International Journal of Computational Methods and Experimental Measurements
Vol. 12, No. 2, June, 2024, pp. 185-190
Journal homepage: http://iieta.org/journals/ijcmem

Enhanced Unsupervised Feature Selection Method Using Crow Search Algorithm and Calinski-Harabasz

Fatima M. Hasan, Talal F. Hussein*, Hanadi D. Saleem, Omar S. Qasim

Department of Mathematics, University of Mosul, Mosul 41002, Iraq

Corresponding Author Email: talal.math@uomosul.edu.iq

https://doi.org/10.18280/ijcmem.120208

Received: 29 May 2023
Revised: 27 August 2023
Accepted: 30 November 2023
Available online: 30 June 2024

Keywords:
crow search algorithm, Calinski-Harabasz index, K-mean clustering, feature selection, data mining

ABSTRACT

This paper proposes enhancing the K-means clustering method by incorporating the Crow Search Algorithm (CSA) and Calinski-Harabasz (CH) index to address the issue of determining the optimal number of clusters and attribute selection. The proposed approach, called Crow Search Algorithm K-mean clustering (CSAK_means), aims to explore the search space more effectively to find the best solutions. The efficiency of the CSAK_means algorithm is evaluated using a comparative experimental study for five datasets from the UCI repositories: Wine, Bodega, Cmc, Zoo, and Abalone. The results confirm that the proposed method outperforms the default algorithms in terms of average feature selection performance and silhouette value.

1. INTRODUCTION

Feature selection is an important process in machine learning and data mining that involves choosing suitable features from many options to create a model that accurately represents the output variables, this step is necessary to improve the performance of the model by reducing overfitting, improving model interpretation, and reducing computational time [1, 2]. When it comes to selecting model features, there are basically three methods that can be used: wrapper-based, filter-based, combined methods [3, 4]. Wrapper-based methods use a classifier to evaluate the performance of an object, while filter-based methods use statistical measures [5]. Features are selected in the modeling process using embedded methods such as decision trees or neural networks. When there are many possibilities in the optimization problem, the feature selection method can be effective [6, 7].

Metaheuristic algorithms such as particle swarm optimization (PSO), artificial bee colony (ABC), differential evolution (DE), and gray wolf algorithm (GWA) have been developed to solve many problems. Feature selection is one of the applications that have some drawbacks despite the advantages of there is on, such as how the factors chosen for all situations or data sets may not be optimal for each, and if selection procedures are not well designed, they can introduce bias, so researchers need to be careful [8].

When observing crow food storage and retrieval behaviors developed an optimization method called Crow Search Algorithm (CSA) [8, 9]. It is a powerful technique that can modify hyperparameters and it is very interesting that be used in various industrial applications thanks to a population- based approach and only two parameters configurable: travel length and awareness probability. Crow swarm intelligence is the basis for CSA technology, which mimics the subtle behavior of swarms of crows. This method was published by Hussien et al. [10] in 2016. This method has been widely used by researchers due to its ease of use, efficiency, and low number of parameters so many modifications of the basic CSA algorithm are not used to overcome this problem. For example, some researchers introduced an adaptive inertial weighting factor and a roulette wheel selection system to enhance detection and control capabilities [11]. Prior work has employed Lévy flight motion to improve the search capabilities of the CSA algorithm, as well as dynamic modifications to the fixed awareness probability value based on the fitness value of individual candidate solutions. To further its efficiency, it has also been coupled with 10 chaotic maps and, under some situations, the CSA algorithm. Others include Differential Evolution (DE) and the BAT algorithm (BA).

Filter-based techniques use statistical measures to ascertain the value of each feature, while wrapper-based approaches use a classifier to evaluate feature performance [12]. Built-in techniques such as decision trees and neural networks use feature selection methods and classifiers together at the same time.

The Crow Search Method (CSA) was created after research was done on how crows store and recover food [13]. Its two adjustable parameters, flight length and awareness probability, along with its processing-based methodology make it a powerful technology that can optimize hyperparameters and is highly appealing for application in a wide range of technological domains. The foundation of CSA is swarm intelligence, which emulates the clever behavior of crow colonies. Since Askarzadeh first presented this approach, many scholars have used it with just a few modifications [14].

