The Third AFSS(1998). 139－142

HANDLING MISSING VALUES IN FUZZY c-MEANS

Sadaaki Miyamoto [a], Osamu Takata [b], and Kazutaka Umayahara [c]

a. Institute of Information Sciences and Electronics, University of Tsukuba
1-1-1 Tennodai,Tsukubashi,Ibaraki,305 Japan
Tel:+81-298-53-5346 Fax:+81-298-53-6471 Email:miyamoto@esys.tsukuba.ac.jp
b. Master's Program in Science and Engineering, University of Tsukuba
1-1-1 Tennodai,Tsukubashi,Ibaraki,305 Japan
Tel:+81-298-53-6553 Fax:+81-298-53-6471 Email:takata@odin.esys.tsukuba.ac.jp
c. Institute of Information Sciences and Electronics, University of Tsukuba
1-1-1 Tennodai,Tsukubashi,Ibaraki,305 Japan
Tel:+81-298-53-6553 Fax:+81-298-53-6471 Email:uma@is.tsukuba.ac.jp

Abstract

Missing values in data for fuzzy c-means clustering is discussed. Two basic methods of fuzzy c-means, i.e., the standard fuzzy c-means and the entropy method are considered and three options of handling missing values are proposed, among which one is to define a new distance between data with missing values, second is to alter a weight in the new distance, and the third is to fill in the missing values by an appropriate numbers. Experimental results are shown.

Keywords

fuzzy c-means, entropy method, missing value

1 Introduction

Missing values have frequently been encountered in data analysis including clustering of data in real applications. In agglomerative clustering, dissimilarity measures including missing values have been considered [1]. Moreover, Jain and Dubes [5] mention general approaches of handling missing values in crisp cluster analysis. Since they refer to definition of distances between data units in the presence of missing values, their methods primarily concern hierarchical clustering.

On the other hand, recent development and applications of fuzzy clustering, which especially concentrate on fuzzy c-means, show significance of this class of techniques.

These observations lead us to the study of missing values in fuzzy clustering, and hence the aim of the present paper is to show how to handle missing values in methods of fuzzy c-means clustering. Two fundamentally different methods of fuzzy c-means, i.e., the standard method by Dunn [3, 4] and Bezdek [2] and the entropy method [6, 7] are considered. In addition, three techniques of handling missing data are studied.

Clustering algorithms are developed for these different combinations of methods and a numerical example is given.

2 Two methods of fuzzy c-means

Two methods here mean the standard fuzzy c-means [2] and the entropy method (e.g., see [7]). Let us briefly review these methods.

2.1 Standard fuzzy c-means

Let $X = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ be set of objects to be clustered. Each $\boldsymbol{x}_i = (x_i^1, ..., x_i^p)$ is a point in p dimensional Euclidean space. Objective function to be minimized is denoted by J and in the standard method we put

$$J = J_m(U, v) = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m d_{ik}, \quad m > 1$$

where $U = (u_{ik})$ is membership matrix with the constraint

$$M = \{(u_{ik}) | \sum_{i=1}^{c} u_{ik} = 1, u_{ik} \in [0,1], \\ k = 1, ..., n\}.$$

and

$$d_{ik} = \|\boldsymbol{x}_k - \boldsymbol{v}_i\|_2^2$$

where $\|\cdot\|$ is the Euclidean norm, and moreover

$$v = (\boldsymbol{v}_1, ..., \boldsymbol{v}_c),$$

namely, v is the abbreviation of all cluster centers.

The following general algorithm of alternative optimization is common to the two methods herein.

Fuzzy c-means algorithm:

CM1. Set initial values for U and v.

CM2. Minimize $J(U,v)$ with respect to $U \in M$ while the previous v is regarded as a fixed parameter.

CM3. Minimize $J(U,v)$ with respect to v while the previous U is regarded as a fixed parameter.

CM4. Check the stopping criterion (the description of the condition is omitted here for simplicity). If the criterion is not satisfied, go to **CM2**.

As is well-known the optimal solution in **CM2** is

$$u_{ik} = [\sum_{j=1}^{c} (\frac{d_{ik}}{d_{jk}})^{\frac{1}{m-1}}]^{-1} \tag{1}$$

and the optimal v in **CM3** is

$$v_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m}. \tag{2}$$

2.2 Entropy method

There are different approaches to the method of entropy [6, 7]. Here we adopt the regularization approach proposed by Miyamoto and Mukaidono [7], since it uses the alternative optimization algorithm **CM** and therefore easier to describe.

The entropy method uses the objective function

$$\begin{aligned} J &= J^{\lambda}(U,v) \\ &= \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik} d_{ik} + \lambda^{-1} \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik} \log u_{ik} \end{aligned}$$

with the same distance and the same constraint M. The parameter λ is compared to the regularizing parameter in ill-posed problems and hence this method is called regularization by entropy [7].

The second term of the entropy function is for fuzzification, while the same role is played by the parameter m in the standard method.

