Simultaneous Fuzzy Clustering with Multiple Correspondence Analysis

Masaki Mitsuhiro[1,3] and Hiroshi Yadohisa[2]
[1]Graduate School of Doshisha University, Kyoto, Japan
[2]Doshisha University, Kyoto, Japan
[3]Corresponding author: Masaki Mitsuhiro, e-mail:dim0009@mail4.doshisha.ac.jp

Abstract

Multiple correspondence analysis is a simple method for analyzing multivariate categorical data. The method has been extended to generate cluster structures between respondents and variable categories by combining multiple correspondence analysis with two-way clustering (Hwang and Dillon, 2010). However, because clear boundaries are not idenrifiable in many real-world clustering problems, hard classification methods such as k-means clustering appear overly restricted. In this study, we propose a method simultaneously combines multiple correspondence analysis with two-way fuzzy c-means clustering, which is an overlapping clustering method. We represent the classification structures of respondents and categorical variables as fuzziness and hardness, respectively. To facilitate the interpretation of the relationships between variable categories and the cluster structure of respondents, the method provides a low-dimensional map that simultaneously displays the object scores of respondents, variable categories, and cluster centroids. The utility of the proposed method in real data is assessed by comparing the results of our simultaneous approach with those of multiple correspondence analysis.

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

1. Introduction

Multiple correspondence analysis (MCA) is a simple method for analyzing multivariate categorical data. Based on categorical principal components analysis, MCA describes interdependencies among categories by assigning coordinates to respondents and to the response categories of dummy-coded multiple categorical data. MCA has been combined with k-means clustering in a unified framework (Hwang et al., 2006). Because it accommodates cluster structures between respondents and variable categories, this approach is useful for categorizing large respondent data sets. Moreover, it jointly displays variable categories and cluster centroids of respondents in a low-dimensional space. Combined MCA and cluster analysis involves tandem analysis, a two-step sequential approach in which objects and variables are clustered following dimensional reduction of variables.

The simultaneous approach has recently been extended by combining MCA with two-way clustering (Hwang and Dillon, 2010), which attempts to classify both respondents and variable categories from multivariate categorical data. Two-way clustering is preferable to one-way clustering, because the former naturally relates the characteristics of respondent clusters to variable categories. The difference between the two-way clustering approaches is visualized in the Figure 1. One-way clustering identifies clusters of respondents only, without regarding the variable categories. Two-way clusteing, on the other hand, attempts to classify both rows (respondents) and columns (variable categories) in a two-way data matrix. However, hard classification methods such as k-means clustering appear to be prohibitively restrictive in combined MCA/two-way clustering, because many real-world problems lack clear cluster boundaries. We consider that fuzzy respondent memberships

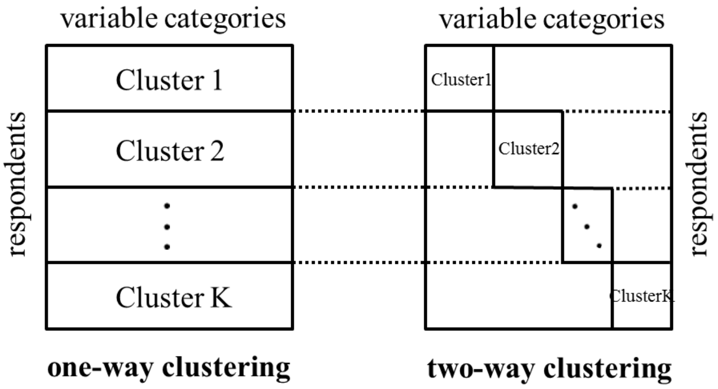

Figure 1: Difference between one-way clustering and two-way clustering.

can be realized by overlapping clustering. Such a classification scheme, in which respondents are assigned to multiple clusters by their degree of fuzzy membership, can potentially provide more insights into the structure of a dataset.

