

On Hard and Fuzzy c -Means Clustering with Conditionally Positive Definite Kernel

Yuchi Kanzawa
Shibaura Institute of Technology
Tokyo, Japan 135-8548
Email: kanzawa@sic.shibaura-it.ac.jp

Yasunori Endo
University of Tsukuba
Ibaraki, Japan

Sadaaki Miyamoto
University of Tsukuba
Ibaraki, Japan

Abstract—In this paper, we investigate three types of c -means clustering algorithms with a conditionally positive definite kernel. One is based on hard c -means, and the others are based on standard and entropy-regularized fuzzy c -means. First, based on a conditionally positive definite kernel describing a squared Euclidean distance between data in the feature space, these algorithms are derived from revised optimization problems of the conventional kernel c -means. Next, based on the relationship between the positive definite kernel and conditionally positive definite kernel, the revised dissimilarity between a datum and a cluster center in the feature space is shown. Finally, it is shown that a conditionally positive definite kernel c -means algorithm and a kernel c -means algorithm with a positive definite kernel derived from the conditionally positive definite kernel are essentially identical to each other. An explicit mapping for a conditionally positive definite kernel is also described geometrically.

Index Terms—Clustering, Fuzzy c -means, Conditionally positive definite kernel

I. INTRODUCTION

Fuzzy c -means (FCM) clustering [1] is a well-known fuzzy clustering technique, and many FCM variants have been proposed. One of the authors has proposed an FCM algorithm based on the concept of entropy regularization [2]; this algorithm is known as the entropy-regularized FCM (eFCM) algorithm. It is a useful algorithm that is mathematically related to other techniques.

Three algorithms [3], [4] have been proposed to cluster data with nonlinear borders via nonlinear transformation from the original pattern space into a higher dimensional feature space with support vector machine (SVM) kernel functions [5]. They are the K-HCM algorithm derived from the hard c -means (HCM) algorithm, the K-sFCM algorithm derived from the standard FCM (sFCM) algorithm, and the K-eFCM algorithm derived from the eFCM algorithm. These three algorithms are known as kernel c -means (K-CM) algorithms. For kernel data analysis, an explicit mapping is unknown, whereas the inner product is known. An explicit mapping has been introduced by one of the authors, and it is used to describe the situation of kernel fuzzy c -means in higher dimensional space via kernel principal component analysis [6], [7].

In kernel data analysis, the kernel must be positive definite so that it describes the inner product between the data in the feature space. On the other hand, there exists a class of conditionally positive definite (cpd) kernel [8]. In this paper, we investigate c -means clustering algorithms with cpd kernel.

First, we develop three c -means clustering algorithms with cpd kernel by solving the revised optimization problem of K-HCM, K-sFCM, and K-eFCM. Accordingly, the cpd kernel is the squared Euclidean distance between the data in the feature space. We call these three algorithms cpdK-HCM, cdK-sFCM, and cpdK-eFCM, respectively; and sum up to cpdK-CM. Next, we show how the cpd kernel is used in K-CM algorithms, with the positive definite kernel obtained by arranging the cpd kernel. Furthermore, we show the connection between the cpdK-CM algorithms and the K-CM algorithms with the positive definite kernel obtained from the cpd kernel, and we interpret a mapping from the original data space to the feature space, which is latent in the cpdK-CM algorithms. Finally, we show a numerical example of the cpdK-CM algorithms.

II. PRELIMINARIES

In this section, we introduce three K-CM clustering algorithms with an explicit mapping from the original data space to the feature space [6],[7]

The data set $X = \{x_i \mid x_i \in \mathbb{R}^p, i \in \{1, \dots, N\}\}$ is given. The membership by which x_i belongs to the j -th cluster is denoted by $u_{i,j}$ ($i \in \{1, \dots, N\}, j \in \{1, \dots, C\}$), and the set of $u_{i,j}$, denoted by $u \in \mathbb{R}^{N \times C}$, is known as the partition matrix. The constraint on u is

$$\sum_{j=1}^C u_{i,j} = 1 \quad (0 \leq u_{i,j} \leq 1).$$

A high-dimensional feature space used in SVM is denoted by $\mathbb{H}$, whereas the original space, which is known as the data space, $\mathbb{R}^p$. $\mathbb{H}$ may be infinite dimensional. Let the inner product be denoted by $\langle \cdot, \cdot \rangle$. The norm of $\mathbb{H}$ for an element $g \in \mathbb{H}$ is given by

