Genetic fuzzy clustering algorithm with greedy crossing-over procedure

Dmitry Stashkov and Lev Kazakovtsev[⋆]

Reshetnev Siberian State Aerospace University, Krasnoyarsk, Russia
stashkov@ngs.ru,levk@bk.ru

Results of testing of the electronic components shipped for the space industry are represented by arrays of data vectors of very high dimensionality, up to hundreds of dimensions. One of the most important problems for increasing the quality of the electronic units is detection of the homogeneous production batches of the electronic devices. We consider the problem of fuzzy clustering of data sources applying new genetic algorithm with greedy agglomerative heuristic based on the EM algorithm for precipitation of mixture of Gaussian distributions.

Each solution (individual) in this algorithm is represented by a pair $< D, W >$ of a set of the distributions $D = \{N(\mu_i, \sigma_i^2, i = \overline{1,k})\}$ and a set of weight coefficients of distributions $W = \alpha_i, i = \overline{1,k}$. During the crossing-over procedure, sets of two selected pairs $< D', W' >$ and $< D'', W'' >$ are joined: $D = D' \cup D''$, $W = W' \cup W''$, then the following greedy procedure [1] runs:

1. Run the EM algorithm with $< D, W >$.

2. If $|D| = k$ then stop.

3. For each $i' \in \{\overline{1, |D|}\}$ do: assign the truncated set $D' = D \setminus \{N(\mu_{i'}, \sigma_{i'}^2)\}$, $W' = W \setminus \{\alpha_{i'}\}$. Run one iteration of the EM-algorithm with D and W. Calculate the objective likehood function L for the received result, store its value in L_i. Continue loop 3.

4. Find index $i'' = \arg\max_{i'=\overline{1,K}} L_{i'}$. Assign the truncated sets $D = D \setminus \{N(\mu_{i''}, \sigma_{i''}^2)\}$, $W = W \setminus \{\alpha_{i''}\}$. Run the EM-algorithm for D and W. Go to Step 2.

Computational experiments with electronic component testing data and classical datasets for clustering problems show that this new algorithm allows to obtain more precise results in comparison with classical EM algorithm and its modifications. For tests of the IC 1526LE2 (number of data vectors$N = 3987$, dimensionality $d = 206$) average result of likelihood function of new algorithm is 443491, average result of EM-algorithm is 350292.

References

1. Orlov, V.,I., Stashkov, D.,V., Kazakovtsev L.A., Stupina A.A.: Fuzzy Clustering of EEE components for Space Industry. IOP Conference Series: Materials Science and Engineering. 155, Article ID 012026, 6 pages (2016) DOI: 10.1088/1757-899X/155/1/012026

[⋆] Results were obtained in the framework of the state task No. 2.5527.2017 (Basic part) of the Ministry of Education and Science of the Russian Federation.