Hindawi Publishing Corporation
Advances in Fuzzy Systems
Volume 2011, Article ID 265170, 10 pages
doi:10.1155/2011/265170

Research Article

A Comparative Study on TIBA Imputation Methods in FCMdd-Based Linear Clustering with Relational Data

Takeshi Yamamoto, Katsuhiro Honda, Akira Notsu, and Hidetomo Ichihashi

Department of Computer Science and Intelligent Systems, Osaka Prefecture University, Osaka 599-8531, Japan

Correspondence should be addressed to Katsuhiro Honda, honda@cs.osakafu-u.ac.jp

Received 9 June 2011; Revised 28 July 2011; Accepted 31 July 2011

Academic Editor: Salvatore Sessa

Relational fuzzy clustering has been developed for extracting intrinsic cluster structures of relational data and was extended to a linear fuzzy clustering model based on Fuzzy *c*-Medoids (FCMdd) concept, in which Fuzzy *c*-Means-(FCM-) like iterative algorithm was performed by defining linear cluster prototypes using two representative medoids for each line prototype. In this paper, the FCMdd-type linear clustering model is further modified in order to handle incomplete data including missing values, and the applicability of several imputation methods is compared. In several numerical experiments, it is demonstrated that some pre-imputation strategies contribute to properly selecting representative medoids of each cluster.

1. Introduction

Relational fuzzy clustering is a relational extension of fuzzy clustering for revealing cluster structures buried in relational data. Relational Fuzzy *c*-Means (RFCM) [1] extended the Fuzzy *c*-Means (FCM) [2] clustering criterion with mutual dissimilarity measures instead of object-type observation in FCM. Although FCM and other variants of *k*-Means [3] use the clustering criterion of the distance between a data point and a cluster prototype, RFCM defines the clustering criterion by using mutual dissimilarities only. When the dissimilarities among objects are measured by squared Euclidean distances, the RFCM criterion is equivalent to the centroid-less formulation of the FCM criterion. Using other dissimilarity measures; however, the RFCM criterion has no clear connection with distances between data points and prototypes. In *k*-Medoids [4], cluster prototypes are selected from data points, and the clustering criterion coincides with one of mutual dissimilar degree among objects. So, *k*-Medoids can be directly extended to relational data analysis even if taking an average cannot be done in non-Euclidean space. Fuzzy *c*-Medoids (FCMdd) [5] is a fuzzy extension of *k*-Medoids and can deal with various dissimilarity measures.

Linear fuzzy clustering models [6, 7] extract linear substructures by modifying the point prototypes of FCM into lines, planes, and linear varieties. Because the subspace learning model in each cluster can be identified with fuzzy principal component analysis (fuzzy PCA) [8], they are often regarded as a kind of local principal component analysis (local PCA) [9]. This paper studies the FCMdd-based linear clustering model [10], which can reveal local linear substructures buried in relational data. In [10], Haga et al. defined each prototypical line by using two representative medoids and demonstrated that the clustering modal can be applied to Euclidean relational data. The FCMdd-type linear clustering model was further modified for dealing with non-Euclidean relational data [11, 12], in which data transformation, called β-spread transformation, was performed before applying the clustering algorithm in a similar manner to Non-Euclidean-type Relational Fuzzy (NERF) *c*-Means [13].

In this paper, a comparative study on the applicability of β-spread transformation is performed in FCMdd-based linear clustering of incomplete relational data. Hathaway and Bezdek [14] proposed several methods for imputing (predicting and substituting) missing elements of incomplete relational data and showed that imputation errors can be revised by β-spread transformation in NERF *c*-Means. This paper demonstrates that the performance of FCMdd-type linear fuzzy clustering for incomplete relational data can also

be improved by β-spread transformation through several comparative experiments including an example of document clustering.

The remaining part of this paper is organized as follows. In Section 2, linear clustering and relational clustering are briefly reviewed. Section 3 introduces FCMdd-type linear clustering model and applies several imputation methods called TIBA. Comparative results are shown in Section 4, and conclusions are given in Section 5.

2. Linear Clustering and Relational Clustering

2.1. FCM-Type Linear Clustering. Assume that we have m-dimensional observations of n patterns $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})^\top$, $i = 1, \ldots, n$. With the goal of partitioning the n patterns into C clusters, the objective function for FCM-type clustering is defined as

$$L_{fcm} = \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci}^{\theta} D_{ci}. \tag{1}$$

u_{ci} is the fuzzy membership degree of pattern i to cluster c, and θ is the fuzzification parameter. The larger the θ, the fuzzier the membership assignment. D_{ci} is the clustering criterion which measures the deviation between pattern i and the prototype of cluster c. In the original FCM clustering [2], cluster prototypes are given by the centroid vectors $\mathbf{b}_c$, and D_{ci} is the squared Euclidean distance as follows:

$$D_{ci} = \|\mathbf{x}_i - \mathbf{b}_c\|^2. \tag{2}$$

The FCM model is reduced to the hard (nonfuzzy) k-Means model [3] when $\theta = 1$, in which cluster memberships are given by the nearest prototype principle.

Besides point-type prototypes $\mathbf{b}_c$ in FCM, Fuzzy c-Lines (FCL) [6] for extracting linear clusters used linear prototypes defined as

$$\mathrm{Line}_c(\mathbf{b}_c, \mathbf{a}_c) = \{\mathbf{x} \mid \mathbf{x} = \mathbf{b}_c + t\mathbf{a}_c;\ t \in R\}, \tag{3}$$

where $\mathbf{a}_c$ is the basis vector of the principal subspace, and $\mathbf{b}_c$ is the centroid, which the linear prototype passes through. The clustering criterion is calculated as

$$D_{ci} = \|\mathbf{x}_i - \mathbf{b}_c\|^2 - \left|\mathbf{a}_c^\top (\mathbf{x}_i - \mathbf{b}_c)\right|^2. \tag{4}$$

The updating rules for membership u_{ci} and the cluster center $\mathbf{b}_c$ are derived as

$$u_{ci} = \left[\sum_{l=1}^{C} \left(\frac{D_{ci}}{D_{li}}\right)^{1/(\theta-1)}\right]^{-1}, \tag{5}$$

$$\mathbf{b}_c = \frac{\sum_{i=1}^{n} u_{ci}^{\theta} \mathbf{x}_i}{\sum_{i=1}^{n} u_{ci}^{\theta}}. \tag{6}$$

The basis vectors $\mathbf{a}_c$ are the principal eigenvectors of the generalized fuzzy scatter matrices:

$$S_{\mathrm{fc}} = \sum_{i=1}^{n} u_{ci}^{\theta} (\mathbf{x}_i - \mathbf{b}_c)(\mathbf{x}_i - \mathbf{b}_c)^\top. \tag{7}$$

This linear clustering model has close relation with local PCA [9]. Indeed, when we consider only a single cluster ($C = 1$), the FCL clustering model is equivalent to the conventional PCA and the basis vector $\mathbf{a}_c$ is reduced to the principal component vector. In this sense, FCL is a type of local PCA, which simultaneously performs membership estimation (local fuzzy group extraction) and fuzzy PCA [8] in each local fuzzy group considering the fuzzy membership degree of u_{ci}^{θ}. The prototypical line Line_c can be identified with principal subspace spanned by fuzzy principal component vector $\mathbf{a}_c$ from the local PCA view point.

