancer original data set (Wbc2), Wine, Iris, Balance, Thyroid, and *E. coli*, for evaluation. Among the selected data sets, Sonar, Ozone and ALL possess relatively high feature dimensionality, i.e. 60, 72, and 80, respectively. The remaining data sets have comparatively smaller feature dimensions (i.e. 9 for Wbc2, 4 for Iris and 5 for Thyroid). Additionally, owing to the fact that data samples are extremely imbalanced between classes in certain data sets, e.g. *E. coli*, we only select those classes with relatively sufficient number of samples for clustering performance comparison. The main characteristics of the employed data sets are illustrated in Table 2.

The employed data sets impose various challenges on clustering analysis. As an example, the ALL data set used in [66, 75] is obtained from the analysis of the ALL-IDB2 microscopic blood cancer images. The essential features, such as colour, shape, and texture details, were extracted from this ALL-IDB2 data set, and a feature vector of 80 dimensions was obtained for each white blood cell image [63]. This image data set poses diverse challenges to classification/clustering models, owing to the complex irregular morphology of nucleus, variations in terms of the nucleus to cytoplasm ratio, as well as the subtle differences between the blast and normal blood cells, which bring in noise and sub-optimal distraction in the follow-on clustering process for lymphoblastic and lymphocyte identification. Other UCI data sets also contain similar challenging factors. Therefore, a comprehensive evaluation of the proposed clustering models can be established owing to diversity of the employed challenging data sets in terms of sample distribution and dimensionality.

4.3. Performance comparison metrics

Five performance indicators are employed to evaluate the clustering performance, namely the sum of intra-cluster distances (i.e. fitness scores), average accuracy, average sensitivity, average specificity, and macro-average F-score (Fscore_M) [72]. The first distance-based metric is used to indicate the convergence speed of the proposed models, while the last four metrics are used as the main criteria for clustering performance comparison. We introduce each performance metric in detail, as follows.

1. Sum of intra-cluster distances: This measurement is obtained by the summation of distances between the data samples and their corresponding centroids, as defined in Eq. (8). The smaller the sum of intra-cluster distances, the more compact

Table 1
Parameter settings of different algorithms.

Algorithms	Parameters
FA [37]	initial attractiveness = 1.0, absorption coefficient = 1.0, randomization parameter = 0.2, adaptive coefficient = 0.97
CFA1 [57]	chaotic component = $\delta \times x^{(n)}$, where $\delta = 1 - \lvert \frac{n-1}{n} \rvert^{0.25}$, $x^{(n)}$ represents chaotic variable generated by Logistic map, and n represents the current iteration number. Other parameters are the same as those of FA.
CFA2 [58]	attractiveness coefficient = Gauss map, with the rest parameters the same as those of FA
NaFA [48]	size of neighbourhood brighter fireflies=3, with other parameters the same as those of FA
VSSFA [59]	adaptive randomization step=$0.4/(1 + exp(0.015 \times (t - max\_iteration)/3))$, where t and $max\_iteration$ represent the current and maximum iteration numbers, respectively. Other search parameters are the same as those of FA.
DA [70]	separation factor = 0.1, alignment factor = 0.1, cohesion factor = 0.7, food factor = 1, enemy factor = 1, inertial weight = $0.9 - m \times ((0.9 - 0.4)/max\_iteration)$, where m and $max\_iteration$ represent the current and maximum iteration numbers, respectively.
SCA [71]	$r_1 = a - t \times a/T$, where $a = 3$ and t and T represent the current and maximum iteration numbers, respectively. $r_2 = 2\pi \times rand$, $r_3 = 2 \times rand$, $r_4 = rand$
MFA [64]	Parameter settings are the same as those of FA
GA [47]	crossover probability=0.8, mutation probability=0.05
ACO [69]	locality of the search process= 10^{-4}, pheromone evaporation rate = 0.85
IIEFA	control matrix $\mu \in (0, 1)$, with other search parameters the same as those of FA
CIEFA	step control matrix $\tau = (1 - t/T_{total}) \times (1 + \mu)$, where $tandT_{total}$ represent the current and maximum iteration numbers, respectively. Other search parameters are the same as those of FA.

Table 2
Ten selected data sets for evaluation.

Data set	Number of attributes	Number of classes	Missing values	Number of instances
Sonar	60	2	No	140
Ozone	72	2	No	196
ALL	80	2	No	100
Wbc1	30	2	No	569
Wbc2	9	2	No	683
Wine	13	3	No	178
Iris	4	3	No	150
Balance	4	2	No	576
Thyroid	5	3	No	90
E. coli	7	3	No	150

the partitioned clusters. Similar to KM clustering, the proposed models employ the sum of intra-cluster distances as the objective function, which is minimized during the search process.

$$f(O, C) = \sum_{i=1}^{k} \sum_{O_l \in C_i} \sqrt{(O_l - Z_i)^2} \tag{8}$$

where C_i and Z_i, represent the ith cluster and the centroid of the ith cluster, while O_l and k denote the data belonging to the ith cluster, and the total number of clusters, respectively.

2. Average accuracy: The mean clustering accuracy is obtained by averaging the accuracy rate of each class, as defined in Eq. (9). The merit of this performance metric is that it treats all classes equally, rather than being dominated by classes with a large number of samples [72].

$$\text{Ave\_accuracy} = \frac{\sum_{i=1}^{k} \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i}}{k} \tag{9}$$

where tp_i, fn_i, fp_i, and tn_i represent true positive, false negative, false positive, and true negative of the ith cluster, respectively.

3. Average sensitivity: As defined in Eq. (10), sensitivity (i.e. recall) is used to measure the proportion of correctly identified positive samples over all positive samples in the data set. Similar to the average accuracy, the macro-average of sensitivity is calculated, in order to ascertain all classes are treated equally for multi-class clustering tasks [72].

$$\text{Ave\_sensitivity} = \frac{\sum_{i=1}^{k} \frac{tp_i}{tp_i + fn_i}}{k} \tag{10}$$

4. Average specificity: Specificity is used to identify the proportion of correctly identified negative samples over all negative samples in the data set [72]. Eq. (11) is used to obtain the macro-average specificity for multiclass tasks.

$$\text{Ave\_specificity} = \frac{\sum_{i=1}^{k} \frac{tn_i}{tn_i + fp_i}}{k} \tag{11}$$

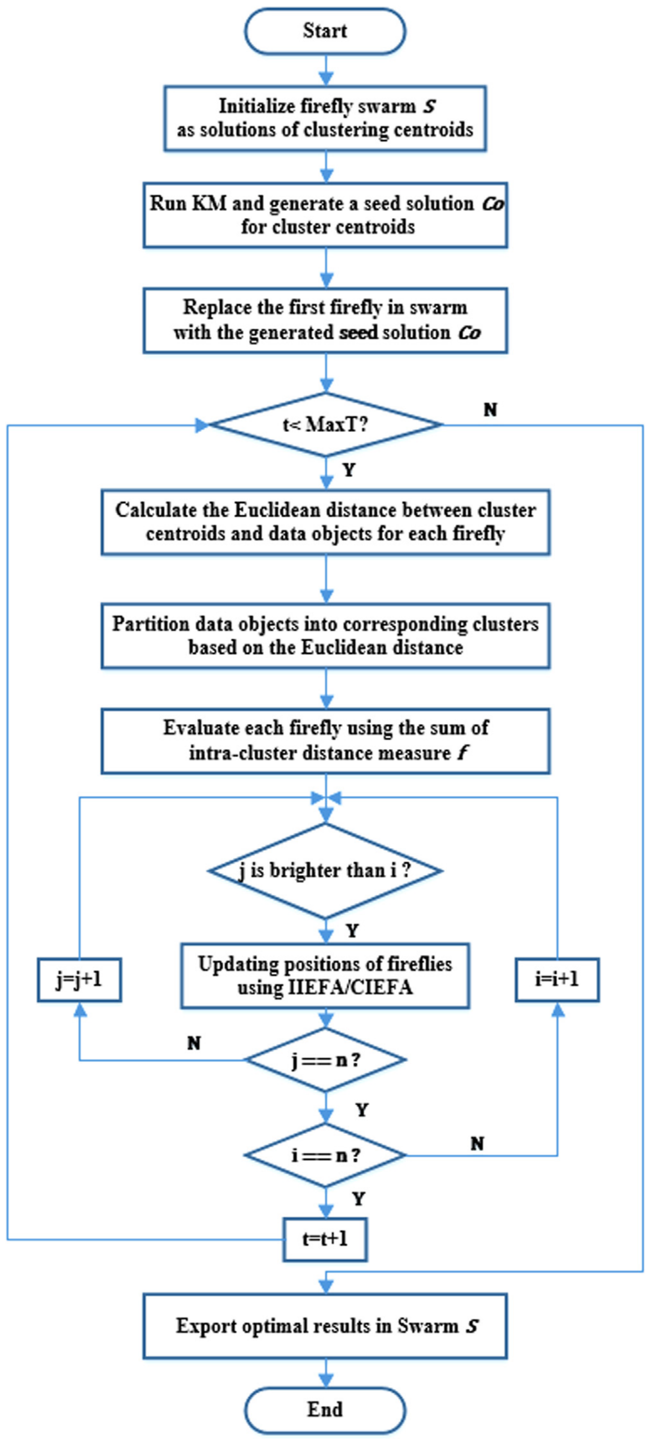


Fig. 4. Flowchart of the proposed clustering method.

5. Macro-average F-score ($Fscore_M$): $Fscore_M$ is a well-accepted performance metric, which is calculated based on the macro-average of precision and recall scores [72], as defined in Eqs. (12)–(14).

$$Fscore_M = \frac{(\sigma^2 + 1) * Precision_M * Recall_M}{\sigma^2 * Precision_M + Recall_M} \quad (12)$$

$$Precision_M = \frac{\sum_{i=1}^{k} \frac{tp_i}{tp_i + fp_i}}{k} \quad (13)$$

$$Recall_M = \frac{\sum_{i=1}^{k} \frac{tp_i}{tp_i + fn_i}}{k} \quad (14)$$

where $\sigma = 1$, in order to obtain equal weightings of precision and recall.

For each data set, a total of 30 runs with each search method integrated with the KM clustering algorithm are conducted. The average performance over 30 runs for each performance metric is calculated and used as the main criterion for comparison.

4.4. Feature selection and clustering performance evaluation

As mentioned earlier, owing to the high dimensionality of Sonar, Ozone, and ALL data sets, and the possibility of the inclusion of redundant features, mRMR [43] is used to conduct feature dimensionality reduction and to investigate its underlying impact on the clustering performance. The clustering results before and after feature selection for each data set are shown in Tables 3–12, respectively. For the three high-dimensional data sets, namely ALL, Sonar, and Ozone, the numbers of selected features are 9, 17, and 22 from the original 80, 60, and 72 features, respectively. These feature sizes are obtained based on trial-and-error, which yield the best performance for nearly all evaluated models. The findings on feature selection are also consistent with those of existing studies [54,66], where the ranges of selected feature numbers are 9–36 [66], 15–20 [54], and 18–25 [54] for ALL, Sonar, and Ozone, respectively, therefore ascertaining efficiency of the mRMR-based feature selection method employed in this research.

The empirical results indicate that in combination with feature selection, the clustering performance is improved for most test cases. As an example, for the ALL data set illustrated in Table 3, the number of features is reduced from the original 80 to 9, while the mean accuracy, sensitivity, specificity, and $Fscore_M$ of the proposed CIEFA model over 30 runs increase significantly, i.e., from 51.23% to 80.4%, 51.67% to 74.67%, 50.8% to 86.13%, and 51.27% to 78.73%, respectively. The selected features include the cytoplasm and nucleus areas, ratio between the nucleus area and the cytoplasm area, form factor, compactness, perimeter and eccentricity, which represent the most significant clinical factors for blood cancer diagnosis [66,75,76]. This in turn also indicates that some redundant or even contradictory features exist in the original data set [66], which may deteriorate the performance of clustering models drastically. Such findings also apply to other data sets, especially the high-dimensional ones [54]. The only exception is the low-dimensional Balance data set, as shown in Table 7, where the full feature set (i.e. a total of only four features) yields the best performance for nearly all the clustering models. In short, it is essential to eliminate redundant and irrelevant features to enhance the clustering performance.

