Fuzzy Clustering for Categorical Multivariate Data *

Chi-Hyon Oh, Katsuhiro Honda and Hidetomo Ichihashi
Graduate School of Engineering, Osaka Prefecture University
Sakai, Osaka, Japan
E-mail:honda@ie.osakafu-u.ac.jp

Abstract

This paper proposes a new fuzzy clustering algorithm for categorical multivariate data. The conventional fuzzy clustering algorithms form fuzzy clusters so as to minimize the total distance from cluster centers to data points. However, they cannot be applied to the case where only cooccurrence relations among individuals and categories are given and the criterion to obtain clusters is not available. The proposed method enables us to handle that kind of data set by maximizing the degree of aggregation among clusters. The clustering results by the proposed method show similarity to those of Correspondence Analysis or Hayashi's Quantification Method Type III. Numerical examples show the usefulness of our method.

1 Introduction

We have been facing an explosive increase in the amount of information or data being stored in the database electrically. In response to the increase of the data storage, importance of extracting useful information, which is even implicit, potential or previously unknown, from the database is getting up-and-coming. The extracted information can be put to use in the areas such as decision support, prediction, forecasting and estimation. Data Mining [1][2] which is also known as Knowledge Discovery from Database (KDD) is this kind of paradigm. It encompasses a number of different technical approaches. Most of them devote themselves to unearth association rules which represent association relationships among different attributes. They find out such rules according to similarity or correlation among data.

There exists a classical technique called Correspondence Analysis [3] or Hayashi's Quantification method type III [4] which analyzes data according to correlation. It is one of the multivariate analysis techniques and can be seen as one of the Data Mining approaches. It deals with categorical multivariate data. The categorical multivariate data set is provided in the form of cross-classification table, contingency table or cooccurrence matrix. Each individual is described by a set of qualitative variables with several categories. The categorical variables are defined by several quantifications of qualitative data: binary indicator, frequency or scaled variable. The Correspondence Analysis quantifies the multi categorical data so as to maximize correlation among data. This technique provides us with useful knowledge from the data set and makes it possible to visualize various criteria of principal component analysis. We can employ the Correspondence Analysis as a way of dimension reduction.

Cluster analysis is a technique which discovers the substructure of a data set by dividing it into several clusters. It is also known as one of the Data Mining approaches. There have been many researches for cluster analysis. Fuzzy clustering is an extension of the cluster analysis, which represents the affiliation of data points to clusters by memberships. Introducing fuzziness to clustering gives us the flexible representations of substructures of the data set. Fuzzy c-Means [5], Fuzzy c-Lines [6], Fuzzy c-Varieties [7] and Fuzzy c-Regression Models [8] are varieties of the fuzzy clustering algorithms. They have different shapes of cluster centers, prototypes of clusters. Most of them conduct clustering in accordance with similarity or dissimilarity derived from distances from cluster centers to data points. They employ Euclidian, Mahalanobis or Manhattan distance as the metric. Few fuzzy clustering approaches, however, exist, which can realize fuzzy clustering where the data vectors are not available and only the similarity or correlation among items is given.

This paper proposes a new fuzzy clustering algorithm when a categorical multivariate data set is given. By the virtue of its capability of handling categorical multivariate data, the proposed method can conduct not only the cluster analysis but also data analysis similar to the Correspondence Analysis. Yamakawa et al. has

*Proc. of Joint 9th IFSA World Congress and 20th NAFIPS International Conference, 2154-2159 (2001)

proposed a hybridization of the fuzzy clustering and the Correspondence Analysis [9]. Their proposed method implement two different data analysis techniques simultaneously and can obtain local relationships among data and scatter diagrams. Though our proposed method also focuses on the same kind of data set, it conducts only fuzzy clustering. Therefore, our method essentially differs from Yamakawa's method. And another thing, Inoue et al. has already proposed a fuzzy clustering algorithm that can also handle with categorical multivariate data [10]. However, as their algorithm form the clusters one after another, the volume of the cluster gradually decreases in turn. Moreover, in the process of assigning data points to clusters, calculation of eigen vectors is needed, which is computationally demanding.

