Families of Triangular Norm-Based Kernel Functions and Their Application to Kernel k-Means

Paper: jc*--**-****:**

Families of Triangular Norm-based Kernel Functions and their Application to Kernel k-Means

Kazushi Okamoto

Department of Informatics, Graduate School of Informatics and Engineering
The University of Electro-Communications
1-5-1 Chofugaoka, Chofu, Tokyo 182-8585, Japan
E-mail: kazushi@uec.ac.jp
[Received 00/00/00; accepted 00/00/00]

Abstract. This study proposes the concept of families of triangular norm (t-norm)-based kernel functions, and discusses their positive-definite property and the conditions for applicable t-norms. A clustering experiment with kernel k-means is performed in order to analyze the characteristics of the proposed concept, as well as the effects of the t-norm and parameter selections. It is evaluated that the clusters obtained in terms of the adjusted rand index and the experimental results suggested the following : (1) the adjusted rand index values obtained by the proposed method were almost the same or higher than those produced using the linear kernel for all of the data sets; (2) the proposed method slightly improved the adjusted rand index values for some data sets compared with the radial basis function (RBF) kernel; (3) the proposed method tended to map data to a higher dimensional feature space than the linear kernel but the dimension was lower than that using the RBF kernel.

Keywords: adjusted rand index; clustering; k-means; kernel method; positive-definite kernel; t-norm

1. Introduction

A kernel method is a fundamental and important pattern analysis approach based on a kernel function, which is used in machine learning tasks such as classification, clustering, and dimension reduction. A kernel function corresponds to a similarity measure between two data points, which map each to a feature space and the inner product on that space.

Many kernel functions have been proposed for various data types such as multidimensional real-valued vectors, strings, and graphs. This study considers kernel functions for multidimensional real-valued vectors, e.g., the linear kernel, polynomial kernel, radial basis function (RBF) kernel, χ^2 kernel[1], and histogram intersection kernel[2]. The χ^2 kernel and histogram intersection kernel are calculated by an element-wise binary operation and their accumulation, where the binary operations are $2xy/(x+y)$ and $\min\{x,y\}$, respectively. This kernel function is called an additive kernel[3]. The linear kernel is also considered to be an additive kernel (its binary operation is xy). In addition, the minimum and product operations are triangular norms (t-norms) [4][5], which generalize intersection operations on fuzzy logic; therefore, I consider various types of additive kernels with t-norms as binary operators.

This study proposes the concept of a t-norm-based additive kernel as well as discussing its positive-definite property and the conditions for applicable t-norms. It is evaluated that the characteristics of the proposed method and the effects of the t-norm and parameter selections in a clustering experiment with kernel k-means. In the experiment, four datasets are applied to the proposed kernel function using nonlinear cluster shapes and eight t-norms, two of which were not parameterized t-norms. The evaluation measures the adjusted rand index (ARI) to quantitatively evaluate the clustering accuracy. In addition, it is measured that the computational time required for 1,000 random vectors with a sparse ratio in order to determine the relationship between the processing time and clustering accuracy. Based on the results obtained, this study validates and discusses the effects of changing the kernel parameters and the t-norms selected.

The remainder of this paper is organized as follows. Section 2 provides definitions of a triangular norm, positive-definite kernel, and kernel k-means. Section 3 describes the t-norm-based additive kernel as well as its positive-definite property and the conditions for applicable t-norms. Section 4 explains the conditions for the clustering experiment, and discusses the characteristics of the proposed method based on the clustering experiment. Section 5 compares each kernel function in terms of the computational time required.

2. Definitions of a triangular norm and positive-definite kernel

2.1. Definition of a triangular norm

A function $T : [0,1] \times [0,1] \longrightarrow [0,1]$ is called a triangular norm (t-norm) if and only if $\forall x, y, z \in [0,1]$,

1. $T(x,1) = x$;

2. $T(x,y) \leq T(x,z)$ if $y \leq z$;
3. $T(x,y) = T(y,x)$;
4. $T(x,T(y,z)) = T(T(x,y),z)$.

From the definition, we get $T(x,0) = 0$. According to fuzzy logic, t-norms represent intersection operations. If T is strictly increasing on $(0,1] \times (0,1]$, then T is called a strict t-norm. A t-norm T_1 weakly dominates another t-norm T_2 if $\forall x,y,z \in [0,1]$, and thus we have

$$T_1(x,T_2(y,z)) \geq T_2(T_1(x,y),z). \quad \ldots\ldots \quad (1)$$

2.2. Definition of a positive-definite kernel

A function $K : \Omega \times \Omega \longrightarrow \mathbb{R}$ is called a positive-definite kernel if and only if $\forall x,y \in \Omega$, $\forall x_1,x_2,\cdots,x_n \in \Omega$,

1. $K(x,y) = K(y,x)$;
2. $\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(x_i,x_j) \geq 0$, $\forall c_i,c_j \in \mathbb{R}$.

A kernel function provides the inner product on the feature space ϕ such that $K(x,y) = \phi^t(x) \cdot \phi(y)$. In addition, for $\forall x,y \in \mathbb{R}^d$, the sum of d positive-definite kernels $K_1,K_2,\cdots,K_d \in f : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$,

$$K'(x,y) = \sum_{k=1}^{d} K_{(k)}(x_i^{(k)},y_i^{(k)}) \quad \ldots\ldots \quad (2)$$

is also a positive-definite kernel on $\mathbb{R}^d$, where $x_i^{(k)}$ is the k-th element of a d-dimensional vector x_i, which is shown as

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K'(x_i,x_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \sum_{k=1}^{d} K_{(k)}(x_i^{(k)},x_j^{(k)}) \\ &= \sum_{k=1}^{d}\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K_{(k)}(x_i^{(k)},x_j^{(k)}) \\ &\geq 0. \quad \ldots\ldots \quad (3)\end{aligned}$$

Kernel K' is called an additive kernel.

