Efficient clustering with a self-adaptive genetic algorithm

Juha Kivijärvi
Turku Centre for Computer Science
Department of Mathematical Sciences
University of Turku
FIN-20014 Turku, Finland
Email: juhkivij@utu.fi
Tel: +358 2 333 8797

Pasi Fränti
Department of Computer Science
University of Joensuu
PB 111
FIN-80101 Joensuu, Finland
Email: franti@cs.joensuu.fi
Tel: +358 13 251 3103

Olli Nevalainen
Turku Centre for Computer Science
Department of Mathematical Sciences
University of Turku
FIN-20014 Turku, Finland
Email: olli.nevalainen@utu.fi
Tel: +358 2 333 8631

We give a *self-adaptive genetic algorithm* (SAGA) for the *clustering problem* (Kaufman and Rousseeuw, 1990). We assume that the number of clusters is fixed and the objects to be clustered are k-dimensional vectors in an Euclidean space.

Genetic algorithms (GAs) have turned out to be very effective in solving the clustering problem (Fränti et al., 1997). However, they have many parameters, the optimal selection of which depends on the problem instance. Furthermore, their optimal values may change during the execution of the algorithm.

Our aim is to give an effective and straightforward algorithm which obtains good solutions for the problem without explicit parameter tuning. The self-adaptation takes place at individual level (Hinterding et al., 1997), which means that each individual of the GA population contains a set of parameters. These are used in the reproduction process. The parameters we adapt are the *crossover method* γ, *mutation probability* ψ and *noise range* ν. Thus, an individual ι is of the form $\iota = (\omega_\iota, \gamma_\iota, \psi_\iota, \nu_\iota)$ where ω_ι is a solution to the clustering problem.

The general structure of SAGA is following:

1. Generate S random individuals to form the initial generation.
2. Iterate the following T times.
   (a) Select S_B surviving individuals for the new generation.
   (b) Select $S - S_B$ pairs of individuals as the set of parents.
   (c) For each pair of parents (ι_a, ι_b) do:
      i. Determine the strategy parameter values $(\gamma_{\iota_n}, \psi_{\iota_n}, \nu_{\iota_n})$ for the offspring ι_n by crossing the strategy parameters of the two parents.
      ii. Mutate the strategy parameter values of ι_n with the probability Ψ.
      iii. Create a new solution ω_{ι_n} by crossing the solutions of the parents. The crossing method is determined by γ_{ι_n}.
      iv. Mutate the solution of the offspring with the probability ψ_{ι_n}.
      v. Add noise to ω_{ι_n}. The maximal noise is ν_{ι_n}.
      vi. Apply *k-means* iterations to ω_{ι_n}.
      vii. Add ι_n to the new generation.
   (d) Replace the current generation by the new generation.
3. Output the best solution of the final generation.

Our experiments show that the results of SAGA are fully comparable with those of a carefully refined non-adaptive genetic algorithm. The major benefit of SAGA is that the refining phase can be omitted completely. The algorithm finds the best or at least a satisfying parameter combination and is able to obtain a very good solution practically always. A drawback of the method is its rather large running time.

References

P. Fränti, J. Kivijärvi, T. Kaukoranta and O. Nevalainen (1997). Genetic algorithms for large scale clustering problems. *The Computer Journal* **40** (9), 547–554.

R. Hinterding, Z. Michalewicz and A.E. Eiben (1997). Adaptation in evolutionary computation: A survey. In *Proceedings of the 4th IEEE International Conference on Evolutionary Computation*, Indianapolis, 65–69.

L. Kaufman and P.J. Rousseeuw (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, New York: John Wiley & Sons.