Our proposed algorithm is effectuated by maximizing a simple objective function. The objective function represents the degree of aggregation of each cluster. The solution algorithm for the objective function is based on iterative procedure through necessary conditions for local minima. We can obtain clusters for the overall data set at a time by using the propose method. In the objective function, we introduce entropy maximization as a regularization proposed by Miyamoto et al. [11] to obtain fuzzy clusters. After the fuzzy clustering, we are supposed to obtain memberships for individuals and categories each and all.

Looking into obtained clusters mixed up with individuals and categories, the similar result of data analysis to the Correspondence Analysis can be derived. Since our proposed method can easily provide fuzzy clusters by solving simple algebraic equations that are far easier than the eigen value problems and doesn't require the calculation of cluster centers, it provides us with useful way to analyze categorical multivariate data. Numerical examples show the usefulness of our method.

2 Fuzzy Clustering for Categorical Multivariate Data (FCCM)

2.1 Categorical Multivariate Data

Let us consider a categorical multivariate data set. M individuals described by a set of qualitative variables d_{ij} with N categories. In many cases, each category consists of some few sub-categories. They, however, are not taken into consideration here. The qualitative variables can be responses to some questionnaires or cooccurrence relations among individuals and categories. The categorical variables are defined by several quantifications, *e.g.*, binary indicator, frequency or scaled variable. The categorical multivariate data set is often given in the form of a table. We show an example of the table in Table 1.

Table 1: an example of categorical multivariate data set

	1	2	$\cdots$	j	$\cdots$	N
1	d_{11}	d_{12}	$\cdots$	d_{1j}	$\cdots$	d_{1N}
2	d_{21}	d_{22}	$\cdots$	d_{2j}	$\cdots$	d_{2N}
$\vdots$	$\vdots$	$\ddots$	$\cdots$	$\ddots$	$\cdots$	$\vdots$
i	d_{i1}	d_{i2}	$\cdots$	d_{ij}	$\cdots$	d_{iN}
$\vdots$	$\vdots$	$\ddots$	$\cdots$	$\ddots$	$\cdots$	$\vdots$
M	d_{M1}	d_{M2}	$\cdots$	d_{Mj}	$\cdots$	d_{MN}

In the table, the rows are the individuals and the columns are the categories. This kind of table is called cross-classification table, contingency table or cooccurrence matrix in a general way.

The conventional technique handling the data set in Table 1, *i.e.*, the Correspondence Analysis, intends to discover relations between individuals and categories. It quantifies the individuals and the categories by solving an eigen value problem. After the quantification, we can plot the individuals and the categories on one or two dimensional space. Then, they are divided in some groups in accordance with the coordinates and characteristics of the data set are detected. The analysis of the categorical multivariate data set is commonly conducted in this manner in the Correspondence Analysis. The division of the data set after the quantification is done by the analysts observing the data plots. Therefore, the Correspondence Analysis is not a clustering technique but a quantification method. Well-known fuzzy clustering algorithms [5]-[8], however, can not handle the data set like in Table 1 because of the clustering criteria they employ. They form clusters according to distances from cluster centers to data points. It is impossible to calculate those distances with respect to the data set in Table

1. In this paper, we propose a new fuzzy clustering algorithm which is able to handle the data set in Table 1. We call it FCCM, Fuzzy Clustering for Categorical Multivariate Data.

2.2 Degree of Aggregation

Firstly, we define two different memberships for the proposed method. One is for the individuals, the other is for the categories. The definitions of two memberships are shown as follows:

$$\sum_{c=1}^{C} u_{ci} = 1, \quad u_{ci} \in [0,1], \quad i = 1, \cdots, M, \tag{1}$$

$$\sum_{j=1}^{N} w_{cj} = 1, \quad w_{cj} \in [0,1], \quad c = 1, \cdots, C, \tag{2}$$

where u_{ci} is the membership of the i-th individual for the c-th cluster and w_{cj} is that of the j-th category for the c-th cluster. C denotes the number of clusters. Though it seems that u_{ci} and w_{cj} have the same constraints since the memberships sum to one, they are different actually. For u_{ci}, the total amount of memberships of the i-th individual to the clusters has to be one. On the other hand, (2) indicates the total membership of the c-th cluster to the categories should be one.

