2010 International Symposium on Nonlinear Theory and its Applications
NOLTA2010, Krakow, Poland, September 5-8, 2010

Hard and Fuzzy c-Means Clustering Algorithms with Geodesic Dissimilarity

Yuchi Kanzawa†, Yasunori Endo‡ and Sadaaki Miyamoto‡

†Shibaura Institute of Technology, Japan
Email: kanzawa@sic.shibaura-it.ac.jp
‡University of Tsukuba, Japan

Abstract—In this paper, the geodesic distance is applied to relational clustering methods. First, it is shown that conventional methods are based on respective three types of relational clustering algorithms among nine ones, and the six rests of the nine ones with the geodesic distance are proposed. Second, geodesic dissimilarity is proposed by assigning the power of the Euclidean distance to the weight of the neighborhood graph of data. Numerical examples show that the proposed geodesic-dissimilarity-based relational clustering algorithms successfully cluster the data that conventional squared-Euclidean-distance-based ones cannot.

1. Introduction

Fuzzy c-means (FCM) [1] is a well-known fuzzy clustering method that is derived from hard c-means (HCM), also called as k-means. Among the many FCM variants proposed thus far, one is the FCM algorithm based on the concept of regularization by entropy [2]. This algorithm is called entropy regularized FCM (eFCM) and is discussed not only because of its usefulness but also because of its mathematical relationships with other techniques. We call the FCM proposed in [1] standard FCM (sFCM) in order to distinguish it from eFCM.

For HCM, sFCM, and eFCM, similar clustering models for relational data can be developed. It should be noted that neither object data nor cluster centers are available in relational clustering. Hence, the distance between data points and cluster centers that appear in HCM, sFCM, and eFCM cannot explicitly be computed. There are two methods to overcome this problem: one is to restrict the solution space and the other is to implicitly compute the object data and cluster centers. The former method is called hard c-medoids (HCMdd), standard fuzzy c-medoids (sFCMdd) [3], or entropy regularized fuzzy c-medoids (eFCMdd), respectively, based on HCM, sFCM, or eFCM. The latter is called relational HCM (RFCM), standard relational FCM (sRFCM) [4], or entropy regularized relational FCM (eRFCM) [5], respectively. The kernelization of HCM, sFCM, and eFCM, called kernel HCM (K-HCM) [6], kernel sFCM (K-sFCM), and kernel eFCM (K-eFCM) [7], can be also applied to relational data if a dissimilarity-based kernel, for example, a Gaussian kernel, is used.

The correct identification of clusters depends on the definition of dissimilarity. The choice of the dissimilarity measure determines the cluster shape, and therefore, it determines the success of a clustering algorithm on the specific application domain. As one such choice, the geodesic distance has been applied to sFCMdd [8], K-eFCM [9], and RHCM [10]. One of our two objectives is to apply the geodesic distance to other six types of relational clustering methods such as HCMdd, eFCMdd, K-HCM, K-sFCM, sRFCM, and eRFCM.

The geodesic distance is computed as the total weight of the shortest weighted path on the neighborhood graph of a data set, where we have the degree of freedom, that is, the number or maximal distance of a neighborhood, and the weight of the edge. While the Euclidean distance is usually assigned as the weight of the edge as in [8] and [9], a density scaling is used in [10]. In this paper, we consider another weight, the power of Euclidean distance, typically the squared-Euclidean distance. The considered measure is no longer the geodesic distance but geodesic dissimilarity because it does not satisfy the triangular inequality. The other of our two objectives is to apply this geodesic dissimilarity to nine types of relational clustering methods such as HCMdd, sFCMdd, eFCMdd, K-HCM, K-sFCM, K-eFCM, RHCM, sRFCM, and eRFCM.

The remainder of this paper is organized as follows. In the second section, we define some notations, and introduce some relational clustering algorithms and the concept of geodesic distance; these are used in our proposed methods. In the third section, we propose applying the geodesic distance to six types of relational clustering methods, and also propose new geodesic dissimilarity that can be applied to nine types of relational clustering algorithms. In the fourth section, we present some numerical examples. In the last section, we conclude this paper.

