Open Access **Baghdad Science Journal** **P-ISSN: 2078-8665**
Published Online First: September 2021 **2022, 19(2): 409-421** **E-ISSN: 2411-7986**

DOI: http://dx.doi.org/10.21123/bsj.2022.19.2.0409

Improved Firefly Algorithm with Variable Neighborhood Search for Data Clustering

Hayder Naser Khraibet Al-Behadili

Computer Science Department, Shatt Al-Arab University College, Basra, Iraq
Email address: hayderkhraibet@sa-uc.edu.iq
ORCID ID: https://orcid.org/0000-0002-5288-4923

Received 6/10/2020, Accepted 14/2/2021, Published Online First 20/9/2021, Published 1/4/2022

Abstract:

Among the metaheuristic algorithms, population-based algorithms are an explorative search algorithm superior to the local search algorithm in terms of exploring the search space to find globally optimal solutions. However, the primary downside of such algorithms is their low exploitative capability, which prevents the expansion of the search space neighborhood for more optimal solutions. The firefly algorithm (FA) is a population-based algorithm that has been widely used in clustering problems. However, FA is limited in terms of its premature convergence when no neighborhood search strategies are employed to improve the quality of clustering solutions in the neighborhood region and exploring the global regions in the search space. On these bases, this work aims to improve FA using variable neighborhood search (VNS) as a local search method, providing VNS the benefit of the trade-off between the exploration and exploitation abilities. The proposed FA-VNS allows fireflies to improve the clustering solutions with the ability to enhance the clustering solutions and maintain the diversity of the clustering solutions during the search process using the perturbation operators of VNS. To evaluate the performance of the algorithm, eight benchmark datasets are utilized with four well-known clustering algorithms. The comparison according to the internal and external evaluation metrics indicates that the proposed FA-VNS can produce more compact clustering solutions than the well-known clustering algorithms.

Keywords: Data clustering, Data mining, Firefly algorithm, Machine learning, Variable neighborhood search.

Introduction:

Data clustering has a root in a number of fields including statistics, bioinformatics, machine learning exploratory data analysis, image segmentation, security, medical image analysis, web handling, and mathematical programming. Its role is to transform data into clusters with high similarity in a common cluster and high dissimilarity in different clusters [1-4]. Clustering can be classified as partitional and hierarchical clustering [5]. The former can be represented as center-, density-, grid-, and model-based clustering, whereas the latter produces a hierarchical tree different from that former which produces spherical groups. In hierarchical clustering, the clustering performs by using ether agglomerative or divisive hierarchic approach. The latter constructs a tree by dividing the dataset into sub-clusters recursively until each data item represents a single cluster. By contrast, in the former, clustering is conducted by computing the similarity among clusters and then combining the two most similar clusters until a single cluster is observed. The advantage of hierarchical clustering is that this approach is suitable for text clustering because it reviews the data as a hierarchy of nested quality clusters. However, this approach is not suitable for big data compared with the partitional clustering that requires lower computational complexity and observed speed.

Center-based clustering is one of the main clustering approaches and uses the central concept to represent the center of clusters. Each cluster has a unique center that represents the minimum intra-clustering distance between the centroid and all members of the cluster [6]. In clustering, the center concept can be represented as an object of the data, which is known as medoid-based clustering, or the

mean of the objects located in the search space of the data, which is known as centroid-based clustering [7]. Centroid-based clustering can be represented as the mean of the objects in a cluster, where each object in the cluster has a minimum distance to the centroid compared with the other cluster centroids in the search space. The minimum intra-clustering within each cluster shows a good clustering quality, which becomes more difficult to obtain when the number of clusters is increased [8]. In addition, the minimum intra-clustering is considered a NP-hard problem when more than three centroids are involved [9,10]. Several distance-based algorithms have been adopted to assign objects to appropriate clusters, such as identifying disease using K-means and artificial neural network (ANN) [11,12], fuzzy C-means, and multi K-means [13]. Nonetheless, finding the best initial clustering centroid and avoiding becoming stuck at the local optima are the challenges of the traditional algorithms [14]. An unsupervised approach using clustering can identify several diseases, which is promising in diagnosing strange diseases or incomprehensible behavior when no enough information is available [11,12]. For instance, in finding the abnormality of a brain tumor, K-means can be used to improve the image and mark the districts in view of their texture feature and ANN to choose the correct object in view of the training of ANN.

Several studies have classified the use of algorithms in solving the clustering problem as distance-based algorithms, including K-means, fuzzy C-means, multi K-means, and other local search algorithms, such as tabu search [15], and population-based algorithms, such as artificial bee colony [16], gray wolves algorithm [17], and firefly algorithm (FA) [18].

Recently, different optimization algorithms have been published to minimize the intra-clustering distance within each cluster, such as iterative simulated annealing [19], randomized local search algorithm [20], and the adaptive acceptance criterion algorithm for optimization problems [21]. Nonetheless, the above algorithms are local search algorithms that intensify only the search process in the neighborhood of the clustering solutions. Thus, the algorithms are limited, and the exploration of the search space for more promising clustering solutions is weak [22]. On the contrary, population-based algorithms have a high exploration capability and are only limited to exploit the neighborhood search space to improve clustering solutions [23]. FA has been incorporating other algorithms as a hybrid algorithm for different optimization algorithms, such as integration into two different clustering techniques, one with K-means and one with K-harmonic [24,25], and other research, such as a hybrid model of FA and fuzzy c-means (FCM). However, the integration is still limited because the initial centroid of the FA algorithm mainly depends on another algorithm to quickly converge to a local optimum.

On these bases, this study enhances the performance of the firefly clustering algorithm by incorporating variable neighborhood search (VNS) [26] as a local search method to overcome their limitations in providing solutions to clustering problems, thereby exhibiting a promising performance in different application domains [27,28]. The FA mainly depends on the initial selection, which causes premature convergence when no neighborhood search strategies are employed to improve the quality of the clustering solutions in the neighborhood region [27] and explore the global regions in the search space [29]. VNS can enhance exploitation capability by improving clustering solutions during the algorithm process. VNS can also enhance the exploration capability using the perturbation operators, thereby avoiding the known premature convergence of the FA [30] and getting stuck at the local optima in the advance stages of the search process [31]. The contribution of VNS is by performing as a local search method with four different operations, which plays an important role in the trade-off between the exploration and exploitation abilities. Indeed, four different operations mean different neighborhoods, and thus, different landscapes can be generated. The concept of VNS with local search generates different local optima, which is local optima for a given neighborhood. Using VNS will enhance the learning process of the FA, which begins with the exploration of the search space. By contrast, research on an optimal clustering solution in the search space of the best clustering solution found during the research process is intensified.

