Data Clustering Using Hybrid Algorithms

Anahita Sabagh Nejad

University of Debrecen
Anahita.sabagh@inf.unideb.hu

Abstract

Clustering is an unsupervised technique for partitioning objects such that all objects in the same group share similar properties. The method that is used here is K-means in combination with three algorithms respectively. This algorithm starts with k clusters for k objects. These k objects are selected randomly as cluster centers or centroids such that each of them belongs to one cluster. In other words,the idea is to keep similar objects in the same cluster. Since this algorithm is iterative, for the next iteration we will assign new cluster centers, by computing the average or mean value of each cluster. Then each object will be reassigned to the cluster that has less distance with or is closer(similar) to the centers. For distance measurement, we use Euclidean distance. For the next iteration, we will have another mean value, so the clusters change During these reassignments by swapping objects. This algorithm will be repeated until meeting a requirement(condition) like converging to the same value, or till the clusters don't change. In this paper, we want to combine K-means with the Artificial Flora optimization algorithm, Artificial Bee Colony, Ant Colony optimization for clustering purposes. Then we will have a comparison between them to find the more accurate one. All these algorithms are agent-based. The agents are Flora seeds (offsprings), bees, and ants. In all of these definitions, swarms are the main core. Each of the swarms has main characteristics that bring intelligence to our algorithms. For example, in flora algorithm, The process of migration of offsprings to find a better environment is one of its main characteristics, finding the information like quality and distance by looking to dance of the bees and finding the shortest path in ant colony by increasing the amount of pheromone are some examples. Individually, Artificial Flora Optimization(AFO) can find a better solution in comparing with Particle swarm optimization(PSO) and Artificial Bee Colony(ABC) regarding research done by some scientists mentioned in this article. Here we want to check the behavior of the Flora algorithm and bee colony when they are used for clustering, and we are added ant colony as another compared algorithm to cluster our data.

Keywords: Swarm Intelligence, k-means clustering, Flora Optimization Algorithm, Ant Colony Optimization, Artificial Bee Colony Optimization

MSC: 90C27, 68W10, 91C20

1. Introduction

The Artificial Flora Optimization(AFO) algorithm is a swarm intelligence algorithm based on migration and reproduction of seeds(offsprings), discovered by Long cheng, Xue-hun Wu and Yan Wang in 2018. The task of Flora is to find an optimal growth environment through the evolution, distinction, and rebirth of Flora. Flora spread its seeds and its distribution area changes because of changes in the environment and competitions [1]. These seeds (offsprings) may fall near the original plant, this way they will have more chance to stay alive and rebirth Flora since the possibility of climate change in a small radius is less than far environment, but some seeds that they go far, may not survive, because of this change in the environment, therefore some of them die that we call it extinction and there is also some that adapt to the new environment and make a new original plant with new characteristics that we call it evolution. The main elements of this algorithm are original plants, offspring plants(seeds), propagation distance and location.

The Artificial Bee Colony(ABC) is an optimization algorithm introduced by Derviş Karaboğa (Erciyes University) in 2005. This algorithm uses the foraging behavior of bees. In the artificial bee colony model, the colony consists of three groups: employed bees, onlookers and scouts. The scout bees are the translation of employed bees [2] initial food sources are produced for all employed bees. The task of an Employed bee is to find the closest food source with higher nectar. Then it and comes back to the hive and dance in this area. This dance that is called waggling dance gives some information like directions, distances and the quality(fitness)of the nectars(food source) to the onlooker bees. If the bees find higher quality foods, they forget the information about the previous one and replace it with the position of the new one that has more quality. So having more quality means having more chances to be selected by onlookers. In this algorithm, position means the possible solutions [2]. The scout bees are translated from a few employed bees, which abandon their food sources and search for new ones [2]. Scout bees start a foraging process cause they search for promising patches. Flower patches with large amounts of nectar or pollen that can be collected with less effort tend to be visited by more bees, whereas patches with less nectar or pollen receive fewer bees [3]. [4] scout bees move randomly from one patch to another.
Ant Colony Optimization algorithm(ACO) proposed by Marco Dorigo in 1992. The idea is to find the shortest path(optimal path) from the nest to the substance. Here a chemical substance called Pheromone helps to find this path. When the amount of pheromone is more in a path, it means that way or track was more interesting for ants, so the population of them is increased until they converge to one line. So the shortest path is the path with the maximum amount of Pheromone. [5] Over time, however, the pheromone trail starts to evaporate, thus reducing its attractive strength. The more time it takes for an ant to travel down the path and back again, the more time the pheromones have to evaporate.

