ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202307_6(7).0008

The Bat Improved K-means Algorithm and Its Application

Na Li[1] and Lipu Zhang[1,*]

[1]College of Media Engineering, Communication University of Zhejiang, Hangzhou, China

* Corresponding Author, E-mail: mathcuz@126. com

Abstract

The k-means algorithm is a typical machine learning algorithm based on divisional clustering, which has the advantages of simplicity and efficiency. Although the k-means algorithm is simple in principle and widely used, the algorithm also has certain drawbacks. To address the shortcomings of the traditional k-means algorithm, which is susceptible to the initial clustering center, we embedded the bat algorithm into the k-means algorithm to optimize the selection of clustering centers and conducted numerical validation through test examples.

Keywords

K-means algorithm, Bat algorithm, Hybrid algorithm, Optimization.

1. Introduction

The k-means clustering algorithm has been widely used in statistics, machine learning, image segmentation, biomedicine, spatial database, etc[1]. Many experts and scholars have worked on improving the k-means clustering algorithm in different aspects. Majhi and Biswal[2] propose a hybrid algorithm of the k-means algorithm and the Ant Lion optimization algorithm to improve the performance of clustering problems. Zahra, et al [3] discuss the recommendation inaccuracy of the recommendation system caused by the random selection of the initial quality center by the K-mean algorithm. Zhang, et al[4] propose a series of novel clustering algorithms by extending the existing k-means class clustering algorithm and combining intra-cluster tightness and separation between clusters. Sakthival, et al[5] present an improved K-means image retrieval algorithm in the image segmentation process, which uses a hierarchical agglomerative clustering algorithm to generate the initial number of clusters and cluster centers. Despite the simplicity of the k-means algorithm in principle and its widespread age, the algorithm also has some drawbacks. First, the value of K needs to be artificially determined in advance. Second, in an unsupervised clustering task, it is impossible to determine how numerous categories are in the dataset, so the value of K is difficult to define. Moreover, the clustering effect of the k-means algorithm depends on the initialization of the cluster centers, which are chosen randomly, and the iteration time of the algorithm can be prolonged if the initial cluster centers are not chosen properly. Thus, the swarm intelligence algorithm can be combined with the k-means algorithm to break the limitations of traditional algorithms and improve clustering performance. Swarm intelligence algorithm algorithms are a class of bio-inspired optimization algorithms that accomplish a given task by a group of simple individuals following a specific interaction mechanism by mimicking genetic evolutionary mechanisms and group collaborative behavior in the biological world. Swarm intelligence algorithms include many bionics algorithms, such as particle swarm optimization algorithms, ant colony algorithms, flower pollination algorithms, firefly algorithms, bat algorithms, whale optimization algorithms, and so on. It has strong self-learning, self-adaptive, and other smart features. Moreover, the algorithm has a simple structure, a fast rate of convergence, and excellent overall convergence. Zhao, et al[6]use a combination of a particle swarm algorithm and a chaotic optimization algorithm for complex coverage optimization in Wireless Sensor

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202307_6(7).0008

Networks (WSN). Farahani, et al[7] propose a better-directed motion firefly algorithm. Stützle and Hoos [8] proposes the maximum minimum ant colony algorithm (MMAS) based on the ant colony algorithm to avoid premature convergence of the algorithm and increase the ability of the algorithm to explore alternative solutions.

To address the shortcomings of the traditional k-means algorithm, which is susceptible to the initial clustering center, the initial clustering center of the k-means algorithm can be optimized by exploiting the strong global search capability and fast convergence of the bat algorithm. This paper verifies the effectiveness and superiority of the algorithm. Firstly, the Bat improved K-means Algorithm(BIKA) is compared with the traditional k-means algorithm on five datasets of standard UCI datasets: Iris, Wine, Seeds, Glass, and Thyroid. Secondly, the BIKA is used to analyze the question C of the 2022 China University Mathematical Modeling Contest, "Composition Analysis and Identification of Ancient Glass Products", in which "choose the appropriate chemical composition to classify the subclasses of high-potassium glass and lead-barium glass".