2. Preliminaries

2.1. Fuzzy Clustering

In this subsection, we introduce nine conventional methods of relational clustering; these are used in our proposed methods that is described in the next section. The introduced methods are classified according to the three original clustering algorithms — HCM, sFCM, or eFCM — or according to how the relational data is used — by using medoids, by transforming the optimization problem, or by using dissimilarity-based kernel function.

For a given data set $x = \{x_i \mid i \in \{1, \ldots, N\}\}$, HCMdd, sFCMdd, eFCMdd, RHCM, sRFCM, and eRFCM assume that the dissimilarity data matrix $R \in \mathbb{R}^{N \times N}$ is given, and K-HCM, K-sFCM, and K-eFCM assume that the kernel matrix $K \in \mathbb{R}^{N \times N}$ is given. The membership by which x_i belongs to the j-th cluster is denoted by $u_{i,j}$ ($i \in \{1, \cdots, N\}, j \in \{1, \cdots, C\}$) and the set of $u_{i,j}$ is denoted by $u \in \mathbb{R}^{N \times C}$; this is called the partition matrix.

Hard c-means (HCM) is the algorithm obtained by solving the following optimization problem:

$$\underset{u,v}{\text{minimize}}\, J_{\mathsf{HCM}}(u, v) \tag{1}$$

$$\text{subject to} \sum_{j=1}^{C} u_{i,j} = 1, \tag{2}$$

where

$$J_{\mathsf{HCM}}(u, v) = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{i,j} d_{i,j} \tag{3}$$

$$d_{i,j} = \|x_i - v_j\|_2^2. \tag{4}$$

Standard fuzzy c-means (sFCM) is the algorithm obtained by solving the following optimization problem:

$$\underset{u,v}{\text{minimize}} \sum_{i=1}^{N}\sum_{j=1}^{C} u_{i,j}^{m} d_{i,j} \tag{5}$$

subject to Eq. (2) and (4). Entropy regularized fuzzy c-means (eFCM) is the algorithm obtained by solving the following optimization problem:

$$\underset{u,v}{\text{minimize}}\, J_{\mathsf{HCM}} + \lambda^{-1}\sum_{i=1}^{N}\sum_{j=1}^{C} u_{i,j}\log(u_{i,j}) \tag{6}$$

subject to Eq. (2) and (4).

Hard c-medoids (HCMdd) is the algorithm obtained by solving the optimization problem (1) subject to Eq. (2), (4), and

$$v_j \in x. \tag{7}$$

Standard fuzzy c-medoids (sFCMdd) is the algorithm obtained by solving the optimization problem (5) subject to Eq. (2), (4), and (7). Entropy regularized fuzzy c-medoids (eFCMdd) is the algorithm obtained by solving the optimization problem (6) subject to Eq. (2), (4), and (7).

Kernel hard c-means (K-HCM) is the algorithm obtained by solving optimization problem (1) subject to Eq. (2) and

$$d_{i,j} = \|\Phi(x_i) - W_j\|_{\mathbb{H}}. \tag{8}$$

Kernel standard fuzzy c-means (K-sFCM) is the algorithm obtained by solving optimization problem (5) subject to Eq. (2) and (8). Kernel entropy regularized fuzzy c-means (K-eFCM) is the algorithm obtained by solving optimization problem (6) subject to Eq. (2) and (8).

Relational hard c-means (RHCM) is the algorithm obtained by solving the following optimization problem:

$$\underset{u}{\text{minimize}}\, J_{\mathsf{RHCM}}, \tag{9}$$

where

$$J_{\mathsf{RHCM}}(u) = \sum_{j=1}^{C}\sum_{i=1}^{N}\sum_{k=1}^{N} u_{i,j}u_{k,j}r_{i,k} \Big/ \left(2\sum_{t=1}^{N} u_{t,j}\right), \tag{10}$$

subject to Eq. (2), and $r_{i,k}$ is given. Standard relational fuzzy c-means (sRFCM) is the algorithm obtained by solving the following optimization problem:

$$\underset{u}{\text{minimize}} \sum_{j=1}^{C}\sum_{i=1}^{N}\sum_{k=1}^{N} u_{i,j}^{m}u_{k,j}^{m}r_{i,k} \Big/ \left(2\sum_{t=1}^{N} u_{t,j}^{m}\right) \tag{11}$$

subject to Eq. (2), and $r_{i,k}$ is given. Entropy regularized relational fuzzy c-means (eRFCM) is the algorithm obtained by solving the following optimization problem:

$$\underset{u}{\text{minimize}}\, J_{\mathsf{RHCM}}(u) + \lambda^{-1}\sum_{i=1}^{N}\sum_{j=1}^{C} u_{i,j}\log(u_{i,j}) \tag{12}$$

subject to Eq. (2), and $r_{i,k}$ is given.

