K-Means Clustering Algorithm Based on Improved Cuckoo Search Algorithm and Its Application

Shuce Ye[1,*], Xiaoli Huang[1,2], Yinyin Teng[1], Yuxia Li[1]
[1]School of electrical engineering and electronic information, Xihua University
[2]Department of physics, Fribourg University, Switzerland
Chengdu, China
[*]e-mail: 350648180@qq.com

***Abstract*—Because the K-Means algorithm is easy to fall into the local optimum and the Cuckoo search (CS) algorithm is affected by the step size, this paper proposes a K-Means clustering algorithm based on improved cuckoo search (ICS-Kmeans). The algorithm is compared with the original K-means, the Kmeans algorithm based on particle swarm optimization (PSO-Kmeans) and the K-Means algorithm based on the cuckoo search (CS-Kmeans). The experimental results show that the proposed algorithm can obtain better clustering effect, faster convergence rate and better accuracy rate through the experimental test on the UCI standard data set. The algorithm is also applied to the clustering of the characteristic parameters of the heart sound MFCC. The results show that a better clustering center can be obtained, the algorithm converges fast.**

Keywords-K-Means;cuckoo search;heart sound characteristics(key words)

I. INTRODUCTION

Data mining is a new type of interdisciplinary, covering statistical techniques, artificial intelligence technology, machine learning technology and database technology. Clustering analysis as one of the methods of data mining is an unsupervised learning problem, which deals with collection of unlabeled data to find a structure. In other words, clustering is the process of organizing similar objects into groups [1]. K-means algorithm is a clustering algorithm based on partition. Because of its simplicity and efficiency, it has become one of the most widely used clustering algorithms [2]. However, there are two shortcomings in the original K-means algorithm: firstly, the difference between each clustering result is greatly affected by the initial class center; secondly, it is easy to fall into the local optimal solution [3]. These deficiencies affect the clustering effect of K-means. Therefore, using the original K-means to cluster the parameters can not obtain the optimal clustering center. In recent years, K-means has been improved by optimizing algorithms, such as K-means based on genetic algorithm [4-8] and K-means algorithm based on particle swarm [9-12]. The effectiveness of a recent technique called cuckoo search (CS) for multi-modal design applications [13–15], and its superiority in benchmark comparisons [16, 17] against PSO and GA makes it an intelligent choice. CS is a search method that imitates obligate brood parasitism of some female cuckoo species specializing in mimicking then color and pattern of few chosen host birds. The parasitic cuckoo often chooses a nest where the host has just laid its own eggs so that when the firstly cuckoo chick hatches, it evicts the host eggs out of the nest to increase its own food share. Specifically, from an optimization standpoint, CS (i) guarantees global convergence, (ii) has local and global search capabilities controlled via a switching parameter (pa), and (iii) uses Levy flights rather than standard random walks to scan the design space more efficiently than the simple Gaussian process [15,18]. In addition, the CS algorithm has the advantages of simple structure, few input parameters, easy realization, random search path and optimization ability is strong. But the CS algorithm search results largely affected by the step factor, when the step size is set too high, will lead to low search accuracy, step length setting is too small, will lead to slow convergence speed[19].

Because the precision of CS algorithm is limited by the step size. In this paper, improved Cuckoo search (ICS) algorithm by combining the iteration times with the step size factor. Then, the ICS algorithm is combined with the K-means clustering algorithm, the K-means clustering algorithm is optimized by changing the search step size factor to obtain the optimal clustering result, and the optimal clustering results are obtained. We selected four sets of data in the UCI data set to test the improved algorithm (ICS-Kmeans), finally, the algorithm also was applied to cluster the MFCC characteristic parameters. Clustering effect is good, can help doctors diagnose heart disease.

II. METHODS AND INPLEMENTATION

A. K-means Clustering Algorithm

The K-mean algorithm is a classifying algorithm based on partition. By assigning a k value, n samples are divided into k clustering subsets, which leads to the high similarity within a cluster, and different clustering objects in lower similarity. The similarity of clustering is the mean measure of samples in subsets. The evaluation function of the commonly used metric clustering results is the square sum function of the error. The formula is as defined in definition (1):

978-1-5386-4794-3/18/$31.00 ©2018 IEEE

$$J = \sum_{i=1}^{k} \sum_{j=1}^{c_i} \| x_i^{(j)} - m_i \| \quad (1)$$

Where x_i^j is the jth sample of the i-th cluster center, m_i representing the i-th cluster center, J is the sum of the squares of errors in all the samples in the data set and the clustering center. When J is smaller, the clustering effect is better.

