九州大学学術情報リポジトリ
Kyushu University Institutional Repository

A Robust K-Means for Document Clustering

Yu, Hengjun
Department of Communication Design Science, Kyushu University

Inoue, Kohei
Department of Communication Design Science, Kyushu University

Hara, Kenji
Department of Communication Design Science, Kyushu University

Urahama, Kiichi
Department of Communication Design Science, Kyushu University

https://hdl.handle.net/2324/1924408

出版情報：Journal of the Institute of Industrial Applications Engineers. 6 (2), pp.54-59, 2018-04-25. 産業応用工学会
バージョン：
権利関係：

Journal of the Institute of Industrial Applications Engineers Vol.6, No.2, pp.60–65, (2018.4.25)
DOI: 10.12792/JIIAE.6.60 Online edition: ISSN 2187-8811 Print edition: ISSN 2188-1758

Paper

A Robust K-Means for Document Clustering

HENGJUN YU* Member, KOHEI INOUE* Member
KENJI HARA* Non-member, KIICHI URAHAMA* Member

(Received November 2, 2017, revised March 18, 2018)

Abstract: We propose a robust K-means clustering algorithm for document clustering, where we suppose that a document-term matrix is given as an input dataset, and the documents in the dataset are clustered on the basis of the frequency of terms that occur in each document. We introduce a robust loss function to K-means clustering to obtain its robust version, and also propose a feature transform method for improving the performance of document clustering. Experimental results show that the proposed method improves the robustness of K-means to outliers and the performance of document clustering demonstrated on one of the BBC datasets originating from the BBC News.

Keywords: Robust clustering, K-means clustering algorithm, Document clustering, Hellinger distance

1. Introduction

Document clustering is a process for finding a number of clusters of similar documents from a large set of textual documents, and has various applications including information retrieval, text mining and automatic document organization. Numerous document clustering methods have been extensively studied in these years. Andrews and Fox overviewed recent developments in document clustering research [1]. Shah and Mahajan reviewed semantic driven document clustering methods [2], and also provided a detailed overview of various document clustering algorithms [3]. Balabantaray et al. compared K-means and K-medoids clustering algorithms for document clustering, and observed that K-means yields better result than K-medoids [4]. Dzogang et al. extended the spherical K-means [5] to an ellipsoidal K-means [6]. Mei and Wang proposed a hyperspherical fuzzy C-means clustering algorithm for online document categorization [7].

The above document clustering methods based on K-means and its variants including K-medoids do not introduce robust techniques directly. In this paper, we propose a robust K-means clustering algorithm for document clustering. In robust statistics, various approaches have been studied to achieve robustness [8]. A reasonable approach to get robustness is to design a robust potential function [9] or loss function [10]. We propose a robust loss function, and apply it to K-means clustering. The proposed loss function has an intermediate form between the l_1 and l_2-norms. Hamza and Brady used a robust cost function similar to the proposed loss function for robust nonnegative matrix factorization for reflectance spectra reconstruction [11]. Their cost function is a special case of the proposed loss function, which is also similar to the Charbonnier loss function [12]. Recently, Barron generalized popular loss functions including the Charbonnier loss function to a two-parameter loss function [10].

We also propose a method for feature transform from a document-term matrix into feature vectors of documents based on the Hellinger distance [13], which is a distance between two probability distributions. Bui et al. [14] also used the Hellinger distance for multi-criteria document clustering with the latent Dirichlet allocation (LDA) proposed by Blei et al. [15]. Our proposed feature transform method is suitable for K-means clustering algorithms. Experimental results show that the proposed robust K-means clustering algorithm can alleviate the influence of outliers being in data, and improve the performance of document clustering.

The rest of this paper is organized as follows: Section 2 briefly summarizes conventional K-means clustering. Section 3 proposes a robust K-means clustering algorithm based on a robust loss function. Section 4 describes a feature transform method for document clustering based on a document-term matrix. Section 5 shows experimental results of clustering of artificial and real datasets. Finally, Section 6 concludes this paper.