The rest of this paper is organized as follows. Section 2 discusses the proposed FA-VNS clustering algorithm. Section 3 shows the benchmark and the evaluation performance, whereas Section 4 presents the results. Finally, Section 5 concludes this research and presents the future research direction.

Proposed FA-VNS:

FA is a population-based algorithm inspired by the nature of fireflies, simulating their flash pattern and characteristics. Owing to its simplicity and the good results obtained in the optimization problem compared with other swarm intelligence algorithms, researchers have applied FA to different optimization problems for several topics in data

mining, including speech recognition, image segmentation, and feature selection. The flash of a firefly is a bioluminescence operator produce by each firefly as a light to attract other fireflies and used for prey. The main purpose of a light flash is threefold, namely, to attract each other, be attractive to less bright ones, and represent the fitness function on the search space. Meanwhile, the two important issues in the main idea of the algorithm include light intensity and attractiveness. The light density reflects the objective function of a particular location, whereas attractiveness is a variable value that changes according to the distance between two fireflies. Figure 1 depicts the flowchart of the standard FA. As showing in Fig. 1, the algorithm starts by initializing all parameters as step 1. In the next step, FA evaluates all fireflies based on the objective function and then ranks them as shown in step 3. In step 4, the algorithm finds the best firefly and compares it with others in the colony, where the firefly with fewer attractiveness moves from its location to another better location as shown in step 5. The algorithm will conduct the above process until all iterations are complete, and then, the algorithm will print the best result produced by the best firefly. The algorithm utilized for unsupervised clustering includes partitional and hierarchical clustering. The FA is used to produce the optimal number of clusters and corresponding optimal centers. The optimal centers minimize the intra-clustering distance between each cluster center and each item in the same cluster. However, the standard FA is limited by its premature convergence when no neighborhood search strategies are employed to improve the quality of clustering solutions in the neighborhood region and explore the global regions in the search space.

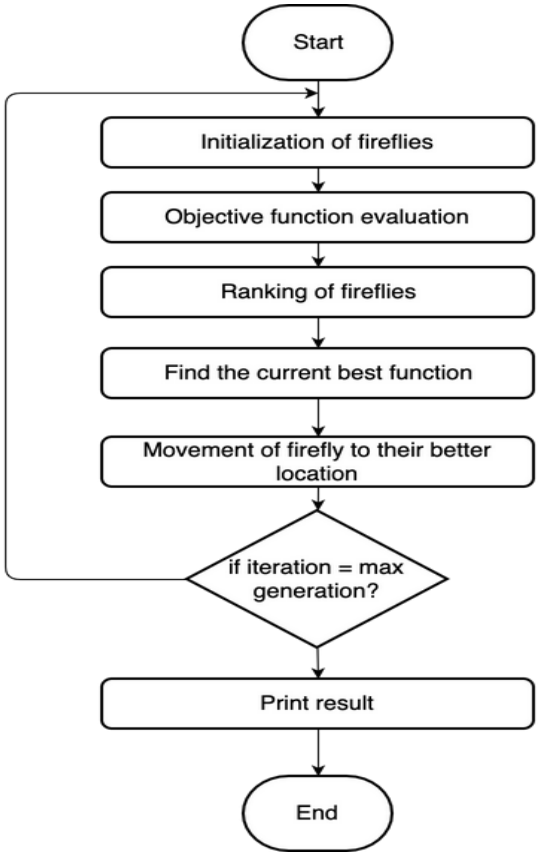

Figure 1. Flowchart of FA.

This study improves the FA by incorporating VNS in each iteration of the algorithm. Therefore, the clustering solution is also improved by using the perturbation operators of VNS to change some parts of the solution with the addition of the ability to find more centroids during the algorithm process. The modification performed based on predefined parameter q_0 was initialized to 0.98 statically, which represents the probability of selecting the current iteration best solution (i_{bs}) for enhancement. In each iteration, a random value is generated [0,1] and compared with the value of q_0; when the values are equal or the random value is greater than q_0, the local search statues become active, and one of the VNS operations is performed in the greedy concept according to a random number that represents one of the VNS operations, such as pair-swap, inversion, insertion, and displacement. The operations are discussed in detail as shown in Fig. 2.

- Pair-swap: Two positions are selected randomly from clustering solution i_{bs} and swapped.
- Inversion: Two positions from the clustering solution are swapped, and the subsequence between the positions is inverted.
- Insertion: Two positions are selected randomly from the clustering solution, and

the front position is inserted before the back position.

- Displacement: A single random position is selected from the clustering solution with a random subsequence of positions, and the selected subsequence is inserted before the selected single position.

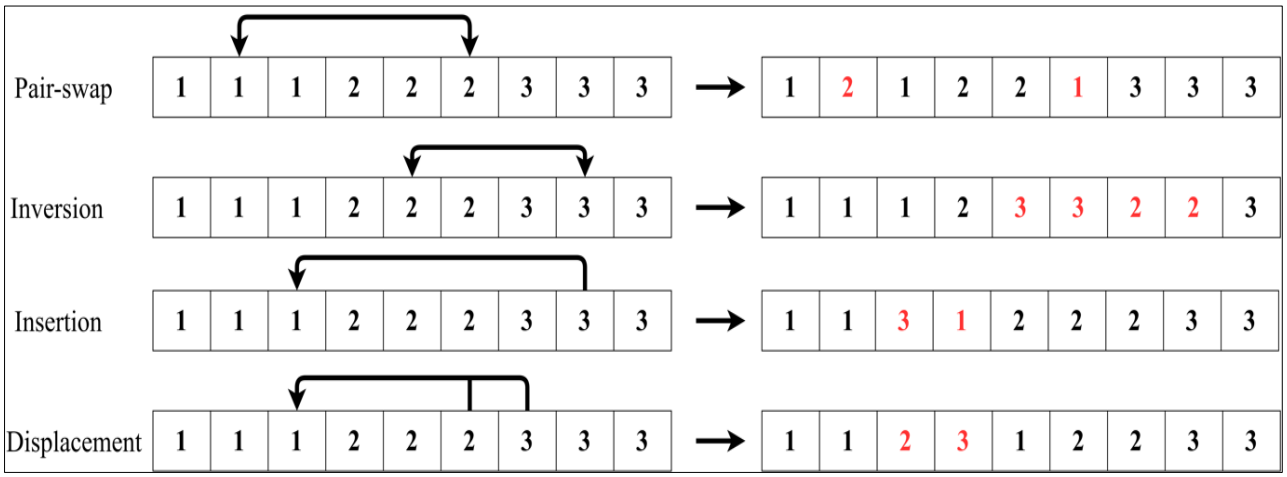

Figure 2. Main operations in the VNS local search algorithm.

