A Hybrid Cuckoo Search and K-Means for Clustering Problem

Abba Suganda Girsang[1]
[1] Master in Computer Science, Bina Nusantara University, Jakarta, Indonesia
Email : agirsang@binus.edu

Ardian Yunanto[2], Ayu Hidayah Aslamiah[3]
[2,3] School of Computer Science, Bina Nusantara University, Jakarta, Indonesia
Email : [2]ayunanto@binus.edu, [3]aaslamia@binus.edu

***Abstract*—Cuckoo search algorithm (CSA) is one of behavior algorithm which is effective to solve optimization problem including the clustering problem. Based on investigation, k-means is also effective to solve the clustering problem specially in fast convergence. This paper combines two algorithms, cuckoo search algorithm and k-means algorithm in clustering problem called FCSA. Cuckoo search is used to build the robust initialization, while K-means is used to accelerate by building the solutions. The result confirms that FCSA's computational time in ten datasets is faster than the compared algorithm**

Keywords— Clustering, Cuckoo search algorithm (CSA), K-means algorithm, Fast algorithm

I. INTRODUCTION

Some heuristic algorithms such as genetic algorithm, particle swarm optimization, ant colony optimization, bee colony optimization, cuckoo search (CS) algorithm and so forth are used to solve various optimization problem such as traveling salesman problem (TSP) [1] repairing inconsistent matrix AHP [2][3], including solving clustering [4][5]. Clustering is a process to divide amount of data in the population based on the centroid point that constantly changing until it formed groups [6]. Many researchers have tried to develop the algorithms that more accurate or fast than the previous research for solving clustering, such as genetic algorithm [7], ant colony optimization [8], artificial bee colony optimization [9].

One of swarm intelligent algorithm is used to solve some optimization problem is CS. Yang and Deb proposed firstly CS based on the obligate brood parasitic behavior of some cuckoo species in combination with the levy flight behavior of some birds and fruit flies [10]. This algorithm was tried to solve some benchmark test functions. Then, many researchers continue this algorithm for solving some optimization problem such as solving traveling salesman problem (TSP) [11], clustering problem [12], multi objective [13], structural optimization problem [14], and so forth. CS and its modified are proposed by some researchers to solve clustering problem [12][15][16][17]. Unfortunately, in our best knowledge, no research is modified for CSA in fast model. Thus, the problem statement in this case is how to use the cuckoo search or its modified to solve the clustering problem with the efficient time. In original CSA for clustering, levy flight is used to generate random solution in each main iteration. Each main iterations is filled with pseudo iteration that calculate fitness from each new solution generated. The best fitness will be generated from the pseudo iteration to find out the best fitness of all cuckoo. Moreover, the centroid will be updated in first and last section of 1 pseudo iteration. In the last section, the algorithm uses random value to choose whether the iteration will generate new solution or not. This algorithm takes too much iteration which causes huge time process. It happens because each main iteration are filled with pseudo iteration that calculate fitness value from each new solutions generated in the pseudo iteration. Therefore, to optimize the redundancy of the process, pseudo iteration is much better to be discarded from the main iteration and replaced by K-means [12]. K-means is used for decreasing its computation time. Thus, the hybrid of cuckoo search and K-means are expected to solve the clustering problem with the short time.

II. RELATED WORK

Cuckoo search algorithm (CSA) is one of the latest nature-inspired metaheuristic algorithm, which is developed in 2009 by Xin-She Yang from Cambridge University and Suash Deb of C.V Raman College of Engineering [10]. Cuckoo search is based on a parasitism of some cuckoo species which chooses a nest where the host bird just laid its own egg. Commonly, the cuckoo eggs hatch earlier than the host eggs. Moreover, once the cuckoo hatches the new chick cuckoo, the cuckoo will evict the host egg by pushing the egg out of the nest. It makes increasing the portion of food for cuckoo chick which is provided by its host bird. The main step of CSA can be described as follows. First, some numerous host nests are generated. Each parasitic cuckoos lays one egg and dump its egg in randomly chosen nest. The best nest with high quality egg will be hatched in the nest and will be used as nest in next generation.