$$\|g\|_{\mathbb{H}} = \sqrt{\langle g, g \rangle}. \quad (1)$$

A transformation $\Phi : \mathbb{R}^p \rightarrow \mathbb{H}$ is employed, whereby x_i is mapped into $\Phi(x_i)$. In general, Φ is not usable; however, the inner product $\langle \Phi(x), \Phi(y) \rangle$ can be expressed by a kernel function

$$\mathcal{K}(x, y) = \langle \Phi(x), \Phi(y) \rangle. \quad (2)$$

The matrix whose (i, j) -element is $\mathcal{K}(x_i, x_j)$ is denoted by K . The cluster center set in $\mathbb{H}$ is denoted by $W = \{W_1, \dots, W_C\}$.

An explicit function Φ corresponding to an introduced kernel function has been proposed as

$$\Phi(x_i) = e_i \quad (3)$$

with

$$\langle e_i, e_j \rangle = e_i^\top K e_j, \quad (4)$$

where $\mathbb{H} = \mathbb{R}^N$ and e_i is the i -th unit vector whose i -th component is unity and remaining components are zero. Note that $\Phi : X \rightarrow \mathbb{R}^N$, i.e., Φ is not defined on $\mathbb{R}^p$ but on finite data X .

Three K-CM algorithms with explicit mapping are obtained by solving the following optimization problem:

$$\underset{u, W}{\text{minimize}} \sum_{i=1}^N \sum_{j=1}^C u_{i,j} \|\Phi(x_i) - W_j\|_{\mathbb{H}}^2, \quad (5)$$

$$\text{subject to } \sum_{j=1}^C u_{i,j} = 1, \quad u_{i,j} \in \{0, 1\}, \quad (6)$$

for K-HCM,

$$\underset{u, W}{\text{minimize}} \sum_{i=1}^N \sum_{j=1}^C u_{i,j}^m \|\Phi(x_i) - W_j\|_{\mathbb{H}}^2, \quad (7)$$

$$\text{subject to } \sum_{j=1}^C u_{i,j} = 1, \quad (8)$$

for K-sFCM, and

$$\underset{u, W}{\text{minimize}} \sum_{i=1}^N \sum_{j=1}^C u_{i,j} \|\Phi(x_i) - W_j\|_{\mathbb{H}}^2 + \lambda^{-1} \sum_{i=1}^N \sum_{j=1}^C u_{i,j} \log(u_{i,j}), \quad (9)$$

$$\text{subject to } \sum_{j=1}^C u_{i,j} = 1. \quad (10)$$

for K-eFCM. These algorithms are described as follows:

Algorithm 1 (K-CM): [6] [7]

STEP 1. Set $m > 1$ for K-sFCM. Set $\lambda > 0$ for K-eFCM. Set the kernel matrix K and initial partition matrix.

STEP 2. Update the cluster centers W :

$$W_j = \left(\sum_{i=1}^N u_{i,j} \right)^{-1} (u_{1,j}, \dots, u_{N,j})^\top, \quad (11)$$

for K-HCM and K-sFCM, and

$$W_j = \left(\sum_{i=1}^N u_{i,j}^m \right)^{-1} (u_{1,j}^m, \dots, u_{N,j}^m)^\top, \quad (12)$$

for K-eFCM.

STEP 3. Update the squared Euclidean distance between the cluster center W_j and data $\Phi(x_i)$ in the

feature space:

$$d_{i,j} = (e_i - W_j)^\top K (e_i - W_j). \quad (13)$$

STEP 4. Update the membership $u_{i,j}$:

$$u_{i,j} = \begin{cases} 1 & (i = \arg \min_k d_{k,j}), \\ 0 & (\text{otherwise}), \end{cases} \quad (14)$$

for K-HCM,

$$u_{i,j} = 1 / \sum_{k=1}^C \left(\frac{d_{i,j}}{d_{i,k}} \right)^{1/(m-1)}, \quad (15)$$

for K-sFCM, and

$$u_{i,j} = \frac{\exp(-\lambda d_{i,j})}{\sum_{k=1}^C \exp(-\lambda d_{i,k})}, \quad (16)$$

for K-eFCM.