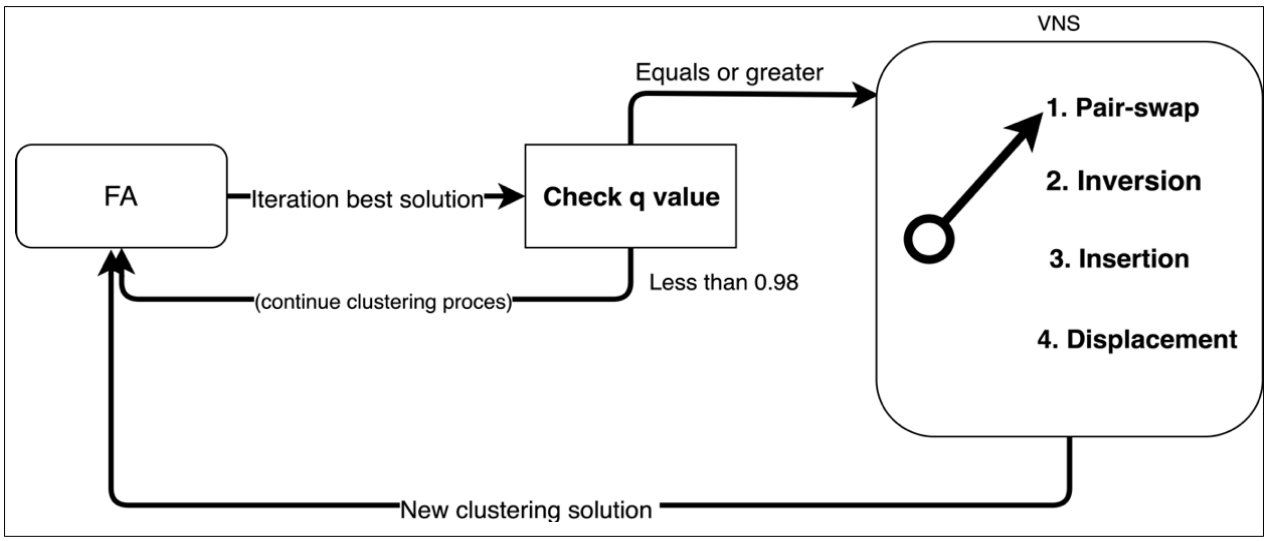

Figure 3. Process flow of the FA-VNS algorithm.

As shown in Fig. 3, the best iteration solution is investigated to improve in a greedy manner to ensure the improvement of the solution using one of the VNS operations. The new clustering solution contains a simple modification of the position (more exploitation), such as pair-swap, or high exploration, such as the use of the displacement operation, which can exploit the benefit in the trade-off between the exploration and exploitation abilities.

The algorithm generates initial centroids randomly for each agent, where the number of centroids is statically predefined [32], such as the example in Table 1 of three centroids carried by a single agent.

Table 1. Example of Centroids Carried by an Agent.

Centroid No.	Attribute 1	Attribute 2	Attribute 3
Centroid 1	3.20	0.53	1.22
Centroid 2	1.23	3.45	1.43
Centroid 3	4.33	3.11	1.12

Each agent attempts to optimize the fitness function by finding the minimum intra-clustering within each cluster, which represents the optimal centroids. The algorithm begins the clustering task by sorting the fireflies according to their fitness function. The two main factors of the FA are the

light intensity (I) and the attractiveness (β), which respectively represents the fitness function and the solution improvement during the algorithm process according to the brightness between the agents and the distance among agents. The improvement relative to the movement of the agents is calculated using Equation (1), where the agent has less bright moves toward the agent with a high density of the brightness, which is calculated using Equation (1).

$$\beta = \beta_0 exp^{(-yr_{ij^2})}, \quad (1)$$

where β_0 represents the initial attractiveness, y is the light absorption, which is usually initialized to 1, and r_{ij^2} represents the Euclidean distance between two agent positions (i, j) on the graph, which represents the distance between the two centroids calculated using Equations (2) and (3).

$$\text{Euclidean distance} = \sqrt{(i-j)^2}, \quad (2)$$

$$\mathrm{x}_{i\,(new)} = \mathrm{x}_{i\,(current)} + \beta * \left(\mathrm{x}_{j\,(current)} - \mathrm{x}_{i\,(current)}\right), \quad (3)$$

where $\mathrm{x}_{i\,(new)}$ is the new position of the agent, $\mathrm{x}_{i\,(current)}$ is the position of current agent i, and $\mathrm{x}_{j\,(current)}$ is the position of current agent j on the search space. In each iteration, the agent with less brightness is treated to improve its centroids by moving it to another position on the search space. The integration of the VNS procedure improves the solution quality as a local search and avoids premature convergence. Combining the FA and VNS improves the solution quality by maintaining the balance between exploration and exploitation. The tiny permutation used in VNS easily improves the neighborhood structure during the search process, moving the centroids of the agent to another position in the neighborhood. Similarly, the large number of permutations enhances the exploration ability, moving the agent centroids to another promising region in the search space. The iteration's best clustering solution is improved using one of the VNS operations according to the q value as shown in Fig. 2 and calculated using Equation (3). If $q > q_0$, then S (VNS activated) and one of the operations are selected according to the value of v, which is randomly generated [1,4] as shown in Equations (4) and (5).

$$p = \begin{cases} S, & if\ q > q_0; \\ VNS\ not\ active, & otherwise. \end{cases} \quad (4)$$

$$s = \begin{cases} \text{Pair} - \text{swap}, & if\ v = 1; \\ \text{Inversion}, & if\ v = 2; \\ \text{Insertion}, & if\ v = 3; \\ \text{Displacement}, & if\ v = 4; \end{cases} \quad (5)$$

The purpose of $q > q_0$ is to improve the solution according to predefined value q_0 set by the user. Those values guide the search toward the objective function following the quality of the clustering solution or toward the neighborhood search, which improves the clustering solution in a stochastic manner. If the q value generated randomly is greater than q_0, then the algorithm selects the iteration's best clustering solution to be continued without any improvement in its local region; otherwise, the iteration's best clustering solution will be a subjected of modification using one of the four operations. The choice of the operation is based on the value of v, which is a random value generated in ranges [1,4] representing one of the available operations, such as pair-swap set by one value, the inversion set by two, and so on.