4.5. Performance comparison and analysis

As mentioned earlier, five metrics are used for clustering performance comparison, namely the fitness scores on the sum of intra-cluster distances, average accuracy, average sensitivity, average specificity, and macro-average F-score ($Fscore_M$). Since the best performances are achieved using the identified significant feature subsets in most test cases for nearly all the methods, we employ the enhanced results obtained in combination with feature selection for further analysis and comparison. The detailed evaluation results over 30 runs for each performance measure after feature selection are shown in Tables 13–17.

With respect to the fitness scores, i.e. the intra-cluster distance measure, as shown in Table 13, IIEFA and CIEFA achieve the minimum distance measures in eight out of ten data sets in total. Specifically, IIEFA yields the minimum intra-cluster measures with five data sets based on the average performance over 30

Table 3
The mean clustering results over 30 independent runs on the ALL data set.

Feature number	Criteria	IIEFA	CIEFA	FA	KM	CFA1	CFA2	NaFA	VSSFA	DA	SCA	MFA	GA	ACO
80 (full set)	Fitness	**293.53**	293.71	294.33	943.13	294.32	294.35	294.33	294.33	294.13	459.26	294.34	294.32	294.33
	Accuracy	0.5137	0.5123	0.5140	0.5157	0.5147	0.5127	0.5133	0.5143	0.513	0.5133	0.5133	**0.5150**	0.5143
	Fscore$_M$	0.5145	0.5127	0.5038	0.5161	0.5115	0.5118	0.5062	0.5053	0.5137	0.3647	**0.5191**	0.5103	0.5187
	Sensitivity	0.5187	0.5167	0.4967	0.5193	0.5113	0.5147	0.5020	0.4993	0.5180	0.5153	**0.5287**	0.5087	0.5267
	Specificity	0.5087	0.5080	**0.5313**	0.5120	0.5180	0.5107	0.5247	0.5293	0.5080	0.5113	0.4980	0.5213	0.5020
9	Fitness	90.481	**89.649**	92.611	96.48	90.519	92.097	93.052	90.883	89.683	111.08	90.782	90.448	91.309
	Accuracy	0.7893	**0.804**	0.7307	0.7693	0.7703	0.7437	0.7197	0.7740	0.7850	0.6267	0.7570	0.7943	0.7527
	Fscore$_M$	0.7767	**0.7873**	0.7063	0.7557	0.7702	0.7130	0.7017	0.7611	0.7763	0.6178	0.7336	0.7841	0.7260
	Sensitivity	0.7593	0.7467	0.6953	0.7427	**0.788**	0.6807	0.7107	0.7527	0.7713	0.7260	0.724	0.7727	0.7067
	Specificity	0.8193	**0.8613**	0.7660	0.7960	0.7527	0.8067	0.7287	0.7953	0.7987	0.5273	0.7900	0.8160	0.7987

Table 4
The mean clustering results over 30 independent runs on the Sonar data set.

Feature number	Criteria	IIEFA	CIEFA	FA	KM	CFA1	CFA2	NaFA	VSSFA	DA	SCA	MFA	GA	ACO
60 (full set)	Fitness	**160.54**	160.73	161.22	195.35	160.85	161.42	160.98	161.31	161.05	242.81	160.92	161.14	160.75
	Accuracy	0.5610	0.5631	**0.5669**	0.5655	0.5624	0.5643	0.5657	0.5645	0.5657	0.5307	0.5629	0.5624	0.5629
	Fscore$_M$	0.5553	**0.5698**	0.5549	0.5664	0.5500	0.5526	0.5583	0.5532	0.5635	0.3944	0.5613	0.5636	0.5671
	Sensitivity	0.5552	**0.5862**	0.5500	0.5781	0.5443	0.5486	0.5590	0.5500	0.5700	0.4324	0.5681	0.5724	0.5814
	Specificity	0.5667	0.5400	**0.5838**	0.5529	0.5805	0.58	0.5724	0.579	0.5614	0.6290	0.5576	0.5524	0.5443
17	Fitness	**75.85**	75.884	76.487	46.251	76.529	76.38	76.381	76.461	76.187	101.95	76.344	75.952	76.470
	Accuracy	0.7100	**0.7110**	0.6733	0.6764	0.6717	0.6669	0.6779	0.6719	0.6760	0.6183	0.6750	0.7088	0.6769
	Fscore$_M$	0.7072	**0.7090**	0.6677	0.6814	0.6546	0.6601	0.6722	0.6461	0.6623	0.5466	0.6772	0.7019	0.6776
	Sensitivity	0.7110	0.7157	0.6829	**0.7224**	0.6538	0.6862	0.6867	0.6243	0.6633	0.5267	0.7048	0.7024	0.7024
	Specificity	0.7090	0.7062	0.6638	0.6305	0.6895	0.6476	0.669	**0.7195**	0.6886	0.7100	0.6452	0.7152	0.6514

Table 5
The mean clustering results over 30 independent runs on the Ozone data set.

Feature number	Criteria	IIEFA	CIEFA	FA	KM	CFA1	CFA2	NaFA	VSSFA	DA	SCA	MFA	GA	ACO
72 (full set)	Fitness	**514.11**	514.38	515.29	1507.7	515.23	515.23	515.23	515.29	514.77	844.99	515.3	515.07	515.44
	Accuracy	0.7333	0.7330	0.7366	**0.7369**	0.7362	0.7361	0.7352	0.7364	0.7337	0.5631	0.7367	0.7367	0.7367
	Fscore$_M$	0.7167	0.7316	0.7221	**0.7543**	0.7374	0.7434	0.7429	0.7065	0.7353	0.4209	0.7312	0.7412	0.7127
	Sensitivity	0.7000	0.7554	0.7136	**0.8313**	0.7701	0.7932	0.7932	0.6565	0.7677	0.4949	0.7463	0.7830	0.6793
	Specificity	0.7667	0.7105	0.7595	0.6425	0.7024	0.6789	0.6772	0.8163	0.6997	0.6313	0.7272	0.6905	**0.7942**
22	Fitness	**301.26**	301.34	302.19	517.76	302.22	302.29	302.22	302.25	301.9	414.87	301.89	301.42	302.24
	Accuracy	**0.7604**	0.7577	0.7490	0.7488	0.7495	0.7497	0.7491	0.7491	0.7500	0.5648	0.7500	0.7531	0.7495
	Fscore$_M$	**0.7524**	0.7466	0.7408	0.7349	0.7438	0.7362	0.7359	0.7433	0.7318	0.3792	0.7419	0.7433	0.7407
	Sensitivity	0.7435	0.7310	0.7391	0.7184	**0.7503**	0.7204	0.7197	0.749	0.7007	0.4173	0.7401	0.7347	0.7381
	Specificity	0.7772	0.7844	0.7588	0.7793	0.7486	0.7789	0.7786	0.7493	**0.7993**	0.7122	0.7599	0.7714	0.7609

Table 6
The mean clustering results over 30 independent runs on the Thyroid data set.

Feature number	Criteria	IIEFA	CIEFA	FA	KM	CFA1	CFA2	NaFA	VSSFA	DA	SCA	MFA	GA	ACO
5 (full set)	Fitness	113.26	**111.65**	115.03	196.65	119.24	116.51	114.97	117.87	114.15	124.5	114.28	114.12	113.54
	Accuracy	0.8235	0.8277	0.8133	0.8215	0.7911	0.8173	0.8205	0.8032	0.8128	**0.8321**	0.8165	0.8126	0.822
	Fscore$_M$	0.7539	0.7688	0.7508	0.7638	0.7090	0.7482	0.7582	0.7256	0.7398	**0.7981**	0.7575	0.7392	0.7667
	Sensitivity	0.7352	0.7415	0.7200	0.7322	0.6867	0.7259	0.7307	0.7048	0.7193	**0.7481**	0.7248	0.7189	0.7330
	Specificity	0.8676	0.8707	0.86	0.8661	0.8433	0.863	0.8654	0.8524	0.8596	**0.8741**	0.8624	0.8594	0.8665
4	Fitness	**96.297**	96.599	99.808	142.21	99.661	99.979	100.49	99.364	97.743	107.3	99.36	96.794	100.56
	Accuracy	**0.8748**	0.8637	0.8101	0.8057	0.8084	0.8069	0.802	0.8116	0.8346	0.841	0.8121	0.8514	0.8044
	Fscore$_M$	**0.8377**	0.8204	0.7628	0.753	0.7611	0.7543	0.7505	0.7719	0.7813	0.8013	0.7649	0.8010	0.7467
	Sensitivity	**0.8122**	0.7956	0.7152	0.7085	0.7126	0.7104	0.703	0.7174	0.7519	0.7615	0.7181	0.7770	0.7067
	Specificity	**0.9061**	0.8978	0.8576	0.8543	0.8563	0.8552	0.8515	0.8587	0.8759	0.8807	0.8591	0.8885	0.8533

runs, i.e., Thyroid, Balance, *E. coli*, Ozone, and Wbc1, while CIEFA achieves the minimum fitness scores with four data sets, i.e., *E. coli*, ALL, Wbc2, and Wine. Moreover, KM clustering produces the minimum intra-cluster measures with the Sonar and Iris data sets in combination with mRMR-based feature selection, although IIEFA and CIEFA achieve the minimum objective function evaluation scores when the full feature sets for both Sonar and Iris data sets are used. Overall, in comparison with the six classical methods i.e. GA, ACO, DA, SCA, FA, KM, and other five FA variants i.e. CFA1, CFA2, NaFA, VSSFA, and MFA, both IIEFA and CIEFA models demonstrate faster convergence rates and great superiority over other methods in identifying enhanced centroids that lead to more compact clusters. The proposed search mechanisms account for the enhanced global exploration capability of IIEFA and CIEFA in comparison with those of other classical methods and FA variants in attaining the global best solutions.

In terms of mean accuracy and Fscore$_M$, as shown in Tables 14–15, the proposed models achieve the best scores for all the data sets over 30 runs. With respect to the mean accuracy rates shown in Table 14, IIEFA achieves the highest average accuracy rates over 30 runs with seven data sets (i.e. Thyroid, Balance,

Table 7
The mean clustering results over 30 independent runs on the Balance data set.

Feature number	Criteria	IIEFA	CIEFA	FA	KM	CFA1	CFA2	NaFA	VSSFA	DA	SCA	MFA	GA	ACO
4 (full set)	Fitness	**1002.9**	1003.1	1003.9	1866.6	1004.2	1003.9	1003.7	1003.6	1003.1	1011.2	1003.3	1003	1003.4
	Accuracy	**0.8047**	0.7923	0.7733	0.7546	0.7494	0.7581	0.7538	0.7725	0.7858	0.7549	0.7993	0.7991	0.7956
	Fscore$_M$	**0.8045**	0.7923	0.7735	0.7518	0.7475	0.758	0.7537	0.7726	0.7857	0.7522	0.7991	0.7991	0.7955
	Sensitivity	**0.8038**	0.7925	0.7749	0.7478	0.7459	0.7574	0.7536	0.7727	0.7855	0.7491	0.7985	0.799	0.7953
	Specificity	**0.8056**	0.7921	0.7718	0.7613	0.7529	0.7588	0.7539	0.7723	0.7860	0.7606	0.8001	0.7993	0.7959
3	Fitness	821.70	821.77	824.55	1300.6	824.56	824.67	824.6	826.58	821.75	826.52	821.86	**821.52**	823.14
	Accuracy	0.7344	0.7344	0.7004	0.7042	0.6939	0.7164	0.7135	0.6747	0.7331	0.7269	**0.7372**	0.7355	0.7217
	Fscore$_M$	0.7342	0.7349	0.7002	0.7073	0.6923	0.7202	0.7134	0.6719	0.7321	0.7303	**0.7377**	0.7356	0.7200
	Sensitivity	0.7338	0.7362	0.7012	0.7126	0.6896	0.7281	0.7162	0.6718	0.7303	**0.7394**	0.7392	0.7359	0.7167
	Specificity	0.735	0.7325	0.6997	0.6957	0.6983	0.7047	0.7109	0.6777	**0.7359**	0.7144	0.7352	0.7352	0.7267

Table 8
The mean clustering results over 30 independent runs on the *E. coli* data set.

