Application of a Genetic Algorithm to Variable Selection in Fuzzy Clustering

Christian Röver and Gero Szepannek

Fachbereich Statistik,
Universität Dortmund,
44221 Dortmund, Germany
roever@statistik.uni-dortmund.de
gero.szepannek@web.de

Abstract. In order to group the observations of a data set into a given number of clusters, an 'optimal' subset out of a greater number of explanatory variables is to be selected. The problem is approached by maximizing a quality measure under certain restrictions that are supposed to keep the subset most representative of the whole data. The restrictions may either be set manually, or generated from the data. A genetic optimization algorithm is developed to solve this problem.
The procedure is then applied to a data set describing features of sub-districts of the city of Dortmund, Germany, to detect different social milieus and investigate the variables making up the differences between these.

References

FRALEY, C. and RAFTERY, A.E. (2002): `mclust`: Software for model-based clustering, density estimation and discriminant analysis. *Technical Report, Department of Statistics, University of Washington.* See `http://www.stat.washington.edu/mclust`.

GOLDBERG, D.E. (1989): *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley, Boston.

HALL, M.A. (1999): Correlation-based feature subset selection for machine learning. *PhD thesis, Department of computer science, University of Waikato.*

IHAKA, R. and GENTLEMAN, R. (1996): R: A language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics 5, Nr. 3, 299-314.* See also `http//:www.r-project.org`

KAUFMAN, L. and ROUSSEEUW, P.J. (1990): *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley, New York.

Keywords

FUZZY CLUSTERING, VARIABLE SELECTION, GENETIC ALGORITHM, CLASSIFICATION ENTROPY