In this paper, the BIKA will be developed in the following way. The principles of the k-means algorithm and the bat algorithm will be introduced in Chapter 2; The comparison between the BIKA and the traditional algorithm will be introduced in Chapter 3; The application of the BIKA in problem C of the 2022 China University Mathematical Modeling Contest will be introduced in Chapter 4.

2. Introduction to Traditional Algorithms

2.1. Introduction of k-means algorithm

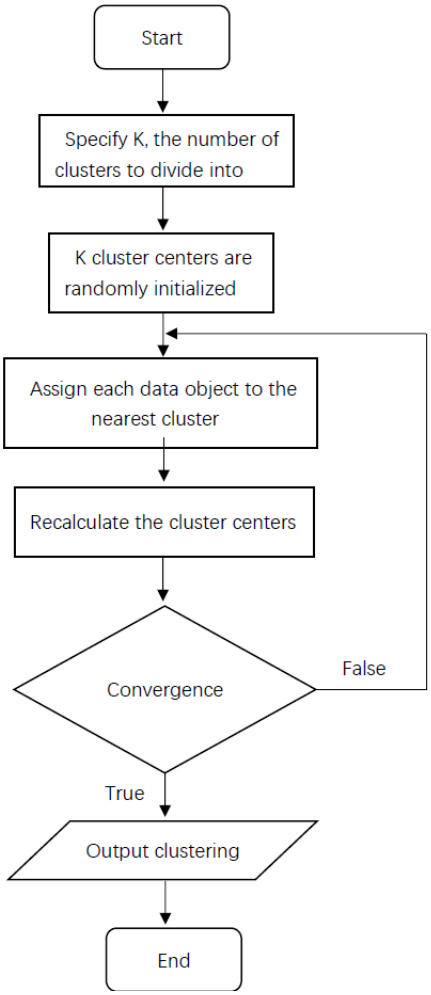

Figure 1. k-means algorithm flow chart

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202307_6(7).0008

The k-means algorithm is an algorithm for unsupervised machine learning. It mainly performs similarity analysis on an unlabeled dataset, which is then divided into classes. The key point of the algorithm is the k-means clustering algorithm which randomly chooses the initial centre of mass, then assigns the other data points to the class closest to them, and then iterates through, placing the center of mass of each class at the mean of the other data points, and then iterates through the same principle.

2.2. Principle of Bat Algorithm

2.2.1. The bionic principle of the bat algorithm

The bat algorithm is a swarm intelligence optimization algorithm based on an iterative search technique inspired by the echolocation predation behavior of bats, which was proposed by Prof.Yang in 2010[9].

Bats emit ultrasound waves as they fly, which are reflected immediately when they encounter an obstacle. Bats use echolocation to detect prey and avoid obstacles. It emits ultrasonic waves and listens to the echoes that bounce back, determining the direction and location of the object based on the different times and intensities of the echoes to both ears.

Thus, the bionic principle of the bat algorithm maps bat individuals with population number m to multiple feasible solutions in a d-dimensional problem space. The optimization process and search are modeled as the process of moving the bat individuals in the population and searching for prey using the value of the fitness function of the solution problem to measure the superiority or inferiority of the position of the bat, and the process of superiority or inferiority of the individuals is analogous to the iterative process of replacing the inferior feasible solution with a superior feasible solution in the optimization and search process.

2.2.2. Model construction of bat algorithm

The following three approximately idealised rules will be used in the model building process to aid in the simulation of the bat algorithm:

(1) All bats use echolocation to sense distance.

(2) The bat flies randomly at position x_i, velocity v_i, with a fixed frequency f, while automatically adjusting the wavelength and loudness of the pulse based on the distance between it and the target prey.

