

# Clustering Evaluation in High-Dimensional Data

Miloš Radovanović \*

May 4, 2019

## Abstract of Invited Presentation

Clustering evaluation plays an important role in unsupervised learning systems, as it is often necessary to automatically quantify the quality of generated cluster configurations. This is especially useful for comparing the performance of different clustering algorithms as well as determining the optimal number of clusters in clustering algorithms that do not estimate it internally. Many clustering quality indexes have been proposed over the years and different indexes are used in different contexts. There is no unifying protocol for clustering evaluation, so it is often unclear which quality index to use in which case.

In this talk, we review existing clustering quality measures and evaluate them in the challenging context of high-dimensional data clustering. High-dimensional data is sparse and distances tend to concentrate, possibly affecting the applicability of various clustering quality indexes. We analyze the stability and discriminative power of a set of standard clustering quality measures with increasing data dimensionality. Our evaluation shows that the curse of dimensionality affects different clustering quality indexes in different ways, and that some are to be preferred when determining clustering quality in many dimensions.

---

\*University of Novi Sad