HCMdd, sFCMdd, eFCMdd, RHCM, sRFCM, eRFCM, K-HCM, K-sFCM, and K-eFCM are given by the following algorithm.

Algorithm 1

STEP 1. *Set C; set m for sFCMdd, K-sFCM, and sRFCM; set λ for eFCMdd, K-eFCM, and eRFCM; set r for RHCM, sRFCM, and eRFCM; set K for K-HCM, K-sFCM, and K-eFCM; and set u.*

STEP 2. *Calculate*

$$v_j = arg\ \min_k \sum_{i=1}^{N} u_{i,j}d_{i,k} \tag{13}$$

for HCMdd and eFCMdd;

$$v_j = arg\ \min_k \sum_{i=1}^{N} u_{i,j}^{m}d_{i,k} \tag{14}$$

for sFCMdd;

$$v_j = \sum_{i=1}^{N} u_{i,j}e_i \Big/ \left(\sum_{i=1}^{N} u_{i,j}\right) \tag{15}$$

for RHCM, eRFCM, K-HCM, and K-eFCM; and

$$v_j = \sum_{i=1}^{N} u_{i,j}^{m}e_i \Big/ \left(\sum_{i=1}^{N} u_{i,j}^{m}\right) \tag{16}$$

for sRFCM and K-sFCM.

STEP 3. *Calculate the membership*

$$u_{i,j} = \begin{cases} 1 & (j = arg\ \min\{d_{i,k}\}), \\ 0 & (otherwise) \end{cases} \tag{17}$$

for HCMdd;

$$u_{i,j} = 1 \Big/ \sum_{k=1}^{C} \left(\frac{d_{i,j}}{d_{i,k}}\right)^{1/(m-1)} \tag{18}$$

for sFCMdd;

$$u_{i,j} = \frac{\exp(-\lambda d_{i,j})}{\sum_{k=1}^{C}\exp(-\lambda d_{i,k})} \tag{19}$$

for eFCMdd;

$$u_{i,j} = 1 \Big/ \sum_{k=1}^{C} \left(\frac{(e_i - W_j)^{\mathsf{T}}K(e_i - W_j)}{(e_i - W_k)^{\mathsf{T}}K(e_i - W_k)}\right)^{1/(m-1)} \tag{20}$$

for K-sFCM;

$$u_{i,j} = 1 \Big/ \sum_{k=1}^{C} \left(\frac{(RW_j)_i - W_j^{\mathsf{T}}RW_j}{(RW_k)_i - W_k^{\mathsf{T}}W_k}\right)^{1/(m-1)} \tag{21}$$

for sRFCM;

$$u_{i,j} = \frac{\exp(-\lambda(e_i - W_j)^{\mathsf{T}}K(e_i - W_j))}{\sum_{k=1}^{C}\exp(-\lambda(e_i - W_k)^{\mathsf{T}}K(e_i - W_k))} \tag{22}$$

for K-eFCM;

$$u_{i,j} = \frac{\exp(-\lambda((RW_j)_i - W_j^{\mathsf{T}}RW_j))}{\sum_{k=1}^{C}\exp(-\lambda((RW_k)_i - W_j^{\mathsf{T}}RW_k))} \tag{23}$$

for eRFCM.

STEP 4. *If (u, v) is convergent, terminate this algorithm. Otherwise, return to* STEP 2.

2.2. Geodesic Distance

In this subsection, the two types of geodesic distance, k-geodesic distance and ε-geodesic distance, are introduced.