(3) The model assumes that the impulse loudness varies from the maximum value of Amax to the minimum value of Amin, and the variation interval can be defined according to the problem. Under the assumption that the bat search space is D-dimensional, the update rules for position xt iand velocity vt i for each bat at each generation are given in Eqs. (1) to (3):

$$f_i = f_{min} + (f_{max} - f_{min}) * \beta, \quad (1)$$

$$v_i^t = v_i^{t-1} + (x_i^t - X_*)f_i, \quad (2)$$

$$x_i^t = x_i^{t-1} + v_i^t, \quad (3)$$

Whereβis a random number generated according to [0, 1], X* is the optimal solution with respect to the current local position in the population, and fi is the bat's acoustic frequency within the range [fmin, fmax].

Once the globally optimal solution has been selected, each local (x_{old}) solution in the current population updates its location using Equation (4).

$$x_{new} = x_{old} + \varepsilon A^t, \quad (4)$$

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202307_6(7).0008

Where ε is a random number between [-1,1] and A^t is the average loudness of the whole population.

The acoustic loudness A and frequency r of the ith bat were updated using equations (5) and (6):

$$A_i^{t+1} = \alpha A_i^t , \tag{5}$$

$$r_i^{t+1} = r_i^0[1 - \exp(-\gamma t)] , \tag{6}$$

Where $\alpha \in (0,1)$ and is the acoustic loudness decay coefficient. $\gamma > 0$ and is the pulse frequency enhancement coefficient. r denotes the initial pulse frequency of the ith bat.

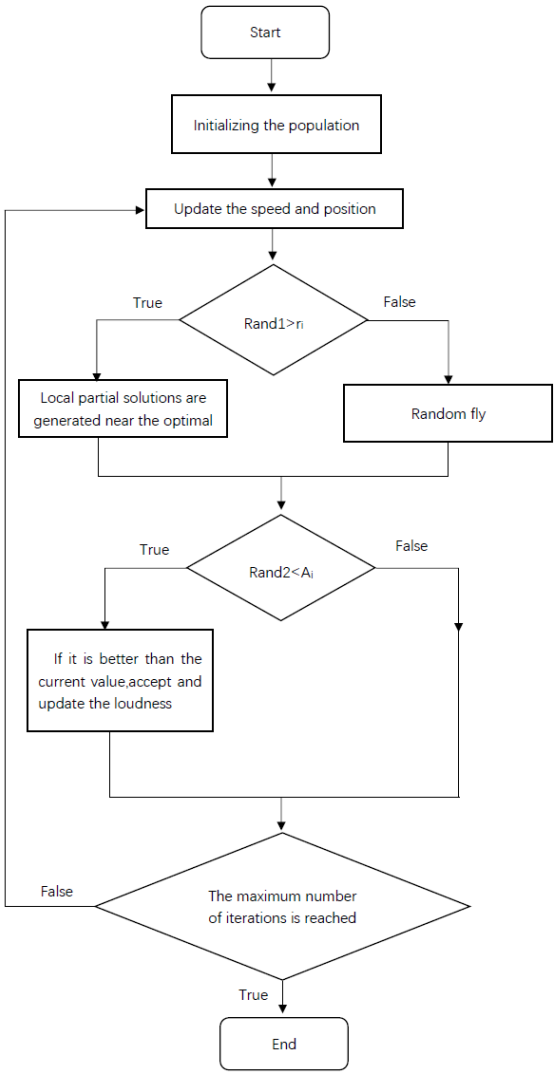

Figure 2. Bat algorithm flow chart

3. The Bat improved K-means Algorithm

3.1. Introduction of the improved algorithm

The core idea of this paper is to take advantage of the strong global search ability and quick convergence of the bat algorithm to find the optimal solution as the initial clustering center of

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202307_6(7).0008

the k-means algorithm to improve the clustering ability of the algorithm and apply it in specific projects. The specific process for BIKA is as follows:

(1) Loading and normalizing data sets

(2) Initializing parameters:the number of clusters K, bat population m, iteration number Tmax, objective function f(x), bat initial position xi (i=1,2,3， … ， m), bat initial velocity vi , acoustic frequency fi, acoustic loudness Ai, frequency ri .