The basic flow of the K-means algorithm is ad follows:

a) Randomly select k samples from the data set as the initial cluster center.

b) Calculate the Euclidean distance between each sample and the K centers, and divide the sample into its nearest subset.

c) Calculate the mean of each subset as a new clustering center.

d) Generate a new cluster center, return to step b, otherwise the algorithm ends.

B. Cuckoo Search Algorithm

Cuckoo search algorithm (Cuckoo Search, CS) is a heuristic algorithm. In 2009, Yang and Deb proposed Cuckoo Search (CS) Algorithm [20].The CS algorithm effectively solves the optimization problem by simulating the parasitic parenting and Levy flight of the cuckoo. Parasitization refers to the cuckoo does not nest during breeding, but laid its own eggs in other nest, with other birds to reproduce. The cuckoo will find hatching and breeding birds which is similar to their ownself [21], and quickly spawn eggs while the bird is out, allowing the bird to replace the offspring. Levy flight is a random walk, this random walk by generating a certain length of the long and shorter steps to balance the local and global optimization.

In order to simplify the process of cuckoo parasitism in nature, the CS algorithm idealizes the process into the following three rules:

1．Each cuckoo only has one egg at a time and chooses a parasitic bird nest for hatching by random walk.

2．In the selected parasitic bird nest, only the best nest can be retained to the next generation.

3．The number of nests that may be parasitic is fixed, the nest owner can find the probability of foreign eggs is found once the foreign eggs will choose to discard the egg or re-nest.

In the above three idealized rules, the search for a new bird's nest location path is as follows:

$$x_i^{(t+1)} = x_i^{(t)} + \alpha \oplus Levy(\lambda); i = 1,2,...,n \quad (2)$$

$x_i^{(t)}$ stands for the ith bird's nest position in the t generation, α is the step size control, $\alpha > 0$, usually, $\alpha = 1$. $Levy(\lambda)$ is Levi's random search path, its expression is as follows:

$$Levy\ (\lambda) = t^{-\lambda}; 1 < \lambda < 3 \quad (3)$$

After the new solution is generated, some solutions are discarded according to a certain probability of discovery, and then the corresponding new solution is generated by the way of random walks, and iteration is completed. The CS algorithm flows as follows:

a) Initialize and set the corresponding parameters of the algorithm.

b) Calculate the fitness value of each nest to find the best nest.

c) Keep the optimal bird's nest and update the other bird's nest according to formula (2).

d) The current fitness value of the nest and the previous generation of nest fitness comparison, if better, then the alternative.

e) Generate random numbers (R) and compare the probability of discovery (P_a), if $R > P_a$, the random walk through the preferred way to generate a new bird's nest, or else retain the nest.

f) If the set stop condition is not reached, step b is returned.

g) The global optimal solution is used as the result output.

C. K - means Based on Improved Cuckoo Algorithm

The original cuckoo algorithm is influenced by step size A and probability of discovery P, and the step size and discovery probability control the accuracy of CS algorithm global and local search, which has great influence on the optimization effect of algorithm. The step size and discovery probability of the CS algorithm are set to a fixed value at initialization and will not change in subsequent iterations. When the step size is set too large, reducing the search accuracy, easy convergence, step length is too small, reducing the search speed, easy to fall into the local optimal. In this paper, the algorithm combines the step size with the number of iterations, sets a longer step length at the beginning of the iteration, and then, as the iteration progresses, the step size is reduced, and the algorithm has a large step in the iteration. Can achieve global optimization, and speed up the iterative speed and in the latter part of the algorithm iteration, with a smaller step size, improve search accuracy, to achieve local optimization. The improved formula is shown in (4).

$$\alpha_i = a_{\max} * \frac{1}{\left(\frac{a_{\max}}{a_{\min}}\right)^{\frac{t}{T}}} * \mathrm{ran}_i * 0.01 \quad (4)$$

$a_{\max}, a_{\min}$ is the maximum step size and the minimum step size, respectively. T is the total number of iterations. t represents the current iteration number. ran_i is the scope of the ith dimension of the data set.