3. t-norm-based positive-definite kernel

If a strict t-norm T satisfies $T(x,y) \geq xy$, where $\forall x,y \in [0,1]$, then T is positive-definite of order two, which was proved by [6]. If T satisfies the positive-definite requirement of order two, then the determinants of the principal minors of the matrix

$$A = \begin{pmatrix} T(x,x) & T(x,y) \\ T(y,x) & T(y,y) \end{pmatrix} \quad \ldots\ldots \quad (4)$$

are all greater than or equal to zero. The principal minors of the matrix A are $A_1 = (T(x,x))$, $A_2 = (T(y,y))$, and A itself. Clearly, the determinants of A_1 and A_2 are greater than or equal to zero by the definition of t-norm. From equation (1) and $T(x,y) \geq xy$, T satisfies $zT(x,y) \leq T(x,zy)$. If $T(0,y) = 0 \leq x \leq y = T(1,y)$ and $z = T(x,y)/T(y,y)$, then w exists such that $x = T(w,y)$ and the inequality above can be rewritten as

$$\begin{aligned}\frac{T(x,y)}{T(y,y)}T(w,T(y,y)) &\leq T(w,\frac{T(x,y)}{T(y,y)}T(y,y)) \\ &= T(w,T(x,y))\end{aligned}$$

$$\begin{aligned}\frac{T(x,y)}{T(y,y)} &\leq \frac{T(w,T(x,y))}{T(w,T(y,y))} \\ &= \frac{T(T(w,y),x)}{T(T(w,y),y)} \\ &= \frac{T(x,x)}{T(x,y)}. \quad \ldots\ldots \quad (5)\end{aligned}$$

This inequality means that $T(x,x)T(y,y) - T^2(x,y) \geq 0$, and this is the determinant of A; therefore, a strict t-norm T, $T(x,y) \geq xy$, is positive-definite of order two.

The t-norm kernel K_{t} is proposed based on the additive kernel:

$$K_{\mathrm{t}}(x,y) = \sum_{k=1}^{d} T(x^{(k)},y^{(k)}), \quad x,y \in [0,1]^d. \quad \ldots \quad (6)$$

The t-norm is positive-definite of order two, so K_t is also positive-definite of order two:

$$\begin{aligned}\sum_{i=1}^{2}\sum_{j=1}^{2} c_i c_j K_t(x_i,x_j) &= \sum_{i=1}^{2}\sum_{j=1}^{2} c_i c_j \sum_{k=1}^{d} T(x_i^{(k)},x_j^{(k)}) \\ &= \sum_{k=1}^{d}\sum_{i=1}^{2}\sum_{j=1}^{2} c_i c_j T(x_i^{(k)},x_j^{(k)}) \\ &\geq 0. \quad \ldots\ldots \quad (7)\end{aligned}$$

In addition, if a symmetric matrix of order n, $S = \{s_{ij}\}$, is positive-definite, then an n-dimensional feature vector $\phi(x) = (\phi_1(x),\phi_2(x),\cdots,\phi_n(x))^t \in \mathbb{R}^n$ exists for any samples $x_1,x_2,\cdots,x_n \in \Omega$, and

$$s_{ij} = \phi^t(x_i) \cdot \phi(x_j) = K(x_i,x_j). \quad \ldots\ldots \quad (8)$$

This means that any symmetric and positive-definite matrix has a corresponding type of kernel function. A strict t-norm T that satisfies $T(x,y) \geq xy$ is positive-definite of order two, so a symmetric matrix (gram matrix) of order two

$$\begin{pmatrix} T(x_1,x_1) & T(x_1,x_2) \\ T(x_2,x_1) & T(x_2,x_2) \end{pmatrix}, \quad \ldots\ldots \quad (9)$$

is also positive-definite. For any samples $x_1,x_2 \in [0,1]$, a two-dimensional feature vector $\phi(x) = (\phi_1(x),\phi_2(x))^t \in \mathbb{R}^2$ exists, and equation (13) can be rewritten as

$$\begin{pmatrix} \phi^t(x_1)\cdot\phi(x_1) & \phi^t(x_1)\cdot\phi(x_2) \\ \phi^t(x_2)\cdot\phi(x_1) & \phi^t(x_2)\cdot\phi(x_2) \end{pmatrix}. \quad \ldots \quad (10)$$

Therefore, a strict t-norm maps two input scalars onto a two-dimensional feature space and calculates the inner product on that space. The t-norms that satisfy $T(x,y) \geq xy$ are as follows (see [4]):

- logical product

$$\min\{x,y\}, \quad \ldots\ldots\ldots\ldots \quad (11)$$

- Mizumoto product

$$\frac{2}{\pi}\cot^{-1}\left(\cot\frac{1}{2}\pi x+\cot\frac{1}{2}\pi y\right), \quad \ldots\ldots \quad (12)$$

- Dombi t-norm, $p\in[1,\infty)$

$$\frac{1}{1+\sqrt[p]{\left(\frac{1-x}{x}\right)^p+\left(\frac{1-y}{y}\right)^p}}, \quad \ldots\ldots\ldots \quad (13)$$

- Dubois t-norm, $p\in[0,1]$

$$\frac{xy}{\max\{x,y,p\}}, \quad \ldots\ldots\ldots\ldots \quad (14)$$

- Frank t-norm, $p\in(0,1)$

$$\log_p\left(1+\frac{(p^x-1)(p^y-1)}{p-1}\right), \quad \ldots\ldots \quad (15)$$

- Hamacher t-norm, $p\in[0,1]$

$$\frac{xy}{p+(1-p)(x+y-xy)}, \quad \ldots\ldots\ldots \quad (16)$$

- Schweizer t-norm 2, $p\in(0,\infty)$

$$\frac{1}{\sqrt[p]{\frac{1}{x^p}+\frac{1}{y^p}-1}}, \quad \ldots\ldots\ldots\ldots \quad (17)$$

- Schweizer t-norm 3, $p\in[1,\infty)$

$$1-\sqrt[p]{(1-x)^p+(1-y)^p-(1-x)^p(1-y)^p}. \quad (18)$$

4. Clustering experiment with kernel k-means

The kernel method can be applied to pattern analysis tasks such as classification (supervised learning), clustering (unsupervised learning), and dimension reduction in nonlinear data analysis. A clustering experiment using kernel k-means (a clustering algorithm in kernel method) is performed in order to analyze the characteristics of the t-norm-based additive kernel as well as the effects of the t-norm and parameter selections.

