Robust Two-mode clustering

Maurizio Vichi
Department of Statistics, Sapienza University of Rome, Italy,
maurizio.vichi@uniroma1.it

Two-mode clustering is the activity of clustering modes (e.g., objects, variables) of an observed two-mode data matrix, simultaneously. This task is required because objects, frequently, are homogeneous only within subsets of variables, while variables may be strongly associated only on subsets of objects. For example, in microarray data analysis groups of genes are generally co-regulated within subsets of samples and groups of samples share a common gene expression pattern only for some subsets of genes. In market basket analysis customers have similar preference patterns only on subsets of products and, vice-versa, classes of products are more frequently consumed and preferred by subgroups of customers. In these situations a classical cluster analysis would cluster one mode (e.g., objects) on the basis of the complete set of the other mode (e.g., variables), thus producing weak results, while this is avoided with a more appropriate two-mode clustering. For *big data,* represented by matrices with a huge number of rows and columns, frequently the main analysis is a two-mode clustering, trying to mine and synthesize the relevant information by reducing the size of the data to a matrix of compact dimensions formed by prototype objects and variables. This is achieved by the simultaneous grouping rows and columns so that results are informative and easy to interpret, denoting compressed, but relevant representation of the big data, while trying to preserve most of the original information. The reduction is generally soft to obtain a light compression of the multivariate data in order to allow the successive application of other multivariate statistical methods that are computationally prohibitive for large data sets. The quality of big data is not always certifiable and frequently they are inflated by many outliers and influential data that have an high impact on the two-mode clustering and successive analyses. Therefore, robust multimode clustering techniques are needed for compressing large data set, while preserving the most relevant information.

A new robust asymmetrical two-mode clustering technique is proposed. A coordinate descent algorithm is developed. The applications on both, synthetic and real datasets, validate the performance and applicability of the new algorithm.

Key Words: Two-mode clustering, double k-means, disjoint principal component analysis, robustness.