STEP 4. If (u, W, d) converge, terminate this algorithm. Otherwise, go to STEP 2.

III. CONDITIONALLY POSITIVE DEFINITE KERNEL c -MEANS

In this section, we investigate K-CM clustering algorithms with a conditionally positive definite kernel. First, we derive conditionally positive definite K-CM algorithms; thus, the conditionally positive definite kernel is the squared Euclidean distance in a feature space. Next, we show how the cpd kernel is used in a K-CM algorithm with a positive definite kernel obtained by arranging the cpd kernel. Furthermore, we determine the relation between the cpdK-CM and K-CM algorithms with a positive definite kernel obtained from a cpd kernel, and we interpret a mapping from the original data space to the feature space, which is latent in cpdK-CM algorithms. Finally, we show a numerical example of the cpdK-CM algorithms.

A. c -Means algorithms with cpd kernel

In this subsection, we develop c -means clustering algorithms with a cpd kernel by solving the revised optimization problems of K-HCM, K-sFCM, and K-eFCM; thus, the cpd kernel is the squared Euclidean distance between the data in the feature space.

The cpd kernel is defined as follows:

Definition 1 (cpd matrix): [8]

A symmetric $N \times N$ matrix K ($N \geq 2$) taking values in $\mathbb{R}$ and satisfying

$$\sum_{i=1}^N \sum_{k=1}^N \zeta_i \zeta_k K_{i,k} \geq 0 \quad (17)$$

for all $\zeta_i \in \mathbb{R}$, with

$$\sum_{i=1}^N \zeta_i = 0, \quad (18)$$

is said to be conditionally positive definite (cpd).

Definition 2 (cpd kernel): [8]

Let $\mathcal{X}$ be a nonempty set. A function $\tilde{\mathcal{K}} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, which yields a conditionally positive definite Gram matrix for all $N \geq 2$, $x_1, \dots, x_N \in \mathcal{X}$, is known as a conditionally positive definite (cpd) kernel.

A cpd kernel describes the squared Euclidean distance in a feature space as follows:

Theorem 1 (Hilbert space representation of cpd kernels): [8]

Let $\tilde{\mathcal{K}}$ be a real-valued conditionally positive definite kernel on $\mathcal{X}$, satisfying $\tilde{\mathcal{K}}(x, x) = 0$ for all $x \in \mathcal{X}$. Then, there exists a Hilbert space $\mathbb{H}$ of real-valued functions on $\mathcal{X}$ and a mapping $\Phi : \mathcal{X} \rightarrow \mathbb{H}$ such that

$$\|\Phi(x) - \Phi(x')\|_{\mathbb{H}}^2 = -\tilde{\mathcal{K}}(x, x'). \quad (19)$$

If we drop the assumption $\tilde{\mathcal{K}}(x, x) = 0$, the Hilbert space representation reads

$$\|\Phi(x) - \Phi(x')\|_{\mathbb{H}}^2 = -\tilde{\mathcal{K}}(x, x') + \frac{1}{2}(\tilde{\mathcal{K}}(x, x) + \tilde{\mathcal{K}}(x', x')). \quad (20)$$

We cannot utilize the squared Euclidean distance between the data in the feature space for K-CM (Algorithm 1) because the optimization problems (5)–(10) include the squared Euclidean distance between the data and the cluster center in the feature space. Therefore, we reformulate the K-CM optimization problems (5)–(10) into the following problems:

$$\underset{u}{\text{minimize}} \sum_{j=1}^C \frac{\sum_{i=1}^N \sum_{k=1}^N u_{i,j} \|\Phi(x_i) - \Phi(x_k)\|_{\mathbb{H}}^2}{\left(\sum_{t=1}^N u_{t,j}\right)^2}, \quad (21)$$

$$\text{subject to } \sum_{j=1}^C u_{i,j} = 1, \quad u_{i,j} \in \{0, 1\}, \quad (22)$$

for K-HCM,

$$\underset{u}{\text{minimize}} \sum_{j=1}^C \frac{\sum_{i=1}^N \sum_{k=1}^N u_{i,j}^m \|\Phi(x_i) - \Phi(x_k)\|_{\mathbb{H}}^2}{\left(\sum_{t=1}^N u_{t,j}\right)^2}, \quad (23)$$

