Improved Cuckoo Search Clustering Algorithm (ICSCA)

Moe Moe Zaw, Ei Ei Mon
University of Technology (Yatanarpon Cyber City)
moemoezaw@gmail.com,eieimonucsy@gmail.com

Abstract

Clustering is a division of data into groups of similar objects. Each group called a cluster consists of objects that are similar between themselves and dissimilar compared to objects of the other groups. Cuckoo Search Clustering Algorithm (CSCA) is a recently developed nature inspired, unsupervised classification method, based on the most recent meta-heuristic algorithm, stirred by the breeding strategy of the parasitic bird, the cuckoo. To better exploit the search space and to enhance the accuracy of this algorithm, an Improved Cuckoo Search Clustering Algorithm (ICSCA) is proposed in this paper. Normally, in the search space, a substantial fraction of the new solutions should be generated by far field randomization and whose locations should be far enough from the current best solution, this will make sure the system will not be trapped in a local optimum. This ICSCA algorithm that is expected to find the global cuckoo solution and exploit the search space more thoroughly than CSCA algorithm is proposed.

Keywords: Data Clusering, Cuckoo Search, Cuckoo Search Clustering Algorithm, Improved Cuckoo Search Clustering Algorithm.

1. Introduction

Clustering is a typical unsupervised learning technique for grouping similar data points. A clustering algorithm assigns a large number of data points to a smaller number of groups such that data points in the same group share the same properties while, in different groups, they are dissimilar. Clustering has many applications such as image image segmentation, information retrieval, market segmentation and scientific and engineering analysis.[9] Clustering techniques have been used successfully to address the scalability problem of machine learning and data mining algorithms, where prior to, and during training, data is clustered and samples from these clusters are selected for training , thereby reducing the computational complexity of the training process, and even improving generalization performance[1][6][5].

Cuckoo search clustering algorithm (CSCA) is a recently developed clustering algorithm which works in an unsupervised way to create clusters of the points of dataset. The CSCA algorithm is conceptually simpler, takes less parameter than other nature inspired algorithms and after some parameter tuning, yields very good results.[10]

In CSCA algorithm, only current best solution is considered for new cuckoo solutions. So, the next solutions are trapped in the current best local optima. Then, as the current best solution is considered in only one iteration, this current best solution will be lost if the next solutions are not as good as the previous best solution. So, this improved cuckoo search clustering algorithm is proposed for finding better new solutions.

This paper is divided into several sections. Section 2 describes the related work on clustering area and cuckoo search. Cuckoo Search Clustering Algorithm (CSCA) is discussed in Section 3. Improved Cuckoo Search Clustering Algorithm (ICSCA) is presented in Section 4. The comparison between CSCA and ICSCA id discussed in Section 5. Section 6 concludes this paper.

2. Related Work

One of the best known and most popular clustering algorithms is the k-means algorithm. The algorithm is efficient at clustering large data sets because its computational complexity only grows linearly with the number of data points. However, the algorithm may converge to solutions that are not optimal [9].

In [3], this paper proposes a clustering method that integrates the simplicity of the k-means algorithm with the capability of the Bees Algorithm, a new population-based search algorithm that is capable of locating near-optimal solutions efficiently, to avoid local optima.

Two new approaches to using Particle Swarm Optimization to cluster data are presenter in [4]. It is shown how PSO can be used to find the centroids of a user specified number of clusters. The algorithm is then extended to use k-means clustering to seed the initial swarm. The letter algorithm basically uses PSO to refine the clusters formed by k-means.

In [8], different data clustering algorithms k-means algorithm, hierarchical clustering algorithm, self-organization maps algorithm, expectation maximization

are compared. They are compared according to size of dataset, number of clusters, types of dataset and type of software used.

In [7], an improved particle swarm optimization based on Gauss chaotic map for clustering is proposed. Gauss chaotic map adopts a random sequence with a random starting point as a parameter, and relies on this parameter to update the positions and velocities of the particles. It provides the significant chaos distribution to balance the exploration and exploitation capability for search process. This easy and fast function generates a random seed processes, and further improve the performance of PSO due to their unpredictability.