The eggs laid by cuckoos which are discovered by the host bird with a probability $pa \in [0, 1]$. The host bird then abandons the cuckoo egg based on the value pa. However the number of nest has to be fix. Therefore, substituting the abandon nest, the host bird creates a new nest so that the nest number is not reduced. This process is repeated until get the best solution. Figure 1 shows the Cuckoo search algorithm for clustering. Each nest represents a solution and a cuckoo egg represents a

978-1-4799-7675-1/17/$31.00 ©2017 IEEE

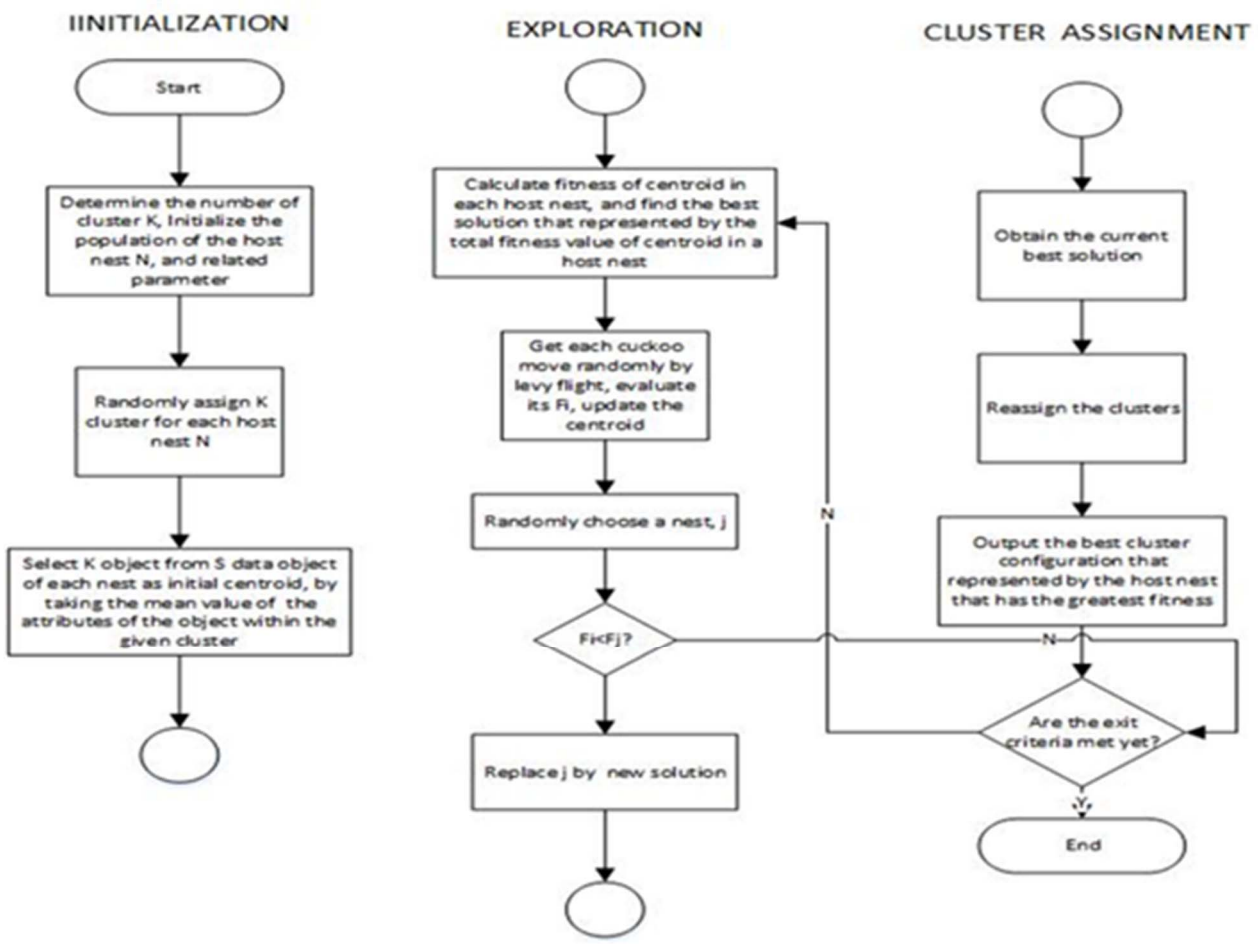

Fig. 1. Cuckoo search algorithm for clustering [10]

new solution. The solution can be built by creating the centroids for each clusters. In the original CSA [10], the new solution of cuckoo is got by using levy flight approach. This new solution then compared to a nest solution which is selected randomly. The new solution is accepted if its solution is better than the nest's solution. Then some of bad solutions (represents bad nest) are abandoned and some of new solutions will be built in order to retain the number solutions. The best solution from all of the cuckoo and nest will be kept. The next process is conducted by repeating build the new solution of cuckoo using levy flight approach. CSA has been applied in many areas of optimization and computational intelligence with promising efficiency. For instance, in computer science, CSA has superior performance over other algorithms for a range of continuous optimization problems, such as for Travelling Salesman Problem and clustering problem.

