A Novel Hybrid Optimization Algorithm for Data Clustering

S.Yuvaraj
PG Student,
Dept of CSE,
Bannari Amman Institute of Technology,
Sathyamangalam.

M.Krishnamoorthi
Assistant Professor (Sr.Gr),
Dept of CSE,
Bannari Amman Institute of Technology,
Sathyamangalam.

ABSTRACT

Clustering is the unsupervised learning in which the data is divided into similar groups (cluster) without any prior knowledge. The emerging swarm-based algorithms become an alternative to the conventional clustering methods to enhance the quality of results. Artificial Bee Colony (ABC) Algorithm is one of the Swarm Intelligent based optimization algorithm that exhibit foraging properties of a Honey Bee Swarm. Bacterial Foraging Optimization (BFO) is another Swarm intelligence algorithm which imitates the foraging properties of the E.coli bacteria. In this paper, we hybridize both ABC and BFO by replacing the Scout bee phase of ABC by BFO to have a minimum Intra cluster distance. From the experimental results, it shows the proposed H-ABFO algorithm outplays the traditional K-means, ABC and BFO algorithms.

Keywords

Clustering, Artificial Bee Colony Algorithm, Bacterial Foraging Algorithm and K-Means Algorithm.

1. INTRODUCTION

Clustering is the important tool in fields like Data mining, Data compression, Statistical analysis. where its aims to gather data into clusters (groups) in manner where objects in same cluster will have a High degree of similarity (Intra cluster) while have low degree of similarity to objects in other clusters (Inter cluster).Swarm intelligence is one of the recent trend that applied in various fields. In this paper, attempt to hybrid the recently introduced swarm intelligent optimization algorithms such as Artificial Bee Colony (ABC) and Bacterial Foraging Optimization (BFO).where both ABC and BFO algorithm tries to find their solution (Food Source) by imitating their properties.

Foraging is a kind of social insect behaviors and can be modeled as an optimization process where an animal seeks to maximize energy intake per unit time spent for foraging. The Swarm Intelligence on Honey Bees is most recently studied field in which their foraging behavior, learning, memorizing are the closely analyzed research areas. ABC algorithm proposed by Karaboga and Basturk describes the its performance analyzed with various heuristics algorithms like genetic algorithm (GA) , differential evolutional algorithms and particle swarm optimization (PSO)[3],optimization algorithm for solving constrained optimization problems[4],optimization algorithm for training feed forward neural networks[5] and an efficient algorithm for numerical function optimization[6]. Karaboga and Ozturk [1] given ABC as a novel clustering approach for numerical optimization problems for solving several benchmark problems (datasets) and compared with results of PSO for same datasets. Changsheng, Jiaxu and Dantong worked on ABC approach for clustering which partition N objects into K clusters in which DEB’s rule is used tested on various data sets and compared with several heuristics algorithms like tabu search, genetic algorithm and ant colony optimization[2].

Bacterial Foraging Optimization (BFO) is another Swarm intelligence algorithm proposed by Passino [8] which imitates the foraging properties of the E.coli bacteria. Miao Wan, Lixiang Li et al., [9] proposed a new Clustering algorithm based on Bacterial foraging optimization in which a group of bacteria forage to converge to certain positions as final cluster centers by minimizing the fitness function. This proposed algorithm is tested with real time datasets for various cluster sizes and densities. Swagatam Das, Sambarta Dasgupta et al., given a mathematical analysis of chemotactic steps in BFOA in view of Gradient Decent Search [11]. In which its simulations result of several benchmark problems shows proposed Adaptive BFOA have better convergence behavior than classical BFOA and also they analyzed the Stability and Convergence of chemotaxis of Bacterium [13] in a continuous time. Morteza Eslamian, Seyed Hossein et al., used a new integer-code algorithm which applied on basis of Bacterial Foraging optimization for solving Unit Commitment problem [12]. By using this integer coding, computation time get decreased and time constrains will be coded directly. Jakob R. Olesen, Jorge Cordero H., and Yifeng Zeng proposed a new algorithm namely Auto-CPB (Auto-Clustering based on particle bacterial foraging) [18], a hybridization of PSO and BFO algorithms. This aims to cluster the data by using simplistic collaboration. The proposed algorithm is compared with traditional K-means algorithm and shows better result then other PSO based approaches. Sibylle D. Müller, Jarno Marchetto et al., proposed an algorithm namely Bacterial Chemotaxis (BC) algorithm [15]. This is an improved optimization technique with features added to the basic algorithm. BC gives similar result as evolutionary strategies but worse when convergence properties enhanced.