In each success prepetition, the centroids of the agent are updated to new centroids if accepted, where the process is applied in a greedy manner, that is, only the best improvement is accepted in either exploration or exploitation. The greedy process improves the clustering solution in the advanced stages by accepting only the best solution and exploring the best clustering solution found in another region of the search space. Figure 4 shows the FA-VNS for clustering problems.

Algorithm (1): FA for clustering (FA-VNS)

Initialize all parameters: Number of fireflies N, $Max\_iteration$ and number of clusters K, q_0
Generate random initial centroids for each firefly $x_i, i = \{1,2,3,\dots,N\}$, where N is the number of fireflies
Calculate the fitness $f(x_i)$ for each firefly and update light intensity I_{xi} of each firefly x_i. where $f(x_i)$ is the intra-clustering distance
    Find the best firefly x_i using $f(x_i)$
    while ($iteration < Max\_iteration$) **do**
        for i = 1 **to** N **do** % all fireflies
            for j = 1 **to** N **do** % all fireflies
                if ($I_{xi} > I_{xj}$) **then**
                    (Move x_i towards x_j) % Attractiveness
                end-if
            end-for
        end-for
Perform VNS according to q value and v value % Select one of the operations
Update centroids in each firefly according to their latest positions
Update the best firefly by ranking the fireflies and find the current best
end-while
Print the best result
end-algorithm

Figure 4. Pseudocode of FA-VNS for the clustering algorithm.

Benchmark and Evaluation Performance:

To verify the performance of the proposed FA-VNS, internal and external evaluation matrices are used. The internal and external matrices are part of the clustering evaluation task and represent the compactness and separation of each cluster. The implementation of the code was in Java with the Weka library to perform the evaluation part, including the internal and external criteria. The analysis of the results will be compared with other clustering algorithms to show how the proposed algorithm can produce better clustering accuracy. The internal evaluation matrices are unsupervised methods that use the distance between the clustering and within the clusters to indicate the quality of clustering, such as the intra-clustering distance (intra) and Calinski–Harabasz (CH). The external evaluation matrices are supervised methods, such as entropy and F-measure, that measure the quality of clustering according to known information, such as the class of each instance, to indicate how many correct classes are placed in the same cluster.

The minimum intra-clustering distance (intra) and the CH metric [33] are the internal evaluation matrices, whereas the entropy and F-measure metrics are the external evaluation matrices used in the research. The intra-clustering distance shown in Equation (6) is the summation of the distance between the cluster centroid and the objects of the cluster. A minimum intra-clustering distance indicates that the clusters have good compactness and are well-separated from each other. k is the number of clusters, nc is the number of objects in cluster, C_i is the centroid of a cluster, and O_j is an object belonging to C_i.

$$Intra = \sum_{i=1}^{k} \sum_{j=1}^{nc} (C_i - O_j). \qquad (6)$$

The CH metric represents the ratio of the sum of the secured error between each cluster B to the within-clustering sum of secured error W, and n is the number of objects in the dataset. The maximum value of CH reflects a high quality of clustering solution where the ratio of B to W is high. The CH metric can be calculated using Equation (7).

$$CH = \frac{\left(\frac{B}{(k-1)}\right)}{\left(\frac{W}{(n-1)}\right)}. \qquad (7)$$

Equation (8) measures the entropy for single clustering w. The entropy of each cluster is first measured. Then, the total entropy of all clusters is calculated using Equation (9).

$$H(w) = -\sum_{o \in C} P(w_o) log_2 P(w_o). \qquad (8)$$

$P(w_o)$ represents the probability of object o in cluster C, whereas H(w) is the entropy of a single cluster. Thus, the sum of all cluster entropies is calculated according to Equation (9), which reflects well-distributed objects in their right clusters if the value of the entropy is small [34].

$$H(\Omega) = \sum_{w \in \Omega} H(w) \frac{N_w}{N}. \quad (9)$$

The F-measure metric requires two other supervised external metrics to be calculated, namely, Precision and Recall, which are used to determine the cluster assignment [35]. The two metrics can be calculated using Equations (10) and (11) respectively, where TP is the true positive, FP is the false positive, and FN is the false negative. The metrics are calculated before calculating the F-measure, and if the value of the F-measure is high, then most of the objects are assigned to the same cluster, as shown in Equation (12).

$$Precision = \frac{TP}{TP+FP}, \quad (10)$$

$$Recall = \frac{TP}{TP+FN}, \quad (11)$$

$$\text{F} - \text{ Measure} = \frac{2*Precision*Recall}{Precision+Recall}. \quad (12)$$

The benchmark used in the research was extracted from UCI and contains eight datasets, which are popularly used for classification and clustering tasks. Table 2 shows the datasets, which cover different application domains, such as life and physical, with different instance sizes, such as small, medium, large, and very large. A comparison was performed against well-known algorithms, including centroid FA (C-FA) [32], genetic algorithm (GA) [36,37], simulated annealing algorithm (SA) [38], and K-means (KM) [39]. The KM is a static algorithm, and its iterations and maximum runs are respectively set to 1,000 and 50 because the algorithm is easily trapped at local optima. Table 3 shows the parameter setting of the above-mentioned algorithm. These parameters are set according to the literature of the clustering and the best known for all algorithms [40].

Table 2. UCI Benchmark Datasets for Clustering Task.

Dataset name	Attribute characteristics	Instance number	Area
Obesity	Multivariate Integer	2,111	Life
Segment	Multivariate Real	2,310	N/A
Hepatitis C Virus	Integer, Real	1,385	Life
Vehicle	Multivariate Real	846	Life
Ecoli	Multivariate Real	336	Life
Glass	Multivariate Real	214	Physical
Contraceptive method choice	Multivariate Categorical, Integer	1,473	Life
Mammographic	Multivariate Integer	961	Life

Table 3. Parameters of Algorithms.

SA	GA	C-FA / FA-VNS	KM
Probability threshold = 0.98	Population size = 50		
Initial temperature = 5	Crossover = 0.8	Initial attractiveness (B_0) = 0	
Final temperature = 0.01			
Temperature multiplier = 0.98	Mutation rate = 0.001	Light absorption (y) = 1.0	
Iterations 1,000/Max Run 10			Max Run = 50

Results:

Experiential Results:

Tables 4–7 show the comparisons between the algorithms using internal and external evaluation metrics. As shown in Table 4, the comparison based on the minimum intra-clustering distance (overall performance) indicates that FA-VNS produced the best results in seven datasets, which is approximately 88% better than the other algorithms. The comparison (algorithm vs. algorithm) indicates that the FA-VNS is better than the SA, GA, and KM in all datasets (100%). The comparison between the FA and FA-VNS shows that FA-VNS produced the best result in seven datasets (approximately 88%), including obesity, segment, hepatitis C virus, vehicle, Ecoli, and mammographic. However, FA produced the best result in only one dataset, namely, the contraceptive method choice.