(3) Generate k cluster centroids at random as initial positions of bat populations. The distance from each sample point to the cluster center is calculated based on the Euclidean distance formula, and each sample point is clustered into the category closest to that sample point.

(4) Based on the initial clustering results obtained in step (3), the best position of bat X* in the current population is found, and the velocity and position are updated according to Eqs. (2-1) to (2-3).

(5) Produce a random Rand1 number, and if Rand1 > ri then generate a local solution close to the optimum solution.

(6) If the Rand2 random number has generated again, then the position is accepted if Rand2 < Ai and the fitness is better than the new solution in step (5) at that point in time.

(7) The fitness values of all individuals in the population are sorted to find the current best fitness X*.

(8) The best position found, X* , is used as the new clustering center and re-clustered.

(9) Determine if the maximum number of iterations is reached, if yes, end, otherwise go back to step (4).

(10) Output final clustering results.

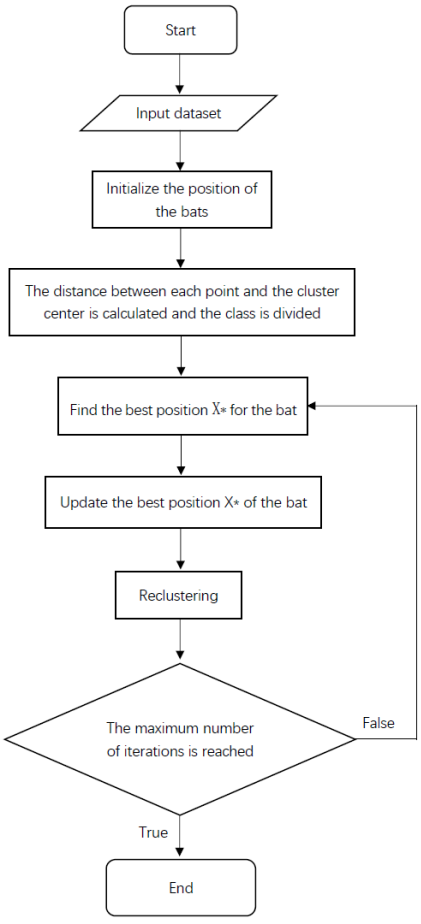

Figure 3. Flowchart for the BIKA

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202307_6(7).0008

3.2. Experiment and Analysis

The datasets were selected from the standard database UCI. The paper test the superiority of the BIKA by comparing experimental results. The running environment of the experiment is Intel(R) Core(TM) i5-10210U, Windows 10, 16G RAM, and Python 3. 9. Datasets used for experiments are given in Table 1.

Table 1. Experimental data set

	Sample	Category	Features
Iris	150	3	4
Wine	178	3	13
Seeds	210	3	7
Glass	214	6	9
Thyroid	215	3	5

The experimental results presented in this paper will be evaluated with respect to the accuracy metric.

(1) Accuracy

$$\mathrm{ACC} = \frac{\mathrm{TP+TN}}{\mathrm{ALL}} \tag{7}$$

The accuracy rate is the ratio of the number of correctly classified samples to the total number of samples. TP represents that the sample is a positive class and is predicted to be a positive class; TN represents that the sample is a negative class and is predicted to be a negative class; ALL represents the total number of samples.