2. K-Means Clustering

Let $\boldsymbol{x}_i \in \mathbb{R}^d$ be a d-dimensional real vector for $i = 1, 2, \ldots, n$. Then K-means clustering aims to partition them into K clusters, and is formulated as the following optimization problem:

$$\min_{\{\boldsymbol{\mu}_k\},\{C_k\}} \sum_{k=1}^{K} \sum_{\boldsymbol{x}\in C_k} \|\boldsymbol{x} - \boldsymbol{\mu}_k\|^2 , \tag{1}$$

where C_k denotes the set of vectors in the kth cluster, $\boldsymbol{\mu}_k \in \mathbb{R}^d$ denotes the centroid in C_k for $k = 1, 2, \ldots, K$, and $\|\cdot\|$ denotes the Euclidean norm. For a fixed set of $\boldsymbol{\mu}_k$, assigning each $\boldsymbol{x}_i$ to the nearest centroid $\boldsymbol{\mu}_k$, we can minimize the ob-

* Department of Communication Design Science, Kyushu University
4-9-1, Shiobaru, Minami-ku, Fukuoka 815-8540, Japan
(k-inoue@design.kyushu-u.ac.jp)

Published by IIAE. 2018

jective function $E(\{\boldsymbol{\mu}_k\},\{C_k\}) = \sum_{k=1}^{K}\sum_{\boldsymbol{x}\in C_k}\|\boldsymbol{x}-\boldsymbol{\mu}_k\|^2$, that is, the ith vector $\boldsymbol{x}_i$ is assigned to the k_ith cluster C_{k_i} for

$$k_i = \arg\min_k \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^2 . \tag{2}$$

After the assignment of all the vectors, we have K clusters $\{C_k\}$ temporarily. Then, for the obtained $\{C_k\}$, we optimize each $\boldsymbol{\mu}_k$ by solving $\partial E/\partial\boldsymbol{\mu}_k = \mathbf{0}$ for $\boldsymbol{\mu}_k$ as follows:

$$\boldsymbol{\mu}_k = \frac{\sum_{\boldsymbol{x}\in C_k}\boldsymbol{x}}{\sum_{\boldsymbol{x}\in C_k}1}. \tag{3}$$

The above procedure for updating $\{C_k\}$ and $\{\boldsymbol{\mu}_k\}$ is repeated until the convergence. The K-means clustering algorithm is summarized as follows:

[K-means clustering algorithm]

0. Assume that a set of vectors $\{\boldsymbol{x}_i\}_{i=1}^n$ and a number of clusters K are given.
1. Initialize the centroid $\boldsymbol{\mu}_k$ in cluster C_k for $k = 1,2,\ldots,K$.
2. Assign each $\boldsymbol{x}_i$ to the nearest cluster determined by (2).
3. Update each $\boldsymbol{\mu}_k$ by (3).
4. If every $\boldsymbol{\mu}_k$ is unchanged by the above update, then halt the procedure. Otherwise, go to the step 2.

3. Robust K-Means Clustering

The above K-means clustering is based on the squared Euclidean distance between a given vector $\boldsymbol{x}_i$ and a centroid $\boldsymbol{\mu}_k$. The squared Euclidean distance can be expressed as follows:

$$\|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^2 = \rho\left(\|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|\right) \tag{4}$$

for $\rho(x) = x^2$. The function like $\rho(x)$ is referred to as the loss function [10]. In this section, we propose an alternative form of $\rho(x)$ as follows:

$$\rho^R(x) = \sqrt{b^2 + x^2} - b \tag{5}$$

for a positive constant b. When $b = 0$, the function $\rho^R(x)$ coincides with the absolute value function $|x|$, which is more robust to outliers than x^2. Although $|x|$ is not differentiable at $x = 0$, $\rho^R(x)$ with $b > 0$ is differentiable everywhere as

$$\frac{d\rho^R}{dx} = \frac{x}{\sqrt{b^2 + x^2}}. \tag{6}$$

We have the following properties of $\rho^R(x)$:

Property 1. $\rho^R(x) \geq 0$, and $\rho^R(0) = 0$.

Property 2. $|d\rho^R/dx| \leq 1$.

Property 3. if $b \geq 1/2$, then $x^2 \geq \rho^R(x)$.