In this study, we propose a method that simultaneously combines MCA with two-way fuzzy c-means clustering. The classification structures of respondents and categorical variables are represented as fuzziness and hardness, respectively. Using this approach, we obtain fuzzy clusters that exclusively relate a subgroup of respondents to a subset of categorical variables. The method provides a low-dimensional map that simultaneously displays the object scores of respondents, the variable categories, and cluster centroids, thereby facilitating interpretation of the relationships between variable categories and the cluster structure of the respondents. This approach reveals clustering of variable categories and respondents. To demonstrate the utility of the proposed method, we compare the results of our simultaneous approach with those of MCA.

2. The proposed method

Let $\boldsymbol{Z}_j$ be an n by p_j matrix of the j-th dummy-coded categorical variable, where n is the number of respondents and p_j is the number of response categories in the variable j $(= 1, \cdots, J)$. This matrix assembled from multiple categorical data. Let $\boldsymbol{F}$ be an n by d matrix of a d-dimensional representation of J categorical variables. Let $\boldsymbol{W}_j$ be a p_j by d matrix of weights, also called category quantifications, assigned to response categories of the j-th variable. Denote the prescribed number of clusters by C, and denote u_{ci} as a membership value for respondent $(i = 1, \cdots, n)$ in the c-th cluster $(c = 1, \cdots, C)$. Let $\boldsymbol{U}_c^m$ be an n by n diagonal matrix of u_{ci}. Denote the prescribed fuzzy weight scalar by m. Let $\boldsymbol{V}_j$ be a p_j by C matrix of indicator variables providing the cluster memberships of response categories of the j-th categorical variable. Let $\boldsymbol{\gamma}_c$ be a d by 1 vector of the centroids of the c-th cluster in d-dimensions. Let $\boldsymbol{\Gamma}$ be a C by d matrix of the centroids of the clusters (where the rows of $\boldsymbol{\Gamma}$ are the $\boldsymbol{\gamma}_c$). Let $\mathbf{1}_n$ be an n by 1 vector, and denote the prescribed nonnegative scalar weights as λ_1, λ_2, and λ_3.

The objective of the proposed method is to classify variable categories and respondents by combining MCA with two-way fuzzy c-means clustering. To this end, our method displays variable categories and centroids of respondent clusters in a low-dimensional space. This problem is equivalent to minimizing the following expression:

$$\Phi = \lambda_1 \sum_{j=1}^{J} ||\boldsymbol{F} - \boldsymbol{Z}_j \boldsymbol{W}_j||^2 + \lambda_2 \sum_{c=1}^{C} ||\boldsymbol{F} - \boldsymbol{1}_n \boldsymbol{\gamma}_c^{'}||^2_{\boldsymbol{U}_c^m} + \lambda_3 \sum_{j=1}^{J} ||\boldsymbol{W}_j - \boldsymbol{V}_j \boldsymbol{\Gamma}||^2$$

where $||\boldsymbol{M}||^2_{\boldsymbol{H}} = \mathrm{tr}(\boldsymbol{M}^{'}\boldsymbol{H}\boldsymbol{M})$ with respect to $\boldsymbol{F}$, $\boldsymbol{W}_j$, $\boldsymbol{U}_c$, $\boldsymbol{V}_j$, $\boldsymbol{\Gamma}$ and $\boldsymbol{\gamma}_c$, subject to $\boldsymbol{F}^{'}\boldsymbol{F} = \boldsymbol{I}_d$, $\sum_{c=1}^{C} u_{ci} = 1$. $\boldsymbol{I}_d$ is a d by d identity matrix. When $\lambda_1 = 1$, the first term reduces to the standard MCA homogeneity criterion. The second and third terms are equivalent to the fuzzy c-means criteria and the standard k-means criteria for $\boldsymbol{F}$, $\boldsymbol{W}_j$, respectively. Our proposed method is solved by minimizing these three terms simultaneously.

The nonnegative scalar weights λ_1, λ_2, and λ_3 can be varied according to the objectives of the analysis, allowing researchers to investigate alternative solutions. We assign heavy weights to clustering terms whose values of loss function are smaller than that of MCA. We obtain a low-dimensional map that simultaneously displays the object scores of respondents, the variable categories, and cluster centroids.

