Multi-objective clustering with Genetic Algorithm

Combining K-Means with a Genetic Algorithm in multiple stages

GenClust++ genetic algorithm-based clustering method, based on GenClust technique, proposed in Rahman & Islam, 2014, is briefly described below (details appear in the original paper (Islam et al., 2018).

Initialization. The numerical attributes of the input dataset are normalized first to the domain [0, 1], followed by the initialization phase of the genetic algorithm. A novel initial population selection is applied producing a good quality and diverse initial population. The starting population is first built in two stages, composed of typically 90 (3**s*, s is size of initial population) chromosomes. In order to enable a discovery of clustering solution with a higher number of clusters, in the first stage chromosome are generated using MK-Means or MK-Means++ (modified versions K-Means and K-Means++ working with categorical attributes) with different number of clusters in range [2…10], as most publicly available datasets have no more than 10 clusters. K-Means with randomly generated values of k taking values in range [2…$\sqrt{n}$], where n is the number of records, is used in the second stage. The initial population of s chromosomes is then chosen from the starting population probabilistically with a higher probability awarded to chromosomes with better fitness. It's worth mentioning that besides different values of k encoded in chromosomes of the initial population, other k values are explored as well over subsequent generations due to the various genetic operations.

Genetic operators. Besides using existing genetic operators, crossover, gene rearrangement, mutation and elitism, utilized by the same authors in Rahman & Islam, 2014 work (details in the original paper), GenClust++ makes use of novel operations as well. A probabilistic cloning operation is performed at every 10th iteration instead of crossover in order to accelerate convergence. Chromosomes are cloned with the probability proportional to the fitness of the chromosomes according to the wheel techniques. Mutation and elitism operations are applied then to introduce diversity. This is further followed by elitism operation and 15 iterations of MK-Means hill-climber applied on all of the population chromosome, repairing potentially damaged by mutation chromosome as a result. In addition, starting from 10th generation, after the probabilistic cloning is performed, chromosome selection operation, a second type of elitism, is carried out at current and every subsequent iteration. The operation merges all the chromosomes of the most recent population and the generation resulting from all genetic operations and chooses the highest fitted s chromosomes from the merged population as the next population.

Termination criteria and final solution. A maximum number of iteration is applied to terminate the algorithm. After the last genetic generation is generated, the best chromosome is used as the initial solution of a final full-length MK-Means to deliver the final clustering solution.

Multi-objective GenClust++

Many conventional nature-inspired clustering algorithms manage to efficiently group linearly separable clusters, but fail on non-linearly separable ones. Therefore, besides the growth of

interest in automatic clustering, a strong tendency in using multi-objective algorithms is found nowadays to address non-linearly separable problems. Thus, a multi-objective version of GenClust++ (MGenClust++) algorithm has been implemented in this study to capture different characteristics of various datasets.

Objective functions. MGenClust++ optimizes two complementary objectives, connectivity and cohesion, instead of only one (i.e. Davies-Bouldin Index) utilized in the originally proposed algorithm.

MaxiMin Strategy. The concept of MaxiMin (Simon, 1958) strategy is applied to determine pareto optimal solutions. The strategy is utilized in chromosome selection operation, which merges all chromosomes between the most recent population and the generation resulting from all genetic operations. MaxiMin is used here to choose the best chromosomes from the merged populations for the next population. First, non-dominated solutions are chosen and added to the next population. At this point, next population may not be filled up completely. Therefore, weakly dominated solutions are sorted by the fitness score and the fittest individuals are added to the next population until it's filled up. The MaxiMin is also used in the process of determining the final best chromosome, used as the initial solution of a final full-length K-Means to deliver the final clustering solution. A set of non-dominated solutions is determined using MaxiMin strategy, out of which the fittest chromosome is then chosen as the initial solution of the final full-length K-Means.

Fitness. To enable the genetic operations such as probabilistic selection that depend on the fitness values of chromosomes, a fitness function has to be calculated. DB Index serves as a fitness function in the originally proposed algorithm. However, connectivity and cohesion objectives are already being optimized, and the use of another objective such as DB Index would cause confusion and inconsistency in the genetic algorithm framework in the sense that one part of the framework would use multiple related conflicting objectives such as connectivity and cohesion and the other part would utilize an unrelated to them single objective such as DB Index. Therefore, the multi-objective clustering problem is transformed in this case into a mono objective one by means of a weighted formula in order to allow the fitness function to be defined using the same multiple conflicting objectives. Given clustering solution x_i of ith particle in the swarm and its corresponding values of objective functions $F_i = \{f_1(x_i), f_2(x_i), \ldots, f_n(x_i)\}$, the implemented weighted sum fitness function is computed using the distances of the objective values from an ideal solution, called utopia point. The fitness is thus defined as follows:

$$Fitness_i = \left[\Sigma\, w_j * (f_j(x_i) - f_j{}^*)^p\right]^{1/p},$$

where w_j is the weight of the distance for jth objective function, $f_j{}^*$ is the best coordinate of utopia point obtained for jth objective function during the optimization process. The value of p is set to 2.

References

M. A. Rahman and M. Z. Islam, "A hybrid clustering technique combining a novel genetic algorithm with K-Means," *Knowledge-Based Systems*, vol. 71, pp. 345–365, Nov. 2014.

M. Z. Islam, V. Estivill-Castro, M. A. Rahman, and T. Bossomaier, “Combining K-Means and a genetic algorithm through a novel arrangement of genetic operators for high quality clustering,” *Expert Systems with Applications*, vol. 91, pp. 402–417, Jan. 2018.

H. A. Simon, “Review of Games and Decisions: Introduction and Critical Survey,” *American Sociological Review*, vol. 23, no. 3, pp. 342–343, 1958.