The k-means clustering algorithm finds the k-partition of n individuals $X=\left\{x_1,x_2,\cdots,x_n\in\mathbb{R}^d\right\}$. The partition minimizes the objective function. Kernel k-means, an extention of k-means with kernel function, partitions n individuals on feature space ϕ to k-clusters with representative points $M=\left\{\mu_1,\mu_2,\cdots,\mu_k\in\mathbb{R}^d\right\}$ and the objective function J

$$J=\min\sum_{\mu\in M}\sum_{x\in C_i}||\phi(x)-\mu||^2, \quad \ldots\ldots\ldots \quad (19)$$

where

$$\mu_i=\frac{1}{|C_i|}\sum_{x\in C_i}\phi(x), \quad \ldots\ldots\ldots\ldots \quad (20)$$

$$C_i=\left\{x|\mu_i=\underset{\mu\in M}{\operatorname{argmin}}||\phi(x)-\mu||^2\right\}. \quad \ldots \quad (21)$$

The $||\phi(x)-\mu_i||^2$ can be rewritten with a positive-definite kernel K as follows:

$$\begin{aligned}||\phi(x)-\mu_i||^2&=||\phi(x)-\frac{1}{|C_i|}\sum_{x'\in C_i}\phi(x')||^2\\&=K(x,x)-\frac{2}{|C_i|}\sum_{x'\in C_i}K(x,x')\\&\quad+\frac{1}{|C_i|^2}\sum_{x'\in C_i}\sum_{x''\in C_i}K(x',x'').\end{aligned} \quad \ldots \quad (22)$$

The RBF kernel performs inner product on higher dimensions than the linear kernel, and it makes expectation of higher accuracy. The t-norm is a generalization of intersection operations, and I guess that its characteristics is close to the linear kernel since its binary operation is algebraic product (= multiplication), one of t-norms. Therefore, this study pays particular attention to the difference in clustering accuracy between the linear kernel and t-norm kernel. The experiment uses data sets with cluster shapes that are not linearly separable.

4.1. Experimental conditions

This experiment uses four data sets, as shown in Fig. 1. According to Fig. 1, data sets A and B comprised two clusters, which were obtained from [7] and [8], respectively. Data sets C and D in Fig. 1 comprised three clusters, which were obtained from [9]. The cluster shapes were intricate and linearly inseparable in each data set. Data sets C and D were more complex compared with data sets A and B according to Fig. 1. The execution parameters for kernel k-means clustering were as follows:

- two clusters, k, for data sets A and B, and three for data sets C and D;
- each clustering process was terminated when the number of iterations reached 1,000, or the difference between the latest and current objective function values was less than 10^{-4};
- one partition that minimized the objective function was determined within 100 attempts using different initial partitions.

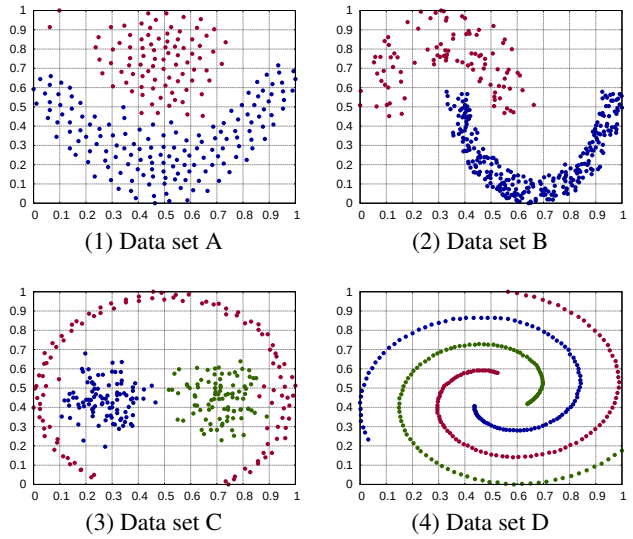


Fig. 1. Four data sets used to evaluate the clustering performance

In this experiment, two standard kernels, i.e., linear kernel K_{lin} and RBF kernel K_{rbf}, were used as the baseline. The definitions of K_{lin} and K_{rbf} are shown with $x, y \in [0,1]^d$:

$$K_{\mathrm{lin}}(x,y) = \sum_{i=1}^{d} x_i y_i$$

$$K_{\mathrm{rbf}}(x,y) = \exp\left(-\frac{\sum_{i=1}^{d}(x_i - y_i)^2}{\sigma^2}\right).$$

The RBF kernel parameter σ was 0.01 to 10 (step size = 0.01). The t-norm kernel parameters, p, were as follows: 1 to 10 (step size = 0.01) for the Dombi t-norm; 0 to 1 (step size = 0.01) for the Dubois t-norm; 0.01 to 0.99 (step size = 0.01) for the Frank t-norm; 0 to 1 (step size = 0.01) for the Hamacher t-norm; 0.01 to 10 (step size = 0.01) for the Schweizer t-norm 2; and 1 to 10 (step size = 0.01) for the Schweizer t-norm 3.

ARI[10] was used to quantitatively evaluate the clustering results. For two partitions $U = \{u_1, u_2, \cdots, u_M\}$ and $V = \{v_1, v_2, \cdots, v_N\}$, the definition of ARI is

$$ARI = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} {}_{n_{ij}}C_2 - \frac{ab}{{}_nC_2}}{\frac{1}{2}(a+b) - \frac{ab}{{}_nC_2}}, \quad \ldots\ldots\ldots (23)$$

$$a = \sum_{i=1}^{M} {}_{n_{i\cdot}}C_2, \quad \ldots\ldots\ldots\ldots (24)$$

$$b = \sum_{j=1}^{N} {}_{n_{\cdot j}}C_2, \quad \ldots\ldots\ldots\ldots (25)$$

where $n_{ij} = |u_i \cap v_j|$, $n_{i\cdot} = \sum_{j=1}^{N} n_{ij}$, $n_{\cdot j} = \sum_{i=1}^{M} n_{ij}$, and $n = \sum_{i=1}^{M}\sum_{j=1}^{N} n_{ij}$. In this evaluation, I assume that $M = N$, U corresponded to a correct partition that was uniquely determined from each data set, and V corresponded to a partition predicted by kernel k-means clustering. ARI was calculated for each of the clustering results.