The rest of the paper is organized in following manner. Section 2 will give detailed description about the basic concepts such as Artificial Bee Colony (ABC) algorithm and Bacterial Foraging Optimization (BFO) algorithms and its principles. Section 3 will describe the proposed Hybrid algorithm in detail. Section 4 will comprise of Cluster Validity Parameters that used to evaluate the Clustering algorithm. Section 5 will be the result and discussions of the proposed algorithm. Section 6 will be the conclusion of the paper.

2. BASIC CONCEPTS

2.1 Clustering Problem

Clustering is the process of grouping objects in multidimensional data based on some similarity measures [1]. Distance measurement is generally used for evaluating similarities between patterns. In particular the problem is stated as follows: given objects say N, allocate each object to one of clusters given as k and minimize the sum of squared Euclidean distances between each object and the center of the cluster belonging to every such allocated object. The clustering problem minimizing is described [1] as in

$$J\ (w, z) = \sum_{i=1}^{N}\sum_{j=1}^{K} W_{ij}\, ||X_i - Z_j||^2$$

Where K is the number of clusters, N the number of patterns, xi (i = 1. . . N) the location of the ith pattern and zj (j = 1, . . . , K) is the center of the jth cluster.

2.2 Principles of K-Means Clustering Algorithm

K-means is one of the simplest unsupervised learning algorithms [17] that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). The idea is to define k centroids, one for each cluster.

This algorithm aims at minimizing an objective function, in this case a squared error function. The objective function is given as,

$$J=\sum_{j=1}^{k}\sum_{i=0}^{k} ||x_{i} - c_{j}||^2$$

2.2.1 Steps for K-means Algorithm:

Step 1: Initially partition cluster into k in random.
Step 2: Find the new cluster partition by placing each object to its closest cluster center.
Step 3: Compute new cluster as centroids of the clusters.
Step 4: Repeat the steps 1 and 2 until no change in the cluster centers.

Although it can be proved that the procedure will always terminate, the k-means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centers. The k-means algorithm can be run multiple times to reduce this effect. K-means is a simple algorithm that has been adapted to many problem domains. As we are going to see, it is a good candidate for extension to work with fuzzy feature vectors.

2.3 Principles of Artificial Bee Colony(ABC) Algorithm

The Artificial Bee Colony algorithm is based on the behavior of honey bee swarms[1-4]. Foraging behavior of honey bees leads to the emergence of social intelligence of honey bee swarms consisting of three essential components: food source, employed foragers and unemployed foragers and two modes of behavior. On various parameters such as richness of energy,proximity to the nest and ease of extracting this energy are depends on food source. When the recruit bee finds and exploits the food source, it will raise to be an employed forager who memorizes the location of the food source. After the employed foraging bee loads and unloads a portion of nectar are send to food source to the hive.If it is low level or exhausted, then nectar amount was decreased , foraging bee abandons the food source and become an unemployed bee. There are two possibilities for an unemployed forager. One is Scout bee, which search spontaneously without any knowledge. According to the information into the nest, the percentage of scout bees varies from 5% to 30% . The mean number of scouts averaged over conditions is about 10%. The other unemployed bee is Recruit, according to knowledge from waggle dance done by some other bee and it will start searching.

