Hard and Fuzzy \( c \)-Medoids for Asymmetric Networks

Yousuke Kaizu\(^1\) Sadaaki Miyamoto\(^2\) Yasunori Endo\(^2\)

\(^1\)Master’s Program in Risk Engineering, University of Tsukuba, Japan
\(^2\)Department of Risk Engineering, University of Tsukuba, Japan

Abstract

Medoid clustering frequently gives better results than those of the \( K \)-means clustering in the sense that a unique object is the representative element of a cluster. Moreover the method of medoids can be applied to nonmetric cases such as weighted graphs that arise in analyzing SNS(Social Networking Service) networks. A general problem in clustering is that asymmetric measures of similarity or dissimilarity are difficult to handle, while relations are asymmetric, e.g., in SNS user groups. In this paper we consider hard and fuzzy \( c \)-medoids for asymmetric graphs in which a cluster has two different centers with outgoing directions and incoming directions. This method is applied to a small illustrative example and real data of a Twitter user network.

Keywords: fuzzy \( c \)-medoids, asymmetric dissimilarity, SNS.

1. Introduction

Clustering [3, 2, 8] is becoming a major tool in data mining with applications to SNS (Social Networking Service) analysis [4]. Two features should be noted in analyzing such networks: first, the basic space is not Euclidean, i.e., an inner product is not defined. Second, asymmetric relations in networks should frequently be analyzed. These two features induce problems in applying standard methods of clustering. Concretely, most clustering techniques are based on symmetric dissimilarity measures and many important methods assume Euclidean spaces. There are two approaches to overcome such problems. First way is to transform an asymmetric relation into symmetric one, and then use a positive-definite kernel to introduce an Euclidean space [14, 12]. Second way is to design a new method in order to handle asymmetric data, which we adopt in this paper.

\( K \)-Medoids [5], which minimize the summation of dissimilarity between the medoid and other points in a cluster, and fuzzy \( c \)-medoid clustering [6] provide a natural idea when non-Euclidean space is given. Medoid clustering is frequently appropriate in the sense that a unique object is the representative element of a cluster. In this paper we extend fuzzy \( c \)-medoids to asymmetric weighted graphs by identifying two centers in a cluster.

The rest of this paper is organized as follows. Section 2 proposes the formulation and algorithms of hard and fuzzy \( c \)-medoids clustering to asymmetric graphs. A parameter is introduced to distinguish three options of handling medoids with outgoing directions and incoming directions. Section 3 shows numerical examples of a small data set for illustration purpose and also a larger data set of a real Twitter network. Finally, Section 4 concludes the paper.

2. \( K \)-medoids and Fuzzy \( c \)-medoids for Asymmetric Networks

We begin with notations. Assume that \((X, d)\) is given in which \(X = \{x_1, x_2, \ldots, x_N\}\) is a set of objects for clustering;

\[
d: X \times X \rightarrow [0, +\infty)
\]

is a dissimilarity measure which is asymmetric, i.e., \(d(x, y) \neq d(y, x)\) in general. We assume also \(d(x, x) = 0\) for simplicity. A cluster is denoted by \(G_i\) (\(i = 1, \ldots, c\)). For hard clusters, \(G_i\) forms a partition of \(X\):

\[
\bigcup_{i=1}^{c} G_i = X, \quad G_i \cap G_j = \emptyset (i \neq j).
\]

In the case of fuzzy clusters, the above relations do not hold but we assume

\[
\sum_{i=1}^{c} \mu_{G_i}(x_k) = 1, \quad \forall x_k \in X
\]

instead.

2.1. Basic \( K \)-medoids clustering

Let us suppose \(d(x, y)\) is symmetric for the moment. To apply \(K\)-means, the squared Euclidean distance between arbitrary vectors in a space has to be calculated. However, \(K\)-medoids can classify data if we can calculate dissimilarity between an arbitrary pair of objects. Thus \(K\)-medoids can be applied to a data set that forms nodes of a network with dissimilarity on edges. In \(K\)-medoids, a cluster center is not a centroid but a representative point in the cluster: a cluster center is thus given by the following:

\[
v_i = \arg \min_{x \in G_i} \sum_{y \in G_i} d(x, y), \quad (1)
\]
where $G_1$ is a crisp cluster.