2. Clustering Methods

The method that is used for hybrid Flora is to start with the N population(original plants) and calculating propagation distance. Each offspring is a solution for our K clusters. The Euclidean distance must be calculated between each plant and cluster centers to decide whether they belong to a cluster or no. For new centroids, this process of the assignment should be repeated. Then select some neighborhood search space that in this algorithm our space is a circle and its radius is the maximum distance. We update the position by setting Flora offsprings for the selected sites in the last step(more offsprings for the best sites)and compute the fitness. In this step, we select the fittest offspring plant(each offspring plant can be an original plant for the next step). Assign other offsprings till meeting an stopping criterion. Deciding about a plant will be alive or not must be calculated by the roulette wheel.

The ABC algorithm starts with n scout bees. Each bee is a solution for k clusters. The Euclidean distance must be computed between each data and cluster centers to decide whether they belong to a cluster or no. For new centroids, this process of assignment repeats. For each site that is visited by a bee, fitness will calculate. M Selected site(sites with higher fitness) will be chosen for the Neighborhood search. In the last step, only the fittest bee will be selected to form part of the population. The remaining bees in the population assign randomly to scout for new solutions until a stopping criterion or condition is met, these steps will repeat. At ant colony, we have some clusters that their probability modifies(updates)by pheromone. Firstly, We consider equal pheromone for each path. Two points i and j are respectively starting points and endpoints(centroids). Each k ant moves from i to j if the value of P is bigger.

3. Clustering Using Methaheuristic algorithms

Here are the results of clustering with Artificial Flora for Iris data set. Setosa is Class Number 1, Versicolor is Class Number 2, Virginica is Class Number 3. These results are the best results obtained with each algorithm.

N	iterations	M	p	c_1	c_2	μ
20	50	10	0.9	0.75	1.25	0.2

Table 1: Parameters of AFO

	1	2	3	total
1	50 (33.3%)	0 (0%)	0 (0%)	100%
2	0 (0%)	49 (32.7%)	0 (0%)	100%
3	0 (0%)	1 (0.7%)	50 (33.3%)	98%
	100%	98%	100%	99.3%(0.7%)

Table 2: Confusion matrix of AFO

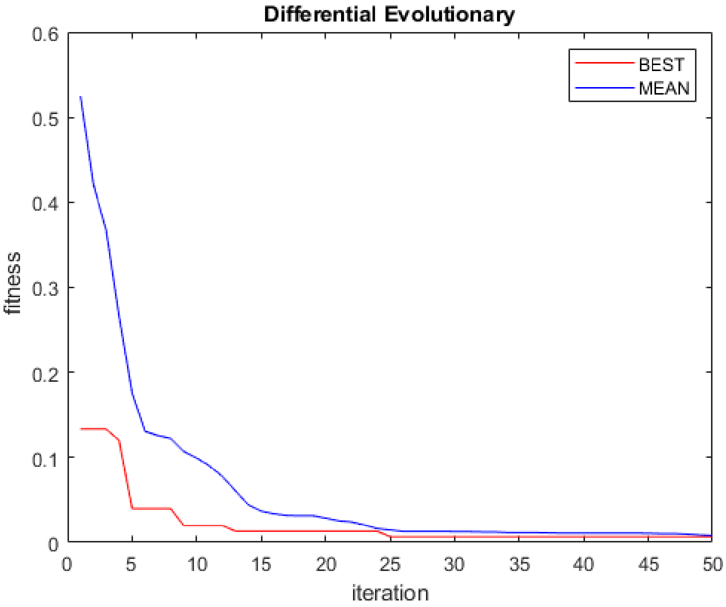

Figure 1: Artificial flora best fitness for 50 iterations

Table 2 shows the percent of data that are correctly predicted in an unsupervised method k-means. For example, in the 1th column of 1th row of table 2, 33.3 percent of our data are placed in the correct group, and for the 2th row and the 3rd row of the same column we hadn't any data to want to be placed, so the result of the 1th column is 100 percent correct. For the 1st row of the 2nd column, we have no data but for the 2nd row, 32.7 percent grouped correctly. For the 3rd row of the same column, we have just 0.7 percent error, which means it is not clustered or grouped well, so the result is 98 percent of our data of the 2nd column are clustered correctly. The 3rd column is the same, and the final result of correct grouping data