Therefore, researchers have proposed several modifications to the original CSA treatment to overcome this problem. For example, some researchers have included a roulette wheel selection mechanism and an adjustable inertia weight factor to enhance exploration and exploitation capabilities. On the fitness score for the individual possible answer.

The CSA technique outperforms several optimization algorithms in six engineering design tasks through the strategies in its design and structure, and with only a few criteria required, this strategy can be effective and easy to implement, leading to positive results in the search for possible solutions through solve the problem [15, 16].

The Calinski-Harabasz (CH) index, also known as the variance criteria ratio, is often used to evaluate the splitting quality using the K-means clustering technique for a specified number of clusters by calculating the ratio of the total between-cluster and inter-cluster dispersion for all classes [10]. Model performance can be evaluated using the CHI score, which is associated with greater clustering performance, in scenarios where ground truth labels are unknown. In addition to being used with other clustering techniques.

This study aims to improve the K-means clustering technique by combining the Crow Search Algorithm (CSA) and the Calinski-Harabasz (CH) to obtain the best selection of the number of clusters or centers as well as to select the best features, where we studied the nuances of the elements Our study has been meticulously designed to give readers a comprehensive understanding and enjoyment of these subjects via in-depth analysis.

The paper is structured as follows: Section 2 provides a thorough synopsis of CSA. In Section 3 we move on to the CH index, which is a measure to evaluate how well classified data is generated using the K-means clustering algorithm. Section 4 focuses on the K-means clustering method. We outline our implementation framework in Section 5, which also presents a proposed action. Research findings for several datasets are presented in Section 6. In Section 7, we conclude the research presented in this article.

2. CROW SEARCH ALGORITHM (CSA)

The Crow Search Algorithm (CSA), developed by Seyadali Mirjalili, is inspired by the instinctive behavior of crows. This new method developed in 2016 has proven to be a useful tool for solving complex optimization problems [17]. A popular optimization strategy in the swarm intelligence family is called CSA. "A swarm intelligence" is a combination of algorithms that are triggered by biological factors or animal emotional reactions. Crows are a unique source of inspiration for CSA as new solutions are sought [18].

Swarm intelligence methods such as ant colony optimization (ACO), firefly algorithm (FA), and particle swarm optimization (PSO) are used for stakeholders to collectively generate intelligence to find answers to optimization problems of crows hiding, food, other birds CSA technology, using food and habit stealing, using the same concept [19].

Crows follow each other when they steal food, they live in groups, they can remember where they concealed their food, and they have the perceptual capacity to change where they are hiding when they sense danger. These are some of the four main principles of the CSA. To use this idea, picture a group of crows in a d-dimensional search space. The group is made up of N crows, each of which represents a possible solution to the problem, and d indicates the number of option variables.

Let $x^{i,iter} = (x_1^{i,iter}, x_2^{i,iter}, x_d^{i,iter})$ using the matching iteration number, indicate the position of the crow I that was determined during the iteration. The crow position may be updated using the following formula:

$$x^{i,iter+1} = \begin{cases} x^{i,iter} + r_i * fl^{i,iter} * (m^{j,iter} - x^{i,iter}), r_j \geq AP^{j,iter} \\ random\ position, \quad otherwise \end{cases} \quad (1)$$

where, $m^{j,iter}$ is the best place for the crow j to hide food after the iteration numbered $iter$, r_i is a random number subject to uniform distribution in [0,1], and $fl^{i,iter}$ represents the flight length of the crow i in the iteration numbered iter, and $AP^{j,iter}$ represents the awareness probability of the crow j in the iteration numbered iter. The likelihood that crow j perceives being monitored is decreased the lower $AP^{j,iter}$ is. As a result, Eq. (1) says that when $r_j \geq AP^{j,iter}$ crow i tracks crow j to a new hiding-food position. When $r_j < AP^{j,iter}$ crow feels that it is being followed, it will fly to a site where there is food that is not hidden. When j is position loses its tracking value, a new position for the crow is created at random.

$$m^{i,iter+1} = \begin{cases} x^{i,iter+1}, f(x^{i,iter+1})\ is\ better\ than\ f(m^{i,iter}) \\ m^{i,iter}, \quad otherwise \end{cases} \quad (2)$$

where, $f(.)$ is the primary objective,

$$m^{i,iter} = (m_1^{i,iter}, m_2^{i,iter}, m_d^{i,iter})$$

Following the numerical iterations, stores the crow i at the best possible storage place. The exploration and exploitation phase of the CSA algorithm can also be learned from the Figure 1.

