Resampling Methods for Exploring Cluster Stability

Friedrich Leisch*
University of Natural Resources and Life Sciences, Vienna, Austria
Friedrich.Leisch@boku.ac.at

Model diagnostics for cluster analysis is still a developing field because of its exploratory nature. Numerous indices have been proposed in the literature to evaluate goodness-of-fit, but no clear winner that works in all situations has been found yet. Derivation of (asymptotic) distribution properties is not possible in most cases. Over the last decade several resampling schemes which cluster repeatedly on bootstrap samples or random splits of the data and compare the resulting partitions have been proposed in the literature. These resampling schemes provide an elegant framework to computationally derive the distribution of interesting quantities describing the quality of a partition. Due to the increasing availability of parallel processing even on standard laptops and desktops these simulation-based approaches can now be used in everyday cluster analysis applications. We give an overview over existing methods, show how they can be represented in a unifying framework including an implementation in R package flexclust, and compare them on simulated and real-world data. Special emphasis will be given to stability of a partition, i.e., given a new sample from the same population, how likely is it to obtain a similar clustering?

Key Words: cluster analysis, resampling methods, bootstrap, R