Gradual Health Improvement in Genetic Algorithm for Clustering

A. H. Beg, and Md Zahidul Islam

School of Computing and Mathematics,
Charles Sturt University, Australia

Current Issues

- **Clustering**
  - Many popular clustering techniques including K-means require a user (data miner) to define the number of clusters *k*.
  - It is often very difficult for a user to guess the number of *k* in advance.
  - Many existing techniques like K-means also have a tendency of getting stuck at local optima.
- **Evolutionary Algorithm based Clustering**
  - In order to addressed the clustering issues various evolutionary algorithm based clustering techniques have been proposed. Typically, they choose the initial population randomly, whereas carefully selected initial population can improve final clustering results.
  - In Genetic Algorithm (GA) based clustering techniques the gene re-arrangement and twin removal are crucial to finally find a good solution.

Motivation

- Automatically find right number of clusters and identify right seeds [1,7].
- High quality chromosomes selection in the initial population [7].
- Maintain gradual health improvement of the chromosomes of a generation.
- Gene re-arrangement and twin removal [1].
- Improve chromosome quality through crossover and mutation.

Proposed Method

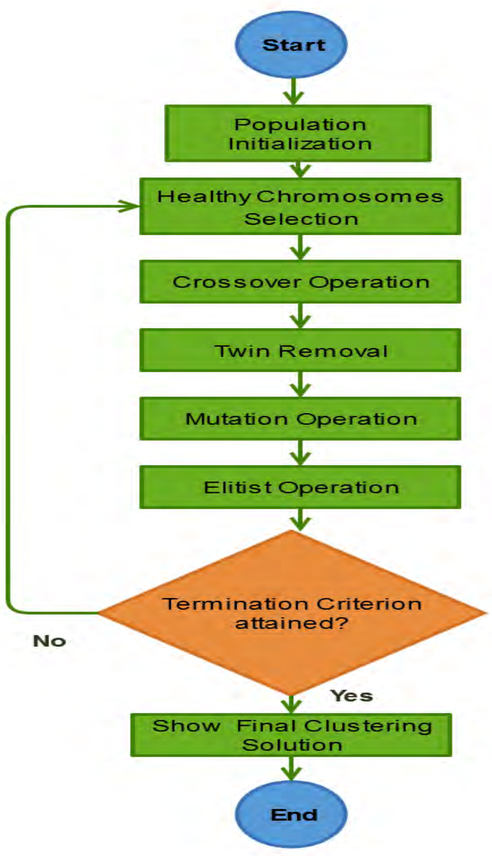

Fig. 1. Flow diagram of GAClust

We propose a genetic algorithm based clustering technique termed as GAClust. GAClust takes a dataset *D* as input. It first normalizes all numerical attributes separately in order to weigh each attributes equally. It then generates high quality chromosomes in the initial population through two phases: deterministic and random.

The Healthy Chromosomes Selection operation is applied from the 2nd iteration. In the healthy chromosomes selection operation GAClust compares the chromosomes between two generation (Current and Previous) and selects a set of chromosomes probabilistically based on their fitness.

The Two Phases of crossover, Twin Removal and Mutation operation are then applied sequentially. In the mutation operation GAClust applies division and absorption operation. At the end of each iteration GAClust applies the elitist operation in order to keeps track of the best chromosome found so far.

Results

We empirically compare our technique (GAClust) with 5 existing techniques called GenClust [1], AGCUK [2], GAGR [3], K-means ++[4] and K-Means [5] on 7 natural data sets that are available in the UCI machine learning repository [6]

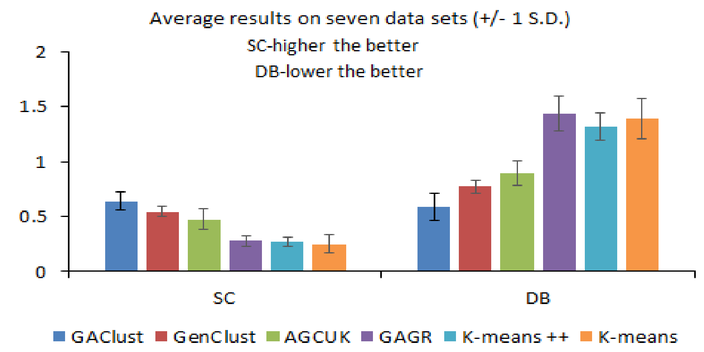

Fig. 2 . Comparative result between GAClust and other techniques

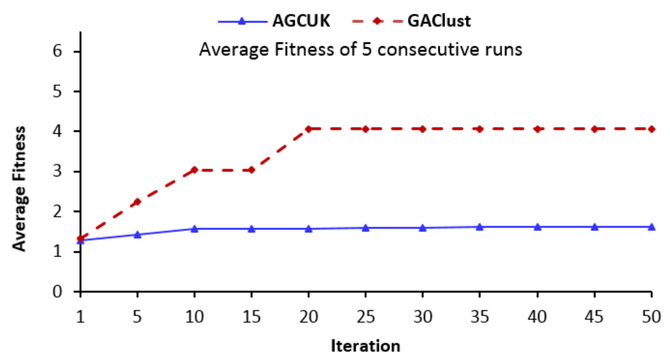

Fig. 3. Average Fitness of best chromosome on PID data set.

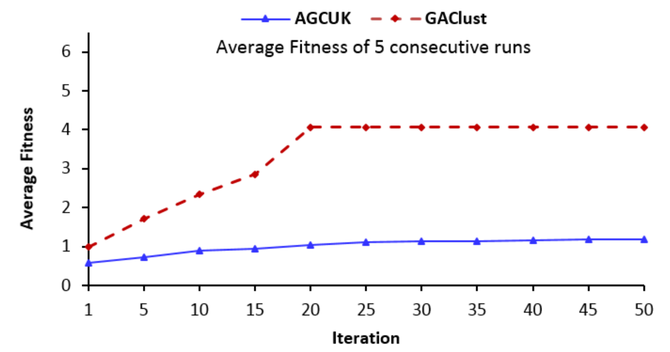

Fig. 4. Average Fitness of all chromosomes on PID data set.

References

1. M. A. Rahman, & M. Z. Islam, A hybrid clustering technique combining a novel genetic algorithm with K-Means, Knowledge Bases Systems. 71(2014) 345-365.
2. Y. Liu, X. Wu, Y. Shen, Automatic clustering using genetic algorithms, Applied Math. and Comp. 218 (2011) 267-1279.
3. D. Chang, X. Zhang, C. Zheng, A genetic algorithm with gene rearrangement for K-means clustering, Pattern Recognition. 42 (2009) 1210-1222.
4. S. P. Lloyd, Least squares quantization in PCM, IEEE Trans. On Information Theory. 28 (1982) 129-13.
5. D. Arthur, & S. Vassilvitskii, k-means++: The Advantages of Careful Seeding, SODA '07 Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, 2007, pp.1027-1035.
6. UCI Machine Learning Repository. <http://archive.ics.uci.edu/ml/datasets.html/> (accessed 22.06.13).
7. A.H. Beg, & M.Z. Islam, Clustering by Genetic Algorithm-High Quality Chromosome Selection for Initial Population. 10th IEEE conference on Industrial Electronics and Applications, Auckland, New Zealand. 2015, pp. 129-134.

++++

Contact details

Abul Hashem Beg

Phone: 02- 6338 4284

Email: abeg@csu.edu.au