When $\theta = 1$, the FCL model is also reduced to the hard (nonfuzzy) local PCA model [15, 16], in which cluster memberships are given by the nearest prototype principle.

2.2. FCM-Type Relational Clustering. RFCM [1] is the relational extension of FCM. When we have relational data composed of mutual relations among patterns $D = \{d_{ij}^2\}$, the FCM-type objective function is redefined as

$$L_{rfcm} = \sum_{c=1}^{C} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{u_{ci}^{\theta} u_{cj}^{\theta} d_{ij}^2}{2 \sum_{t=1}^{n} u_{ct}^{\theta}}. \tag{8}$$

d_{ij} can be any type of dissimilarity between patterns i and j but is assumed to be Euclidean-like one in RFCM. Indeed, this model is equivalent to FCM only when d_{ij}^2 is the squared Euclidean distance, and the clustering model derives only poor results if the relational information is highly non-Euclidean.

In order to modify RFCM for handling non-Euclidean distance metrics, Hathaway and Bezdek [13] considered NERF, which includes the following β-spread transformation:

$$D_\beta = D + \beta \times (M - I), \tag{9}$$

where β is added to off-diagonal elements of non-Euclidean relational data D. I is a unit matrix, and β is a suitably chosen scalar. M is a matrix whose elements are all 1. Hathaway and Bezdek discussed that D_β is Euclidean if $PD_\beta P$ with $P = I - (1/n)M$ is negative semidefinite; that is, β is greater than or equal to the largest eigenvalue of PDP. By the way, the basic RFCM iteration can be continued when clustering criteria are all nonnegative. In NERF, β is gradually increased from 0 to a certain value by considering the negative elements of clustering criteria.

3. FCMdd-Type Linear Clustering and TIBA Imputation

3.1. FCMdd-Type Linear Clustering. Assume that d_{ij} is the mutual Euclidean distance such that

$$d_{ii} = 0, \quad d_{ij} \geq 0, \qquad d_{ij} = d_{ji}, \qquad i, j = 0, \ldots, n. \tag{10}$$

FCMdd [5] is a fuzzy extension of k-medoids [4], which performs an FCM-like clustering by selecting $\mathbf{b}_c$ from patterns $\mathbf{x}_i$, $i = 1, \ldots, n$. The representative objects are

called "medoids" and are given by solving combinatorial optimization problems. Haga et al. [10] applied the idea to linear fuzzy clustering, in which each linear prototype is spanned by two representative medoids $\mathbf{x}_{c_1}$ and $\mathbf{x}_{c_2}$ as

$$\mathrm{Line}_c(\mathbf{x}_{c_1}, \mathbf{x}_{c_2}) = \{\mathbf{x} \mid \mathbf{x} = \mathbf{x}_{c_1} + t(\mathbf{x}_{c_2} - \mathbf{x}_{c_1});\ t \in R\}. \quad (11)$$

The squared Euclidean distance between object i and the prototypical line Line_c is given as

$$D_{ci} = d_{i,c_1}^2 - \frac{\left(d_{i,c_1}^2 - d_{i,c_2}^2 + d_{c_1,c_2}^2\right)^2}{4d_{c_1,c_2}^2}. \quad (12)$$

With fixed fuzzy memberships u_{ci}, the optimal medoids are derived by the following combinatorial optimization problem:

$$(c_1, c_2) = \underset{\substack{(k_1,k_2)\\ 1\le k_1,k_2\le n\\ k_1\neq k_2}}{\arg\min} \sum_{i=1}^{n} u_{ci}^{\theta} D_{ci}. \quad (13)$$

The optimal medoid set of (c_1, c_2) is searched by enumerating all pairs of objects. In order to reduce the computational cost, a simplified medoid search process was also proposed, in which medoids are selected from a subset X_c of objects:

$$(c_1, c_2) = \underset{\substack{(k_1,k_2)\\ \mathbf{x}_{k_1},\mathbf{x}_{k_2}\in X_c\\ k_1\neq k_2}}{\arg\min} \sum_{i=1}^{n} u_{ci}^{\theta} D_{ci}, \quad (14)$$

where $X_c = \{\mathbf{x}_i : u_{ci} > M_{\min}\}$.

This linear fuzzy clustering model was also extended to the 2D prototype case by spanning 2D prototypical planes using three medoids [10].

Although non-Euclidean relational data may bring negative values for the clustering criteria of (12), from the practical view point, we have no trouble in operating the conventional FCMdd-type linear clustering algorithm if all clustering criteria are not negative.

Yamamoto et al. [11] proposed a procedure for β-spread transformation so as to avoid negative criterion values in FCMdd-type linear clustering. Because a negative criterion value implies a non-Euclidean situation, relational data should be revised so that the criterion value is always nonnegative. In the previous research [12], it was shown that the clustering criterion D_{ci} is always nonnegative if triangle inequality ($d_{c_1,c_2} \le d_{i,c_1} + d_{i,c_2}$) is satisfied. Then, β-spread transformation should be performed so that the following triangle inequality is satisfied for all objects:

$$d_{c_1,c_2} + \beta \le d_{i,c_1} + \beta + d_{i,c_2} + \beta. \quad (15)$$

A plausible value of $\Delta\beta$ in an iteration step is obtained as

$$\Delta\beta = \max\left\{\max_i\{d_{c_1,c_2} - d_{i,c_1} - d_{i,c_2} - \beta\}, 0\right\}. \quad (16)$$

Here, $\Delta\beta$ is positive when some D_{ci} are negative, while $\Delta\beta$ is zero when all D_{ci} are nonnegative. Then, β is monotonically increasing.

