RESAMPLING FOR STABILITY ESTIMATION VS. CLUSTER VALIDATION VIA DATA SPLITTING AND SUBSAMPLING. WHICH APPROACH IS BETTER IN DETECTION OF CLUSTERS IN TAXONOMY?

Dorota Rozmus[1]

[1] Department of Economic and Financial Analysis, University of Economics in Katowice,
(e-mail: dorota.rozmus@ue.katowice.pl)

ABSTRACT: Due to the fact that there are no labels or gold standards by which performance of clustering can be measured, the problem of determining the right number of clusters (k) has not been solved to this day. However, new methods are proposed to ensure the best possible clustering performance.

KEYWORDS: Clustering, cluster stability, clustering performance.

1 Cluster stability

Clustering algorithms seek to partition data into groups, according to certain similarity measures. The overall goal is to place similar data points in the same cluster, and dissimilar data points in different clusters.

Due to the fact that there are no labels or gold standards by which performance can be measured, the problem of determining the right number of clusters (k) has not been solved to this day.

The concept of stability has emerged as a strategy for assessing the performance and reproducibility of data clustering. The underlying premise is that a good clustering of the data will be reproduced over perturbed datasets that are nearly identical to the original data.

Several methods have been developed for measuring of cluster stability. According to (Liu, Yu, Blair, 2021), these methods can be broken down into the following three categories: resampling for stability estimation, cluster validation via data splitting and subsampling, and alternative methods that do not adhere to these classic approaches. In this study only the first two approaches will be taken into consideration.

The aim of the research will be to compare these two approaches in the context of indicating the value of the *k* parameter (number of clusters). The study will be conducted on benchmark data sets, which are usually used in comparative studies.

References

DUDOIT, S., & FRIDLYAND, J. 2002. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, **3(7)**, 1–21.

FANG, Y., & WANG, J. 2012. Selection of the number of clusters via the bootstrap method. *Computational Statistics and Data Analysis*, **56**, 468–477.

HENNING, C. 2007. Cluster-wise assessment of cluster stability. *Computational Statistics and Data Analysis*, **52**, 258–271.

LIU, T., YU, H., & BLAIR, R. H. 2022. Stability estimation for unsupervised clustering: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, **14(6)**.

ŞENBABAOĞLU, Y., MICHAILIDIS, G., & LI, J. Z. 2014. Critical limitations of consensus clustering in class discovery. *Scientific Reports 4*.

TIBSHIRANI, R., & WALTHER, G. 2005. Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, **14(3)**, 511–528.

YU, H., CHAPMAN, B., DI FLORIO, A., EISCHEN, E., GOTZ, D., JACOB, M., & BLAIR, R. H. 2019. Bootstrapping estimates of stability for clusters, observations and model selection. *Computational Statistics*, **34(1)**, 349–372.