Table 2. Clustering accuracy

	Traditional k-means algorithm	BIKA
Iris	0. 8400	0. 9400
Wine	0. 8652	0. 9607
Seeds	0. 8563	0. 9288
Glass	0. 4872	0. 5629
Thyroid	0. 7035	0. 7854

From Table 2, it is clear that the BIKA is superior to the traditional k-means algorithm in terms of accuracy. Compared to the traditional k-means algorithm, the BIKA improves by 10%, 9. 55%, 7. 25%, 7. 57%, and 8. 19% in accuracy metrics. Figures 4-8 show the clustering of the BIKA on the five Iris, Wine, Seeds, Glass, and Thyroid datasets.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202307_6(7).0008

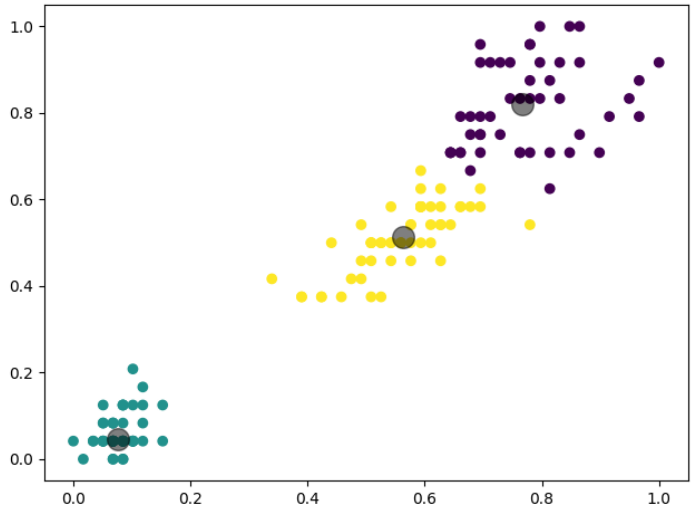

Figure 4. Clustering of the BIKA on the iris dataset

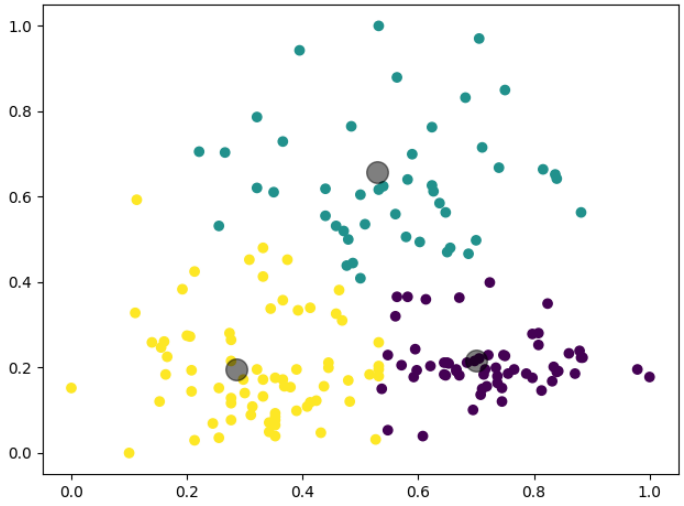

Figure 5. Clustering of the BIKA on the wine dataset

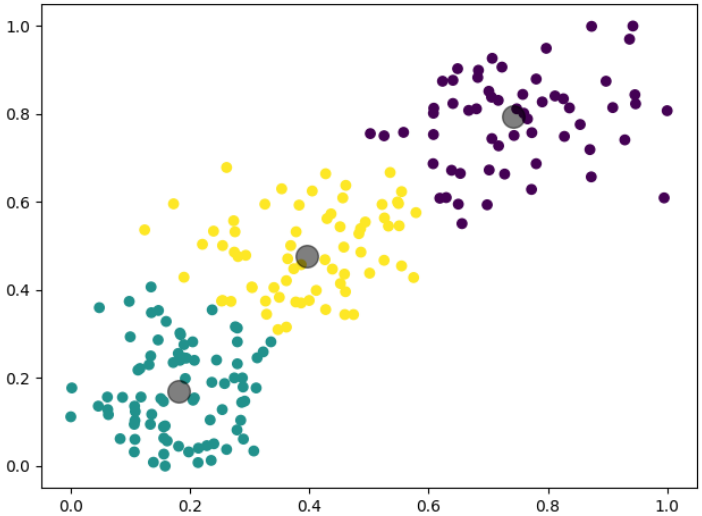

Figure 6. Clustering of the BIKA on the seeds dataset

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202307_6(7).0008

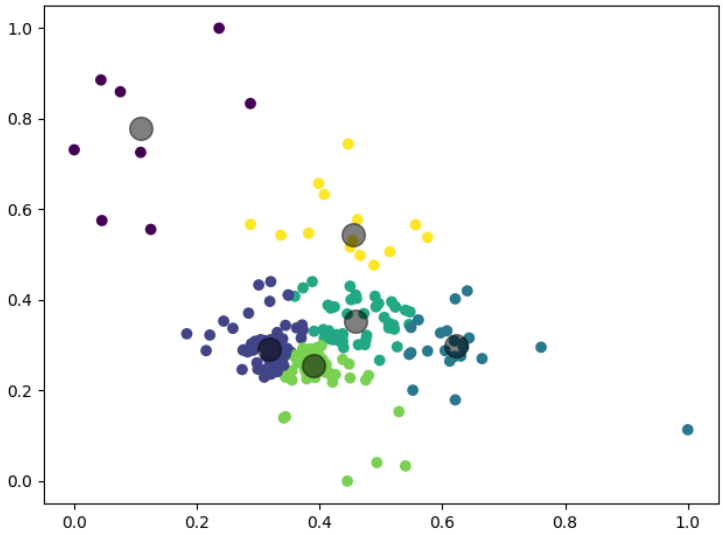

Figure 7. Clustering of the BIKA on the glass dataset

