Clustering Using Improved Cuckoo Search Algorithm

Jie Zhao[1], Xiujuan Lei[1,2], Zhenqiang Wu[1], and Ying Tan[2]

[1] School of Computer Science, Shaanxi Normal University, Xi'an, 710062, China
[2] School of Electronics Engineering and Computer Science, Peking University, Beijing, 100871, China
xjlei168@163.com

Abstract. Cuckoo search (CS) is one of the new swarm intelligence optimization algorithms inspired by the obligate brood parasitic behavior of cuckoo, which used the idea of Lévy flights. But the convergence and stability of the algorithm is not ideal due to the heavy-tail property of Lévy flights. Therefore an improved cuckoo search (ICS) algorithm for clustering is proposed, in which the movement and randomization of the cuckoo is modified. The simulation results of ICS clustering method on UCI benchmark data sets compared with other different clustering algorithms show that the new algorithm is feasible and efficient in data clustering, and the stability and convergence speed both get improved obviously.

Keywords: Clustering, cuckoo search, Lévy flights, swarm intelligence optimization algorithm.

1 Introduction

Clustering is the process of separating similar objects or multi-dimensional data vectors into a number of clusters or groups. It is an unsupervised problem. Clustering techniques have been used successfully in data analysis, image analysis, data mining and other fields of science and engineering [1].

Many algorithms have been developed for clustering. The traditional clustering methods can be classified into four categories: partitioning methods, hierarchical methods, density-based methods and grid-based methods [2].

Swarm intelligence optimization algorithm such as genetic algorithms (GA) [3], ant colony optimization [4], particle swarm optimization (PSO) [5, 6], artificial bee colony (ABC) [7, 8], bacteria foraging optimization algorithm (BFO) [9], firefly algorithm (FA) [10] has been widely used in the clustering in recent years. Cuckoo Search (CS) algorithm is a new intelligence optimization algorithm which has been successfully applied to the global optimization problem [11], economic dispatch [12], clustering [13-16] and other fields [17]. However, the cuckoo search clustering algorithm has several drawbacks such as slow convergence speed and vibration of the convergence.

Y. Tan et al. (Eds.): ICSI 2014, Part I, LNCS 8794, pp. 479–488, 2014.

In this paper, we propose an improved cuckoo search (ICS) algorithm for clustering, in which the movement of cuckoo and random disturbance was modified to find optimal cluster center. The algorithm was tested on four UCI benchmark datasets, and its performance was compared respectively with K-means, PSO, GA, FA and CS clustering algorithm. The simulation results illustrated that this algorithm not only own higher convergence performance but also can find out the optimal solution than the other algorithms.

2 Cuckoo Search Algorithm

Cuckoo search algorithm is a novel metaheuristic optimization algorithm developed by Xin-she Yang and Suash Deb in 2009 [18], which is based on the obligate brood parasitic behaviour of some cuckoo species in combination with the Lévy flights behavior of some birds and fruit flies.

2.1 Cuckoo Brood Parasitic Behaviour

Cuckoos are fascinating birds not only because of the beautiful sounds they can make, but also because of their aggressive reproduction strategy they share [19]. Quite a number of species engage the obligate brood parasitism by laying their eggs in the nests of other host birds, which may be different species. They may remove others' eggs to increase the hatching probability of their own eggs [20]. If a host bird discovers that the eggs are not their own' eggs, they will either throw these alien eggs away or simply abandon its nest and build a new nest elsewhere.

Studies also indicated that the cuckoo eggs hatch slightly earlier than their host eggs. Once the first cuckoo chick is hatched, the first instinct action it will take is to evict the host eggs by blindly propelling the eggs out of the host, which increases the cuckoo chick's share of food provided by its host bird. In addition, a cuckoo chick can also mimic the call of host chicks to gain access to more feeding opportunity.

