Flexible Implementations of Cluster Analysis and Mixture Models

Friedrich Leisch
Institut für Statistik, Ludwig-Maximilians-Universität München
Ludwigstraße 33, 80539 München, Germany
`Friedrich.Leisch@stat.uni-muenchen.de`

Cluster analysis and latent class regression are popular methods for grouping observations into unobserved segments. Two major groups of algorithms are traditional distance/similarity-based clustering algorithms like K-means, and model-based methods like finite mixture models. We present the general design principles of the R packages `flexclust` and `flexmix` which implement popular algorithms from these two groups in an extensible way, such as generalized K-means, QTclust and several variants of EM for finite mixtures.

At first sight the only commonality of these algorithms is that they try to find unobserved groups in data. However, another important common aspect is that the optimization techniques used are rather general and not bound to a particular distance measure or stochastic model. Using a driver concept similar to `glm()` families, both packages allow the user to extend the provided functionality and easily program new methods. Both partitioning clustering methods and mixture models can be fitted much faster and with highter precision if a priori grouping information is available and used for model fitting.

Finally we discuss several new features of the packages, like clustering with group constraints, mixture models with varying and fixed effects, and visualization techniques for model parameters, cluster validity, and cluster separation. Examples are taken from application domains of market segmentation and the analysis of microarray data.

References

Bettina Grün and Friedrich Leisch. Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics & Data Analysis*, 2006. Accepted for publication.

Friedrich Leisch. Neighborhood graphs and shadow plots for cluster visualization. Department of Statistics, University of Munich, Germany, *Submitted*, 2006a.

Friedrich Leisch. A toolbox for k-centroids cluster analysis. *Computational Statistics & Data Analysis*, 2006b. Accepted for publication.

Friedrich Leisch and Bettina Grün. Extending standard cluster algorithms to allow for group constraints. In Alfredo Rizzi and Maurizio Vichi, editors, *Compstat 2006—Proceedings in Computational Statistics*, pages 885–892. Physica Verlag, Heidelberg, Germany, 2006. ISBN 3-7908-1708-2.