Artificial Bee Colony Algorithm is More Effective on Small Size Datasets as Compared to Large Size Datasets in Data Clustering

Zeeshan Danish
University of Malakand
Chakdara, Pakistan

Ahmed Hassan
Humboldt University
Berlin, Germany

Akhtar Badshah
University of Malakand
Chakdara, Pakistan

ABSTRACT

Data clustering is a widespread data compression, vector quantization, data analysis and data mining technique. The principle objective of data clustering is to make clusters (or groups) such that data having high degree of similarity is gathered in the same cluster while data having high degree of dissimilarity is gathered in the different clusters and plays a key role for users to organize, summarize, and steer the data adequately. In this work Artificial Bee Colony (ABC) algorithm is applied to different size datasets. Results clearly show that ABC when applied on small size datasets were more effective than those of large size datasets in terms of intra- cluster distance, computation cycles and time required to complete those cycles.

Keywords

Artificial bee colony algorithm; Data clustering.

1. INTRODUCTION

Clustering is one of the most complicated tasks in pattern Recognition [1], Image analysis [2]**.** The most popular class of clustering algorithms is K-means algorithm [3] which is a centre based, simple and fast algorithm but has the insufficiencies that it highly depends on the initial states and is easily trapped in local minima from the starting position of the search and global solutions of large problems cannot find with reasonable amount of computation effort [4]. In order to overcome local optima problem, the researchers from diverse fields are applying hierarchical clustering, partition-based clustering, density-based clustering, and artificial intelligence based clustering methods, such as: statistics [5], graph theory [6], expectation-maximization algorithms [7], artificial neural networks [8], evolutionary algorithms[9], swarm intelligence algorithms [10-13].

Simulated Annealing approach were discussed and it has been proved theoretically by Selim and Al-Sultan that a clustering problem's global solution can be reached [14]. The algorithm does not "stick" to a local optimal solution, rather it obtains the optimal solution. A disadvantage of the simulated annealing approach is that no characterization of a stopping point is computationally available. Another disadvantage is that verifying that a set of data is Standard Data is as difficult a task as that of solving the clustering problem itself. A new algorithm for solving this problem based on a TS technique. The algorithm obtained results that are better than the well-known k-means and the SA algorithms for many test problems [15].

Another Swarm Intelligence Ant colony clustering algorithm employs distributed agents who mimic the way real ants find a shortest path from their nest to food source and back. Its performance was compared with GA, TS and SA [12]. They showed that their algorithms are better than other algorithms in performance and time.

Mualik and Bandyopadhyay [16] proposed a genetic algorithm based method to solve the clustering problem and experiment on synthetic and real life datasets to evaluate the performance. The results showed that GA-based method might improve the final output of K-means.

The PSO which simulates bird flocking was used for clustering [17]. In order to improve its performance further, the PSO algorithm is hybridized with K-means and N–M simplex search method. Results reveal that K–NM–PSO is superior to the K-means, PSO, and N–M simplex search method [17].

Cuckoo Search Algorithm (CSA) is another Swarm Intelligence algorithm used to cluster data. It is shown how CSA can be used to find the optimally clustering N object into K clusters. The CSA is tested on various data sets, and its performance is compared with those of K-Means, Fuzzy C-Means, Fuzzy PSO and Genetic K-Means clustering. The simulation results show that the new method carries out better results than the K-Means, Fuzzy C-Means, Fuzzy PSO and Genetic K-Means [18].

2. ARTIFICIAL BEE COLONY ALGORITHM

It is one of the newly presented swarm-based Algorithms [19]. Artificial bee colony algorithm behaves the same way as honeybee swarm in its intelligent foraging deeds. Basturk and Karaboga presented an artificial bee colony (ABC) algorithm, which depends upon the foraging conduct of honey-bees for the purpose of numerical function optimization problems [20]. ABC got remarkable advantages over traditional population-based algorithms such as Particle Swarm Optimization (PSO) algorithm, Evolution Algorithm (EA), Genetic Algorithm (GA), Evolution Strategies (ES) and Differential Evolution (DE) due to using less control parameters, ease of implementation and simplicity [19].