Feature number	Criteria	IIEFA	CIEFA	FA	KM	CFA1	CFA2	NaFA	VSSFA	DA	SCA	MFA	GA	ACO
7 (full set)	Fitness	257.63	251.13	260.17	473.53	259.33	257.07	257.73	260.05	253.28	252.78	261.52	**244.09**	257.85
	Accuracy	0.7739	0.7945	0.7769	0.8883	0.7756	0.771	0.7704	0.7641	0.7893	**0.929**	0.7633	0.8154	0.7769
	Fscore$_M$	0.6992	0.7248	0.6692	0.8341	0.6687	0.6605	0.6574	0.6503	0.6935	**0.8971**	0.6498	0.7491	0.6701
	Sensitivity	0.6609	0.6918	0.6653	0.8324	0.6633	0.6564	0.6556	0.6462	0.684	**0.8936**	0.6449	0.7231	0.6653
	Specificity	0.8304	0.8459	0.8327	0.9162	0.8317	0.8282	0.8278	0.8231	0.842	**0.9468**	0.8224	0.8616	0.8327
5	Fitness	**196.23**	**196.23**	198.08	321.05	198.03	196.53	198.2	197.91	197.72	238.63	198.00	197.64	197.96
	Accuracy	**0.9644**	**0.9644**	0.9564	0.9406	0.9563	0.961	0.9557	0.9575	0.9556	0.931	0.9566	0.9566	0.9536
	Fscore$_M$	**0.9474**	**0.9474**	0.9352	0.9109	0.9349	0.9421	0.934	0.9368	0.9337	0.9005	0.9355	0.9354	0.9308
	Sensitivity	**0.9467**	**0.9467**	0.9347	0.9109	0.9344	0.9416	0.9336	0.9362	0.9333	0.8964	0.9349	0.9349	0.9304
	Specificity	**0.9733**	**0.9733**	0.9673	0.9554	0.9672	0.9708	0.9668	0.9681	0.9667	0.9482	0.9674	0.9674	0.9652

Table 9
The mean clustering results over 30 independent runs on the Wbc1 data set.

Feature number	Criteria	IIEFA	CIEFA	FA	KM	CFA1	CFA2	NaFA	VSSFA	DA	SCA	MFA	GA	ACO
30 (full set)	Fitness	**2280.8**	2281.5	2293.9	11575	2293.8	2293.9	2293.7	2293.8	2286	2800.3	2293.5	2285.9	2293.8
	Accuracy	**0.9147**	0.9145	0.9114	0.9097	0.9108	0.9113	0.9105	0.9111	0.9129	0.7230	0.9100	0.9142	0.9110
	Fscore$_M$	**0.9092**	0.9039	0.9047	0.9051	0.9081	0.899	0.8963	0.8949	0.909	0.6082	0.8978	0.9064	0.8986
	Sensitivity	0.9056	0.8953	0.8986	0.9016	**0.9067**	0.8846	0.8813	0.8759	0.9066	0.6242	0.8845	0.8986	0.8844
	Specificity	0.8990	0.9088	0.8894	0.8845	0.8804	0.9032	0.9058	**0.9119**	0.8925	0.6558	0.9022	0.8984	0.9031
20	Fitness	**1761.4**	1761.6	1768.7	6887.6	1768.7	1768.7	1768.7	1768.7	1764.7	2220.2	1768.7	1762.1	1768.7
	Accuracy	**0.9461**	0.9448	0.9332	0.9332	0.9332	0.9332	0.9332	0.9332	0.9385	0.813	0.9332	0.9393	0.9332
	Fscore$_M$	0.9361	**0.9394**	0.9230	0.9276	0.9215	0.9261	0.9291	0.9184	0.9295	0.7795	0.9276	0.9372	0.9322
	Sensitivity	0.9141	0.9290	0.9028	0.9185	0.8976	0.9133	0.9237	0.8871	0.9110	0.7803	0.9185	0.9358	0.9341
	Specificity	**0.9470**	0.9285	0.9237	0.908	0.9289	0.9133	0.9028	0.9394	0.9298	0.7290	0.9080	0.9069	0.8924

Table 10
The mean clustering results over 30 independent runs on the Wbc2 data set.

Feature number	Criteria	IIEFA	CIEFA	FA	KM	CFA1	CFA2	NaFA	VSSFA	DA	SCA	MFA	GA	ACO
9 (full set)	Fitness	1092.1	**1092**	1098.3	2724.4	1102.7	1102.9	1100	1102.8	1093.7	1327.3	1093.5	1092.1	1093.1
	Accuracy	**0.9693**	0.9692	0.9629	0.9560	0.9563	0.9559	0.9604	0.9562	0.9683	0.9542	0.9684	0.9679	0.9680
	Fscore$_M$	**0.9662**	0.9661	0.9588	0.9562	0.9538	0.9489	0.9555	0.9525	0.9623	0.9462	0.9646	0.9640	0.9637
	Sensitivity	**0.9667**	0.9666	0.9561	0.9568	0.9522	0.9421	0.9525	0.9495	0.962	0.9365	0.9646	0.9639	0.9646
	Specificity	0.9667	0.9666	0.9580	0.9367	0.9423	0.9511	0.9539	0.9446	**0.9682**	0.9500	0.9660	0.9655	0.9649
7	Fitness	**931.67**	**931.67**	933.94	1819.8	934.25	935.78	933.42	933.96	932.27	1104.8	932.17	931.68	**931.67**
	Accuracy	**0.9649**	**0.9649**	0.9647	**0.9649**	0.9644	**0.9649**	0.9648	0.9647	**0.9649**	0.949	**0.9649**	**0.9649**	**0.9649**
	Fscore$_M$	0.9629	0.9629	0.9619	0.9629	0.9607	0.9613	0.9597	0.9572	0.9621	0.9451	0.9652	0.9629	**0.9660**
	Sensitivity	0.9624	0.9624	0.9613	0.9624	0.9603	0.9604	0.9583	0.9554	0.9614	0.9408	0.9654	0.9624	**0.9663**
	Specificity	0.9584	0.9584	0.9593	0.9584	0.9599	0.9604	0.9624	**0.9652**	0.9594	0.9305	0.9555	0.9584	0.9545

E. coli, Ozone, Wbc1, Wbc2 and Iris), while CIEFA achieves the best results with five data sets (i.e. Sonar, *E. coli*, ALL, Wine, and Iris). Both IIEFA and CIEFA demonstrate a clear advantage over other methods with four data sets i.e. Thyroid, Sonar, Balance, and ALL. Pertaining to the Fscore$_M$ measure shown in Table 15, IIEFA and CIEFA achieve the best average scores over 30 runs with six data sets, i.e. Thyroid, Balance, *E. coli*, Ozone, Wbc2, and Iris for IIEFA and Sonar, *E. coli*, ALL, Wbc1, Wine, and Iris for CIEFA, respectively. Similar to the accuracy indicator, a clear performance distinction can be observed between the proposed models and other methods with respect to the Fscore$_M$ results.

Moreover, the observed advantages of IIEFA and CIEFA are further reinforced by the results of sensitivity and specificity, as shown in Tables 16–17. With respect to sensitivity and specificity, IIEFA achieves the highest scores for both metrics with five data sets (i.e. Thyroid, Balance, *E. coli*, Wbc2, and Iris), while CIEFA achieves the best results for both metrics with three data sets (i.e. *E. coli*, Wine, and Iris). This indicates that both CIEFA and IIEFA outperform other baseline models with most of the employed data sets. They are capable of clustering and recognizing data samples from different classes effectively.

Table 11
The mean clustering results over 30 independent runs on the Wine data set.

Feature number	Criteria	IIEFA	CIEFA	FA	KM	CFA1	CFA2	NaFA	VSSFA	DA	SCA	MFA	GA	ACO
13 (full set)	Fitness	456.78	452.06	461.45	1282.9	451.84	453.8	453.8	461.93	**451.34**	580.06	449.81	451.65	451.75
	Accuracy	0.9485	0.9705	0.9372	0.9654	0.9669	0.9607	0.9598	0.9301	0.9689	0.7876	**0.9747**	0.9692	0.9683
	Fscore$_M$	0.9295	0.9577	0.921	0.9544	0.9561	0.9492	0.9447	0.9099	0.9566	0.7048	**0.9649**	0.9586	0.9579
	Sensitivity	0.9318	0.9617	0.9198	0.9567	0.9585	0.9507	0.9492	0.9110	0.9610	0.7041	**0.9682**	0.9613	0.9603
	Specificity	0.9618	0.9784	0.9546	0.9749	0.9762	0.9718	0.9711	0.9493	0.9777	0.8383	**0.9816**	0.9781	0.9772
9	Fitness	348.48	**339.84**	342.74	744.96	344.93	345.01	342.78	342.76	346.88	431.28	342.6	342.05	340.52
	Accuracy	0.9484	**0.98**	0.9663	0.9665	0.9578	0.9587	0.9649	0.9664	0.9518	0.9109	0.9665	0.9676	0.9735
	Fscore$_M$	0.9242	**0.9713**	0.9519	0.9538	0.9393	0.942	0.9499	0.9536	0.9313	0.8818	0.9517	0.953	0.9622
	Sensitivity	0.9286	**0.9749**	0.9563	0.9573	0.9442	0.9466	0.9546	0.9572	0.9368	0.8790	0.9561	0.9562	0.9662
	Specificity	0.9619	**0.9858**	0.9751	0.9759	0.9687	0.9697	0.9741	0.9758	0.9638	0.9325	0.9755	0.9764	0.9809

Table 12
The mean clustering results over 30 independent runs on the Iris data set.

Feature number	Criteria	IIEFA	CIEFA	FA	KM	CFA1	CFA2	NaFA	VSSFA	DA	SCA	MFA	GA	ACO
4 (full set)	Fitness	130.24	131.37	133.09	150.75	132.49	131.94	133.62	133.09	131.57	161.79	132.18	**129.71**	130.04
	Accuracy	0.8818	0.8744	0.8677	0.8653	0.8735	0.8714	0.8659	0.8704	0.8742	0.8739	0.8738	**0.8876**	0.8855
	Fscore$_M$	0.8228	0.8117	0.8018	0.7987	0.8106	0.8080	0.7993	0.8061	0.8116	0.8253	0.8110	**0.8315**	0.8284
	Sensitivity	0.8227	0.8116	0.8016	0.7980	0.8102	0.8071	0.7989	0.8056	0.8113	0.8109	0.8107	**0.8313**	0.8282
	Specificity	0.9113	0.9058	0.9008	0.899	0.9051	0.9036	0.8994	0.9028	0.9057	0.9054	0.9053	**0.9157**	0.9141
2	Fitness	42.932	42.932	43.226	**17.927**	43.243	43.296	43.199	43.225	42.942	57.223	42.956	42.932	42.992
	Accuracy	**0.9733**	**0.9733**	**0.9733**	**0.9733**	**0.9733**	**0.9733**	**0.9733**	**0.9733**	**0.9733**	0.9587	**0.9733**	**0.9733**	**0.9733**
	Fscore$_M$	**0.9602**	**0.9602**	**0.9602**	**0.9602**	**0.9602**	**0.9602**	**0.9602**	**0.9602**	**0.9602**	0.9432	**0.9602**	**0.9602**	**0.9602**
	Sensitivity	**0.9600**	**0.9600**	**0.9600**	**0.9600**	**0.9600**	**0.9600**	**0.9600**	**0.9600**	**0.9600**	0.9573	**0.9600**	**0.9600**	**0.9600**
	Specificity	**0.9800**	**0.9800**	**0.9800**	**0.9800**	**0.9800**	**0.9800**	**0.9800**	**0.9800**	**0.9800**	0.9787	**0.9800**	**0.9800**	**0.9800**

Table 13
The mean results of the minimum intra-cluster distance measure over 30 runs.