A sample procedure including the automated β-spread transformation can be summarized as follows:

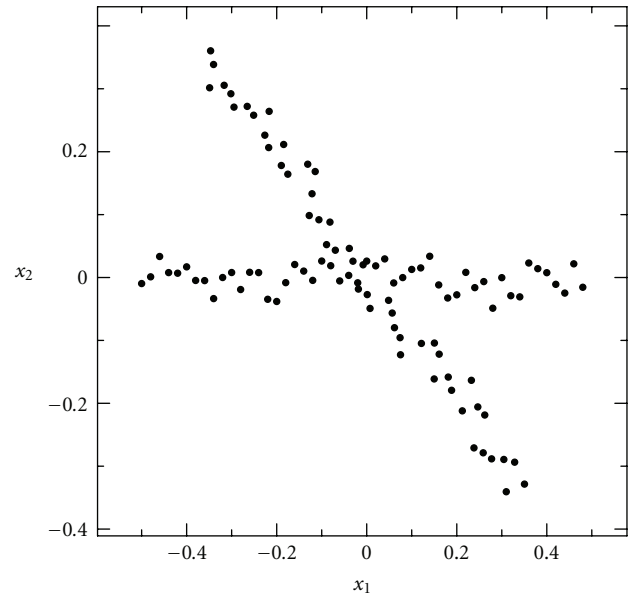


Figure 1: 2D plots of artificial data set.

Step 1. Set $\beta = 0$. Randomly initialize the prototypical medoids (two representative objects) of each cluster.

Step 2. Calculate the clustering criteria D_{ci} by (12).

Step 3. If there is at least one object that has $D_{ci} < 0$, update $\beta = \beta + \Delta\beta$ by (16).

Step 4. Update fuzzy memberships by (5).

Step 5. Search medoids in each cluster.

Step 6. Repeat Steps 2–5 until a certain stopping criterion is satisfied.

In Step 6, such a stopping criterion as $\min |u_{ci}^{(\mathrm{new})} - u_{ci}^{(\mathrm{old})}| < \varepsilon$ is used where ε is a small positive value.

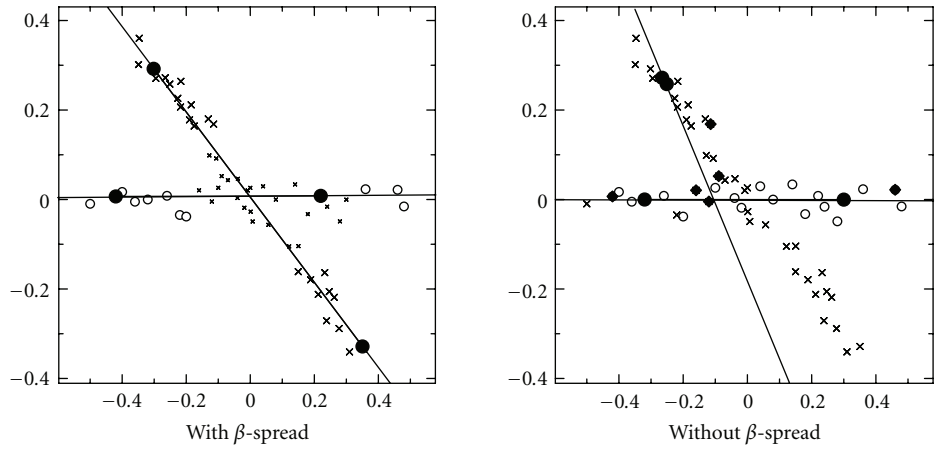
Although the proposed model is in the fuzzy clustering category, it is easily seen that a hard (non-fuzzy) version can be covered when $\theta = 1$, in which cluster memberships are given by the nearest prototype principle.

3.2. Missing Value Imputation by TIBA. Hathaway and Bezdek [14] demonstrated that the β-spread transformation is also useful for handling missing elements in relational data matrices. Although preimputation of missing elements may cause imputation errors and bring illegal effects in clustering process, β-spread transformation can decrease the illegal effects.

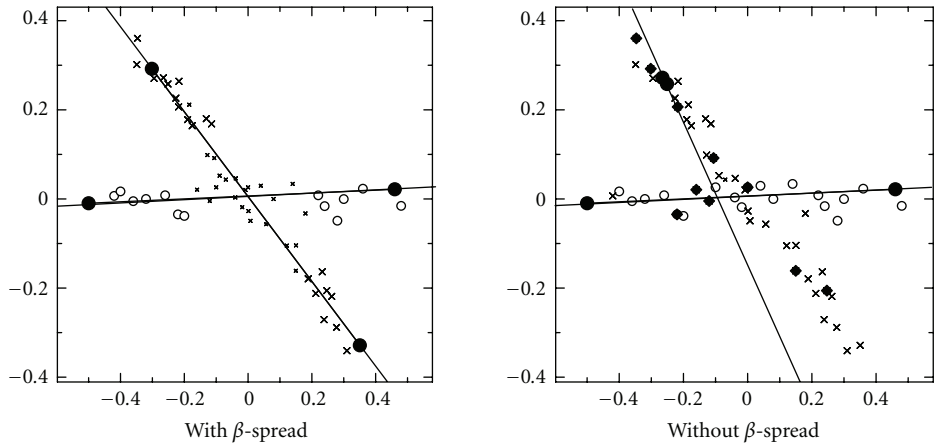
This paper considers the applicability of several imputation techniques in FCMdd-type linear clustering.