The ε-neighborhood of a point $x \in X$ is defined as $N_\varepsilon(x) = \{z \in X \mid \|x - z\|_2 \le \varepsilon\}$. The k-neighborhood of a point $x \in X$ is the set of k closest points to x in the ℓ_2 norm sense: $N_k(x) \subset X$ such that $|N_k(x)| = k$ and $\max_{z\in N_k(x)} \|x - z\|_2 \le \min_{z\in Z\setminus N_k(x)} \|x - z\|_2$. The k-neighborhood graph of X is an undirected graph whose the vertices are X and whose the edges $(x_i, x_{\tilde{i}})$ exist if $x_i \in N_k(x_{\tilde{i}})$ or $x_{\tilde{i}} \in N_k(x_i)$. The ε-neighborhood graph of X is an undirected graph whose the vertices are X and whose the edges $(x_i, x_{\tilde{i}})$ exist if $x_i \in N_\varepsilon(x_{\tilde{i}})$ or $x_{\tilde{i}} \in N_\varepsilon(x_i)$. Assume a symmetric matrix $D \in \mathbb{R}^{N\times N}$ of non-negative weights on the edges of G_k. The k-geodesic distance from any point x to any point $\tilde{x}$ is defined as the

total weight of the shortest weighted path from x to $\tilde{x}$ on G_k, which is denoted by $\delta_{G_k,D}(x,\tilde{x})$. The ε-geodesic distance from any point x to any point $\tilde{x}$ is defined as the total weight of the shortest weighted path from x to $\tilde{x}$ on G_ε, which is denoted by $\delta_{G_\varepsilon,D}(x,\tilde{x})$.

3. Proposed Method

One of our two objectives in this paper is to show that conventional geodesic-distance-based clustering methods are based on respective three types of relational clustering algorithms among nine ones, and to propose the six rests of the nine ones with the geodesic distance. The other is to propose a new dissimilarity based on the geodesic distance, and to apply conventional relational clustering methods, as described in the second section.

3.1. Six types of Geodesic-Distance-based Clustering

The geodesic distance has been applied to some relational clustering algorithms [8]–[10]. sFCMdd in [8], K-eFCM in [9], and RHCM in [10] were used with the geodesic distance. The survey of relational clustering methods with the geodesic distance is summarized in Table 1. The symbol "–" in this table indicates that we could not find relevant literatures. Therefore, we propose these methods by applying the geodesic distance, that is, HCMdd, eFCMdd, K-HCM, K-sFCM, sRFCM, and eRFCM (Algorithm 1).

Table 1: Conventional Geodesic-Distance-based Clustering Methods

	Medoids	Relational	Kernel
HCM	–	[10]	–
sFCM	[8]	–	–
eFCM	–	–	[9]

3.2. New Dissimilarity based on Geodesic Distance

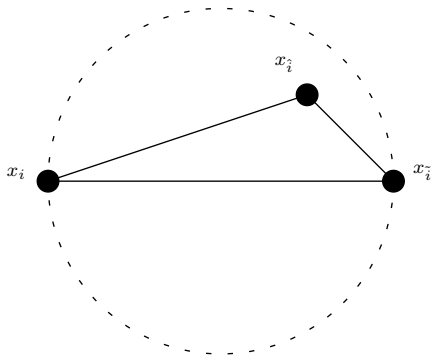

Figure 1: $\delta_{G_\infty,D^{(2)}}(x_i,x_{\tilde{i}})$ is less than the squared-Euclidean distance between them if there exists a point $x_{\hat{x}}$ in the sphere with center $(x_i+x_{\tilde{i}})/2$ and radius $\|x_i-x_{\tilde{i}}\|_2/2$, as shown in Fig. 1

The geodesic distance is computed as the total weight of the shortest weighted path on the neighborhood graph of a data set, where we have the degree of freedom, that is, the number or the maximal distance of the neighborhood, and the weight of the edge. While the Euclidean distance is usually assigned as the weight of the edge in [8] and [9] as

$$D_{i,\tilde{i}} = \|x_i - x_{\tilde{i}}\|_2, \tag{24}$$

a density scaling is used in [10].