Property 4. $d^2\rho^R/dx^2 \geq 0$, i.e., $\rho^R(x)$ is a convex function.

Based on the above loss function, we formulate a robust K-means clustering as follows:

$$\min_{\{\boldsymbol{\mu}_k\},\{C_k\}} \sum_{k=1}^{K}\sum_{\boldsymbol{x}\in C_k}\rho^R\left(\|\boldsymbol{x}-\boldsymbol{\mu}_k\|\right). \tag{7}$$

Let $E^R(\{\boldsymbol{\mu}_k\},\{C_k\})$ be the objective function in (7). Then we have the necessary condition for optimality as follows:

$$\frac{\partial E^R}{\partial\boldsymbol{\mu}_k} = \sum_{\boldsymbol{x}\in C_k}\frac{\boldsymbol{x}-\boldsymbol{\mu}_k}{\sqrt{b^2 + \|\boldsymbol{x}-\boldsymbol{\mu}_k\|^2}} = \mathbf{0}, \tag{8}$$

where $\mathbf{0}$ is a d-dimensional zero vector having all components equal to zero. From (8) we have

$$\boldsymbol{\mu}_k = \frac{\sum_{\boldsymbol{x}\in C_k}\frac{\boldsymbol{x}}{\sqrt{b^2 + \|\boldsymbol{x}-\boldsymbol{\mu}_k\|^2}}}{\sum_{\boldsymbol{x}\in C_k}\frac{1}{\sqrt{b^2 + \|\boldsymbol{x}-\boldsymbol{\mu}_k\|^2}}}, \tag{9}$$

which is used for updating $\boldsymbol{\mu}_k$ instead of (3) in the proposed robust K-means clustering algorithm.

4. Feature Transform for Document Clustering

Assume that a document-term matrix is given as a dataset for document clustering. Let $A = [a_{ij}]$ for $i = 1,2,\ldots,n$ and $j = 1,2,\ldots,d$ be a document-term matrix, where a_{ij} is the (i, j) element of A, and denotes the frequency of the jth term that occurs in the ith document. Then we compute the probability of occurrence of each term in each document by $p_{ij} = a_{ij}/\sum_{j'=1}^{d} a_{ij'}$ which satisfies $\sum_{j=1}^{d} p_{ij} = 1$, and stack them in a vector as $\boldsymbol{p}_i = [p_{i,1}, p_{i,2}\ldots, p_{i,d}]$ for $i = 1,2,\ldots,n$. Then the Hellinger distance between $\boldsymbol{p}_i$ and $\boldsymbol{p}_{i'}$ is defined by [13]

$$D_H(\boldsymbol{p}_i,\boldsymbol{p}_{i'}) = \frac{1}{\sqrt{2}}\sqrt{\sum_{j=1}^{d}\left(\sqrt{p_{ij}} - \sqrt{p_{i'j}}\right)^2}. \tag{10}$$

The square of $D_h(\boldsymbol{p}_i,\boldsymbol{p}_{i'})$ can be written as

$$\begin{aligned} D_H^2(\boldsymbol{p}_i,\boldsymbol{p}_{i'}) &= \frac{1}{2}\sum_{j=1}^{d}\left(\sqrt{p_{ij}} - \sqrt{p_{i'j}}\right)^2 && (11)\\ &= \sum_{j=1}^{d}\left[\frac{1}{\sqrt{2}}\left(\sqrt{p_{ij}} - \sqrt{p_{i'j}}\right)\right]^2 && (12)\\ &= \sum_{j=1}^{d}\left(\sqrt{\frac{p_{ij}}{2}} - \sqrt{\frac{p_{i'j}}{2}}\right)^2 && (13)\\ &= D_E^2\left(\tilde{\boldsymbol{p}}_i,\tilde{\boldsymbol{p}}_{i'}\right), && (14) \end{aligned}$$

where $\tilde{\boldsymbol{p}}_i = [\sqrt{p_{i,1}/2}, \sqrt{P_{i,2}/2},\ldots,\sqrt{p_{i,d}/2}]$, and $D_E^2(\tilde{\boldsymbol{p}}_i,\tilde{\boldsymbol{p}}_{i'})$ denotes the square of the Euclidean distance between $\tilde{\boldsymbol{p}}_i$ and $\tilde{\boldsymbol{p}}_{i'}$. That is, the Hellinger distance between $\boldsymbol{p}_i$ and $\boldsymbol{p}_{i'}$ is equivalent to the Euclidean distance between $\tilde{\boldsymbol{p}}_i$ and $\tilde{\boldsymbol{p}}_{i'}$. As a result, we obtain a means of (robust) K-means clustering