$K$-means minimize the summation of the squared Euclidean distance between the centroid of a cluster and points in the cluster. In contrast, $K$-medoids minimize the summation of dissimilarity between the medoid and points in the cluster.

A basic $K$-medoid clustering can be described as the alternate optimization [2, 9]:

$$J_1(U, V) = \sum_{j=1}^{c} \sum_{k=1}^{N} u_{kj} d(x_k, v_j)$$

with the constraint on $U$:

$$M_U = \{ U = (u_{ki}) : \sum_{j=1}^{c} u_{kj} = 1, \forall k; u_{kj} \geq 0, \forall k, j \}$$

and another constraint on $V = (v_1, \ldots, v_c)$:

$$M_V = \{ V = (v_1, \ldots, v_c) : v_j \in X, \forall j \}.$$

The alternate optimization algorithm is as follows:

**$K$-medoid algorithm.**

**KMED1:** Give an initial value for $\bar{V}$.

**KMED2:** Fix $\bar{V}$ and find

$$\bar{U} = \arg \min_{U \in M_U} J_1(U, \bar{V}).$$

**KMED3:** Fix $\bar{U}$ and find

$$\bar{V} = \arg \min_{V \in M_V} J_1(\bar{U}, V).$$

**KMED4:** If the solution $(\bar{U}, \bar{V})$ is convergent, stop. Else go to KMED2.

End KMED.

The main difference of this algorithm from that of $K$-means is that $V \in M_V$ is imposed.

We immediately have

$$\bar{u}_{ki} = 1 \iff i = \arg \min_{1 \leq j \leq c} d(x_k, \bar{v}_j),$$

where $\bar{v}_j$ is given by (1). The optimal solutions $\bar{u}_{ki}$ and $\bar{v}_i$ are also written as $u_{ki}$ and $v_i$ for simplicity without confusion.

### 2.2. Fuzzy $c$-medoids for asymmetric measures

Let us assume that $d(x, y)$ is asymmetric. We introduce a new measure having three variables and a parameter $\alpha \in [0, 1]$:

$$D(x, v, w; \alpha) = \alpha d(x, v) + (1 - \alpha)d(w, x).$$

We moreover assume that either $\alpha = 0$, $\alpha = 1$, or $\alpha = \frac{1}{2}$ for simplicity.

We proceed to describe fuzzy $c$-medoids for asymmetric measures. For this purpose the following objective function is considered:

$$J(U, V, W) = \sum_{j=1}^{c} \sum_{k=1}^{N} (u_{ki})^m D(x_k, v_i, w_i, \alpha), \quad m > 1,$$

where $V = (v_1, \ldots, v_c)$ and $W = (w_1, \ldots, w_c)$.

Since $J(U, V, W)$ has three variables, the following alternate optimization is used:

**Asymmetric fuzzy $c$-medoid algorithm.**

**AFCMED1:** Give an initial value for $\bar{V}$ and $\bar{W}$.

**AFCMED2:** Fix $\bar{V}$, $\bar{W}$ and find

$$\bar{U} = \arg \min_{U \in M_U} J(U, \bar{V}, \bar{W}).$$

**AFCMED3:** Fix $\bar{U}$, $\bar{W}$ and find

$$\bar{V} = \arg \min_{V \in M_V} J(\bar{U}, V, \bar{W}).$$

**AFCMED4:** Fix $\bar{U}$, $\bar{V}$ and find

$$\bar{W} = \arg \min_{W \in M_W} J(\bar{U}, \bar{V}, W).$$

**AFCMED5:** If the solution $(\bar{U}, \bar{V}, \bar{W})$ is convergent, stop. Else go to AFCMED2.

End AFCMED.