Data set	Feature size	IIEFA	CIEFA	FA	KM	CFA1	CFA2	NaFA	VSSFA	DA	SCA	MFA	GA	ACO
Thyroid	4	**96.297**	96.599	99.808	142.21	99.661	99.979	100.49	99.364	97.743	107.3	99.36	96.794	100.56
Sonar	17	75.85	75.884	76.487	**46.251**	76.529	76.38	76.381	76.461	76.187	101.95	76.344	75.952	76.47
Balance	4	**1002.9**	1003.1	1003.9	1866.6	1004.2	1003.9	1003.7	1003.6	1003.1	1011.2	1003.3	1003	1003.4
E. coli	5	**196.23**	**196.23**	198.08	321.05	198.03	196.53	198.2	197.91	197.72	238.63	198	197.64	197.96
Ozone	22	**301.26**	301.34	302.19	517.76	302.22	302.29	302.22	302.25	301.9	414.87	301.89	301.42	302.24
ALL	9	90.481	**89.649**	92.611	96.48	90.519	92.097	93.052	90.883	89.683	111.08	90.782	90.448	91.309
Wbc1	20	**1761.4**	1761.6	1768.7	6887.6	1768.7	1768.7	1768.7	1768.7	1764.7	2220.2	1768.7	2285.9	1768.7
Wbc2	9	1092.1	**1092.0**	1098.3	2724.4	1102.7	1102.9	1100.0	1102.8	1093.7	1327.3	1093.5	1092.1	1093.1
Wine	9	348.48	**339.84**	342.74	744.96	344.93	345.01	342.78	342.76	346.88	431.28	342.6	342.05	340.52
Iris	2	42.932	42.932	43.226	**17.927**	43.243	43.296	43.199	43.225	42.942	57.223	42.956	42.932	42.992

Table 14
The mean results of average accuracy after feature selection over 30 runs.

Data set	Feature size	IIEFA	CIEFA	FA	KM	CFA1	CFA2	NaFA	VSSFA	DA	SCA	MFA	GA	ACO
Thyroid	4	**0.8748**	0.8637	0.8101	0.8057	0.8084	0.8069	0.802	0.8116	0.8346	0.841	0.8121	0.8514	0.8044
Sonar	17	0.71	**0.711**	0.6733	0.6764	0.6717	0.6669	0.6779	0.6719	0.676	0.6183	0.675	0.7088	0.6769
Balance	4	**0.8047**	0.7923	0.7733	0.7546	0.7494	0.7581	0.7538	0.7725	0.7858	0.7549	0.7993	0.7991	0.7956
E. coli	5	**0.9644**	**0.9644**	0.9564	0.9406	0.9563	0.961	0.9557	0.9575	0.9556	0.931	0.9566	0.9566	0.9536
Ozone	22	**0.7604**	0.7577	0.749	0.7488	0.7495	0.7497	0.7491	0.7491	0.75	0.5648	0.75	0.7531	0.7495
ALL	9	0.7893	**0.804**	0.7307	0.7693	0.7703	0.7437	0.7197	0.774	0.785	0.6267	0.757	0.7943	0.7527
Wbc1	20	**0.9461**	0.9448	0.9332	0.9332	0.9332	0.9332	0.9332	0.9332	0.9385	0.813	0.9332	0.9142	0.9332
Wbc2	9	**0.9693**	0.9692	0.9629	0.956	0.9563	0.9559	0.9604	0.9562	0.9683	0.9542	0.9684	0.9679	0.968
Wine	9	0.9484	**0.98**	0.9663	0.9665	0.9578	0.9587	0.9649	0.9664	0.9518	0.9109	0.9665	0.9676	0.9735
Iris	2	**0.9733**	**0.9733**	**0.9733**	**0.9733**	**0.9733**	**0.9733**	**0.9733**	**0.9733**	**0.9733**	0.9587	**0.9733**	**0.9733**	**0.9733**

Table 15
The mean results of Fscore$_M$ after feature selection over 30 runs.

Data set	Feature size	IIEFA	CIEFA	FA	KM	CFA1	CFA2	NaFA	VSSFA	DA	SCA	MFA	GA	ACO
Thyroid	4	**0.8377**	0.8204	0.7628	0.753	0.7611	0.7543	0.7505	0.7719	0.7813	0.8013	0.7649	0.801	0.7467
Sonar	17	0.7072	**0.709**	0.6677	0.6814	0.6546	0.6601	0.6722	0.6461	0.6623	0.5466	0.6772	0.7019	0.6776
Balance	4	**0.8045**	0.7923	0.7735	0.7518	0.7475	0.758	0.7537	0.7726	0.7857	0.7522	0.7991	0.7991	0.7955
E. coli	5	**0.9474**	**0.9474**	0.9352	0.9109	0.9349	0.9421	0.934	0.9368	0.9337	0.9005	0.9355	0.9354	0.9308
Ozone	22	**0.7524**	0.7466	0.7408	0.7349	0.7438	0.7362	0.7359	0.7433	0.7318	0.3792	0.7419	0.7433	0.7407
ALL	9	0.7767	**0.7873**	0.7063	0.7557	0.7702	0.713	0.7017	0.7611	0.7763	0.6178	0.7336	0.7841	0.726
Wbc1	20	0.9361	**0.9394**	0.923	0.9276	0.9215	0.9261	0.9291	0.9184	0.9295	0.7795	0.9276	0.9064	0.9322
Wbc2	9	**0.9662**	0.9661	0.9588	0.9562	0.9538	0.9489	0.9555	0.9525	0.9623	0.9462	0.9646	0.964	0.9637
Wine	9	0.9242	**0.9713**	0.9519	0.9538	0.9393	0.942	0.9499	0.9536	0.9313	0.8818	0.9517	0.953	0.9622
Iris	2	**0.9602**	**0.9602**	**0.9602**	**0.9602**	**0.9602**	**0.9602**	**0.9602**	**0.9602**	**0.9602**	0.9432	**0.9602**	**0.9602**	**0.9602**

Table 16
The mean results of average sensitivity after feature selection over 30 runs.

Data set	Feature size	IIEFA	CIEFA	FA	KM	CFA1	CFA2	NaFA	VSSFA	DA	SCA	MFA	GA	ACO
Thyroid	4	**0.8122**	0.7956	0.7152	0.7085	0.7126	0.7104	0.703	0.7174	0.7519	0.7615	0.7181	0.777	0.7067
Sonar	17	0.711	0.7157	0.6829	**0.7224**	0.6538	0.6862	0.6867	0.6243	0.6633	0.5267	0.7048	0.7024	0.7024
Balance	4	**0.8038**	0.7925	0.7749	0.7478	0.7459	0.7574	0.7536	0.7727	0.7855	0.7491	0.7985	0.799	0.7953
E. coli	5	**0.9467**	**0.9467**	0.9347	0.9109	0.9344	0.9416	0.9336	0.9362	0.9333	0.8964	0.9349	0.9349	0.9304
Ozone	22	0.7435	0.731	0.7391	0.7184	**0.7503**	0.7204	0.7197	0.749	0.7007	0.4173	0.7401	0.7347	0.7381
ALL	9	0.7593	0.7467	0.6953	0.7427	**0.788**	0.6807	0.7107	0.7527	0.7713	0.726	0.724	0.7727	0.7067
Wbc1	20	0.9141	0.929	0.9028	0.9185	0.8976	0.9133	0.9237	0.8871	0.911	0.7803	0.9185	**0.9358**	0.9341
Wbc2	9	**0.9667**	0.9666	0.9561	0.9568	0.9522	0.9421	0.9525	0.9495	0.962	0.9365	0.9646	0.9639	0.9646
Wine	9	0.9286	**0.9749**	0.9563	0.9573	0.9442	0.9466	0.9546	0.9572	0.9368	0.879	0.9561	0.9562	0.9662
Iris	2	**0.9600**	**0.9600**	**0.9600**	**0.9600**	**0.9600**	**0.9600**	**0.9600**	**0.9600**	**0.9600**	0.9573	**0.9600**	**0.9600**	**0.9600**

Table 17
The mean results of average specificity after feature selection over 30 runs.

Data set	Feature size	IIEFA	CIEFA	FA	KM	CFA1	CFA2	NaFA	VSSFA	DA	SCA	MFA	GA	ACO
Thyroid	4	**0.9061**	0.8978	0.8576	0.8543	0.8563	0.8552	0.8515	0.8587	0.8759	0.8807	0.8591	0.8885	0.8533
Sonar	17	0.709	0.7062	0.6638	0.6305	0.6895	0.6476	0.669	0.7195	0.6886	**0.71**	0.6452	0.7152	0.6514
Balance	4	**0.8056**	0.7921	0.7718	0.7613	0.7529	0.7588	0.7539	0.7723	0.786	0.7606	0.8001	0.7993	0.7959
E. coli	5	**0.9733**	**0.9733**	0.9673	0.9554	0.9672	0.9708	0.9668	0.9681	0.9667	0.9482	0.9674	0.9674	0.9652
Ozone	22	0.7772	0.7844	0.7588	0.7793	0.7486	0.7789	0.7786	0.7493	**0.7993**	0.7122	0.7599	0.7714	0.7609
ALL	9	0.8193	**0.8613**	0.766	0.796	0.7527	0.8067	0.7287	0.7953	0.7987	0.5273	0.79	0.816	0.7987
Wbc1	20	**0.947**	0.9285	0.9237	0.908	0.9289	0.9133	0.9028	0.9394	0.9298	0.729	0.908	0.9069	0.8924
Wbc2	9	0.9667	0.9666	0.958	0.9367	0.9423	0.9511	0.9539	0.9446	**0.9682**	0.95	0.966	0.9655	0.9649
Wine	9	0.9619	**0.9858**	0.9751	0.9759	0.9687	0.9697	0.9741	0.9758	0.9638	0.9325	0.9755	0.9764	0.9809
Iris	2	**0.9800**	**0.9800**	**0.9800**	**0.9800**	**0.9800**	**0.9800**	**0.9800**	**0.9800**	**0.9800**	0.9787	**0.9800**	**0.9800**	**0.9800**

Overall, the average accuracy, sensitivity, specificity and Fscore_M results evidently indicate the superiority of IIEFA and CIEFA over other search methods, in terms of robustness and flexibility, for both high- and low-dimensional clustering problems in combination with feature selection. In particular, the proposed models outperform five other FA variants significantly in nearly all the test cases. Moreover, CIEFA demonstrates an evident advantage on the Wine data set than IIEFA on all five performance metrics, while attaining results similar to those of IIEFA with the rest of the data sets. Besides that, nearly all methods achieve similar scores on all five performance measures on the Iris data set (except for SCA). Since only two significant features are identified and remained after feature selection for the Iris data set, the complexity of this clustering task is significantly reduced.

The underlying reasons for the advantage demonstrated by IIEFA and CIEFA can be ascribed to the enhanced capability of exploration and exploitation contributed by the proposed search strategies. The first proposed mechanism is to intensify inward exploration by replacing the attractiveness coefficient with a random search matrix. The diversity of search directions, scales, and spaces is enhanced significantly, therefore improving the exploration ability and mitigating the constraints of biological laws. The second strategy is to intensify outward exploration by relocating the 'ineffective fireflies' to a greater and extended space out of the neighbourhoods of fireflies in comparison in the early stage of the search process. The search territory of firefly swarms is further expanded, therefore facilitating the ability of global exploration. With intensified neighbouring and global exploration from the above two strategies plus the advantages of automatic subdivision inherited from the original FA model [47], the probability of being trapped in local optima is reduced effectively, while the diversity of movement is enhanced significantly for the proposed FA models. Evidenced by the experimental and statistical results, these advantages enable the proposed FA models to undertake challenging clustering tasks with high dimensionality, noise, and less separable clusters, e.g. the ALL data set.