Totally, four different selection methods are used by ABC:

i) A global probabilistic selection method: For discovering promising areas, the onlookers calculate the probability value calculated by eq. 3.1

$$\mathrm{Pi} = \frac{\mathrm{fiti}}{\sum_{\mathrm{n=1}}^{\mathrm{SN}} \mathrm{fitn}} \tag{3.1}$$

Where n = 1, 2, 3 …… SN and SN is the number of food sources which is equal to the number of onlooker bees or employed bees and fit_i is the fitness value of the solution I which is related to the nectar amount of the food source in the position i [21].

ii) A local probabilistic selection method: For determining a food source near the source in the memory, the onlookers and the employed foragers carry out a local probabilistic selection method in an area which depends on the visual information for instance the fragrance color & shape of blossoms (sources) (the type of nectar source will not be identified by bees till it reach at the accurate place and distinguish between different sources growing there, which depends upon their scent) is presented in eq. 3.2.

$$V_{ij}= X_{ij}+ \Phi_{i\,j}+ (X_{ij} - X_{kj}) \qquad (3.2)$$

Where j϶ {1, 2, 3..., D} and k϶ {1, 2, 3 ..., SN} are arbitrarily selected index. $\Phi_{i\,j}$ is a random value in the range [1, 1].

iii) Greedy selection process: Local selection called greedy selection process is carried out by employed bees and onlookers in which if the food quantity of the candidate source is better than that of the present one, then bee memorizes candidates source produced by e.q 3.2 and forgets the present one. Else, the present one in the memory is memorized by the bee.

iv) The Scouts carry out a random selection process as defined in eq. 3.3.Canonical ABC contains three control parameters: The value of limit, number of nectar sources that is either the number of onlooker bees or employed represented by SN, and finally MCN represents maximum cycle number.[19]

Where j϶ {1, 2, 3..., D} and x_i is abandoned source, then x_i replace novel food source discovered by scout.

$$X^j_i = X^j_{\,min} + rand\,[0,1]\,(X^j_{max} - X^j_{min)} \qquad (3.3)$$

3. ABC APPROACH FOR DATA CLUSTERING

The approach has been successfully applied on clustering problems on several renowned real data sets and the results were related with other famous heuristic algorithms, for instance TS, SA, GA, ACO and the newly presented K–NM–PSO technique. Results gained illustrated that ABC outer class the aforesaid algorithms in calculations of processing time required and the quality of solution [22]. Fig 1 represents the flow chart for ABC algorithm applied on data clustering.