The loss function is minimized by an alternating least squares (ALS) algorithm, which sequentially updates each unknown parameters (with other parameters fixed) until convergence is reached. The ALS algorithm is detailed below:

Algorithm

Step0: Randomly select initial values for $\boldsymbol{F}$, $\boldsymbol{W}_j$, $\boldsymbol{U}_c$, $\boldsymbol{V}_j$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\gamma}_c$.

Step1: Update $\boldsymbol{F}$ for fixed $\boldsymbol{W}_j$, $\boldsymbol{U}_c$, $\boldsymbol{V}_j$, and $\boldsymbol{\gamma}_c$. This is equivalent to maximizing $\mathrm{tr}(\boldsymbol{F}^{'}(\lambda_1 \sum_{j=1}^{J} \boldsymbol{Z}_j\boldsymbol{W}_j + \lambda_2 \sum_{c=1}^{C} \boldsymbol{U}_c^m \boldsymbol{1}_n \boldsymbol{\gamma}_c))$ as follows: Let $\mathrm{SVD}(\lambda_1 \sum_{j=1}^{J} \boldsymbol{Z}_j \boldsymbol{W}_j + \lambda_2 \sum_{c=1}^{C} \boldsymbol{U}_c^m \boldsymbol{1}_n \boldsymbol{\gamma}_c) = \boldsymbol{P}\boldsymbol{D}\boldsymbol{Q}^{'}$. Then, $\hat{\boldsymbol{F}} = \boldsymbol{P}\boldsymbol{Q}^{'}$.

Step2: Update $\boldsymbol{W}_j$ for fixed $\boldsymbol{F}$, $\boldsymbol{V}_j$, and $\boldsymbol{\Gamma}$. $\boldsymbol{I}_{p_j}$ is a p_j by p_j identity matrix.

$$\hat{\boldsymbol{W}}_j = \left(\lambda_1 \boldsymbol{Z}_j^{'} \boldsymbol{Z}_j + \lambda_3 \boldsymbol{I}_{p_j}\right)^{-1} \left(\lambda_1 \boldsymbol{Z}_j^{'} \boldsymbol{F} + \lambda_3 \boldsymbol{V}_j \boldsymbol{\Gamma}\right)$$

Step3: Update membership parameter u_{ci} for fixed $\boldsymbol{F}$, $\boldsymbol{W}_j$, and $\boldsymbol{\gamma}_c$. Insert the updated $\hat{u}_{ci}$ into $\boldsymbol{U}_c$. Define $d_{ci} = (\boldsymbol{f}_i - \boldsymbol{\gamma}_c)^{'}(\boldsymbol{f}_i - \boldsymbol{\gamma}_c)$. Then, u_{ci} is updated as follows:

$$\hat{u}_{ci} = \left(\sum_{k=1}^{C} \left(\frac{d_{ci}}{d_{ki}} \right)^{\frac{1}{m-1}} \right)^{-1}$$

Step4: Update $\boldsymbol{V}_j$ for fixed $\boldsymbol{F}$, $\boldsymbol{W}_j$, and $\boldsymbol{\Gamma}$. This is equivalent to separately minimizing the third terms of the objective function via the standard k-means algorithm.

Step5: Uptate $\boldsymbol{\Gamma}$ and γ_c for fixed $\boldsymbol{F}$, $\boldsymbol{W}_j$, $\boldsymbol{U}_c$, $\boldsymbol{V}_j$, $\boldsymbol{\Gamma}$ and γ_c.

$$\hat{\gamma}_c = \left(\lambda_2 \mathbf{1}_n^{'} \boldsymbol{U}_c^m \mathbf{1}_n + \lambda_3 \sum_{j=1}^{J} \boldsymbol{V}_{jc}^{'} \boldsymbol{V}_{jc} \right)^{-1} \left(\lambda_2 \mathbf{1}_n^{'} \boldsymbol{U}_c^m \boldsymbol{F} + \lambda_3 \sum_{j=1}^{J} \boldsymbol{V}_{jc} \boldsymbol{W}_j \right)$$

$$\hat{\boldsymbol{\Gamma}} = \sum_{c=1}^{C} \hat{\gamma}_c$$

Step6: Based on the result of Step5, return to Step1 or exit the algorithm (if the process has converged).

