International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:7, No:7, 2013

Quantity and Quality Aware Artificial Bee Colony Algorithm for Clustering

U. Idachaba, F. Z. Wang, A. Qi, and N. Helian

***Abstract*—**Artificial Bee Colony (ABC) algorithm is a relatively new swarm intelligence technique for clustering. It produces higher quality clusters compared to other population-based algorithms but with poor energy efficiency, cluster quality consistency and typically slower in convergence speed. Inspired by energy saving foraging behavior of natural honey bees this paper presents a Quality and Quantity Aware Artificial Bee Colony (Q^2ABC) algorithm to improve quality of cluster identification, energy efficiency and convergence speed of the original ABC. To evaluate the performance of Q^2ABC algorithm, experiments were conducted on a suite of ten benchmark UCI datasets. The results demonstrate Q^2ABC outperformed ABC and K-means algorithm in the quality of clusters delivered.

***Keywords*—**Artificial bee colony algorithm, clustering.

I. Motivation

CLUSTERING, a process that aims to group observed sample of data or objects into homogeneous classes based on similarity of their observed attributes has found application in many areas such as web mining, image segmentation, security, textual document collection, artificial intelligence, pattern recognition, oncology, paleontology, pathology, psychiatry, geology, geography, psychology, sociology, archaeology, marketing segmentation and business strategy [1], [2]. Cluster analysis has drawn the interests of researchers from various disciplines in the creation, use and modification of its underlying methods [3].

II. Problems Identified

Artificial bee colony (ABC) algorithm is a relatively new swarm intelligence technique for clustering [4]. It produces higher quality clusters compared to other population based algorithms but with poor energy efficiency, consistency and typically slower in convergence speed [5]. ABC recruitment activity is probability based which creates the likelihood of having occurrences of disproportionate match between required exploitation effort and provided exploitation effort. The probability based recruitment activity also creates inconsistency in the quality of solution delivered. ABC provides limited scope for exploitation. This limit in scope leads to repeated exploitation hence an ineffective and inefficient use of exploitation efforts. ABC limit of abandonment approach exposes the algorithm to abandoning good solutions and the exploitation efforts used in identifying the solutions.

U. Idachaba is with the Future Computing Group, University of Kent, Kent, CT2 7NF UK (phone: +44 (0)1227 823192; e-mail: usi2@ kent.ac.uk).

F. Z. Wang, is a Professor with University of Kent and Head of the School of Computer Science, University of Kent, Kent, CT2 7NF UK (phone: +44 (0)1227 823192; e-mail: F.Z.Wang@kent.ac.uk).

A. Qi is a Professor with the Future Computing Group, University of Kent, Kent, CT2 7NF UK (phone: +44 (0)1227 823192; e-mail: qiailing1@126.com).

N. Helian is a Senior Lecturer with the University of Hertfordshire, Hertfordshire, AL10 9AB UK (phone: +44 (0)1707 284000; e-mail: n.helian@hertz.ac.uk).

III. Approach

Inspired by energy saving foraging behavior of natural honey bees this paper presents a quality and quantity aware artificial bee colony (Q^2ABC) algorithm for clustering to improve quality of cluster identification, energy conservation and convergence speed of ABC. We modified three main foraging activities: Recruitment, Exploitation and Abandonment. We introduced a structured approach using a new equation relating to quality and quantity for recruitment to ensure lateral proportion based exploitation efforts. We also introduced a repellent aware approach to avoid already failed choices and, a quality and quantity abandonment approach to avoid false positive abandonments.

