Cluster Stability: Concepts and Determinants

Abstract: Andreas Baumgart

The sensitivity of clustering algorithms with respect to small changes in the input is a well-known problem in data analysis – though “stability” is hardly made precise anywhere. Its reflection leads to the necessity to define clustering for arbitrary probability measures-and not just empirical ones. The next step to do is to specify metrics among finite partitions in order to quantify changes in the output. For that purpose a further look at the various cluster similarity indices proposed so far is required. To deal with additional aspects of stability, the need for laws of large numbers is stressed. In a first step, a reformulation of this problem is proposed in form of “adjacency-functions”. Finally, the role of data preprocessing is emphasized.