In this paper, we propose another weight, the power of Euclidean distance:

$$D^{(q)}_{i,\tilde{i}} = \|x_i - x_{\tilde{i}}\|_2^q \tag{25}$$

with a parameter q. If $q = 2$, the weight is the squared-Euclidean distance

$$D^{(2)}_{i,\tilde{i}} = \|x_i - x_{\tilde{i}}\|_2^2. \tag{26}$$

The proposed measure is no longer the geodesic distance but geodesic dissimilarity because it does not satisfy the triangular inequality. The proposed geodesic dissimilarity $\delta_{G_\infty,D}$ has the following properties. $\delta_{G_\infty,D^{(2)}}(x_i,x_{\tilde{i}})$ is less than the squared-Euclidean distance between them if there exists at least one point in the sphere with center $(x_i+x_{\tilde{i}})/2$ and radius $\|x_i-x_{\tilde{i}}\|_2/2$, as shown in Fig. 1. This sphere, the border being where $\delta_{G_\infty,D^{(q)}}(x_i,x_{\tilde{i}})$ is less than $\|x_i-x_{\tilde{i}}\|_2^q$, is inflated in the direction orthogonal to the segment x_i–$x_{\tilde{i}}$ with $q > 2$, deflated with $q < 2$, and degenerate with $q \le 1$, as shown in Fig. 2. $\delta_{G_\infty,D^{(q)}}(x_i,x_{\tilde{i}})$ is minimal if the point $x_{\hat{i}}$ is at the midst between x_i and $x_{\tilde{i}}$. Furthermore, if there exist s points on the segment x_i–$x_{\tilde{i}}$, as shown in Fig. 3, $\delta_{G_\infty,D^{(q)}}(x_i,x_{\tilde{i}})$ tends to zero as $s \to \infty$. We propose this geodesic dissimilarity for use in clustering algorithms (Algorithm 1).

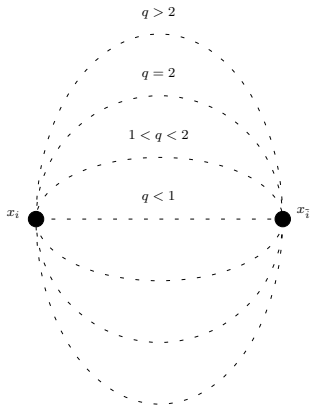

Figure 2: This sphere, the border being where $\delta_{G_\infty,D^{(q)}}(x_i,x_{\tilde{i}})$ is less than $\|x_i-x_{\tilde{i}}\|_2^q$, is inflated in the direction orthogonal to the segment x_i–$x_{\tilde{i}}$ with $q > 2$, deflated with $q < 2$, and degenerate with $q \le 1$.

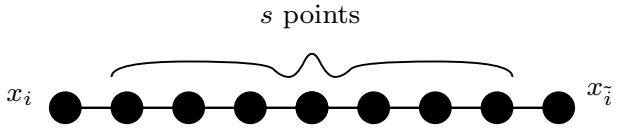

Figure 3: $\delta_{G_\infty,D}(x_i,x_{\tilde{i}})$ tends to zero as $s \to \infty$ if there exist s points on the segment x_i–$x_{\tilde{i}}$.

4. Numerical Example

In this section, we show some examples of clustering using our proposed methods. The fuzzifier parameters are fixed as $m = 2$ in sFCM-based methods and $\lambda = 1/2$ in eFCM-based methods. As a kernel function, a Gaussian kernel is selected with the kernel parameter $\sigma^2 = 0.002$. In each example, 100 trials for the proposed algorithm with randomly different initializations are tested and the solu-

tion with the minimal objective function value is selected as the final result. We consider clustering the data shown in Fig. 4 into two moon-shaped clusters. This data set is constructed using 300 elements in the two-dimensional Euclidean space. First, we see that all algorithms (Algorithm 1) with the squared-Euclidean distance as the dissimilarity fail to cluster correctly; the result of sRFCM-based methods is shown in Fig. 5, where the plus symbol indicates one cluster and the cross symbol indicates the other. This figure indicates that the each inside edges of the moons are mis-clustered with each other.

