Unsupervised learning for multiview data

Ndèye Niang

Unsupervised learning for multiview data

Ndèye Niang[1]
[1] Cédric-CNAM, Paris, France, ndeye.niang_keita@cnam.fr

With the data explosion more and more data are collected from multiple sources represented by multiple views, where each describes a perspective of the data. To deal with this kind of data in the context of unsupervised learning, one can rely on factorial approaches and clustering. Depending on the objective, these two types of methods can be used separately, successively in a two step approach or simultaneously leading to subspace clustering. In this presentation, we will review, discuss and illustrate different unsupervised approaches from the most classical to the most recent.

Keywords: unsupervised learning, multiview data, clustering, RV coefficient

---

Due to the increasing ease with which measurements can be taken and stored, more and more data are collected. This lead to high-dimensional data in which the numerous variables can be structured into homogeneous blocks representing multiple views, where each describes a perspective of the data. Data can be collected according to several criteria associated with different themes, different devices or measurement protocols defining the block structure. Examples of such data can be found in genomics, sensory analysis and chemical or food industry. In all these cases, the data obtained constitute a set of several homogeneous blocks of variables, referred to as multiple tables, multiblock or multiview data.

To deal with these multiview data in the context of unsupervised learning, it is usual to rely on dimension reduction approaches based on the two classical families of factorial methods and clustering. Depending on the objective, these two type of methods can be used separately [4], successively in a two step approach known as tandem approach or simultaneously leading to subspace clustering [6] . Clustering concerns the classic aspect of grouping individuals described here by several blocks of variables using consensus methods [4] or subspace clustering approaches [2]. Proposed methods also address the less classical aspect of variable clustering [5], which has been extended to clustering of blocks of variables using the RV coefficient. [3].

Methods for analyzing multiple tables are based on relationship measures and on comparison between the different data tables. The general idea is to take into account the natural or specific correlation that exists between variables in the same block. The approaches generally proposed can be grouped into two main families. The first one of multiblock component methods are based on a summary of each block through latent variables, known as canonical variables or components and are generally linear combinations of the variables

[1] . In the case of clustering, the summary is obtained as a qualitative variable resulting from a partitioning of each data table [4]. Methods differ according to the criteria used to obtain the block summary. We consider this approach to be vector-based, as opposed to the second family of methods, which we consider to be matrix-based. These later methods directly study the relationships and similarities between tables (without using components) through a distance measure or overall linkage between blocks.
In this talk, we will review, discuss and illustrate different unsupervised approaches from the most classical to the most recent.

References

[1] S. Bougeard, H. Abdi, G. Saporta, and N. Niang. Clusterwise analysis for multiblock component methods. *Advances in Data Analysis and Classification*, 12(2):285–313, 2018.

[2] X. Chen, X. Ye, X. Xu, and J. Z. Huang. A feature group weighting method for subspace clustering of high-dimensional data. *Pattern Recognition*, 45:434–446, 2012.

[3] F. Llobell, E. Vigneau, and El. M. Qannari. Clustering datasets by means of clustatis with identification of atypical datasets. application to sensometrics. *Food Quality and Preference*, 75:97–104, 2019.

[4] N. Niang and M. Ouattara. Weighted consensus clustering for multiblock data. In *SFC 2019*, Paris, France, September 2019.

[5] N. Niang, M. Ouattara, and G. Saporta. A comparison of some methods for clustering of variables of mixed types. In *XXX Meeting of the Portuguese Association for Classification and Data Analysis (JOCLAD 2023)*, pages 85–86, 2023.

[6] M. Yamamoto and H. Hwang. A general formulation of cluster analysis with dimension reduction and subspace separation. *Behaviormetrika*, 41(1):115–129, 2014.