

Како валидирати ненадгледано учење - кластеризација података

МЕТРИКЕ

Wang, Kaijun, Baijie Wang, and Liuqing Peng. "CVAP: Validation for cluster analyses." Data Science Journal 0 (2009): 0904220071 – имплементација урађена у MATLAB-у. [ЛИНК](#)

"To measure the quality of clustering results, there are two kinds of validity indices: external indices and internal indices.

An external index is a measure of agreement between two partitions where the first partition is the a priori known clustering structure, and the second results from the clustering procedure (Dudoit et al., 2002).

Internal indices are used to measure the goodness of a clustering structure without external information (Tseng et al., 2005).

For external indices, we evaluate the results of a clustering algorithm based on a known cluster structure of a data set (or cluster labels).

For internal indices, we evaluate the results using quantities and features inherent in the data set. The optimal number of clusters is usually determined based on an internal validity index."

Литература:

- Dudoit, S. & Fridlyand, J. (2002) *A prediction-based resampling method for estimating the number of clusters in a dataset*. Genome Biology, 3(7): 0036.1-21.
 - Thalamuthu, A, Mukhopadhyay, I, Zheng, X, & Tseng, G. C. (2006) *Evaluation and comparison of gene clustering methods in microarray analysis*. Bioinformatics, 22(19):2405-12.
1. За почетак препоручујем да прочитате пост на сајту [stats.stackexchange](https://stats.stackexchange.com). Дато је доста описа и набројано доста метрика за оба случаја, када су доступни индекси (ground-truth) и када нису.
 2. Одличан преглед метрика на једном месту - званични сајт **scikit-learn** библиотеке. [ЛИНК](#). За већину метрика су побројане како предности тако и мане.