Note that when $\alpha = 1$, $W$ is not used, and if $\alpha = 0$, $V$ is not used. In these cases, respective steps of AFCMED should be skipped.

The optimal solution in AFCMED2 is the following:

$$u_{ki} = \left[ \sum_{j=1}^{c} \frac{D(x_k, v_j, w_i; \alpha)^{-\frac{1}{m-1}}}{\sum_{j=1}^{c} D(x_k, v_j, w_i; \alpha)^{-\frac{1}{m-1}}} \right]^{-1},$$

(2)

$$u_{ki} = 1, \quad (x_k \neq v_i \text{ or } x_k \neq w_i)$$

(3)

while the solutions for $V$ and $W$ are respectively given as follows:

$$v_i = \arg \min_{z_i \in X} \sum_{k=1}^{N} (u_{ki})^m D(x_k, z_i, w_i)$$

(4)

$$w_i = \arg \min_{y_i \in X} \sum_{k=1}^{N} (u_{ki})^m D(x_k, v_i, y_i)$$

(5)

### 2.3. Theoretical properties

We can prove the following theoretical properties of the solutions of fuzzy $c$-medoids. First property is almost trivial and the proof is omitted.
Proposition 1 Cluster centers \( v_i \) and \( w_j \) given respectively by (4) and (5) satisfy the following:

\[
\begin{align*}
    v_i &= \arg \min_{z_j \in X} \sum_{k=1}^{N} (u_{ik})^m d(x_k, z_j), \\
    w_i &= \arg \min_{y_j \in X} \sum_{k=1}^{N} (u_{jk})^m d(y_j, x_k).
\end{align*}
\]

Second property is on the convergence of the algorithm. The convergence criterion in AFCMED5 is roughly written in terms of \((\bar{U}, \bar{V}, \bar{W})\). We can also use the value of objective function, i.e., we stop the algorithm when the objective function value is not decreased. Note that the objective function value is monotonically nonincreasing.

Proposition 2 Suppose we stop the algorithm when the objective function value is not decreased. Then algorithm AFCMED necessarily stops, and the upper bound of the number of iterations is \((N_c)^2\).

The proof is easy when we observe that choice for all combinations of \((v_i, w_i) \in X \times X\) is finite, and the objective function is monotone nonincreasing. However, \((N_c)^2\) is generally huge and hence it gives an unrealistic upper bound.

Third property is on the membership value \( u_{ki} \).

We suppose an object \( x_k \) is "movable" to the infinity in the sense that \( D(x_k, v_i; w_i; \alpha) \to \infty \) for all \( 1 \leq i \leq c \).

Proposition 3 Suppose \( x_k = v_i = w_i \). When \( \alpha = 1 \), \( x_k = v_i \) and when \( \alpha = 0 \), \( x_k = w_i \). We then have

\[
    u_{ki} = \max_{1 \leq i \leq N} u_{ki} = 1.
\]

Suppose \( x_k \) moves to the infinity in the sense that \( D(x_k, v_i; w_i) \to \infty \) for all \( 1 \leq i \leq c \). Then we have

\[
    u_{ki} \to \frac{1}{c}, \quad \forall 1 \leq i \leq c.
\]

The proof is not difficult when we observe the form of \( u_{ki} \) in (2). If \( x_k = v_i = w_i \), then \( D(x_k, v_i, w_i; \alpha) = 0 \) and \( u_{ki} \) takes its maximum value of \( u_{ki} = 1 \). If \( D(x_k, v_i, w_i; \alpha) \to \infty \), then it is easy to see \( u_{ki} \to \frac{1}{c} \).

