the net are considered the fittest young spiders based on this concept. the black widow optimization algorithm is developed according to this concept [1].

Data mining techniques include three components: preference criterion, model, and search algorithm. In data mining techniques, the most common functions are classification, clustering, association rule mining, regression, sequence, dependency modelling and link analysis. Model representation determines the model flexibility for representing the underlying data and the interpretability in human terms. This comprises includes linear and nonlinear models, decision trees and rules, probabilistic graphical dependency models, example-based techniques and relational attribute models [2]. Choosing which model to use for mining is determined by the preference criterion; and this depends on the underlined data set; by associating some measure of goodness with the model functions. It tries to generate a model function with a large number of degrees of freedom. Specification of the search algorithm is defined after the model and the preference criterion are selected [3].

2. Cluster Analysis

Clustering is the process that organizes objects into self-similar groups via discovering the boundaries between these groups algorithmically using a group of different statistical algorithms and methods. Cluster analysis doesn't make any difference between dependent and independent variables. It examines the set of interdependent relationships entirely to detect the similarity relationships between the objects so as to identify the clusters. We can also utilize cluster analysis as a dimension reduction method in which the number of objects are grouped into a set of clusters; and then a reduced set of variables are used for predictive modeling [4].

2.1 Unsupervised clustering

The aim of unsupervised clustering is maximizing the intra-cluster similarity and minimizing the similarity of the intra-cluster, given a similarity/dissimilarity measure. A specific objective function is used: (e.g., a function that minimizes the interclass distances to find tight clusters). It uses a data set which doesn't have a target variable. K-means and hierarchical clustering are the most widely used unsupervised clustering techniques in segmentation [5].

2.1.1 K-means Algorithm

Because of its simplicity and speed, K-means is one of the most widely used clustering techniques. It partitions the data into k clusters through assigning each object to its closest cluster centroid (the mean value of the variables for

all objects in that particular cluster) based on the distance measure used. It is more robust to different types of variables [6].
The basic algorithm for k -means works as in figure 1
1. determine the number of clusters, k.
2. Select centroids for k cluster.
3. After select centroids assign each object to the nearest cluster centroid.
4. Recompute the new cluster centroid.
5. Repeat step three and four until the convergence criterion is met or maximum iteration is reached.

