Clustering Based on Normal Mixture Model for Aggregated Symbolic Data

Nobuo Shimizu
Institute of Statistical Mathematics, Japan

Junji Nakano
Institute of Statistical Mathematics, Japan

Abstract

Symbolic Data Analysis (SDA) handles symbolic data (SD), in which values of a variable can be more complex than the traditional data such as real numbers and categorical values. Typical SD take intervals, histograms or bar charts as variable values, which describe information about the marginal distribution of each variable (Billard and Diday, 2006).

SDA provides techniques for handling such SD, including several extensions of clustering analysis. Hierarchical methods based on several definitions of dissimilarity between two SD have been mainly studied (Billard and Diday, 2006). As mixture model-based clustering methods for classical data are becoming popular recently (Everitt et al., 2011), we investigate clustering method based on normal mixture model in SD framework.

We consider the case where individuals of classical data are divided into some naturally defined groups and each group is considered to be SD, and call it aggregated symbolic data (ASD). ASD may be represented by some information about its joint distribution.

EM algorithm (Dempster et al., 1977) is often used in model-based clustering for classical data, and various extensions to some mixture models of EM algorithm are studied (McLachlan and Peel, 2000). We derive simplified EM algorithm in clustering based on normal mixture model for ASD by using mean and variance of each variable in ASD and also covariance information among variables in ASD. We apply our method to artificial and real data examples.