2.4. Hard \( c \)-medoids

The function \( J(U, V, W) \) can be used for \( m = 1 \) to derive solutions for \( K \)-medoids alias hard \( c \)-medoids for asymmetric dissimilarity measures. The solutions are reduced to the following.

\[
    u_{ki} = \begin{cases} 
    1, & i = \arg \min_{1 \leq j \leq c} D(x_k, v_j, w_j) \\
    0, & \text{otherwise}
    \end{cases}
\]

\[
    v_i = \arg \min_{z_j \in X} \sum_{x_k \in G_i} d(x_k, z_j),
\]

\[
    w_i = \arg \min_{y_j \in X} \sum_{x_k \in G_i} d(y_j, x_k).
\]

Proposition 2 holds also for hard \( c \)-medoids, since the argument is the same as that for fuzzy \( c \)-medoids.

3. Examples

3.1. A small example of travelers among countries

We show a small example for illustrating how the algorithm works. In this example, The data of foreign traveler in major 19 countries in 2001 from World Tourism Organization [16] are used. The countries are South Africa, America, Canada, China, Taiwan, Hong Kong, Korea, Japan, India, Indonesia, Singapore, Australia, New Zealand, England, France, Switzerland, Italy, Thailand, and Malaysia. The number of travelers \( n_{ij} \) from country \( x_i \) to country \( x_j \) is given and normalized to similarity

\[
    s(x_i, x_j) = \frac{n_{ij}}{\sum_l n_{il}}.
\]

Then \( s(x_i, x_j) \) is transformed to dissimilarity

\[
    d(x_i, x_j) = 1 - s(x_i, x_j),
\]

except that \( d(x_i, x_i) \) is set to zero for all \( x_i \).

Figures 1, 2, and 3 respectively show graphs with two, three, and four clusters. Clusters are distinguished by different colors and we used the hard \( c \)-medoids with \( m = 1 \) and \( \alpha = \frac{1}{2} \) for two medoids in a cluster. In these figures, some edges are thicker and others are thinner. Where an edge has higher similarity, the edge is shown by a thicker arrow. The two centers \( v_i \) and \( w_i \) in each cluster are expressed by a pentagram and a hexagram. Thus a hexagram means outer-direction medoid and a pentagram means inner-direction medoid. We used Gephi [17], an interactive visualization software for graph data.

More details about clusters are as follows.

Two clusters:

\[
\{ \text{ China, Taiwan, Hong Kong, Korea } \},
\]

with medoids: \( v = \text{ China, w = Korea } \);

\{ \text{ South Africa, America, Canada, Japan, India, Indonesia, Singapore, Australia, New Zealand, England, France, Switzerland, Italy, Thailand, Malaysia } \},

with medoids: \( v = \text{ America, w = Australia } \).

Three clusters:

\[
\{ \text{ China, Taiwan, Hong Kong, Korea } \},
\]

with medoids: \( v = \text{ China, w = Korea } \);

\{ \text{ France, Switzerland, Italy } \},

with medoids: \( v = \text{ France, w = Switzerland } \);

\{ \text{ South Africa, America, Canada, Japan, India, Indonesia, Singapore, Australia, New Zealand, England, Thailand, Malaysia } \},

with medoids: \( v = \text{ America, w = Australia } \).

Four clusters:

\[
\{ \text{ China, Taiwan, Hong Kong, Korea, Japan } \},
\]

with medoids: \( v = \text{ China, w = Korea } \);

\{ \text{ France, Switzerland, Italy } \},
with medoids: $v = \text{France}$, $w = \text{Switzerland}$.

{America, Canada},

with medoids: $v = w = \text{Canada}$.

{ South Africa, India, Indonesia, Singapore, Australia, New Zealand, England, Thailand, Malaysia };

with medoids: $v = \text{Singapore}$, $w = \text{Australia}$.

Thus the cluster {China, Taiwan, Hong Kong, Korea} is stable in the three figures except that Japan is included in the case of four clusters. Other large cluster in Figure 1 are subdivided into two and three clusters in the next two figures, reflecting geometrical nearness.

Figure 1: classification result of nineteen countries to two clusters. Clusters are distinguished by different colors. Pentagram implies $v_i$ and hexagram means $w_i$.