III. THE PROPOSED METHOD

The proposed method is divided into 2 main process: exploration (using CSA) and converging (using K-Means). First, initialization is conducted to adjust parameter used such as number nests, number generations, and so forth. The exploration is then done in step 2-6. Based on Yang's research [8], bird has an egg in some nests. An egg represents solution created which is built randomly. In clustering, the solution might be the centroids. Then a cuckoo is selected based on levy flight randomly. This cuckoo's solutions is then compared to one of nest's solution which is chosen randomly. The best one is kept, while the bad one is moved out. Using the coefficient fraction of *pa*, some worst of all solutions are discarded and then build the new ones to substitute the discarded solutions. The best of them is kept as initialization for k-means process. This k-means algorithm generates the new centroid as new solution. Then this solution is used for repeating the CSA process until the maximum iteration met.

IV. EXPERIMENT RESULTS

Matlab is used as a processing tools to develop and compile the proposed algorithm and comparison method. FCSA algorithm is then compared with original K-means algorithm for clustering problem [12,13]. The experiment

method is implemented by using dataset retrieved from Uci Dataset, namely Iris, Wine, Yeast, Abalone, Breast cancer, Glass, E.coli, Haberman, Sonar and Parkinson [14]. Table 1 describes the data type, the number of iteration and the number cuckoo nest

Algorithm of FCSA

1. Initialize the number of nests, maximum generation for Levy flight

//2-6 Exploration by CSA

2. Generate random centroid for solutions corresponding of each the host egg nests.
3. Generate some solutions (make centroid) by some cuckoo egg.
4. Get a cuckoo randomly by Levy flights evaluate its quality/fitness
5. Choose a nest solution and comparing to the solution of step 4, keep the best one, and discard the other.
6. Based on the fraction *pa*, some of solutions are discarded and some news are built, keep the best one solution.//This solution contains some centroids of solutions

// 7-11 Converging by K- Means

7. Each points are clustered to the respective centroids.
8. Calculate the distance based on solution that already set by *K*-means algorithm.
9. The smallest fitness will be stored as the best fitness value.
10. The centroids are recalculated and storing as best solution
11. Back to step 4 to generate the new solutions until generation for Levy flight maximum achieved

Table 1. Data set and setting number cuckoo

Dataset	Iteration	Cluster	Cuckoo Nest
Wine	300	3	10
Iris	300	3	10
Yeast	300	10	10
Abalone	300	29	10
Breast cancer	300	2	10
Glass	300	7	10
E.coli	300	8	10
Haberman	300	2	10
Sonar	300	2	10
Parkinson	300	2	10

The performance of FCSA is evaluated in two focus, quality and computation time. Quality means determining values taken from the sum of squared errors as described Eq. (1).

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{nj} \left\| x_{ij} - c_i \right\|^2 \quad (1)$$

Quality is then observed by calculate the mean value and standard deviation from 10 experiment in the proposed methods. The result then compared by 10 experiments from K-means method. The second focus is the computation process through generating the result. The computation proccess indicates the best computation from its 10 experiment, standard deviation and mean value. The the result will be compared with K-means method.

Fig. 2 and 3 describe about the optimal fitness value reached between 50 iteration. From those figure, it can be inferred that the proposed method can achieve the optimal fitness less than 50 iteration on those dataset. The proposed method get the lower fitness result than the comparison at first iteration, but achieve better result at next iteration. Table 2 describes the calculation of mean and standard deviation quality.