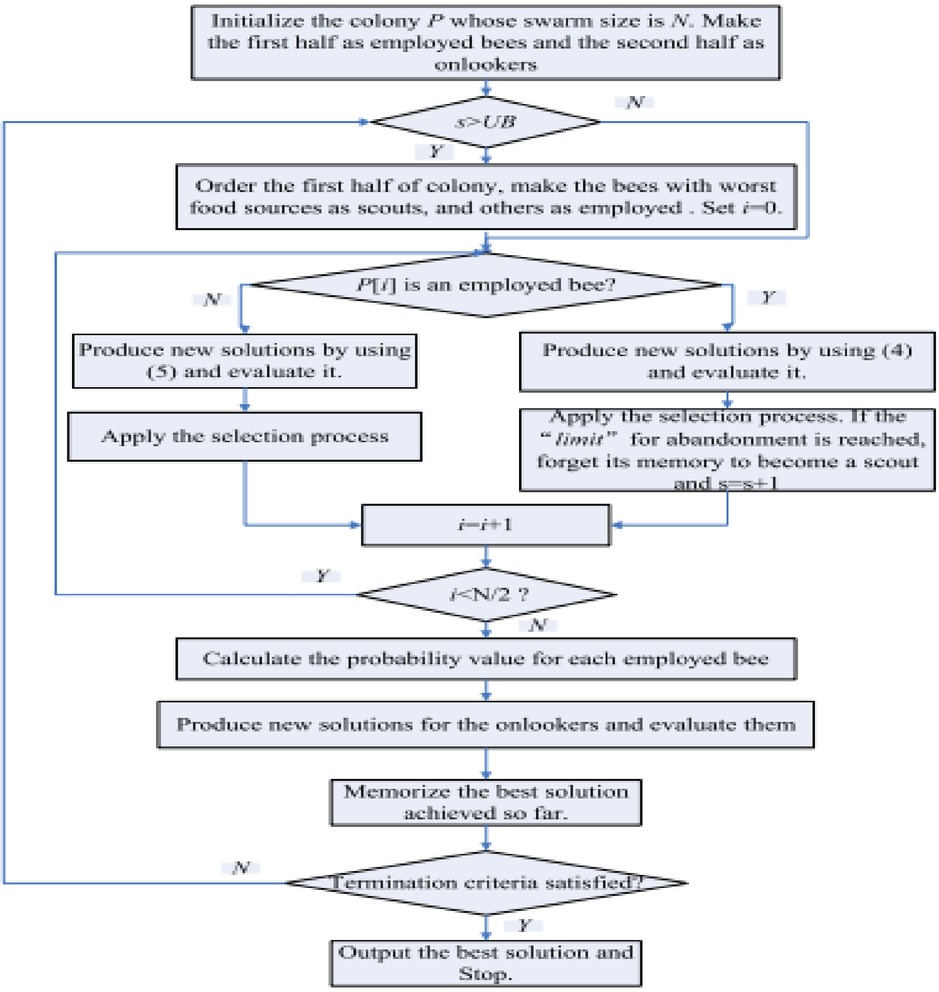
Initialize the colony P whose swarm size is N. Make the first half as employed bees and the second half as onlookers
s>UB
N
Y
Order the first half of colony, make the bees with worst food sources as scouts, and others as employed . Set i=0.
P[i] is an employed bee?
N
Y
Produce new solutions by using (5) and evaluate it.
Produce new solutions by using (4) and evaluate it.
Apply the selection process
Apply the selection process. If the "limit" for abandonment is reached, forget its memory to become a scout and s=s+1
i=i+1
Y
i<N/2 ?
N
Calculate the probability value for each employed bee
Produce new solutions for the onlookers and evaluate them
Memorize the best solution achieved so far.
N
Termination criteria satisfied?
Y
Output the best solution and Stop.

Fig 1: ABC algorithm for data clustering[19]

4. RESULTS AND ANALYSIS

To assess the performance of ABC algorithm for data clustering on different size datasets such as contraceptive Method Choice, wisconsin Breast Cancer, iris and thyroid datasets chosen from the standard UCI machine learning database[23] are implemented on the system having specifications i.e 2.8GHZ processor and 512GB RAM embedded on Pentium 4, and outcomes are compared with the estimated intra-cluster distances, number of cycles required and time required to complete the operation calculated from the smallest iris data set. The results in the table Following sequence is adopted by the datasets. The estimated calculations in terms of intra-cluster distance, number of cycles, time required to complete the operation is lagging behind the actual calculation which shows that with the increasing the size of the dataset decreases the searching process resulting in the extra overhead of the estimated intra-cluster distances, number of cycles required and time required to complete the operation as shown in table1-2 and fig1-3. For the following datasets below Number of instances of data records are shown by N, for each record number of characters are represented by P and total number of clusters to be formed is represented by K.

Contraceptive Method Choice (n=1473, p=10, k=3): the CMC dataset is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. The samples are married women who were either not pregnant or do not know if they were at the time of interview. The problem is to predict the current contraceptive method choices (including no use, long-term methods, or short-term methods) of a woman based on her demographic and socioeconomic characteristics.