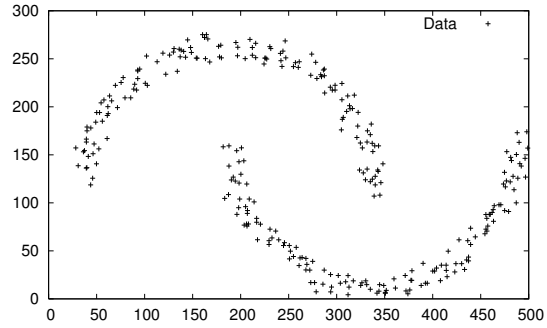

Figure 4: Data

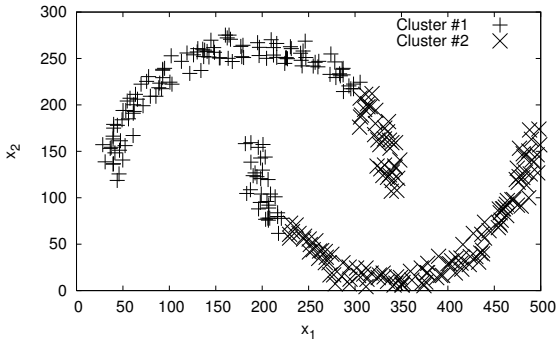

Figure 5: Mis-clustering result by Algorithm 1 with the squared-Euclidean distance

Next, we use the proposed geodesic dissimilarity with $q = 2$ instead of the squared-Euclidean distance in Algorithm 1 and we obtain the desired clustering results, as shown in Fig. 6. Thus, this example shows that the proposed geodesic-dissimilarity-based algorithms achieve the successful clustering results for the data for which the conventional squared-Euclidean-distance-based algorithms fail.

5. Conclusion

In this paper, we considered applying the geodesic distance to clustering methods. First, we showed that conventional methods were based on respective three types of relational clustering algorithms among nine ones, and we proposed the six rests of the nine ones with the geodesic distance. Second, we proposed the geodesic dissimilarity by assigning the power of the Euclidean distance to the weight of the neighborhood graph of data. Through numerical examples, we found that the proposed geodesic-dissimilarity-based relational clustering algorithms achieve successful clustering results for the data for which the conventional squared-Euclidean-distance-based algorithms fail. In future works, (1) we intend to investigate the property of the parameter in the proposed geodesic dissimilarity, and (2) test the differences between clustering features based on the differences among relational clustering algorithms.

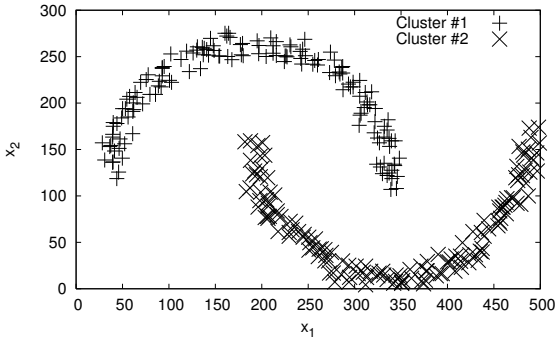

Figure 6: Successful clustering Result

References

[1] Bezdek, J.P.: 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] Krishnapuram, R., Joshi, A, Nasraoui, O. and Yi, L.: "Low-complexity Fuzzy Relational Clustering Algorithm for Web Mining", IEEE Transactions on Fuzzy Systems, Vol.9, No.4, pp.595–607, 2001.

[4] 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.

[5] Filippone, M.: "Dealing with Non-metric Dissimilarities in Fuzzy Central Clustering Algorithms", Int. J. of Approximate Reasoning, Vol.40, No.2, pp.363–384, 2009.

[6] Miyamoto, S. and Nakayama, Y.: "Algorithms of Hard c-Means Clustering Using Kernel Functions in Support Vector Machines", JACIII, Vol.7, No.1, pp.19–24, 2003.

[7] 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.

[8] Feil, B. and Abonyi, J.: "Geodesic Distance Based Fuzzy Clustering", Lecture Notes in Computer Science, Soft Computing in Industrial Applications, Vol.39, pp.50–59, 2007.

[9] Kim, J., Shim, K.-H. and Choi, S.: "Soft Geodesic Kernel K-means", Proc. ICASSP2007, Vol.2, pp.429–432, 2007.

[10] Asgharbeygi, N. and Maleki, A.: "Geodesic K-means Clustering", Proc. ICPR08, pp.1–4, 2008.