In contrast, some limitations related to search diversity and search efficiency can be identified in classical search methods according to empirical studies. As an example, Radcliffe and Surry [77] indicated that the GA-based clustering algorithms in some cases suffered from degeneracy resulted from the phenomenon of multiple chromosomes representing the same or very similar solutions [77]. Such degeneracy could lead to inefficient coverage of the search space, since the centroid solutions with the same or very similar configurations are repeatedly explored [78]. Moreover, multiple occurrences of the strongly favourable individuals in the GA can lead to the reproduction of many highly correlated offspring solutions, therefore reducing diversity of the population and resulting in premature convergence. Similarly, in ACO, the effect of emphasizing short paths diminishes, and search stagnation emerges when the quality of solutions becomes closer as the differences between individuals decrease [79]. Premature convergence can also occur in ACO as the sub-optimal solutions dominate the search process at an early stage, and the parameter of trail persistence is not tuned properly [69,80,81]. Consequently, owing to the potential local optima traps (GA) and search stagnation (ACO) without proper counteracting strategies, classical evolutionary algorithms such as GA and ACO are less competitive in comparison with the proposed CIEFA and IIEFA models based on results from the abovementioned five metrics including intra-cluster distances, accuracy, Fscore_M, sensitivity and specificity, as illustrated in Tables 13–17. Similar limitations are also applied to other FA variants. As an example, in the MFA model [64], each firefly not only moves towards all brighter fireflies in its neighbourhood, but also moves towards the swarm leader at the same time. The search diversity and exploration capability of the firefly swarm are obstructed owing to the continuous exposure to attraction of the global best solution during the search process. Consequently, the firefly swarm is more likely to converge prematurely, and be trapped in local optima.

Overall, owing to the assistance of the two proposed strategies, CIEFA and IIEFA are able to overcome local optima traps and outperform classical search methods, i.e. GA, ACO, FA, DA and SCA. They also outperform advanced FA variants employed in this study i.e. CFA1, CFA2, NaFA, VSSFA, and MFA. Additionally, the merits of the proposed strategies also indicate that a strict adherence to biological laws imposes certain constraints on the exploration ability of heuristic search algorithms. As a result, the original biological laws from nature need to be further extracted

Table 18
The mean ranking results based on the Friedman test for the CIEFA model.

Algorithms	Mean ranking based on distance measure	Algorithms	Mean ranking based on 1/Accuracy	Algorithms	Mean ranking based on 1/Fscore$_M$
CIEFA	1.40	CIEFA	1.80	CIEFA	1.80
GA	3.25	GA	3.95	GA	4.20
DA	4.05	MFA	4.70	MFA	4.90
MFA	4.70	DA	5.25	ACO	5.80
ACO	6.20	ACO	6.10	VSSFA	6.45
VSSFA	6.80	VSSFA	6.50	DA	6.50
FA	7.45	FA	7.25	FA	6.80
NaFA	7.65	CFA2	7.65	KM	7.25
CFA1	7.75	CFA1	7.90	CFA1	7.70
CFA2	7.85	KM	8.00	CFA2	7.80
KM	9.70	NaFA	8.10	NaFA	8.00
SCA	11.20	SCA	10.80	SCA	10.80

Table 19
The mean ranking results based on the Friedman test for the IIEFA model.

Algorithms	Mean ranking based on distance measure	Algorithms	Mean ranking based on 1/Accuracy	Algorithms	Mean ranking based on 1/Fscore$_M$
IIEFA	2.40	IIEFA	2.60	IIEFA	2.60
GA	3.10	GA	3.85	GA	4.10
DA	3.90	MFA	4.70	MFA	4.90
MFA	4.60	DA	5.15	ACO	5.80
ACO	6.10	ACO	6.10	VSSFA	6.35
VSSFA	6.70	VSSFA	6.40	DA	6.40
FA	7.35	FA	7.15	FA	6.70
NaFA	7.55	CFA2	7.55	KM	7.15
CFA1	7.65	CFA1	7.80	CFA1	7.60
CFA2	7.75	KM	7.90	CFA2	7.70
KM	9.70	NaFA	8.00	NaFA	7.90
SCA	11.20	SCA	10.80	SCA	10.80

and refined to best facilitate the effectiveness and discard potential restrictions in the development of metaheuristic algorithms. Furthermore, there is other insightful research on metaheuristic algorithms, which provides promising directions for future investigation [67].

4.6. Statistical tests

To examine the significance of the performance difference between the proposed models and baseline methods, both Friedman and Wilcoxon rank sum tests are conducted.

4.6.1. The Friedman test

In the Friedman test, a test statistic Q is constructed based on the mean rankings of test treatments, which can be approximated by a chi-squared distribution. Then, the null hypothesis that K treatments come from the same population is tested according to the p-values given by $P(\chi^2_{k-1} > Q)$ [82,83]. The Friedman test is conducted with respect to three main comprehensive performance metrics (intra-cluster distance measures, average clustering accuracy, and Fscore$_M$) for IIEFA and CIEFA. Tables 18–19 show the mean ranking results of the three performance metrics for the CIEFA and IIEFA models, respectively. For each metric, the mean ranking of each method is obtained by averaging its rankings over ten data sets based on the results shown in Tables 13–15. The significance level is set to 0.05 (i.e. $\alpha = 0.05$) as the confidence level in all test cases. Tables 20–21 show the details of statistical test results for the CIEFA and IIEFA models, respectively.

As indicated in Tables 18–19, the proposed CIEFA and IIEFA models dominate the highest rankings, and demonstrate clear advantages in the performance metrics of intra-cluster distance measure, accuracy, and Fscore$_M$ with the Friedman test. In comparison with the five FA variants, i.e. VSSFA, NaFA, CFA1, CFA2, and MFA, the proposed models achieve significant improvements in all three performance metrics, indicating the advantages of the proposed search mechanisms. The proposed FA models also outperform KM and five classical search methods, i.e. GA, ACO, FA DA, and SCA. Comparatively, the CIEFA model achieves a better ranking than that of the IIEFA model in overall evaluation based on the experimental results. Furthermore, as indicted in Tables 20–21, the p-values of the Friedman test are all lower than 0.05 with respect to each metric for both the IIEFA and CIEFA models, which suggest an overall statistically significant difference between the mean ranks of IIEFA and CIEFA as compared with those of other test algorithms.

Table 20
Statistical results of the Friedman test for the CIEFA model.

Algorithms	Chi-Square	p-value	Hypothesis
Fitness	65.948929	7.1418E−10	Rejected
1/Accuracy	52.348099	2.3578E−07	Rejected
1/Fscore$_M$	45.847933	3.0000E−06	Rejected

Table 21
Statistical results of the Friedman test for the IIEFA model.

Algorithms	Chi-Square	p-value	Hypothesis
Fitness	59.698571	1.0547E−08	Rejected
1/Accuracy	46.035280	3.0000E−06	Rejected
1/Fscore$_M$	39.724308	4.0000E−05	Rejected

4.6.2. The Wilcoxon rank sum test

The Wilcoxon rank sum test is conducted based on the mean accuracy rates of all the methods to further indicate the statistical distinctiveness of the proposed FA models against each baseline method. As indicated in Tables 22–23, the majority of the test results are lower than 0.05 for both CIEFA and IIEFA models, which indicate the proposed FA models significantly outperform 11 baseline algorithms with respect to most of data sets from the statistical perspective. The Iris data set is an exception since all the algorithms except for SCA achieve the same highest accuracy of 97.33% with feature selection. Moreover, as shown in Tables 22–23, in comparison with CIEFA, IIEFA demonstrates higher frequencies of insignificant difference in clustering accuracy as compared with those of the baseline models. This tendency becomes more evident on the ALL data set, since IIEFA does not show statistically significant differences as compared with seven baseline methods, i.e. KM, CFA1, VSSFA, DA, MFA, GA, and ACO, while for CIEFA, a similar case only occurs to two baseline methods, i.e. GA and VSSFA. This phenomenon may be attributed to the challenging factors of the ALL data set, owing to its high dimensionality and highly inseparable data distributions caused by the subtle differences between the normal and blast cases. On the other hand, the advantage demonstrated by CIEFA over IIEFA on the ALL data set can be ascribed to the proposed dispersing mechanism, which further enhances the exploration capability on the basis of IIEFA and reduces the probability of being trapped in local optima. Therefore, CIEFA is capable of delivering better clustering performances than those of IIEFA in tackling data samples with complex distributions and narrow class margins.

In summary, the proposed IIEFA and CIEFA models outperform other algorithms in clustering problems from two perspectives, i.e. (1) constructing more compact clusters with fast convergence rates, and (2) improving clustering performance in terms of accuracy, sensitivity, specificity and Fscore_M measurements, with fewer parameter settings. Moreover, CIEFA demonstrates more advantages than IIEFA especially with data sets containing inseparable and less compact clusters (i.e. ALL) owing to its enhanced exploration capability. Despite a wide variety of characteristics demonstrated by the above ten data sets, real-life clustering tasks can pose greater challenges to the proposed clustering models owing to the elevated feature dimensionalities and more complex cluster distributions. As such, we further examine the efficiency of the proposed models against other baseline methods by undertaking three additional clustering tasks with higher dimensionalities and larger numbers of classes.

5. Evaluation on high-dimensional clustering tasks with complex cluster distributions

After examining the proposed IIEFA and CIEFA models both theoretically and experimentally in the above section, we further extend our evaluation to more challenging clustering tasks with both high dimensionalities and complex cluster distributions, as an attempt to ascertain the performance of the proposed methods more comprehensively. The extended evaluation is conducted with three additional high-dimensional UCI data sets [74], i.e. Drivface, Micromass, and Gas Sensor Array Drift (Sensor). The dimensionalities of Driveface, Micromass, and Sensor data sets are 6400, 1300, and 128, while the numbers of classes are 3, 5, and 5, respectively. The details of the data sets are provided in Table 24. These data sets pose significant challenges to any clustering models owing to the considerable expansion of data dimensionalities, as well as the large numbers of classes and more complex cluster distributions, as compared with those in our previous experiments and related studies. We employ the same experimental settings as those provided in Section 4, i.e. the maximum number of function evaluations=population (50) $\times$ the maximum number of iterations (200), and runs=30. The clustering results of intra-cluster distance, accuracy, and Fscore_M over 30 runs are illustrated in Tables 25–27, respectively.

As shown in Tables 25–27, the advantages of the proposed CIEFA and IIEFA models are further ascertained by the empirical results on these high-dimensional data sets. Specifically, as indicated in Table 25, with respect to the distance measure, CIEFA yields the smallest intra-cluster distances on both the Micromass and Sensor data sets, while IIEFA produces the most compact clusters with the smallest intra-cluster distances on the Drivface data set. Moreover, as depicted in Tables 26–27, the proposed CIEFA model achieves the best accuracy rates as well as Fscore_M results on all the three data sets, followed closely by those of IIEFA. Both proposed models show better performances than those of all the baseline methods.

Furthermore, we conduct the Wilcoxon rank sum test based on the accuracy rates. The statistical test results of CIEFA and IIEFA are provided in Tables 28–29 respectively. As shown in Tables 28–29, the majority of test results are lower than 0.05 for both CIEFA and IIEFA models. This indicates the statistical advantage of the proposed models over 11 baseline methods by delivering significantly better clustering results on these three high-dimensional data sets. Some baseline algorithms also exhibit competitive performances in some test cases, i.e. CFA1 and GA show similar result distributions to those of CIEFA on Drivface and Micromass, respectively, while CFA1 and CFA2 have similar performances as those of IIEFA on Drivface and Micromass, respectively, with the GA obtaining similar results to those of IIEFA for Micromass and Sensor, respectively.

It is inevitable to encounter the challenge of the curse of dimensionality when dealing with such high-dimensional data sets. With respect to clustering analysis in our study, the curse of dimensionality exacerbates the adversity of search for optimal centroids during the clustering process from two perspectives. Firstly, the diagonal-based search prescribed in the original FA model is likely to miss the potential promising areas due to the constraints in its search directions and scales. The situation becomes worse on high-dimensional data sets owing to the considerable expansion of the search space as the problem dimensionality increases, as well as the resulting oversight of larger search sub-spaces. Therefore, the exploration capabilities of the search methods in directions and scales have great impact on model performance for the clustering tasks on such high-dimensional data sets. The proposed models are able to release the search operation from the diagonal-based search in the original FA model to the region-based multi-dimensional exploration with a greater variety of directions and scales, in order to address these challenges.

Secondly, as analysed in [84], when tackling high-dimensional clustering tasks, the sparse distribution of high-dimensional data sets makes the computation of the intra-cluster distance measures less respondent to the shift of initial centroids. This could result in pre-mature convergence and early stagnation. To overcome such barriers, the CIEFA model employs a dispatching mechanism to scatter fireflies with high similarities to other unexploited territories, in order to increase search diversity and avoid local optima traps. Such search capabilities become vital when dealing with high-dimensional, large search spaces with sparse data distributions.