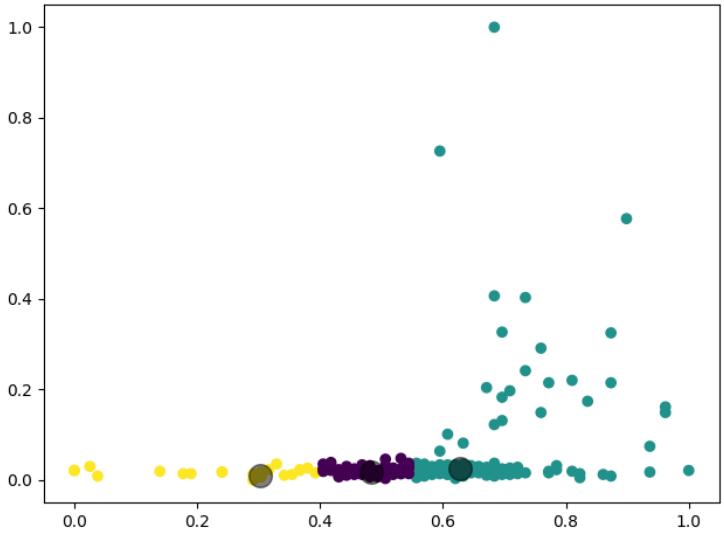

Figure 8. Clustering of the BIKA on the thyroid dataset

4. Application of BIKA in Mathematical Modeling

The BIKA is used to analyze question C of the 2022 China University Mathematical Modeling Contest, "Composition Analysis and Identification of Ancient Glass Products", in which "choose the appropriate chemical composition to classify the subclasses of high-potassium glass and lead-barium glass". (http://www.mcm.edu.cn/)

The data set selected for this paper is the data in Annex to Question C. The data set was normalized to ensure validity and reliability of the experiment.

(1) min-max normalization

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}, \tag{8}$$

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202307_6(7).0008

Normalizing the data will reduce the oscillation of the data model and achieve optimality as quickly as possible. The following is an example of the BIKA for clustering in the high potassium glass dataset.

