2016 Joint 8th International Conference on Soft Computing and Intelligent Systems and 2016 17th International Symposium on Advanced Intelligent Systems

On Kernelized Sequential Hard Clustering

Yukihiro Hamasuna
School of Sci. and Eng.,
Kindai University
3-4-1 Kowakae, Higashi-osaka,
Osaka 577-8502, Japan
Phone: +81-6-4307-3518;
Fax: +81-6-6730-5880;
Email: yhama@info.kindai.ac.jp

Yasunori Endo
Faculty of Eng., Info. and Sys.,
University of Tsukuba,
1-1-1 Tennodai, Tsukuba,
Ibaraki 305-8573, Japan
Phone: +81-29-853-5135;
Fax: +81-29-853-5135;
email: endo@risk.tsukuba.ac.jp

***Abstract*—A method of sequential clustering extracts a cluster sequentially without determining the number of clusters. The sequential hard clustering is based on noise clustering and one of the typical sequential clustering methods. A kernelized sequential hard clustering is proposed by introducing the kernel method to sequential hard clustering to handle datasets which consists non-linear clusters and execute robust clustering. The performance of the proposed method is evaluated with a typical dataset which consists non-linear cluster boundary. Negative results are obtained through numerical examples and those show that the proposed method can not extract non-linear clusters sequentially.**

***Index Terms*—sequential cluster extraction, sequential hard c-means, noise clustering, kernel function.**

1. Introduction

Clustering is one of the data analysis methods that divides a set of objects into several clusters. Objects classified in the same cluster are considered similar, while those in different clusters are considered dissimilar. Hard c-means (HCM) which is known as k-means [9], variants of fuzzy c-means (FCM) [1], [11], are the most well-known and basic clustering methods. Various clustering methods are widely studied and paid much attention to handle large scale complex data sets to extract value from them in recent years.

Typical clustering methods such as HCM and FCM need to determine the number of clusters in advance. On the other hand, sequential clustering methods extract a cluster sequentially without determining the number of clusters. The word *sequential cluster extraction* means that an algorithm extracts 'one cluster at a time'. Various sequential clustering algorithms have been proposed and discussed to execute robust clustering [4], [7], [8], [10], [14]. Determining the suitable parameter is required in these sequential clustering algorithms instead of the number of clusters. A method of noise clustering [2], [3] based sequential clustering algorithms [4], [8], [10] uses a noise parameter which is the dissimilarity between an object and the pseudo-representative of the noise cluster. The noise clustering based algorithm divides a set of objects into two clusters by using the noise parameter. One is an extracting cluster, and the other is a noise cluster. The larger the noise parameter, the larger the size or area of the extracting cluster.

The noise parameter is considered as a radius of extracting cluster. Then a hyper-spherical cluster is extracted by using noise parameter based algorithms [4], [8]. Kernel method is actively used in pattern analysis to handle complex datasets and extract important properties [13]. In the field of clustering, kernel function is well-known as the significant method to handle complex datasets which consists non-linear cluster boundary [5], [6], [11]. The dissimilarity between object and cluster center is considered in high dimensional feature space in the kernelized clustering method. The sequential cluster extraction is quite important to handle complex dataset and execute robust clustering as above mentioned. It is, however, unclear the effect and usefulness of kernel method in noise clustering based sequential clustering method. We consider that the kernelized sequential clustering methods are quite important to handle massive and complex datasets with robustness. We will first consider the kernelized sequential hard clustering by introducing kernel function to conventional sequential clustering method. We will show the negative results by the proposed method through simple dataset with non-linear cluster boundary.

This paper is organized as follows: In section 2, we introduce symbols, hard c-means clustering, and sequential clustering methods. In section 3, we propose kernelized sequential hard clustering by introducing the kernel function. In section 4, we show numerical examples to evaluate the performance of the proposed method. In section 5, we conclude this paper.