Hathaway and Bezdek [14] used three imputation techniques based on triangle inequality-based approximation (TIBA). The triangle inequality, which Euclidean relational data always satisfy, is represented as follows:

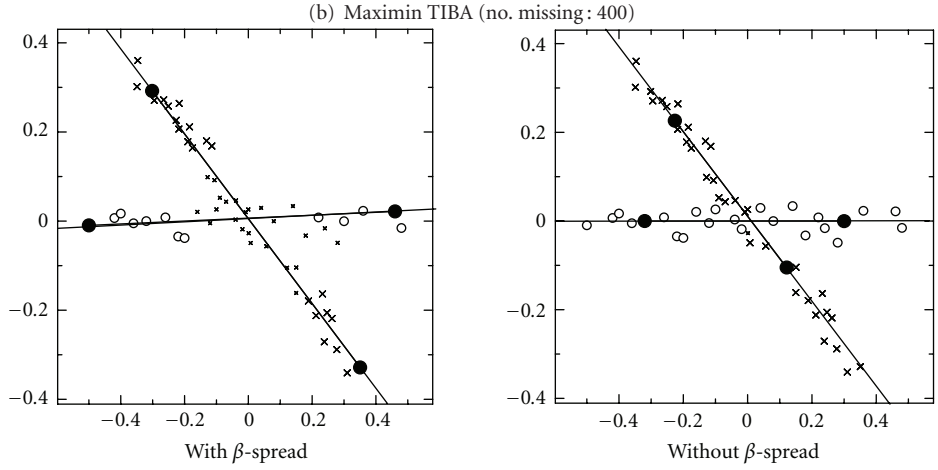
$$d_{ij} \le d_{ik} + d_{kj}. \quad (17)$$



(a) Minimax TIBA (no. missing : 500)



(b) Maximin TIBA (no. missing : 400)



(c) Average TIBA (no. missing : 500)

FIGURE 2: Comparison of cluster partitions from relational data imputed by three TIBAs, with (left)/without (right) β-spread transformation (Euclidean norm).

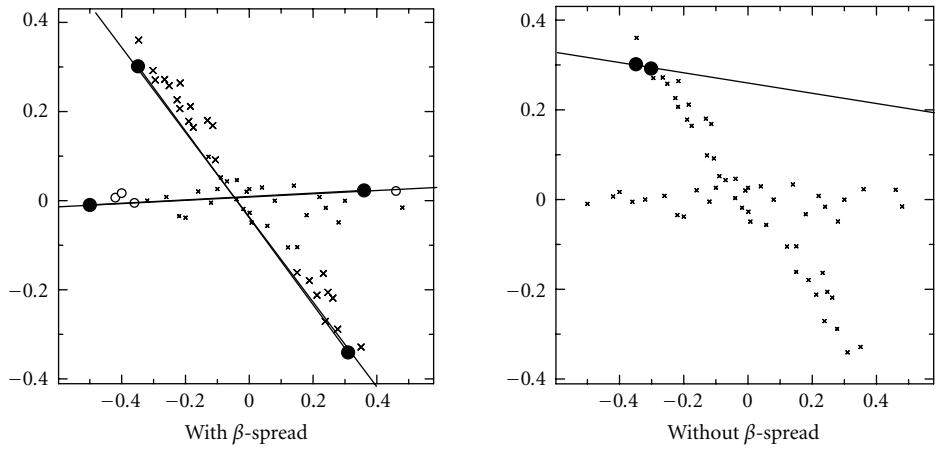
Assume that an element of relational matrix $\widetilde{d_{ij}}$ is missing and is to be preimputed before applying the clustering algorithm. Let K_{ij} be the corresponding index set as

$$K_{ij} = \left\{k \mid d_{ik} \quad \text{and} \quad d_{kj} \quad \text{observed}\right\}. \tag{18}$$

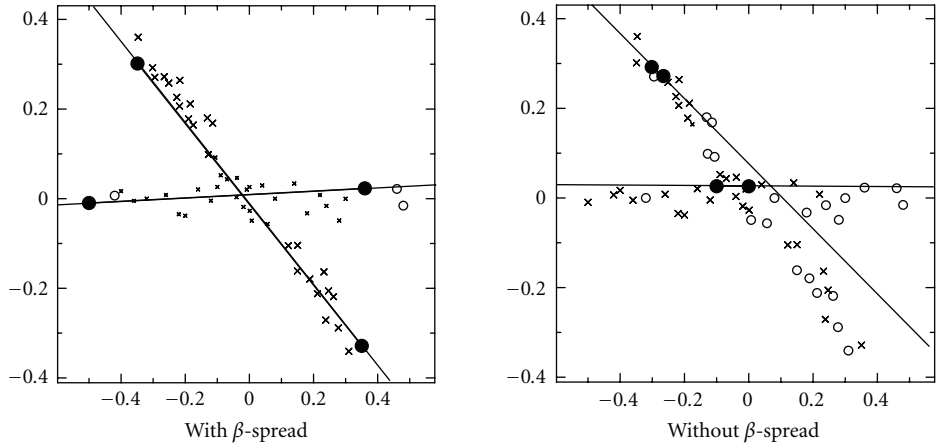
For each $k \in K_{ij}$, the triangle inequality (17) is given as the upper bound of $\widetilde{d_{ij}}$. Missing elements are replaced with the minimum upper bound of $\widetilde{d_{ij}}$:

$$\widetilde{d_{ij}} = \widetilde{d_{ji}} = \min_k \left\{d_{ik} + d_{kj}\right\}, \tag{19}$$

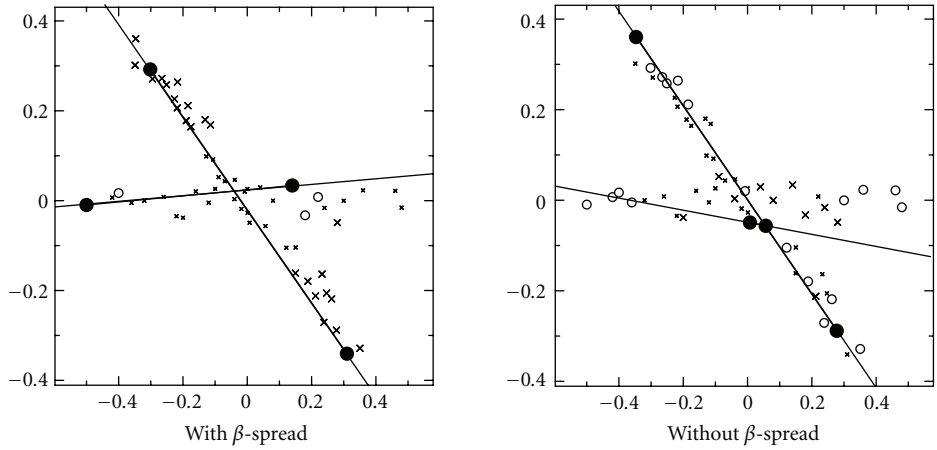
which is called minimax TIBA. By the way, $\widetilde{d_{ij}}$ is imputed by zero value if K_{ij} is empty.



(a) Minimax TIBA (no. missing : 1000)



(b) Maximin TIBA (no. missing : 700)



(c) Average TIBA (no. missing : 1100)

FIGURE 3: Comparison of cluster partition from relational data imputed by three TIBAs, with (left)/without (right) β-spread transformation (L_1 norm).