2.2 Lévy Flights

In nature, animals search for food in a random or quasi-random manner. Various studies have shown that the flight behavior of many animals and insects demonstrates the typical characteristics of Lévy flights [21]. Lévy flights comprise sequences of randomly orientated straight-line movements. Frequently occurring but relatively short straight-line movement randomly alternate with more occasionally occurring longer movements, which in turn are punctuated by even rarer, even longer movements, and so on with this pattern repeated at all scales. As a consequence, the straight-line movements have no characteristic scales, and Lévy flights are said to be scale-free, the distribution of straight-line movement lengths have a power-law tail [22]. Fig. 1 shows the path of Lévy flights of 60 steps starting from (0, 0).

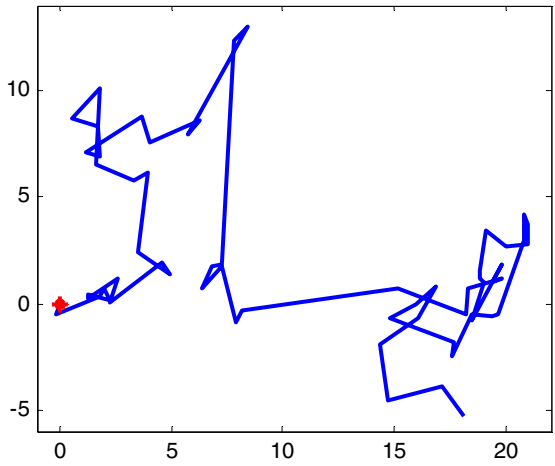

Fig. 1. Lévy flights in consecutive 60 steps starting at the origin (0, 0) which marked with " * "

2.3 Cuckoo Search Algorithm

Here we simply describe the cuckoo search algorithm as follows which contain three idealized rules [19]:

(1) Each cuckoo lays one egg at a time, and dumps its egg in a randomly chosen nest;

(2) The best nest with high quality of eggs will carry over to the next generations;

(3) The number of available host nests is fixed, and then the egg laid by a cuckoo is discovered by the host bird with a probability $p_a \in [0,1]$. In this case, the host bird can either throw the eggs away or abandon the nest, and build a completely new nest. For simplicity, this rule can be approximated by the fraction p_a of the n nests are replaced by new nests (with new random solutions).

For a maximization problem, the quality of fitness of a solution can be proportional to the objective function. Other forms of fitness can be defined in a similar way to the fitness function in genetic algorithm and other optimization algorithms [23].

Based on these three rules, the basic steps of the cuckoo search can be summarized as the pseudo code as Table 1 [19]:

Table 1. Pseudo code of the cuckoo search

```
begin
  Objective function f(x), x = (x_1, ..., x_d)^T
  Generate initial population of n host nests x_i (i = 1,2, ..., n)
  while (t < MaxGeneration) or (stop criterion)
    Get a cuckoo randomly by Lévy flights
      evaluate its quality/fitness F_i
    Choose a nest among n (say, j) randomly
```

```
    if (F_i > F_j),
       replace j by the new solution;
    end
    A fraction (p_a) of worse nests are abandoned and new ones are built;
    Keep the best solutions (or nests with quality solutions);
    Rank the solutions and find the current best
  end while
  Postprocess results and visualization
end
```

When generating new solutions x (t+1) for a cuckoo i, a Lévy flight is performed using the following equation:

$$x_i^{(t+1)} = x_i^{(t)} + \alpha \oplus \text{Lévy}(\beta), \ (i = 1,2,...,n) \ . \tag{1}$$

where $\alpha > 0$ is the step size which should be related to the scales of the problem of interests. The product $\oplus$ means entry-wise multiplications. The Lévy flight essentially provides a random walk while the random step length is drawn from a Lévy distribution which has an infinite variance with an infinite mean [19].

$$\text{Lévy}(\beta) \sim u = t^{-1-\beta}, \quad (0 < \beta \le 2) \ . \tag{2}$$

Mantegna puts forward a most efficient and yet straightforward ways to calculate Lévy distribution [18, 24].

$$\text{Lévy}(\beta) \sim \frac{u}{|v|^{1/\beta}} \ . \tag{3}$$

$$u \sim N(0, \ \sigma_u^2) \ , \quad v \sim N(0, \ \sigma_v^2) \ . \tag{4}$$

$$\sigma_u = \left\{ \frac{\Gamma(1+\beta)\sin(\pi\beta/2)}{\Gamma[(1+\beta)/2]\beta 2^{(\beta-1)/2}} \right\}^{1/\beta} , \qquad \sigma_v = 1 \ . \tag{5}$$