Figure 2: Classification result of nineteen countries to three clusters. Clusters are distinguished by different colors. Pentagram implies $v_i$ and hexagram means $w_i$.

3.2 Twitter user network

A real Twitter user network of Japanese political parties was used of which the data have been taken by the authors. There are five parties in these tweets: Liberal Democratic Party of Japan, Democratic Party of Japan, Japanese Communist Party, New Komeito, and Social Democratic Party. The data are network among 1,745 users of these parties held in the form of asymmetric adjacency matrix $A = (a_{ij})$ which consists of 0 and 1: user $i$ following $j$ is represented by $a_{ij} = 1$. The matrix is transformed into asymmetric dissimilarity $D = (d_{ij})$ using $S = (s_{ij})$ as follows:

$$S = A + \frac{1}{2} A^2,$$
$$d_{ij} = 1 - \frac{1}{\max_{k,l} s_{k,l}} s_{ij}, \forall i, j,$$
$$d_{ii} = 0, \forall i.$$

We tested three algorithms of the proposed method using $v_i$ and $w_i$ ($\alpha = \frac{1}{2}$), the method using $v_i$ alone ($\alpha = 1$), and the method using $w_i$ alone ($\alpha = 0$). All the three methods are with $m = 1$ (hard c-medoids). One hundred trials with different random initial values are made and the Rand Index (RI) [11] was calculated against the right party belongingness. Table 1 summarizes the results. The user groups are well-separated and all the three methods give rather good RI values. In particular, the proposed method using both $v_i$ and $w_i$ gives the best results among these three methods.

4. Conclusion

The formulation and algorithms of hard and fuzzy c-medoids for asymmetric dissimilarity measures have
Figure 3: Classification result of nineteen countries to four clusters. Clusters are distinguished by different colors. Pentagram implies $v_i$ and hexagram means $w_i$. A ‘blue’ cluster is with $v_i = w_i$ at Canada.

Table 1: The maximum and the average of the Rand indexes (RI) for three methods: the proposed method using $v_i$ and $w_i$, the method using $v_i$ alone, and the method using $w_i$ alone.

<table>
<thead>
<tr>
<th>Method</th>
<th>Max RI</th>
<th>Ave. RI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_i$ and $w_i$</td>
<td>0.95</td>
<td>0.9</td>
</tr>
<tr>
<td>$v_i$ alone</td>
<td>0.89</td>
<td>0.8</td>
</tr>
<tr>
<td>$w_i$ alone</td>
<td>0.94</td>
<td>0.88</td>
</tr>
</tbody>
</table>

been proposed and tested using numerical examples. The method includes three options of $\alpha = 0$, 1, and $\frac{1}{2}$. The values of $\alpha = 0$, or 1 mean that a cluster has only one medoid of $v_i$ or $w_i$, while $\alpha = \frac{1}{2}$ implies that a cluster should have two medoids of $v_i$ and $w_i$.

A fundamental problem is that which of $\alpha$ should be adopted. This problem has no general solution and dependent on an application domain. It hence needs further investigation in a specific application such as SNS networks.

Another problem is that larger computation is needed than $K$-means. For such problems, we need to consider multistage clustering (e.g., [13]) whereby we can effectively reduce computation. Moreover an algorithm of $k$-medoid++ should be considered which is a variation of $k$-means++ [1].

As other future studies, the present method should be compared with other existing methods (e.g.,[10, 15]) that can handle asymmetric measures of dissimilarity using large-scale real examples. Moreover cluster validity criteria (e.g.,[2]) should be developed for medoid clustering.

Acknowledgment

This study has partly been supported by the Grant-in-Aid for Scientific Research, JSPS, Japan, no.26330270.

References

Figure 5: Classification result of nineteen countries to four clusters using $w_i$ alone. Clusters are distinguished by different colors. Hexagram implies $w_i$.

[17] https://gephi.org/