Clustering Algorithms Based on Tolerance Vector Concept

ENDO Yasunori†, HASEGAWA Yasushi‡, HAMASUNA Yukihiro‡

†Department of Risk Engineering,
Faculty of Systems and Information Engineering,
University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan
Email: endo@risk.tsukuba.ac.jp
‡Graduate School of Systems and Information Engineering,
University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan

Abstract—This paper provides new clustering algorithms for data with tolerance. Tolerance includes wide meanings, e.g., calculation errors and loss of attribute of data. The tolerance is modified using by new concept of tolerance vector. First, the concept is explained and optimization problems of clustering are formulated using the vectors. Second, the problems are solved using some different ways. Third, the new clustering algorithms are constructed by using the solutions of the problems. Moreover, the effectiveness of proposed algorithms is verified through some numerical examples.

1. Introduction

Clustering is one of the unsupervised classification and fuzzy c-means (FCM)[1] is one of the typical technique of fuzzy clustering.

Information on a real space is transformed to data in a pattern space and analyzed in clustering. Therefore, there are some problems that should be considered when transforming, for example, measurement error margin, data that cannot be regarded as one point, and missing values in data. In the past, these uncertainties of data have been represented as interval range and many clustering algorithms for these interval ranges of data have been constructed[2, 3] and one of the authors have also proposed one of such algorithms[4, 5]. In these algorithms, nearest neighbor distance, furthest neighbor distance and Hausdorff distance have been used to calculate the dissimilarity between the target data in clustering. However, the guideline to select the available distance in each case has not been shown so that this problem is difficult. When we consider such a situation, it is more desirable to calculate the dissimilarity between such interval ranges of data without introducing a particular distance, e.g., nearest neighbor one and so on.

One of the authors introduced the new concept of tolerance which includes the above-mentioned uncertainties of data and is different from the interval from the viewpoint of introduction of tolerance vectors, and proposed two clustering algorithms, one is based on Euclidean norm[6] and the other is L_1-norm[7]. The tolerance is defined as hypersphere in these algorithms.

In this paper, we consider new optimization problems in which the tolerance is defined as hyper-rectangle and we construct new clustering algorithms based on sFCM (Standard Fuzzy c-means)[1] on L_1-norm and Euclidean (L_2) norm for data with tolerance through solving the optimization problems.

2. Theory

In this section, we discuss about optimization problems for clustering.

We define some notations at the beginning. $X = \{x_1, \dots, x_n\}$ is a subset on p dimensional vector space $\mathbf{R}^p$ and we write $x_k = (x_{k1}, \dots, x_{kp})^T \in \mathbf{R}^p$. Here, we consider classifying the data set X into clusters $C_i (i = 1, \dots, c)$. Let $v_i = (v_{i1}, \dots, v_{ip})^T \in \mathbf{R}^p$ be the cluster center C_i and $V = \{v_1, \dots, v_c\}$ be the set of cluster centers. Moreover, μ_{ki} is the membership grade belonging x_k to C_i and we denote the partition matrix $U = [\mu_{ki}]$. Fuzzy c-means calculates V and U which minimize a objective function by alternative optimization.

2.1. Tolerance Vector

Here, we define tolerance vector $\varepsilon_k = (\varepsilon_{k1}, \dots, \varepsilon_{kp})^T \in \mathbf{R}^p$, and $E = \{\varepsilon_1, \dots, \varepsilon_n\}$ is the set of tolerance. The absolute value of ε_{kj} is restricted by the maximum tolerance κ_{kj}, and the constraint condition is shown by the following expression.

$$|\varepsilon_{kj}| \leq \kappa_{kj}, \ (\kappa_{kj} > 0). \tag{1}$$

Fig.1 shows the concept of tolerance in $\mathbf{R}^2$.

