2003 International Symposium on Advanced Intelligent Systems
September 25-28, 2003, Jeju, Korea

Simultaneous Approach to Fuzzy Clustering and Quantification of Categorical Data with Missing Values

Katsuhiro Honda, Yoshihito Nakamura, Hidetomo Ichihashi
Graduate School of Engineering, Osaka Prefecture University
1-1 Gakuen-cho, Sakai, Osaka, 599-8531, Japan
honda@ie.osakafu-u.ac.jp

Abstract— **This paper proposes a simultaneous application of homogeneity analysis and fuzzy clustering with incomplete data. Taking the similarity between the loss of homogeneity in homogeneity analysis and the least squares criterion in principal component analysis into account, the new objective function is defined in a similar formulation to the linear fuzzy clustering with missing values. Numerical experiment shows the characteristic properties of the proposed method.**

Keywords— **Fuzzy clustering, homogeneity analysis, missing values.**

I. Introduction

Simultaneous approaches to multivariate data analysis and fuzzy clustering have been applied to knowledge discovery from large scale databases because the local model derived in each cluster effectively reveals the local features of non-linearly distributed high-dimensional data sets. Fuzzy c-Varieties (FCV) clustering proposed by Bezdek *et al.* [1] [2] is regarded as a simultaneous approach to principal component analysis (PCA) and fuzzy clustering since FCV partitions a data set into several linear clusters using linear varieties as prototypes of clusters and the basis vectors of the prototypical linear varieties are often identified with local principal component vectors.

In spite of the usefulness, however, they often suffer from missing observations in real world applications. Honda *et al.* [3], [4] proposed a modified linear fuzzy clustering algorithm in which the objective function was regarded as the least squares criterion for local PCA. While the objective function of the FCV algorithm is composed of the distances between data points and prototypical linear varieties, the same solution can be derived from the least squares criterion that achieves "component-wise" lower rank approximation of the data matrix. However, the algorithm is available only when the data matrix consists of numerical variables.

This paper proposes a new approach to the quantification of incomplete categorical data, which constructs multiple uni-dimensional scales by partitioning a set of samples into clusters. Homogeneity analysis [5], [6] is a quantification technique for representing the structure of non-numerical multivariate data and tries to minimize departures from perfect homogeneity that are measured by the Gifi loss function. The minimization of the Gifi loss function is based on the approximation of matrix, so the algorithm is similar to that of PCA with least squares criterion. In the proposed method, samples are partitioned by introducing membership values into the Gifi loss function and missing values are ignored by multiplying "0" weights to corresponding deviations in component-wise approximation.

Finally, the characteristic features of our method are shown in numerical examples.

II. Linear Fuzzy Clustering with Missing Values

Let $X = (x_{ij})$ denote an $(n \times m)$ data matrix consisting of m dimensional observation of n samples. The matrix is often denoted as $X = (\tilde{\boldsymbol{x}}_1, \cdots, \tilde{\boldsymbol{x}}_n)^\top$ using m dimensional column vectors $\tilde{\boldsymbol{x}}_i$ composed of the i-th row elements of X. In the following, column vectors are shown in bold.

FCV is a clustering method that partitions a data set into C linear fuzzy clusters. The objective function of FCV consists of distances from data points to p dimensional prototypical linear varieties spanned by linearly independent vectors $\boldsymbol{a}_{ck}$ as follows [1], [2]:

$$L_{fcv} = \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}^{\theta}\Big\{||\tilde{\boldsymbol{x}}_i - \boldsymbol{b}_c||^2 - \sum_{k=1}^{p} \boldsymbol{a}_{ck}^\top R_{ci}\boldsymbol{a}_{ck}\Big\}, \quad (1)$$

$$R_{ci} = (\tilde{\boldsymbol{x}}_i - \boldsymbol{b}_c)(\tilde{\boldsymbol{x}}_i - \boldsymbol{b}_c)^\top, \quad (2)$$

where u_{ci} denotes the membership degree of the data point $\tilde{\boldsymbol{x}}_i$ to the c-th cluster and $\top$ represents the transpose of the vector. $\boldsymbol{b}_c$ is the center of the c-th cluster. The weighting exponent θ is added for fuzzification. The larger θ is, the fuzzier the membership assignments are. Because the optimal $\boldsymbol{a}_{ck}$ are eigenvectors corresponding to the largest eigenvalues of the generalized fuzzy scatter matrix, the vectors are regarded as the fuzzy principal component vectors extracted in each cluster considering the memberships [7].