In CS Algorithm, the worst nest is abandoned with a probability p_a and a new nest is built with random walks [19].

$$x_{worst}^{(t+1)} = x_{worst}^{(t)} + \alpha * rand() \ . \tag{6}$$

3 Clustering Using ICS Algorithm

3.1 The Clustering Criterion

Clustering is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters. Dissimilarities and similarities are assessed based on the attribute values describing the objects and often involve distance measures. The most popular distance measure is Euclidean distance [1].

Let $i=(x_{i1},x_{i2},...,x_{ip})$ and $j=(x_{j1},x_{j2},...,x_{jp})$ be two objects described by p numeric attributes, the Euclidean distance between object i and j is define as:

$$d(i,j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\cdots+(x_{ip}-x_{jp})^2} \quad . \tag{7}$$

For a given N objects the clustering problem is to minimize the sum of squared Euclidean distances between each object and allocate each object to one of k cluster centers [1]. The main goal of the clustering method is to find the centers of the clusters by minimizing the objective function. The clustering objective function is the sum of Euclidean distances of the objects to their centers as given in Eq. (8) [2]

$$J_c = \sum_{k=1}^{m} \sum_{X_i \in C_k} d(X_i, Z_k) \quad . \tag{8}$$

where m denotes the number of clusters, C_k denotes the *kth* cluster, $d(X_i,Z_k)$ denotes the Euclidean distance between object X_i and cluster center Z_k.

3.2 Clustering Using ICS Algorithm

Some researchers [13-16] have been designed CS for clustering. The Lévy flight is more efficient because the step length is heavy-tailed and any large step is possible, which makes the whole search space to be covered [14]. However, Lévy flight often leads to slow convergence rate and vibration when clustering using CS. We propose ICS algorithm for data clustering, in the algorithm, each egg in a nest represents a cluster center, the cuckoo searches for a new nest in line with Eq. (9)

$$x_i^{(t+1)}=x_i^{(t)}+\alpha\oplus \text{Lévy}(\beta)+w*rand*(Z_{best}-x_i^{(t)}),\ (i=1,2,...,n) \quad . \tag{9}$$

where *rand* is random number, w denotes disturbance constant, Z_{best} denotes the cluster center of the best cluster.

The ICS algorithm for data clustering is as the following steps:

Step 1: Generate initial population of n host nests randomly, the host nests position denote the cluster centers. Initialize the iterations *iter*, maximum iteration *maxiter* and cluster number *nc*;

Step 2: Clustering and calculate the clustering objective function *Fold* using Eq. (8) to find the best nest *bestnest*;

Step 3: Generate n-1 new nests using Eq. (9) except the *bestnest*, clustering and calculate the clustering objective function *Fnew*,

Step 4: Compare *Fnew* with *Fold*, if *Fnew* < *Fold*, replace the old nests by the new ones;

Step 5: A fraction (p_a) of worst nests are abandoned and new ones are built using Eq. (6);

Step 6: Find the best solution, set *iter* = *iter* + 1;

Step 7: If *iter* < = *maxiter*, goto *Step* 3, otherwise output the clustering result.

4 Implementation and Results

In order to test the accuracy and the efficiency of ICS data clustering algorithm, experiments have been performed on four datasets including Iris, Glass, Wine and Sonar selected from standard set UCI [25]. All algorithms are implemented in Matlab R2011b and executed on a Pentium dual-core processor 3.10GHz PC with 4G RAM. The parameter values used in our algorithm are n=15, p_a=0.25, α =0.01 and w=*0.06*.

4.1 Data Set Description

The four clustering data sets Iris, Glass, Wine and Sonar are well-known and popular-used benchmark datasets. Table 2 shows the characteristics of the datasets.

Table 2. Summary of the characteristics of the considered data sets

Name of data set	Number of classes	Number of features	Size of data set (size of classes)
Iris	3	4	150(50,50,50)
Glass	6	9	214(29,76,70,17,13,9)
Wine	3	13	178(59,71,48)
Sonar	2	60	208(111,97)

4.2 Analysis of Algorithm Convergence

To evaluate the convergence performance, we have compared the ICS algorithm with traditional K-means, PSO and CS clustering algorithm on Iris data set.