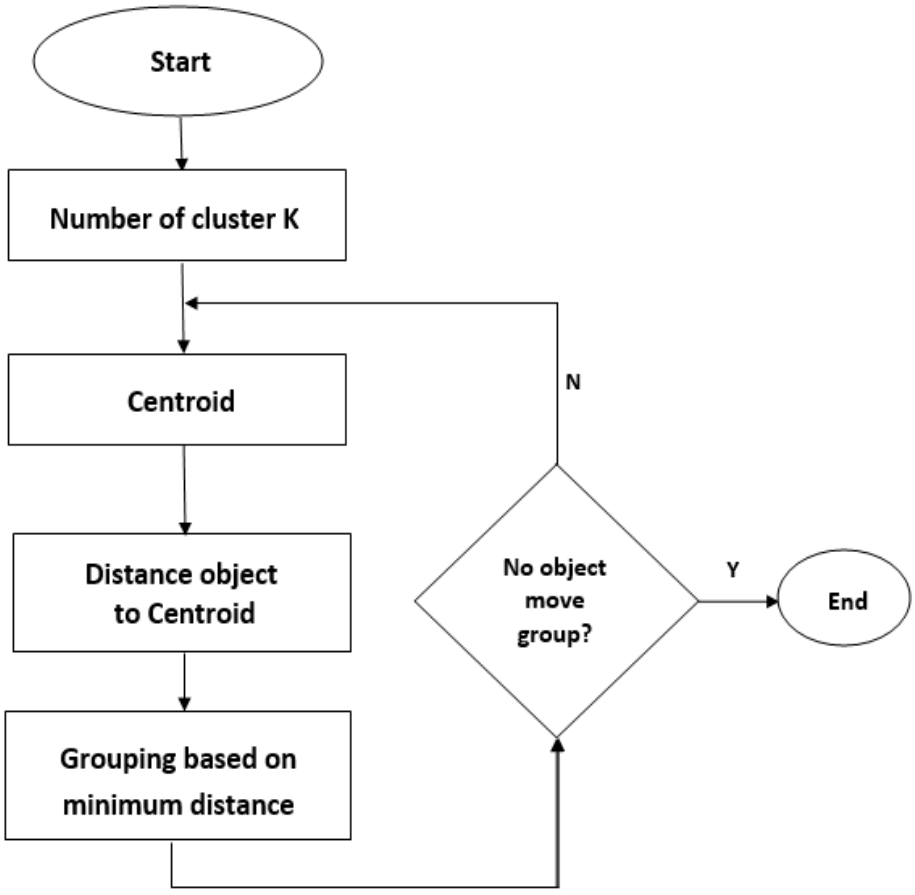

Figure 1: Flowchart of the k-means algorithm [3]

2.1.2 Hierarchical Clustering

Clusters generated by hierarchical clustering are organized into a hierarchical structure. We can use visualizing this hierarchical structure to understand the structure of clusters in the data set as well as the clusters themselves. A measure of similarity between objects is only required for this technique. Your specification of the number of clusters is not required. Obtaining any number of clusters can be done by cutting the hierarchical structure at a proper level [7].

2.2 Supervised clustering

The aim of supervised clustering is identifying clusters that have high probability densities with respect to individual classes (class-uniform clusters). We use it when there is a target variable and a training set including the variables to cluster [8].

3. Literature Review

A Genetic Algorithm (GA) was proposed by Maulik and Bandyopadhyay to select cluster centroids for k-means clustering. A chromosome is considered a string of real numbers where a cluster centroid is represented by each real number. After creating clusters based on the centroids in the chromosome, the clusters determine the new centroids and these newly determined centroids replace the centroids of the chromosome. The sum of Euclidian distances from the centroid to each data point is used as the fitness function. The sum of Euclidian distances is reported as an experimental result [9].
Another GA for k-means clustering was proposed by Lin et al. where data points from the data set are used as the cluster centroids. The chromosome is considered a binary string, which contains k number of 1s and remaining 0s. The data points corresponding to the 1s are the cluster centroids [10]. Soni and Patel use clustered Iris data set [11] using classical k-means and k-medoids methods and reported their clustering accuracies [12]. Wei et el. used k-means method for clustering along with mutual information-based unsupervised feature transformation. Clustering are done for several data sets from [11] and their clustering accuracies are reported [12].

4. Black Widow Optimization Algorithm (BWOA)

Figure 2 shows the BWOA flowchart. The BWOA starts with an initial population of spiders like other evolutionary algorithms, so that every spider represents a potential solution. The initial spiders, in pairs, try to reproduce the new generation. Female black widow eats the male during or after mating. Then, it carries stored sperms in its sperm thecae and releases them into egg sacs. They cohabit on the maternal web for several days to a week, during which time sibling cannibalism is observed [15].

4.1 Initial population

The black widow optimization begins with a random initial black widow spider population which has male and female black widow spiders to generate offspring for the next generation. The widow fitness is obtained by evaluation of fitness function f at a widow, the initial population of black widow spiders can be expressed as:

$$X_{N,d} = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} \ldots\ldots x_{1,d} \\ x_{2,1} & x_{2,2} & x_{2,3} \ldots\ldots x_{2,d} \\ x_{N,1} & x_{N,2} & x_{N,3} \ldots\ldots x_{N,d} \end{bmatrix} \quad (1)$$