In [2], a hybrid clustering algorithm based on K-mean and K-harmonic mean (KHM) is described. Its performance is compared with the traditional K-means & KHM algorithm. The result obtained from proposed hybrid algorithm is much better than the traditional K-mean & KHM algorithm.

Cuckoo search clustering algorithm (CSCA) is a recently developed clustering algorithm which is capable to group a set of input samples (data points) into clusters with similar features. The developed framework is applied to a benchmark dataset IRIS. The results obtained are highly accurate. Inspired by the results, this Clustering Algorithm have been further applied to extract water from a real time multispectral remote sensing image [10].

3. Cuckoo Search Clustering Algorithm (CSCA)

Cuckoo Search Clustering Algorithm (CSCA) Algorithm has been designed as a clustering algorithm, so it is capable to group a set of input samples (data points) into clusters with similar features. It works in an unsupervised way, without considering the class of the input patterns during the process. The Davies-Bouldin index (DBI) is used to represent the fitness of each nest, i.e., of each solution. The goal for achieving a proper clustering is to minimize the DBI [8].

As new cuckoo solutions, the cuckoo laid eggs which correspond to a new solution set. The new set of clusters is created using the current best solution. For finding better solutions a pseudorandom proportional rule is used. In this rule, $q \in [0,1]$ is the standard CSCA parameter and *q1* is a random value in [0,1]. This rule helps to exhibit exploration and exploitation way to search the solutions.

New Cuckoo xnew is created as,

q$\in\{0.1\}$

q1=random number $\in\{0,1\}$

if(q1<q)

For each cluster centroid $m_{new,k}$

For each dimension n_d

$$x_{new}(m_{new,k,n_d}) = x_{best} + pow(-1, n_d) * rand(1)$$

where, x_{best} is the best solution of current iteration

end

end

else

select another set of clusters randomly from the search space

end

4. Improved Cuckoo Search Clustering Algorithm (ICSCA)

A substantial fraction of the new solutions should be generated by far field randomization and whose locations should be far enough from the current best solution, this will make sure the system will not be trapped in a local optimum. In CSCA, the equation for new cuckoo introduced is the searching of around the current best solution. The system can be trapped in local optima near the best solution i.e., the search space will be only near the best solution. The distance between the current position and the current best solution should be considered. Then, for the next solution set, the current best solution should be remained unchanged because that current best solution may be a better solution than the next solution set. So, in this improved cuckoo search clustering algorithm, a new cuckoo function is introduced as follows:

For each cuckoo

$$x_{new} = x_i + (x_i - x_{best}) * rand(1)$$

end

where,

x_{new} = new solution

x_{best} = the best solution of current iteration

x_i = current solution

In this improved Cuckoo Search Clustering Algorithm, the system cannot be trapped because new cuckoo solution can be considered by using current best solution and the distance between the current best solution and current solution. Moreover, all the nests

except the best one will be replaced by new cuckoo solutions. This can avoid the more iterations to reach optimal solution because the current best solution which is better than the next new solution cannot be lost.

The improved cuckoo search clustering algorithm (ICSCA) is as shown in Fig(1).

Algorithm : ICSCA(no of clusters, no of host nests)

1. Consider NH host nests containing 1 egg (solution) each
2. For each solution of host i
3. Initialize x_i to contain k randomly selected cluster centroids (corresponding to k clusters), as $x_i = (m_{i,1}, \ldots, m_{i,j}, \ldots m_{i,k})$ where $m_{i,k}$ represents the kth cluster centroid vector of ith cluster centroid vector of i^{th} host.

[End for loop]

4. For **t** iterations
5. For each solution of host i of the population
6. For each data document z_p
7. Calculate distance $d(z_p,m_{j,k})$ from all cluster centroids $C_{i,k}$ by using eq-1
8. Assign z_p to $C_{i,k}$ by $d(z_p,m_{j,k}) = \min_{k=1\ldots k} \{ d(z_p,m_{j,k}) \}$

[End for loop]

9. Calculate fitness functionf(x_i) for each host nest i by eq-2

[End for loop]

10. Replace all the nests **except for the best one** by **new Cuckoo eggs generated by New Cuckoo Solution** from their positions
11. A fraction pa of worse nests are abandoned and new ones are built randomly
12. Keep the best solutions (or nests with quality solutions)
13. Find the current best solution