Honda *et al.* [3], [4] proposed to modify the objective function using least squares criterion and applied them to the analysis of incomplete data sets. Introducing memberships u_{ci}, the least squares criterion for fuzzy local PCA is defined as

$$L_{lsc} = \sum_{c=1}^{C} \mathrm{tr}\Big\{(X - Y_c)^\top U_c^{\theta}(X - Y_c)\Big\}, \quad (3)$$

where $U_c = \mathrm{diag}(u_{c1}, \cdots, u_{cn})$ and tr represents the trace of the matrix (the sum of the diagonal entries). $Y_c = (y_{cij})$ denotes the lower rank approximation of the data matrix X in the c-th cluster,

$$Y_c = F_c A_c^\top + \boldsymbol{1}_n \boldsymbol{b}_c^\top, \quad (4)$$

where $F_c = (f_{cik})$ is the $(n \times p)$ score matrix and $A_c = (\boldsymbol{a}_{c1}, \cdots, \boldsymbol{a}_{cp})$ is the $(m \times p)$ principal component matrix of the c-th fuzzy cluster. $\mathbf{1}_n$ is n dimensional vector whose elements are all 1. The objective function achieves the lower rank approximation of the data matrix and derives the same solution as the FCV algorithm because the principal component vectors $\boldsymbol{a}_{c1}, \cdots, \boldsymbol{a}_{cp}$ and the cluster center $\boldsymbol{b}_c$ span the same prototypical linear varieties.

By the way, Eq.(3) can also be expressed as

$$L_{lsc} = \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci}^{\theta} \sum_{j=1}^{m} (x_{ij} - \sum_{k=1}^{p} f_{cik} a_{cjk} - b_{cj})^2. \quad (5)$$

This formulation means that the clustering criterion is composed of the component-wise approximation of the data matrix. So, we can handle missing values in the data matrix by considering the approximation of the observed elements only.

In [3], [4], missing values in the data matrix are ignored by multiplying "0" weights over the corresponding reconstruction errors. Considering binary weights d_{ij},

$$d_{ij} = \begin{cases} 1 & ; x_{ij} \text{ is observed.} \\ 0 & ; x_{ij} \text{ is missing.} \end{cases} \quad (6)$$

the objective function of FCV with missing values is defined as

$$\begin{aligned} L_{fcvm} &= \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci} \sum_{j=1}^{m} d_{ij} (x_{ij} - \sum_{k=1}^{p} f_{cik} a_{cjk} - b_{cj})^2 \\ &+ \lambda \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci} \log u_{ci}, \quad (7) \end{aligned}$$

where the entropy term is added for fuzzification instead of the weighting exponent in the standard FCV algorithm. The fuzzification technique is called "Regularization by entropy" [8]. The larger λ is, the fuzzier the membership assignments are.

To obtain a unique solution, the objective function is minimized under the constraints that

$$F_c^{\top} U_c F_c = I \quad ; \quad c = 1, \cdots, C, \quad (8)$$

$$F_c^{\top} U_c \mathbf{1}_n = \mathbf{0} \quad ; \quad c = 1, \cdots, C, \quad (9)$$

$$\sum_{c=1}^{C} u_{ci} = 1 \quad ; \quad i = 1, \cdots, n, \quad (10)$$

and $A_c^{\top} A_c$ is orthogonal. The optimal solution is derived based on the alternating least squares.

However, this approach is useful only when the data matrix is composed of numerical variables and we cannot apply the algorithm to the analysis of categorical data. In the next section, we enhance the idea to the simultaneous approach to fuzzy clustering and quantification of categorical data.

