Fuzzy Set-based Distant Cluster Identification

KEON MYUNG LEE, SUN A LEE
Department of Computer Science, College of Electrical and Computer Engineering
Chungbuk National University, PT-ERC
Cheongju, Chungbuk
REPUBLIC OF KOREA
kmlee@cbnu.ac.kr

Abstract: This presents a method to identify distant clusters for a data set using fuzzy techniques. The proposed method first applies a fuzzy clustering algorithm to the data set. It introduces a fuzzy membership function called **distance membership function** which transforms metric distance between two objects into the degree of farness. In order to measure the degree of farness between two fuzzy clusters, it defines the **cluster farness measures** which incorporate distance membership function and fuzzy clusters. It uses the measures to construct a fuzzy distant relation for clusters from which the distant clusters are identified.

Key-Words: distant cluster identification, clustering, fuzzy clustering, fuzzy set theory, data analysis

1 Introduction
Clustering is a typical data analysis technique to partition a collection of data into clusters in a way to maximize within-cluster similarities and to minimize between-cluster similarities. Clustering has been employed for the purpose of identifying natural structures embedded in data set or picking up a subset of data for further analysis.[3] It has been widely studied in various fields such as pattern recognition, statistical data analysis, data mining.[3,8] Clustering studies have been mainly focused on cluster formation, determination of the number of clusters, quality evaluation of clustering results.[3-8]

In this study we have been concerned with how to find pairs of cluster which could be considered as distant ones. It is of interest to determine pairs of distant clusters in some applications: For example, in microarray data analysis, distant clusters of samples or genes could be some candidates for further analysis to determine possible markers.

The metric distances between clusters give unbounded distance values and thus it is not convenient to directly use them for determining pairs of distant clusters each other. In some situations, it would be more useful to give a degree ranged on the domain [0,1] rather than to give a value on the unbounded range like metric distance, when we compare the clusters. In addition, it is desirable to take into account the extent of overlapping of clusters under consideration. This paper proposes a method to provide for clusters the degree of farness which considers both distance and overlapping between clusters using fuzzy set techniques. The proposed method can be used to select some subset of data which show contrasting characteristics against a specific cluster, or to find groups of distant clusters over the entire data set.

2 Related Works
In this study we have interest in some measures to give a degree value on [0,1] which tells how they are distant each other. To this notion, the distance and dissimilarity measures are related in some sense. Although distance and dissimilarity are interchangeably used in some situations, here we distinguish them each other. A distance measure gives a unbounded distance value, yet a dissimilarity measure tells a value bounded on [0,1]. The dissimilarity between clusters may consider either their degree of overlapping or their distance, or both. When all attributes of objects take values from continuous numeric domains, the distance between two objects could be calculated by various measures such as Euclidian distance, Manhattan distance, Minkowski distance, and cosine distance. If attributes are categorical or discrete, specialized measures are employed like Jaccard’s coefficient, Kullback-Leibler divergence.[9]

The distance measures between clusters have been studied in agglomerative hierarchical clustering algorithms. The algorithms take an approach to merge clusters based on the distance between them. Each agglomerative hierarchical clustering algorithms are usually classified into average linkage clustering, complete linkage clustering, and single linkage clustering.

In average linkage algorithms, the distance between clusters is determined by calculating the average distance between all pairs of objects in the two different clusters. In complete linkage clustering, the distance between two clusters is defined as the greatest distance between a member of one cluster and a member of the other cluster. In single linkage clustering, the distance of two clusters is measured as the minimum distance between members of...
the two clusters. All these distance measures between clusters give a unbounded value.

In this study we are concerned with a sort of dissimilarity measures for clusters of which data are expressed in a multidimensional numeric space. The proposed measures are called cluster farness measures since they take into both account distance between clusters and how much they overlap each other. Using the measures, we want to identify the distant clusters in a set of data.

