DSSV 2023
Data Science, Statistics & Visualisation
ECDA 2023
European Conference on Data Analysis

Spectral clustering on association-based distances for mixed data

F. Palumbo[1] and C. Tortora[2]

[1]Università di Napoli Federico II, Dept. of Political Sciences, Naples, Italy
[2]San José State University, Dept. of Mathematics and Statistics, San José (CA), USA

keywords: spectral clustering, association-based distance, mixed data

Data clustering aims to find homogeneous groups in data. Dealing with continuous variables, statistical units are represented as points in the R^p metric space, where p indicates the number of considered variables, and the homogeneity in data is measured in terms of distances among the units within groups. Then, starting from a pairwise distance matrix, partitioning methods find groups according to the distances among units, and the choice The aim of the paper is to propose a spectral clustering implementation that can be suitably applied to mixed-type data.

A commonly used one SC procedure is the NJW algorithm, that takes as input the Euclidean distance matrix $\mathbf{S}$; then an affinity matrix $\mathbf{A}$ is computed as a weighted negative exponential of $\mathbf{S}$ after which the diagonal entries A_{ii} are set to zero. A second matrix is then computed, the diagonal matrix $\mathbf{D}$ with D_{ii} equal to the sum of the elements of row i of $\mathbf{A}$. The graph Laplacian matrix $\mathbf{L}$ can then be calculated from $\mathbf{A}$ and $\mathbf{D}$ as $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{1/2}$. The following step is to create the $n \times k$ matrix $\tilde{\mathbf{Y}}$ using the eigenvectors corresponding to the k largest eigenvalues obtained with the spectral decomposition of the Laplacian matrix $\mathbf{L}$. Each row of $\tilde{\mathbf{Y}}$ is re-normalized to unit length to give $\mathbf{Y}$. The matrix $\mathbf{Y}$ is characterized by well-separated clusters and, therefore, many different clustering algorithms can be used to partition the data. The most commonly used is k-means clustering.

The definition of the starting distance matrix is key to extend the SC to non-continuous attributes: some recent work focused on extending SC for mixed-type data. Some approaches automatically transform the data into categorical values and then applies a dimension reduction version of SC. SC for mixed data [1], instead, uses Euclidean distance for continuous variables, matching coefficient for categorical, and a tuning algorithm to determine the weights. Recently, [2] proposed a unified framework that includes the so-called association-based distances for categorical data. The mis-match between each category pair is weighted proportionally to the divergence between the the conditional distributions of the other variables, given the two categories in the pair: in case of little to no divergence in the distributions, then the mis-match in question is not emphasized. In this proposal we extend association based distances to include continuous variables and we use the new distance in SC.

[1] F. Mbuga and C. Tortora, "Spectral clustering of mixed-type data," *Stats*, vol. 5, no. 1, pp. 1–11, 2021.

[2] M. van de Velden, A. Iodice D'Enza, A. Markos, and C. Cavicchia, "A general framework for implementing distances for categorical variables," *submitted to Pattern Recognition*, pp. 1–21, 2023.