The triangle inequality is also represented as follows:

$$\begin{aligned} d_{ik} &\leq \widetilde{d_{ij}} + d_{jk}, \\ d_{jk} &\leq \widetilde{d_{ji}} + d_{ik}, \end{aligned} \tag{20}$$

and brings the following inequalities:

$$\begin{aligned} \widetilde{d_{ij}} &\geq d_{ik} - d_{jk}, \\ \widetilde{d_{ji}} &\geq d_{jk} - d_{ik}. \end{aligned} \tag{21}$$

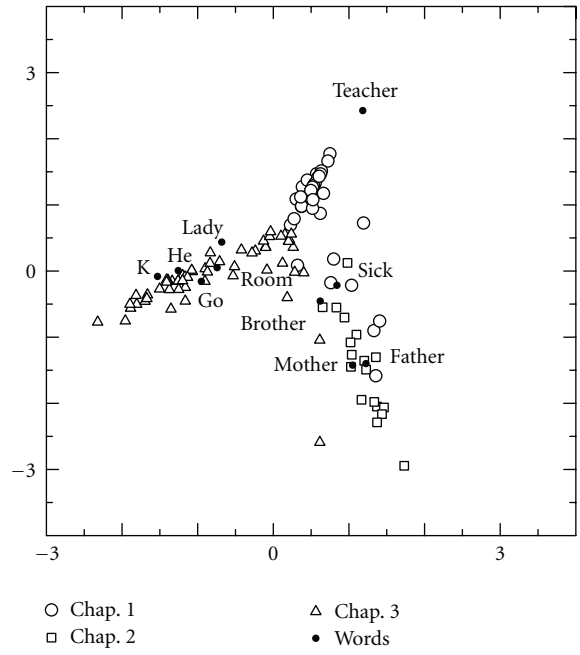


Figure 4: Document-keyword biplots [18].

So, the lower bound of $\widetilde{d}_{ij}$ is given as

$$\widetilde{d_{ij}} \geq \left| d_{ik} - d_{jk} \right|. \tag{22}$$

Missing elements are replaced with the maximum lower bound of $\widetilde{d}_{ij}$:

$$\widetilde{d_{ij}} = \widetilde{d_{ji}} = \max_k \left\{ \left| d_{ik} - d_{jk} \right| \right\}, \tag{23}$$

which is called maximin TIBA.

It is also possible to combine the previous two imputation values for predicting a reasonable estimation of missing values. The average values of minimax TIBA and maximin TIBA are used for imputing missing elements. This TIBA is called average TIBA.

These imputation techniques based on triangle inequalities can be easily applied to relational clustering problems. In the next section, these three imputation approaches are compared in FCMdd-type linear clustering tasks in conjunction with β-spread transformation.

4. Numerical Experiments

Two experimental results are shown in order to consider the applicability of the three TIBA imputation techniques in FCMdd-type linear clustering with β-spread transformation.

In previous researches, it has been shown that "soft" clustering models outperformed "hard" ones in local PCA tasks [15–17], and "fuzzy" models can be more useful than probabilistic ones [9]. Therefore, in this paper, the characteristics of the fuzzy version are investigated.

4.1. Artificial Data Set. An artificial relational data set composed of 60 patterns was generated from a 2D data set shown in Figure 1, in which patterns form two line-shaped clusters. It is obvious that the local linear structures cannot be extracted by the conventional point-prototype models such as FCM-like models and FCMdd. We made two relational data matrices. The first relational data matrix was generated by Euclidean norm, and the second one was generated by L_1 norm, which is non-Euclidean measure. The iterative algorithm was performed until the medoids became unchanged, and the model parameters were set as $(C, \theta) = (2, 2)$. In order to demonstrate the characteristics of the algorithm, the initial memberships were given in a supervised manner; that is, $(u_{1i}, u_{2i}) = (0.9, 0.1)$ for the first visual cluster and $(u_{1i}, u_{2i}) = (0.1, 0.9)$ for the second one.

In the previous research [12], it was demonstrated that the two linear substructures can be successfully revealed by the FCMdd-based linear clustering algorithm without β-spread transformation for Euclidean relational data while it can be done only with β-spread transformation for L_1 norm.

First, Euclidean incomplete relational data matrices were generated by removing a part of off-diagonal elements where K_{ij} was not empty. In order to protect tridiagonal parts of relational data, the maximum number of missing elements was set as $n^2 - 3n + 2$.

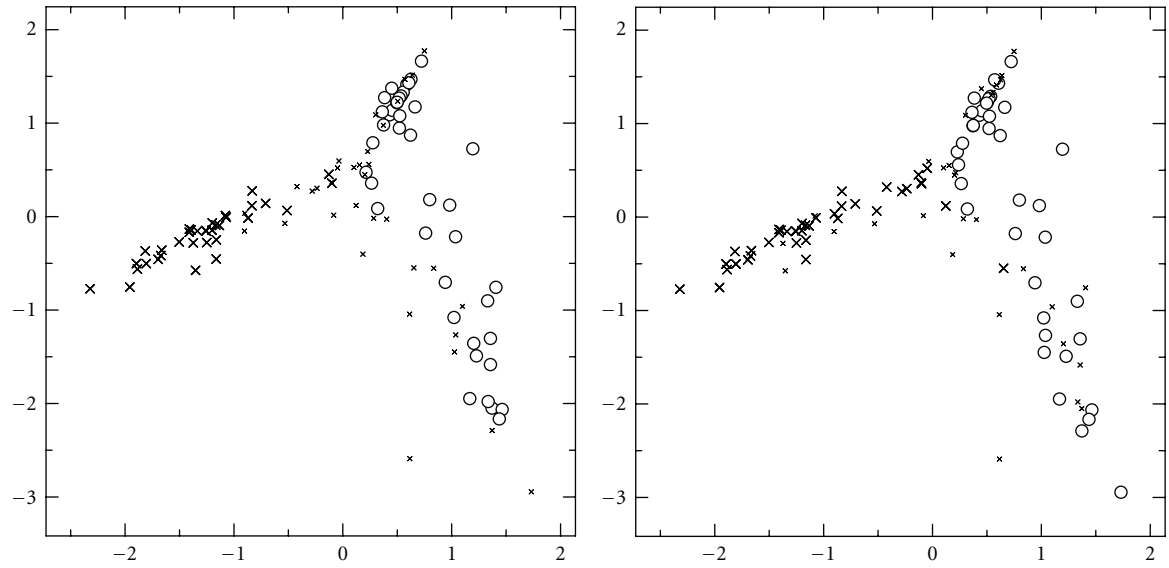
Clustering results are compared with those without β-spread transformation in Figure 2. Objects were partitioned into two clusters of circles and times, and smaller times mean that the patterns were shared almost equally by the two clusters. Medoids and prototypical lines are indicated by black circles and lines, respectively.

