Cluster validation based on the COSA framework

Maarten M.D. Kampert[1], Jacqueline J. Meulman[1,2], and Jerome H. Friedman[2]

1. Leiden University, 2. Stanford University

Abstract. The validation of clusters is a vast and ongoing area of research. In 2004 Friedman and Meulman proposed clustering objects on subsets of attributes (COSA), a framework that outputs a distance matrix that can be as input for a variety of proximity mapping methods. In this paper we show how the COSA framework can be adapted for cluster validation. We introduce a self-tuning algorithm (COSA validation) that takes as input a clustering structure, and returns optimal distances and a set of optimal attribute weights within each cluster. These attribute weights can be seen as importance measures indicating the attributes on which the cluster has a low variance (cluster homogeneity). We will show that when we list the observed attribute weights in descending order, and compare them with random ordered lists, a visual validation of each cluster can be obtained. Moreover, we will use this resampling strategy to derive a permutation test for cluster validation. In this paper we introduce the properties of the COSA validation framework and compare its performance with that of other methods using simulated and real-life data.

References

FRIEDMAN J.M., and MEULMAN , J.J. (2004): Clustering objects on subsets of attributes. *Journal of Royal Statistics Society Series B, 66, 815–849.*

Kampert, M.M.D. (2019): Improved strategies for distance based clustering of objects on subsets of attributes in high-dimensional data. *Monograph, Faculty of Science, Leiden University.*

Max 5 keywords

COSA, CLUSTER VALIDATION, FEATURE SELECTION