III. Simultaneous Approach to Fuzzy Clustering and Homogeneity Analysis

A. Homogeneity Analysis

Suppose that we have collected data on n objects on m categorical variables with $K_j, j = 1, \cdots, m$ categories. The categories of each variable are often nominal, i.e., only the classes formed by the objects play a role. These non-numerical variables are represented by indicator matrices. Let G_j denote the $n \times K_j$ indicator matrix corresponding to variable j and its entries be the binary variables as follows:

$$g_{ijk} = \begin{cases} 1 & ; \text{ if object } i \text{ belongs to category } k. \\ 0 & ; \text{ otherwise.} \end{cases}$$

$$G_j = \begin{pmatrix} g_{1j1} & \cdots & g_{1jk} & \cdots & g_{1jK_j} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ g_{ij1} & \cdots & g_{ijk} & \cdots & g_{ijK_j} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ g_{nj1} & \cdots & g_{njk} & \cdots & g_{njK_j} \end{pmatrix}, \quad j = 1, \cdots, m.$$

These matrices can be collected in an $(n \times K)$ partitioned matrix $G = [G_1, G_2, \cdots, G_m]$, where $K = \sum_{j=1}^{m} K_j$ is the total number of categories.

The goal of the quantification of categorical data is to represent these objects in a p dimensional space $(p < m)$. Homogeneity analysis [5], [6] is the basic technique of non-linear multivariate analysis and aims at the representation of the structure of non-numerical multivariate data by assigning scores to the objects and the categories of variables. Let W_j denote the $(K_j \times p)$ matrix containing the multiple category quantification of variable j and Z be an $(n \times p)$ matrix containing the resulting p object scores as follows:

$$Z = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{pmatrix}, \quad (11)$$

$$W_j = \begin{pmatrix} w_{j11} & w_{j12} & \cdots & w_{j1p} \\ w_{j21} & w_{j22} & \cdots & w_{j2p} \\ \vdots & \vdots & \ddots & \vdots \\ w_{jK_j1} & w_{jK_j2} & \cdots & w_{jK_jp} \end{pmatrix}. \quad (12)$$

Homogeneity analysis is based on the principle that a scale consisting of nominal variables is homogenizable if all variables can be quantified in such a way that the resulting scale is homogeneous, i.e., all the variables in the scale are linearly related. The departures from perfect homogeneity are measured by the Gifi loss function.

$$\sigma = \frac{1}{m} \sum_{j=1}^{m} \mathrm{tr}\{(Z - G_j W_j)^{\top} (Z - G_j W_j)\}. \quad (13)$$

In order to avoid the trivial solution, the loss function is minimized under the conditions,

$$\begin{aligned} \mathbf{1}_n^\top Z &= \mathbf{0}^\top, &(14)\\ Z^\top Z &= nI. &(15)\end{aligned}$$

Here, the Gifi loss function can be represented as

$$\sigma = \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n}\sum_{l=1}^{p}\Big\{z_{il}-\sum_{k=1}^{K_j}g_{jik}w_{jkl}\Big\}^2, \qquad (16)$$

and is similar to the least squares criterion for PCA based on the component-wise approximation. Then, in the next subsection, we propose a new simultaneous approach to fuzzy clustering and quantification of incomplete categorical data in the same manner as the previous section.

B. Simultaneous Approach to Fuzzy Clustering and Homogeneity Analysis with Missing Values

In this subsection, we propose a new approach that performs fuzzy clustering and homogeneity analysis simultaneously. The fuzzy partitioning of n objects is performed by introducing memberships $u_{ci}, c = 1,\cdots,C, i = 1,\cdots,n$ and u_{ci} denotes the membership degree of the i-th object to the c-th cluster. The objective function with regularization by entropy is defined as follows:

$$\begin{aligned} \sigma^* &= \frac{1}{m}\sum_{c=1}^{C}\sum_{j=1}^{m}\mathrm{tr}\{(Z_c-G_jW_{cj})^\top U_cM_j\\ &\quad \times(Z_c-G_jW_{cj})\}+\lambda\sum_{c=1}^{C}\sum_{i=1}^{n}u_{ci}\log u_{ci}, \qquad (17)\end{aligned}$$

where M_j is the $(n \times n)$ diagonal matrix,

$$M_j = \mathrm{diag}\Big(\sum_{k=1}^{K_j}g_{1jk},\cdots,\sum_{k=1}^{K_j}g_{njk}\Big), \qquad (18)$$

If the i-th object answered to the j-th question, $\sum_{k=1}^{K_j}g_{ijk}$ is 1. Otherwise, 0. Then the i-th diagonal element of M_j is the binary variable that indicates whether the i-th object (individual) answered to the j-th question. Therefore, the minimization of Eq.(17) implies that local quantification is performed by ignoring the missing values of indicator matrix in the same way as linear fuzzy clustering with missing values.