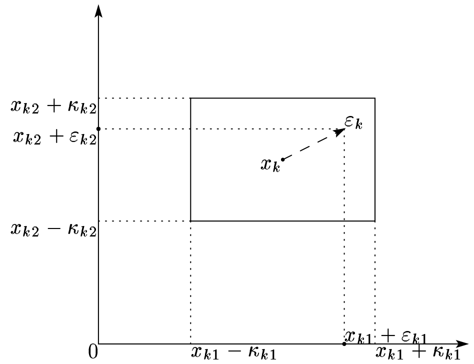

Figure 1: An example of the tolerance vector in $\mathbf{R}^2$.

2.2. sFCM for data with tolerance on L_1-norm

The objective function of sFCM based on L_1-norm is defined by Jajuga[9].

$$J_{\mathrm{sFCM}\text{-}L_1} = \sum_{i=1}^{c}\sum_{k=1}^{n}\sum_{j=1}^{p}\mu_{ki}^{m}|x_{kj} - v_{ij}|, \tag{2}$$

under the constraint

$$\sum_{i=1}^{c}\mu_{ki} = 1,\ (\mu_{ki} \geq 0). \tag{3}$$

We define the following objective function based on the above equation.

$$J_{\mathrm{sFCMT_R}\text{-}L_1} = \sum_{i=1}^{c}\sum_{k=1}^{n}\mu_{ki}^{m}d_{ki}, \tag{4}$$

where

$$d_{ki} = \sum_{j=1}^{p}|x_{kj} + \varepsilon_{kj} - v_{ij}|.$$

The following optimal solution is obtained by using the Lagrange function.

$$\mu_{ki} = \left(\sum_{s=1}^{c}\left(\frac{d_{ki}}{d_{ks}}\right)^{\frac{1}{m-1}}\right)^{-1}. \tag{5}$$

Here we propose two methods to obtain v_{ij}. One is based on Ref.[10], called **Method 1**. The other is based on Ref.[9], called **Method 2**.

Method 1

From (4), semi-objective function is

$$J_{ij}(v_{ij}) = \sum_{k=1}^{n}\mu_{ki}^{m}|x_{kj} + \varepsilon_{kj} - v_{ij}|. \tag{6}$$

If this equation is minimized, the objective function is also minimized. The optimal solution of v_{ij} is calculated according to the following procedures. The advantage of this method is to obtain the exact optimum solutions.

Step 1 Data is sorted in ascending order in each dimension.

$$x_{1j} + \varepsilon_{1j}, \ldots, x_{nj} + \varepsilon_{nj}$$
$$\downarrow \text{Sorting}$$
$$x_{q(1)j} + \varepsilon_{q(1)j} \leq \ldots \leq x_{q(n)j} + \varepsilon_{q(n)j}$$

where $q(k)$ is substitution of $(1, \ldots, n)$.

Step 2 We calculate as follows.

$$S = -\frac{1}{2}\sum_{k=1}^{n}(\mu_{ki})^{m}.$$

Step 3 It starts from $r = 0$ and the following calculations are repeated between $S < 0$.

$$r := r + 1;$$
$$S := S + (\mu_{q(r)i})^{m};$$

Step 4 From the above calculation, we obtain

$$v_{ij} = x_{q(r)j} + \varepsilon_{q(r)j}. \tag{7}$$

Method 2

From (4), semi-objective function is

$$J_{ij}(v_{ij}) = \sum_{k=1}^{n}w_{ki}(x_{kj} + \varepsilon_{kj} - v_{ij})^2, \tag{8}$$

where

$$w_{ki} = \frac{\mu_{ki}^{m}}{|x_{kj} + \varepsilon_{kj} - v_{ij}|}.$$

From (8),

$$\frac{\partial J_{ij}}{\partial v_{ij}} = -2\sum_{k=1}^{n}w_{ki}(x_{kj} + \varepsilon_{kj} - v_{ij}) = 0. \tag{9}$$

Then, we have

$$v_{ij} = \frac{\sum_{k=1}^{n}w_{ki}(x_{kj} + \varepsilon_{kj})}{\sum_{k=1}^{n}w_{ki}}. \tag{10}$$

Next, we consider the way to obtain ε_{kj}. The procedure is as same as v_{ij}.