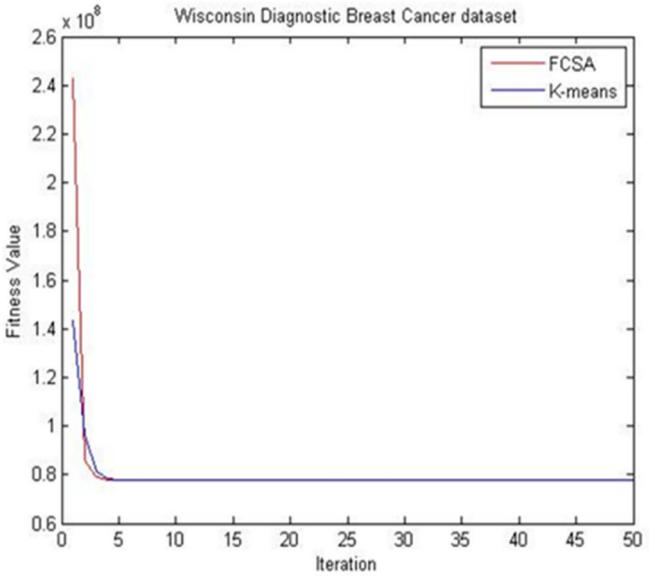

Fig. 2. Plot Fitness Wisconsin Diagnostic Breast Cancer.

Table 2. Mean and standard deviation Quality of FCSA and *K*-means

Dataset Name	FCSA		*K*-means	
	Mean	Std. Dev	Mean	Std. Dev
Wine	**2.4e+06**	**0**	2.4e+06	8.04e+04
Iris	**78.94**	**0**	85,33	20,21
Yeast	**468,78**	**94,06**	489,92	154,71
Abalone	**27,73**	**0,62**	27,93	0,74
Breast Cancer	**7.8e+07**	**0**	**7.8e+07**	**0**
Glass	**317.61**	**15.94**	343,66	17,40
E.coli	**15,04**	**0,68**	15,47	0,70
Haberman	**23404**	**17,18**	30517	18,16
Sonar	**237,43**	**0,14**	280,61	0,17
Parkinson	**3.4e+05**	**0**	5.6e+05	**0**

The result shows that from 10 dataset, almost all of dataset shows that our proposed method is more robust than compared

method. Table 3 describes about the computational process from FCSA and K-means. From table 2 and 3, it can be concluded that FCSA have best processing fitness and computational time in almost all dataset

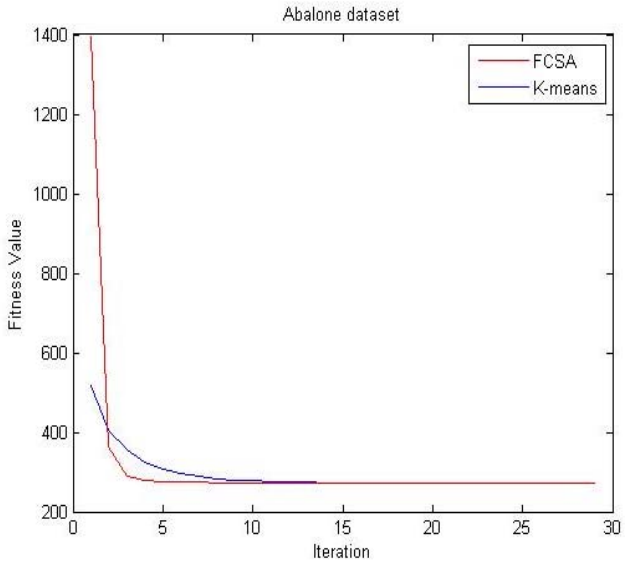

Fig. 3. Plot Fitness Abalone.

Table 3. Mean and standard deviation Computation Process of FCSA and *K*-means

Dataset Name	FCSA		*K*-means	
	Mean	Std. Dev	Mean	Std. Dev
Wine	**0,014**	**0,003**	0,19	0,02
Iris	**0,0075**	**0**	0,16	0,002
Yeast	**0,08**	**0,06**	1,25	0,11
Abalone	**1,22**	**0,13**	4,71	0,78
Breast Cancer	**0,04**	**0,005**	0,41	0,02
Glass	**0,02**	**0**	0,3	0,002
E.coli	**0,02**	**0,01**	0,35	0,003
Haberman	0.14	**0**	**0,014**	0
Sonar	**0,04**	**0**	0,31	0,01
Parkinson	**0,02**	**0**	0,023	0,01