3 Distant Cluster Identification

For the convenience of description, the following notations are used:

\[ D = \{d_1, d_2, \ldots, d_n\} \] : a data set of size \( n \) for which distant clusters are searched

\[ d_i = (d_{i1}, d_{i2}, \ldots, d_{ip}) \] : the \( i \) th data with \( p \) attribute values

\[ C = \{C_1, C_2, \ldots, C_m\} \] : a set of fuzzy clusters constructed over \( D \)

\[ C_i = \{(c_{i1}, \mu_{i1}), (c_{i2}, \mu_{i2}), \ldots, (c_{ip}, \mu_{ip})\} \] : the \( i \) th cluster of \( C \) where \( \mu_{ij} \) is the membership degree of data \( c_{ij} \) to cluster \( C_i \)

\[ O_i = (o_{i1}, o_{i2}, \ldots, o_{ip}) \] : the center of cluster \( C_i \)

\[ \|d_i - d_j\| \] : distance between data \( d_i \) ad \( d_j \) evaluated by a distance measure such as Euclidean distance, Manhattan distance

\[ LS = \{\alpha_i | i = 1, \ldots, q \quad 0 \leq \alpha_1 < \alpha_2 < \cdots < \alpha_q \leq 1\} \] : a level set with which \( \alpha \)-cut sets are defined

\[ IS_k(C_i) = \{c_{ij} | \alpha_k \leq \mu_{ij} < \alpha_{k+1}\} \] : the subset of \( C_i \) whose data membership degree is in the interval \([\alpha_k, \alpha_{k+1}]\)

\( \theta_c \) : a threshold for cluster membership

\[ K(C_i, C_j) \] : a cluster farness measure to tell how far the clusters \( C_i \) and \( C_j \) are

\( \gamma_c \) : cluster radius which is the maximum distance from cluster center in order for an object to be considered as a member of the cluster

\( \theta_d \) : distance threshold which is the minimum distance in order for an object to be regarded as a distant one from a cluster center

\[ C^m_i \] : the distinct cluster member set of cluster \( C_i \) which consists of object whose membership degree is not less than threshold \( \theta_a \)

\[ C_i^\alpha = \{c_{ir} | \mu_{ir} \geq \theta_a\} \]

The distant cluster identification problem is to find distant cluster pairs for a data set. For a given data set \( D \) and a fuzzy cluster set \( C \), it is to find pairs of clusters \((C_i, C_j)\) for which \( K(C_i, C_j) \) is greater than a given threshold.

4 Distant Cluster Identification

In order to find out distant clusters, we introduce the notion of cluster farness measures. The measures are one to give for a pair of clusters a value on the range [0, 1] which indicates how much they are distant each other. The higher the measure gives, the farther they are. Compared to distance measures which could produce unbounded values, the range of cluster farness measures is bounded on [0, 1]. Thus this property allows to easily identify distant clusters just by choosing pairs of clusters for which the degree of farness is greater than a specified threshold.

4.1 Cluster Farness Measures

The proposed measures make use of a fuzzy membership function so-called distance membership function \( \mu_{\delta}(d) \) which maps distance to a degree on [0, 1]. Figure 1 shows a distance membership function.

\[ \mu_{\delta}(d) \]

![Fig.1 A distance membership function \( \mu_{\delta}(d) \)](image)

The membership function \( \mu_{\delta}(d) \) has the following constraints:

\[ (a) \quad 0 \leq \mu_{\delta}(d) \leq 1 \]

\[ (b) \quad \mu_{\delta}(d) = 0 \text{ if } d < \theta_d \]

\[ (c) \quad \theta_d > \gamma_c \]

\[ (d) \quad \mu_{\delta}(d_i) \leq \mu_{\delta}(d_j) \text{ if } d_i \leq d_j \]

The constraint (a) is required for it to be a membership function. The constraint (b) is required due to the definition of the threshold \( \theta_d \) which is the minimum distance to be considered as somewhat distant cluster. The constraint (c) is imposed to avoid the situations that an object belonging to a cluster is treated as a distant one.
The constraint (d) enforces the monotonicity on the membership function.