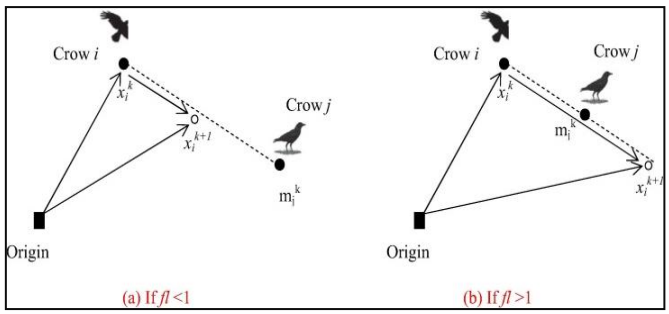

Figure 1. The crow position update diagram in CSA case i shows the effect of fl in searching process [17]

The main mechanism on which the CSA algorithm relies in finding solutions while searching the solution space in order to obtain the optimal solution can be summarized through the following sequential steps:

Algorithm: Crow Search Algorithm
Step 1. Begin by initializing a population of crows $x^{i,iter}$ with random positions.
Step 2. Evaluate the fitness of each crow using a fitness function. This helps to determine how well the crow is performing in solving the optimization problem.
Step 3. Generate new positions for the crows $x^{i,iter+1}$ using

a specified algorithm.
Step 4. Check the feasibility of the new positions. This ensures that the crows are not moving to illegal or impossible locations.
Step 5. Analyze the new positions' fitness function. This aids to evaluate the crows' performance following their creation of their new places.
Step 6. Update the memory of the algorithm $m^{i,iter+1}$. This involves keeping track of the best position found so far.
Step 7. Check the criteria for dismissal. Steps 3-6 are performed until the maximum number of iterations (itermax) has been accomplished. When the termination requirement is met, an ideal memory place in respect to the objective function's value is given as the optimization problem's solution.

3. THE CALINSKI-HARABASZ INDEX (CH)

Focusing on integration, with the K-means clustering technique, the CH index is a commonly used statistic for evaluating clustering schemes [20, 21]. This index, which shows how effectively data is separated into a specified number of clusters by algorithms, is essential for evaluating the effectiveness of clustering algorithms.

This effectiveness is statistically evaluated using contrast ratio criteria, sometimes referred to as the CH index. Its calculation involves dividing the total dispersal within and between groups by the total dispersal, where "dispersal" is defined as the sum of the squared distances. In simple terms, it contrasts the degree of dispersion of data points between groups with the amount of distribution within each group [22]:

$$CH = \frac{\frac{BGSS}{K-1}}{\frac{WGSS}{N-K}} = \frac{BGSS}{WGSS} * \frac{N-K}{K-1} \quad (3)$$

where,

N: the total amount of observations.

K: the overall cluster count.

The formula for calculating the between-group sum of squares inter-cluster dispersion is as follows:

$$BGSS = \sum_{k=1}^{K} n_k * \|C_k - C\|^2 \quad (4)$$

where,

n_k: how many observations there are in cluster *k*.

C_k : the cluster *K* centroid.

C: the barycenter, or centroid, of the dataset.

K: how many clusters there are enter an equation here.

The following equation is used to determine WGSS or within-group sum of squares intra-cluster dispersion.

$$WGSS_k = \sum_{i=1}^{n_k} \|X_{ik} - C_k\|^2 \quad (5)$$

where,

n_k: the quantity of data in cluster *k*.

C_k : the cluster *k* centroid.

X_{ik}: the i-th observation of cluster *k*.

Then add up each individual square sum within a group:

$$\text{WGSS} = \sum_{k=1}^{K} WGSS_k \quad (6)$$

where,

$WGSS_k$: the within group sum of squares of cluster *k*.

The big values of the Calinski-Harabasz index, according to the above calculation, indicate superior clustering.