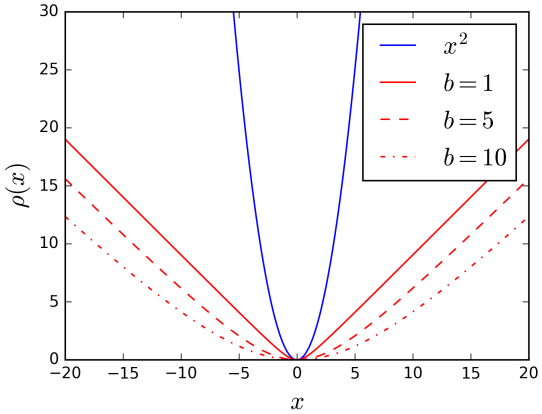

Figure 1: Graphs of $\rho(x)$.

based on the Hellinger distance by the feature vector transform given by

$$\tilde{\boldsymbol{p}} = \sqrt{\frac{\boldsymbol{p}}{2}}, \tag{15}$$

where the square root of a vector denotes the elementwise operation.

5. Experimental Results

In this section, we first show the experimental results on artificial two-dimensional data, and then show the results on a real document-term dataset.

5.1 Artificial Two-Dimensional Data Fig. 1 shows the graphs of $\rho(x) = x^2$ and $\rho^R(x)$ in (5), where the horizontal and vertical axes denote x and $\rho(x)$ or $\rho^R(x)$, respectively. The blue line denotes $\rho(x) = x^2$ which is not robust to outliers because the value of $\rho(x)$ increases rapidly with the increase in $|x|$, which means that the data being far apart from a point of attention have a big impact on results. On the other hand, the red lines denote $\rho^R(x)$ in (5) for $b = 1, 5$ and 10 with solid, broken and dashed dotted lines, respectively. These red lines have gentler slopes than the blue line; the slopes do not exceed 45° as described in Property 2. Additionally, $\rho^R(x)$ is under $\rho(x)$ if $b \geq 1/2$ by Property 3.

Fig. 2 shows an artificial two-dimensional (2-D) dataset which consists of three main clusters on the left side and three outlying points on the right side. Fig. 3 shows a zoomed part in Fig. 2 including the three clusters, where three centroids given by conventional K-means clustering algorithm with $K = 3$ are denoted by red circles, one of which at the upper right deviates toward the outliers on the right side. Fig. 4 shows the same region as Fig. 3 and three centroids given by the proposed robust K-means clustering algorithm with $b = 1/2$, where the deviation of the upper right centroid toward the outliers is alleviated compared with Fig. 3. The condition for halting the algorithm is that

$$\max_k \left\{ \max \left(\left| \boldsymbol{\mu}_k^{(t+1)} - \boldsymbol{\mu}_k^{(t)} \right| \right) \right\} \leq \epsilon, \tag{16}$$

where the superscript t on $\boldsymbol{\mu}_k$ denotes the number of iterations of the updating procedure, i.e., $t = 0, 1, 2, \ldots$, and ϵ is

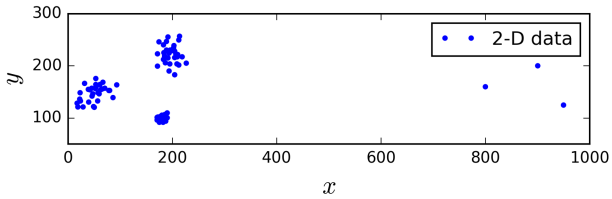

Figure 2: Two-dimensional data.