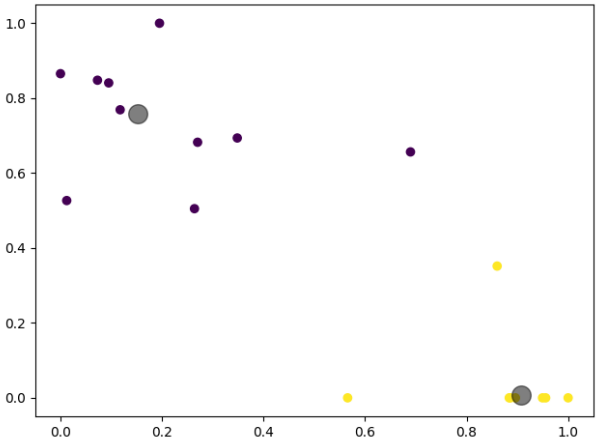

Figure 9. Clustering of the BIKA on the high potassium glass dataset

From Figure 9 , it can be seen that the high potassium glass can be subdivided into high potassium low silica glass and low potassium high silica glass. In this paper, the hybrid model achieves 100% accuracy on the high-potassium glass dataset.

The clustering of BIKA in the high potassium glass data set is shown in the following figure:

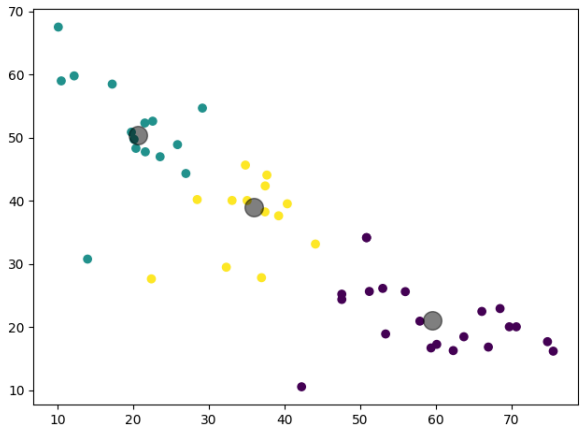

Figure 10. Clustering of the BIKA on the lead-barium glass dataset

As can be seen from Figure 10 , lead barium glass can be subdivided into low lead barium glass, medium lead barium glass, and high lead barium glass. On the lead-barium glass dataset, the BIKA achieved 98. 6% accuracy.

5. Summary

The paper focuses on finding the initial clustering centers of the k-means algorithm by the bat algorithm in the swarm intelligence algorithm, and the experiments are conducted on the dataset of the standard database UCI in comparison with the traditional k-means algorithm. While the BIKA has some improvement in clustering performance, the BIKA also has some limitations in dealing with high dimensional data sets and needs to be further improved.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202307_6(7).0008

References

[1] Jain AK . Data clustering: 50 years beyond k-means[C]// International Conference on Pattern Recognition. North-Holland, 2010.

[2] Majhi S K , Biswal S . Optimal cluster analysis using hybrid k-means and Ant Lion Optimizer[J]. Karbala International Journal of Modern Science, 2018.

[3] Zahra S , Ghazanfar M A , Khalid A , et al. Novel centroid selection approaches for KMeans-clustering based recommender systems[J]. Information Sciences, 2015.

[4] Zhang, Haijun, Huang, et al. Extensions of Kmeans-Type Algorithms: A New Clustering Framework by Integrating Intracluster Compactness and Intercluster Separation[J]. IEEE transactions on neural networks and learning systems, 2014.

[5] Sakthivel K , Abinaya R , Nivetha I , et al. Region Based Image Retrieval using k-means and Hierarchical Clustering Algorithms[J]. Research and Reviews, 2014(1).

[6] Zhao Qiang,Li Changwei,Zhu Dong,Xie Chunli. Coverage Optimization of Wireless Sensor Networks Using Combinations of PSO and Chaos Optimization[J]. Electronics,2022,11(6).

[7] Farahani S M , Abshouri A A , Nasiri B , et al. An Improved Firefly Algorithm with Directed Movement. 2011.

[8] T Stützle, Hoos H H . MAX-MIN Ant system[J]. Future Generation Computer Systems, 2000, 16(8):889-914.

[9] Yang X S , Gandomi A H . Bat Algorithm: A Novel Approach for Global Engineering Optimization[J]. Engineering Computations, 2012, 29(5):464-483.