The above algorithm monotonically decreases the loss function. The membership matrix renders this algorithm rather sensitive to local optima. To ensure convergence to the global optimum, the implementation of many randomly started runs is recommended.

3. Numerical example

The utility of the proposed method is compared to that of MCA combined with one-way clustering and sequential two-way clustering approaches by numerical example. The input data are television program preference data comprising three multiple-choice variables taken from a survey of 100 Japanese undergraduates (Adachi, 2000). Respondents were asked to choose their one preferred category at each of three time points, $t = 1$ (first year of junior high school), $t = 2$ (first year of high school), and $t = 3$ (first year of university). Programs were divided into six categories: (1) animation, (2) cinema, (3) drama, (4) music, (5) sport, and (6) variety. Thus, the dataset comprises preferences selected from six television program categories at three time points.

First, the performance of our simultaneous approach is compared with that of MCA. Figure 2 displays the two-dimensional map of the object scores of respondents and the variable categories (denoted by labels such as “A1,” “C1,” and “D1;” for example, “A1” denotes “animation” at the point $t = 1$). Next, we apply our proposed method to the same data. Figure 3 displays the two-dimensional map of the object scores of respondents, the variable categories, and the centroids of three clusters for $d = 2$, $c = 3$, and $m = 2$. We assigned three categorical variables, each containing six response categories $(J = 3, p_j = 6)$. The nonnegative scalar weights were set to $\lambda_1 = 0.2$, $\lambda_2 = 0.4$, and $\lambda_3 = 0.4$.

In Figure 2, the object scores of respondents are scattered across the map. Because no particular cluster structure emerges from MCA, the relationships between respondents and variable categories are not readily interpreted. On the other hand, the proposed method produces clear cluster structures between respondents and variable categories (Figure 3). The graphical representation in Figure 3 is consistent with the clustering information provided in Table 1.

The first (majority) cluster, whose centroid is represented by “C1,” contains data from 66 respondents and 9 categories. The respondents in cluster 1 show a preference for drama and music programs. The second cluster, whose centroid is represented by “C2,” is constructed from 14 respondents and 9 categories. The respondents in cluster 2 are likely to watch animation and sport programs. Finaly, the third cluster, whose centroid is represented by “C3,” is generated from 20 respondents and 9 categories. The respondents in this cluster prefer variety programs. In addition, the extent of fuzzy membership of respondents is visible in Figure 3.

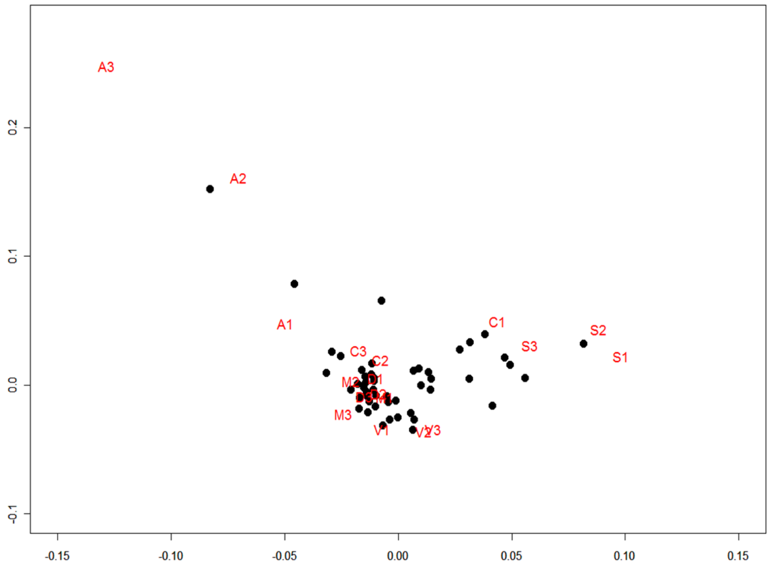

Figure 2: Two-dimensional solution of MCA. Letters and numbers on the plot refer to program categories and time points, respectively (respondents selected one of the six program categories at each of three different time points).