Method 1

Step 1 Data is sorted in ascending order in each dimension.

$$v_{1j} - x_{kj}, \ldots, v_{cj} - x_{kj}$$
$$\downarrow \text{Sorting}$$
$$v_{q(1)j} - x_{kj} \leq \ldots \leq v_{q(c)j} - x_{kj}$$

where $q(i)$ is substitution of $(1, \ldots, c)$.

Step 2 We calculate as follows.

$$S = -\frac{1}{2}\sum_{i=1}^{c}(\mu_{ki})^{m}.$$

Step 3 It starts from $r = 0$ and the following calculations are repeated between $S < 0$.

$$r := r + 1;$$
$$S := S + (\mu_{kq(r)})^{m};$$

Step 4 From the above calculation, we obtain

$$\varepsilon_{kj} = \mathrm{sign}(v_{q(r)j} - x_{kj}) \times \min\{|v_{q(r)j} - x_{kj}|, \kappa_{kj}\}. \tag{11}$$

Method 2

From (4), semi-objective function is

$$J_{kj}(\varepsilon_{kj}) = \sum_{i=1}^{c}w_{ki}(x_{kj} + \varepsilon_{kj} - v_{ij})^2. \tag{12}$$

We partially differentiate (12) with respect to ε_{kj} and we have

$$\varepsilon_{kj} = \frac{\sum_{i=1}^{c}w_{ki}(v_{ij} - x_{kj})}{\sum_{i=1}^{c}w_{ki}}. \tag{13}$$

From (1) and (13), we obtain

$$\varepsilon_{kj} = \mathrm{sign}\left(\frac{\sum_{i=1}^{c}w_{ki}(v_{ij} - x_{kj})}{\sum_{i=1}^{c}w_{ki}}\right) \times \min\left\{\left|\frac{\sum_{i=1}^{c}w_{ki}(v_{ij} - x_{kj})}{\sum_{i=1}^{c}w_{ki}}\right|, \kappa_{kj}\right\}. \tag{14}$$

2.3. sFCM for data with tolerance on Euclidean norm

We introduce the tolerance into the objective function of sFCM by Bezdek[1].

$$J_{\mathrm{sFCMT_R}\text{-}L_2} = \sum_{i=1}^{c}\sum_{k=1}^{n}\mu_{ki}^{m}\|x_k + \varepsilon_k - v_i\|^2, \tag{15}$$

where

$$\|x_k + \varepsilon_k - v_i\|^2 = \sum_{j=1}^{p}(x_{kj} + \varepsilon_{kj} - v_{ij})^2,$$

under the constraints

$$\sum_{i=1}^{c}\mu_{ki} = 1,\ (\mu_{ki} \geq 0), \tag{16}$$

$$\varepsilon_{kj}^2 \leq \kappa_{kj}^2,\ (\kappa_{kj} > 0). \tag{17}$$

We introduce the following Lagrange function to solve the optimization problem,

$$L_s = \sum_{k=1}^{n}\sum_{i=1}^{c}\mu_{ki}^{m}\|x_k + \varepsilon_k - v_i\|^2 + \sum_{k=1}^{n}\gamma_k(\sum_{i=1}^{c}\mu_{ki} - 1) + \sum_{k=1}^{n}\sum_{j=1}^{p}\delta_{kj}(\varepsilon_{kj}^2 - \kappa_{kj}^2). \tag{18}$$

