An Improved Artificial Bee Colony Algorithm for Clustering

Qiuhang Tan*, Hejun Wu+, Biao Hu* and Xingcheng Liu+
School of Information Science and Technology, Sun Yat-sen University,
Higher Education Mega Center, Guangzhou, Guangdong Province, P.R. China
+{wuhejun, isslxc}@mail.sysu.edu.cn, *{tanqh3, hubiao}@mail2.sysu.edu.cn

ABSTRACT
Artificial Bee Colony (ABC) algorithm, which was initially proposed for numerical function optimization, has been increasingly used for clustering. However, when it is directly applied to clustering, the performance of ABC is lower than expected. This paper proposes an improved ABC algorithm for clustering, denoted as EABC. EABC uses a key initialization method to accommodate the special solution space of clustering. Experimental results show that the evaluation of clustering is significantly improved and the latency of clustering is sharply reduced. Furthermore, EABC outperforms two ABC variants in clustering benchmark data sets.

Categories and Subject Descriptors
I.2.8 [Artificial Intelligence]: Problem Solving, Control Methods, and Search – *Heuristic methods*

General Terms
Algorithms.

Keywords
artificial bee colony (ABC); clustering; swarm intelligence.

1. INTRODUCTION
Clustering is to classify patterns into groups without supervision [1]. Clustering methods are able to find the interrelationship among data points without prior information about the data. Therefore, clustering has been a common and useful method in many computing applications.

In clustering, Swarm Intelligence (SI) methods [2][3] are promising, as they are efficient in multivariable optimization while clustering can be transformed to multivariable problems. As a SI algorithm, Artificial Bee Colony (ABC) algorithm [4][5] computes an optimal solution to the clustering evaluation function when it is applied to clustering. The general optimization process is briefly described in the following. After data normalization and solutions initialization, ABC searches for optimal solution in iterations. The iterative search is mainly divided into three steps, which are employed bee phase, onlooker bee phase and scout bee phase. The ABC iterates through these three phases until the termination criterion is met.

GECCO'14, July 12–16, 2014, Vancouver, BC, Canada.
ACM 978-1-4503-2881-4/14/07.
http://dx.doi.org/10.1145/2598394.2598464

In ABC-clustering, the initial solutions are randomly generated in initialization phase. Through extensive experiments, we found that such an approach often encounters poor clustering results. The reason, we conjecture, is as follows. Since the initial centers of clusters are the initial solutions from the random initialization phase of ABC, the probability of obtaining relatively good initial solutions is low. Moreover, the initialization phase does not address the special need of clustering, which requires the cluster centers to be relatively far from each other.

In this paper, an improved ABC (EABC) algorithm is proposed. To improve the evaluation and stability of clustering, EABC generates the initial centers by several data points among the data set, which are relatively far away from each other by the similarity measure of clustering. The centers are determined one by one with the reference of chosen data points until the number of centers reaches the given amount of clusters. We have performed a series of experiments, and the experimental results show that the evaluation of clustering is significantly improved and the latency of clustering is greatly reduced. EABC also outperforms two ABC variants in clustering benchmark data sets.

The rest of this paper is organized as follows. Section 2 describes ABC-clustering method. Section 3 proposes our new initialization method of EABC. Section 4 describes the discussions on the results and parameter. Section 5 summarizes the paper and proposes future work.

2. ABC-clustering Method
ABC-clustering method is described as below.

Step 1) Input the data set and the number of clusters *K*.

Step 2) Normalize all data.

Step 3) Randomly initialize *SN* food sources.

Step 4) Repeat (until the termination criterion is met)

a) Employed bee phase. Update the solutions and compare the new solution with the old one based on the evaluation function (clustering validity index). Only if the new solution is better will it replace the old one.

b) Calculate fitness and probabilities.

c) Onlooker bee phase. Choose the solutions with relatively higher fitness and probabilities to update, and use the same update strategy as that in the employed bee phase.

d) Memorize the best solution.

e) Scout bee phase. If the trial number of the solution reaches the trial limit, reinitialize the solution randomly.

Table 1. Results (mean and standard deviation over 30 runs) on different data sets

Data set	EABC (IRP=0.4)	ABC	GABC	MABC
Iris	31.31±0.11	43.51±0.32	44.51±0.70	45.31±0.64
Wine	102.86±0.42	147.36±0.49	146.48±1.07	148.76±1.03
Glass	54.28±0.16	85.86±0.58	84.99±1.33	86.16±1.42
WDBC	354.11±2.20	644.89±1.65	631.37±6.92	639.12±7.00
Dermatology	407.38±0.73	477.44±1.73	425.31±1.26	476.22±2.45

3. IMPROVED ABC FOR CLUSTERING

In order to improve the performance of ABC-clustering method, we propose a new initialization method which generates more available initial solutions to ABC-clustering method. The new method and its parameter will be introduced in this section.