Table 4. Average Results of Intra-cluster Distance for All Clustering Algorithms.

Dataset	SA	GA	C-FA	FA-VNS	KM
Obesity	35,362.6275	27,715.2199	12,901.9814	12,855.6326	13,607.0667
Segment	246,926.4910	218,774.7740	148,842.9455	148,828.2497	151,277.6737
Hepatitis C Virus	80,236.6915	76,201.3042	73,917.1591	73,829.3409	75,710.9380
Vehicle	62,711.0780	50,757.7549	45,182.2347	45,092.9024	46,715.0926
Ecoli	70.4205	69.2617	68.8236	67.8215	69.2823
Glass	221.4013	227.0986	257.8070	213.1613	226.0174
Contraceptive method choice	7,766.9707	6,255.1688	5,541.0280	5,541.5492	2,761.8924
Mammographic	7,231.3300	7,033.2570	7,029.3091	7,015.4579	7,035.10513

The comparison (overall performance) using the internal CH metric shown in Table 5 indicates that the FA-VNS performed better than the other algorithms. The proposed algorithm generated the best results in five datasets, which is approximately 63% better than the SA, GA, C-FA, and KM. The KM algorithm ranks second, obtaining the best results only in three datasets (approximate 27%). FA-VNS is better than the SA, GA, and FA in all datasets (100%). The comparison between KM and FA-VNS shows that FA-VNS produced the best results in five datasets (approximately 63%), including obesity, segment, hepatitis C virus, contraceptive method choice, and mammographic. However, KM produced the best result only in three datasets, including vehicle, Ecoli, and glass (approximately 27%).

Table 5. Average Results of CH for All Clustering Algorithms.

Dataset	SA	GA	C-FA	FA-VNS	KM
Obesity	260.7551	682.7948	688.6069	695.5994	605.2164
Segment	112.1491	196.1889	563.2220	586.2384	522.9889
Hepatitis C Virus	126.5040	151.0178	172.6397	178.0231	176.0981
Vehicle	1,940.3073	1,549.3331	2,072.2440	2,075.1880	2,150.9243
Ecoli	138.5564	147.8499	159.1714	161.5269	165.2268
Glass	73.6235	68.16510	89.7760	91.6285	99.5874
Contraceptive method choice	1,803.3643	1,826.8024	2,780.9023	2,782.9139	2,761.8924
Mammographic	1,512.8451	1,695.7252	1,700.6901	1,705.4735	1,700.6901

Table 6 shows the comparison between the algorithms using the F-measure metric, which indicates that FA-VNS performed better than the other algorithms in six datasets (75%). The C-FA ranks second, obtaining the best results in two datasets (25%). The comparison (algorithm vs. algorithm) indicates that FA-VNS is better than the SA, GA, and KM in all datasets (100%). The comparison between the C-FA and FA-VNS shows that FA-VNS produced the best results in six datasets (75%), including obesity, segment, hepatitis C virus, glass, and mammographic. However, C-FA produced the best results in two datasets only, namely, Ecoli and contraceptive method choice (approximately 25%).

Table 6. Average Results of F-measure for All Clustering Algorithms.

Dataset	SA	GA	C-FA	FA-VNS	KM
Obesity	0.2260	0.2317	0.2707	0.2889	0.2841
Segment	0.2178	0.2520	0.4765	0.4811	0.4377
Hepatitis C Virus	0.3347	0.3368	0.3797	0.4331	0.4226
Vehicle	0.3067	0.3365	0.3604	0.3691	0.3679
Ecoli	0.4989	0.5300	0.6070	0.6027	0.6021
Glass	0.3905	0.3745	0.4325	0.4375	0.4103
Contraceptive method choice	0.3530	0.3590	0.3666	0.3604	0.3158
Mammographic	0.5599	0.5676	0.56359	0.5730	0.5635

The last comparison is based on the entropy metric, which shows how the objects are assigned to their clusters. Table 7 shows that FA-VNS is better than the other algorithms on five datasets by approximately 63%. The SA, C-FA, and KM produced the best results in only one dataset, and GA did not produce any good results (overall performance). The comparison (algorithm vs. algorithm) indicates that FA-VNS is better than the SA in all datasets (100%) and better than the GA in seven datasets (approximately 88%), including obesity, segment, hepatitis C virus, vehicle, Ecoli, contraceptive method choice, and mammographic. However, the GA produced the best result only in the glass dataset. The comparison between FA-VNS and the C-FA indicates that FA-VNS is better than the C-FA in six datasets (75%), namely, obesity, segment, hepatitis C virus, Ecoli, glass, and mammographic, whereas the C-FA obtained the best results in only two datasets (25%), including vehicle and contraceptive method choice. The comparison between KM and FA-VNS shows that FA-VNS produced the best results in six datasets (75%), including obesity, segment, hepatitis C virus, vehicle, Ecoli, and mammographic, whereas KM produced the best results only in two datasets, namely, glass and contraceptive method choice (approximately 25%).

Table 7. Average Results of Entropy for All Clustering Algorithms.

Dataset	SA	GA	C-FA	FA-VNS	KM
Obesity	0.98975	0.9723	0.8769	0.8712	0.8744
Segment	2.4803	2.3107	1.3652	1.2052	1.47566
Hepatitis C Virus	0.64284	0.6136	0.5656	0.5484	0.5665
Vehicle	1.8048	1.7188	1.62771	1.6317	1.6440
Ecoli	0.8247	0.8037	0.6570	0.6459	0.7417
Glass	1.3584	1.3812	1.4602	1.4423	1.3878
Contraceptive method choice	1.5138	1.5038	1.4900	1.5031	0.9983
Mammographic	0.9058	0.8938	0.9013	0.9006	0.9018

The experiments shown in Fig. 5 indicate that FA-VNS produced minimum intra-clustering in (88%) of the datasets, which is better than that produced by other algorithms. This finding indicates that the clustering results are more compact to the cluster center and well-separated according to the CH metric. The CH metric shows a high ratio between the clusters (approximately 63%), which is better than that in other algorithms. The results also show that the results of FA-VNS are more accreted according to the F-measure results, which are (approximately 75%) better than those of the other algorithms. The entropy shows the distribution of the objects to the right clusters, capturing the count of similar objects assigned in different clusters. The entropy is approximately 63% better than those of the other algorithms. The results indicate that VNS enhances the algorithm to find better clustering assignments during the algorithm process by finding promising regions in the search space and simultaneously improving the clustering solutions during the search process. Fig. 5 concludes the results of the comparisons according to the internal and external metrics.

