A New Algorithm for Data Clustering Based on Cuckoo Search Optimization

Ishak Boushaki Saida[1], Kamel Nadjet[2], and Bendjeghaba Omar[3]

[1] University M'hamed Bougara of Boumerdes (UMBB) and LRIA (USTHB), Algeria
saida_2005_compte@yahoo.fr
[2] University Farhat Abbes of Setif (UFAS) and LRIA (USTHB), Algeria
nkamel@usthb.dz
[3] LREEI, University M'hamed Bougara of Boumerdes (UMBB), Algeria
benomar75@yahoo.fr

Abstract. This paper presents a new algorithm for data clustering based on the cuckoo search optimization. Cuckoo search is generic and robust for many optimization problems and it has attractive features like easy implementation, stable convergence characteristic and good computational efficiency. The performance of the proposed algorithm was assessed on four different dataset from the UCI Machine Learning Repository and compared with well known and recent algorithms: K-means, particle swarm optimization, gravitational search algorithm, the big bang–big crunch algorithm and the black hole algorithm. The experimental results improve the power of the new method to achieve the best values for three data sets.

Keywords: Data Clustering, Cuckoo Search, Metaheuristic, Optimization.

1 Introduction

Clustering is an unsupervised classification technique of data mining [1] [2]. It divides a set of data into groups or clusters based on the similarity between the data objects, such that similar objects fall in the same cluster and different objects in different clusters.

Clustering is used in several applications like document clustering [3], image segmentation [4] and pattern recognition [5]. For each application we have to select and extract a set of features to represent the data objects and also we have to define measuring proximity between these data objects.

Many clustering methods have been proposed. They are classified into several major algorithms: hierarchical clustering, partitioning clustering, density based clustering and graph based clustering.

One of the most popular and famous partitioning algorithm is K-means because its efficiency and simplicity [6]. Unfortunately, the K-means algorithm suffers from two problems: It needs to define the number of clusters before starting and in addition, its performance depends strongly on the initial centroids and may get trapped in local optimal solutions.

J.-S. Pan et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 238,
DOI: 10.1007/978-3-319-01796-9_6,

Recently, nature inspired approaches have received increased attention from researchers dealing with data clustering problems [7].

To avoid the inconvenience of K-means, we propose in this paper to use a new metaheuristic approach. It is mainly based on the cuckoo search (CS) algorithm which was proposed by Xin-She Yang and Suash Deb in 2009 [8] [9]. Cuckoo search is based on the interesting breeding behaviour such as brood parasitism of certain species of cuckoos and typical characteristics of Lévy flights. The CS is generic and robust for many optimization problems [10] [11]. It is a population based and this algorithm overcomes the problem of local optimum to global one.

The efficiency of the proposed algorithm is tested on four different data sets issued from literature [12] and the obtained results are compared with some recent well known algorithms reported in [13].

The remaining of this paper is organized as follows: In section 2, related works is presented. In section 3, we present cluster analysis. In section 4, we describe the basics of cuckoo search algorithm. The proposed approach for data clustering is explained in section 5. Numerical experimentation and comparisons are provided in Section 6. Finally, conclusions and our future work are drawn in Section 7.

2 Related Works

Several metaheuristic were developed to overcome the disadvantage of K-means. Most of them are evolutional and population based. For instance the genetic algorithm is evolutionary population optimization based; it uses natural genetics and evolution: selection, mutation, and crossover [14]. It is still suffers from the difficulty of coding modelling. Also, the operation of crossover and mutation are too expensive. More over it needs much parameter to handle. The ant colony algorithm is another metaheuristic inspired from the behaviour of the real ants to find the shortest path between a food source and their nest [15] [16]. Particle Swarm Optimization (PSO) incorporates swarming behaviours observed in flocks of birds, schools of fish, or swarms of bees, and even human social behaviour, from which the idea is emerged [17][18]. Like the genetic algorithm, it needs much parameter to manipulate. A data clustering algorithm based on the gravitational search algorithm (GSA) was proposed in [19] [20]. It is based on the law of gravity and the notion of mass interactions. The Big Bang–Big Crunch (BB–BC) algorithm was also applied for resolving the problem of clustering [21]. It is an optimization method that is based on one of the theories of the evolution of the universe namely the Big Bang and Big Crunch theory. Another heuristic algorithm namely the black hole algorithm was defined to resolve the problem of clustering, which is inspired from the black hole phenomenon [13].