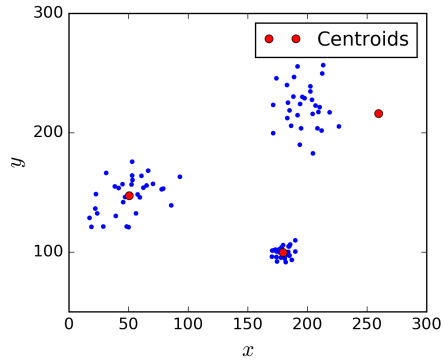

Figure 3: Centroids given by conventional K-means.

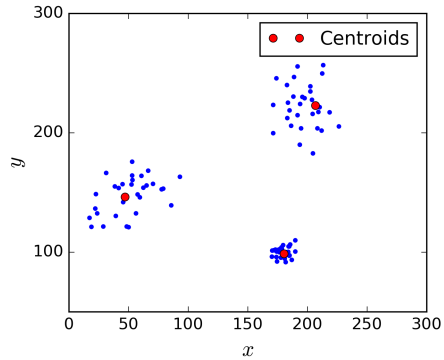

Figure 4: Centroids given by robust K-means.

a positive constant which is set to $\epsilon = 10^{-6}$ in this example. All centroids $\boldsymbol{\mu}_k$ for $k = 1, 2, \ldots, K$ are randomly initialized and used commonly in both the conventional and proposed algorithms.

Table 1 shows the numerical data of the coordinates of the centroids and the numbers of iterations for compared algorithms. The first and second rows in the table show the (x, y) coordinates of the centroid of the upper right cluster in Fig. 3 or 4, from the left to right columns, for outlier-free result (denoted by 'No outliers'), conventional K-means (Conv.), and the proposed robust K-means for $b = 1$, $1/2$ and $1/10$, respectively. The x-coordinate for the conventional K-means is quite different from that for the outlier-free result. On the other hand, the proposed method for every value of b achieved nearer x-coordinates to 'No outliers' than the conventional K-means. Better y-coordinates are given when $b = 1/2$ and $1/10$ in the proposed method. The

Table 1: Coordinates of the Centroid of a Cluster in Fig. 2 and the Numbers of Iterations.

	No outliers	Conv.	$b = 1$	$b = 1/2$	$b = 1/10$
x	197.3723	259.7324	**206.7222**	**206.7222**	**206.7222**
y	221.5227	216.0813	222.9668	**222.9667**	**222.9667**
Iter.	-	**2**	**10**	**10**	11

Table 2: Matching matrix by K-means.

	athletics	cricket	football	rugby	tennis	total
athletics	94	0	6	0	1	101
cricket	0	12	112	0	0	124
football	3	0	253	2	7	265
rugby	0	2	73	72	0	147
tennis	2	0	68	0	30	100

third row in the table shows the number of iterations (Iter.), where the conventional K-means had the smallest number of 2. In the proposed method, $b = 1$ and $1/2$ had smaller iteration number of 10 than 11 for $b = 1/10$. These results suggest that $b = 1/2$ is a good choice of the value of b in the proposed method because it gives better (x, y) coordinates with a smaller number of iterations than other settings of b.

5.2 Real Document-Term Data Next, we show the results on the BBC Datasets [16], from which we selected the BBCSport dataset which is a collection of 737 documents from the BBC Sport website corresponding to sports news articles in five topical areas with 4613 terms, i.e., the size of the document-term matrix A is $n \times d = 737 \times 4613$, and the five topics are athletics, cricket, football, rugby and tennis. We evaluate the performance of clustering algorithms with matching matrices (or confusion matrices) [17].