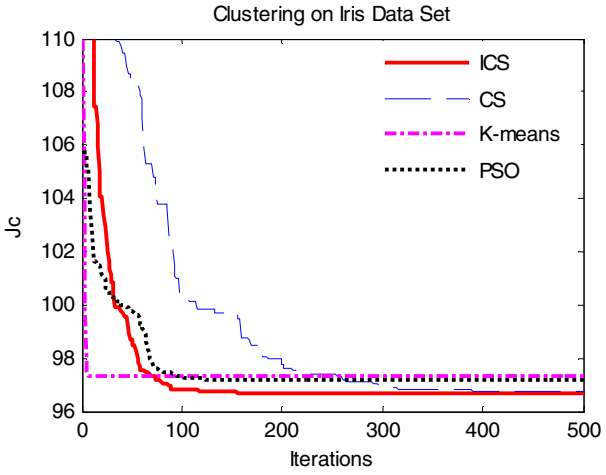

Fig. 2. Convergence curve of clustering on Iris data set

Fig. 2 illustrates that the ICS clustering algorithm has achieved the best convergence performance in the terms of the clustering objective function. K-means algorithm is easily to fall into local optimum due to the premature convergence [26]. The disadvantage of CS clustering algorithm is the slow convergence rate and vibration of the convergence; the convergence rate is insufficient when it searches global optimum.

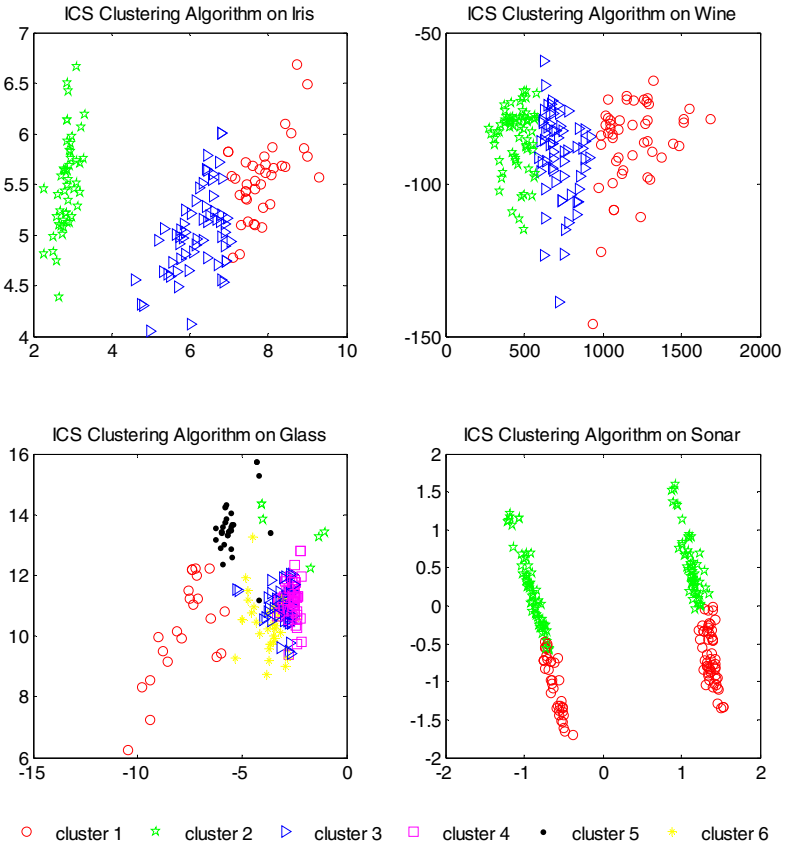

Fig. 3. Results of ICS clustering algorithms on Iris, Wine, Glass and Sonar data sets

The results of ICS clustering algorithms on Iris, Wine, Glass and Sonar data sets is given in Fig. 3 which can make it visualized clearly. Principal component analysis was utilized to reduce the dimensionality of the data set. It can be seen from Fig. 3 clearly that ICS clustering algorithm possess better effect on Iris data set and Wine data set. The Glass data set has six classes and the Sonar data set has sixty features, so the higher data complexity leads to the larger clustering error.