$$\text{subject to } \sum_{j=1}^C u_{i,j} = 1, \quad (24)$$

for K-sFCM, and

$$\underset{u}{\text{minimize}} \sum_{j=1}^C \frac{\sum_{i=1}^N \sum_{k=1}^N u_{i,j} \|\Phi(x_i) - \Phi(x_k)\|_{\mathbb{H}}^2}{\left(\sum_{t=1}^N u_{t,j}\right)^2} + \lambda^{-1} \sum_{i=1}^N \sum_{j=1}^C u_{i,j} \log(u_{i,j}), \quad (25)$$

$$\text{subject to } \sum_{j=1}^C u_{i,j} = 1. \quad (26)$$

for K-eFCM. The form $\|\Phi(x_i) - \Phi(x_k)\|_{\mathbb{H}}^2$ in the above objective functions can be translated into $-K_{i,k}$ if $\tilde{K}_{i,i} = 0$, and $-\tilde{K}_{i,k} + 1/2(\tilde{K}_{i,i} + \tilde{K}_{k,k})$ otherwise, with the cpd kernel matrix $\tilde{K}$. Such optimization problems are identical to those for RHCM and RFCM [9] if the corresponding relational data matrix is $\tilde{K}$, and they are solved using the following algorithms:

Algorithm 2 (cpdK- $\{H, sF, eF\}$ CM):

STEP 1. Set the fuzzifier parameter m and λ for cpdK-sFCM and cpdK-eFCM, respectively. Set the cpd kernel matrix $\tilde{K}$ and the initial membership $u_{i,j}$. Set the N -dimensional squared matrix R as

$$R_{i,k} = -\tilde{K}_{i,k} + 1/2(\tilde{K}_{i,i} + \tilde{K}_{k,k}) \quad (27)$$

STEP 2. Update the intermediate variable $w_j \in \mathbb{R}^N$ as

$$w_j = \left(\sum_{i=1}^N u_{i,j}\right)^{-1} (u_{1,j}, \dots, u_{N,j})^T, \quad (28)$$

for cpdK-HCM,

$$w_j = \left(\sum_{i=1}^N u_{i,j}^m\right)^{-1} (u_{1,j}^m, \dots, u_{N,j}^m)^T, \quad (29)$$

for cpdK-sFCM, and

$$w_j = \left(\sum_{i=1}^N u_{i,j}\right)^{-1} (u_{1,j}, \dots, u_{N,j})^T. \quad (30)$$

for cpdK-eFCM.

STEP 3. Update the intermediate variable $\tilde{d}_{i,j}$ as

$$\tilde{d}_{i,j} = (Rw_j)_i - \frac{1}{2}w_j^T R w_j. \quad (31)$$

STEP 4. Update the membership $u_{i,j}$ as

$$u_{i,j} = \begin{cases} 1 & (i = \arg \min_k \tilde{d}_{k,j}), \\ 0 & (\text{otherwise}), \end{cases} \quad (32)$$

for cpdK-HCM,

$$u_{i,j} = 1 / \sum_{k=1}^C \left(\frac{\tilde{d}_{i,j}}{\tilde{d}_{i,k}} \right)^{1/(m-1)}, \quad (33)$$

for cpdK-sFCM, and

$$u_{i,j} = \frac{\exp(-\lambda \tilde{d}_{i,j})}{\sum_{k=1}^C \exp(-\lambda \tilde{d}_{i,k})}, \quad (34)$$

for cpdK-eFCM.

STEP 5. If a stopping criterion is satisfied, terminate this algorithm. Otherwise, go to STEP 2.

B. *c-Means with pd kernel obtained from cpd kernel*

In this subsection, we show how the cpd kernel is used in K-CM algorithms with a positive definite kernel obtained by arranging the cpd kernel. First, we show the relation between the cpd kernel and the pd kernel. Next, for K-CM (Algorithm 1) with a pd kernel obtained by reformulating the cpd kernel, we show how the cpd kernel is used to calculate the squared Euclidean distance between the data and the cluster center in the feature space.