4. K-MEANS CLUSTERING

The sum of the squared distances between data points and the cluster centroids they correspond to is minimized using the K-means algorithm, a powerful unsupervised clustering tool. By dividing the data set into K distinct clusters, this method maximizes the similarity within each cluster while facilitating appropriate separation between clusters. The main steps of the CSA flowchart can be illustrated in Figure 2 [23]:

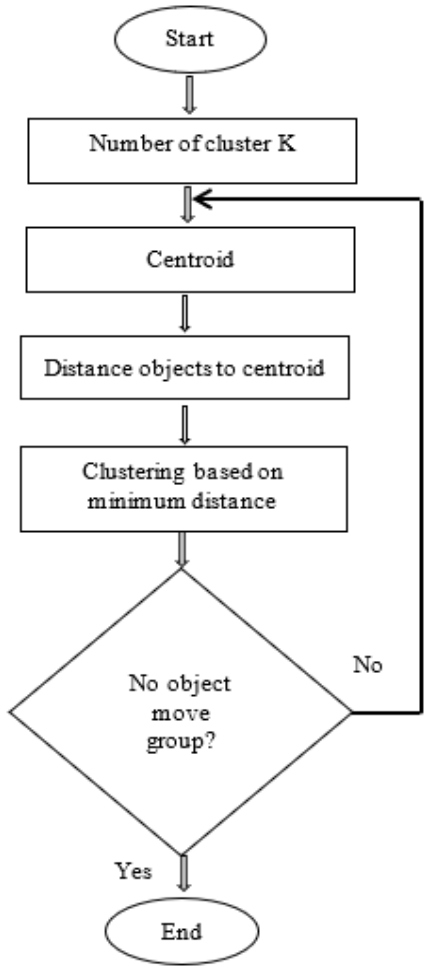

Figure 2. Flowchart of K-means algorithm [24]

The method randomly selects K data points as initial centroids before starting the iterative process of mapping each data point to its closest centroid. Once convergence is achieved, these centroids are later calculated again, often by averaging the data points within each group.

K-means is a popular choice in fields such as machine learning and image processing due to its known ability to rank large data sets according to underlying patterns and structures. To measure the similarity between data points and focal points [24].

$$D(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^n} \quad (7)$$

5. THE PROPOSED ENHANCEMENT

The proposed strategy aims to increase the performance of K-means clustering by simultaneously increasing the number of clusters and adding feature selection. The Crow Search Algorithm (CSA) is used as a population-based search strategy, with the K value and feature selection expressed as binary strings. Performance in K-means clustering is strongly influenced by the selection of K, where K stands for the number of clusters. While numerous methods, including some inspired by natural algorithms, have been employed to enhance K-means clustering performance through suitable K selection, none have attempted to choose multiple features at once. The suggested search technique incorporates feature selection using the CSA and tries to maximize the number of clusters in K-means clustering.

Input data set X of size $N \times D$, where D represents the number of features and N represents the number of instances. Set the number of clusters, $Kmax$, to be considered in the search.

Initialize the population of crows, C, with a set of K and feature selection pairs.

Determine each crow's fitness in C by applying the K-means clustering algorithm with the suitable K and feature selection.

To create a new population $\acute{C}$, choose the most fit crows from C based on their fitness scores.

Use the search operators of CSA, including the levy flight and crossover operators, to generate new solutions for $\acute{C}$.

Evaluate the fitness of the new solutions in $\acute{C}$ using the K-means clustering algorithm with the corresponding K and feature selection.

Select the best crows from $\acute{C}$ to form a new population $\acute{C}$ based on their fitness values.

Continue from step 6 through step 8 until a termination condition is met, such as when the maximum number of iterations is achieved or the best fitness value converges.

Output the solution with the highest fitness value from the final population as the selected K and feature subset for K-means clustering.

The suggested technique concurrently optimizes the number of clusters and feature selection for K-means clustering using a population-based search strategy with the CSA operators. The K-means clustering technique, which assesses the effectiveness of the clustering solution, serves as the foundation for the fitness evaluation. By changing and merging the already-existing solutions in the population, the CSA operators (such as levy flight and crossover) are employed to produce new solutions. When a stopping requirement is satisfied, such as a maximum number of iterations or convergence of the best fitness value.

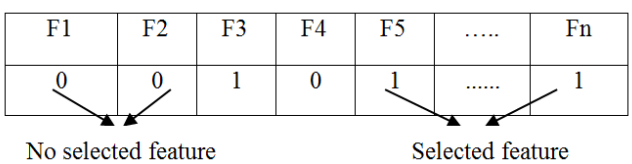

Figure 3. Representation of feature selection in the proposed algorithm [5]

Figure 3 is an illustration of the solution representation, which shows how the K and feature selection are encoded in a binary string format. The first part of the string represents the K value, and the second part represents the feature selection, where a 1 indicates that the corresponding feature is selected and a 0 indicates that it is not selected.