Each approximation method with β-spread transformation could estimate cluster medoids for capturing the two visual linear prototypes until the numbers of missing elements are less than about 30% although patterns having ambiguous memberships increased more than complete relational data. β-spread transformation performed on each approximation, minimax TIBA: $\beta = 0.04867$, maximin TIBA: $\beta = 0.044286$, average TIBA: $\beta = 0.051706$. Here, the maximum eigenvalues of PDP after imputation were, minimax TIBA: 0.053808, maximin TIBA: 0.1711612, average TIBA: 0.083694. So, the TIBA imputation brought a slightly non-Euclidean situation, and β-spread transformation successfully modified the data set.

On the other hand, without β-spread transformation, only average TIBA made it possible to extract linear substructures while minimax TIBA and maximin TIBA brought inappropriate results where some patterns depicted by black diamonds in Figure 2 had negative clustering criterion values.

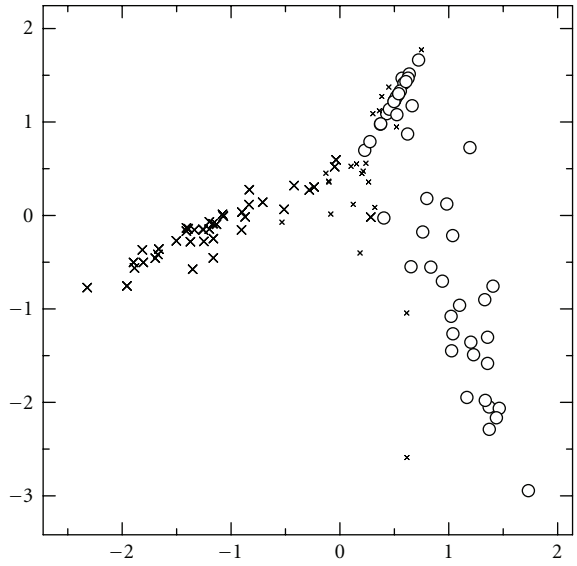
These results imply that the FCMdd-type linear clustering can successfully extract linear substructures of incomplete Euclidean relational data using β-spread transformation although the three imputation techniques cause non-Euclidean relational matrices.

Second, FCMdd-type linear fuzzy clustering was applied to non-Euclidean relational data.



(a) Minimax TIBA (no. missing : 6200)

(b) Maximin TIBA (no. missing : 8000)



(c) Average TIBA (no. missing : 7200)

FIGURE 5: Comparison of cluster partition from incomplete "Kokoro" text data derived with Jaccard's coefficient imputed by three TIBAs.

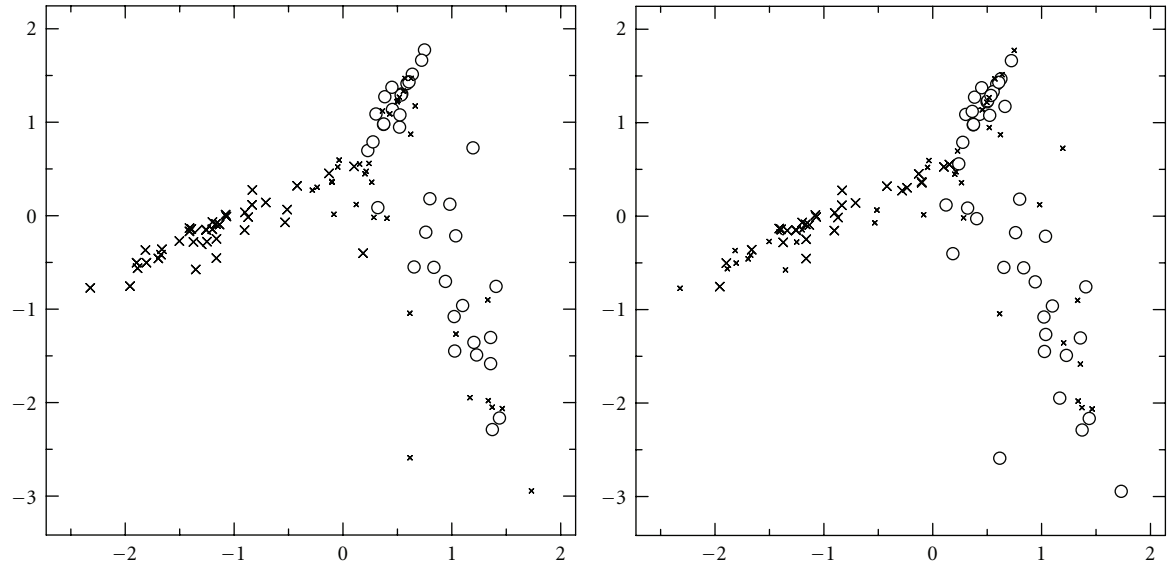
Incomplete relational data matrices were generated in the same manner with the Euclidean case. Clustering results are depicted in Figure 3.

With β-spread transformation, minimax TIBA and average TIBA could extract linear prototypes until the numbers of missing elements are less than about 60%, and maximin TIBA also could until about 40%. The parameters β in β-spread transformation were minimax TIBA: $\beta = 0.23344$, maximin TIBA: $\beta = 0.137577$, average TIBA: $\beta = 0.165504$. The derived β values are still smaller than the maximum eigenvalues of PDP, minimax TIBA: 0.655306, maximin TIBA: 0.427750, average TIBA: 0.619776.

Without β-spread transformation; however, all the three TIBAs brought inappropriate partitions because many patterns had negative clustering criterion values.