$$lb \leq X_i \leq ub$$

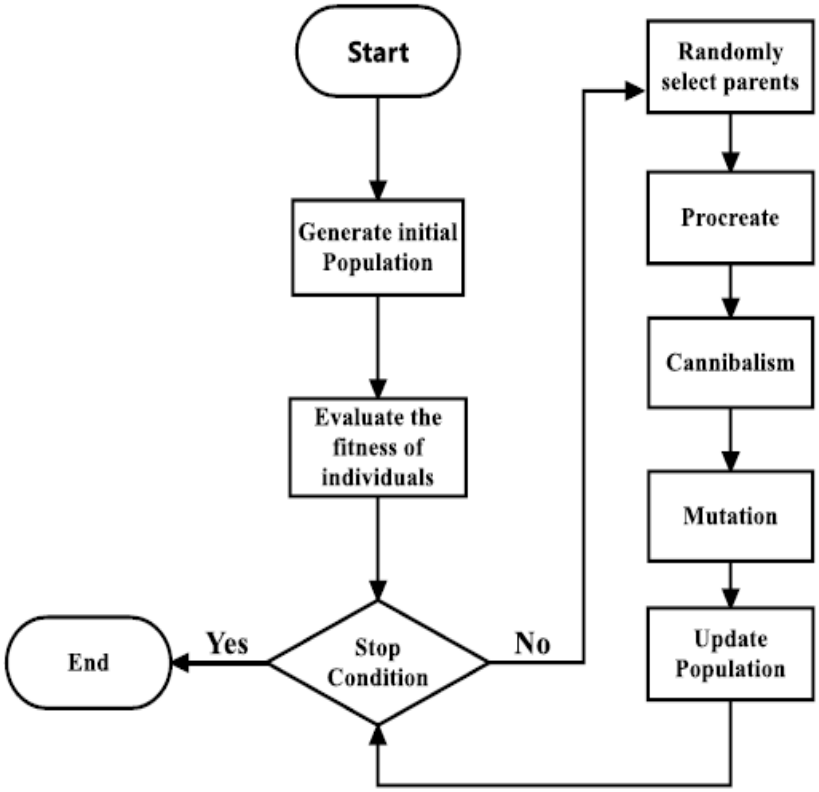

Figure 2. Flowchart of the black widow optimization algorithm [1]

$X_{N,d}$ is the black widow spiders population, d is the number of decision variables of the problem, N is the number of population, l_b is the population lower bound, and u_b is the population upper bound. The potential solution populations ($X_{N,d}$) are used for minimizing or maximizing the following objective function represented in Equation (2):

$$\text{Objective function} = (X_{N,d}) \quad (2)$$

4.2 Procreate

As the pairs don't depend on each other, they start to mate to reproduce a new generation, in parallel, as well in nature, each pair mate in its web, separately from other spiders. In real world, about 1000 eggs are produced in each mating, but in the end, some of the strongest spider babies are survived. Now, here in this algorithm so as to reproduce, an array called mio should also be created as long as widow array with random numbers containing, then

offspring is produced by using μ with the following equation, in which x_1 and x_2 are parents, y_1 and y_2 are offspring [17].

$$y_1 = \mu \times x_1 + (1 - \mu) \times x_2 \quad (3)$$

$$y_2 = \mu \times x_2 + (1 - \mu) \times x_1 \quad (4)$$

y_1, and y_2 are the young spiders from reproduction, i and j are a random number between 1 to N and μ is the random number between 0 and 1.

4.3 Cannibalism

Here we have three kinds of cannibalism: The first one is sexual cannibalism; where the female black widow eats her husband during or after mating. In this algorithm, we can recognize any female and male by their fitness values. Another kind is sibling cannibalism where the strong spider lings eat their weaker siblings. In this algorithm, a cannibalism rating is set (CR) where we can determine the number of survivors according to. In some cases, the third kind of cannibalism is often observed in which the baby spiders eat their mother. The fitness value is used to determine strong or weak spider lings [15].

4.4 Mutation

In this stage, we randomly select Mute pop number of individual form population. As Figure 3 illustrates, each of the chosen solutions randomly exchanges two elements in the array. Mute pop is calculated by the mutation rate [1].