6. RESULTS AND DISCUSSION

We used the proposed CSAK_means method on five distinct publicly accessible datasets to evaluate its effectiveness. We also examined how well it performed compared to the traditional K-means algorithm that uses CSA to determine the optimal number of clusters. Through this comparison, we can confirm that the proposed algorithm outperforms the traditional method in terms of results.

Table 1 provides a brief overview of the datasets used in the experiments, which were obtained from the UCI Machine Learning Repository. The experiments used five real-world datasets that differed in dimensionality, number of observations, and clusters (represented by C). Each dataset is briefly described in the table to provide an overview of its characteristics.

Table 1. Describe the characteristics of the dataset

The Datasets	Instances (N)	Dimensions (D)
Data1 (Wine)	178	13
Data2 (Biodeg)	1055	41
Data3 (Cmc)	1473	9
Data4 (Zoo)	101	17
Data5 (Abalone)	4177	8

The silhouette score is used as a statistic to evaluate the effectiveness of the algorithms put into practice. A cluster validation approach is used to obtain this result, to minimize the distance between points within each cluster and maximize the gap between clusters. It measures the quality of the clustering results. The average sampling distance from every other point in the same group is less than the average sampling distance from every other point in the closest group. High values indicate a strong match between the sample and its group, while low or negative values reflect the presence of too many or too few clustering. The final score ranges from -1 to +1. The aggregation configuration is appropriate when most of the elements have high degrees of silhouettes, where the silhouette width $s(i)$ is determined as follows:

$$s(i) = \frac{b(i) - a(i)}{max\{b(i)\ ,\ a(i)\}} \quad (8)$$

The Silhouette width $s(i)$ is a metric that pertains to a specific data point (i). It is calculated using the average distance between (i) and all other data points in the same cluster (ci), which is represented by $a(i)$, as well as the average distance between (i) and all other data points in a different cluster (ci), represented by $b(i)$.

$$\boldsymbol{b(i) = \min_{I \neq J} \frac{1}{|c_j|} \sum_{j \in J} d(i,j)} \quad (9)$$

where, *d* stands for the separation between *i* and *j*.

Based on the results in Table 2, it is clear, as evidenced by the silhouette value, that the CSAK-means method outperforms K-means for all datasets when clustering accuracy is taken into account. In contrast to K-means, CSAK-means is a better method for selecting and analyzing the number of clusters with a reasonable size. Furthermore, CSAK methods are exceptionally compatible with high-dimensional data and can easily handle complex data. However, data sets with multiple dimensions can cause problems with standard clustering techniques. As a result, CSAK-means provides a useful method for clustering high-dimensional data and outperforms K-means in terms of clustering accuracy.

Table 2. Cumulative accuracy of CSAK-means and K-means algorithms based on silhouette value results

The Datasets	CSAK-Means	K-Means
Data1 (Wine)	0.4225	0.3051
Data2 (Biodeg)	0.3062	0.2344
Data3 (Cmc)	0.4915	0.3481
Data4 (Zoo)	0.5619	0.4819
Data5 (Abalone)	0.4957	0.4457

Table 3. Comparison between CSAK-means and K-means algorithms in terms of feature selection

The Datasets	CSAK-Means	K-Means
Data1 (Wine)	7.2	13
Data2 (Biodeg)	20.6	41
Data3 (Cmc)	3.2	9
Data4 (Zoo)	8.2	17
Data5 (Abalone)	4.8	8