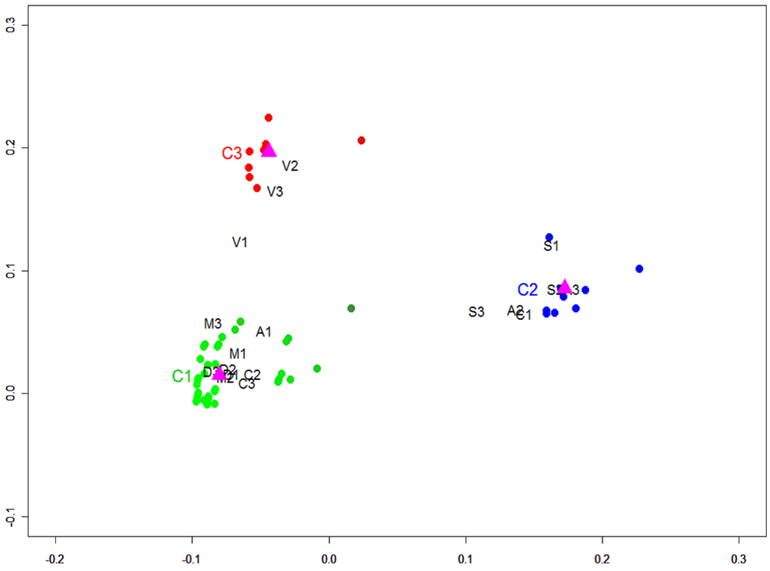

Figure 3: Two-dimensional solution of our proposed method. The object scores (values of fuzzy membership) are indicated by RGB colors. Triangles indicate the centroids of three clusters (labeled “C1,” “C2,” and “C3”).

Table 1: Cluster memberships of variable categories

Cluster1 $(n_1 = 66)$	Cluster2 $(n_2 = 14)$	Cluster3 $(n_3 = 20)$
A1	C1	V1
D1	S1	V2
M1	A2	V3
C2	S2	
D2	A3	
M2	S3	
C3		
D3		
M3		

The proposed method generates a graphical solution that is more easily interpretable than that of MCA. In particular, the object scores of respondents and the weights of variable categories are clearly clustered.

4. Conclusions

We propose a method that simultaneously combines MCA with two-way fuzzy c-means clustering in multivariate categorical data. This method provides a low-dimensional map that simultaneously displays variable categories and cluster centroids, thereby facilitating the interpretation of the relationships between variable categories and the cluster structure of respondents.

References

Adachi, K. (2000) “Optimal scaling of a longitudinal choice variable with time-varying representation of individuals,”*British Journal of Mathematical and Statistical Psychology, 53*, 233–253.

Greenacre, M. (1993) *Correspondence Analysis in Practice,* London: Academic Press.

Greenacre, M. and Blasius, J. (2007) *Multiple Correspondence Analysis and Related Methods,* London: Academic Press.

Hwang, H., Dillon, W. R., and Takane, Y. (2006) “An extension of multiple correspondence analysis for identifying heterogeneous subgroups of respondents,”*Psychometrika, 71*, 161–171.

Hwang, H., Dillon, W. R., and Takane, Y. (2010) “Fuzzy cluster multiple correspondence analysis,”*Behaviormetrika, 37(2),* 111–133.

Hwang, H. and Dillon, W. R. (2010) “Simultaneous two-way clustering of multiple correspondence analysis,”*Multivariate Behavioral Research, 45*, 186–208.