Table 2 shows a matching matrix given by conventional K-means, where the row vectors of the document-term matrix are directly used for the feature vectors of documents. The initial centroids are selected from the documents as $\boldsymbol{\mu}_1 = \boldsymbol{a}_{15}$, $\boldsymbol{\mu}_2 = \boldsymbol{a}_{168}$, $\boldsymbol{\mu}_3 = \boldsymbol{a}_{325}$, $\boldsymbol{\mu}_4 = \boldsymbol{a}_{513}$ and $\boldsymbol{\mu}_5 = \boldsymbol{a}_{713}$, where $\boldsymbol{a}_i$ denotes the ith row vector of A, and commonly used in the following experiments. The top row of athletics shows that the actual documents on athletics are divided into three clusters, athletics, football and tennis. The rightmost column in the table shows the total number of documents in each category. In the matching matrix, if the

Table 3: Matching matrix by robust K-means ($b = 1/2$).

	athletics	cricket	football	rugby	tennis	total
athletics	98	0	3	0	0	101
cricket	0	73	51	0	0	124
football	0	0	251	8	6	265
rugby	0	0	64	83	0	147
tennis	2	0	60	0	38	100

Table 4: Matching matrix by K-means with feature transform.

	athletics	cricket	football	rugby	tennis	total
athletics	101	0	0	0	0	101
cricket	2	119	2	1	0	124
football	2	0	204	3	56	265
rugby	0	1	12	131	3	147
tennis	4	0	6	0	90	100

Table 5: Matching matrix by robust K-means ($b = 1/2$) with feature transform.

	athletics	cricket	football	rugby	tennis	total
athletics	101	0	0	0	0	101
cricket	2	119	2	1	0	124
football	2	0	204	3	56	265
rugby	0	0	12	133	2	147
tennis	4	0	6	0	90	100

Table 6: Averaged table of confusion by K-means.

True Positive 92.2	False Negative 55.2
False Positive 55.2	True Negative 534.4

Table 7: Averaged table of confusion by robust K-means ($b = 1/2$).

True Positive 108.6	False Negative 38.8
False Positive 38.8	True Negative 550.8

Table 8: Averaged table of confusion by K-means with feature transform.

True Positive 129	False Negative 18.4
False Positive 18.4	True Negative 571.2

Table 9: Averaged table of confusion by robust K-means ($b = 1/2$) with feature transform.

True Positive 129.4	False Negative 18
False Positive 18	True Negative 571.6

diagonal elements are larger and the off-diagonal elements are closer to zero, then the obtained clusters show a better match with the true categories of documents.

Table 3 shows the matching matrix given by the proposed robust K-means with $b = 1/2$. The diagonal elements in Table 3 are larger than that in Table 2 except for the central element corresponding to the football category.

Tables 4 and 5 show the matching matrices given by K-means and robust K-means with the proposed feature transform described in Section 4, respectively. In both methods, the performance is improved by the feature transform.

Tables 6 to 9 show the averaged tables of confusion [18] computed from Tables 2 to 5, respectively. In these tables, the larger diagonal elements (True Positive (TP) and True Negative (TN)) and the smaller off-diagonal elements (False Negative (FN) and False Positive (FP)) are, the better the performance is.

Finally, Table 10 summarizes the F-measures for compared methods, where an F-measure is the harmonic mean of precision and recall given by

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{17}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{18}$$

That is, the F-measure is given by

$$F = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{19}$$

In Table 10, the proposed robust K-means with feature transform achieves the highest value of the F-measure among the compared methods.

Table 10: F-measures.

K-means	Robust K-means	K-means (Feature transform)	Robust K-means (Feature transform)
0.626	0.737	0.875	**0.878**

6. Conclusion

In this paper, we proposed a robust K-means clustering algorithm for document clustering, where a document-term matrix is given as an input dataset. Initially, each document is expressed by the corresponding row vector in the document-term matrix. To improve the performance of document clustering, we also proposed a feature transform method based on the Hellinger distance between two probability distributions to have better feature vectors than the row vectors of the document-term matrix. Experimental results showed that the proposed method improved the robustness to outliers and the performance of document clustering compared with conventional K-means.

Future work will include the development of the fuzzy versions of the proposed robust K-means for improved document clustering.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP16H03019.

References

[1] N. O. Andrews, E. A. Fox, "Recent Developments in Document Clustering," Technical Report TR-07-35, Computer Science, Virginia Tech., 2007.