Wisconsin Breast Cancer (n=683, p=9, k=2): the WBC dataset n is consists of 683 objects characterized by nine features: clump thickness, cell size uniformity, cell shape uniformity, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses. There are two categories in the data: malignant (444 objects) and benign (239 objects).

The Iris dataset (n=150, p=4, k=3): the iris database contains one hundred and fifty arbitrary specimens of blossoms from its class versicolor, virginica and setosa gathered by Anderson. For every class, it contains fifty perceptions for petal (length), petal (width), sepal (length) and sepal (width) measured in centimeter.

The thyroid gland dataset (n=215, p=5, k=3): This informational index consists of 215 specimens of patients experiencing 3 humanoid thyroid maladies: hypothyroidism, euthyroidism, and hyperthyroidism where thirty patients are experienced hyperthyroidism thyroid, one hundred and fifty people are tried euthyroidism thyroid, while thirty five patients are experienced hypothyroidism thyroid. Every individual was portrayed by five highlights of research facility tests: add up to Ser-um thyroxin as measured by the isotopic uprooting technique, T3-resin uptake-up test, basal thyroid-invigorating hormone (TSH) as measured by radioimmuno test, add up to serum triiodothyronine as measured by radioimmuno test, maximal total contrast of TSH esteem after infusion of two hundred smaller scale grams of thyrotropin discharging hormone when contrasted with the basal esteem.

Table 1. Comparison of ABC algorithm when applied on different size datasets based on intra- cluster distance

Dataset	Iris	Thyroid	Wisconsin Breast Cancer	Contraceptive Method Choice
Average Worst Best	78.94 78.94 78.94	10104.03 10108.24 10100.31	25120.99 25129.55 25111.65	30205.91 30211.85 30190.11
Expected increase in intra-cluster distance as compared to iris data set.	78.94	113.147	359.44	775.19
Extra overhead	0	9995.093	24770.11	29436.66

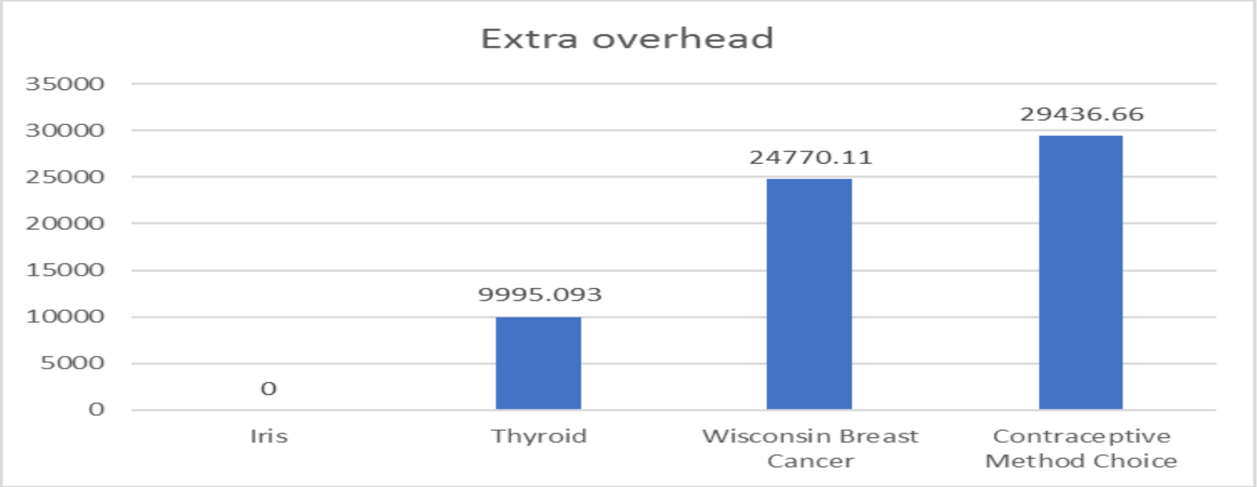

Fig 2: Comparison of ABC algorithm when applied on different size datasets based on intra- cluster distance