The cpd kernel and the pd kernel have the following relation:

Theorem 2 (Connection cpd — pd): [8]

Let K be an N -dimensional symmetric matrix; $\mathbf{1} \in \mathbb{R}^N$, an N -dimensional real-valued vector whose elements are all 1; $E \in \mathbb{R}^{N \times N}$, the identity matrix; and $\eta \in \mathbb{R}^N$, a vector satisfying

$$\mathbf{1}^\top \eta = 1. \quad (35)$$

Then,

$$K = (E - \mathbf{1}\eta^\top) \tilde{K} (E - \eta \mathbf{1}^\top) \quad (36)$$

is positive definite if and only if $\tilde{K}$ is conditionally positive definite.

Given the cpd kernel $\tilde{K}$ for a data set X , we can obtain the pd kernel

$$K = (E - \mathbf{1}\mathbf{1}^\top/N) \tilde{K} (E - \mathbf{1}\mathbf{1}^\top/N) \quad (37)$$

from Theorem 2 and use it K-CM (Algorithm 1). Then, STEP 3 of Algorithm 1 is rewritten as

STEP 3. Update the intermediate variable $\tilde{d}_{i,j}$ as

$$d_{i,j} = (e_i - W_j)^\top \tilde{K} (e_i - W_j) \quad (38)$$

from that

$$(E - \eta \mathbf{1}^\top)(e_i - W_j) = (e_i - W_j). \quad (39)$$

C. Discussion

In this subsection, we show the relation between the K-CM algorithm with a pd kernel obtained from the cpd kernel and the cpdK-CM algorithm, and we geometrically interpret the explicit mapping from the data space to the feature space, hiding in cpdK-CM.

First, the updating equations (28) and (29) of the intermediate variable w_j in cpdK-CM (Algorithm 2) and the updating equations (11) and (12) of the cluster center W_j in K-CM (Algorithm 1) are identical. Next, the updating equations (32), (33), and (34) of the membership $u_{i,j}$ in cpdK-CM (Algorithm 2) and the updating equations (14), (15), and

(16) of the membership $u_{i,j}$ in K-CM (Algorithm 1) are also identical. Furthermore, the updating equation of $d_{i,j}$ in K-CM (Algorithm 1) with a pd kernel obtained by reformulating the cpd kernel is expressed as Eq. (38). On the other hand, the updating equation of $\tilde{d}_{i,j}$ in cpdK-CM (Algorithm 2) can be expressed as

$$\begin{aligned} \tilde{d}_{i,j} &= (Rw_j)_i - \frac{1}{2} w_j^\top R w_j \\ &= \frac{1}{2} (e_i - w_j)^\top \tilde{K} (e_i - w_j) \end{aligned} \quad (40)$$

from $R = -\tilde{K}$, with $\tilde{K}_{i,i} = 0$. Therefore, $\tilde{d}_{i,j}$ in cpdK-CM (Algorithm 2) and $d_{i,j}$ in K-CM (Algorithm 1) with a pd kernel obtained by reformulating the cpd kernel have the following relation:

$$\tilde{d}_{i,j} = \frac{1}{2} d_{i,j}. \quad (41)$$

Comparing these relations with that between the pd kernel and the cpd kernel, cpdK-CM (Algorithm 2) for the cpd kernel $\tilde{K}$ with $\tilde{K}_{i,i} = 0$ and K-CM (Algorithm 1) for the pd kernel K with

$$K = 2(E - \mathbf{1}\eta^\top) \tilde{K} (E - \eta \mathbf{1}^\top) \quad (42)$$

are identical.

Next, we geometrically interpret the explicit mapping from the data space to the feature space, hiding in cpdK-CM. Since cpdK-CM (Algorithm 2) and K-CM (Algorithm 1) with a pd kernel obtained from the cpd kernel are identical, we interpret cpdK-CM (Algorithm 2) from the geometrical interpretation of K-CM (Algorithm 1) with a pd kernel obtained from the cpd kernel.

