Feature group clustering

Ndeye Niang

Conservatoire National des Arts et Metiers, France ndeye.niang_keita@cnam.fr

Mounir Bendali-Braham

2 Laboratoire d'Oceanographie et du Climat : Approches et Experimentations Numeriques (LOCEAN), France

Sylvie Thiria

2 Laboratoire d'Oceanographie et du Climat : Approches et Experimentations Numeriques, France

Keywords

Clustering, Multiblock data, RV index

The addressed problem is clustering features which are divided into several homogeneous and meaningful blocks. Preserving this homogeneity in data blocks would help to exhibit the underlying structure. The proposed method, called CLUSTABS, aims at finding clusters of feature blocks through a K-means like algorithm based on the RV correlation index. CLUSTABS is an extension of the clustering analysis of variable around latent components method proposed by Vigneau and Qanari. CLUSTAB method is similar to CLUSTATIS, another extension of CLV proposed by Llobell et al. in the same time than our proposal. RV index is a measure of the relationship between two sets of variables based on their associated scalar products matrices. It is non-negative and scaled between 0 and 1; the closer to 1, the more similar the matrices. The agglomerated feature blocks are the most similar according to the RV correlation coefficient. The centroid of a feature blocks cluster is the so-called compromise matrix, weighted average of the scalar products matrices associated to the feature blocks data. The compromise matrix has to be the most similar to these matrices according to the RV index. The solution is obtained through weights equal to the coordinates of the first standardized eigenvector of the matrix whose elements are RV coefficients between pairs of scalar products matrices. These weights represent the agreement between data tables and the compromise. The interest of CLUSTABS is illustrated on synthetic data as well as on a publicly available dataset.