	1	2	3	total
1	50 (33.3%)	0 (0%)	0 (0%)	100%
2	0 (0%)	48 (32.0%)	0 (0%)	100%
3	0 (0%)	2 (1.3%)	50 (33.3%)	96.2%(3.8%)
	100%	96%(4.0%)	100%	98.7%(1.3%)

Table 3: Confusion matrix of ABC

for the 3rd column is 100 percent, and the total of correct clustering with hybrid k-means with AFO is 99.3 percent. This work is done for the other algorithms(ABC and ACO) too. For example, in the above table(table 3) the percent of the total correct grouping is 98.7 percent and for wrong grouping is 1.3. ...

The results of Artificial bee colony and Genetic Algorithm combination using UniformCrossOver with this parameter MCN=50 that is the maximum number of cycles, NB=20 that is the number of bees is illustrated in figure 2 as below:

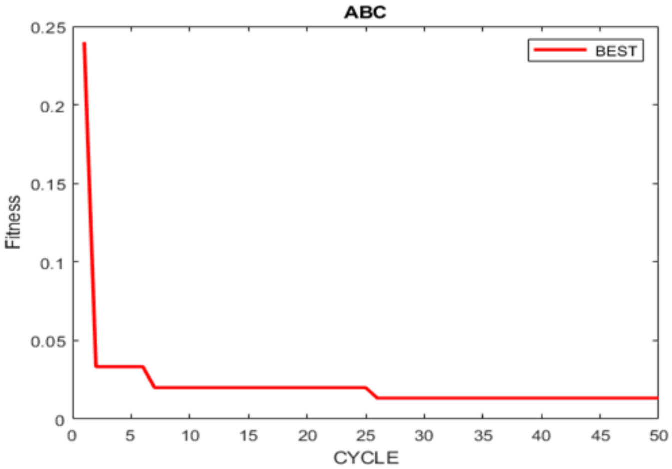

Figure 2: Fitness cycle for artificial bee colony

The results of clustering by Ant colony optimization algorithm are illustrated with a maximum iteration of 50 and 20 populations. Maxiter=50 npop=20 Alpha=1.

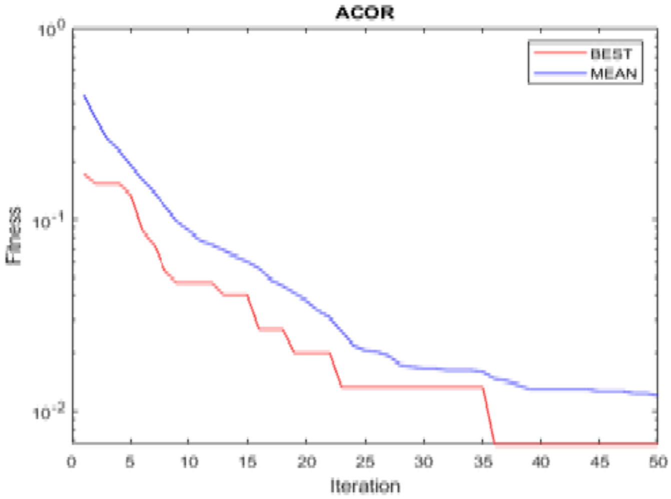

Figure 3: Fitness iteration of ant colony

	1	2	3	total
1	50 (33.3%)	0 (0%)	0 (0%)	100%
2	0 (0%)	49 (32.7%)	0 (0%)	100%
3	0 (0%)	1 (0.7%)	50 (33.3%)	98%(2.0%)
	100%	98%(2.0%)	100%	99.3%(0.7%)