The new algorithm proposed in this paper is based on the cuckoo search optimization. In this metaheuristic no much parameters is used. We need only to define the worse nests probability which does not really affect in the results of clustering. More over, the research of the optimal solution is done by a mathematical function. In each generation we select the best solution and the next generation is

calculated by the cuckoo search function using the best solution. Thereby, we always convert to the optimal solution.

3 Cluster Analysis

The main goal of the clustering process is to group the most similar objects in the same cluster or group. Each object is defined by a set of attributes or measurements.

To determine the similar objects, we use the measure of similarity between them. Several similarity measures are defined in the literature. In this paper we use the Euclidean distance to calculate the similarity between the objects. It is the most popular metric done by the formula (1):

$$distance(o_i, o_j) = \left(\sum_{p=1}^{m} \left|o_{ip} - o_{jp}\right|^{\frac{1}{2}}\right)^2 \quad (1)$$

Where: m is the number of attributes and o_{ip} is the value of the attribute number p of the object number i (o_i).

The result of a clustering algorithm must be evaluated and validated. This is done by using validity indexes. They are classified into internal and external one [22]. In this paper we use the sum intra cluster (SSE) which is an internal validity index, and the error rate which is an external validity index. These indexes are defined by the formula (2) and (3).

$$\text{SSE} = \sum_{i=1}^{k} \sum_{j=1}^{n} W_{ij} * \sqrt{\sum_{p=1}^{m} \left(o_{jp} - c_{ip}\right)^2} \quad (2)$$

Where: $W_{ij} = 1$ if the object is in the cluster and 0 otherwise. k is the number of clusters, n is the number of objects, m is the number of attributes and c_{ip} is the value of the attribute number p of the centroid of the cluster number i (c_i).

$$ER = \frac{number\ of\ misplaced\ objects}{total\ of\ objects\ whithin\ dataset} * 100 \quad (3)$$

4 Basics of Cuckoo Search Algorithm

The Cuckoo search (CS) is a new metaheuristic optimisation algorithm, proposed by Xin-She Yang and Suash Deb [8] [9]. The algorithm is based on the obligate brood parasitic behaviour of some cuckoo species in combination with the Lévy flight behaviour of some birds and fruit flies. In fact, the algorithm has three particular idealized rules [8]:

1. Each cuckoo lays one egg at a time, and dumps its egg in randomly chosen nest;
2. The best nests with high quality of eggs will carry over to the next generations;

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

Based on these three rules, the basic steps of the Cuckoo Search (CS) can be summarized by the pseudo code shown in Figure 1.

Cuckoo Search via Lévy Flights
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
Post process results and visualization
End

Fig. 1. Pseudo code of the standard Cuckoo Search (CS) [8]

5 Clustering with Cuckoo Search

For solving the data clustering problem, the standard cuckoo search algorithm is adapted to reach the centroids of the clusters. For doing this, we suppose that we have n objects, and each object is defined by m attributes. In this work, the main goal of the CS is to find k centroids of clusters which minimize the formula (2). Knowing that the problem is multi-dimensional, the data set must be represented by a matrix (n, m), such as the row i corresponds to the object number.

In cuckoo search mechanism, the solutions are the nests and each nest is represented by a matrix with k rows and m columns, where, the matrix rows are the centroids of clusters.

We propose a cuckoo search algorithm for the data clustering throught the following steps:

1. Generate randomly the initial population of *nb_nest* host nests;
2. Calculate the fitness of these solutions and find the best solution;
   While (t < MaxGeneration) or (stop criterion);
3. Generate *nb_nest new* solutions with the cuckoo search;
4. Calculate the fitness of the new solutions;
5. Compare the new solutions with the old solutions, if the new solution is better than the old one, replace the old solution by the new one ;
6. Generate a fraction (p_a) of new solutions to replace the worse nests;
7. Compare these solutions with the old solutions. If the new solution is better than the old solution, replace the old solution by the new one;
8. Find the best solution;
   End while;
9. Print the best nest and fitness;

6 Implementation and Results

In order to test the validity and the efficiency of the proposed approach, we have selected four data sets from the literature [12]. We have used an internal and an external quality measure in order to evaluate and compare this method with the other ones cited in [13].

For the Internal quality measure, we consider the sum of intra-cluster distances represented by the formula (2). The goal is to minimize this function called fitness function. For the External quality measure, we calculate the error rate (ER), which represents the percentage of misplaced objects as given in formula (3). It is the same one used in [13], thereby we can compare the performance of the CS algorithm to the most recent algorithms reported in [13]: K-means, particle swarm optimization (PSO), gravitational search algorithm (GSA) the big bang–big crunch algorithm (BB-BC) and the black hole algorithm (BH).