Table 2. Comparison of ABC algorithm when applied on different size datasets based on computation time and numbers

Dataset	Iris	Thyroid	Wisconsin Breast Cancer	Contraceptive Method Choice
Actual Time(s)	29.68	85.26	286.99	640.80
Actual Numbers	8658	24136	72888	155438
Estimated Time(s)	29.68	42.54	135.14	291.46
Estimated Numbers	8658	12410	39423	85022
Overhead Time(s)	0	42.72	151.85	349.34
Overhead Numbers	0	11726	33465	70416

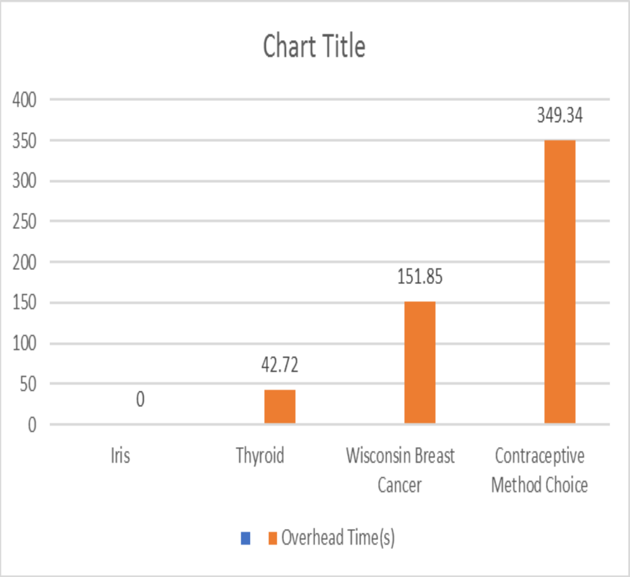

Fig 3: Comparison of ABC algorithm when applied on different size datasets based on computation time.

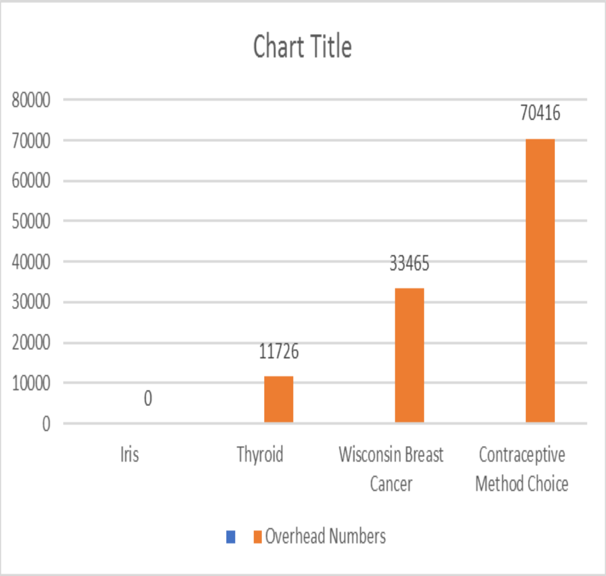

Fig 4: Comparison of ABC algorithm when applied on different size datasets based on computation numbers.

5. CONCLUSION

With the increase in the dimension and size of the problem, the convergence speed of ABC algorithm decreases. This face can be describe as: When the size increases, then in the process of searching a food source, in canonical ABC algorithm, the information exchange among the bees is very difficult as compared to small size datasets. Thus the process of exploitation is effected in large size datasets. Larger the size of datasets less effective will be the ABC algorithm. To overcome this overhead for large size datasets several modified forms of ABC can be applied to improve the information exchange among the bees during the searching process.

6. REFERENCES

[1] M. S. Kamel and S. Z. Selim, "New algorithms for solving the fuzzy clustering problem," Pattern recognition, vol. 27, pp. 421-428, 1994.