Secondly, we give a definition of the clustering criterion of the FCCM to obtain fuzzy clusters. It should be provided so as to group the individuals and the categories which have high correlations each other. In this sense, we regard the following degree of aggregation as the clustering criterion of the FCCM.

$$\sum_{i=1}^{M}\sum_{j=1}^{N} u_{ci} w_{cj} d_{ij} \quad c = 1, \cdots, C. \tag{3}$$

The degree of aggregation for each cluster is the total amount of products of qualitative variables d_{ij} and memberships for individuals and categories, u_{ci} and w_{cj}. We maximize the degree of aggregation in (3) to form fuzzy clusters by assigning memberships to individuals and categories.

Furthermore, if we define the total amount of memberships w_{cj} of the j-th category to the clusters as one, in such a way as in (4), we are unable to obtain proper clusters.

$$\sum_{j=1}^{C} w_{cj} = 1, \quad w_{cj} \in [0,1], \quad j = 1, \cdots, N. \tag{4}$$

The degree of aggregation will be maximized by allocating individuals and categories to only one cluster under the constraints provided in (4). That is why we employ (2).

2.3 Objective Function

The FCCM can be driven by optimization of an objective function to maximize the degree of aggregation. We use Lagrange's method of indeterminate multiplier to derive the objective function for the FCCM. The objective function can be written as follows:

$$\begin{aligned}
\max \ L = & \sum_{c=1}^{C}\sum_{i=1}^{M}\sum_{j=1}^{N} u_{ci} w_{cj} d_{ij} \\
& - T_u \sum_{c=1}^{C}\sum_{i=1}^{M} u_{ci} \log u_{ci} - T_w \sum_{c=1}^{C}\sum_{j=1}^{N} w_{cj} \log w_{cj} \\
& + \sum_{i=1}^{M} \lambda_i \Big(\sum_{c=1}^{C} u_{ci} - 1 \Big) \\
& + \sum_{c=1}^{C} \gamma_c \Big(\sum_{j=1}^{N} w_{cj} - 1 \Big),
\end{aligned} \tag{5}$$

where λ_i and γ_c are Lagrangian multipliers respectively. The second and third terms in (5) represent entropy maximization as a regularization which was introduced in Fuzzy c-Means by Miyamoto *et al.* [11] for the first time. It enables us to obtain fuzzy clusters. T_u and T_w are the weighting parameters which specify the degree of fuzziness. The remaining terms describe the constraints of memberships, *i.e.*, (1) and (2), respectively.

From the necessary conditions for the optimality of the objective function L, *i.e.*, $\partial L/\partial u_{ci} = 0$ and $\partial L/\partial w_{cj} = 0$, we have the following equations.

$$u_{ci} = \frac{\exp(\sum_{j=1}^{N} w_{cj} d_{ij}/T_u)}{\sum_{c=1}^{C} \exp(\sum_{j=1}^{N} w_{cj} d_{ij}/T_u)}, \tag{6}$$

$$w_{cj} = \frac{\exp(\sum_{i=1}^{M} u_{ci} d_{ij}/T_w)}{\sum_{j=1}^{N} \exp(\sum_{i=1}^{M} u_{ci} d_{ij}/T_w)}. \tag{7}$$

The optimization algorithm is based on Picard iteration through necessary conditions for local minima of the objective function. Therefore, the proposed algorithm can be written as follows:

The FCCM Algorithm

Step 1 Set values of parameters C, T_u, T_w and ϵ. Initialize memberships u_{ci} randomly.

Step 2 Update membership w_{cj} using (7).

Step 3 Update memberships u_{ci} using (6).

Step 4 If $\max |u_{ci}^{NEW} - u_{ci}^{OLD}| < \epsilon$, then stop. Otherwise, return to *Step 2*.