Table 1. Best ARI values for each kernel and data set

	data set			
	A	B	C	D
linear kernel	0.4535	0.5767	0.4650	-0.0054
RBF kernel	0.4880 [σ=8.52]	0.5767 [σ=9.99]	**0.7611** [σ=0.28]	**0.1375** [σ=0.33]
t-norm kernel logical product	0.0240	0.5146	0.2990	-0.0037
t-norm kernel Mizumoto product	0.4997	0.5528	0.4650	-0.0050
t-norm kernel Dombi t-norm	**0.5237** [p=1.98]	0.5612 [p=3.81]	0.4717 [p=7.42]	0.0462 [p=8.95]
t-norm kernel Dubois t-norm	0.5117 [p=0.82]	**0.5853** [p=0.76]	0.4757 [p=0.37]	0.0315 [p=0.19]
t-norm kernel Frank t-norm	0.4880 [p=0.01]	0.5767 [p=0.99]	0.4688 [p=0.63]	-0.0049 [p=0.43]
t-norm kernel Hamacher t-norm	0.4880 [p=0.33]	0.5767 [p=1.00]	0.4650 [p=1.00]	-0.0046 [p=0.11]
t-norm kernel Schweizer t-norm 2	**0.5237** [p=3.29]	0.5767 [p=0.55]	0.4717 [p=9.46]	0.0477 [p=4.32]
t-norm kernel Schweizer t-norm 3	**0.5237** [p=2.48]	0.5767 [p=1.97]	0.4717 [p=9.99]	0.0445 [p=8.77]

4.2. Evaluation of the clustering results in terms of the ARI and cluster shapes

The ARI values for each kernel and data set are shown in Table 1. The performance was better when the ARI value was higher. In Table 1, the best ARI values for each data set are indicated by the underlined bold font.

The parameterized t-norms were the Dombi t-norm, Dubois t-norm, Frank t-norm, Hamacher t-norm, Schweizer t-norm 2, and Schweizer t-norm 3. According to Table 1, the ARI values for the parameterized t-norm kernels were almost the same or higher than those for the linear kernel with all of the data sets. The differences in the ARI values for the linear kernel and parameterized t-norm kernels were 0.0345 (Frank t-norm and Hamacher t-norm) to 0.0702 (Dombi t-norm, Schweizer t-norm 2, and Schweizer t-norm 3) for data set A, –0.0155 (Dombi t-norm) to 0.0086 (Dubois t-norm) for data set B, 0 (Hamacher t-norm) to 0.0107 (Dubois t-

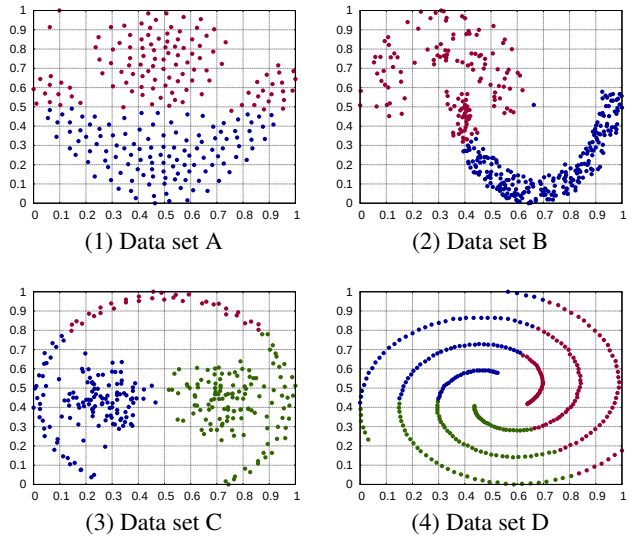


Fig. 2. Clustering results obtained with the linear kernel

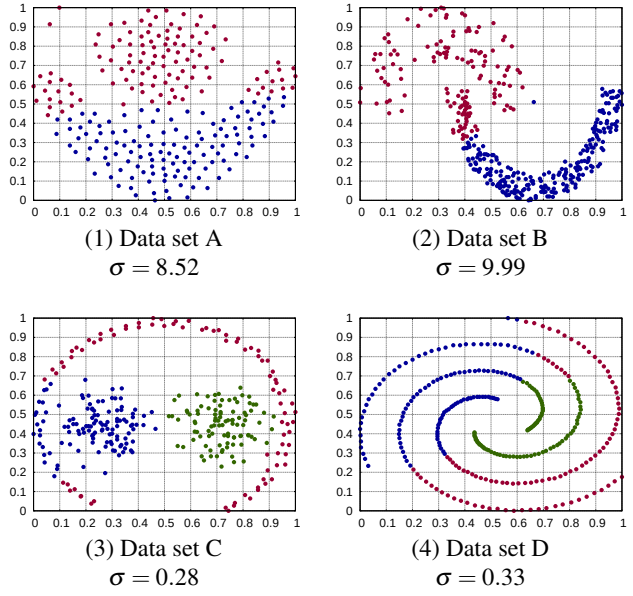


Fig. 3. Clustering results obtained with the RBF kernel using the parameters for each data set that achieved the best ARI values