We should note that, for all datasets the population size and the probability of worse nests were set to 20 and 0.25 respectively.

6.1 Iris Dataset

The Iris dataset contains 150 objects with four attributes. They are unscrewed into 3 classes of 50 instances, where each class represents a type of iris plant. The best obtained solution using the cuckoo search for the iris data is given in Table 1, where the row i represents the value of all the attributes of the centroid of the cluster number i. The variation graph of the fitness function according to the number of generations is represented in Figure 2.

Table 1. The best solution for Iris data by CS

Center 1	5.9347	2.7979	4.4179	1.4171
Center 2	6.7336	3.0664	5.6301	2.1055
Center 3	5.0130	3.4040	1.4710	0.2358

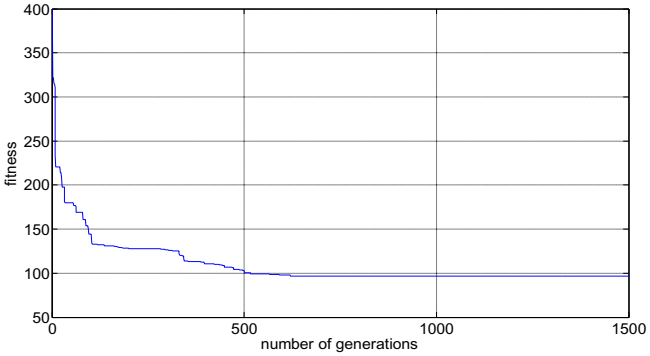

Fig. 2. Fitness function for Iris dataset

6.2 Wine Dataset

The Wine dataset describes the quality of wine from physicochemical properties. There are 178 instances with 13 features grouped into 3 classes.

The best obtained solution using the cuckoo search for the wine data is given in Table 2, and the variation graph of the fitness function according to the number of generations is represented in Figure 3.

Table 2. The best solution for Wine data by CS

Center 1	Center 2	Center 3
13.69227	12.48561	12.80205
1.82475	2.29206	2.52757
2.52063	2.42019	2.43197
16.89081	21.29222	19.61010
105.29640	92.54816	98.89915
2.81080	2.04322	2.10036
3.15359	1.73150	1.45264
0.30168	0.36522	0.45469
2.02348	1.42479	1.41032
5.73206	4.43186	5.75083
1.08947	1.02753	0.86257
3.13876	2.41258	2.19260
1137.21760	463.65681	687.03916

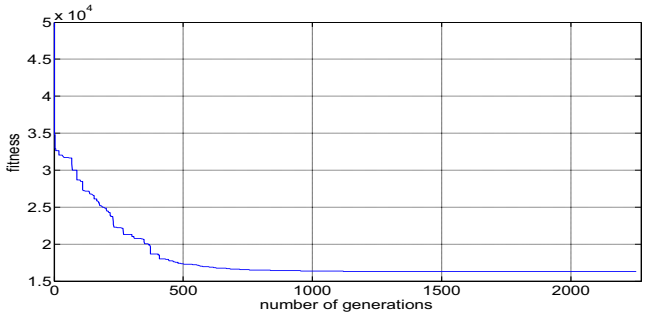

Fig. 3. Fitness function for Wine dataset

6.3 Cancer Dataset

The Cancer dataset represents the Wisconsin breast cancer databases. The dataset contains 683 instances with 9 features. Each instance has one of two possible classes: benign or malignant.

The best obtained solution using the cuckoo search for the cancer data is shown in Table 3, and the variation graph of the fitness function according to the number of generations is represented in Figure 4.

Table 3. The best solution for cancer data by CS

Center 1	2.88848	1.12802	1.20058	1.16359	1.99256	1.11893	2.00507	1.10059	1.03144
Center 2	7.11641	6.64365	6.62561	5.61432	5.24276	8.10514	6.07841	6.02254	2.32808

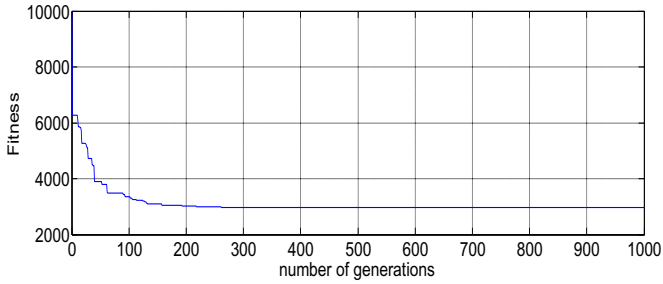

Fig. 4. Fitness function for Cancer dataset

6.4 Vowel Dataset

The Vowel dataset consists of 871 instances. Each point are represented by 3 features. These instances are grouped into 6 classes.

