

Probability Metrics Standpoint on the Cluster Stability Problem

Zeev Volkovich, Zeev Barzily, Dvora Toledano-Kitay and Renata Avros

*Software Engineering Department,
ORT Braude College of Engineering,
Karmiel 21982, Israel
E-mail: vlvolkov@braude.ac.il / zbarzily@braude.ac.il
dvora@braude.ac.il / r_avros@braude.ac.il*

Key words: Cluster Analysis, Clustering, Partitioning, Unsupervised Learning, Cluster Stability, Two sample test

The estimation of the suggested number of clusters in the considered dataset symbolizes an ill posed problem of an essential relevance in cluster analysis. High stability in partitions, obtained from the same data source, is interpreted as a high consistency of the clustering process. Thus, the number of clusters that maximizes cluster stability can serve as an estimator for the "true" number of clusters. We offer a probabilistic model for the cluster stability. Kernel based probability metrics with appropriate two sample tests attend in the model's constructions. We claim that this sequence of clustered samples can be interpreted within the clusters as i.i.d. samples drawn from the same population if the number of clusters is chosen correctly. Thus, the problem of determining the "true" number of clusters is reduced to a hypothesis testing problem. Data outliers and clustering algorithm's shortcomings make the hypothesis rejection to be hardly expected. So, we consider an empirical distribution of the test statistics which is concentrated mostly at the origin, if the number of clusters is chosen correctly. We present numerous modifications of this approach and discuss several known stability methods from the provided methodology point of view.