Fig. 1 shows the feature space $\mathbb{H}$ used in K-CM (Algorithm 1). The mapping $e_i = \Phi(x_i)$ from each data x_i by an explicit mapping Φ is expressed along the i -th axis. From the updating equation of the cluster center W_j in K-CM (Algorithm 1), every cluster center W_j is in the $N - 1$ -dimensional subspace spanned by each data e_i . Hence $e_i - W_j$ for all $i \in \{1, \dots, N\}$ and $j \in \{1, \dots, C\}$ is orthogonal to $\mathbf{1}$. Therefore, it is guaranteed that $d_{i,j}$ is non-negative, even if $\tilde{K}$ is conditionally positive definite (and not positive definite) in K-CM (Algorithm 1).

IV. NUMERICAL EXAMPLE

In this section, we present an example of cpdK-CM. Fig. 2 shows 147 data points in two-dimensional space, which form doughnut-shaped and ball-shaped clusters. Such data cannot be clustered correctly using the usual HCM, sFCM, and eFCM algorithms because these clusters have a nonlinear border. Introducing the cpd kernel

$$\tilde{K}(x, x') = \log(1 + \|x - x'\|_2^{0.5}), \quad (43)$$

cpdK-sFCM (Algorithm 2) with the fuzzifier parameter $m = 1.1$ yielded a successful result, as shown in Fig. 3.

V. CONCLUSION

In this paper, we investigated a hard and two fuzzy c -means clustering algorithms with a conditionally positive definite