The best obtained solution using the cuckoo search for the Vowel data is given in Table 4, and the variation graph of the fitness function versus to the number of generations is represented in Figure 5.

Table 4. The best solution for Vowel data by CS

Center 1	360.48780	2290.36553	2976.77797
Center 2	437.21449	993.57547	2658.09454
Center 3	407.75777	1011.99989	2310.59089
Center 4	507.54437	1839.75796	2555.73930
Center 5	374.96411	2149.78838	2678.23073
Center 6	622.99723	1308.62969	2332.83028

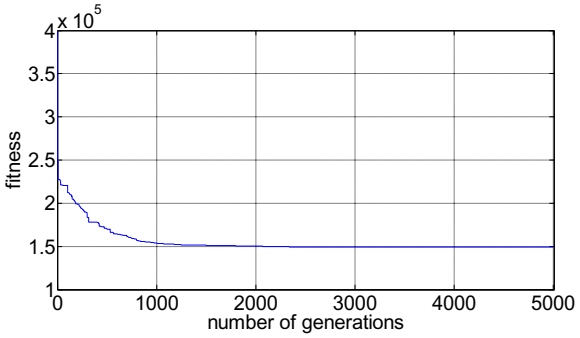

Fig. 5. Fitness function for Vowel dataset

For the considered dataset, the best value of fitness functions given by the cuckoo search are compared with those obtained by the different algorithms: K-means, particle swarm optimization (PSO), gravitational search algorithm (GSA), big bang–big crunch algorithm (BB-BC) and the black hole algorithm (BH). The comparison results are shown in Table 5.

From Table 5, it is obvious that the Cuckoo Search algorithm can reach very important results. In the case of Iris data, the value of the best fitness function obtained by the Cuckoo Search is 96.65564, which is better than all other ones. Also, for the Wine data the value of the best fitness function given by the Cuckoo Search is 16292.24388, which is significantly better than the all the other ones. As we can see in table 5, the value of the best fitness function achieved by the Cuckoo Search for the Cancer data is 2964.38839, which is the best one. However, the value of the best fitness function found by the Cuckoo Search for Vowel data is 148990.15884, which is the best one after the black hole algorithm (BH).

Table 6 compares the error rate obtained by the clustering with the CS on the four dataset with different clustering algorithms (K-means, PSO, GSA, BB-BC and BH). As shown in the Table 6, CS gives a minimum error rate for all the datasets except the Vowel data.

Table 5. Best fitness obtained by different algorithms on Iris, Wine, Cancer and Vowel dataset

Approach	Iris	Wine	Cancer	Vowel
K_means	97.32592	16,555.67942	2986.96134	149 394.80398
PSO	96.87935	16,304.48576	2974.48092	152 461.56473
GSA	96.68794	16,313.87620	2965.76394	151 317.56392
BB-BC	96.67648	16,298.67356	2964.38753	149 038.51683
BH	96.65589	16,293.41995	2964.38878	148 985.61373
CS	**96.65564**	**16292.24388**	**2964.38839**	**148990.15884**

Table 6. The error rate (ER) of clustering algorithm on the different dataset

Dataset	K-means	PSO	GSA	BB-BC	BH	CS
Iris	13,42	10,06	10,04	10,05	10,02	**10,01**
Wine	31,14	28,79	29,15	28,52	28,47	**27,07**
Cancer	4,39	3,79	3,74	3,70	3,70	**3,52**
Vowel	43,57	42,39	42,26	41,89	41,65	**42,45**

7 Conclusion

In this paper, we have presented a new approach for solving the data clustering problem. The approach is principally based on the cuckoo search algorithm. The proposed algorithm is applied to four different data sets. Simulation experiments show that the proposed approach gives better results compared to some other more frequently used clustering approaches. The cuckoo search algorithm is useful for solving the data clustering problem. Moreover it is easy to implement and it manipulates a few parameters. In order to improve the obtained results and as a future work, we plan to hybridize the proposed approach with other algorithms.

