Supplement S2: K-means clustering and validation indices

K-means clustering with missing data

The objective function for K-means clustering can be defined as:

$$\arg\min_{\mathbf{C}} \sum_{\mathbf{x}\in\mathbf{X}} d^2(\mathbf{x}, \mathbf{c}_k), \tag{1}$$

where $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}, \mathbf{x}_i \in \mathbb{R}^n$, refers to a set of N observations, and $\mathbf{C} = \{\mathbf{c}_k\}_{k=1}^{K}$ are obtained cluster profiles. $d()$ denotes modified version of the l_2-norm. Since partially incomplete data, the modified norm is needed for clustering. The main idea of the modified approach is to use pairwise available components and scale the result to the missing components (Gower, 1971) .

K-means clustering method consists of two main steps: an initialization and local refinement steps (see Algorithm 1). These steps are usually performed using multiple restarts and the result with the smallest clustering error will be selected. In an initialization step a local partition of data is decided. The quality of clustering depends on the initialization step since clustering acts locally. A local refinement step perform local search which improve quality of initial partition. The aim of this step is to minimize clustering error, that is, summed distance of observations to the nearest prototypes. The step is performed in an iterative way assigning observations to the nearest prototypes and updating prototype locations. An advantage of K-means with K-means++ type of initializations is that it has only a linear time complexity and comparable fast convergence since K-means++ favors distinct prototypes in a data space (Arthur and Vassilvitskii, 2007).

Algorithm 1: Prototype-based clustering with K-means++ initialization

Input: Data set $\mathbf{X}$ and given number of profiles K

Output: Obtained cluster profiles, which minimize the objective function (1)

Select the first profile, $\mathbf{c}_1$, as an average value of observations in $\mathbf{X}$

for $j = 2, j = j + 1, j \leq K$ **do**

Select $\mathbf{c}_j$ randomly from $\mathbf{X}$ with probability:

$\min d^2(\mathbf{x}, \{\mathbf{c}_k\}_{k=1}^{j-1}) / \sum_{\mathbf{x}\in\mathbf{X}} \min d^2(\mathbf{x}, \{\mathbf{c}_k\}_{k=1}^{j-1})$

repeat

1. Assign each observation to the closest profile using $\min d^2(\mathbf{x}, \{\mathbf{c}_k\}_{k=1}^{j})$
2. Recompute the profiles as average values of the assigned observations.

until *The partition does not change*

end

Internal cluster validation indices

In K-means setting the number of clusters is essential to be determined. Internal cluster validation indices (CVIs) identify the number of clusters such that any external/prior information is not needed in the calculations. The most of the CVIs are defined by compactness and separability of the clustering result. The validity index provides a measure for each number of clusters. Depending on the used index formula, the lowest

or the highest measure is usually selected as the final number of clusters. Further, the number of clusters can be also selected using the speed of improvement of the cluster validation measures, for example, using a classical knee-point method (Thorndike, 1953).

Our previous study (Niemelä et al., 2018) presented the most commonly used validation indices. The reduced formulas were used since constant terms and monotone functions offered in the original formulas do not affect to the final solutions. In addition, the used formulas were extended for the general similarity measure. In the study, compactness was defined by Intra and separability by Inter. Compactness is usually defined by using summed variances of observations around prototypes in different clusters. Separability indicates how well distinct clusters are for each other. Minimum or maximum values of distances of all prototypes or variance of prototypes are popularly used variables. The study proposed formulas in the form where Intra was divided by Inter and thus they were attempted to be minimized.

In general, the decision of the number of clusters by using CVIs involves the following procedure:

1) Repeat clustering iteratively ranging K from $K_{\min}$ to $K_{\max}$. Obtain calculated cluster profiles and data partitions for each value of K based on Algorithm 1.
2) Calculate index measures using CVIs for each value of K. Form index curves based on the measured values.
3) Select the optimal number of clusters according to some decision criteria, for example, minimum/maximum values of cluster validation index curves or using speed of improvements of index measures.

Regarding to the described methods, the source codes are available online: http://users.jyu.fi/~mapeniem/BJET/Kmeans/

References

Arthur, D. & Vassilvitskii, S. (2007). K-means++: the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (pp. 1027–1035). New Orleans, Louisiana, United States: Society for Industrial and Applied Mathematics.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* (pp. 857-871). Washington, United States: International Biometric Society.

Niemelä, M., Äyrämö, S., & Kärkkäinen, T. (2018). Comparison of cluster validation indices with missing data. In *Proceedings of the 26th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning* (pp. 461–466). Bruges, Belgium: ESANN.

Thorndike, R. L. (1953). Who belongs in the family. *Psychometrika*, *18*(4): 267–276.