[2] M. Omran, A. Salman, and A. P. Engelbrecht, "Image classification using particle swarm optimization," in Proceedings of the 4th Asia-Pacific conference on simulated evolution and learning, 2002, pp. 18-22.

[3] S. Z. Selim and M. A. Ismail, "K-means-type algorithms: a generalized convergence theorem and characterization of local optimality," IEEE Transactions on pattern analysis and machine intelligence, pp. 81-87, 1984.

[4] M. Fathian, B. Amiri, and A. Maroosi, "Application of honey-bee mating optimization algorithm on clustering," Applied Mathematics and Computation, vol. 190, pp. 1502-1513, 2007.

[5] E. W. Forgy, "Cluster analysis of multivariate data: efficiency versus interpretability of classifications," Biometrics, vol. 21, pp. 768-769, 1965.

[6] C. T. Zahn, "Graph-theoretical methods for detecting and describing gestalt clusters," IEEE Transactions on computers, vol. 100, pp. 68-86, 1971.

[7] T. Mitchell, "Machine Learning, McGraw-Hill Higher Education," New York, 1997.

[8] J. West, J. M. Fitzpatrick, M. Y. Wang, B. M. Dawant, C. R. Maurer Jr, R. M. Kessler, et al., "Comparison and evaluation of retrospective intermodality brain image registration techniques," Journal of computer assisted tomography, vol. 21, pp. 554-568, 1997.

[9] S. Paterlini and T. Minerva, "Evolutionary approaches for cluster analysis," in Soft Computing Applications, ed: Springer, 2003, pp. 165-176.

[10] C.-H. Tsang and S. Kwong, "Ant colony clustering and feature extraction for anomaly intrusion detection," in Swarm Intelligence in Data Mining, ed: Springer, 2006, pp. 101-123.

[11] R. Younsi and W. Wang, "A new artificial immune system algorithm for clustering," in International Conference on Intelligent Data Engineering and Automated Learning, 2004, pp. 58-64.

[12] P. Shelokar, V. K. Jayaraman, and B. D. Kulkarni, "An ant colony approach for clustering," Analytica Chimica Acta, vol. 509, pp. 187-195, 2004.

[13] M. Omran, A. P. Engelbrecht, and A. Salman, "Particle swarm optimization method for image clustering," International Journal of Pattern Recognition and Artificial Intelligence, vol. 19, pp. 297-321, 2005.

[14] S. Z. Selim and K. Alsultan, "A simulated annealing algorithm for the clustering problem," Pattern recognition, vol. 24, pp. 1003-1008, 1991.

[15] K. S. Al-Sultan, "A tabu search approach to the clustering problem," Pattern Recognition, vol. 28, pp. 1443-1451, 1995.

[16] U. Maulik and S. Bandyopadhyay, "Genetic algorithm-based clustering technique," Pattern recognition, vol. 33, pp. 1455-1465, 2000.

[17] Y.-T. Kao, E. Zahara, and I.-W. Kao, "A hybridized approach to data clustering," Expert Systems with Applications, vol. 34, pp. 1754-1762, 2008.

[18] P. Manikandan and S. Selvarajan, "Data clustering using cuckoo search algorithm (CSA)," in Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012, 2014, pp. 1275-1283.

[19] D. Karaboga and B. Akay, "A comparative study of artificial bee colony algorithm," Applied mathematics and computation, vol. 214, pp. 108-132, 2009.

[20] D. Karaboga and B. Basturk, "On the performance of artificial bee colony (ABC) algorithm," Applied soft computing, vol. 8, pp. 687-697, 2008.

[21] L. N. De Castro and F. J. Von Zuben, "Artificial immune systems: Part I–basic theory and applications," Universidade Estadual de Campinas, Dezembro de, Tech. Rep, vol. 210, 1999.

[22] C. Zhang, D. Ouyang, and J. Ning, "An artificial bee colony approach for clustering," Expert Systems with Applications, vol. 37, pp. 4761-4767, 2010.

[23] M. Lichman, "{UCI} Machine Learning Repository," 2013.