3.1 Initialization Using Typical Data Points

In ABC, generating the initial solutions randomly makes the clustering result unstable. What is worse, if the initial solutions happen to be extremely bad, the clustering result will be unacceptable.

In clustering, data point belongs to the cluster whose center is the closest in terms of the similarity measure. Hence, if the initial centers are close to the ideal cluster centers respectively and far away from each other, the clustering process will be much more efficient.

Inspired by the above idea, this proposed initialization method uses several typical data points among the data set to be the initial cluster centers (initial solutions). Assuming that the given number of clusters is *K*, the initial cluster centers of one solution will be determined one by one according to the rules below.

1) The 1st center is a data point randomly selected from the data set.
2) The *N*th ($N \in \mathrm{Z}^+, 1 < N \leq K$) center is a data point randomly selected from *IRN* (*Initialization Reference Number*) ($IRN \in \mathrm{Z}^+$) data points which are farthest away from the centroid of the former *N-1* centers in terms of the similarity measure.

After the first initial center is randomly selected from the data set, the rest centers are selected one after another by the reference of former selected ones.

Applying this new initialization method, the selected data points can reflect the distribution of clusters in a certain degree. Such an initialization method can reduce the instability of ABC-clustering method and help ABC to find even more optimal solution.

3.2 Initialization Reference Proportion (IRP)

When clustering different data sets, the *Initialization Reference Number* (*IRN*) should make adjustment to the data sizes. Hence, *IRN* can't be set as a constant parameter. Another parameter *Initialization Reference Proportion (IRP)* is thus introduced. *IRN* can be calculated as Equation (1).

$$IRN = data\_size \times IRP \quad (1)$$
$$(IRN \in \mathrm{Z}^+, IRP \in (0,1])$$

With *IRP, IRN* can accommodate data sets with different sizes. Thus *IRP* is the most important and the only parameter in the new initialization method. Since one cluster center is selected from *IRN* data points, the *IRN* surely relates to size of that cluster. In other words, *IRP* can be determined by the proportion of clusters in the data set.

4. RESULTS AND DISCUSSION

The results of the four algorithms used for clustering are shown in Table 1. The proposed EABC outperforms the other three algorithms in clustering all five benchmark data sets. The best improvement is that the value is 45% better than the original ABC when clustering on WDBC. The other two ABC variants shows competitive performance to ABC, but still are much worse than EABC.

Due to the typical data points initialization method, EABC stands out in the beginning of the clustering process. In most cases, the other algorithms can't even reach the EABC's initial evaluations when they finish their 20000 evaluations. In real world clustering, such an initialization method is an efficient timesaver, and it generates more available initial solutions which help the ABC mechanism to find even more optimal solution.

IRP is the most important parameter in EABC. According to our experimental results and analysis, the suggested valid range of *IRP* is [0.3, 0.7], which can achieve more available initial solutions for clustering.

5. CONCLUSION

In this literature, we have proposed an initialization method using typical data points for the ABC-clustering process. Lots of experiments demonstrate that such an initialization mechanism leads to better clustering evaluation, and at the same time, much less time consumption. To improve the clustering applicability of this new initialization method, its crucial parameter *Initialization Reference Proportion* (*IRP*) has been explored. A valid range [0.3, 0.7] of *IRP* is finally proposed.

We plan to extend EABC to support automatic clustering and apply the new initialization method to other clustering algorithms.

6. ACKNOWLEDGEMENTS

This work is supported by grants 61272397 and 61173018 from the National Natural Science Foundation of China (NSFC).

7. REFERENCES

[1] Jain A K, Murty M N, Flynn P J. Data clustering: a review. ACM computing surveys (CSUR), 1999, 31(3): 264-323.

[2] Van der Merwe D W, Engelbrecht A P. Data clustering using particle swarm optimization, The 2003 IEEE Congress on Evolutionary Computation (CEC'03), Canberra, Australia. 2003, 1: 215-220.

[3] Shelokar P S, Jayaraman V K, Kulkarni B D. An ant colony approach for clustering. Analytica Chimica Acta, 2004, 509(2): 187-195.

[4] Karaboga D, Basturk B. A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. Journal of global optimization, 2007, 39(3): 459-471.

[5] Karaboga D, Ozturk C. A novel clustering approach: Artificial Bee Colony (ABC) algorithm. Applied Soft Computing, 2011, 11(1): 652-657.