2.3.1 Steps for ABC Algorithm:

Step 1: Initialize the populations of Solutions i = 1…..SN and evaluate them.
Step 2: Produce a new solution for employed bee , evaluate them and apply greedy selection process.
Step 3: Calculate the Probabilities of sources based on which the onlookers is preferred.
Step 4: Assign onlooker bees to employed bees based on the Probabilities of solution, then produce new solution and apply greedy selection process.
Step 5: Stop the process if there is abondoned in source and send scout bee to the search area for discovering new source in random.
Step 6: memorize the best food source found.
Step 7:If the termination condition satisfied,then stop the algorithm else go to step 2.

In this algorithm, initialization with feasible solution is impossible for some cases and it is also time consuming . Scout bee that selects the solution in random will choose new solution from population which most probobaly an infeasible one. So ABC , does not consider the initialization to be feasible.

2.4 Principles of Bacterial Foraging Optimization (BFO) Algorithm

Bacterial Foraging Optimization (BFO) Algorithm [8][9] is a global optimization algorithm for the distributed optimization and control. It is nature inspired optimization technique proposed by Passino. Since other optimization algorithms like evolutionary programming ,evolutionary strategies ,genetic Algorithm and also recent swarm based algorithms like Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO) are found way efficiently and proved their effectiveness. Based on the same strategy of swarm optimization BFO algorithm is proposed. BFO is inspired by the behavior of Escherichia coli.

The general idea behind BFO is Application of group foraging strategy of a swarm of bacteria in multi-optimal function optimization.To maximize energy obtained per unit time so the bacteria search for nutrients. Individual bacterium also communicates with others by sending signals.This algorithm is efficient in solving various real world applications which are all in the basis of optimization problems. The kind of biology behind and its foraging behavior is emulated in extraordinary manner and used as a simple optimization problem. This bacterial foraging optimization methodology is been used in clustering problems in which a group of bacteria forage to converge to certain positions as final cluster centers by minimizing the fitness function.

2.4.1 Steps for BFO Algorithm:

Step 1: input the number of clusters k.
Step 2: Initialize the k centers for $c_1,c_2,………,c_k$ and generate random positions for each cluster centers $\mathbf{Z_c}$.
Step 3: start the Chemotatic $\mathbf{N_c}$ for the population of S bacteria at the $\mathbf{J}^{th}$ step and this iteration (**Ns**-Number of steps)

will be continue until it continues to reduce the objective function.

Step 4: After Chemotatic Step **N_c**, the Reproduction Phase N_{re} will be taken and the bacteria that are healthier will be retained and it will be split into two bacteria in same location.

Step 5: Next phase is Elimination and Dispersal **N_{ed}**, in which Bacteria either get killed (or) dispersed to a new environment.

Step 6: once the condition satisfied , calculate the distance **d_c** as the bacteria will converge to the certain place as its final cluster centers .otherwise, go to step 3.

Step 7: finally, allocate the data **d** to the final clusters that are closest in the clusters.

In BFO, during the Elimination and Dispersal only less probability of bacteria will be assigned randomly over search space. Which are dispersed to a new environment is the primary merit that this algorithm.

3. H-ABFO ALGORITHM

From the study made on the above algorithms, we clarified that in ABC algorithm initialization of feasible solution is impossible in some cases and also scout bee made its search in random which probably leads to an infeasible solution. So in order to overcome this problem we hybridize this ABC along with BFO algorithm. The H-ABFO algorithm is done by replacing the Scout bee phase of ABC algorithm with BFO.

3.1 Steps for H-ABFO Algorithm:

Step 1: Input the populations of Solutions i= 1…..SN,number of clusters and evaluate them.

Step 2: Produce a new solution for employed bee , evaluate them and apply greedy selection process.

Step 3: Calculate the Probabilities of sources based on which the onlookers is preferred. Assign onlooker bees to employed bees based on the Probabilities of solution, then produce new solution and apply greedy selection process.

Step 4: stop the process if there is abondoned in solution and apply bacterial Chemotaic **N_c** with population of S bacteria in multimodal plane for feasible solution in an iteration and Reprocduce **N_{re},** the bacteria which holds better solution (Healthier) and bacteria will disperse to new environment.

Step 5: calculate the distance **d_c** as the bacteria will converge to the certain place as its final cluster centers .otherwise, go to step 4.