In summary, the proposed CIEFA and IIEFA models demonstrate significant advantages in dealing with high-dimensional data sets over the baseline methods owing to wider and more effective exploration of the search space and the improved population and search diversity. Nevertheless, clustering on high-dimensional data sets remains a challenging topic owing to the possible presence of the redundant, noisy, and irrelevant features that can severely affect the model performance. Other search strategies and feature selection models will be further explored to enhance model efficiency in dealing with high-dimensional clustering tasks in future studies.

Table 22
The Wilcoxon rank sum test results of the proposed CIEFA model.

Data set	FA	KM	CFA1	CFA2	NaFA	VSSFA	DA	SCA	MFA	GA	ACO
Thyroid	2.02E−06	7.99E−07	1.45E−06	9.95E−07	5.99E−07	1.90E−06	1.56E−02	4.01E−02	3.70E−06	9.33E−01	1.32E−06
Sonar	1.07E−08	6.06E−06	1.51E−06	3.81E−08	6.03E−06	7.21E−09	9.00E−08	8.32E−11	2.55E−07	**6.49E−01**	4.03E−06
Balance	2.89E−05	5.54E−07	4.18E−08	7.98E−08	1.72E−09	3.84E−06	**1.41E−01**	6.85E−03	3.91E−03	1.80E−02	1.14E−03
E. coli	2.15E−02	2.84E−05	1.10E−02	6.45E−04	2.77E−03	**1.61E−01**	1.37E−03	4.43E−12	1.10E−02	1.32E−03	1.42E−04
Ozone	2.88E−11	3.21E−11	1.80E−11	1.45E−11	2.53E−11	2.53E−11	8.57E−12	1.73E−11	8.57E−12	2.42E−06	1.80E−11
ALL	1.18E−04	1.30E−02	4.02E−02	5.81E−04	3.58E−04	**9.88E−01**	1.52E−02	1.48E−09	2.67E−02	**6.04E−01**	3.64E−03
Wbc1	5.01E−13	5.01E−13	5.01E−13	5.01E−13	5.01E−13	5.01E−13	6.11E−11	1.54E−11	5.01E−13	6.83E−12	5.01E−13
Wbc2	3.10E−10	3.44E−13	5.47E−13	4.39E−13	5.65E−11	3.41E−13	3.33E−04	5.94E−11	1.45E−05	2.96E−11	**6.48E−02**
Wine	3.49E−08	5.59E−09	1.13E−09	7.21E−09	4.17E−10	2.62E−09	3.93E−05	6.38E−12	2.68E−07	6.67E−08	2.28E−09
Iris	**1.00E+00**	**1.00E+00**	**1.00E+00**	**1.00E+00**	**1.00E+00**	**1.00E+00**	**1.00E+00**	4.91E−04	**1.00E+00**	**1.00E+00**	**1.00E+00**

Table 23
The Wilcoxon rank sum test results of the proposed IIEFA model.

Data set	FA	KM	CFA1	CFA2	NaFA	VSSFA	DA	SCA	MFA	GA	ACO
Thyroid	3.42E−08	1.99E−08	3.03E−08	2.67E−08	1.39E−08	4.85E−08	9.74E−04	3.09E−03	6.11E−08	**8.32E−02**	2.77E−08
Sonar	1.19E−08	1.79E−05	4.80E−06	9.18E−08	1.48E−05	7.40E−09	2.83E−07	9.03E−11	6.19E−07	**8.06E−01**	1.10E−05
Balance	2.55E−05	6.22E−07	5.33E−08	9.90E−08	2.81E−09	5.25E−06	7.79E−02	3.41E−03	4.40E−03	1.58E−02	1.32E−03
E. coli	2.15E−02	2.84E−05	1.10E−02	6.45E−04	2.77E−03	**1.61E−01**	1.37E−03	4.43E−12	1.10E−02	1.32E−03	1.42E−04
Ozone	1.34E−12	1.65E−12	6.09E−13	4.40E−13	1.06E−12	1.06E−12	2.05E−13	7.73E−12	2.05E−13	8.43E−11	6.09E−13
ALL	5.87E−03	**3.38E−01**	**5.44E−01**	2.56E−02	7.92E−03	**4.48E−01**	**2.79E−01**	5.78E−09	**2.93E−01**	**4.64E−01**	**1.18E−01**
Wbc1	6.47E−13	6.47E−13	6.47E−13	6.47E−13	6.47E−13	6.47E−13	2.20E−11	1.86E−11	6.47E−13	8.41E−12	6.47E−13
Wbc2	4.99E−11	1.58E−13	2.59E−13	2.05E−13	1.20E−11	1.57E−13	2.47E−05	3.31E−11	5.19E−07	7.15E−13	5.56E−03
Wine	3.52E−04	1.43E−04	4.12E−05	1.21E−04	2.83E−05	9.41E−05	1.15E−02	4.99E−07	9.77E−04	5.90E−04	8.17E−05
Iris	**1.00E+00**	**1.00E+00**	**1.00E+00**	**1.00E+00**	**1.00E+00**	**1.00E+00**	**1.00E+00**	4.91E−04	**1.00E+00**	**1.00E+00**	**1.00E+00**

Table 24
Three high-dimensional data sets with multiple classes.

Data set	Number of attributes	Number of classes	Missing values	Number of instances
Drivface	6400	3	No	81
Micromass	1300	5	No	180
Sensor	128	5	No	415

Table 25
The mean results of the minimum intra-cluster distance measure on high-dimensional data sets over 30 runs.

Data set	IIEFA	CIEFA	FA	KM	CFA1	CFA2	NaFA	VSSFA	DA	SCA	MFA	GA	ACO
Drivface	**4849.4**	4869.6	4857.9	322247	4948.2	4908.6	4919.2	4938.4	4926.9	6161.3	5015.5	4942.9	4922.1
Micromass	656.91	**653.41**	673.38	2626.3	677.28	664.91	667.21	671.38	671.86	813.75	672.43	666.89	674.51
Sensor	426.26	**422.13**	446.94	534.03	439.08	435.31	435.34	449.97	440.72	1171.9	438.28	443.84	440.46

Table 26
The mean results of average accuracy on high-dimensional data sets over 30 runs.

Data set	IIEFA	CIEFA	FA	KM	CFA1	CFA2	NaFA	VSSFA	DA	SCA	MFA	GA	ACO
Drivface	0.7687	**0.7748**	0.7561	0.7583	0.7558	0.7424	0.7484	0.7536	0.7556	0.6593	0.7479	0.7605	0.7588
Micromass	0.8582	**0.8644**	0.819	0.831	0.8101	0.8381	0.8316	0.8256	0.8177	0.833	0.8221	0.8387	0.8177
Sensor	0.8118	**0.8187**	0.7928	0.8006	0.7959	0.7965	0.8003	0.7882	0.7922	0.7652	0.799	0.7968	0.7944

Table 27
The mean results of $Fscore_M$ on high-dimensional data sets over 30 runs.

Data set	IIEFA	CIEFA	FA	KM	CFA1	CFA2	NaFA	VSSFA	DA	SCA	MFA	GA	ACO
Drivface	0.6524	**0.6618**	0.6502	0.6573	0.6401	0.6403	0.6436	0.6484	0.6433	0.5526	0.6355	0.6535	0.6551
Micromass	0.6332	**0.6402**	0.5397	0.5607	0.5158	0.5638	0.5535	0.54	0.512	0.6082	0.5529	0.5631	0.5289
Sensor	0.6089	**0.6255**	0.531	0.556	0.5399	0.5472	0.5602	0.5182	0.5259	0.5125	0.5573	0.5508	0.5373

6. Further comparison analysis between IIEFA and CIEFA

We further explore the performance distinction between CIEFA and IIEFA using challenging clustering tasks with noise, complicated data distributions, and non-compact and less separable clusters. Specifically, to better distinguish between CIEFA and IIEFA, both models are further evaluated with another four challenging high-dimensional data sets, i.e. a skin lesion data set (denoted as Lesion) [85], as well as three UCI data sets [74], i.e. Human Activity (Activity), Libras Movements (Libras), and Mice Protein Expression (Protein). The skin lesion data set is used in [85], which extracted shape, colour, and texture features of 660 dermoscopic skin lesion images from the Edinburgh Research and Innovation (Dermofit) lesion data set [86]. A 98-dimension feature vector for each skin lesion image was then obtained to represent the lesion information for subsequent clustering analysis. Moreover, the dimensionalities of the Human Activity, Libras, and Mice Protein data sets are 560, 90, and 77, respectively. In this research, we employ three classes for the Libras data set and two classes for the Skin Lesion, Human Activity and Mice

Table 28
The Wilcoxon rank sum test results of the proposed CIEFA model on high-dimensional data sets.

Data set	FA	KM	CFA1	CFA2	NaFA	VSSFA	DA	SCA	MFA	GA	ACO
Drivface	3.44E−03	1.74E−02	**7.51E−02**	4.09E−03	1.23E−03	3.15E−02	3.15E−02	4.94E−10	4.24E−03	1.38E−02	2.67E−02
Micromass	3.32E−04	2.66E−03	6.60E−04	4.18E−02	2.69E−02	2.90E−03	7.93E−04	9.29E−05	3.53E−03	**7.67E−02**	1.14E−04
Sensor	4.57E−06	6.13E−04	4.15E−04	6.02E−05	1.24E−04	1.88E−06	1.96E−05	5.68E−11	7.21E−04	2.12E−04	8.80E−05

Table 29
The Wilcoxon rank sum test results of the proposed IIEFA model on high-dimensional data sets.

Data set	FA	KM	CFA1	CFA2	NaFA	VSSFA	DA	SCA	MFA	GA	ACO
Drivface	4.06E−03	2.21E−02	**1.17E−01**	5.99E−03	1.40E−03	4.54E−02	4.62E−02	4.35E−10	5.61E−03	1.74E−02	3.65E−02
Micromass	1.49E−04	4.60E−03	9.77E−04	**6.58E−02**	3.57E−02	1.96E−03	4.83E−04	6.52E−05	5.02E−03	**9.20E−02**	8.36E−05
Sensor	4.18E−04	4.33E−02	3.42E−02	1.91E−03	1.84E−02	5.71E−05	1.63E−03	3.64E−11	3.94E−02	**5.49E−02**	3.97E−03

Protein data sets respectively. Details of the data sets are shown in Table 30. For each high-dimensional data set, a total of 30 runs are conducted for each proposed model. In order to fully evaluate the model efficiency, no feature selection is applied. The detailed clustering results are provided in Table 31.

As illustrated in Table 31, the empirical results of the CIEFA model for these high-dimensional data sets demonstrate sufficient advantages over those of IIEFA according to five performance metrics, i.e. intra-cluster distances, accuracy, sensitivity, specificity, and Fscore_{M}, over 30 runs. As an example, the CIEFA model achieves higher average accuracy rates of 67.12%, 80.20%, 76.62%, and 79.07% for the Human Activity, Skin Lesion, Mice Protein, and Libras data sets, respectively, while maintaining lower intra-cluster distances with these data sets. In contrast, the IIEFA model produces comparatively slightly lower accuracy rates of 64.36%, 78.54%, 72.38%, and 78.01% for the Human Activity, Skin Lesion, Mice Protein, and Libras data sets, respectively, while producing slightly higher intra-cluster distances. A similar observation can be obtained for the other three performance metrics, i.e. sensitivity, specificity, and Fscore_{M}, for both models on most of the test cases. This indicates that the CIEFA model offers a better option, as compared with IIEFA, to undertake high-dimensional clustering tasks. This finding is also in agreement with that obtained by the experimental studies using three high-dimensional data sets as discussed in Section 5.

As discussed above, complexity of clustering tasks is significantly increased on these high-dimensional data sets owing to a higher probability of inclusion of noise and redundant or contradictory features. The clustering tasks could be even more challenging especially when the data samples are not well-separated, and their distributions are far different from compact spherical. As an example, the skin lesion data set [85] consists of two types of lesions, benign and malignant. The appearance difference between these two types of lesions in terms of shape, colour and texture can be very subtle, which sometimes causes confusion even to dermatologists, therefore posing great challenges on the clustering tasks. In other words, this high-dimensional skin lesion data set contains highly inseparable and non-compact clusters. The enhanced exploration capability acquired from the additional dispersing mechanism in CIEFA accounts for its efficiency in identifying optimal centroids for this challenging lesion problem, as well as other UCI data sets, as compared with IIEFA.