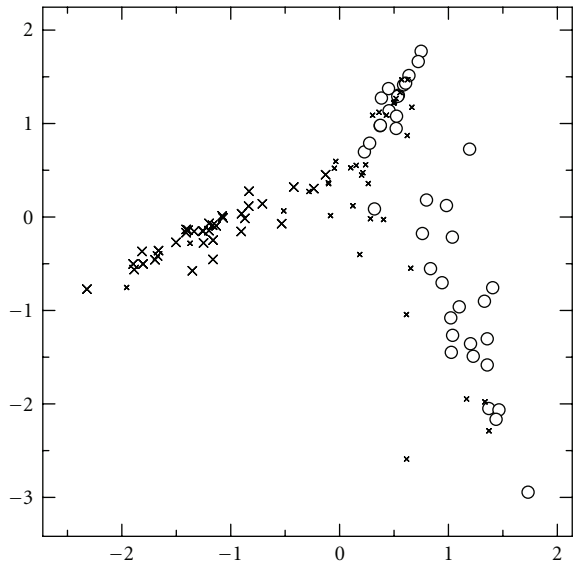
In this way, β-spread transformation also works well in incomplete situations.

4.2. Document Clustering. In the second experiment, TIBA imputation methods are compared in a document classification task. A relational data set was generated using a famous Japanese novel "Kokoro" by Soseki Natsume. The novel is composed of 3 chapters (Sensei and I, My Parents and I, Sensei and His Testament), and the chapters include 36, 18, 56 sections, respectively. The text data (Japanese language) can be downloaded from Aozora Bunko (http://www.aozora.gr.jp/). The sections were used as individual text documents ($n = 110$), which should be partitioned without the chapter information. The text documents



(a) Minimax TIBA (no. missing : 5600)

(b) Maximin TIBA (no. missing : 7400)



(c) Average TIBA (no. missing : 6600)

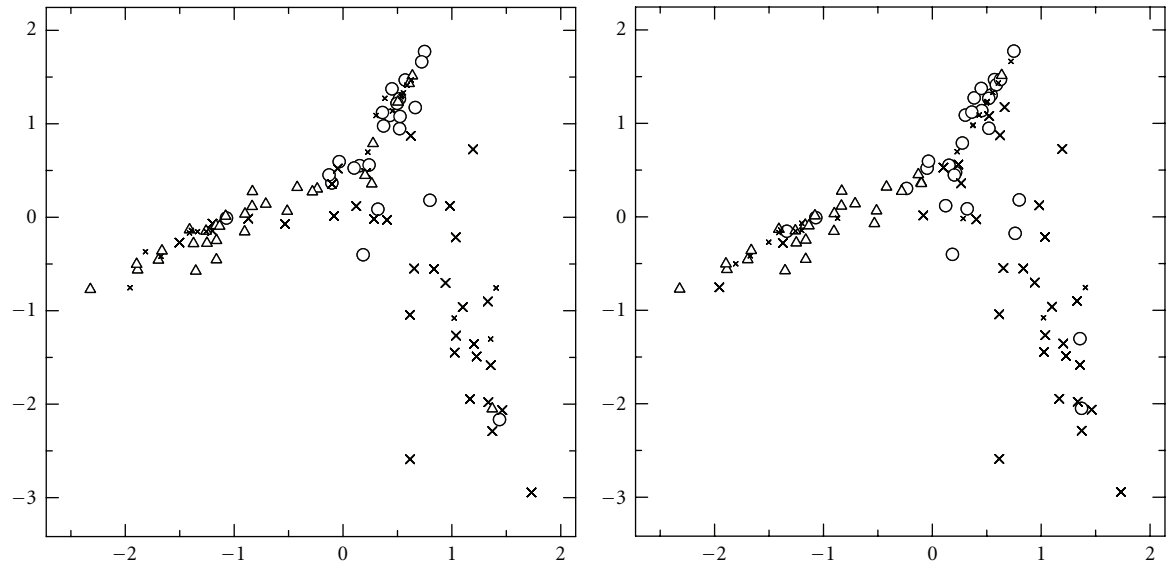
Figure 6: Comparison of cluster partition from incomplete "Kokoro" text data derived with Dice's coefficient imputed by three TIBAs.

were preprocessed using "Chasen" morphological analysis system software (http://chasen.naist.jp/hiki/ChaSen/), which segments the Japanese text string into morphemes. Wada et al. [18] performed a PCA-based structural analysis with the 83 most frequently used substantives and verbs with their tf-idf weights and revealed that the chapter structure can be emphasized by using 10 meaningful keywords as is shown in Figure 4, which is 2D biplots of principal components. Chapters 2 and 3 form two linear clusters in 10D data space, and chapter 1 exists on their intersection. In this experiment, parameters were set as $C = 2$, $\theta = 2.0$ with the goal of revealing the two linear substructures.

Two relational data matrices were generated considering co-occurrence information of the 10 keywords. Jaccard coefficient and Dice coefficient are the similarity measures for asymmetric information on binary variables [19]. Assume that the cooccurrence information of keywords among two text documents are summarized in a 2×2 contingency table as shown in Table 1 where "1" means occurrence of the keyword.

Table 1: 2×2 contingency table for text documents.

	keyword B		
keyword A	1	0	Total
1	a	b	a + b
0	c	d	c + d
Total	a + c	b + d	



(a) Jaccard (no. missing : 8000) (b) Dice (no. missing : 7400)

Figure 7: Comparison of cluster partition of Fuzzy c-Medoids for incomplete "Kokoro" text data derived with Jaccard's and Dice's coefficients imputed by maximin TIBA.

Jaccard's coefficient is the similarity represented as

$$s_{ij} = \frac{a}{a+b+c}. \tag{24}$$

Dice's coefficient is also the similarity represented as

$$s_{ij} = \frac{2a}{2a+b+c}. \tag{25}$$

Because the linear clustering model uses distance (dissimilarity) measures, the similarity measures s_{ij} were transformed into dissimilarity ones d_{ij}.

$$d_{ij} = \max_{k,l}\{s_{kl}\} - s_{ij}. \tag{26}$$

Before applying the FCMdd-based linear fuzzy clustering, randomly selected elements were withheld from the relational matrix with 11,772 elements and were imputed by the three TIBA methods. Then, the cluster partitions for Jaccard's index were derived as shown in Figure 5. Two clusters are depicted by circles and times, and small times mean ambiguous assignment. Documents were properly partitioned into two clusters considering linear substructures.

Minimax TIBA allowed with 50% missing values or fewer. Average TIBA tolerated 60% missing values or fewer. Maximin TIBA resulted in a good partition with 68% missing values or fewer. The parameters β in β-spread transformation were given as minimax TIBA: $\beta = 0.960$, maximin TIBA: $\beta = 0.411$, average TIBA: $\beta = 0.435$. The derived β values are still smaller than the maximum eigenvalues of PDP without missing elements, minimax TIBA: 5.662, maximin TIBA: 4.548, average TIBA: 2.427.