To derive unique solution, Eq.(17) is minimized under the following conditions.

$$\begin{aligned} \boldsymbol{u}_c^\top M_*Z_c &= \mathbf{0}^\top, &(19)\\ Z_c^\top U_cM_*Z_c &= m\Big\{\sum_{i=1}^{n}u_{ci}\Big\}I, &(20)\end{aligned}$$

where $M_* = \sum_{j=1}^{m}M_j$.

The optimal solution is derived based on iterative least squares technique. From the necessary condition for the optimality $\partial\sigma^*/\partial W_{cj} = O$, the updating rule for W_{cj} is derived as

$$\hat{W}_{cj} = D_{cj}^{-1}G_j^\top U_cZ_c, \qquad (21)$$

where $D_{cj} = G_j^\top U_cG_j$. Consequently, from $\partial\sigma^*/\partial Z_c = O$ and $\partial\sigma^*/\partial u_{ci} = 0$, we have

$$\hat{Z}_c = M_*^{-1}\sum_{j=1}^{m}G_jW_{cj}, \qquad (22)$$

and

$$\begin{aligned} \hat{u}_{ci} &= \exp(B_{ci}-1), &(23)\\ B_{ci} &= \frac{-1}{\lambda m}\sum_{j=1}^{m}\Big\{\sum_{l=1}^{K_j}g_{ijl}\Big\}\sum_{h=1}^{p}(z_{cih}-\sum_{k=1}^{K_j}g_{ijk}w_{cjkh})^2, &(24)\end{aligned}$$

respectively. When we consider the "probabilistic constraint" [9] for memberships ($\sum_{c=1}^{C}u_{ci} = 1$), the new membership is calculated as

$$\hat{u}_{ci} = \frac{\exp(B_{ci})}{\sum_{l=1}^{C}\exp(B_{li})}. \qquad (25)$$

The proposed algorithm can be written as follows.

Step1 Initialize Z_c and U_c randomly and normalize them so that the probabilistic constraint and Eqs.(19), (20) hold.
Step2 Calculate W_{cj} using Eq.(21).
Step3 Calculate Z_c using Eq.(22).
Step4 Normalize Z_c so that Eqs.(19), (20) hold.
Step5 Calculate u_{ci} using Eq.(25).
Step6 If

$$\max_{c,i} \mid u_{ci}^{NEW} - u_{ci}^{OLD} \mid < \epsilon,$$

then stop. Otherwise, return to Step 2.

IV. Numerical Experiments

In this section, we present the result of analysis for finding the relationship between interests and tastes. Table I shows a result of questionnaire about interests and tastes of 7 youngsters. In the table, the answer of responder 6 for "interests" is missing and the corresponding elements are all "0". We applied the proposed algorithm to this categorical data set to derive 2 dimensional plots ignoring the missing value. Table II shows the membership of responders to each cluster and Fig. 1 shows the 2 dimensional plots derived in each cluster. In the plots, ○ and • indicate the categories and the responders respectively. The responders are partitioned by the tastes for pasta and the responders who like spaghetti are included only in 2nd cluster. Then Fig. 1-(b) emphasizes the features of persons who like pasta while Fig. 1-(a) shows the general features. In this way, the proposed method is useful for finding local features of incomplete categorical data sets.