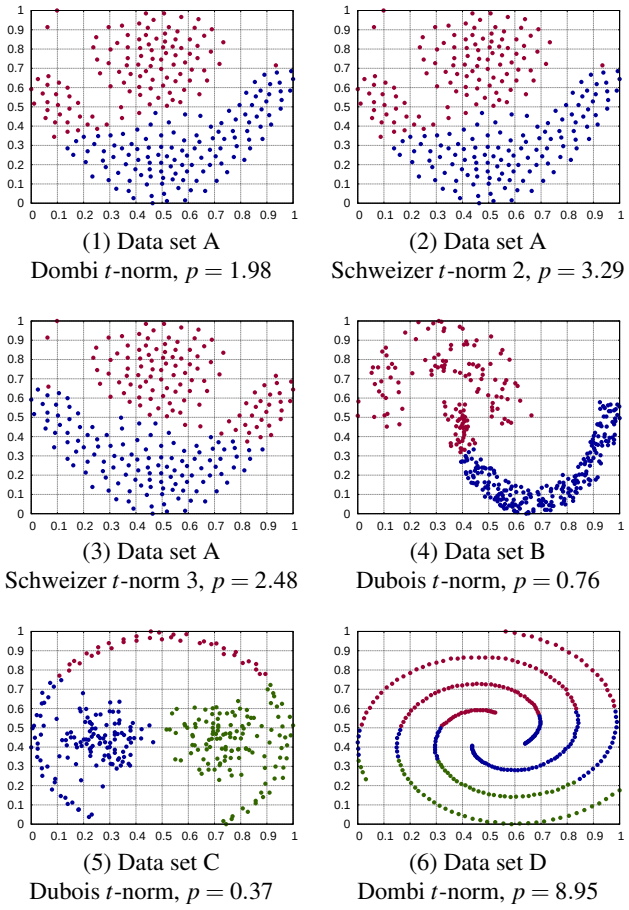
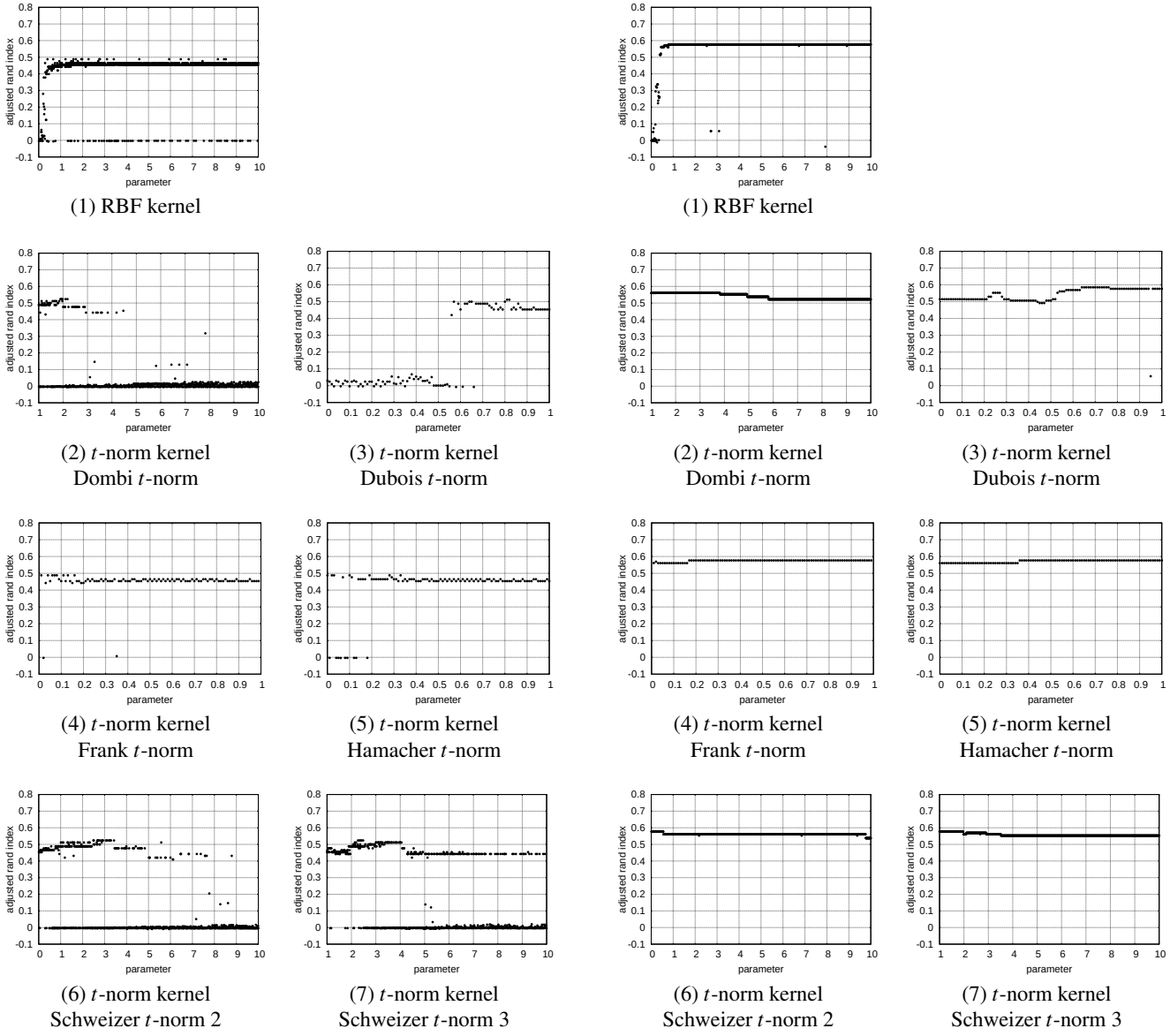


Fig. 4. Clustering results obtained with t-norm kernels using the combination of the t-norm and parameter for each data set that achieved the best ARI values

norm) for data set C, and 0.0008 (Hamacher t-norm) to 0.0531 (Schweizer t-norm 2) for data set D. The ARI values of the parameterized t-norm kernels for data sets A and B were better than those for the RBF kernel, whereas the values for data sets C and D are lower than those for the RBF kernel.

The cluster shapes obtained with the linear kernel, RBF kernel, and t-norm kernels are shown in Fig. 2, Fig. 3, and Fig. 4, respectively. In Fig. 3, the best parameters that maximized the ARI for each data set were determined from Table 1. In Fig. 4, the best combinations of the t-norm and its parameters that maximized the ARI for each data set were also determined from Table 1. According to Fig. 2 (1), Fig. 3 (1), and Fig. 4 (1) (2) (3), the cluster shapes were almost the same for the linear kernel and RBF kernel, whereas they differed for the t-norm kernels. According to Fig. 2 (2), Fig. 3 (2), and Fig. 4 (4), the cluster shapes were extremely similar for the linear kernel, RBF kernel, and t-norm kernel with the Dubois t-norm.

