DSSV 2023 Data Science, Statistics & Visualisation
EC DA 2023 European Conference on Data Analysis

Quantifying variable importance in cluster analysis

C. Hennig[1] and K. Murphy[2]

[1]Department of Statistical Science "Paolo Fortunati", University of Bologna, Italy
[2]Hamilton Institute, Maynooth University, Ireland

The quantification of variable importance in cluster analysis is of interest in order to interpret and understand the impact of the variables on a clustering, and potentially also for variable selection. For general clustering methods, it can be measured by comparing a clustering with all variables with a clustering in which a variable has been left out or permuted ([1, 2]). We compare these two approaches regarding their ability to tell apart meaningful from noise variables.

A potential concern regarding clustering mixed continuous/categorical variables is that certain methods may be unduly dominated by either the continuous or the categorical variables ([3]). We address this by comparing methods such as latent class model-based clustering, distance-based clustering using Gower's distance with various weighting/standardisation schemes, or KAMILA regarding the relative importance of the continuous and categorical variables using a comprehensive simulation study and real data.

[1] C. Hennig and T. F. Liao, "Comparing latent class and dissimilarity based clustering for mixed type variables with application to social stratification," *Journal of the Royal Statistical Society, Series C*, vol. 62, pp. 309–369, 2013.
[2] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.
[3] A. H. Foss, M. Markatou, and B. Ray, "Distance metrics and clustering methods for mixed-type data," *International Statistical Review*, vol. 87, pp. 80–109, 2019.