Simultaneous Fuzzy Clustering with Multiple Correspondence Analysis

Masaki Mitsuhiro*
Doshisha University, Kyoto, Japan dim0009@mail4.doshisha.ac.jp

Hiroshi Yadohisa
Doshisha University, Kyoto, Japan hyadohis@mail.doshisha.ac.jp

Multiple correspondence analysis is a simple method for analyzing multivariate categorical data. It is a categorical principal components analysis that assigns coordinates to respondents and the response categories of dummy-coded multiple categorical data to describe interdependencies among categories. To accommodate cluster structures between respondents and variable categories, the method has been extended by combining multiple correspondence analysis with two-way clustering, which attempts to classify both respondents and variable categories from the multivariate categorical data (Hwang and Dillon, 2010). Two-way clustering is preferred for interpreting analysis results because with one-way clustering, it is not easy to describe the characteristics of each respondent cluster based on its relationship with variable categories. This two-way clustering of multiple correspondence analysis is an alternative method involving the tandem approach, which is the application of clustering of objects and variables after applying a dimensional reduction of variables. However, hard classification methods such as k-means clustering appear to be too restricted because it is often difficult to identify clear boundaries between clusters in real-world problems. This paper proposes an extension of the multiple correspondence analysis with two-way k-means clustering in a unified framework, and our method simultaneously combines multiple correspondence analysis with two-way fuzzy c-means clustering, which is an overlapping clustering method. We represent the classification structure of respondents as fuzziness and the classification structure of categorical variables as hardness. By using our method, we obtain fuzzy clusters that exclusively relate a subgroup of respondents to a subset of categorical variables. The method can provide a low-dimensional map that simultaneously displays the object scores of respondents, variable categories, and cluster centroids in order to facilitate the interpretation of the relationships between variable categories and the cluster structure of respondents underlying the data. This approach provides cluster memberships of variable categories as well as respondents. Furthermore, we showed the usefulness of the proposed method for real data by comparing multiple correspondence analysis with our simultaneous approach.

Key Words: Categorical data, fuzzy c-means, ALS