Finding the cluster centers based on a two-dimensional feature space requires that the cluster shapes are separable in terms of drawing a circle around each cluster center. Separation of the cluster shapes is possible in Fig. 2 (3) and (4), but not in Fig. 3 (3) and (4) because kernel k-means using the RBF kernel maps the data points onto a higher dimensional feature space and calculates the cluster centers based on this feature space. Separation of the cluster shapes is possible for the t-norm kernels in Fig. 4 (5) but not for those in Fig. 4 (6). Therefore, I consider that t-norm kernels also map data points onto a higher dimensional feature space but the dimension is lower than that used by the RBF kernel because the t-norm kernel could not correctly detect the curve lined cluster shape in data set C.



(1) RBF kernel

(2) t-norm kernel Dombi t-norm

(3) t-norm kernel Dubois t-norm

(4) t-norm kernel Frank t-norm

(5) t-norm kernel Hamacher t-norm

(6) t-norm kernel Schweizer t-norm 2

(7) t-norm kernel Schweizer t-norm 3

Fig. 5. Kernel parameter - ARI value graphs for data set A

(1) RBF kernel

(2) t-norm kernel Dombi t-norm

(3) t-norm kernel Dubois t-norm

(4) t-norm kernel Frank t-norm

(5) t-norm kernel Hamacher t-norm

(6) t-norm kernel Schweizer t-norm 2

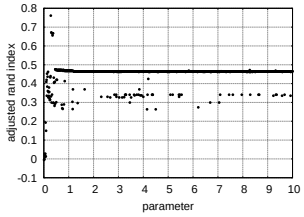
(7) t-norm kernel Schweizer t-norm 3

Fig. 6. Kernel parameter - ARI value graphs for data set B

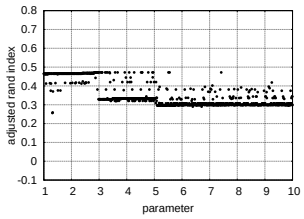
4.3. Evaluation of the kernel parameter selection based on the ARI

The input-output relationship of parameterized t-norm depends on the parameter, and we should treat it as a different t-norm if the parameter is different even with the same t-norm. Hence, it is necessary to validate how to set parameters. The Fig. 5, Fig. 6, Fig. 7, and Fig. 8 show the ARI values for each kernel and parameter, where these figures present the kernel parameter ARI graphs corresponding to the seven parameterized kernels for data set A, data set B, data set C, and data set D, respectively. In Fig. 5, Fig. 6, Fig. 7, and Fig. 8, the term "parameter" means σ for RBF kernel and p for parameterized t-norms. A plotted point in each graph corresponds to an ARI value for a parameter, and the number of plotted points differs for each graph since the number of parameters differs for each parameterized t-norm (please refer section 4.1). The used kernel parameters are same as section 4.1.

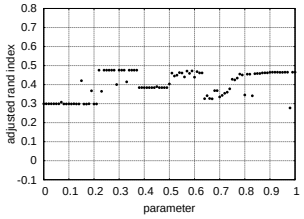
The Fig. 5 (1), Fig. 6 (1), Fig. 7 (1), and Fig. 8 (1) confirm that the RBF kernel achieved the highest ARI values when the parameter selection was a success, but these parameters were limited. In addition, even when the parameter selection was a failure, the ARI values for the RBF kernel were almost equal to the best results for the t-norm kernels. Thus, the results suggest that the RBF kernel performed better than the t-norm kernels in terms of the clustering accuracy. In addition, I consider that the utility value of t-norm kernels depends on the specific application and task because: (1) the t-norm does not need to be calculated if either input is zero; (2) some of the t-norms, such as the Dubois t-norm and Hamacher t-norm, require simple, low-cost operations and calculations so a low computational cost is expected; and (3) t-norm kernels perform better than linear kernels.



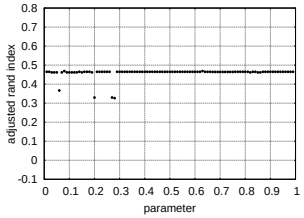
(1) RBF kernel



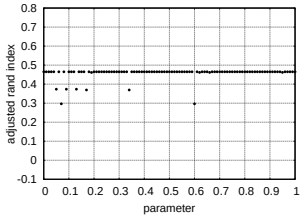
(2) t-norm kernel
Dombi t-norm



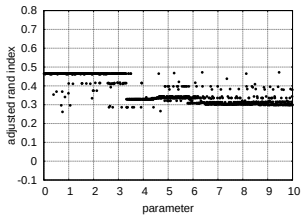
(3) t-norm kernel
Dubois t-norm



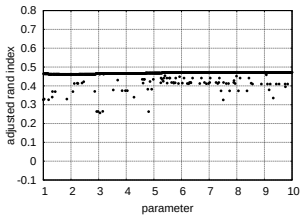
(4) t-norm kernel
Frank t-norm



(5) t-norm kernel
Hamacher t-norm

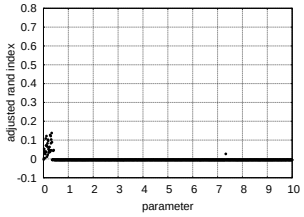


(6) t-norm kernel
Schweizer t-norm 2

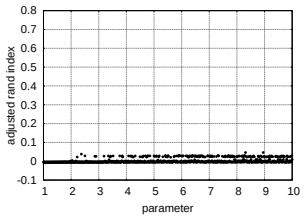


(7) t-norm kernel
Schweizer t-norm 3

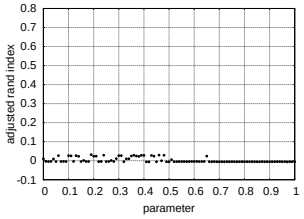
Fig. 7. Kernel parameter - ARI value graphs for data set C