Step 6: else memorize the best food source found.

Step7: If the termination condition satisfied,then stop the algorithm else go to step 2.

From this Hybridization algorithm, we can able to initialize the feasible solution when there is an abondoned in the solution.

4. CLUSTER VALIDITY PARAMETERS

As this paper deals optimization of clustering it is important to evaluate the clustering result to find the partitioning that best fits the data. The procedure to evaluate the clustering algorithm is known as Cluster Validity [8]. Two kinds of Cluster validity measures have been taken. One is External Criteria, which evaluate based on Pre-specified class label information of dataset. The second one is Internal Criteria, which evaluate without any prior knowledge. The external criteria's are

a) Rand Coefficient:
It computes the degree of similarity between the known correct cluster structure and the results obtained by a clustering algorithm and it is defined as,

$$R = \frac{SS + DD}{SS + SD + DS + DD}$$

where SS, DD, SD, DS denotes the data points belongs to same and different clusters.

b) Jaccard coefficient (J):
It is similar as rand coefficient except that it excludes DD and is defined as ,

$$J = \frac{SS}{SS + SD + DS}$$

where SS, SD, DS denotes the data points belongs to same and different clusters.

The internal criteria's are

c) Beta index (β):
It computes the ratio of total variation and within class variation and is defined as ,

$$\beta = \frac{\sum_{i=1}^{C}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}\right)^2}{\sum_{i=1}^{C}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}_i\right)^2}$$

d) Distance Index:
It determines the ratio of average intra-cluster distance and average inter-cluster distance .where intra cluster distance is the distance between points and its center. It is defined by,

$$Intra = \frac{1}{n}\sum_{i=1}^{K}\sum_{x_j \varepsilon C_i} \| X_j - Z_i \|^2$$

Where Inter cluster distance is the distance between two cluster centers and its given by,

$$Inter = \frac{1}{K}\sum \| z_i - z_j \|^2$$

The cluster result can be measured by combining the average Intra-cluster distance and the average Inter-cluster distance and it is defined by the ratio,

$$Dis = \frac{Intra}{Inter}$$

As a result, we want to minimize the value of measure Dis.

Table 1. Comparison table for K-MEANS,ABC,BFO and HYBRID Algorithms

DATASETS	METHODS	RAND COEFFICIENT	JACCARD COEFFICIENT	BETA INDEX	DISTANCE INDEX	INTRA CLUSTER DISTANCE
ZOO	K-MEANS	0.7965	0.3766	4.1003	0.0182	0.4023
	ABC	0.9056	0.6712	5.0234	0.0153	0.1900
	BFO	0.9124	0.6924	5.9624	0.0152	0.2030
	H-ABFO	0.8956	0.5985	4.1072	0.0159	0.1200
WINE	K-MEANS	0.7418	0.4005	7.0011	0.2651	0.4135
	ABC	0.7314	0.4367	7.1241	0.0611	0.1091
	BFO	0.7502	0.4429	7.9358	0.0243	0.2130
	H-ABFO	0.6123	0.3567	7.0127	0.5699	0.1003

5. RESULT AND DISCUSSIONS

In this paper, we proposed a new Hybrid algorithm for data clustering with help of Artificial Bee Colony [1-3] and Bacterial foraging optimization algorithms [8][9] and compared with traditional k-means algorithm[19][20].For each problem, we have used cluster validity measures to evaluate the clustering algorithms. Both External and Internal criteria is been considered and following table shows the values obtained for the existing algorithms and also for newly proposed one.

For Zoo Dataset,

1) It is evident that in terms of Cluster validity Measures (Distance Index) and for its objective (Intra cluster Distance) the performance of the proposed H-ABFO algorithm works better for Zoo data set (as in Fig 1).
2) Whereas in terms of its measure (Beta Index) ,BFO and ABC is better than the proposed H-ABFO algorithm.
3) When considering the other Cluster Validity measures (Rand and Jaccard Coeficients), ABC and BFO algorithms gives better result but H-ABFO outplays K-means.
4) On whole , for zoo dataset the proposed H-ABFO Algorithm either works better nor close enough to the earlier values obtained for the given Cluster validity measures.