Clustering results for Dice coefficient are depicted in Figure 6. Our approach also extracted linear substructure from incomplete relational data of Dice coefficient. Minimax TIBA allowed with 48% missing values or fewer. Average TIBA tolerated 55% missing values or fewer. Maximin TIBA resulted in a good partition with 63% missing values or fewer. The parameters β in β-spread transformation were given as, minimax TIBA: $\beta = 0.760$, maximin TIBA: $\beta = 0.383$, average TIBA: $\beta = 0.608$. The derived β values are still smaller than the maximum eigenvalues of PDP, minimax TIBA: 6.169, maximin TIBA: 4.288, average TIBA: 2.219.

In the experiments, it was demonstrated that the TIBA imputation methods work well for incomplete non-Euclidean relational data in conjunction with β-spread transformation.

Finally, comparison with other methods is discussed. Although we have already many clustering algorithms, some of which are used in document clustering tasks [20], most of them are designed for finding groups composed of similar pattern from the view point of "point prototype" or "hierarchical aggregation". For example, Fuzzy c-Medoids (FCMdd) [5], which is a representative method of point-prototype models, can be applied to the relational data set of this subsection. Figure 7 shows the clustering results of finding three chapter structures of circles, times, and triangles. Small times mean ambiguous assignment as well. The conventional clustering methods are useful for finding such document groups considering mutual similarity among documents (or sometime keyword groups).

On the other hand, the proposed method is designed for a different purpose of finding "local linear structures" from the view point of local PCA, which is useful for cluster-wise information summarization such as local feature map construction. In this sense, the proposed method has different future application area from the conventional clustering tools.

5. Conclusion

This paper compared the applicability of TIBA imputation methods and β-spread transformation for handling incomplete relational data in FCMdd-type linear clustering. In numerical experiments, three imputation techniques of minimax TIBA, maximin TIBA, and average TIBA were compared using two data sets. The experimental results indicated that β-spread transformation still works well for incomplete data in conjunction with β-spread transformation. All the three TIBAs are useful for imputing incomplete non-Euclidean relational data.

From the view point of local PCA concept, the proposed method can be used for local information summarization or local feature map construction where data structures are visually summarized in low-dimensional space in conjunction with data clustering. The application is remained in future works. Another potential future work is an extension to the case of multidimensional prototype models, which is useful for constructing 2D feature map.

Acknowledgment

This work was supported in part by the Ministry of Education, Culture, Sports, Science and Technology, Japan, under Grant-in-Aid for Scientific Research (23500283).

References

[1] R. J. Hathaway, J. W. Davenport, and J. C. Bezdek, "Relational duals of the *c*-means clustering algorithms," *Pattern Recognition*, vol. 22, no. 2, pp. 205–212, 1989.

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

[3] J. B. MacQueen, "Some methods of classification and analysis of multivariate observations," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, University of California Press, Berkeley, Calif, USA, 1967.

[4] L. Kaufman and P. J. Rousseeuw, *Finding Groups In Data: An Introduction To Cluster Analysis*, Wiley-Interscience, 1990.

[5] R. Krishnapuram, A. Joshi, O. Nasraoui, and L. Yi, "Low-complexity fuzzy relational clustering algorithms for web mining," *IEEE Transactions on Fuzzy Systems*, vol. 9, no. 4, pp. 595–607, 2001.

[6] J. C. Bezdek, C. Coray, R. Gunderson, and J. Watson, "Detection and characterization of cluster substructure 1. Linear structure fuzzy *c*-lines," *SIAM Journal on Applied Mathematics*, vol. 40, no. 2, pp. 339–357, 1981.

[7] 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 Journal on Applied Mathematics*, vol. 40, no. 2, pp. 358–372, 1981.

[8] Y. Yabuuchi and J. Watada, "Fuzzy principal component analysis and its application," *Biomedical Fuzzy and Human Sciences*, vol. 3, pp. 83–92, 1997.

[9] K. Honda and H. Ichihashi, "Regularized linear fuzzy clustering and probabilistic PCA mixture models," *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 4, pp. 508–516, 2005.

[10] N. Haga, K. Honda, A. Notsu, and H. Ichihashi, "Local sub-space learning by extended fuzzy *c*-medoids clustering," *International Journal of Knowledge Engineering and Soft Data Paradigms*, vol. 2, no. 2, pp. 169–181, 2010.

[11] T. Yamamoto, K. Honda, A. Notsu, and H. Ichihashi, "An FCMdd based linear clustering model for non-Euclidean relational data," in *Proceedings of the 5th International Conference on Soft Computing and Intelligent Systems*, pp. 243–247, 2010, 11th International Symposium on Advanced Intelligent Systems.

[12] T. Yamamoto, K. Honda, A. Notsu, and H. Ichihashi, "Non-Euclidean extension of FCMdd-based linear clustering for relational data," *Journal of Advanced Computational Intelligence and Intelligent Informatics*. In press.

[13] R. J. Hathaway and J. C. Bezdek, "Nerf *c*-means: non-Euclidean relational fuzzy clustering," *Pattern Recognition*, vol. 27, no. 3, pp. 429–437, 1994.

[14] R. J. Hathaway and J. C. Bezdek, "Clustering incomplete relational data using the non-Euclidean relational fuzzy *c*-means algorithm," *Pattern Recognition Letters*, vol. 23, no. 1–3, pp. 151–160, 1994.

[15] N. Kambhatla and T. K. Leen, "Dimension reduction by local principal component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1493–1516, 1997.

[16] G. E. Hinton, P. Dayan, and M. Revow, "Modeling the manifolds of images of handwritten digits," *IEEE Transactions on Neural Networks*, vol. 8, no. 1, pp. 65–74, 1997.

[17] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

[18] H. Wada, K. Honda, A. Notsu, and H. Ichihashi, "Document map construction and keyword selection based on local PCA," in *Proceedings of the 4th International Conference on Soft Computing and Intelligent Systems*, pp. 682–685, 2008, 9th International Symposium on Advanced Intelligent Systems.

[19] M. R. Anderberg, *Cluster Analysis for Applications*, Academic Press, 1973.

[20] A. Hotho, A. Nürnberger, and G. Paaß, "A brief survey of text mining," *Journal for Computational Linguistics and Language Technology*, vol. 20, pp. 19–62, 2005.