From the Kuhn-Tucker condition, we get as follows:

$$\mu_{ki} = \left(\sum_{l=1}^{c}\left(\frac{\|x_k + \varepsilon_k - v_i\|^2}{\|x_k + \varepsilon_k - v_l\|^2}\right)^{\frac{1}{m-1}}\right)^{-1}. \tag{19}$$

$$v_{ij} = \frac{\sum_{k=1}^{n}\mu_{ki}^{m}(x_{kj} + \varepsilon_{kj})}{\sum_{k=1}^{n}\mu_{ki}^{m}}. \tag{20}$$

$$\varepsilon_{kj} = -\alpha_{kj}\sum_{i=1}^{c}\mu_{ki}^{m}(x_{kj} - v_{ij}), \tag{21}$$

where

$$\alpha_{kj} = \min\left\{\frac{\kappa_{kj}}{|\sum_{i=1}^{c}\mu_{ki}^{m}(x_{kj} - v_{ij})|}, \frac{1}{\sum_{i=1}^{c}\mu_{ki}^{m}}\right\}.$$

Note that we can develop the same discussion for entropy regularized FCM (eFCM) as sFCM in which the following objective function is considered:

$$J_{\mathrm{eFCMT_R}} = \sum_{i=1}^{c}\sum_{k=1}^{n}\mu_{ki}d_{ki} + \lambda^{-1}\sum_{i=1}^{c}\sum_{k=1}^{n}\mu_{ki}\log\mu_{ki}. \tag{22}$$

3. Algorithms

The algorithms derived in the above section are called sFCMT$_{\mathbf{R}}$-L_1-1, sFCMT$_{\mathbf{R}}$-L_1-2, and sFCMT$_{\mathbf{R}}$-L_2 in turn.

Each algorithm is calculated according to the following procedure. Eqs. **A**, **B** and **C** used in each algorithm follow Table 1.

Algorithm

```
Step 1 Give the values of $m$ and $\kappa_{kj}$, and set initial
    values of $E$ and $V$.
Step 2 Calculate $U = \mu_{ki}$ by Eq. A.
Step 3 Calculate $V = v_{ij}$ by Eq. B.
Step 4 Calculate $E = \varepsilon_{kj}$ by Eq. C.
Step 5 If (U,E,V) is convergent, stop. Otherwise, go
    back to Step 2.
```

Table 1: This table shows optimal solutions of each algorithm.

Algorithm	Eq. **A**	Eq. **B**	Eq. **C**
sFCMT$_\mathrm{R}$-L_1-1	(5)	(7)	(11)
sFCMT$_\mathrm{R}$-L_1-2	(5)	(10)	(14)
sFCMT$_\mathrm{R}$-L_2	(19)	(20)	(21)

4. Numerical examples

In this section, we show some examples of classification by the above-mentioned six algorithms. The classified data set is diagnosis of heart disease[11]. The result of the diagnosis is known. We choose five attributes from 13 ones of original data referring to the advice of a specialist. The number of data is 866 and 560 data contains missing values in some attributes. Please refer to Table 2 for the explanation of each attribute and the number of missing values.

Table 2: The explanation of each attribute and the number of missing values.

Attribute	Number of missing values
Resting blood pressure	5
Maximum heart rate achieved	1
ST depression induced by exercise relative to rest	8
The slope of the peak exercise ST segment	255
Number of major vessels colored by fluoroscopy	557

To treat missing values as tolerance, we give the average of maximum value and minimum one to the missing one of each attribute, and set the maximum tolerance κ_{kj} on the absolute value of difference between the average and the minimum value.

In all algorithms, the convergence condition is

$$\max_{i,j}|v_{ij} - \bar{v}_{ij}| < 10^{-6},$$

where $\bar{v}_{ij}$is the previous optimal solution. In addition, $m = 2$ in sFCMT$_\mathrm{R}$.

In each algorithm, we give initial cluster centers at random and classify the data set into two clusters. We run this trial 1000 times and show the average of ratio of correctly classified results. Please refer to Table 3 for the results of only 306 data without missing values, Table 4 for the results of the classification by using the algorithms proposed in this paper and Table 5 for the results of the classification by using the algorithms which treat missing value as interval data and uses nearest neighbor distance to calculate the dissimilarity.

Table 3: The results of classifying only 306 data without missing values.