The proposed measures assume that the clusters are formed using a fuzzy clustering algorithm and thus an object may have a membership degree to each cluster. Here we propose three cluster farness measures $K(C_i, C_j)$ as follows:

$$K_1(C_i, C_j) = \frac{\sum_{c_i \in C_i, c_j \in C_j} \min\{\mu_{ij}, \mu_{ji}\} \cdot \mu_{ij} \cdot |c_i - c_j|}{\sum_{c_i \in C_i, c_j \in C_j} \min\{\mu_{ij}, \mu_{ji}\}}$$

$$K_2(C_i, C_j) = \mu_{ij} |O_i - O_j|$$

$$K_3(C_i, C_j) = \frac{\sum_{k=1}^{q} \sum_{c_k \in C_{i,j}} \min\{\mu_{ik}, \mu_{jk}\} \cdot \mu_{ik} \cdot \min\{|c_k - c_{i,j}|\}}{\sum_{c_k \in C_{i,j}} \sum_{c_k \in C_{i,j}} \min\{\mu_{ik}, \mu_{jk}\}}$$

In $K_1(C_i, C_j)$, when an object $c_{ij}$ has for a cluster $C$, a membership degree not less than a specified threshold, i.e., $\mu_{ij} \geq \theta_C$, it is regarded as a member of the cluster, i.e., $c_{ij} \in C_i$. In order to define the degree of farness, $K_1(C_i, C_j)$ takes into account both the membership degrees of compared objects to their corresponding cluster and the membership degree of their distance to the distance membership function. For all pairs of objects $(c_{ij}, c_{ij})$ to be considered, $\min\{\mu_{ij}, \mu_{ji}\}$ plays the role of a weighting factor.

$K_2(C_i, C_j)$ just pays attention to the distance between the centers of clusters. It allows fast computation for the degree of cluster farness, but could not reflect the distribution of objects in clusters.

$K_3(C_i, C_j)$ is an approximation to $K_1(C_i, C_j)$ developed to reduce the computation burden for $K_1(C_i, C_j)$. It first stratifies the objects of clusters into the subsets according to the membership degrees to cluster as follows:

$$IS_k(C_i) = \{c_k | \alpha_k \leq \mu_{ij} < \alpha_{k+1}\}$$

$$LS = \{\alpha_k, i = 1, \ldots, q \ 0 \leq \alpha_1 < \alpha_2 < \cdots < \alpha_q \leq 1\}$$

Then the objects at the same level set are evaluated to determine the degree of cluster farness. Due to this characteristics, $K_3(C_i, C_j)$ could provide similar values to $K_1(C_i, C_j)$ in a faster way.

### 4.2 Characteristics of the Cluster Farness Measures

The above measures satisfy the following properties:

(a) $0 \leq K(C_i, C_j) \leq 1$

(b) $K(C_i, C_j) = K(C_j, C_i)$

(c) $K(C_i, C_j) = 0$ when the distance threshold $\theta_d$ is set to be large enough to cover the farthest pair of objects in the same cluster.

The property (a) is met for $K_1(C_i, C_j)$ because it is a weighted sum of $\mu_{ij} |c_{ir} - c_{ij}|$, and $\mu_{ij} |c_{ir} - c_{ij}|$ is in the interval $[0, 1]$. $K_2(C_i, C_j)$ holds the property (a) due to the property of distance membership function. $K_3(C_i, C_j)$ satisfies the property (a) in the same context of $K_1(C_i, C_j)$. The reflexivity property of (b) holds because the minimum operation and the distance operation are reflexive. The property (c) is satisfied because an object is regarded as a (fuzzy) member of cluster $C_i$ only when its membership degree not less than a specified threshold, i.e., $\mu_{ij} \geq \theta_C$, and the threshold $\theta_d$ is greater than two times of the cluster radius $\gamma_C$ which is set to cover the cluster members.

### 4.3 Fuzzy Distant Relation

The proposed cluster farness measures allow to build a fuzzy relation which shows the degree of farness between clusters. The fuzzy distant relation $R_{FC}$ is defined as follows:

$$R_{FC} = \{(C_i, C_j, \mu_{ij}) | C_i \in C, C_j \in C, \mu_{ij} = K(C_i, C_j)\}$$

The relation $R_{FC}$ has the following properties:

(a) $R_{FC}$ is symmetric.