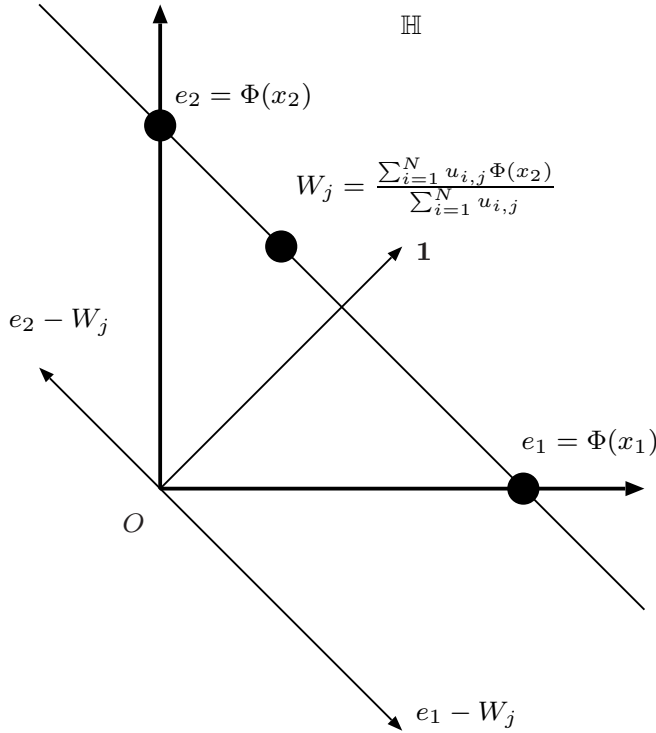


Fig. 1. Data and cluster centers in feature space $\mathbb{H}$

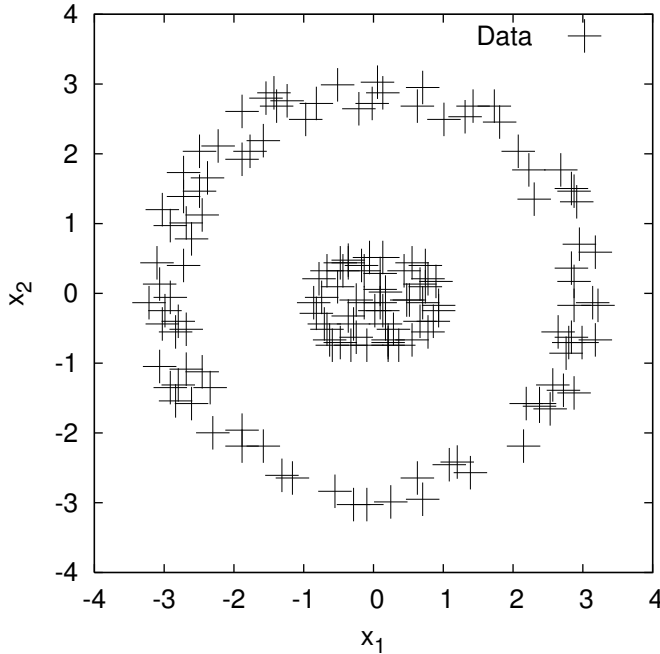


Fig. 2. Data

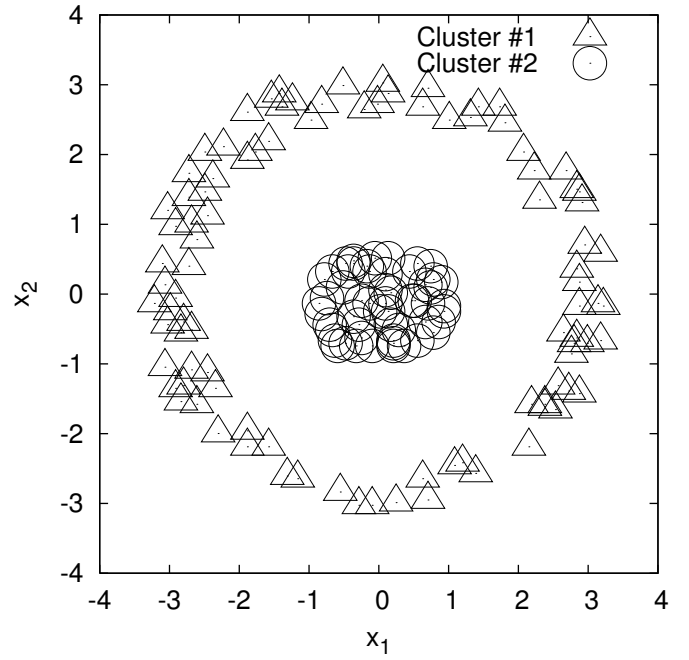


Fig. 3. Successful result ($\tilde{\mathcal{K}}(x, x') = \log(1 + \|x - x'\|_2^{0.5})$, $m = 1.1$)

kernel. First, we developed c -means clustering algorithms using cpd kernel by solving the revised optimization problem of K-HCM, K-sFCM, and K-eFCM; thus, the cpd kernel is the squared Euclidean distance between the data in the feature space. Next, we showed how the cpd kernel is used in K-CM algorithms with a positive definite kernel obtained by arranging the cpd kernel. Furthermore, we showed the relation between the cpdK-CM algorithms and K-CM algorithms with a positive definite kernel obtained from the cpd kernel, and we interpreted a mapping from the original data space to the feature space, which is latent in cpdK-CM algorithms. Finally, we showed a numerical example of the cpdK-CM algorithms.

REFERENCES

- [1] Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum, New York, (1981).
- [2] Miyamoto, S. and Umayahara, K.: "Methods in Hard and Fuzzy Clustering", in: Liu, Z.-Q. and Miyamoto, S. (eds), Soft computing and human-centered machines, Springer-Verlag Tokyo, (2000).
- [3] Miyamoto, S. and Nakayama, Y.: "Algorithms of Hard c -Means Clustering Using Kernel Functions in Support Vector Machines", J. Advanced Computational Intelligence and Intelligent Informatics, Vol.7, No.1, pp.19–24, 2003.
- [4] Miyamoto, S. and Suizu, D.: "Fuzzy c -Means Clustering Using Kernel Functions in Support Vector Machines", J. Advanced Computational Intelligence and Intelligent Informatics, Vol.7, No.1, pp.25–30, 2003.
- [5] Vapnik, V., N.: Statistical Learning Theory, Wiley, New York, 1998.
- [6] Miyamoto, S., Kawasaki, Y. and Sawazaki, K.: "An Explicit Mapping for Kernel Data Analysis and Application to Text Analysis", Proc. IFSA-EUSFLAT 2009, pp.618–623, 2009.
- [7] Miyamoto, S. and Sawazaki, K.: "An Explicit Mapping for Kernel Data Analysis and Application to c -Means Clustering", Proc. NOLTA 2009, pp.556–559, 2009.
- [8] Schoölkopf, B.: "The Kernel Trick for Distances", In Advances in Neural Information Processing System 13, MIT Press, 2001.
- [9] Hathaway, R.J., Davenport, J.W. and Bezdek, J.C.: "Relational Duals of the c -means Clustering Algorithms", Pattern Recognition, Vol.22, No.2, pp.205–212, 1989.