Model-based Co-clustering

Mohamed Nadif
LIPADE, University Paris Descartes, 45 rue des Saint-Pères, 75006 Paris, France
mohamed.nadif@parisdescartes.fr

Let $X := \{x_{ij}; i \in I, j \in J\}$ be a data matrix where I is the set of objects and J the set of variables. Clustering methods of I are for the most part heuristic techniques derived from empirical methods, usually optimizing measurement criteria. In recent years, what used to be an algorithmic, heuristic and geometric focus has tended to give way to a more standard approach in statistics: the definition of probabilistic clustering models to formalize the intuitive notion of a natural class and the estimation of the model parameters from the sample. This approach allows precise analysis and can provide a statistical interpretation of many metrical criteria whose different variants are not always explicit. These criteria include the within-group sum of squares criterion, a criterion which gives rise to new variants corresponding to precise hypotheses. It is an approach that also provides a formal framework for tackling difficult problems such as determining the number of classes or validating the obtained clustering structure. This approach has become popular and makes it possible to propose different variants of Expectation-Maximization algorithm for clustering. Since these last years, this approach is extended to the case of co-clustering that consists in partitioning simultaneously the sets I and J. The latent block model was proposed and different co-clustering algorithms were derived. According to the type of data, Bernoulli, Gaussian and Poisson latent block models were proposed. They assume that there is a partition $\mathbf{z}$ into g row clusters on I and a partition $\mathbf{w}$ into m columns clusters on J, such that the random variables x_{ij} are conditionally independent given $\mathbf{z}$ and $\mathbf{w}$. In this work, these models are reviewed, all algorithms are discussed and evaluated in terms of estimation and clustering. Furthermore, different connections among objective functions that are commonly used with various co-clustering algorithms, and the complete data-likelihood of observed data while considering the classification maximum likelihood approach, are presented.

Keywords: co-clustering; contingency table; latent block model.