In summary, the dispersing mechanism in CIEFA is able to boost the exploration capability by dispatching fireflies with high similarities in fitness values to the extended and unexploited search space. As such, the probability of identifying optimal centroids closer to the global optima is increased with the assistance of intensified local exploration as well as the expanded search territory. Therefore, CIEFA offers a better option, as compared with IIEFA, to deal with challenging clustering tasks such as data samples with high dimensionality, noise, and complicated distributions.

7. Conclusion

In this research, we have proposed two FA variants, namely IIEFA and CIEFA, to undertake the problems associated with initialization sensitivity and local optima traps of the conventional KM clustering algorithm. Two new strategies have been proposed in IIEFA and CIEFA to increase search diversification and efficiency. Firstly, the attractiveness coefficient in the original FA model is substituted by a randomized control matrix, therefore the one-dimensional search strategy in the original FA model is elevated to a multi-dimensional search mechanism with greater search scales and directions for exploration in the neighbourhood. Secondly, in the early stage of the search process, a firefly solution sharing a high similarity with another is relocated to a new position outside the scope between the two fireflies in comparison. As such, the chances of identifying global optima and avoiding local optima are enhanced, owing to the fact that fireflies with high similarities are dispersed and the distribution of the whole swarm is more diversified. Therefore, the search efficiency is improved with the guarantee of sufficient variance between fireflies in comparison at the early convergence stage. The performances of IIEFA- and CIEFA-enhanced KM clustering methods are first investigated with ALL and 9 other UCI data sets, which include both high-dimensional and low-dimensional problems. In combination with mRMR-based feature selection, the proposed methods show superiority over the KM clustering algorithm, five classical search methods, and five other FA variants in terms of the convergence speed and clustering performance with respect to average accuracy rates, sensitivity, specificity, and macro-average F-score (Fscore_M) over 30 runs. The results have been ascertained using Friedman and Wilcoxon rank sum tests. In short, the proposed search strategies account for the improved efficiency in enhancing the cluster centroids of original KM clustering, which in turn overcome the local optima traps. Moreover, we have conducted a further evaluation using three additional high-dimensional UCI data sets, and the results reinforce the effectiveness and advantage of the proposed models over the baseline methods in dealing with high-dimensional clustering tasks. Lastly, a dedicated comprehensive study has also been conducted to further identify the distinctiveness between IIEFA and CIEFA using four additional high-dimensional data sets. The empirical results indicate that CIEFA outperforms IIEFA in dealing with challenging clustering tasks with noise, complicated data distributions, and non-compact and less separable clusters, owing to its enhanced exploration capability and expanded search territory.

For future research, other objective functions with the consideration of both inter- and intra-cluster measurements will be employed to enhance the proposed models for dealing with complex and irregular data distribution problems. The proposed FA variants will also be evaluated using other optimization tasks,

Table 30
Four additional high-dimensional data sets for further comparison between IIEFA and CIEFA.

Data set	Number of attributes	Number of classes	Missing values	Number of instances
Activity	560	2	No	600
Lesion	98	2	No	660
Protein	77	2	No	300
Libras	90	3	No	72

Table 31
The mean clustering results over 30 independent runs with four high-dimensional data sets.

Criteria	Human Activity (560 dim)		Skin Lesion (98 dim)		Mice Protein (77 dim)		Libras (90 dim)	
	IIEFA	CIEFA	IIEFA	CIEFA	IIEFA	CIEFA	IIEFA	CIEFA
Fitness	12785	**12582**	5399.8	**5352.8**	2345.0	**2307.1**	466.05	**462.95**
Accuracy	0.6436	**0.6712**	0.7854	**0.802**	0.7238	**0.7662**	0.7801	**0.7907**
Fscore_M	0.6056	**0.6841**	0.8036	**0.8103**	0.7086	**0.7523**	0.6883	**0.6994**
Sensitivity	0.6303	**0.7123**	**0.7898**	0.7750	**0.7562**	0.7180	0.6693	**0.6861**
Specificity	**0.6568**	0.6300	0.7800	**0.8344**	0.6913	**0.8144**	0.8342	**0.8431**

such as discriminative feature selection [39,85,87], image segmentation [88], and evolving deep neural network generation [89, 90].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

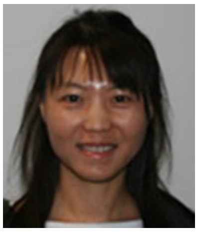
We appreciate the funding support received from Northumbria University for this research.

References

[1] A.K. Jain, Data Clustering: 50 Years beyond K-Means, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.
[2] A. Elazab, C. Wang, F. Jia, J. Wu, G. Li, Q. Hu, Segmentation of brain tissues from magnetic resonance images using adaptively regularized kernel-based Fuzzy c-means clustering, Comput. Math. Methods Med. 2015 (2015) 485495.
[3] M. Gong, Y. Liang, J. Shi, W. Ma, J. Ma, Fuzzy C-Means clustering with local information and kernel metric for image segmentation, IEEE Trans. Image Process. 22 (2) (2013) 573–584.
[4] G.B. Kande, P.V. Subbaiah, T.S. Savithri, Unsupervised fuzzy based vessel segmentation in pathological digital fundus images, J. Med. Syst. 34 (5) (2010) 849–858.
[5] A. Mekhmoukh, K. Mokrani, Improved Fuzzy C-means based particle swarm optimization (PSO) initialization and outlier rejection with level set methods for MR brain image segmentation, Comput. Methods Programs Biomed. 122 (2) (2015) 266–281.
[6] R.G.L. Balasubramanian, Macula segmentation and fovea localization employing image processing and heuristic based clustering for automated retinal screening, Comput. Methods Programs Biomed. 160 (2018) 153–163.
[7] X. Tu, J. Gao, C. Zhu, J.Z. Cheng, Z. Ma, X. Dai, M. Xie, MR Image segmentation and bias field estimation based on coherent local intensity clustering with total variation regularization, Med. Biol. Eng. Comput. 54 (12) (2016) 1807–1818.
[8] J. Wang, J. Kong, Y.H. Lu, M. Qi, B.X. Zhang, A modified FCM algorithm for MRI brain image segmentation using both local and non-local spatial constraints, Comput. Med. Imaging Graph. 32 (8) (2008) 685–698.
[9] A. Chitra, A. Rajkumar, Paraphrase extraction using fuzzy hierarchical clustering, Appl. Soft Comput. 34 (C) (2015) 426–437.
[10] M. Mahdavi, H. Abolhassani, Harmony k-means algorithm for document clustering, Data Min. Knowl. Discov. 18 (3) (2009) 370–391.
[11] L.M. Abualigah, A.T. Khader, M.A. Al-Betar, O.A. Alomari, Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering, Expert Syst. Appl. 84 (C) (2017) 24–36.
[12] G. Iván, V. Grolmusz, On dimension reduction of clustering results in structural bioinformatics, Biochim. Biophys. Acta 1844 (12) (2014) 2277–2283.
[13] Triguero S. del Río, V. López, J. Bacardit, J.M. Benítez, F. Herrera, ROSEFW-RF: The winner algorithm for the ECBDL'14 big data competition: An extremely imbalanced big data bioinformatics problem, Knowl.-Based Syst. 87 (2015) 69–79.
[14] C.-M. Liu, C.-H. Lee, L.-C. Wang, Distributed clustering algorithms for data-gathering in wireless mobile sensor networks, J. Parallel Distrib. Comput. 67 (11) (2007) 1187–1200.
[15] J. Zhu, C.-H. Lung, V. Srivastava, A hybrid clustering technique using quantitative and qualitative data for wireless sensor networks, Ad Hoc Netw. 25 (2015) 38–53.
[16] Y. Marinakis, M. Marinaki, M. Doumpos, C. Zopounidis, Ant colony and particle swarm optimization for financial classification problems, Expert Syst. Appl. 36 (7) (2009) 10604–10611.
[17] A. Likas, N. Vlassis, J.J. Verbeek, The global k-means clustering algorithm, Pattern Recognit. 36 (2) (2003) 451–461.
[18] G. Gan, M.K.-P. Ng, K-means clustering with outlier removal, Pattern Recognit. Lett. 90 (2017) 8–14.
[19] Y.P. Raykov, A. Boukouvalas, F. Baig, M.A. Little, What to do when K-means clustering fails: A simple yet principled alternative algorithm, PLOS ONE 11 (9) (2016) e0162259.
[20] K.S. Al-Sultan, A Tabu search approach to the clustering problem, Pattern Recognit. 28 (9) (1995) 1443–1451.
[21] C.S. Sung, H.W. Jin, A tabu-search-based heuristic for clustering, Pattern Recognit. 33 (5) (2000) 849–858.
[22] S.Z. Selim, K. Alsultan, A simulated annealing algorithm for the clustering problem, Pattern Recognit. 24 (10) (1991) 1003–1008.
[23] U. Maulik, S. Bandyopadhyay, Genetic algorithm-based clustering technique, Pattern Recognit. 33 (9) (2000) 1455–1465.
[24] D. Karaboga, C. Ozturk, A novel clustering approach: Artificial bee colony (ABC) algorithm, Appl. Soft Comput. 11 (1) (2011) 652–657.
[25] P. Das, D.K. Das, S. Dey, A modified bee colony optimization (MBCO) and its hybridization with k-means for an application to data clustering, Appl. Soft Comput. 70 (2018) 590–603.
[26] P.S. Shelokar, V.K. Jayaraman, B.D. Kulkarni, An ant colony approach for clustering, Anal. Chim. Acta 509 (2) (2004) 187–195.
[27] T. Niknam, B. Amiri, An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis, Appl. Soft Comput. 10 (1) (2010) 183–197.
[28] C.Y. Chen, F. Ye, Particle swarm optimization algorithm and its application to clustering analysis, in: Proceedings of IEEE International Conference on Networking, Sensing and Control, 2004.
[29] A. Bouyer, A. Hatamlou, An efficient hybrid clustering method based on improved cuckoo optimization and modified particle swarm optimization algorithms, Appl. Soft Comput. 67 (2018) 172–182.
[30] S.I. Boushaki, N. Kamel, O. Bendjeghaba, A new quantum chaotic cuckoo search algorithm for data clustering, Expert Syst. Appl. 96 (2018) 358–372.
[31] J. Senthilnath, S.N. Omkar, V. Mani, Clustering using firefly algorithm: Performance study, Swarm Evol. Comput. 1 (3) (2011) 164–171.
[32] L. Zhou, L. Li, Improvement of the Firefly-based K-means Clustering Algorithm, in: Proceedings of the 2018 International Conference on Data Science, 2018, pp. 157-162.
[33] A. Hatamlou, S. Abdullah, H. Nezamabadi-pour, A combined approach for clustering based on k-means and gravitational search algorithms, Swarm Evol. Comput. 6 (2012) 47–52.
[34] X.H. Han, L. Quan, X.Y. Xiong, M. Almeter, J. Xiang, Y. Lan, A novel data clustering algorithm based on modified gravitational search algorithm, Eng. Appl. Artif. Intell. 61 (2017) 1–7.
[35] A. Hatamlou, Black hole: A new heuristic optimization approach for data clustering, Inform. Sci. 222 (2013) 175–184.