Because of the shortage of the original K-means clustering algorithm itself. In this paper, improve the K-means algorithm by optimizing the global searching ability of the cuckoo algorithm to obtain the optimal solution.

Suppose that the original data needs to be clustered into k classes, and each sample has D dimensional features, and the K set of D dimension data is chosen as a bird nest solution. That is, the location of each bird's nest is the $k \times d$ dimension matrix. In K-means, the square error and function are used as the basis of the clustering result, so the formula (1) is used as the fitness function of the algorithm.

The K-means algorithm based on the improved CS algorithm is as follows:

a) Given the data set T and the number of clusters k, initialize the relevant parameters.

b) The clustering of a given sample is calculated and the fitness function value of the nest is calculated by formula (1).

c) Keep the optimal bird's nest and update the other bird's nest by formula (4).

d) The new nest will be clustered again and the fitness function value will be calculated. The result will be compared with the fitness value of the previous generation.

e) According to the probability of discovery to abandon the nest, randomly generate a new solution.

f) If the algorithm termination condition is not reached, return to step 2 to continue, otherwise, the optimal solution is output.

D. Experimental Results and Analysis

In order to verify the effectiveness of the proposed algorithm, we selected four sets of data from the UCI data set (http://archive.ics.uci.edu/ml/datasets.html)to test. And the algorithm of our proposed algorithm is compared with the original Kmeans algorithm, PSO-based Kmeans algorithm, CS-based Kmeans algorithm. Select the Iris, Wine, Seeds, Haberman four data sets as experimental test data, the basic information as shown in Table I:

TABLE I. DATA SET INFORMATION

Data set name	Data number	Feature number	Classification number
Iris	150	4	3
Wine	178	13	3
Seeds	210	7	3
Haberman	306	2	2

The simulation environment for the experiment is Windows 10 operating system, the simulation software is Matlab 2016. In addition to the traditional kmeans algorithm, the remaining algorithms use a fixed population size value of 20, the number of iterations 200 times. On this basis, the above four algorithms were calculated on the four data sets of fitness values, as shown in Table II:

TABLE II. EXPERIMENTAL RESULTS ON 4 DATASETS

Data Set	Objective function value	Kmeans	PSO_Kmeans	CS_Kmeans	ICS-Kmeans
Iris	Best	78.9408	79.6912	78.9483	78.9408
	Worst	142.8592	91.6121	78.9601	78.9408
	Average	112.9553	84.2511	78.9523	78.9408
Wine	Best	2370689	2374759	2370695	2370689
	Worst	2647028	2386796	2370745	2370693
	Average	2512257	2377370	2370721	2370691
Seeds	Best	587.3	587.3	587.63	587.31
	Worst	588.7	587.7	587.79	587.31
	Average	588.1	587.5	587.72	587.31
Haberman	Best	30507.02	30533.2	30509	30507.02
	Worst	43425.3	30722.8	30519.4	30507.02
	Average	35332.5	30617.3	30513.7	30507.02

In Table II, list the best, worst, and average fitness values for the four algorithms on each of the four datasets. We use the squared sum function of the error to evaluate the quality of the clustering results. It is clear that the smaller the value of this fitness values, the better the effect of clustering. It can be seen from the experimental results that the fitness values of this algorithm are superior to the other three algorithms, both in low-dimensional Iris, mid-dimensional Seeds, high-dimensional Wine and Haberman datasets, and the optimization ability is stable, each run result tends to the optimal solution.

Further take Wine, Haberman data set, the performance of the three algorithms were compared, get the results shown in Fig. 1and Fig. 2.