V. CONCLUSION

In this research, a fast algorithm is proposed to decrease the time process. The concept is to omit Levy flight feature and change it with K-means method to produce the best solution. Cuckoo Search will be used to avoid local optima that often happen in K-means weakness. K-means is well known as solving of clustering problem and has ability to solve the clustering problem faster. The experiment result showed that the proposed method has ability to solve the clustering problem faster meanwhile the fitness value is almost equal with the compared method based on the same tools processing. The future work that can be taken from this research is to reduce the computation time of the proposed algorithm while at the same time optimize the quality of the results regardless of the

REFERENCES

[1] A. S. Girsang, C.-W. Tsai, and C.-S. Yang, “A Fast Bee Colony Optimization for Traveling Salesman Problem,” in Innovations in Bio-Inspired Computing and Applications (IBICA), 2012 Third International Conference on, 2012, pp. 7–12.

[2] A. S. Girsang, C.-W. Tsai, and C.-S. Yang, “Rectifying the Inconsistent Fuzzy Preference Matrix in AHP Using a Multi-Objective BicriterionAnt,” Neural Process. Lett., vol. 44, no. 2, pp. 519–538, 2016.

[3] A. S. Girsang, C.-W. Tsai, and C.-S. Yang, “Multi-objective particle swarm optimization for repairing inconsistent comparison matrices." International Journal of Computers and Applica-tions 36.3, pp 101-109, 2014.

[4] I. B. Saida, K. Nadjet, and B. Omar, “A new algorithm for data clustering based on cuckoo search optimization,” in Genetic and Evolutionary Computing, Springer, 2014, pp. 55–64.

[5] S. Irvan, G. Robin Solala, and G. Abba Suganda, “An Adaptive Cat Swarm Optimization Based on Particle Swarm Optimization Approach (ACPSO) for Clustering,” Int. Rev. Comput. Softw., vol. 11, no. 1, pp. 20–26, 2016.

[6] Z. Huang, “A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining.,” in DMKD, 1997, p. 0.

[7] J. C. Bezdek, S. Boggavarapu, L. O. Hall, and A. Bensaid, “Genetic algorithm guided clustering,” in Evolutionary Computation, 1994. IEEE World Congress on Computational Intelligence., Proceedings of the First IEEE Conference on, 1994, pp. 34–39.

[8] P. S. Shelokar, V. K. Jayaraman, and B. D. Kulkarni, “An ant colony approach for clustering,” Anal. Chim. Acta, vol. 509, no. 2, pp. 187–195, 2004.

[9] D. Karaboga and C. Ozturk, “A novel clustering approach: Artificial Bee Colony (ABC) algorithm,” Appl. Soft Comput., vol. 11, no. 1, pp. 652–657, 2011.

[10] X.-S. Yang and S. Deb, “Cuckoo search via L{é}vy flights,” in Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on, 2009, pp. 210–214.

[11] X. Ouyang, Y. Zhou, Q. Luo, and H. Chen, “A novel discrete cuckoo search algorithm for spherical traveling salesman problem,” Appl. Math. Inf. Sci., vol. 7, no. 2, p. 777, 2013.

[12] R. Tang, S. Fong, X.-S. Yang, and S. Deb, “Integrating nature-inspired optimization algorithms to K-means clustering,” in Digital Information Management (ICDIM), 2012 Seventh International Conference on, 2012, pp. 116–

123.
[13] X.-S. Yang and S. Deb, "Multiobjective cuckoo search for design optimization," Comput. Oper. Res., vol. 40, no. 6, pp. 1616–1624, 2013.
[14] A. H. Gandomi, X.-S. Yang, and A. H. Alavi, "Cuckoo search algorithm: a metaheuristic approach to solve structural optimization problems," Eng. Comput., vol. 29, no. 1, pp. 17–35, 2013.
[15] N. M. Nawi, A. Khan, and M. Z. Rehman, "A new back-propagation neural network optimized with cuckoo search algorithm," in Computational Science and Its Applications--ICCSA 2013, Springer, 2013, pp. 413–426.
[16] A. Hashmi, D. Gupta, Y. Upadhyay, and S. Goel, "Swarm intelligence based approach for data clustering," Int. J. Innov. Res. Stud, vol. 2, pp. 572–589, 2013.
[17] S. Arora and I. Chana, "A survey of clustering techniques for big data analysis," in Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference-, 2014, pp. 59–65.