3 Numerical Example

3.1 Literature Retrieval Data Set

Table 2: literature retrieval data set

	Key1	Key2	Key3	Key4	Key5	Key6	Key7	Key8	Key9	Key10	Key11	Key12
Lit.1	1	1	1	0	0	0	0	0	0	0	0	0
Lit.2	0	0	1	1	1	1	1	0	1	0	0	0
Lit.3	0	1	0	1	1	0	0	1	0	0	0	0
Lit.4	1	0	0	0	2	0	0	1	0	0	0	0
Lit.5	0	0	0	1	0	1	1	0	0	0	0	0
Lit.6	0	0	0	0	0	0	0	0	0	1	0	0
Lit.7	0	0	0	0	0	0	0	0	0	1	1	0
Lit.8	0	0	0	0	0	0	0	0	0	1	1	1
Lit.9	0	0	0	0	0	0	0	0	1	0	1	1

In numerical example, we apply our proposed method to literature retrieval data set used in [10] and [12]. We also apply the Correspondence Analysis to the data set and compare it with the proposed method. The data set is shown in Table 2. The rows represent the literatures and the columns are the key words. The data set shows the cooccurrence relations among the literatures and the key words. Each entry denotes the number of appearances of the key word in the corresponding literature. For example, the key word 5 appears twice in

the literature 4. The retrieval of literatures would be done in a system as to the cooccurrence relations. For instance, the literatures 6, 7 and 8 might be retrieved if the key word 10 is entered into the retrieval system according to Table 2. However, the literature 9 should be retrieved in focusing the attention on the cooccurrence relations to other key words, *i.e.*, key words 11 and 12.

3.2 Numerical Results

We used the following values of parameters for the FCCM.

The number of clsuters C: 2
The degree of fuzziness T_u: 0.1
The degree of fuzziness T_w: 1.5
Stopping condition of the FCCM ϵ: 0.0001

The results are shown in Table 3 and Table 4. In Table 3 and Table 4, we underlined larger memberships of literatures and key words. we assume that literatures and key words are more likely to belong to the cluster to which they have larger memberships. From Table 3, we can see that literatures are divided into $\{1, 2, 3, 4, 5\}$ and $\{6, 7, 8, 9\}$. On the one hand, key words are partitioned into $\{1, 2, 3, 4, 5, 6, 7, 8\}$ and $\{9, 10, 11, 12\}$. These results are reasonable in accordance with Table 2.

Table 3: menberships of literatures

Literature	Cluster 1	Cluseter 2
1	0.338	0.662
2	0.011	0.989
3	0.011	0.989
4	0.002	0.998
5	0.141	0.859
6	0.894	0.106
7	0.988	0.012
8	0.996	0.004
9	0.973	0.027

Table 4: memberships of key words

Key word	Cluster 1	Cluster 2
1	0.044	0.066
2	0.044	0.066
3	0.044	0.066
4	0.039	0.146
5	0.035	0.311
6	0.038	0.075
7	0.038	0.075
8	0.035	0.083
9	0.067	0.043
10	0.237	0.024
11	0.250	0.023
12	0.129	0.022

Figure 1 shows the result of the Correspondence Analysis applied to Table 2 and represents the scatter diagram of literatures and key words after quantification. The values corresponding to the first and second eigen values were plotted on the diagram. The horizontal axis corresponds to the first eigen value and the