Thus the algorithm **CM** with

$$J(U,v) = J^{\lambda}(U,v)$$

is used in the entropy method.

The optimal solutions in **CM2** is given by

$$u_{ik} = \frac{e^{-\lambda d_{ik}}}{\sum_{j=1}^{c} e^{-\lambda d_{jk}}}. \tag{3}$$

and that in **CM3** is

$$v_i = \frac{\sum_{k=1}^{n} u_{ik} x_k}{\sum_{k=1}^{n} u_{ik}}. \tag{4}$$

Theoretical properties of the two methods are compared using fuzzy classification functions in which x_k are replaced by the variable x in (1) and (3) [8].

3 Handling missing values

Without loss of generality, assume for the moment that the first component of x_1 is missing:

$$x_1 = (*, x_1^2, ..., x_1^p).$$

Obviously, the distance $\|x_1 - v_i\|^2$ cannot be calculated. There are two major approaches of dealing with missing data. First, objects with missing values are simply deleted and the rest is considered for the clustering. Second approach is to define the distance d_{i1} between x_1 and v_i in more or less heuristic manner.

Since the first way is uninteresting, the second approach alone is studied here, in which three different definitions of d_{i1} are introduced:

1. Simply ignore the missing value and calculate the distance from the resulting coordinate:

$$d_{i1} = \sum_{\ell=2}^{p} (x_1^{\ell} - v_i^{\ell})^2. \tag{5}$$

2. Ignore the missing coordinate and multiply $p/(p-1)$:

$$d_{i1} = \frac{p}{p-1} \sum_{\ell=2}^{p} (x_1^{\ell} - v_i^{\ell})^2. \tag{6}$$

(If q coordinates are missing, multiply $p/(p-q)$.)

3. Replace $*$, the missing value, by the weighted average:

$$x_1^1 = \frac{\sum_{i=1}^{c} (u_{i1})^m v_i^1}{\sum_{i=1}^{c} (u_{i1})^m}. \tag{7}$$

(In the entropy method, put $m = 1$.)

It is easy to see that d_{ik}'s by (5) and (6) lead to the same solutions, since the multiplier $p/(p-1)$ is cancelled out in (1) and (3). We therefore should consider only two methods of (5) and (7) in the present class of fuzzy c-means. (It cannot generally be proved that (5) and (6) lead to the same solution in variations of the fuzzy c-means, and hence they should be considered to be different methods of handling missing values.)

4 A numerical example

Figure 1 shows a set of points scattered on a plane. Some points have missing values in the vertical coordinates, and they are shown on the horizontal axis, as if their vertical coordinates were zero. The result of clustering by fuzzy c-means is shown in Fig.2, which is obtained from the cut of $\alpha = 0.5$.

Figures 3 and 4 show the projections of the points onto the horizontal axis, with the vertical coordinate of the membership values. The dissimilarity by (5) is used in the upper figure while (7) is used in the lower. The standard fuzzy c-means (i.e., (1) and (2)) is used. Those points without the missing value are shown by dots while the points with the missing value are shown by circles. The circles are interpolated by the curves to emphasize them.

Figures 5 and 6 show the projections of the points in the same way, except that the entropy method is used. The upper uses (5) and the lower uses (7).

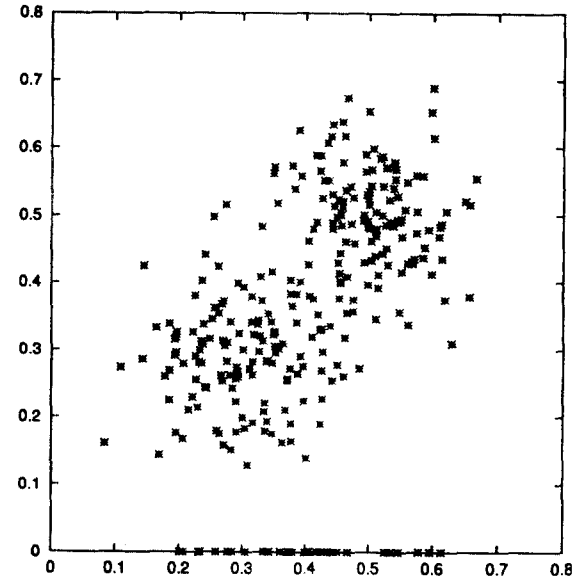

Figure 1: A set of points scatterd on a plane.

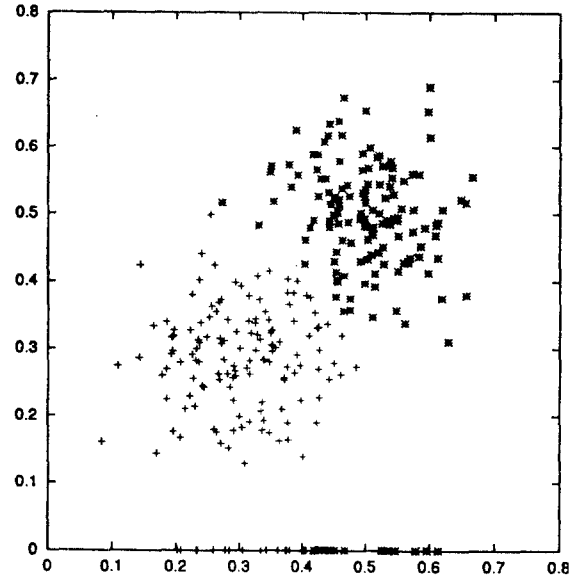

Figure 2: Result by fuzzy c-means with the cut of $\alpha = 0.5$.