4.3 Clustering Results

The best clustering objective function, mean clustering objective function and clustering error for K-means, PSO, GA, FA, CS and the proposed algorithm of ICS on different data set including Iris, Glass, Wine and Sonar are shown in Table 3. Experiments were repeated 30 times.

Table 3. Comparison of clustering results via the four algorithms

Data set	Algorithm	Best Jc	Mean Jc	Clustering Error
Iris	K-means [27]	97.32	102.57	16.05±10.10
	PSO [27]	97.10	102.26	10.64±4.50
	GA[7]	113.98	125.19	—
	FA	99.06	103.18	10.3±3.61
	CS	96.89	97.67	10.2±1.1
	ICS	96.66	96.68	9.6±0.6
Glass	K-means [27]	213.42	241.03	48.30±3.14
	PSO [27]	230.54	258.02	48.72±1.34
	CS	212.74	215.22	52.34±2.3
	ICS	210.95	213.84	43.93±1.89
Wine	K-means [27]	16555.68	17662.73	34.38±6.08
	PSO [27]	16307.16	16320.67	28.74±0.39
	GA[7]	16530.53	16530.53	—
	FA	16714.00	18070.59	31.46±3.45
	CS	16298.79	16309.24	29.21±1.34
	ICS	16295.67	16302.40	27.64±1.08
Sonar	K-means [27]	234.77	235.06	44.95±0.97
	PSO [27]	271.83	276.68	46.60±0.42
	FA	239.75	245.71	45.34±4.67
	CS	271.52	282.70	46.63±0.53
	ICS	232.20	238.58	44.23±0.24

As shown in Table 3, it is obvious that the ICS clustering algorithm could find the optimal clustering objective function value, and the mean clustering objective function value close to the best clustering objective function value, which illustrates the new algorithm has good stability. The results are exactly same as the phenomenon showed in Fig. 3.

5 Conclusion and Discussion

The ICS algorithm to solve clustering problems has been developed in this paper. To evaluate the performance of the ICS, it is compared with K-means, PSO and CS clustering algorithms on four well known UCI data sets. The experimental results indicated that the ICS clustering algorithm has best convergence performance,

stability and better clustering effect. In order to improve the obtained results, we plan to apply the proposed approach into other clustering areas as our future work.

Acknowledgement. This paper is supported by the National Natural Science Foundation of China (61100164, 61173190), Scientific Research Start-up Foundation for Returned Scholars, Ministry of Education of China ([2012]1707) and the Fundamental Research Funds for the Central Universities, Shaanxi Normal University (GK201402035, GK201302025).