Algorithm	The average of correctly classified ratio
sFCM-L_1-1	70.0
sFCM-L_1-2	71.9
sFCM-L_2	75.2

Table 4: The results of the classification by using the proposed algorithms in this paper.

Algorithm	The average of correctly classified ratio
sFCMT$_R$-L_1-1	68.6
sFCMT$_R$-L_1-2	67.4
sFCMT$_R$-L_2	73.4

Table 5: The results of the classification by using the algorithms which treat missing value as interval data and use nearest neighbor distance to calculate dissimilarity.

Algorithm	The average of correctly classified ratio
sFCM-L_1-1	69.0
sFCM-L_1-2	68.9
sFCM-L_2	67.2

To compare the results for all data by the proposal algorithms(Table 4) with the results for only data without missing values(Table 3), the latter is a little better than the former. However, this is very natural. From these examples, we can not find significant difference between the algorithms, using the tolerance and nearest neighbor distance. The important point is that ε is calculated only by the proposed algorithms. The meaning of ε depends on the data set.

5. Conclusion

In this paper, we considered the optimization problems for data with tolerance and solved the optimal solutions. Using the results, we have constructed new six algorithms. Moreover, we shown the usefulness of the proposed algorithms through some numerical examples.

In these algorithms, data with tolerance is not regarded as interval data. The reason is that the algorithms is more appropriate because we can use the former dissimilarities based on normal distance in the frame of the optimization. Moreover, we can use the proposed algorithms for the data with tolerance defined as hyper-rectangle in more cases than the algorithms for the data with tolerance defined as hyper-sphere[6, 7] because more certainties should be represented as hyper-rectangle than hyper-sphere.

In the forthcoming paper, we will consider to apply the concept of tolerance to regression analysis and support vector machine.

Acknowledgment

This study is partly supported by the Grant-in-Aid for Scientific Research (C) (project No.18500170) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

[1] J.C.Bezdek : “Pattern Recognition with Fuzzy Objective Function Algorithms”, Plenum(1981).

[2] Osamu Takata, Sadaaki Miyamoto : “Fuzzy clustering of Data with Interval Uncertainties”, Journal of Japan Society for Fuzzy Theory and Systems, Vol.12, No.5, pp.686-695 (2000) (in Japanese)

[3] M. Sato-Ilic and J. Oshima : “On Weighted Principal Component Analysis for Interval-Valued Data and Its Dynamic Feature”, International Journal of Innovative Computing, Information and Control, Vol.2, No.1, pp.69-82 (2006).

[4] Yasunori Endo : “Clustering Algorithm Using Covariance for Fuzzy Data,”In Proc. 1998 International Symposium on Nonlinear Theory and Its Applications, pp.511-514 (1998.9).

[5] Yasunori Endo, Kazuo Horiuchi : “On Clustering Algorithm for Fuzzy Data,” In Proc. 1997 International Symposium on Nonlinear Theory and Its Applications, pp.381-384 (1997.11).

[6] Ryuichi Murata, Yasunori Endo, Hideyuki Haruyama, Sadaaki Miyamoto : “On Fuzzy c-Means for Data with Tolerance”, Journal of Advanced Computational Intelligence and Intelligent Informatics Vol.10, No.5, pp.673-681 (2006).

[7] Hiromi Toyoda : “L_1-Norm based Fuzzy Clustering for Data with Tolerance”, Graduation thesis, College of Engineering Systems, University of Tsukuba (2005) (in Japanese).

[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 (IFSA’97), June 25-30, (1997), Prague, Chech ,Vol.2, pp.86-92 (1997).

[9] Krzysztof Jajuga : “L_1-norm based fuzzy clustering”, Fuzzy Sets and Systems, Vol.39, pp43-50 (1991).

[10] Takashi Koga : “Clustering Algorithm based on L_1-norm space”, Master’s thesis, Graduate School of Systems and Information Engineering, University of Tsukuba (2002) (in Japanese).

[11] UCI Machine Learning Databases `http://www.ics.uci.edu/˜mlearn/databases/heart-disease/`