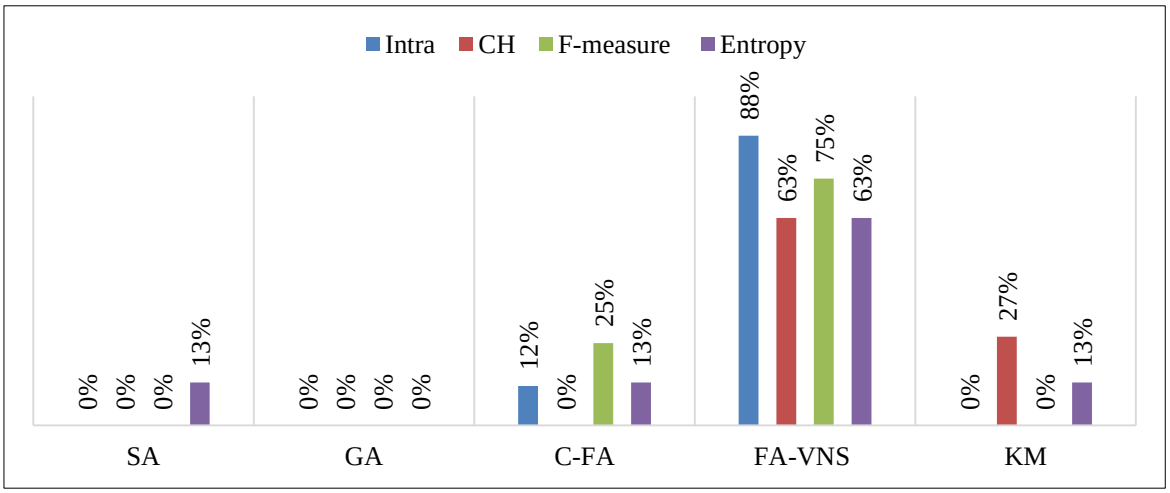

Figure 5. Quality metrics of clustering (internal and external) of FA-VNS vs. best-known algorithms.

Figure 6 shows that the behavior of the two algorithms is C-FA and FA-VNS during the iteration process from 1 to 100. The FA-VNS algorithm starts with high exploration and then moves to a different region using the four operations in VNS. The FA-VNS algorithm modifies the neighbored structure of the best iteration solution to find a better quality of solutions, such as showing in iteration 61, where the algorithm moves the search process to other regions, such as the new region at iteration 67. The C-FA algorithm showing other behavior produced the same results during the algorithm run time. This result means that FA-VNS can effectively explore more regains on the search space, which in the end increases the probability to find deeper regains to have high-quality clustering solutions.

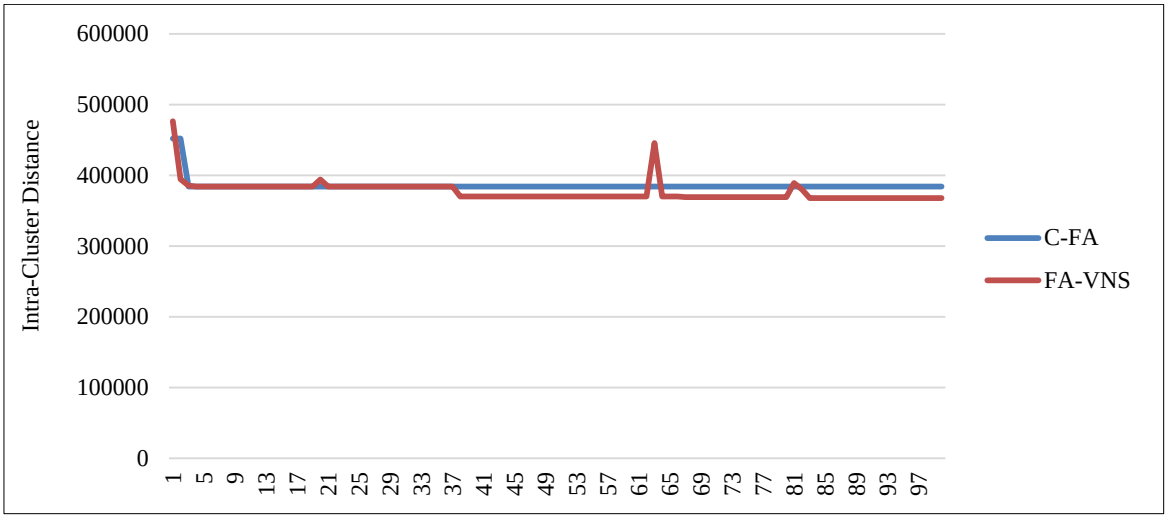

Figure 6. Behavior pattern of C-FA and FA-VNS.

The statistical analysis of paired samples T-test is performed to test the difference in the average mean of the internal and external metrics. The p-value indicates a sufficient difference between two means of the algorithms when the p-value is less than 0.05. Table 8 shows the statistical analysis of the paired samples T-test between FA-VNS and C-FA, which shows evidence to reject the null hypothesis. Most of the p-values are less than 0.05, except for the entropy metric in C-FA, which indicates that no substantial difference exists in the means of both algorithms.

Table 8. P-value between FA-VNS vs. C-FA based on Different Evaluation Criteria.

Algorithm	p-Value
FA-VNS vs. C-FA based on Intra metric	0.0237
FA-VNS vs. C-FA based on CH metric	0.04256
FA-VNS vs. C-FA based on F-measure	0.0381
FA-VNS vs. C-FA based on Entropy	0.2552

Discussion:

One of the interesting facts is meta-heuristic performance is controlled by trade-off between the exploration and exploitation abilities. This process controls the improvement in this aspect has a high quality of solutions, as the search process is treated in sequence stages from time to time. This sequence is adapted to produce high-quality solutions through the exploration process initially, then deeply exploitation through the advanced search process in the neighbourhood region. FA as an optimization algorithm for clustering followed the same procedure of the meta-heuristic algorithms. However, FA is limited in terms of its premature convergence when no neighborhood search strategies are employed to improve the quality of clustering solutions in the neighborhood region and exploring the global regions in the search space. To improve the FA performance, neighborhood search is required, which has a benefit of the trade-off between the exploration and exploitation abilities through the perturbation operators. This research proposed to use VNS as a local search method to maintain the diversity of the clustering solutions using the perturbation operators and improve the quality of clustering solutions in the neighborhood region using the benefit of local search method. Two aspects can be indicated for this improvement. The first aspect is to generate several landscapes with different quality of solutions. Meanwhile, the second aspect generates more diversity of the clustering solutions during the search process. Both aspect aspects control the trade-off between the

neighborhood region for better exploitation and the global regions for better exploration in the search space.