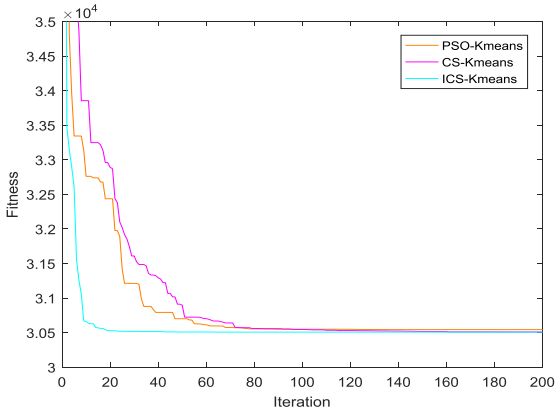

Figure 1. the convergence curves of the three algorithms on the Haberman dataset

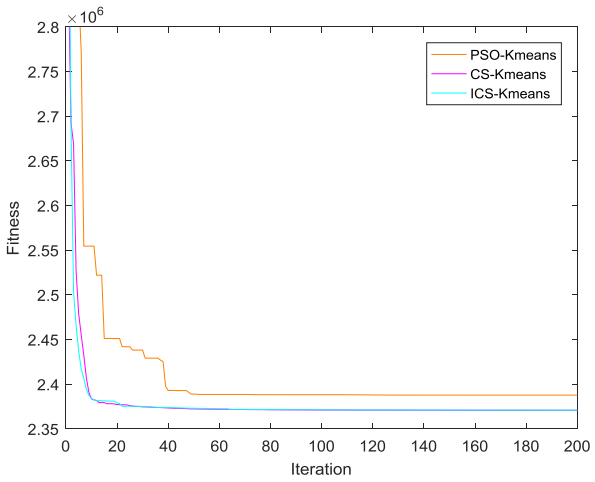

Figure 2. the convergence curves of the three algorithms on the Wine dataset

It can be seen from Fig. 1and Fig. 2 that the algorithm can converge to the optimal solution in the iteration of 10 times, which can converge to the optimal solution faster than that of PSO-Kmeans and CS-Kmeans algorithm, while CS-kmeans Since the step size is initially set to length, it can converge quickly in the Wine data set and can not converge quickly near the optimal solution on the Haberman dataset. In this paper, we combine the step size with the number of iterations to achieve fast optimization and accelerate convergence in the early stage of the algorithm. With the increase of the number of iterations, the step size is reduced and the local optimization is achieved. Therefore, this algorithm can find the optimal solution faster than the contrast algorithm.

Due to the instability of the original K-means clustering effect, the accuracy of the calculator can not be verified accurately, only the accuracy of the three algorithms is verified on the data set. The experimental results are shown in Table III:

TABLE III. COMPARISON OF CLUSTERING ACCURACY OF THREE ALGORITHMS

	Iris	**Seed**	**Wine**	**Haberman**
PSO_kmeans	88.7	90	70.8	51.4
CS_Kmeans	89.3	90	72.1	52.3
ICS- Kmeans	89.3	90	72.5	52.6

It can be seen from Table III that the improved K-means algorithm is relatively accurate, mainly because the improved K-means algorithm has a certain global optimization ability, improve the accuracy of clustering. The CS-Kmeans and the algorithm converge to the optimal solution and obtain the same accuracy rate in the low-dimensional Iris data set and the mid-dimensional Seed dataset. On the high-dimensional Wine data set and the Haberman data set, the algorithm has better accuracy than the contrast algorithm, mainly from the algorithm to balance the size of step in the early and late stages of the algorithm. .Thus, our improved ICS-kmeans algorithm. It is not limited by the step size factor, and it is further optimized on the basis of CS-Kmeans, which improves the accuracy of the algorithm.

In order to verify the clustering effect of the proposed algorithm on the parameters of the heart sound, select a heart sound characteristic parameter data set D, the data set D contains 255 sample data, each data is a 13-dimensional vector. Table III shows the results of the four algorithms for the data set.

TABLE IV. TEST RESULTS OF HEART SOUND CHARACTERISTIC PARAMETER DATA

Objective function value	**PSO_kmeans**	**CS-Kmeans**	**ICS- Kmeans**
Best	18.6953	18.6283	18.1787
Worst	20.5647	19.2077	18.2152
Average	19.2213	18.7362	18.1944

By comparing the data in Table IV, the proposed algorithm has better searching ability and clustering effect than PSO-Kmeans algorithm and CS-Kmeans algorithm in clustering of vocal MFCC characteristic parameters. Therefore, the algorithm can obtain the optimal clustering center, and the clustering center can be used as a codebook to describe the heart sound characteristics, which can help improve the accuracy of heart sound recognition.

III. CONCLUSION

In this paper, a K-Means clustering algorithm based on adaptive step-size cuckoo search is proposed, and the search precision and convergence speed of the algorithm are proposed before and after the algorithm is based on the shortcomings of K-Means algorithm and the influence of CS algorithm on the step size. The experimental results show that the algorithm has better clustering effect, faster convergence rate and better accuracy than the original K-means, PSO-Kmeans clustering algorithm and CS-Kmeans clustering algorithm. Applied to the heart sound characteristic parameter, the optimal clustering center can be obtained.