2. Preliminaries

A set of objects to be clustered is given and denoted by $X = \{x_1, \dots, x_n\}$ in which $x_k\,(k = 1, \dots, n)$ is an object. In most cases, each object x_k is a vector in p-dimensional

978-1-5090-2678-4/16 $31.00 © 2016 IEEE
DOI 10.1109/SCIS&ISIS.2016.165

Euclidean space $\Re^p$, that is, an object $x_k \in \Re^p$. A cluster is denoted by G_i and a collection of clusters is given by $\mathcal{G} = \{G_1, \dots, G_c\}$. A cluster center of G_i is denoted by $v_i \in \Re^p$ and a set of v_i is given by $V = \{v_1, \dots, v_c\}$. A membership degree of x_k belonging to G_i and a partition matrix is denoted as u_{ki} and $U = (u_{ki})_{k=1\sim n,\ i=1\sim c}$.

2.1. Sequential Hard c-Means

Sequential hard c-means (SHCM) is based on the noise clustering [2] and known as a typical sequential clustering method [4]. The objective function of SHCM J_{sh} is as follows:

$$J_{sh}(U,V) = \sum_{k=1}^{n} u_{k1}\|x_k - v_1\|^2 + \sum_{k=1}^{n} u_{k0} D.$$

$D > 0$ is a noise parameter. Constraint on membership degree $\mathcal{U}_{sh}$ is as follows:

$$\mathcal{U}_{sh} = \left\{ (u_{ki}) : u_{ki} \in \{0,1\}\,,\ \sum_{i=0}^{1} u_{ki} = 1,\ ^{\forall} k \right\}. \quad (1)$$

u_{k1} is a membership degree for extracting cluster and u_{k0} is one for noise cluster.

Optimal solutions for v_1 and u_{ki} are as follows:

$$v_1 = \frac{\sum_{k=1}^{n} u_{k1} x_k}{\sum_{k=1}^{n} u_{ki}}, \quad (2)$$
$$u_{ki} = \begin{cases} i & \left(\|x_k - v_1\|^2 \le D\right) \\ 1-i & \text{(otherwise)} \end{cases} \quad (i = 0, 1)\,.$$

SHCM divides a set of objects into two clusters by comparing $\|x_k - v_1\|^2$ to D. One is an extracting cluster satisfying $u_{k1} = 1$ and the other is a noise cluster satisfying $u_{k0} = 1$.

The algorithm of SHCM is summarized as follows:

Algorithm 1 SHCM

SHCM 1	Give initial values v_1 and set parameter D.
SHCM 2	Repeat updating u_{ki} and v_1 until convergence.
SHCM 3	Extract $\{x_k \mid u_{k1} = 1\}$ from X.
SHCM 4	If $X = \emptyset$, stop. Otherwise, go back to **SHCM1**.

The size of extracting cluster is depended on D in each iteration, so a different D can be given or updating method is proposed in [2]. We then extract different sizes of clusters by determining the suitable parameter.

2.2. New Sequential Hard c-Means

We have proposed new sequential hard c-means (νSHCM) by considering parameter D to be a variable to be optimized by introducing new parameter ν instead of D [7]. The objective function of νSHCM is as follows:

$$J_{\nu sh}(U,V,D) = \sum_{k=1}^{n} u_{k1}\|x_k - v_1\|^2 + \sum_{k=1}^{n} u_{k0} D - \nu D \quad (3)$$

The constraint on u_{ki} is the same as (1). The difference between SHCM and νSHCM is that noise parameter D is a given parameter in SHCM, whereas it is a variable to be optimized in νSHCM. D is considered the radius of extracted cluster in the conventional method. Without the third term in (3), $J_{\nu sh}(U,V,D)$ is minimized with $D = 0$ and all $u_{k0} = 1$. By introducing the third term, $J_{\nu sh}(U,V,D)$ is minimized with $D > 0$. Parameter ν increases the value of D, which affects the size of an extracted cluster.