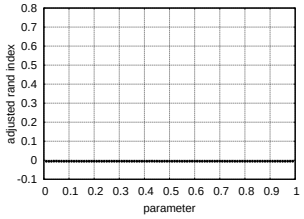
(1) RBF kernel



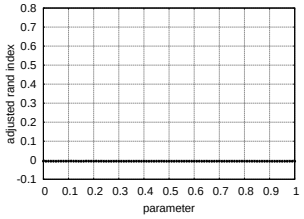
(2) t-norm kernel
Dombi t-norm



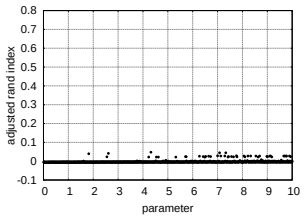
(3) t-norm kernel
Dubois t-norm



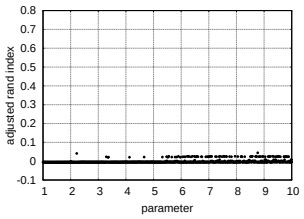
(4) t-norm kernel
Frank t-norm



(5) t-norm kernel
Hamacher t-norm



(6) t-norm kernel
Schweizer t-norm 2



(7) t-norm kernel
Schweizer t-norm 3

Fig. 8. Kernel parameter - ARI value graphs for data set D

According to the parameter selections for the t-norm kernels, Fig. 5, Fig. 6, Fig. 7, and Fig. 8 indicate the following:

- t-norm kernels with the Frank t-norm and Hamacher t-norm are robust to the parameters selected;
- t-norm kernels with the Dombi t-norm and Schweizer t-norm (2) (3) are sensitive to the parameters selected;
- it seems better to set the parameter of t-norm kernel with Dombi t-norm and Schweizer t-norm (2) (3) to three or less;
- the recommended parameters for the t-norm kernel with the Dubois t-norm are [0.7, 0.9].

5. Computational times required for kernel function calculation

The property of t-norm, $T(0,x) = 0$, makes an advantage that t-norm kernel calculation requires only summation of t-norm outputs whose inputs are not both 0. Thus, it is effective to apply t-norm kernel to L1 normalized sparse histograms in terms of computation costs. In this study, the computational times were measured for each kernel function with $V = \{v_1, v_2, \cdots, v_{1000}\}$, a set of 1,000 random vectors. $v_i \in V$ is a 500-dimensional vector, which was randomly generated with a sparse ratio s that represents the ratio of the number of zeroes in the vector. The computational time of

$$\sum_{i=1}^{999} \sum_{j=i+1}^{1000} K(v_i, v_j) \quad \cdots \quad (26)$$

Table 2. Average computational times for each kernel when applied to a set of 1,000 random vectors generated with a sparse ratio s

	time [s]									
	$s=0.1$	$s=0.2$	$s=0.3$	$s=0.4$	$s=0.5$	$s=0.6$	$s=0.7$	$s=0.8$	$s=0.9$	$s=1.0$
linear kernel	3.2×10^{-9}	3.0×10^{-9}	2.8×10^{-9}	3.5×10^{-9}	3.6×10^{-9}	3.4×10^{-9}	3.4×10^{-9}	2.9×10^{-9}	2.6×10^{-9}	2.9×10^{-9}
RBF kernel	8.4×10^{-2}	8.4×10^{-2}	8.4×10^{-2}	8.4×10^{-2}	8.4×10^{-2}	8.4×10^{-2}	8.5×10^{-2}	8.4×10^{-2}	8.4×10^{-2}	8.4×10^{-2}
t-norm kernel logical product	2.5×10^{-9}	2.5×10^{-9}	2.5×10^{-9}	2.5×10^{-9}	2.4×10^{-9}	2.5×10^{-9}	2.5×10^{-9}	2.5×10^{-9}	2.5×10^{-9}	2.5×10^{-9}
t-norm kernel Mizumoto product	2.5×10^{-9}	2.5×10^{-9}	2.5×10^{-9}	2.5×10^{-9}	2.5×10^{-9}	2.5×10^{-9}	2.5×10^{-9}	2.5×10^{-9}	2.5×10^{-9}	2.5×10^{-9}
t-norm kernel Dombi t-norm	0.13	0.32	0.61	1.0	1.5	2.1	2.7	3.4	4.3	5.2
t-norm kernel Dubois t-norm	2.5×10^{-9}	2.4×10^{-9}	2.4×10^{-9}	2.4×10^{-9}	2.3×10^{-9}	2.3×10^{-9}	2.1×10^{-9}	2.1×10^{-9}	2.1×10^{-9}	2.1×10^{-9}
t-norm kernel Frank t-norm	0.1	0.22	0.4	0.65	0.9	1.3	1.8	2.1	2.7	3.1
t-norm kernel Hamacher t-norm	2.5×10^{-9}	2.5×10^{-9}	2.4×10^{-9}	2.4×10^{-9}	2.3×10^{-9}	2.2×10^{-9}	2.1×10^{-9}	2.2×10^{-9}	2.1×10^{-9}	2.2×10^{-9}
t-norm kernel Schweizer t-norm (2)	0.13	0.32	0.63	1.0	1.6	2.2	3.1	3.8	4.7	5.7
t-norm kernel Schweizer t-norm (3)	0.13	0.31	0.61	1.0	1.5	2.1	2.8	3.5	4.3	5.2

was measured on a workstation with a dual Intel Xeon CPU E5-2650 v3 (2.30GHz $\times$ 10 cores), 64 GB RAM, and Ubuntu Linux 16.04. The C++ language and GNU Compiler Collection (version 5.4.0) are used for the implementation. The computational times for equation (30) were measured for 100 repeats of the three non-parameterized kernels, which were measured using all of the parameters shown in Section 4.1 for the seven parameterized kernels. The purpose of measuring computational time is to evaluate the position of the proposal in terms of computational cost. In kernel k-means case, measuring computational time corresponds to measure the calculation speed of a kernel function K in equation (26), and the measurement helps to select which kernel (or t-norm) should be used.