References

1. Han, J.W., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers (2011)
2. Lei, X.J.: Swarm Intelligent Optimization Algorithms and their Applications. Science Press (2012)
3. Maulik, U., Bandyopadhyay, S.: Genetic Algorithm-based Clustering Technique. Pattern Recognition 33, 1455–1465 (2000)
4. Kao, Y., Cheng, K.: An ACO-based clustering algorithm. In: 5th International Workshop on Ant Colony Optimization and Swarm Intelligence, pp. 340–347 (2006)
5. Van Der Merwe, D.W., Engelbrecht, A.P.: Data Clustering Using Particle Swarm Optimization. In: Congress on Evolutionary Computation (CEC 2003), pp. 215–220 (2003)
6. Zhang, Q., Lei, X.J., Huang, X., Zhang, A.D.: An Improved Projection Pursuit Clustering Model and its Application Based on Quantum-behaved PSO. In: 2010 Sixth International Conference on Natural Computation (ICNC 2010), vol. 5, pp. 2581–2585 (2010)
7. Zhang, C.S., Ouyang, D.T., Ning, J.X.: An Artificial Bee Colony Approach for Clustering. Expert Systems with Applications 37, 4761–4767 (2010)
8. Lei, X.J., Tian, J.F., Ge, L., Zhang, A.D.: The Clustering Model and Algorithm of PPI Network Based on Propagating Mechanism of Artificial Bee Colony. Information Sciences 247, 21–39 (2013)
9. Lei, X.J., Wu, S., Ge, L., Zhang, A.D.: Clustering and Overlapping Modules Detection in PPI Network Based on IBFO. Proteomics 13, 278–290 (2013)
10. Senthilnath, J., Omkar, S.N., Mani, V.: Clustering Using Firefly Algorithm: Performance study. Swarm and Evolutionary Computation 1, 164–171 (2011)
11. Ghodrati, A., Lotfi, S.: A Hybrid CS/PSO Algorithm for Global Optimization. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ACIIDS 2012, Part III. LNCS, vol. 7198, pp. 89–98. Springer, Heidelberg (2012)
12. Basu, M., Chowdhury, A.: Cuckoo Search Algorithm for Economic Dispatch. Energy 60, 99–108 (2013)
13. Saida, I.B., Nadjet, K., Omar, B.: A New Algorithm for Data Clustering Based on Cuckoo Search Optimization. Genetic and Evolutionary Computing 238, 55–64 (2014)
14. Senthilnath, J., Das, V., Omkar, S.N., Mani, V.: Clustering Using Lévy Flight Cuckoo Search. In: Bansal, J.C., Singh, P., Deep, K., Pant, M., Nagar, A. (eds.) Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications, (BIC-TA 2012). AISC, vol. 202, pp. 65–75. Springer, Heidelberg (2013)
15. Goel, S., Sharma, A., Bedi, P.: Cuckoo Search Clustering Algorithm: A Novel Strategy of Biomimicry. In: World Congress on Information and Communication Technologies, pp. 916–926 (2011)

16. Manikandan, P., Selvarajan, S.: Data Clustering Using Cuckoo Search Algorithm (CSA). In: Babu, B.V., Nagar, A., Deep, K., Pant, M., Bansal, J.C., Ray, K., Gupta, U. (eds.) Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012). AISC, vol. 236, pp. 1275–1283. Springer, Heidelberg (2014)
17. Bulatović, R.R., Đorđević, S.R., Đorđević, V.S.: Cuckoo Search Algorithm: A Metaheuristic Approach to Solving the Problem of Optimum Synthesis of a Six-bar Double Dwell Linkage. Mechanism and Machine Theory 61, 1–13 (2013)
18. Yang, X.S.: Nature-Inspired Metaheuristic Algorithms, 2nd edn. Luniver Press (2010)
19. Yang, X.S., Deb, S.: Cuckoo Search via Lévy Flights. In: World Congress on Nature & Biologically Inspired Computind, pp. 210–214. IEEE Publications, USA (2009)
20. Payne, R.B., Sorenson, M.D., Klitz, K.: The Cuckoos. Oxford University Press (2005)
21. Valian, E., Mohanna, S., Tavakoli, S.: Improved Cuckoo Search Algorithm for Feedforward Neural Network Training. International Journal of Artificial Intelligence & Applications 2, 36–43 (2011)
22. Reynolds, A.M., Rhodes, C.J.: The Lévy Flight Paradigm: Random Search Patterns and Mechanisms. Concepts & Synthesis 90, 877–887 (2009)
23. Gandomi, A.H., Yang, X.S., Alavi, A.H.: Cuckoo Search Algorithm: a Metaheuristic Approach to Solve Structural Optimization Problems. Engineering with Computers 29, 17–35 (2013)
24. Mantegna, R.N.: Fast, Accurate Algorithm for Numerical Simulation of Lévy Stable Stochastic Processes. Physical Review E 49, 4677–4689 (1994)
25. UCI Machine Learning Repository, `http://archive.ics.uci.edu/ml/`
26. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: An Efficient k-Means Clustering Algorithm. In: Analysis and Implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 881–892. IEEE Press, New York (2002)
27. Hassanzadeh, T., Meybodi, M.R.: A New Hybrid Approach for Data Clustering using Firefly Algorithm and K-means. In: CSI International Symposium on Artificial Intelligence and Signal Processing, pp. 7–11. IEEE Press, New York (2012)