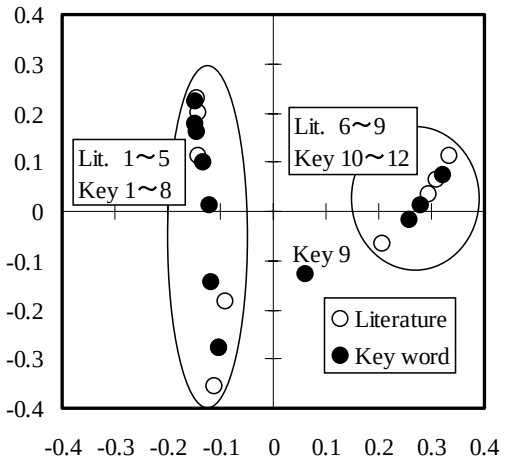

Figure 1: the result of the Correspondence Analysis

vertical axis does the second one. In Figure 1, ∘ and • indicate literatures and key words respectively. We can divide literatures and key words into two groups according to the observation of Figure 1. One is literatures $\{1, 2, 3, 4, 5\}$ and key words $\{1, 2, 3, 4, 5, 6, 7, 8\}$ and the other is literatures $\{6, 7, 8, 9\}$ and key words $\{10, 11, 12\}$. Only the key word 9 belongs to neither group. The two groups are circled in Figure 1. Comparing the result of the Correspondence Analysis with that of the proposed method, we can observe that the similar results are obtained except for the key word 9.

4 Conclusions

In this paper, we proposed a new fuzzy clustering algorithm, the FCCM, for categorical multivariate data to which the conventional fuzzy clustering algorithms could not be applied. The FCCM was applied to the literature retrieval data set which was a kind of categorical multivariate data set in the numerical example. The Correspondence Analysis was also applied to the data set and compared with the FCCM. The FCCM showed the similar result to that of the Correspondence Analysis. While the Correspondence Analysis requires solving eigen value problem which is computationally demanding, the FCCM needs simple algebraic calculations. Therefore, we can conclude that the FCCM is not only a fuzzy clustering algorithm handling categorical multivariate data but also a simple alternative of the Correspondent Analysis.

Besides, if we modify the definition of memberships in (2), we can apply our proposed method to the case where only similarities among data are given. In that case, we redefine the constraint (2) as in (8).

$$\sum_{j=1}^{M} w_{cj} = 1, \quad w_{cj} \in [0, 1], \quad c = 1, \cdots, C, \tag{8}$$

In (8), w_{cj} is not for categories but individuals. This modification leads to the similar result of Hayashi's Quantification Method Type IV [4], which also handles the same kind of similaritiy data set.

References

[1] P. Adriaans, and D. Zantinge, *Data Mining*, Addison Wesley Longman, 1996.

[2] M. J. A. Berry, and G. S. Linoff, *Data Mining Techniques*, John Wiley & Sons, 1997.

[3] M. Tenenhaus, and F. W. Young, "An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data," *Psychometrika*, Vol.50, No.1, 1985, pp 91-119.

[4] C. Hayashi, "On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematical statistical point of view," *Annals of the Institute of Statistical Mathematics*, Vol.3, 1952, pp 69-98.

[5] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*, Plenum Press, New York, 1981.

[6] J. C. Bezdek, C. Coray, R. Gunderson, and J. Watson, "Detection and characterization of cluster substructure. I. linear structure, fuzzy c-lines," *SIAM J. Appl. Math.*, Vol.40, No.2, 1981, pp 339-357.

[7] J. C. Bezdek, C. Coray, R. Gunderson, and J. Watson, "Detection and characterization of cluster substructure. II. fuzzy c-varieties and convex combinations thereof," *SIAM J. Appl. Math.*, Vol.40, No.2, 1981, pp 358-372.

[8] R. J. Hathaway, and J. C. Bezdek, "Switching regression models and fuzzy clustering," *IEEE Trans. on Fuzzy Systems*, Vol.1, No.3, 1993, pp 195-204.

[9] A. Yamakwa, Y. Kanaumi, H. Ichihashi, and T. Miyoshi "Simultaneous Application of Clustering and Correspondence Analysis," *Proc. of IJCNN'99*, Paper #625, 1999, pp 1-6.

[10] K. Inoue, and K. Urahama, " Fuzzy Clustering Based on Cooccurrence Matrix and Its Application to Data Retrieval," *Trans. of IEICE D-II*, Vol.J-81-DII, No.12, 2000, pp 957-966 (in Japanese).

[11] S. Miyamoto, and M. Mukaidono, "Fuzzy c-means as a regularization and maximum entropy approach," *Proc. of IFSA'97*, Vol.II, 1997, pp 86-92.

[12] T. K. Landauer, and S. T. Dumais: "The latent semantic analysis theory of acquisition, induction and representation of knowledge," *Psychol. Rev.*, Vol.104, No.2, 1997, pp 211-240.