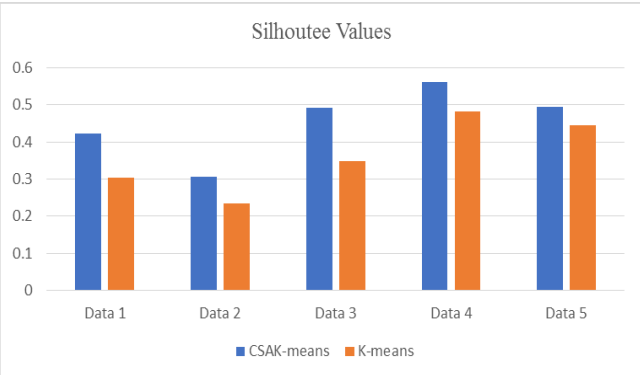

Figure 4. Silhouette values compared between CSAK-means and K-means algorithms

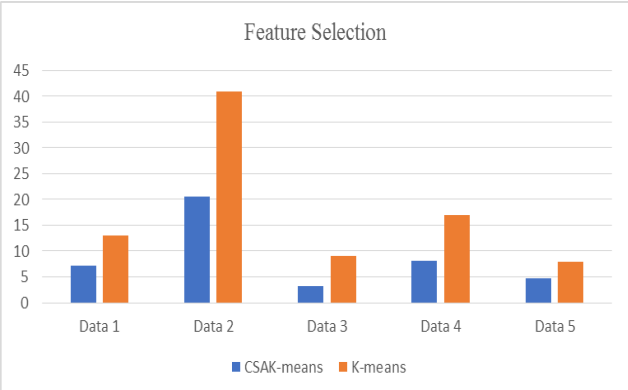

Figure 5. A comparison of feature selection comparison between CSAK-means and K-means algorithms

Based on the data in Table 3, it appears that the CSAK-means algorithm performs better than other methods in determining the optimal number of clusters and selecting features for each dataset, and this shows that the use of the CSAK-means algorithm is usually close to what is expected, compared to the K-means. In general, the CSAK-means algorithm has high statistical efficiency compared to the traditional method in all tests performed on the data, which is evident in Tables 2 and 3, and Figures 4 and 5.

7. CONCLUSIONS

In this work, we presented a proposed algorithm that combines the crow search algorithm (CSA) and the Calinski-Harabasz (CH) index, where CSA is used to find the best feature selection, and CH is used to determine the optimal number of clusters. The proposed CSAK-means is used to improve and provide practical and efficient solutions to clustering problems. The proposed CSAK algorithm outperforms standard methods in terms of silhouette value index and average feature selection, and this is demonstrated by the tests conducted on five sets of data. The proposed algorithm can have many potential applications in analyzing complex and real-world data and can be used to analyze various optimization problems.

REFERENCES

[1] Braik, M., Al-Zoubi, H., Ryalat, M., Sheta, A., Alzubi, O. (2023). Memory based hybrid crow search algorithm for solving numerical and constrained global optimization problems. Artificial Intelligence Review, 56: 27-99. https://doi.org/10.1007/s10462-022-10164-x

[2] Chandrashekar, G., Shin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1): 16-28. https://doi.org/10.1016/j.compeleceng.2013.11.024

[3] Ismael, O.M., Qasim, O.S., Algamal, Z.Y. (2021). A new adaptive algorithm for v-support vector regression with feature selection using Harris hawks optimization algorithm. Journal of Physics: Conference Series, 1897: 012057. https://doi.org/10.1088/1742-6596/1897/1/012057

[4] Nguyen, B.H., Xue, B., Zhang, M.J. (2020). A survey on swarm intelligence approaches to feature selection in data mining. Swarm and Evolutionary Computation, 54: 100663. https://doi.org/10.1016/j.swevo.2020.100663

[5] Qasim, O.S., Mahmoud, M.S., Hasan, F.M. (2020). Hybrid binary dragonfly optimization algorithm with statistical dependence for feature selection. International Journal of Mathematical, Engineering and Management Sciences, 5(6): 1420-1428. https://doi.org/10.33889/IJMEMS.2020.5.6.105

[6] Jović, A., Brkić, K., Bogunović, N. (2015). A review of feature selection methods with applications. In 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, pp. 1200-1205. https://doi.org/10.1109/MIPRO.2015.7160458

[7] Meraihi, Y., Gabis, A.B., Ramdane-Cherif, A., Acheli, D. (2021). A comprehensive survey of Crow Search Algorithm and its applications. Artificial Intelligence

Review, 54: 2669-2716. https://doi.org/10.1007/s10462-020-09911-9

[8] Islam, M.R., Jenny, I.J., Nayon, M., Islam, M.R., Amiruzzaman, M., Abdullah-Al-Wadud, M. (2021). Clustering algorithms to analyze the road traffic crashes. In 2021 International Conference on Science & Contemporary Technologies (ICSCT), Dhaka, Bangladesh, pp. 1-6. https://doi.org/10.1109/ICSCT53883.2021.9642542