Table 2 shows the average computational time for each kernel when applied to a set of 1,000 random vectors V generated with a sparse ratio of s. According to Table 2, the linear kernel and t-norm kernels with the logical product, Mizumoto product, Dubois t-norm, and Hamacher t-norm had approximately the same average computational time. Similarly, the t-norm kernels with the Dombi t-norm, Frank t-norm, Schweizer t-norm (2), and Schweizer t-norm (3) had approximately the same average computational times. The RBF kernel is used widely and it also achieved better performance in our experiment using kernel k-means, but the average computational time was reduced by 2.5×10^{-8} to 3×10^{-8} when using the t-norm kernel with the Dubois t-norm. Moreover, the average computational times for the t-norm kernel with the Dombi t-norm, Frank t-norm, Schweizer t-norm (2), and Schweizer t-norm (3) increased with a larger sparse ratio s. These kernels are unsuitable for dense vectors.

According to the discussions in section 4.2, section 4.3, and this section, the characteristics of t-norm kernel are validated in terms of clustering accuracy, parameter selection, and computational times, but after all what t-norm should we use? In my opinion, while the proper t-norm should be considered and selected after applying it to data, I recommend the Dubois t-norm to use initially, because of its low computational cost, better clustering accuracy compared with the linear kernel, and the simple parameter selection process. Then, the t-norm kernel with Dubois t-norm can be a useful option when the clustering accuracy of the linear kernel is insufficient and data size is huge such as big data case.

6. Conclusion

This study proposed the concept of the t-norm-based additive kernel, as well as discussing its positive-definite property and the conditions for applicable t-norms. A clustering experiment with kernel k-means was performed to analyze the characteristics of the proposed method as well as the effects of the t-norm and parameter selections, where the clustering results obtained were evaluated in terms of the ARI. The experiment used four data sets with nonlinear cluster shapes and it was applied that eight t-norms to the proposed kernel function, two of which were non-parameterized t-norms. In addition, it was measured that the computational times for 1,000 random vectors with a sparse ratio to determine the relationship between the processing time and clustering accuracy. The results of the clustering experiment suggested that: (1) the ARI

values obtained by the proposed method were almost the same or higher than those by the linear kernel with all of the data sets; (2) the proposed method slightly improved the ARI values for some data sets compared with the RBF kernel; and (3) the proposed method maps data to a higher dimensional feature space than the linear kernel but the dimension is lower than that of the RBF kernel. The t-norm kernel with the Dubois t-norm had a low calculation cost compared with the RBF kernel, and it obtained good ARI values for some data sets with the evaluated kernel functions; therefore, I consider that this kernel is useful depending on the application and task. Then, I recommend the Dubois t-norm to use initially, because of its low computational cost, better clustering accuracy compared with the linear kernel, and the simple parameter selection process.

This study only performed clustering experiments using data sets with two dimensions. In future research, I will perform clustering experiments using multi-dimensional data sets or data sets with an extremely high number of dimensions compared with the number of instances, and the characteristics of the proposed method should also be analyzed in other pattern analysis tasks such as classification and dimension reduction.

References:

[1] Y. Rubner, J. Puzicha, C. Tomasi, J. M. Buhmann, *Empirical evaluation of dissimilarity measures for color and texture*, Computer Vision and Image Understanding, 84(1), 25–43, 2001.

[2] A. Barla, F. Odone, and A. Verri, *Histogram intersection kernel for image classification*, Proceedings of 2003 International Conference on Image Processing, 2, III-513-16, 2003.

[3] A. Vedaldi and A. Zisserman, *Efficient additive kernels via explicit feature maps*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(3), 480–492, 2012.

[4] M. Mizumoto, *Pictorial representations of fuzzy connectives, part I: cases of t-norms, t-conorms and averaging operators*, Fuzzy Sets and Systems, 31(2), 217–242, 1989.

[5] E. P. Klement, R. Mesiar, and E. Pap, *Triangular norms. position paper I: basic analytical and algebraic properties*, Fuzzy Sets and Systems, 143(1), 5–26, 2004.

[6] C. Alsina and M. S. Tomas, *On positive semidefinite strict t-norms*, General Inequalities 6, 215–225, 1992.

[7] L. Fu and E. Medico, *FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data*, BMC Bioinformatics, 8(1), 1–15, 2007.

[8] A. K. Jain and M. H. C. Law, *Data clustering: a user's dilemma*, Lecture Notes in Computer Science, 3776, 1–10, 2005.

[9] H. Chang and D. Y. Yeung, *Robust path-based spectral clustering*, 41(1), 191–203, 2008.

[10] L. Hubert and P. Arabie, *Comparing partitions*, Journal of Classification, 2(1), 193–218, 1985.

Name:
Kazushi Okamoto

Affiliation:
Department of Informatics, Graduate School of Informatics and Engineering, The University of Electro-Communications

Address:
1-5-1 Chofugaoka, Chofu, Tokyo 182-8585, Japan

Brief Biographical History:
2002-2006 B.E., Kochi University of Technology, Japan
2006-2008 M.E., Kochi University of Technology, Japan
2008-2011 Ph.D., Tokyo Institute of Technology, Japan
2011-2015 Assistant Professor, Chiba University, Japan
2015- Assistant Professor, The University of Electro-Communications, Japan

Main Works:
- K. Okamoto, F. Dong, S. Yoshida, and K. Hirota, Content-Based Image Retrieval via Combination of Similarity Measures, Journal of Advanced Computational Intelligence and Intelligent Informatics, 15(6), 687-697, 2011.
- K. Okamoto, K. Kawamoto, F. Dong, S. Yoshida, and K. Hirota, An Evaluation Strategy for Visual Key Image Retrieval on Mobile Devices, Journal of Advanced Computational Intelligence and Intelligent Informatics, 16(5), 713-722, 2012.

Membership in Academic Societies:
- Japan Society for Fuzzy Theory and Systems (SOFT)
- The Institute of Electronics, Information and Communication Engineers (IEICE)
- Information Processing Society of Japan (IPSJ)