References

1. Jain, K., Murthy, M.N., Flynn, P.J.: Data Clustering: A Review. ACM Computing Surveys 31(3), 264–323 (1999)
2. Xu, R., Wunsch, D.C.: Clustering, 2nd edn., pp. 1–13. IEEE Press, John Wiley and Sons, Inc. (2009)
3. Verma, H., Kandpal, E., Pandey, B., Dhar, J.: A Novel Document Clustering Algorithm Using Squared Distance Optimization Through Genetic Algorithms. (IJCSE) International Journal on Computer Science and Engineering 02(5), 1875–1879 (2010)
4. Hsuan-Ming, F., Ji-Hwei, H., Shiang-Min, J.: Bacterial Foraging Particle Swarm Optimization Algorithm Based Fuzzy-VQ Compression Systems. Journal of Information Hiding and Multimedia Signal Processing 3(3), 227–239 (2012)
5. Wong, K.C., Li, G.C.L.: Simultaneous Pattern and Data Clustering for Pattern Cluster Analysis. IEEE Transaction on Knowledge and Data Engineering Los Angeles 20, 911–923 (2008)

6. Jain, A.K.: Data clustering: 50 Years beyond K-means. Pattern Recognition Letters 31, 651–666
7. Colanzi, T.E., Assunção, W.K.K.G., Pozo, A.T.R., Vendramin, A.C.B.K., Pereira, D.A.B., Zorzo, C.A., de Paula Filho, P.L.: Application of Bio-inspired Metaheuristics in the Data Clustering Problem. Clei Electronic Journal 14(3) (2011)
8. Yang, X.-S., Deb, S.: Cuckoo Search via Levy Flights. In: Proc. of World Congress on Nature and Biologically Inspired Computing (NaBIC 2009), India, pp. 210–214. IEEE Publications, USA (2009)
9. Yang, X.-S., Deb, S.: Engineering Optimisation by Cuckoo Search. International Journal of Mathematical Modelling and Numerical Optimisation 1(4-30), 330–343 (2010)
10. Jothi, R., Vigneshwaran, A.: An Optimal Job Scheduling in Grid Using Cuckoo Algorithm. International Journal of Computer Science and Telecommunications 3(2), 65–69 (2012)
11. Noghrehabadi, A., Ghalambaz, M., Ghalambaz, M., Vosough, A.: A hybrid Power Series – Cuckoo Search Optimization Algorithmto Electrostatic Deflection of Micro Fixed-fixed Actuators. International Journal of Multidisciplinary Sciences and Engineering 2(4), 22–26 (2011)
12. Merz, C.J., Blake, C.L.: UCI Repository of Machine Learning Databases, http://www.ics.uci.edu/-mlearn/MLRepository.html
13. Hatamlou, A.: Black hole: A New Heuristic Optimization Approach for Data Clustering. Information Sciences 222, 175–184 (2013)
14. Auffarth, B.: Clustering by a Genetic Algorithm with Biased Mutation Operator. In: IEEE Congress Evolutionary Computation (CEC), pp. 1–8 (July 2010)
15. Shelokar, P.S., Jayaraman, V.K., Kulkarni, B.D.: An Ant Colony Approach for Clustering. Analytica Chimica Acta 509, 187–195 (2004)
16. Liu, X., Fu, H.: An Effective Clustering Algorithm with Ant Colony. Journal of Computers 5(4), 598–605 (2010)
17. Premalatha, K., Natarajan, A.M.: A New Approach for Data Clustering Based on PSO with Local Search. Computer and Information Science 1(4), 139–145 (2008)
18. Chuang, L.-Y., Lin, Y.-D., Yang, C.-H.: An Improved Particle Swarm Optimization for Data Clustering. In: Proceedings of the International Multiconference of Engineers and Computer Scientists Hong Kong, vol. I, pp. 440–445 (March 2012)
19. Hatamlou, A., Abdullah, S., Nezamabadi-pour, H.: Application of Gravitational Search Algorithm on Data Clustering. In: Yao, J., Ramanna, S., Wang, G., Suraj, Z. (eds.) RSKT 2011. LNCS, vol. 6954, pp. 337–346. Springer, Heidelberg (2011)
20. Hatamlou, A., Abdullah, S., Nezamabadi-pour, H.: A Combined Approach for Clustering Based on K-means and Gravitational Search Algorithms. Swarm and Evolutionary Computation 6, 47–52 (2012)
21. Hatamlou, A., Abdullah, S., Hatamlou, M.: Data Clustering Using Big Bang–Big Crunch Algorithm. In: Pichappan, P., Ahmadi, H., Ariwa, E. (eds.) INCT 2011. CCIS, vol. 241, pp. 383–388. Springer, Heidelberg (2011)
22. Rendón, E., Abundez, I., Arizmendi, A., M Quiroz, E.: Internal Versus External Cluster Validation Indexes. International Journal 5(1), 27–34 (2011)