In general, improved algorithm contrasts with the other three algorithms, both on the UCI dataset and on the MFCC characteristic parameters of the heart sound, showing better and stable clustering effects, faster convergence rates, and better accuracy, which effectively improves the shortcomings of K-means algorithm.

ACKNOWLEDGMENT

This work is supported by the ChunHui plan project of Ministry of Education, China (Grant No. z2011089), Graduate Innovation Fund of Xihua University, China, (Grant No. ycjj2017061).

REFERENCES

[1] P. Berkhin. A Survey of Clustering Data Mining Techniques. Springer Link. 2006; 25-71.

[2] Anil K. Jain. Data clustering: 50 years beyond K-means. Pattern Recognition Letters. 2010; 31(8):651-666.

[3] Simon Fong, Suash Deb. Towards Enhancement of Performance of K-Means Clustering Using Nature-Inspired Optimization Algorithms. The Scientific World Journal. 2014;2014(2014)564829.

[4] LAI Yu-xia1, LIU Jian-ping1, YANG Guo-xing, K-Means Clustering Analysis Based on Genetic Algorithm, Computer Engineering . 2008; 20: 200-202.

[5] D.-X. Chang, X.-D. Zhang, C.-W. Zheng. A genetic algorithm with gene rearrangement for K-means clustering. Pattern Recognition. 2009: 42 (7):1210–1222.

[6] Swee. Chuan Tan, Kai. Ming Ting, Shyh. Wei Teng, A general stochastic clustering method for automatic cluster discovery. Pattern Recognition. 2011; 44 (10-11): 2786–2799.

[7] J. Xiao, Y. Yan, J. Zhang, Y. Tang. A quantum-inspired genetic algorithm for k-means clustering, Expert Syst. Appl. 2010; 37 (7): 4966–4973.

[8] Md Anisur Rahman, Md Zahidul Islam. A hybrid clustering technique combining a novel genetic algorithm with K-Means. Knowledeg-based Systems. 2014; 71: 345-365.

[9] Taher Niknam, Babak Amiri. An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis. Applied Soft Computing. 2010; 10 (1):183-197.

[10] Chen CY, Fun Y. Particle swarm optimization algorithm and its application to clustering analysis, Proceedings of 17th conference on electrical power distribution networks (EPDC), (2012) 789–794.

[11] Chuang LY, Hsiao CJ, Yang CH. Chaotic particle swarm optimization for data clustering. Expert Syst Appl. 2011;38 (12): 14555–14563.

[12] Ahmed AAE, Rodrigo AC, Stan M. A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data. Artif Intell Rev. 2015; 44:23–45.

[13] J.Z. Wang, H. Jiang, Y.J. Wu, Y. Dong. Forecasting solar radiation using an optimized hybrid model by Cuckoo Search algorithm. Energy2015; 81 (1) :627–644.

[14] X.S. Yang, S. Deb. Cuckoo search: recent advance and applications. Neural Comput. & Applic. 2014;24 (1) :169–174.

[15] M. Jamil, H.J. Zepernick, X.S. Yang. Levy Flight Based Cuckoo Search Algorithm for ´ Synthesizing Cross-Ambiguity Functions. IEEE Military Communications Conference (Milcom), San Diego, CA, 2013; 823–828.

[16] L.D. Coelho, C.E. Klein, S.L. Sabat, V.C. Mariani. Optimal chiller loading for energy conservation using a new differential cuckoo search approach. Energy. 2014; 75 (1) :237–243.

[17] A. Natarajan, S. Subramanian, K. Premalatha. A comparative study of cuckoo search and bat algorithm for Bloom filter optimisation in spam filtering. Int. J. Bio-Inspir. Comp. 2012;4 (2): 89–99.

[18] X.-S. Yang, Nature-inspired Optimization Algorithm, first ed. Elsevier, MA, USA, 2014.

[19] YANG X-S,DEB S. Cuckoo search:recent advances and applications, Neural Computing and Applications ,2014,24(1):169-174.

[20] YANG X-S,DEB S. Cuckoo search via Levy flights[C]//Nature & Biologically Inspired Computing,2009.NaBIC 2009.World Congress n.Coimbatore:IEEE,2009:210-214.

[21] Liyu, Mliang. New Meta-heuristic Cuckoo Search Algorithm. Systems engineering. 2012 (08),64-69.