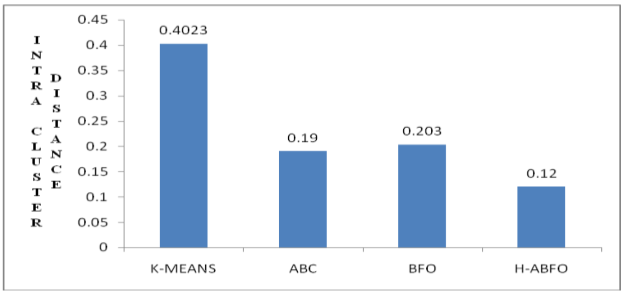

Fig 1: Comparison chart for Intracluster distance using Zoo dataset

For Wine Dataset,

1) When testing with Wine dataset, the Intra Cluster Distance for proposed H-ABFO gives better result than other existing algorithms(as in Fig 2).
2) When comparing other internal criteria (Beta index) for H-ABFO is very much close enough to the existing algorithms.
3) On considering the external criteria's (Rand and Jaccard Coeficients), the existing algorithm performs better than the new H-ABFO algorithm.
4) For wine dataset, the proposed H-ABFO algorithm will be better for its Intra Cluster Distance but not much for other Cluster validity parameters.

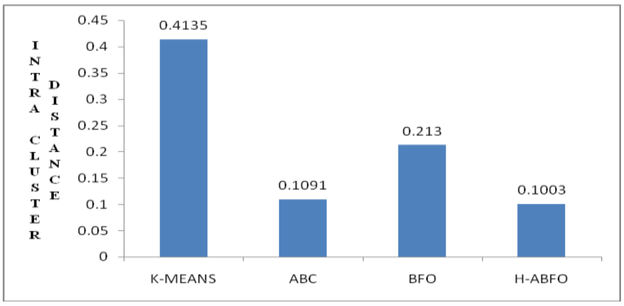

Fig 2 : Comparison chart for Intracluster distance using Wine dataset,

To sum it up , the proposed Hybrid algorithm is efficient and works better for various real time datsets and it can be used to solve real life problems.

6. CONCLUSION

Swarm intelligence is one of the recent trends that applied in various fields. This paper presents the new algorithm which hybridizes the ABC and BFO algorithms. The cluster validity measures are used to evaluate the performance of these algorithms. Thus the performance of the proposed algorithm is compared with other well-known optimization algorithms like ABC, BFO and also with traditional K-Means Algorithm. This algorithm was tested by using real time datasets such as zoo, wine and the result obtained shows that the Intra Cluster Distance and Distance Index obtained is better than other

Existing Swarm intelligence algorithm like ABC and BFO and for other Cluster Validity measures, the values are close enough to the Compared algorithms.

7. REFERENCES

[1] Dervis Karaboga and Celal Ozturk, "A novel clustering approach:Artificial Bee Colony (ABC) algorithm", Applied Soft Computing, pp.652-657,2011.

[2] Changsheng Zhang, Dantong Ouyang and Jiaxu Ning, "An artificial bee colony approach for clustering", Expert Systems with Applications, pp. 4761–4767,2010.

[3] Dervis Karaboga and Basturk B, "On the performance of artificial bee colony (ABC) algorithm". Applied Soft Computing, Vol. 8, pp. 687–697,2008.

[4] Dervis Karaboga and Basturk B, "Artificial Bee Colony (ABC) optimization algorithm for solving constrained optimization problems", LNCS: Advances in Soft Computing: Foundations of Fuzzy Logic and Soft Computing, vol. 4529, Springer–Verlag, pp.789–798, 2007.

[5] Dervis Karaboga and Basturk B, "Artificial Bee Colony (ABC) optimization algorithm for training feed-forward neural networks", LNCS: Modeling Decisions for Artificial Intelligence, MDAI, vol. 4617, Springer–Verlag, pp. 318–329,2007.