[36] A. Hatamlou, S. Abdullah, M. Hatamlou, Data clustering using big bang–big crunch algorithm, in: Innovative Computing Technology, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.
[37] X.S. Yang, Firefly algorithm, levy flights and global optimization, in: Research and Development in Intelligent Systems, 2010, pp. 209–218.
[38] N, Nekouie, A new method in multimodal optimization based on firefly algorithm, Artif. Intell. Rev. 46 (2) (2016) 267–287.
[39] L. Zhang, W. Srisukkham, S.C. Neoh, C.P. Lim, D. Pandit, Classifier ensemble reduction using a modified firefly algorithm: An empirical evaluation, Expert Syst. Appl. 93 (2018) 395–422.
[40] X.-S. Yang, Firefly algorithms for multimodal optimization, in: Stochastic Algorithms: Foundations and Applications, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.
[41] A.H. Gandomi, X.-S. Yang, A.H. Alavi, Mixed variable structural optimization using firefly algorithm, Comput. Struct. 89 (23) (2011) 2325–2336.
[42] X.-S. Yang, Multiobjective firefly algorithm for continuous optimization, Eng. Comput. 29 (2) (2013) 175–184.
[43] H. Peng, F. Long, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.
[44] H.M. Moftah, A.T. Azar, E.T. Al-Shammari, N.I. Ghali, A.E. Hassanien, Adaptive k-means clustering algorithm for MR breast image segmentation, Neural Comput. Appl. 24 (7) (2014) 1917–1928.
[45] X.S. Peng, C. Zhou, D.M. Hepburn, M.D. Judd, W.H. Siew, Application of k-means method to pattern recognition in on-line cable partial discharge monitoring, IEEE Trans. Dielectr. Electr. Insul. 20 (3) (2013) 754–761.
[46] K. Wagstaff, S. Rogers, Constrained k-means clustering with background knowledge, in: Proceedings of the Eighteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc, 2001, pp. 577–584.
[47] X.S. Yang, Nature-Inspired Metaheuristic Algorithms, 2010.
[48] H. Wang, W. Wang, X. Zhou, H. Sun, J. Zhao, X. Yu, Z. Cui, Firefly algorithm with neighborhood attraction, Inform. Sci. 382–383 (2017) 374–387.
[49] I. Fister, I. Fister Jr, X.S. Yang, J. Brest, A comprehensive review of firefly algorithms, Swarm Evol. Comput. 13 (2013) 34–46.
[50] F.B. Ozsoydan, A. Baykasoğlu, Quantum firefly swarms for multimodal dynamic optimization problems, Expert Syst. Appl. 115 (2019) 189–199.
[51] A. Banerjee, D. Ghosh, S. Das, Modified firefly algorithm for area estimation and tracking of fast expanding oil spills, Appl. Soft Comput. 73 (2018) 829–847.
[52] A. Baykasoğlu, F.B. Ozsoydan, An improved firefly algorithm for solving dynamic multidimensional knapsack problems, Expert Syst. Appl. 41 (8) (2014) 3712–3725.
[53] A.K. Sadhu, A. Konar, T. Bhattacharjee, S. Das, Synergism of firefly algorithm and q-learning for robot arm path planning, Swarm Evol. Comput. 43 (2018) 50–68.
[54] L. Zhang, K. Mistry, C.P. Lim, S.C. Neoh, Feature selection using firefly optimization for classification and regression models, Decis. Support Syst. 106 (2018) 64–85.
[55] M. Alweshah, S. Abdullah, Hybridizing firefly algorithms with a probabilistic neural network for solving classification problems, Appl. Soft Comput. 35 (2015) 513–524.
[56] O.P. Verma, D. Aggarwal, T. Patodi, Opposition and dimensional based modified firefly algorithm, Expert Syst. Appl. 44 (2016) 168–176.
[57] A. Kazem. E. Sharifi, F.K. Hussain, M. Saberic, O.K. Hussain, Support vector regression with chaos-based firefly algorithm for stock market price forecasting, Appl. Soft Comput. 13 (2) (2013) 947–958.
[58] A.H. Gandomi, X.S. Yang, S. Talatahari, A.H. Alavi, Firefly algorithm with chaos, Commun. Nonlinear Sci. Numer. Simul. 18 (1) (2013) 89–98.
[59] S.H. Yu, S.L. Zhu, Y. Ma, D.M. Mao, A variable step size firefly algorithm for numerical optimization, Appl. Math. Comput. 263 (2015) 214–220.
[60] H. Wang, W.J. Wang, H. Sun, S. Rahnamayan, Firefly algorithm with random attraction, Int. J. Bio-Inspired Comput. 8 (1) (2016) 33–41.
[61] L. He, S. Huang, Modified firefly algorithm based multilevel thresholding for color image segmentation, Neurocomputing 240 (2017) 152–174.
[62] H. Wang, W. Wang, L. Cui, H. Sun, J. Zhao, Y. Wang, Y. Xue, A hybrid multi-objective firefly algorithm for big data optimization, Appl. Soft Comput. 69 (2018) 806–815.
[63] S. Das, S. Maity, B.-Y. Qu, P.N. Suganthan, Real-parameter evolutionary multimodal optimization — A survey of the state-of-the-art, Swarm Evol. Comput. 1 (2) (2011) 71–88.
[64] T. Hassanzadeh, M.R. Meybodi, A new hybrid approach for data clustering using firefly algorithm and K-means, in: The 16th CSI International Symposium on Artificial Intelligence and Signal Processing, AISP 2012, 2012.
[65] S.J. Nanda, G. Panda, A survey on nature inspired metaheuristic algorithms for partitional clustering, Swarm Evol. Comput. 16 (2014) 1–18.
[66] W. Srisukkham, L. Zhang, S.C. Neoh, S. Todryk, C.P. Lim, Intelligent leukaemia diagnosis with bare-bones PSO based feature optimization, Appl. Soft Comput. 56 (2017) 405–419.
[67] K. Sörensen, Metaheuristics—the metaphor exposed, Int. Trans. Oper. Res. 22 (1) (2015) 3–18.
[68] J.H. Holland, Genetic algorithms, Sci. Am. 267 (1) (1992) 66–73.
[69] K. Socha, M. Dorigo, Ant colony optimization for continuous domains, European J. Oper. Res. 185 (3) (2008) 1155–1173.
[70] S. Mirjalili, Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems, Neural Comput. Appl. 27 (4) (2015) 1053–1073.
[71] S. Mirjalili, SCA: A Sine cosine algorithm for solving optimization problems, Knowl.-Based Syst. 96 (2016) 120–133.
[72] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Inf. Process. Manage. 45 (4) (2009) 427–437.
[73] R.D. Labati, V. Piuri, F. Scotti, ALL-IDB: The acute lymphoblastic leukemia image database for image processing, in: Proceedings of the 18th IEEE International Conference on Image Processing, 2011.
[74] C. Blake, C. Merz, UCI Repository of Machine Learning Databases, University of California, Department of Information and Computer Science, Irvine, CA, 1998.
[75] S.C. Neoh, W. Srisukkham, L. Zhang, S. Todryk, B. Greystoke, C.P. Lim, M.A. Hossain, N. Aslam, An intelligent decision support system for leukaemia diagnosis using microscopic blood images, Sci. Rep. 5 (2015) 14938.
[76] L. Putzu, G. Caocci, C. Di Ruberto, Leucocyte classification for leukaemia detection using image processing techniques, Artif. Intell. Med. 62 (3) (2014) 179–191.
[77] N.J. Radcliffe, P.D. Surry, Fitness variance of formae and performance prediction, in: L.D. Whitley, M.D. Vose (Eds.), Foundations of Genetic Algorithms, Elsevier, 1995, pp. 51–72.
[78] D.X. Chang, X.D. Zhang, C.W. Zheng, A genetic algorithm with gene rearrangement for k-means clustering, Pattern Recognit. 42 (7) (2009) 1210–1222.
[79] B. Bullnheimer, R.F. Hartl, C. Strauss, A new rank based version of the ant system - a computational study, CEJOR Cent. Eur. J. Oper. Res. 7 (1999) 25–38.
[80] M. Mavrovouniotis, S.X. Yang, Ant colony optimization with self-adaptive evaporation rate in dynamic environments, in: 2014 IEEE Symposium on Computational Intelligence in Dynamic and Uncertain Environments, CIDUE, 2014.
[81] V.K. Ojha, A. Abraham, V. Snášel, ACO for continuous function optimization: A performance analysis, in: 2014 14th International Conference on Intelligent Systems Design and Applications, 2014.
[82] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, J. Amer. Statist. Assoc. 32 (200) (1937) 675–701.
[83] D.G. Pereira, A. Afonso, F.M. Medeiros, Overview of friedman's test and post-hoc analysis, Comm. Statist. Simulation Comput. 44 (10) (2015) 2636–2653.
[84] S. Michael, L. Ertöz, V. Kumar, The challenges of clustering high dimensional data, in: New Directions in Statistical Physics, Springer, 2004, pp. 273–309.
[85] T.Y. Tan, L. Zhang, S.C. Neoh, C.P. Lim, Intelligent skin Cancer detection using enhanced particle swarm optimization, Knowl.-Based Syst. 158 (2018) 118–135.
[86] L. Ballerini, R.B. Fisher, B. Aldridge, J. Rees, A color and texture based hierarchical k-NN approach to the classification of non-melanoma skin lesions, in: M.E. Celebi, G. Schaefer (Eds.), Color Medical Image Analysis, Springer Netherlands, Dordrecht, 2013, pp. 63–86.
[87] D. Pandit, L. Zhang, S. Chattopadhyay, C.P. Lim, C.Y. Liu, A scattering and repulsive swarm intelligence algorithm for solving global optimization problems, Knowl.-Based Syst. 156 (2018) 12–42.
[88] L. Zhang, K. Mistry, S.C. Neoh, C.P. Lim, Intelligent facial emotion recognition using moth-firefly optimization, Knowl.-Based Syst. 111 (2016) 248–267.
[89] T. Tan, L. Zhang, C.P. Lim, B. Fielding, Y. Yu, E. Anderson, Evolving ensemble models for image segmentation using enhanced particle swarm optimization, IEEE Access 7 (2019) 34004–34019.
[90] B. Fielding, L. Zhang, Evolving image classification architectures with enhanced particle swarm optimisation, IEEE Access 6 (2018) 68560–68575.

Hailun Xie received his B.Sc. in Building Environment and Energy Engineering, and M.Sc. in Building Energy Conservation at Shenyang Jianzhu University in China. He is currently a Ph.D. student in the Department of Computer and Information Sciences, Northumbria University. His research interests include machine learning, evolutionary computation and intelligent robotics.



Li Zhang is currently an Associate Professor (Reader) in Computer Science in Northumbria University, UK. She holds expertise in deep learning, machine learning, intelligent robotics and evolutionary computation. She gained her Ph.D. and postdoctoral experience from University of Birmingham previously. Dr. Zhang is an Associate Editor for Decision Support Systems.

Chee Peng Lim received the Ph.D. degree in intelligent systems from the University of Sheffield, Sheffield, U.K., in 1997. He is currently Professor of Computational Intelligence with the Institute for Intelligent Systems Research and Innovation, Deakin University, Australia. His research interests include computational intelligence-based systems for data analytics, condition monitoring, optimization, and decision support. He has authored/co-authored over 450 technical papers in journals, conference proceedings, and books.

Yonghong Yu is an Associate Professor at TongDa College, Nanjing University of Posts and Telecommunications, in China. He received the B.Sc. and M.Sc. degrees in Computer Science from Wuhan University, and Ph.D. in Computer Application Technology from Nanjing University. His main research interests include machine learning, recommender systems and social networks applications.

Chengyu Liu received the B.S. and Ph.D. degrees in biomedical engineering from Shandong University, China, in 2005 and 2010, respectively. He has completed the post-doctoral trainings at Shandong University, China, from 2010 to 2013, Newcastle University, U.K., from 2013 to 2014, and Emory University, USA, from 2015 to 2017. He is currently a Professor with the Southeast University, China. He is also the Director of the Southeast-Lenovo Wearable Heart-Sleep-Emotion Intelligent Monitoring Lab. He was the PI on over 10 awarded grants. He has published over 130 papers. His research topics include intelligent monitoring, and machine learning for cardiovascular signals.

Han Liu is currently a Research Associate in Data Science in the School of Computer Science and Informatics at the Cardiff University. He has previously been a Research Associate in Computational Intelligence in the School of Computing at the University of Portsmouth. He received a B.Sc. in Computing from University of Portsmouth in 2011, an M.Sc. in Software Engineering from University of Southampton in 2012, and a Ph.D. in Machine Learning from University of Portsmouth in 2015. His research interests include data mining, machine learning, rule based systems, intelligent systems, fuzzy systems, big data, granular computing and computational intelligence.

Julie Walters is currently an Associate Head of Department in the Department of Computer and Information Sciences, Faculty of Engineering and Environment, Northumbria University. Her research interests include machine learning, natural language processing and data mining.