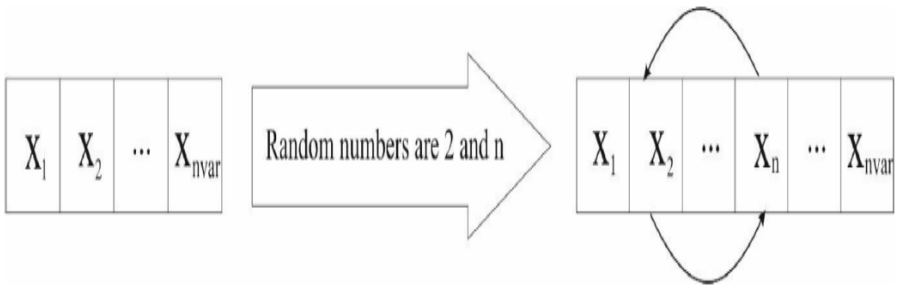

Figure 3. Mutation [1]

4.5 Convergence

Like other evolutionary algorithms, three stop conditions can be considered:
(A) a predefined number of iterations.
(B) Observance of no change in the fitness value of the best widow for several iterations.
(C) Reaching to the specified level of accuracy.

5. BWOA for Clustering Problems

The K-means cluster algorithm is one of the most widely used clustering algorithms and has been applied in many fields of science and technology. A black widow optimization with K-means cluster algorithm that efficiently eliminates the drawback of empty cluster is presented in this paper.

5.1. Phase I: K-Means Algorithm

Step 1: K initial cluster centers $z_1, z_2, \ldots\ldots, z_K$ are chosen randomly from the n observations $\{x_1, x_2, \ldots\ldots, x_n\}$.
Step 2: A point x_1, $i = 1, 2, \ldots\ldots, n$ is assigned to cluster C_j, $j \in \{1, 2, , k\}$ *iff*

$$\|x_i - z_j\| \leq \|x_i - z_p\|, \text{ P } = 1,2,\ldots\ldots,\text{K } \& \text{ j} \neq \text{P} \qquad (5)$$

Step 3: New cluster centers z_1, z_2, z_K are computed as follows:

$$z_j = \frac{1}{n_i} \sum_{\text{xi} \in c_j}^{\text{n}} x_i \quad i = 1,2, \ldots\ldots, K \qquad (6)$$

where n_i is the number of elements belonging to cluster C_j.
Step 4: If $z_i^* = z_i$ $i=1,2,\ldots.,K$ then terminate, otherwise continue from step 2. After this phase we get an initial centre for all predetermined clusters.

5.2. Phase II: Black Widow Optimization Algorithm

Step 1: Initial population

Each individual represents a row-matrix $1 \times n$ where n is the number of observation, each widow contain integer $[1, K]$ which represent the cluster which this observation belongs to it. e.g., let there is ten observations $\{x_1, x_2, \ldots., x_{10}\}$ which must be assigned to four cluster $k = 4$. Table 1 shows the structure of the individual. Also Figure 4 represents the four clusters classification.

Table 1. The structure of the individual

Individual 1	3	1	2	3	2	1	3
Observations	x_1	x_2	x_3	x_4	x_5	x_6	x_7

Evaluate the desired objective function; where the task is to searching for appropriate cluster classifications such that the fitness function is minimized. The clustering fitness function for the K clusters $C_1, C_2, \ldots\ldots, C_K$ is given by

$$f\ (C_1, C_2, \ldots\ldots, C_k)\ = \sum_{i=1}^{k}\ \sum_{j\in c_i}^{n}\ \|x_j\ +\ z_i\| \qquad (7)$$

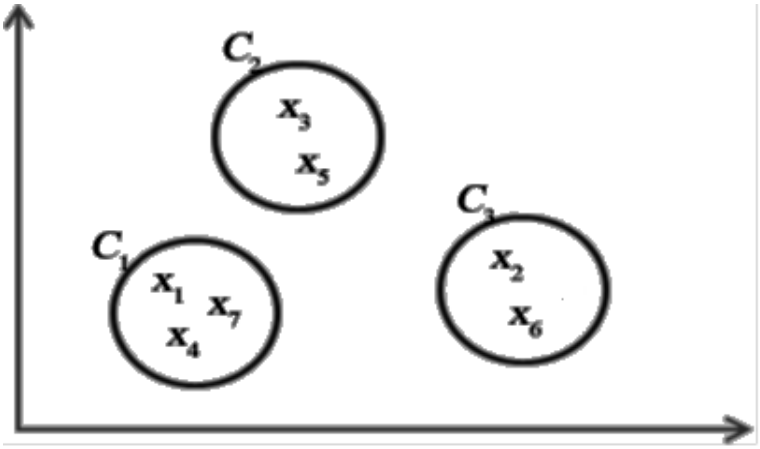

Figure 4. Classification of the four clusters.