Conclusion:

This study addresses the problem of improving the clustering solution through FA by using VNS as the local search method (i.e., FA-VNS). The improvement is achieved by intensifying the search process during the algorithm process and moving the search process using the VNS premutation to find more promising regions in the search space. VNS with several operations, such as pair-swap, inversion, insertion, and displacement, provide different neighborhoods. Different neighborhoods generate several landscapes and support the algorithm to find more clustering solutions and avoid being stuck at local optima. Therefore, the result of the performance of FA-VNS has been compared with well-known clustering algorithms on UCI Machine Learning Repository datasets. The proposed algorithm produces a better clustering solution than the other clustering algorithms using internal and external evaluation metrics. The reason is that the learning process of the proposed FA-VNS algorithm that can find a promising region on the search space increased as the algorithm begins with the high exploration looking for global regions. By contrast, the search toward an optimal clustering solution in the search space of the best clustering solution found during the search process is intensified. The premature convergence of FA pushed the author of this research to contribute to the utilization of neighborhood search strategies of VNS to improve the quality of clustering solutions by finding global regions in the search space and avoiding a local optima problem.

Furthermore, the advantage of using VNS with the FA is twofold. The first aspect is to generate several landscapes with different quality clustering solutions while making more improvements to the local search to find deeper local regions. Meanwhile, the second aspect is to maintain more diversity of the clustering solutions during the search process according to the perturbation in VNS, which allows a high probability to improve the quality of clustering solutions in the neighborhood region and explore the global regions in the search space.

The proposed FA-VNS algorithm has produced better clustering results compared with other algorithms in terms of internal and external evaluation criteria. However, the FA-VNS algorithm still has a limitation in some parts. For instance, the choice of operations is based on a randomly generated number and does not provide efficient information on the best operation for a particular dataset. Another limitation is the time complexity where the operations in the neighborhoods require more time to find more solutions. Furthermore, the algorithm cannot find the right number of clusters that is required by users as a predefined parameter.

Future research should focus on evaluating the proposed algorithm on other datasets using other evaluation criteria. An online parameter adaption is used to optimize the parameter in VNS to select the best operation for a particular dataset, including self-adaptive strategy, adaptive strategy, and search-based strategy. Other suggestions for future research exploring other search methods include guided and iterated local search with additional comparisons to find its effect on the performance of the algorithm in terms of accuracy and time complexity.

Author's declaration:

- Conflicts of Interest: None.
- I hereby confirm that all the Figures and Tables in the manuscript are mine. Besides, the Figures and images, which are not mine, have been given the permission for re-publication attached with the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee in Shatt Al-Arab University Collage.

References:

1. Al-behadili HNK, Ku-Mahamud KR, Sagban R. Hybrid Ant Colony Optimization and Iterated Local Search for Rules-Based Classification. J Theor Appl Inf Technol. 2020;98(04):657–71.
2. Kuo RJ, Zulvia FE. An improved differential evolution with cluster decomposition algorithm for automatic clustering. Soft Comput. 2019;23(18):8957–73.
3. Xu D, Tian Y. A Comprehensive Survey of Clustering Algorithms. Ann Data Sci. 2015;2(2):165–93.
4. Mandala SR, Kumara SRT, Rao CR, Albert R. Clustering social networks using ant colony optimization. Oper Res. 2013;13(1):47–65.
5. Jabbar AM, Ku-Mahamud KR, Sagban R. Modified ACS Centroid Memory for Data Clustering. J Comput Sci. 2019;15(10):1439–49.
6. Gupta A, Datta S, Das S. Fast automatic estimation of the number of clusters from the minimum inter-center distance for k-means clustering. Pattern Recognit Lett. 2018;116(September):72–9.
7. Budiaji W. Medoid-based shadow value validation and visualization. Int J Adv Intell Informatics. 2019;5(2):76–88.
8. Tung AKH, Han J, Lakshmanan LVS, Ng RT.