(b) $R_{FC}$ is irreflexive.

The properties (a) and (b) follow from the properties of cluster farness measures shown in Section 3.2. Once a fuzzy distant relation $R_{FC}$ is constructed, it is possible to identify distant clusters by taking $\alpha$-cut operation for the relation. The higher $\alpha$-cut operations are applied, the more distant clusters are determined.

### 5 An Application Example

In order to identify distant clusters for a data set, the following procedure is applied:

**procedure** Find_Distant_Clusters

**input**: a set of data $D$

...
output: a set of fuzzy clusters $C$

a fuzzy distant relation $R_{FC}$

begin

Apply the fuzzy $k$-means algorithm to $D$.

Let the constructed fuzzy clusters be $C$.

For each cluster $C_i$

Apply an farness measure $K$ to all other clusters $C_j$.

Set $\left( (C_i, C_j), K(C_i, C_j) \right)$ for $R_{FC}$.

Apply an $\alpha$-cut operation to $R_{FC}$ to choose the pairs of clusters.

end.

In order to show the applicability of the proposed method, a data set of Fig. 2 was generated using the Gaussian function centered at $(0.5, 0.5, -0.5), (1.5, -2.0, -1.5), (0.5, -1.0, -1.5), (-1.1, -1.5), (1.5, -0.5, 1.5)$, for each center 10 data was generated with standard deviation 0.3. For the data set, the fuzzy k-means clustering algorithm was applied to get fuzzy clusters. The colors in the figure indicate the cluster labels for which membership degree is the highest.

![Fig. 2 Data used in the experiment](image)

Table 1 shows the fuzzy relation constructed for the data by the $K_1$ measure. We can see that the relation is symmetric and irreflexive. From the relation, we can get the distant cluster pairs by taking $\alpha$-cut operations. For example, if $\alpha$ is set to 0.7, the following pairs are chosen as the distant clusters: $\{(C_1, C_2), (C_1, C_3), (C_1, C_4), (C_1, C_5), (C_2, C_5), (C_4, C_5)\}$. When $\alpha$ is set to a higher value like 0.9, the more distant cluster pairs are identified as follows: $\{(C_1, C_2), (C_1, C_3), (C_1, C_4), (C_2, C_5)\}$

### Table 1. The fuzzy distant relation $R_{FC}$ for data of Fig. 2

<table>
<thead>
<tr>
<th></th>
<th>$C_1$</th>
<th>$C_2$</th>
<th>$C_3$</th>
<th>$C_4$</th>
<th>$C_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_1$</td>
<td>0.0</td>
<td>1.0</td>
<td>0.99</td>
<td>0.82</td>
<td>0.99</td>
</tr>
<tr>
<td>$C_2$</td>
<td>1.00</td>
<td>0.0</td>
<td>0.64</td>
<td>0.24</td>
<td>0.98</td>
</tr>
<tr>
<td>$C_3$</td>
<td>0.97</td>
<td>0.64</td>
<td>0.0</td>
<td>0.23</td>
<td>0.17</td>
</tr>
<tr>
<td>$C_4$</td>
<td>0.82</td>
<td>0.24</td>
<td>0.23</td>
<td>0.00</td>
<td>0.79</td>
</tr>
<tr>
<td>$C_5$</td>
<td>0.99</td>
<td>0.98</td>
<td>0.17</td>
<td>0.79</td>
<td>0.0</td>
</tr>
</tbody>
</table>

6 Conclusion

The paper introduced the distant cluster identification problem for clustering. It also proposed the notion of cluster farness measure and presented three cluster farness measures. Some interesting properties for the measures have been investigated. In addition, it showed that the measures could be used to define a fuzzy relation to indicate the degree of farness between clusters, and that the pairs of distant clusters could be found out from such a fuzzy relation. The proposed method is a useful tool to easily identify distant clusters for a data set.

References:


