Influences of Clustering modifications on the performance of the Genetic Algorithm driven Clustering algorithm

Dirk Devogelaere, Marcel Rijckaert

K.U.Leuven, Chemical Engineering Department
De Croylaan 46, B-3001 Leuven,Belgium
Email: {dirk.devogelaere}@cit.kuleuven.ac.be
Tel: +16 32 23 68

SUMMARY

One way to look at basic modeling approaches is to split them up into mechanistic and data based models. A few years ago we developed our own data based model approach [1], called **G**enetic **A**lgorithm **d**riven **C**lustering (GAdC). The proposed methodology relies on semi-supervised clustering with a generative floating-point genetic algorithm and local learning. In this contribution we investigate the influence of clustering modification on the performance of the prediction of a real world application [2]. The task is the prediction of alga frequency distributions on the basis of the measured concentrations of the chemical substances, the global information concerning the season when the sample was taken, the river size and its flow velocity.

We deal with an evolutionary algorithm (EA) by implementing the GAdC as a generative floating-point genetic algorithm. An EA acts on a set of individuals. An individual is a representation of a point within the search space of the EA. In its simplest form, this individual is represented as a one-dimensional string of variables, called a chromosome. Each chromosome of the EA represents the coordinates of the cluster centers and a scaling factor for each dimension. If the dimensionality of the data is D, and there are K cluster centers, there will be D*K genes for the cluster centers and D genes for the scaling factors. The chromosome can be evaluated. This means that a certain fitness value (based on the objective value) is assigned to the individual depending on the problem at hand. In our case, the chromosome is decoded into a solution of the clustering. This solution is evaluated ("goodness of prediction") and the value is assigned to the chromosome in the EA.

The research in this paper is focused on how the performance of prediction is influenced by choosing a representative for the cluster centers. In all the variants, the genetic algorithm (GA) determines the cluster centers. By replacing the value determined by the GA in the non-empty clusters, we influence the mapping realized between the search space and the solution space of the GA. Different cluster centers in the search space might resolve to the same distribution of the cases over the different clusters, resulting in the same solution in the solution space. The EA is not aware of the similarity of these individuals. To overcome redundancy in the individuals' space, the EA has to update the offspring . Two variants for replacing the value proposed by the GA for the cluster centers were implemented. In the first variant (centroid), we replace the GA value by the mean value of the cluster calculated based on the elements in the cluster. In the second variant (closest) we replace the cluster center by the closest element of the cluster to the GA value. The variant without replacement is called "standard".

For each variant we predicted the outcome of all the seven algae distributions 30 times. For each algae distribution the mean squared error on the test set and the standard deviation were calculated depending on the variant used to update the cluster centers. Contrary to what was expected neither the centroid nor the closest variant performs better on the test set in general. Another way to present the results is plotting the error on the test set versus the ultimate fitness value obtained during training. There is a general tendency that training stops at lower fitness values in the case of the closest variant, while training stops at the higher fitness values in the case of the standard variant. As a consequence of the way of mapping the two variants "closest" and "centroid" cover a subspace of the solution space of the "standard" variant. This might indicate that less generations of the GA are necessary to achieve the same error on the test set. Secondly it might be important, as it is in training of neural networks, that training should be stopped at the right moment. If not, a kind of over training occurs and worse results on the test set are obtained. The results of a preliminary run indicate that the test error indeed seems to decrease but after obtaining a minimum at about 250 generations the error smoothly increases again as a function of the number of generations.

References

[1] D. Devogelaere, P. Van Bael, and M. Rijckaert (1999). Regression Through Genetic Algorithm driven Clustering, *Proceedings of the European Conference on Intelligent Techniques and Soft Computing (EUFIT),* September 7-10, Aachen, Germany.
[2] URL (1999) ftp.mitgmbh.de/pub/problem.zip