We carried-out experiments in two phases to evaluate the performance of Q^2ABC algorithm. In the first phase, we compared the performance of ABC algorithm to both K-means and the proposed Q^2ABC algorithm in terms of cluster quality. We chose K-means for comparison because it is popular and shares same cluster representation (centroid) as ABC. In the second phase we carried-out a further comparative analysis between the ABC algorithm and proposed Q^2ABC in terms of the respective quality of clusters identified and associated processing speed and energy efficiency. We used a suite of ten randomly chosen UCI dataset (Amazon Commerce Reviews Set, Blood Transfusion Service Centre, Breast Cancer Wisconsin Original, Congressional Voting Records, Contraceptive Method Choice, Dermatology, Flags, MSNBC.com Anonymous Web Data, Post-Operative Patient, and Statlog Heart.) for the comparison [6]. We used *objective function* (*Euclidean distance*), summation of *mutation counts* and *execution time* (*minutes*) representing quality, energy efficiency and processing speed respectively as metrics for the comparison. Performance of these algorithms is represented by how small the objective function and execution time values are and how large the mutation count is. All three algorithms had maximum cluster count set at three and run against each dataset thirty times except for the Amazon data which had sixty runs against it. The ABC and proposed Q2-ABC algorithms had three hundred iterations in each run with limit

of abandonment set to the product of dimension and given maximum cluster count. A variation in cluster count output was observed in the results of the experiment. Consequently a selective comparison comparing results with same cluster count output was used.

IV. Results

Figs. 1-6 show results of the experiments.[1]