[9] Krams, I. (2013). Gifts of the crow: How perception, emotion, and thought allow smart birds to behave like humans. The Auk, 130(3): 556. https://doi.org/10.1525/auk.2013.130.3.556

[10] Hussien, A.G., Amin, M., Wang, M.J., Liang, G.X., Alsanad, A., Gumaei, A., Chen, H.L. (2020). Crow search algorithm: Theory, recent advances, and applications. IEEE Access, 8: 173548-173565. https://doi.org/10.1109/ACCESS.2020.3024108

[11] Necira, A., Naimi, D., Salhi, A., Salhi, S., Menani, S. (2022). Dynamic crow search algorithm based on adaptive parameters for large-scale global optimization. Evolutionary Intelligence, 15: 2153-2169. https://doi.org/10.1007/s12065-021-00628-4

[12] Shi, Z.J., Li, Q.S., Zhang, S., Huang, X.J. (2017). Improved crow search algorithm with inertia weight factor and roulette wheel selection scheme. In 2017 10th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, pp. 205-209. https://doi.org/10.1109/ISCID.2017.140

[13] Bouguettaya, A., Yu, Q., Liu, X.M., Zhou, X.M., Song, A. (2015). Efficient agglomerative hierarchical clustering. Expert Systems with Applications, 42(5): 2785-2797. https://doi.org/10.1016/j.eswa.2014.09.054

[14] Braik, M.S. (2021). Chameleon Swarm Algorithm: A bio-inspired optimizer for solving engineering design problems. Expert Systems with Applications, 174: 114685. https://doi.org/10.1016/j.eswa.2021.114685

[15] Abu-Jamous, B., Kelly, S. (2018). Clust: Automatic extraction of optimal co-expressed gene clusters from gene expression data. Genome Biology, 19: 172. https://doi.org/10.1186/s13059-018-1536-8

[16] Lakshmi, K., Visalakshi, N.K., Shanthi, S. (2018). Data clustering using k-means based on crow search algorithm. Sdhanā, 43: 190. https://doi.org/10.1007/s12046-018-0962-3

[17] Sarker, I.H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN Computer Science, 2: 160. https://doi.org/10.1007/s42979-021-00592-x

[18] Yang, J., Li, Y., Liu, Q., Li, L., Feng, A., Wang, T., Zhen, S., Xu, A.D. Lyu, J. (2020). Brief introduction of medical database and data mining technology in big data era. Journal of Evidence-Based Medicine, 13(1): 57-69. https://doi.org/10.1111/jebm.12373

[19] Delavari, N., Beikzadeh, M.R., Phon-Amnuaisuk, S. (2005). Application of enhanced analysis model for data mining processes in higher educational system. In 2005 6th International Conference on Information Technology Based Higher Education and Training, Santo Domingo, Dominican Republic. https://doi.org/10.1109/ITHET.2005.1560303

[20] Yang, K., Haddad, C.A., Yannis, G., Antoniou, C. (2022). Classification and evaluation of driving behavior safety levels: A driving simulation study. IEEE Open Journal of Intelligent Transportation Systems, 3: 111-125. https://doi.org/10.1109/OJITS.2022.3149474

[21] Hassan, M., Seidel, T. (2017). Using internal evaluation measures to validate the quality of diverse stream clustering algorithms. Vietnam Journal of Computer Science, 4: 171-183. https://doi.org/10.1007/s40595-016-0086-9

[22] Wang, X., Xu, Y.S. (2019). An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. IOP Conference Series: Materials Science and Engineering, 569(5): 052024. https://doi.org/10.1088/1757-899X/569/5/052024

[23] Li, L., Wang, J., Li, X.T. (2020). Efficiency analysis of machine learning intelligent investment based on K-means algorithm. IEEE Access, 8: 147463-147470. https://doi.org/10.1109/ACCESS.2020.3011366

[24] Nagpal, A., Jatain, A., Gaur, D. (2013). Review based on data clustering algorithms. In 2013 IEEE Conference on Information & Communication Technologies, Thuckalay, India, pp. 298-303. https://doi.org/10.1109/CICT.2013.6558109