[End for loop]

14. Consider the clustering solution represented by the best solution
15. Exit

Figure1. Improved Cuckoo Search Clustering Algorithm

$$d(z_p, m_{i,k}) = \sqrt{\sum_{j=1}^{N_d} (z_{pj} - m_{(i,k),j})^2 \Big/ N_d} \qquad (1)$$

where

z_p and $m_{i,k}$ = two data vectors

N_d = dimension of vector space

Z_{pj} and $m_{(i,k),j}$ = the data z_p and $m_{(i,k)}$'s weight values in dimension j

$$F = \frac{\sum_{i=1}^{N_c} \{ \frac{\sum_{j=1}^{P_j} d(o_i, m_{i,j})}{p_i} \}}{N_c} \qquad (2)$$

where

$m_{i,j}$ = j^{th} document vector which belong to cluster i;

o_i= the centroid vector of i^{th} cluster

$d(o_i,m_{i,j})$= distance between document $m_{i,j}$ and the cluster centroid o_i

p_i = the number of documents which belongs to cluster C_i

N_c= number of clusters

4.1. New Cuckoo Solution

The cuckoo laid eggs which correspond to a new solution set. The movement of cuckoo to the new position is determined by using the current best solution and the current solution.

For each cuckoo

$$x_{new} = x_i + (x_i - x_{best}) * rand(1)$$

end

where,

x_{new} = new solution

x_{best} = the best solution of current iteration

x_i = current solution

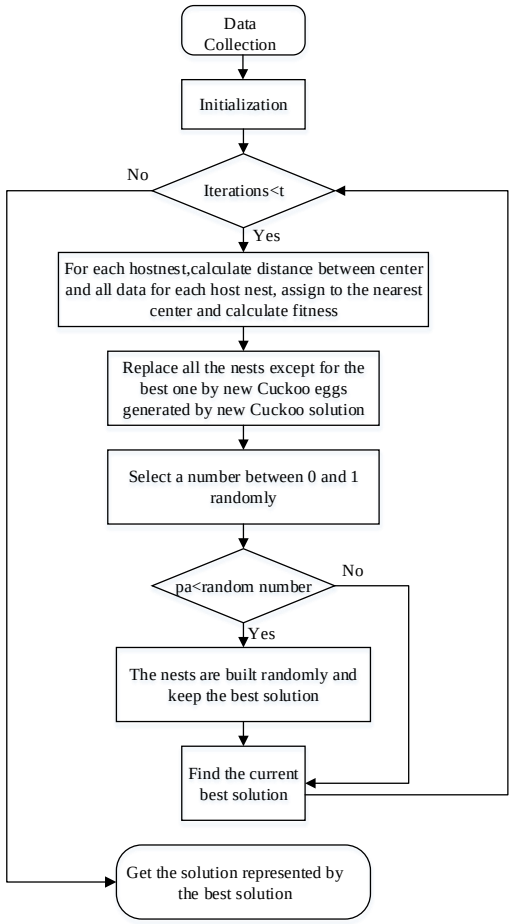

Figure 2. Flowchart of ICSCA Algorithm

According to Fig(2), the data to be clustered will be collected first. The user must intilize the number of clusters(the clusters to be formed), number of host nests(different solutions). Each host nest will contain the user defined number of clusters.For t iterations, the algorithm will be performed. For each host nest, the distance between the centers and all data for each host nest will be calculated and assign the data to the nearest cluster.Then, the fitness of each host nest will be calculated. All host nests except the one with best fitness will be replaced with new Cuckoo solutions calculated by $\mathbf{x_{new} = x_i + (x_i - x_{best}) * rand(1)}$ i.e., for each host nest with no good fitness, the new centers will be calculated by new Cuckoo equation. Then, a random number between 0 and 1 will be selected to decide whether the host bird will find that their solutons have been replaced by the Cuckoo solutions. If the random number is greater than the probability pa, all the nests will be destroyed and replaced by random solution i.e., when the host bird has known that their nests have been replaced by Cuckoo, they will destroy all nests and will build their own nests. If the random number is less than pa, the nests will not be destroyed and they will be carried on to the next iteration. At the end of each iteration, find the host nest with the current best solution (best fitness colution). At the end of all iterations, the current best solution of the last iteration is the best solution.