Considering the necessary condition for variable D. From $\frac{\partial J_{\nu sh}}{\partial D} = 0$, we obtain,

$$\nu = \sum_{k=1}^{n} u_{k0}. \quad (4)$$

ν is the number of objects classified into noise cluster, i.e., νSHCM classifies $(n-\nu)$ objects into extracting cluster and ν objects into noise cluster.

Substituting (4) into (3), we obtained simplified objective function $J'_{\nu sh}(U,V)$ as follows:

$$J'_{\nu sh}(U,V) = \sum_{k=1}^{n} u_{k1}\|x_k - v_1\|^2.$$

This objective function should be minimized under the constraint (1) and condition (4). The optimal solution of v_1 is the same as for (2).

Considering the constraint (1) and condition (4), the optimal solution to u_{ki} is obtained by sorting dissimilarity $d_{k1} = \|x_k - v_1\|^2$ in descending order. At the head of $(n-\nu)$ objects take $u_{k1} = 1$ and the rest of ν objects take $u_{k0} = 1$.

The algorithm of νSHCM is constructed based on the above discussions as follows:

Algorithm 2 νSHCM

νSHCM 1	Give initial value v_1 and set parameter ν.
νSHCM 2	Repeat updating u_{ki} and v_1 until convergence.
νSHCM 3	Extract $\{x_k \mid u_{k1} = 1\}$ from X.
νSHCM 4	If $X = \emptyset$, stop. Otherwise, go back to **νSHCM1**.

The value of ν depends on the number of extracting objects in each iteration, so a different ν can be given in each extracting iteration. We then extract different sizes of clusters by adjusting the value of ν. νSHCM is a variant of SHCM.

3. Kernelized SHCM

We propose a kernelized sequential hard c-means (KSHCM) by introducing kernel function to SHCM. We have proposed the kernelized νSHCM (KνSHCM) as the same procedure [7]. νSHCM and KνSHCM are strongly depended on the parameter ν, that is, $n-\nu$ objects with small dissimilarity are classified into the same cluster. Then, the shape of clusters and the cluster boundary between extracting cluster and noise cluster is less considered in νSHCM and KνSHCM. We apply the kernel method to SHCM to consider non-linear cluster boundary and structure.

First, we define symbols to introduce kernel functions. $\phi : \Re^p \to \Re^s (p \ll s)$ means mapping from input space $\Re^p$ to high dimensional feature space $\Re^s$. An object in feature space is denoted by $\phi(x_k) \in \Re^s$. Kernel function $K : \Re^p \times \Re^p \to \Re$ satisfies the following relation:

$$K(x, y) = \langle \phi(x), \phi(y) \rangle.$$

Note that ϕ is not explicit. The inner product of object in high dimensional feature space is calculated easily by using kernel function K and known as kernel trick [5], [6], [11].

The objective function of KSHCM is as follows:

$$J_{ksh}(U, W) = \sum_{k=1}^{n} u_{k1} \|\phi(x_k) - W_1\|^2 + \sum_{k=1}^{n} u_{k0} D.$$

W_1 is the cluster center in $\Re^s$. The optimal solution for W_1 is derived as follows:

$$W_1 = \frac{\sum_{k=1}^{n} u_{k1} \phi(x_k)}{\sum_{k=1}^{n} u_{k1}}. \tag{5}$$

This formula cannot be used directly because $\phi(x_k)$ is implicit. To calculate the dissimilarity in the high dimensional feature space, substitute (5) into $\|\phi(x_k) - W_1\|^2$ then d_{k1} is calculated by the following formula:

$$\begin{aligned} d_{k1} =& \|\phi(x_k) - W_1\|^2 \\ =& \langle \phi(x_k), \phi(x_k) \rangle - 2\langle \phi(x_k), W_1 \rangle + \langle W_1, W_1 \rangle \\ =& K(x_k, x_k) - \frac{2}{U_1} \sum_{s=1}^{n} u_{s1} K(x_k, x_s) \\ &+ \frac{1}{(U_1)^2} \sum_{s=1}^{n} \sum_{t=1}^{n} u_{s1} u_{t1} K(x_s, x_t), \end{aligned}$$

where,

$$U_1 = \sum_{s=1}^{n} u_{s1} = |G_1|.$$

The following gaussian kernel is a typical kernel function in kernel data analysis:

$$K(x, y) = \exp\left(-\beta \|x - y\|^2\right). \tag{6}$$

β is a kernel parameter for the gaussian kernel. The algorithm of KSHCM is constructed from updating u_{ki} and d_{k1}. Based on the above discussions, the algorithm of KSHCM is summarized as follows:

Initial dissimilarity in **KSHCM 1** is calculated by selecting an object x_k from X in each iteration.

Algorithm 3 KSHCM

KSHCM 1	Give initial dissimilarity d_{k1}, D, and β.
KSHCM 2	Repeat updating u_{ki} and d_{k1} until convergence.
KSHCM 3	Extract $\{x_k \mid u_{k1} = 1\}$ from X.
KSHCM 4	If $X = \emptyset$, stop. Otherwise, go back to **KSHCM 1**.

4. Experiments

We show numerical examples with simple artificial dataset to show the performance of KSHCM. First, we describe calculation conditions. Second, we show clustering results by KSHCM. Third, we summarize the features of the proposed method.

We use one typical dataset which consists non-linear clusters. Fig. 1 is an illustrative example of artificial dataset classified into adequate clusters. This dataset consists 150 objects and is classified into two clusters. It is well known that HCM and FCM with kernel method can classify this dataset into adequate two clusters (inside ball and outside circle) [11].

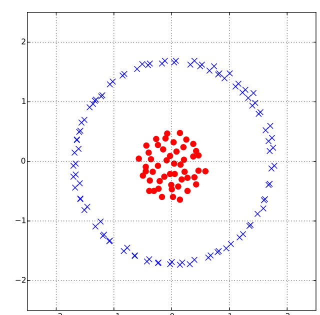

Figure 1. Artificial data 1($n = 150$, $p = 2$, $c = 2$).

We use the gaussian kernel (6) as a kernel function. Then, the dissimilarity is described as follows:

$$\begin{aligned} d_{k1} =& \|\phi(x_k) - W_1\|^2 \\ =& 1 - \frac{2}{|G_1|} \sum_{s=1}^{n} u_{s1} \exp\left(-\beta \|x_k - x_s\|^2\right) \\ &+ \frac{1}{|G_1|^2} \sum_{s=1}^{n} \sum_{t=1}^{n} u_{s1} u_{t1} \exp\left(-\beta \|x_s - x_t\|^2\right). \end{aligned}$$

We show experimental results to show the performance of the proposed method. The following Figs. 2–4 are the illustrative examples of clustering results.

In these figures, the value displayed at each object means the cluster number. Figs. 2–4 are all negative results in the viewpoint of extracting non-linear cluster structure. The proposed method KSHCM is strongly depended on the parameter D. Fig. 2 is better result in our experiments.

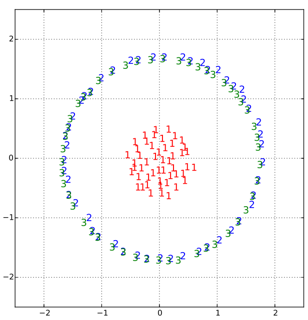

Figure 2. Result 1: Theree clusters are obtained ($D = 0.9$, $\beta = 1.0$).

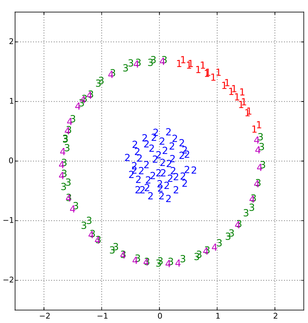

Figure 3. Result 2: Four clusters are obtained ($D = 0.8$, $\beta = 1.0$).