Step 2: Procreate

According to equations 3 and 4 we get offspring using parents to procreate the young spiders from reproduction process. To avoid random duplication selection of pairs, the reproduction process is carried out for d/2 times.

Step 3: Cannibalism

sibling cannibalism in which the strong spider lings eat their weaker siblings. In this algorithm, we set a cannibalism rating (CR) according to which the number of survivors is determined.

Step 4: Mutation

For each individual, mutation operator is implemented as follows, first select two columns randomly from i_{th} individual and then generate two new columns as shown in Table 2.

Table 2. The structure of the individual

	*					*	
Individual 1	3	1	2	3	2	1	3
Observations	x_1	x_2	x_3	x_4	x_5	x_6	x_7

Step 5: Convergence

In this step keep the best solutions of the existing population; where BWOA keeps the best solutions found so far during the process.

6. Experimental Results

Our BWOA with K-Means Algorithm are implemented with parameters in table 3 using Python and run on a personal computer with Intel Core i7-10510U 1.80GHz processor, 8GB RAM, and Windows 10 Pro (64-bit) operating system. We have experimented with data sets namely Iris, and Seeds. The Iris data set has 4 real attributes (Sepal length, Sepal width, Petal length, Petal width) cm. It has 150 data points with three data categories, namely Setosa, Versicolor and Virginica. Each category has 50 data points. For our experiment, we take the number of clusters k equal to the number of data categories mentioned in the data set. Hence, for Iris data set, k = 3 is used. The confusion matrix generated. We have calculated the clustering accuracy (ACC) using the formula in (8).

$$ACC = \frac{\text{No.of data points clustered correctly}}{Number\ of\ data\ points\ in\ the\ dataset} * 100\% \quad (8)$$

Table 3: Parameters values of BWOA.

procreate rate	cannibalism rate	mutation rate
O.64	**0.47**	**0.42**

Table 4: confusion matrix generated by our BWOA with k-means based method for Iris data set

Data Category	Cluster ID		
	Setosa	Versicolor	Virginica
Setosa	**50**	0	0
Versicolor	0	**47**	3
Virginica	0	2	**48**

Table 5: Clustering accuracy comparison of our BWOA for Iris data set with previous work

ACC %		
k-means [12]	Genetic Algorithm [14]	Results of our BWOA
88.70 %	93.33%	**96.67%**

The Seeds data set has 7 real attributes (area A, perimeter P, compactness, length of kernel, width of kernel, asymmetry coefficient and length of kernel groove) and 210 data points. It has three data categories, namely Kama, Rosa, and Canadian. Each category has 70 data points. For our experiment, we take

the number of clusters k equal to the number of data categories mentioned in the data set. Hence, for Seeds data set, k = 3 is used.

Table 6: confusion matrix generated by our BWOA with k-means based method for Seeds data set

Data Category	Cluster ID		
	Kama	Rosa	Canadian
Kama	**62**	8	0
Rosa	7	**63**	0
Canadian	0	0	**70**

Table 7: Clustering accuracy comparison of our BWOA for Seeds data set with previous work

ACC %		
k-means [12]	Genetic Algorithm [14]	Results of our BWOA
89.05%	92.86%	**92.86%**

7. Conclusion

Data clustering is regarded as one of the most useful and important data mining techniques. K-means algorithm is usually used for data clustering. Data clustering can be considered a combinatorial optimization problem and metaheuristic algorithm like BWOA which may be a good choice for data clustering. This paper proposes K-means clustering algorithm with BWOA. The algorithm maintains all important characteristic features of the K-means algorithm and at the same time removes the possibility of generation of empty clusters. Experimental results show that the proposed clustering algorithm together with BWOA are able to solve the clustering problem. Results of simulation experiments using data sets called Iris and Seeds are used to prove our claim.