A. Phase 1 Experiment

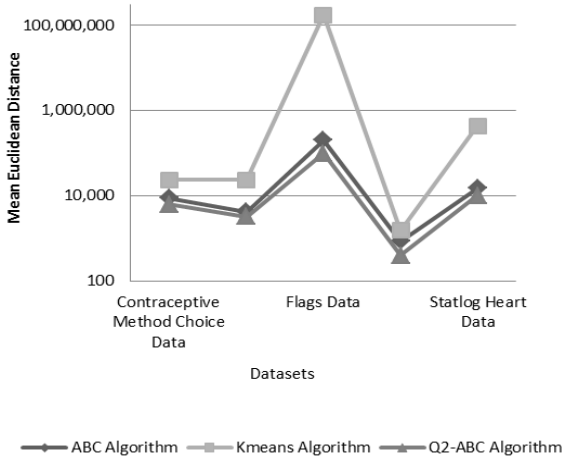

Fig. 1 Cluster quality results for ABC, K-means and Q^2ABC

B. Phase 2 Experiment

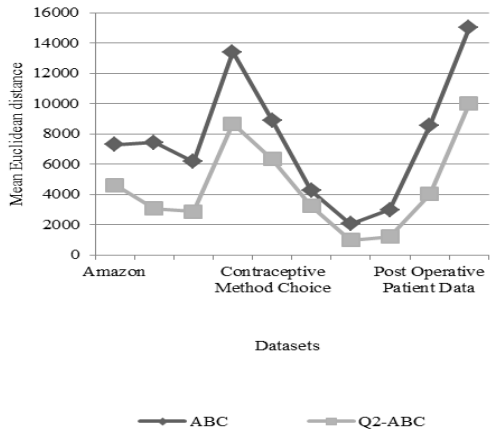

Fig. 2 ABC and Q^2ABC cluster quality

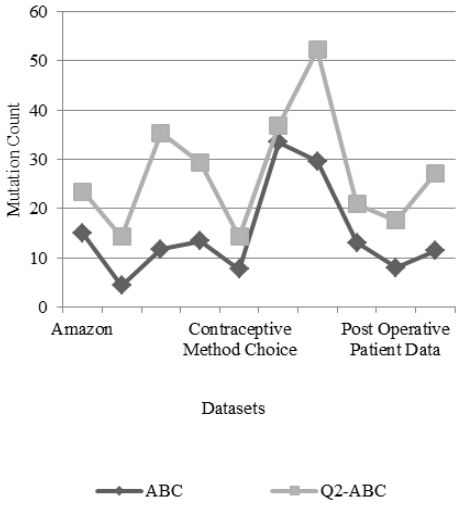

Fig. 3 ABC and Q^2ABC execution energy efficiency

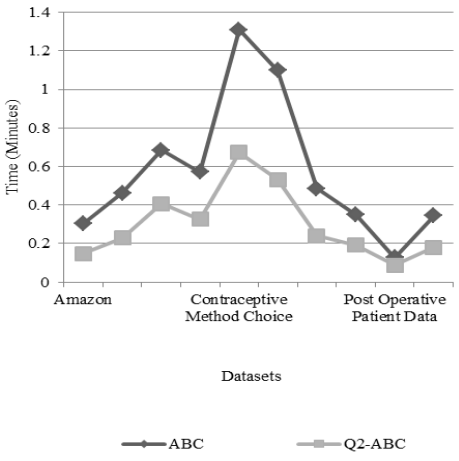

Fig. 4 ABC and Q^2-ABC execution time

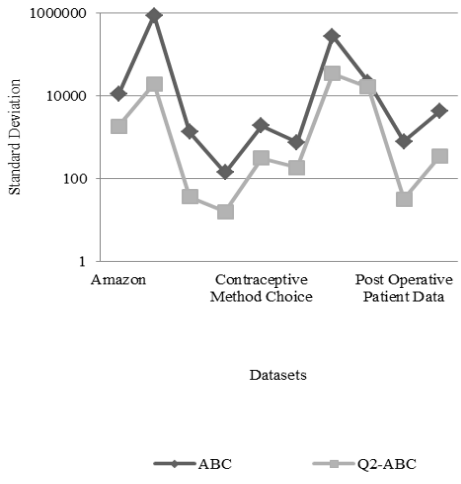

Fig. 5 ABC and Q^2ABC comparison of resultant cluster quality consistency

[1] Fig. 3 normalized data: Amazon $x10^1$, Blood Transfusion Centre $x10^2$, Breast Cancer Wisconsin Original $x10^2$, MSNBC $x10^{-1}$, Post-Operative Patient Data $x10^2$ and Statlog Heart Data Set $x10^{-1}$. Fig. 4 normalized data: Amazon x 10^1 and MSNBC $x10^1$. Fig. 5 normalized data: Amazon x 10^2 and MSNBC $x10^2$.

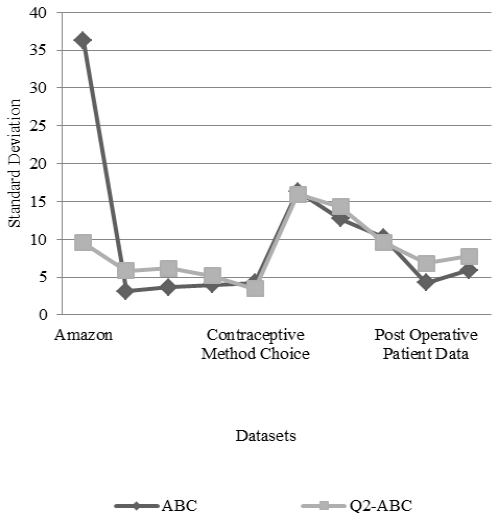

Fig. 6 ABC and Q^2ABC comparison of energy efficiency consistency

In Fig. 1 our experiment shows both ABC algorithm delivered better quality clusters compared to K-means. In Fig. 2 the mean value for quality of clusters identified by Q^2ABC was at least 24% better than the ABC and in five of the ten test cases over 52% better. This was achievable as a result of energy conservation through our repellent scent awareness approach to effectively ensure the optimal use of energy within individual iterations to promote a faster convergence to optimum cluster quality. Our quantity and quality based approach for food source abandonment avoids the likelihood of abandoning potentially good food sources a weakness inherent in the ABC. This does not only protect potentially good solutions from abandonment, it also protects efforts already used to generate them. The stochastic nature of ABC produces inconsistency in the quality of solutions it generates. Using standard deviation, Fig. 5 shows Q^2ABC to have over 83% better consistency compared to ABC on all datasets except for MSNBC dataset. Substituting our proportion based quality and quantity recruitment approach for the probability base recruitment approach of the ABC produced stable and more reliable solutions. The combination of proportionality based distribution of labour and the circumvention of repetition realized through our repellent scent awareness approach reduced execution time. Fig. 4 shows a mean execution time reduction by over 41% in Q^2ABC for nine of the ten test cases. Fig. 3 shows a mean mutation count improvement of at least 10% with nine of the ten test cases showing over 57% increase. This demonstrates efficient use of energy. Both algorithms had same number of iterations and where required to produce same number of clusters. Q^2ABC repellent scent awareness ensured avoidance of effort waste on already tried solutions conserving energy for profitable efforts. However, consequence to the heavy dependence of this approach on random selection the ABC showed better mutation count stability on most of the datasets compared to Q^2ABC. See Fig. 6. From Tables I-III we see that the better performance of Q^2ABC over ABC has statistical significance.