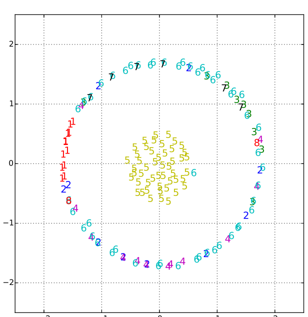

Figure 4. Result 3: Eight clusters are obtained ($D = 0.5$, $\beta = 1.0$).

Although we set several parameters D and β, we can not obtain suitable classified result. It is considered that several different approaches such as determining the suitable parameter, applying different kernel function, and visualization by principal component analysis are required to obtain better results and verify the properties of KSHCM.

5. Conclusions

In this paper, we introduce kernel function into SHCM and construct a new algorithm of KSHCM. we expect that KSHCM extracts non-linear clusters sequentially. However, it is not achieved and the effectiveness of KSHCM is not verified.

In future works, we will conduct numerical experiments widely by using other kernel functions with various datasets. We will visualize cluster structure in each iteration by using kernel principal component analysis to determine the suitable parameters and consider the cluster structure.

Acknowledgment

This work was partly supported by JSPS KAKENHI Grant Numbers JP26330271, JP16K16128.

References

[1] Bezdek J. C., 'Pattern Recognition with Fuzzy Objective Function Algorithms', Plenum Press, New York, 1981.

[2] Davé R. N., Krishnapuram R., Robust clustering methods : A unified view, *IEEE Transactions on Fuzzy Systems*, Vol. 5, No. 2, pp. 270–293, 1997.

[3] Davé R. N., Characterization and detection of noise in clustering, *Pattern Recognition Letters*, Vol. 12, No. 11, pp. 657–664, 1991.

[4] Endo Y., Taniguchi A., and HamasunaY., Hard c-Means using Quadratic Penalty-Vector Regularization for Uncertain Data, *Journal of Advanced Computational Intelligence and Intelligent Informatics* (JACIII), Vol. 16, No. 7, pp. 831–840, 2012.

[5] Endo Y., Haruyama H., Okubo T., On some hierarchical clustering algorithms using kernel functions, *Proc. of IEEE International Conference on Fuzzy Systems*(FUZZ-IEEE2004), pp. 1513–1518, 2004.

[6] Girolami M., Mercer kernel-based clustering in feature space, *IEEE Transactions Neural networks*, Vol. 13, No. 3, pp. 780–784, 2002.

[7] Hamasuna Y. and Endo Y., On A Family of New Sequential Hard c-Means, *Journal of Advanced Computational Intelligence and Intelligent Informatics* (JACIII), Vol. 19, No. 6, pp. 759–765, 2015.

[8] Hamasuna Y. and Endo Y., On Sequential Cluster Extraction Based on L_1-Regularized Possibilistic c-Means, *Journal of Advanced Computational Intelligence and Intelligent Informatics* (JACIII), Vol. 19, No. 5, pp. 655–661, 2015.

[9] Jain A. K., Data clustering: 50 years beyond K-means, *Pattern Recognition Letters*, Vol. 31, No. 8, pp. 651–666, 2010.

[10] Miyamoto S., Arai K., Different Sequential Clustering Algorithms and Sequential Regression Models, *Proc. of 2009 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2009)*, pp. 1107–1112, 2009.

[11] Miyamoto S., Ichihashi H., and Honda K., 'Algorithms for Fuzzy Clustering', Springer, Heidelberg, 2008.

[12] Miyamoto S., Kuroda Y., Arai K., Algorithms for Sequential Extraction of Clusters by Possibilistic Method and Comparison with Mountain Clustering, *Journal of Advanced Computational Intelligence and Intelligent Informatics* (JACIII), Vol. 12, No. 5, pp. 448–453, 2008.

[13] Vapnik V. N., 'The nature of Statistical Learning Theory', Wiley, New York, 1998.

[14] Yager R. R., Filev D., " Approximate clustering via the mountain method ", *IEEE Transactions, on Systems, Man, and Cybernetics*, Vol.24, No.8, pp. 1279–1284, 1994.