REFERENCES

[1] Hayyolalam, V. Kazem, A.A.P. “Black Widow Optimization Algorithm: A novel meta-heuristic approach for solving engineering optimization problems,” Eng. Appl. Artif. Intel., 87, 103249, (**2020**).

[2] Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P. and Uthurusamy, R. “Advances in Knowledge Discovery and Data Mining,” MIT Press, Menlo Park, (**1996**).

[3] Han, J. and Kamber, M. "Data Mining: Concepts and Techniques," Morgan Kaufmann, San Francisco, (**2000**).

[4] Chawla, S. "A Novel Approach of Cluster Based Optimal Ranking of Clicked URLs Using Genetic Algorithm for Effective Personalized Web Search," Applied Soft Computing, 46, (**2016**), pp.90-103.

[5] Abedini, M., Moradi, M.H. and Hosseinian, S.M. "Optimal Clustering of MGs Based on Droop Controller for Improving Reliability Using a Hybrid of Harmony Search and Genetic Algorithms," ISA Transactions, 61, (**2016**), pp.119-128.

[6] Binu, D. "Cluster Analysis Using Optimization Algorithms with Newly Designed Objective Functions," Expert Systems with Applications, 42, (**2015**), pp. 5848-5859.

[7] Kaczmarowski, A., Yang, S., Szlufarska, I. and Morgan, D. "Genetic Algorithm Optimization of Defect Clusters in Crystalline Materials," Computational Materials Science, 98, (**2015**), pp.234-244.

[8] Brito, P., Bertrand, P., Cucumel, G., de Carvalho, F, "Selected contributions in data analysis and classi_cation," A Festschrift for E. Diday. Springer, Heidelberg, (**2007**).

[9] U. Maulik and S. Bandyopadhyay, "Genetic algorithm-based clustering technique," Pattern Recognition, vol. 33, (**2000**), pp. 1455 – 1465.

[10] H.-J. Lin, F.-W. Yang, and Y.-T. Kao, "An efficient GA-based clustering technique," Tamkang Journal of Science and Engineering, vol. 8, no. 2, (**2005**), pp. 113 – 122.

[11] K. G. Soni and A. Patel, "Comparative analysis of k-means and k-medoids algorithm on iris data," International Journal of Computational Intelligence Research, vol. 13, no. 5, (**2017**), pp. 899 – 906.

[12] UCI Machine Learning Repository, Last accessed on 10 June, (**2019**).

[13] M. Wei, T. W. S. Chow, and R. H. M. Chan, "Clustering heterogeneous data with k-means by mutual information-based unsupervised feature transformation," Entropy, vol. 17, (**2015**), pp. 1535–1548.

[14] K. Musharrat, K. B. Pappu, B, Priom, and T. I. Md, "Data Clustering Using Hybrid Genetic Algorithm with k-Means and k-Medoids Algorithms," Conference Paper, (**2019**). doi: 10.1109/ICSEC47112.2019.8974797.

[15] K. Premkumar, M. Vishnupriya, T. S., Babu, B. V. Manikandan and A. Z. Kouzani, "Black Widow Optimization-Based Optimal PI-Controlled Wind Turbine Emulator," Sustainability, 12, (**2020**). doi:10.3390/su122410357.

[16] Gallegos, M.T, "Maximum likelihood clustering with outliers," In: K. Jajuga, A. Sokolowski, H.-H. Bock (eds.): Classi_cation, clustering, and data analysis. Springer, Heidelberg, Gallegos, M.T., Ritter, G (**2002**), pp.248-255.

[17] Scott, Catherine, Kirk, Devin, McCann, Sean, Gries, Gerhard, "Web reduction by courting male black widows renders pheromone-emitting females' webs less attractive to rival males," Anim. Behav. 107, Retrieved July 16, (**2015**), pp.71–78.