[6] Dervis Karaboga and Basturk B, "A powerful and efficient algorithm for numerical function optimization: Artificial Bee Colony (ABC) algorithm", Journal of Global Optimization, Vol.39,pp. 459-471,2007.

[7] Dervis Karaboga and Bahriye Akay, "A Comparative Study of Artificial Bee Colony Algorithm", Applied Mathematics and Computation, vol. 214,pp. 108-132,2009.

[8] Passino, K. M, "Biomimicry of bacterial foraging for distributed optimization and control", IEEE Control Systems Magazine, vol.22,pp. 52–67,2002.

[9] Miao Wan , Lixiang Li , Jinghua Xiao , Cong Wang and Yixian Yang , "Data clustering using bacterial foraging optimization" , Journal of Intell Information System,vol. 38, pp.321–341,2011.

[10] Dervis Karaboga and Bahriye Akay, "A modified Artificial Bee Colony algorithm for real-parameter optimization", Journal of Information Sciences, vol.192,pp.120–142,2010.

[11] Sambarta Dasgupta, Swagatam Das, Ajith Abraham and Arijit Biswas, "Adaptive Computational Chemotaxis in Bacterial Foraging Optimization: An Analysis", IEEE Transactions on Evolutionary Computation, vol. 13,pp:919-941,2009.

[12] Morteza Eslamian, Seyed Hossein Hosseinian, and Behrooz Vahidi, "Bacterial Foraging-Based Solution to the Unit-Commitment Problem ", IEEE Transactions on Power Systems, vol. 24, pp.1478-1488,2009.

[13] Swagatam Das, Sambarta Dasgupta , Arijit Biswas , Ajith Abraham and Amit Konar, "On Stability of the Chemotactic Dynamics in Bacterial-Foraging Optimization Algorithm" , IEEE Transactions on Systems, Man, and Cybernetics—part a: Systems and Humans, vol. 39, pp.670-679.2009.

[14] Mishra S and Bhende C.N, "Bacterial Foraging Technique-Based Optimized Active Power Filter for Load Compensation", IEEE Transactions on Power Delivery, vol. 22, pp. 457-466,2007.

[15] Sibylle D. Müller, Jarno Marchetto, Stefano Airaghi, and Petros Koumoutsakos,"Optimization Based on Bacterial Chemotaxis" IEEE Transactions on Evolutionary Computation, vol. 6,pp. 16-29,2002.

[16] Rui Xu and Donald Wunsch II, "Survey of Clustering Algorithms", IEEE Transactions on Neural Networks, vol. 16, pp.645-678,2005.

[17] Krishna K and Narasimha Murty M, "Genetic K-Means Algorithm", IEEE Transactions on Systems, Man, and Cybernetics—part b: Cybernetics, vol. 29,pp. 433-439,1999.

[18] Jakob R. Olesen, Jorge Cordero H., and Yifeng Zeng, "Auto-Clustering Using Particle Swarm Optimization and Bacterial Foraging", L. Cao et al. (Eds.): ADMI 2009, LNCS 5680, pp. 69–83,2009.

[19] Juntao Wang and Xiao long Su,"An improved K-Means clustering algorithm", IEEE 3rd International Conference on Communication Software and Networks (ICCSN), pp. 44 – 46,2011.

[20] Shuhua Ren and Alin Fan, "K-means clustering algorithm based on coefficient of variation", 4th International Congress on Image and Signal Processing (CISP), vol.4,pp. 2076 – 2079,2011.

[21] Kanungo T, Mount D.M, Netanyahu N.S, Piatko C.D, Silverman R, and Wu A.Y, "An efficient k-means clustering algorithm: analysis and implementation", IEEE Transactions on Pattern Analysis and Machine Intelligence , vol.24,pp. 881 – 892,2002.

Xiaojun Bi and Yanjiao Wang ," An Improved Artificial Bee Colony Algorithm", 3rd International Conference on Computer Research and Development (ICCRD),vol.2,pp. 174 – 177,2011