5 Conclusion

Data with missing values have been considered and two methods of fuzzy c-means have been applied. Three options of handling missing values have been proposed, but two of which are identical. Still another method of fuzzy c-means is being developed [9], for which the three options are all different.

The method of handling missing values can be generalized to data with uncertainties of more general form, which is under development.

Moreover, real data with missing values should be analyzed which is a main subject in future studies.

Acknowledgment

This research has partly been supported by Tsukuba Advanced Research Alliance, University of Tsukuba.

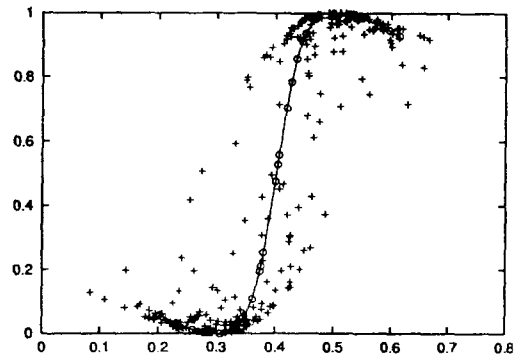

Figure 3: Standard fuzzy c-means by option 1.

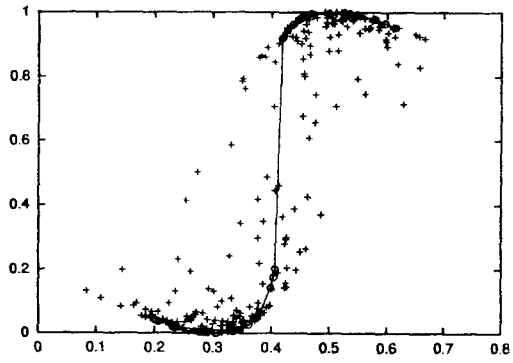

Figure 4: Standard fuzzy c-means by option 3.

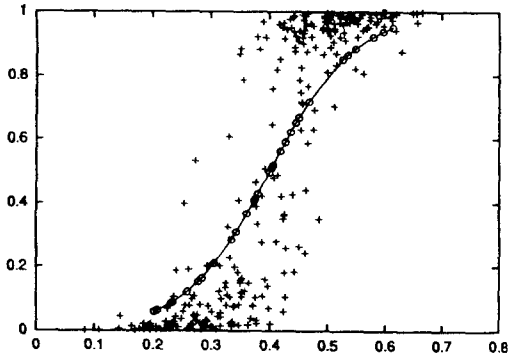

Figure 5: Entropy method by option 1.

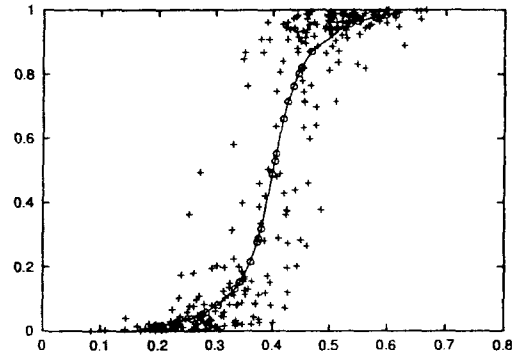

Figure 6: Entropy method by option 3.

References

[1] M. R. Anderberg, *Cluster Analysis for Applications*, Academic Press, New York, 1973.

[2] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum, New York, 1981.

[3] J. C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, *J. of Cybernetics*, Vol.3, pp. 32–57, 1974.

[4] J. C. Dunn, Well-separated clusters and optimal fuzzy partitions, *J. of Cybernetics*, Vol.4, pp. 95–104, 1974.

[5] A. K. Jain, R. C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, NJ, 1988.

[6] R.-P. Li and M. Mukaidono, A maximum entropy approach to fuzzy clustering, *Proc. of the 4th IEEE Intern. Conf. on Fuzzy Systems (FUZZ-IEEE/IFES'95)*, Yokohama, Japan, March 20-24, 1995, pp. 2227–2232.

[7] 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 (IFSA'97)*, June 25-30, 1997, Prague, Chech, Vol.II, 86-92, 1997.

[8] S. Miyamoto, Two methods of fuzzy c-means and classification functions, *Proc. of IFCS'98*, Rome, Italy, July 21-24, 1998, to appear.

[9] S. Miyamoto and K. Umayahara, Fuzzy clustering by quadratic regularization, *Proc. of FUZZ-IEEE'98*, Anchorage, USA, May 4-9, 1998, to appear.