[2] N. Shah, S. Mahajan, "Semantic based Document Clustering: A Detailed Review," *International Journal of Computer Applications*, Vol. 52, No. 5, pp.42-52, 2012. DOI: 10.5120/8202-1598

[3] N. Shah, S. Mahajan, "Document Clustering: A Detailed Review," *International Journal of Applied Information Systems*, Vol. 4, No. 5, pp.30-38, 2012. DOI: 10.5120/ijais12-450691

[4] R. C. Balabantaray, C. Sarma, M. Jha, "Document Clustering using K-Means and K-Medoids," *International Journal of Knowledge Based Computer Systems*, Vol. 1, No. 1, 2013.

[5] I. S. Dhillon, D. S. Modha, "Concept Decompositions for Large Sparse Text Data Using Clustering," *Machine Learning*, Vol. 42, No. 1-2, pp.143-175, 2001. DOI: https://doi.org/10.1023/A:1007612920971

[6] F. Dzogang, C. Marsala, M. Lesot, M. Rifqi, "An Ellipsoidal K-Means for Document Clustering," *International Conference on DataMining*, 2012. DOI: 10.1109/ICDM.2012.126

[7] J.-P. Mei, Y. Wang, "Hyperspherical Fuzzy clustering for online document categorization," *IEEE International Conference on Fuzzy Systems*, 2016. DOI: 10.1109/FUZZ-IEEE.2016.7737866

[8] P. J. Huber, "Robust Statistics", *Wiley*, 1981. DOI: 10.1002/0471725250

[9] K. Lange, "Convergence of EM image reconstruction algorithms with Gibbs smoothing," *IEEE Trans. Medical Imaging*, Vol. 9, No. 4, pp.439-446, 1990. DOI: 10.1109/42.61759

[10] J. T. Barron, "A More General Robust Loss Function," Cornell University Library, arXiv:1701.03077v3 [cs.CV], 2017.

[11] A. B. Hamza and D. J. Brady, "Reconstruction of reflectance spectra using robust nonnegative matrix factorization," *IEEE Trans. Signal Process.*, Vol. 54, No. 9, pp.3637-3642, 2006. DOI: 10.1109/TSP.2006.879282

[12] P. Charbonnier, L. Blanc-Feraud, G. Aubert, M. Barlaud, "Two deterministic half-quadratic regularization algorithms for computed imaging," *Proceedings of 1st International Conference on Image Processing*, 1994. DOI: 10.1109/ICIP.1994.413553

[13] P. Harsha, "Hellinger distance," *Lecture notes on Communication Complexity*, 2011.

[14] Q. V. Bui, K. Sayadi, M. Bui, "A multi-criteria document clustering method based on topic modeling and pseudoclosure function," *Infomatica*, Vol.40, No.2, pp.169-180, 2016.

[15] D. M. Blei, A. Y. Ng, M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, Vol.3, pp.993-1022, 2003.

[16] D. Greene, P. Cunningham, "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering," *Proceedings of the 23rd International Conference on Machine Learning*, pp.377-384, 2006. DOI: 10.1145/1143844.1143892 `http://mlg.ucd.ie/datasets/bbc.html`

[17] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, Vol.27, No.8, pp.861-874, 2006. DOI: 10.1016/j.patrec.2005.10.010

[18] Confusion matrix, Wikipedia. `https://en.wikipedia.org/wiki/Confusion_matrix`

Hengjun Yu (Member) He received B.Eng. degree from University of Electronic Science and Technology of China in 2014. He is currently a graduate student in Kyushu University. His research interests include image processing and gesture recognition.

Kohei Inoue (Member) He received B.Des., M.Des. and D.Eng. degrees from Kyushu Institute of Design in 1996, 1998 and 2000, respectively. He is currently an Associate Professor in Kyushu University. His research interests include pattern recognition and image processing.

Kenji Hara (Non-member) He received the BE and ME degrees from Kyoto University in 1987 and 1989, respectively, and the PhD degree from Kyushu University in 1999. He is currently an Associate Professor in Kyushu University. His research interests include physics-based vision and geometric modeling.

Kiichi Urahama (Member) He received M.Eng. and D.Eng. degrees from Kyushu University in 1976 and 1980. From 1980 to 1995 he was an Associate Professor in Kyushu Institute of Technology. He is now a Professor in Kyushu University. His research interests include pattern recognition, image processing and computer graphics.