Studying complexity of model-based clustering

Volodymyr Melnykov and Semhar Michael *

Cluster analysis is a popular statistics and computer science technique commonly used in various areas of research. In this paper, we investigate factors that can influence clustering performance in the model-based clustering framework. The four factors chosen for consideration are the level of overlap, number of clusters, number of dimensions, and sample size. Through a comprehensive simulation study, we investigate model-based clustering in different settings. As a measure of clustering performance, we employ three popular classification indices capable of measuring the degree of agreement in two partitioning vectors, thus making the comparison between the true and estimated classification vectors possible. In addition to studying clustering complexity, the performance of the three classification measures is also evaluated.

Keywords: clustering complexity, model-based clustering, overlap, adjusted Rand index, Fowlkes and Mallows index, CARP

*Volodymyr Melnykov is an Assistant Professor of Statistics in The University of Alabama, Tuscaloosa, AL 35487, email: vmelnykov@ua.edu; Semhar Michael is a graduate student in The University of Alabama, Tuscaloosa, AL, 35487.