TABLE I
CLUSTER QUALITY: STATISTICAL SIGNIFICANCE

	ABC	*Q^2-ABC*
Mean	7597.206549	4496.549
Variance	17408203.94	8995759
Observations	10	10
Pearson Correlation	0.963079338	
Hypothesized Mean Difference	0	
df	9	
t Stat	6.465278862	
P(T<=t) one-tail	0.0000580259	
t Critical one-tail	1.833112933	
P(T<=t) two-tail	0.000116052	
t Critical two-tail	2.262157163	

TABLE II
ENERGY EFFICIENCY: STATISTICAL SIGNIFICANCE

	ABC	*Q^2-ABC*
Mean	14.78876	27.16296
Variance	88.64001	141.2135
Observations	10	10
Pearson Correlation	0.820403	
Hypothesized Mean Difference	0	
df	9	
t Stat	-5.75202	
P(T<=t) one-tail	0.000138	
t Critical one-tail	1.833113	
P(T<=t) two-tail	0.000276	
t Critical two-tail	2.262157	

TABLE III
EXECUTION TIME: STATISTICAL SIGNIFICANCE

	ABC	*Q^2-ABC*
Mean	0.573078859	0.300777
Variance	0.136306753	0.034103
Observations	10	10
Pearson Correlation	0.991419596	
Hypothesized Mean Difference	0	
df	9	
t Stat	4.588283786	
P(T<=t) one-tail	0.000656116	
t Critical one-tail	1.833112933	
P(T<=t) two-tail	0.001312233	
t Critical two-tail	2.262157163	

V. CONCLUSION

Bio Inspired clustering techniques optimizes quality but at the expense of cost and time. Our proposed approach improves the quality with reduced cost and time by avoiding repetitive exploitation, abandonment of good solutions and mismatch between required exploitation effort and provided exploitation effort. In our typical scenario, the quality of clusters identified by Q^2ABC was between 24% and 52% better than ABC; the mutation count for Q^2ABC was between 10% and 57% better than ABC and the execution time by Q^2ABC was between 30% and 52% better than ABC. However, Q^2ABC unlike ABC requires memory to support its repellent scent feature hence our future work would focus on investigating repellent scent life span influence on performance pattern and, effective ways to be proactive and

not just reactive to repellent scent.

REFERENCES

[1] E. Bonabeau, and C. Meyer, "Swarm intelligence," Harvard Business Review, vol. 79, no. 5, pp. 106-114, 2001.
[2] D. T. Pham, S. Otri, A. Afify,M. Mahmuddin, and H. Al-Jabbouli, "Data Clustering Using the Bee Algorithm," in . In Proc. 40th CIRP International Manufacturing Systems Seminar, 2007.
[3] A. Jain, M.N. Murty, and P.J. Flynn, "Data clustering: a review," ACM computing surveys (CSUR), vol. 31, no. 3, pp. 264-323, 1999.
[4] D. Karaboga, and C. Ozturk, "A novel clustering approach: Artificial Bee Colony (ABC) algorithm," Applied Soft Computing, vol. 11, no. 1, pp. 652-657, 2011.
[5] W. Gao, and S. Liu, "Improved artificial bee colony algorithm for global optimization," Information Processing Letters, vol. 111, pp. 871-882, 2011.
[6] A. Frank, and A. Asuncion, "UCI Machine Learning Repository," University of California, California, 2010.