5. Comparison between CSCA and ICSCA

In CSCA, new cuckoo solution is generated as follows:

For each cluster centroid $m_{new,k}$

For each dimension n_d

$$x_{new}(m_{new,k,n_d}) = x_{best} + pow(-1, n_d) * rand(1)$$

where, x_{best} is the best solution of current iteration

end

end

The new solution is generated from near the current best solution. The distance from the current best solution is pow(-1, n_d)*rand(1). The position of the current solution is not considered. If the best solution is a little far from the current best solution, the best solution will be missed or more iterations will be needed to to reach the best solution since the new cuckoo solution considers only the best solution. Moreover, according to the pseudorandom proportional rule, if q1<q, all of the current solutions will be replaced by the new cuckoo solutions. At this point, if the new cuckoo solution is not as good as the current best solution, the current best solution will be lost because of the replacement of all current solutions. So, ICSCA considers these points of CSCA.

In ICSCA, the new cuckoo solution is considered as follows.

$$x_{new} = x_i + (x_i - x_{best}) * rand(1)$$

The new cuckoo solution will be generated to replace all the less fitness host nests except the best one. The new cuckoo solution is calculated by using both current solution and the distance between the current solution and current best solution. Since the distance

between the current solution and the current best solution are considered, the new solution will not be trapped only near the current best solution. Moreover, the other solutions except the current best solution will be replaced by the new cuckoo solution. So, the current best solution will not be lost if the current best solution is better than the new cuckoo solution.In ICSCA, the new solution will be moved from the current position to the distance of $(x_i - x_{best}) * rand(1)$.

6. Conclusion

In this paper, an improved cuckoo search clustering algorithm is proposed. It is proposed to generate the new cuckoo solutions to better exploit the search space. It can also be expected to decrease the iteration of the Cuckoo Search Clustering Algorithm. Therefore, Improved Cuckoo Search Clustering Algorithm is expected to achieve the better clustering accuracy. As our future work, this Improved Cuckoo Search Clustering Algorithm (ICSCA) can be tested using benchmark datasets. Other new equations for new cuckoo solution can also be generated to achieve better performance.

References

[1] D Fisher. "Knowledge Acquisition via Incremental Conceptual Clustering"., Machine Learning, Vol. 2, pp 119-172. 1987.

[2] Devi Ahilya Vishwavidyalaya, "A Hybrid Clustering Algorithm for Data Mining", CCSEA, SEA, CLOUD, DKMP, CS & IT 05, pp. 387–393, 2012.

[3] DT.Pham, S. OTri, A.Afify, M.Mahmuddin , H.AI-Jabbouli, "Data Clustering Using the Bees Algorithm", Proc, 40th CIRP International Manufacturing Systems Seminar 2007.

[4] DW van der Merwe , AP Englbercht, "Data Clustering using Particle Swarm Optimization", 2003 IEEE.

[5] G Potgieter, "Mining Continuous Classes using Evolutionary Computing", M.Sc Thesis. Department of Computer Science, University of Pretoria, Pretoria, South Africa. 2002.

[6] JR Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, 1993.

[7] Li-Yeh Chuang, Yu-Da Lin, and Cheng-Hong Yang, " An Improved Particle Swarm Optimization for Data Clustering", Procedding of the International MultiConference of Engineers and Computer Scientists, 2012 Vol I, Hong Kong.

[8] Osama Abu Abbas, "Comparisons between Data Clustering Algorithms", *The International Arab Journal of Information Technology* Vol. 5, No. 3, July 2008.

[9] Pham, D.T. and Afify, A.A. "Clustering techniques and their applications inengineering". Submitted to Proceedings of the Institution of Mechanical Engineers,*Part C*: *Journal of Mechanical Engineering Science, 2006.*

[10] Samiksha Goel, Arpita Sharma, Punam Bedi, "Cuckoo Search Clustering Algorithm: A novel strategy of biomimicry", Delhi University, Delhi, India