Table 4: Confusion matrix of ACO

	Min	AVG	max	SUM	std dev
AF	0.88871	0.969359	0.9912	29.08078	0.023425
ABC	0.92436	0.969814	0.98258	29.09443	0.014943
ACO	0.97413	0.984028	0.9912	29.52084	0.004549

Table 5: Statistical measurements for accuracy

Table 5 shows the statistical measurements for the best or the more accurate time that the algorithms converged to the same fitness value. The program was run for 30 times. . . . The class results of the implementation of this algorithm for clustering are shown in table 6:

	Af	ABC+GA	ACO
MR 1	0.0088	0.017422	0.0088
ECM2	0.0132	0.026133	0.0132
NECM3	0.0066	0.013067	0.0066
Sensitivity	0.99333	0.98667	0.99333
Specificity	0.98693	0.9744	0.98693
Accuracy	0.9912	0.98258	0.9912
Precision	0.98667	0.97336	0.98667
Recall	0.98693	0.9744	0.98693
F-measure	0.9868	0.97388	0.9868
Consistency	0.9804	0.9616	0.9804
ROC4	0.99013	0.98053	0.99013
B5	0.98963	0.97959	0.98963
Time	23.2546	1.756	1.573
BEST fitness	0.0066667	0.013333	0.0066667

Table 6: Class Results

B^1	$1-\sqrt{\frac{1}{2}\left(\left(\frac{FN}{TP+FN}\right)^2+\left(\frac{FP}{TN+FP}\right)^2\right)}$
MR^2	$\frac{FP+FN}{TP+TN+FP+FN}$
Err_1	$\frac{FP}{TN+FP}$
Err_2	$\frac{FP}{TP+FN}$
ECM^3	$C_1Err_1P_{ndf}+C_2Err_2P_{df}$
$NECM^4$	$Err_1P_{ndf}+\frac{C2}{C1}Err_2P_{df}$
$Consistency$	$\frac{dn-k^2}{k(n-k)}$
F	$\frac{2*R*P}{R+P}$
$Recall$	$\frac{TP}{TP+FN}$
$Precision$	$\frac{TP}{TP+FP}$
$Accuracy$	$\frac{TP+TN}{TP+TN+FP+FN}$
$Specificity$	$\frac{Numberofmodulescorrectlypredictedasnon-faultprone}{totalnumberofactualnonfaultymodules}\times 100$
$Sensitivity$	$\frac{Numberofmodulescorrectlyasfaultprone}{totalnumberofactualfaultymodules}\times 100$

Table 7: Statistical measurements for accuracy

4. Conclutions

This paper has presented a new clustering method based on the Flora Algorithm. The method employs the flora seeds(offsprings)to search for the set of cluster centers that minimizes the distance. One of the advantages of the proposed method is that it does not become trapped at locally optimal solutions. The proposed method used for clustering data (iris dataset) by using different swarm intelligence algorithms. In the summarization of the class results of the previous algorithms, Accuracy is the most important factor and here Artificial Flora and Ant colony are more accurate than the Artificial Bee Colony. Experimental results show that this algorithm has good performance to cluster data. On base of these tables, AF and ACO have maximum accuracy, but the standard deviation of AFO is less than ACO. All 3 algorithms are robust, but the average accuracy of ACO was higher on one side, and on another side, the standard deviation of ACO was less than AF and ABC.

Acknowledgements. This work was supported by the University of Debrecen, Faculty of Informatics.

References

[1] Long Cheng, Xue-han Wu, and Yan Wang. Artificial flora (af) optimization algorithm. *Applied Sciences*, 8(3):329, 2018.

[2] D KARABOGA. An idea based on honey bee swarm for numerical optimization. 2005.

[3] Karl Von Frisch. *Bees: their vision, chemical senses, and language.* Cornell University Press, 2014.

[4] D. Pham, Ebubekir Koç, Afshin Ghanbarzadeh, and Sameh Otri. Optimisation of the weights of multi-layered perceptrons using the bees algorithm. *Proceedings of 5th International Symposium on Intelligent Manufacturing Systems*, pages 38–46, 06 2006.

[5] Scott Camazine, Jean-Louis Deneubourg, Nigel R Franks, James Sneyd, Eric Bonabeau, and Guy Theraula. *Self-organization in biological systems.* Princeton university press, 2003.