Constraint-based clustering in large databases. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2001.
9. Panasenko A, Khandeev V. Exact Algorithm for the One-Dimensional Quadratic Euclidean Cardinality-Weighted 2-Clustering with Given Center Problem. Int Conf Math Optim Theory Oper Res. 2020;30–5.
10. Davidson I, Gourru A, Ravi S, Davidson I, Gourru A, The SR, et al. The Cluster Description Problem - Complexity Results , Formulations and Approximations To cite this version : HAL Id : hal-02060574 The Cluster Description Problem - Complexity Results , Formulations and Approximations. 2019;
11. Arunkumar N, Mohammed MA, Abd Ghani MK, Ibrahim DA, Abdulhay E, Ramirez-Gonzalez G, et al. K-Means clustering and neural network for object detecting and identifying abnormality of brain tumor. Soft Comput. 2019;23(19):9083–96.
12. Obaid OI, Mohammed MA, Abd Ghani MK, Mostafa SA, Al-Dhief FT. Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer. Int J Eng Technol. 2018;7(4.36 Special Issue 36):160–6.
13. Reddy TN. Optimization of K-Means Algorithm: Ant Colony Optimization. In: International Conference on Computing Methodologies and Communication (ICCMC). 2017. p. 530–5.
14. Maghawry AM, Omar Y, Badr A. Initial Centroid Selection Optimization for K-Means with Genetic Algorithm to Enhance Clustering of Transcribed Arabic Broadcast News Documents. In: Silhavy R, Silhavy P, Prokopova Z, editors. Applied Computational Intelligence and Mathematical Methods. Cham: Springer International Publishing; 2018. p. 86–101.
15. Ghany KKA, AbdelAziz AM, Soliman THA, Sewisy AAEM. A hybrid modified step Whale Optimization Algorithm with Tabu Search for data clustering. J King Saud Univ - Comput Inf Sci. 2020;(February).
16. Zulvia RJKFE. Automatic clustering using an improved artificial bee colony optimization for customer segmentation. Knowl Inf Syst. 2018;(43).
17. Kumar V, Chhabra JK, Kumar D. Grey Wolf Algorithm-Based Clustering Technique. J Intell Syst. 2017;26(1):153–68.
18. Ezugwu AE. Nature-inspired metaheuristic techniques for automatic clustering: a survey and performance study. Vol. 2, SN Applied Sciences. Springer International Publishing; 2020.
19. Kittaneh R, Abdullah S, Abuhamdah A. terative Simulated Annealing for Medical Clustering Problems. Trends Appl Sci Res. 2012;(7):103–17.
20. Franzin A, Stützle T. Comparison of acceptance criteria in randomized local searches. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2018;10764 LNCS(August):16–29.
21. Abuhamdah A. Adaptive Acceptance Criterion (AAC) algorithm for optimization problems. J Comput Sci. 2015;11(4):675–91.
22. Zhao F, He X, Yang G, Ma W, Zhang C, Song H. A hybrid iterated local search algorithm with adaptive perturbation mechanism by success-history based parameter adaptation for differential evolution (SHADE). Eng Optim. 2020;52(3):367–83.
23. Ayvaz D, Topcuoglu H, Gurgen F. Performance evaluation of evolutionary heuristics in dynamic environments. Appl Intell. 2012 Jul 1;37(1):130-44.
24. Xie H, Zhang L, Lim CP, Yu Y, Liu C, Liu H, et al. Improving K-means clustering with enhanced Firefly Algorithms. Appl Soft Comput J. 2019;
25. Abshouri AA, Bakhtiary A. A new clustering method based on Firefly and KHM. J Commun Comput. 2012;9:387–91.
26. Hansen P, Mladenović N. Variable neighborhood search. In: Handbook of Heuristics. 2018.
27. Wang H, Cui Z, Sun H, Rahnamayan S, Yang XS. Randomly attracted firefly algorithm with neighborhood search and dynamic parameter adjustment mechanism. Soft Comput. 2017;21(18):5325–39.
28. Huang KW, Girsang AS, Wu ZX, Chuang YW. A hybrid crow search algorithm for solving permutation flow shop scheduling problems. Appl Sci. 2019;9(7).
29. Jensen JH. A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space. Chem Sci. 2019;10(12):3567–72.
30. Yu S, Yang S, Su S. Self-adaptive step firefly algorithm. J Appl Math. 2013 Nov;2013.
31. Fan C, Fu Q, Long G, Xing Q. Hybrid artificial bee colony algorithm with variable neighborhood search and memory mechanism. J Syst Eng Electron. 2018;29(2):405–14.
32. Tang R, Fong S, Yang XS, Deb S. Integrating nature-inspired optimization algorithms to K-means clustering. In: 7th International Conference on Digital Information Management, ICDIM 2012. 2012.
33. Gao X, Wu S. CUBOS: An Internal Cluster Validity Index for Categorical Data. Teh Vjesn - Tech Gaz. 2019;26(2):486–94.
34. Lubis MDS, Mawengkang H, Suwilo S. Performance Analysis of Entropy Methods on K Means in Clustering Process. J Phys Conf Ser. 2017;930(1).
35. Nizam T, Hassan SI. Exemplifying the effects of distance metrics on clustering techniques: F-measure, accuracy and efficiency. In: Proceedings of the 7th International Conference on Computing for Sustainable Global Development, INDIACom 2020. 2020.
36. Zabihi F, Nasiri B. A Novel History-driven Artificial Bee Colony Algorithm for Data Clustering. Appl Soft Comput J. 2018;71:226–41.
37. Maulik U, Bandyopadhyay S. Genetic algorithm-based clustering technique. Pattern Recognit. 2000;33(9):1455–65.
38. Selim SZ, Alsultan K. A simulated annealing algorithm for the clustering problem. Pattern Recognit. 1991;24(10):1003–8.
39. Das P, Das DK, Dey S. A modified Bee Colony Optimization (MBCO) and its hybridization with k-means for an application to data clustering. Appl Soft

Comput J. 2018;70:590–603.
40. Niknam T, Amiri B. An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis. Appl Soft Comput J. 2010;10(1):183–97.

تحسين خوارزمية اليراعة باستخدام البحث المتغير المحلي في الجوار لتجميع البيانات

حيدر ناصر خريبط البهادلي

قسم علوم الحاسبات، كلية شط العرب الجامعة، البصرة، العراق.

الخلاصة:

من بين الخوارزميات الأدلة العليا (الميتاهيورستك)، تعد الخوارزميات القائمة على البحوث المتعددة (المجتمع) خوارزمية بحث استكشافية متفوقة كخوارزمية البحث المحلية من حيث استكشاف مساحة البحث للعثور على الحلول المثلى العالمية. ومع ذلك، فإن الجانب السلبي الأساسي للخوارزميات القائمة على البحوث المتعددة (المجتمع) هو قدرتها الاستغلالية المنخفضة، مما يمنع توسع منطقة البحث عن الحلول المثلى. خوارزمية اليَرَاعَة المضيئة (Firefly (FA هي خوارزمية تعتمد على المجتمع والتي تم استخدامها على نطاق واسع في مشاكل التجميع. ومع ذلك، فإن FA مقيد بتقاربها السابق لأوانه عندما لا يتم استخدام استراتيجيات بحث محلي لتحسين جودة حلول المجموعات في منطقة المجاورة واستكشاف المناطق العالمية في مساحة البحث. على هذا الأساس، فإن الهدف من هذا العمل هو تحسين FA باستخدام البحث المتغير في الأحياء (VNS) كطريقة بحث محلية (FA-VNS)، وبالتالي توفير فائدة VNS للمفاضلة بين قدرات الاستكشاف والاستغلال. يسمح FA-VNS المقترح لليراعات بتحسين حلول التجميع مع القدرة على تعزيز حلول التجميع والحفاظ على تنوع حلول التجميع أثناء عملية البحث باستخدام مشغلي الاضطراب في VNS. لتقييم أداء الخوارزمية، يتم استخدام ثماني مجموعات بيانات معيارية مع أربع خوارزميات تجميع معروفة. تشير المقارنة وفقًا لمقاييس التقييم الداخلية والخارجية إلى أن FA-VNS المقترحة يمكن أن تنتج حلول تجميع أكثر إحكاما من خوارزميات التجميع المعروفة.

الكلمات المفتاحية: تجميع البيانات، التنقيب عن البيانات، خوارزمية اليراعة، تعلم الآلة، بحث الجوار المتغير.