TABLE I

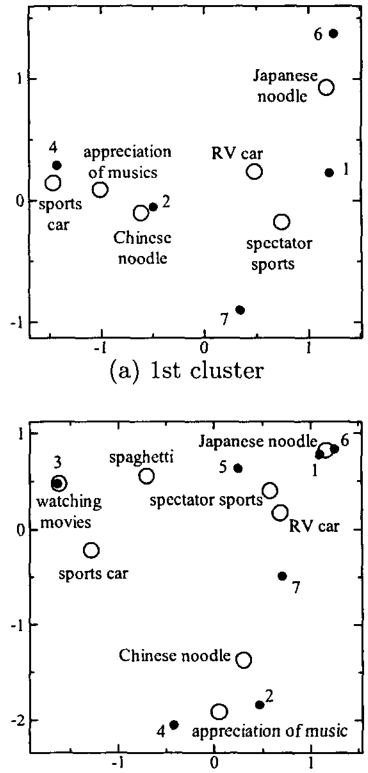
Cross-classification table of interests and tastes of youngsters

responder	interests			tastes for pasta			liking for car	
	appreciation of music	watching movies	spectator sports	Chinese noodle	spaghetti	Japanese noodle	recreation vehicle	sports car
1	0	0	1	0	0	1	1	0
2	1	0	0	1	0	0	1	0
3	0	1	0	0	1	0	0	1
4	1	0	0	1	0	0	0	1
5	0	0	1	0	1	0	1	0
6	0	0	0	0	0	1	1	0
7	0	0	1	1	0	0	1	0

TABLE II

Membership values of responders

responder	$c = 1$	$c = 2$
1	0.488	0.512
2	0.587	0.413
3	0.000	1.000
4	0.617	0.383
5	0.000	1.000
6	0.464	0.536
7	0.462	0.538



(a) 1st cluster

(b) 2nd cluster

Fig. 1. Combined category quantifications and object scores plots

V. Conclusion

In this paper, we proposed a new local quantification method that can handle missing observations. The objective function was defined by introducing memberships to the Gifi loss function of homogeneity analysis. Because the minimization of the loss function is based on the component-wise approximation of matrix, missing values can be ignored by multiplying "0" weights to the corresponding errors in the same way as linear fuzzy clustering with incomplete data. We have formulated the problem so as to classify individuals into several clusters but categories can be also classified. Tsuchiya [10] proposed a method for the construction of multi uni-dimensional scales by classifying a set of qualitative variables into groups. The comparative study is left for future works.

Acknowledgments

This work was supported by the Ministry of Education, Culture, Sports, Science and Technology, Japan, under Grant-in-Aid for Scientific Research (#15700196).

References

[1] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, 1981.

[2] J. C. Bezdek, C. Coray, R. Gunderson and J. Watson, Detection and Characterization of Cluster Substructure 2. Fuzzy c-Varieties and Convex Combinations Thereof, SIAM J. Appl. Math., vol.40, no.2, pp 358-372, 1981.

[3] K. Honda, N. Sugiura, H. Ichihashi and S. Araki, Collaborative Filtering Using Principal Component Analysis and Fuzzy Clustering, Web Intelligence: Research and Development, Lecture Notes in Artificial Intelligence 2198, Sprnger, pp. 394-402, 2001.

[4] K. Honda and H. Ichihashi, Linear Fuzzy Clustering Techniques with Missing Values and Their Application to Local Principal Component Analysis, IEEE Trans. on Fuzzy Systems, to appear.

[5] A. Gifi, Nonlinear Multivariate Analysis, Wiley, 1990.

[6] J. Bond and G. Michailidis, Homogeneity Analysis in Lisp-Stat, Journal of Statistical Software, vol.1, issue 2, 1996.

[7] Y. Yabuuchi and J. Watada, Fuzzy Principal Component Analysis and its Application, Biomedical Fuzzy and Human Sciences, vol.3, pp.83-92, 1997.

[8] S. Miyamoto and M. Mukaidono, Fuzzy c-Means as a Regularization and Maximum Entropy Approach, Proc. of the 7th International Fuzzy Systems Association World Congress, vol. 2, pp. 86-92, 1997.

[9] F. Höppner, F. Klawonn, R. Kruse and T. Runkler, Fuzzy Cluster Analysis, Jhon Wiley & Sons, 1999.

[10] T. Tsuchiya, A Quantification Method for Classification of Variables, The Japanese Journal of Behaviormetrics, vol.22, no.2 pp.95-109, 1995 (in Japanese).