An Indiscernibility-Based Clustering Method with Iterative Refinement of Equivalence Relations

Shoji Hirano and Shusaku Tsumoto

Department of Medical Informatics, Shimane Medical University, School of Medicine
89-1 Enya-cho, Izumo, Shimane 693-8501, Japan
shirano@ieee.org, tsumoto@computer.org

Abstract. In this paper, we present an indiscernibility-based clustering method that can handle relative proximity. The main benefit of this method is that it can be applied to proximity measures that do not satisfy the triangular inequality. Additionally, it may be used with a proximity matrix – thus it does not require direct access to the original data values. In the experiments we demonstrate, with the use of partially mutated proximity matrices, that this method produces good clusters even when the employed proximity does not satisfy the triangular inequality.

1 Introduction

Clustering is a powerful tool for revealing underlying structure of the data. A number of methods, for example, hierarchical, partial, and model-based methods, have been proposed and have produced good results on both artificial and real-life data [1].

In order to assess the quality of clusters being produced, most of the conventional clustering methods employ quality measures that are associated with centroids of clusters. For example, the internal homogeneity of a cluster can be measured as the sum of differences from objects in the cluster to their centroid, and it can be further used as a component of the total quality measure for assessing a clustering result. Such centroid-based methods work well on datasets in which the proximity of objects satisfies the natures of distance that are, positivity ($d(x, y) \geq 0$), identity ($d(x, y) = 0$ iff $x = y$), symmetry ($d(x, y) = d(y, x)$), and triangular inequality ($d(x, z) \leq d(x, y) + d(y, z)$), for any objects x, y and z. However, they have a potential weakness in handling relative proximity. Relative proximity is a class of proximity measures that is suitable for representing subjective similarity or dissimilarity such as the degree of likeness between people. It may not satisfy the triangular inequality because the proximity $d(x, z)$ of x and z is allowed to be independent of y. Usually, the centroid c of objects x, y and z is expected to be in their convex hull. However, if we use relative proximity, the centroid can be out of x, y, and z's convex hull because proximity between c and other objects can be far greater (if we use dissimilarity as proximity) or smaller (if we use similarity) than $d(x, y)$, $d(y, z)$ and $d(x, z)$. Namely, a centroid does

N. Lavrač et al. (Eds.): PKDD 2003, LNAI 2838, pp. 192–203, 2003.

not hold its geometric properties under these conditions. Thus another criterion should be used for evaluating the quality of the clusters.

In this paper, we present a new clustering method based on the indiscernibility degree of objects. The main benefit of this method is that it can be applied to proximity measures that do not satisfy the triangular inequality. Additionally, it may be used with a proximity matrix – thus it does not require direct access to the original data values.

2 The Method

2.1 Overview

Our method is based on iterative refinement of N binary classifications, where N denotes the number of objects. First, an equivalence relation, that classifies all the other objects into two classes, is assigned to each of N objects by referring to the relative proximity. Next, for each pair of objects, the number of binary classifications in which the pair is included in the same class is counted. This number is termed the indiscernibility degree. If the indiscernibility degree of a pair is larger than a user-defined threshold value, the equivalence relations may be modified so that all of the equivalence relations commonly classify the pair into the same class. This process is repeated until class assignment becomes stable. Consequently, we may obtain the clustering result that follows a given level of granularity, without using geometric measures.

2.2 Assignment of Initial Equivalence Relations

When dissimilarity is defined relatively, the only information available for object x_i is the dissimilarity of x_i to other objects, for example to x_j, $d(x_i, x_j)$. This is because the dissimilarities for other pairs of objects, namely $d(x_j, x_k)$, $x_j, x_k \neq x_i$, are determined independently of x_i. Therefore, we independently assign an initial equivalence relation to each object and evaluate the relative dissimilarity observed from the corresponding object.

Let $U = \{x_1, x_2, ..., x_N\}$ be the set of objects we are interested in. An equivalence relation R_i for object x_i is defined by

$$U/R_i = \{P_i,\ U - P_i\}, \tag{1}$$

where

$$P_i = \{x_j |\ d(x_i, x_j) \leq Th_{di}\}, \quad \forall x_j \in U. \tag{2}$$

$d(x_i, x_j)$ denotes dissimilarity between objects x_i and x_j, and Th_{di} denotes an upper threshold value of dissimilarity for object x_i. The equivalence relation, R_i classifies U into two categories: P_i, which contains objects similar to x_i and $U - P_i$, which contains objects dissimilar to x_i. When $d(x_i, x_j)$ is smaller than Th_{di}, object x_j is considered to be indiscernible to x_i. U/R_i can be alternatively written as $U/R_i = \{\{[x_i]_{R_i}\}, \{\overline{[x_i]_{R_i}}\}\}$, where $[x_i]_{R_i} \cap \overline{[x_i]_{R_i}} = \phi$ and $[x_i]_{R_i} \cup \overline{[x_i]_{R_i}} = U$ hold.

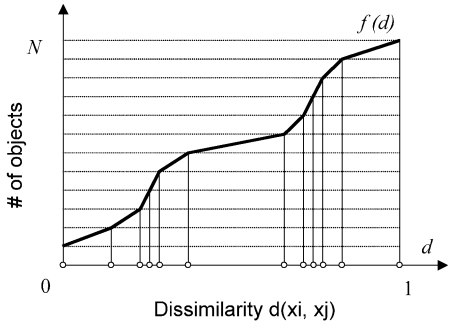

Fig. 1. An example of function $f(d)$ generated by $d(x_i, x_s)$.

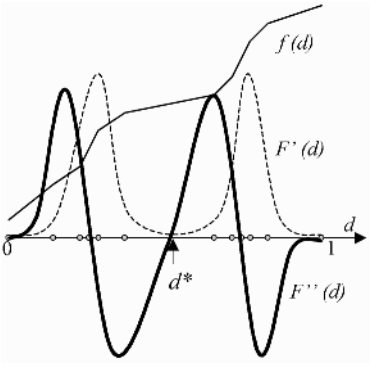

Fig. 2. Relations between $f(d)$ and its smoothed first- and second-order derivatives $F'(d)$ and $F''(d)$.

Definition of the dissimilarity measure $d(x_i, x_j)$ is arbitrary. If all the attribute values are numerical, ordered, and independent of each other, conventional Minkowski distance

$$d(x_i, x_j) = \left(\sum_{a=1}^{N_a} |x_{ia} - x_{j_a}|^p \right)^{\frac{1}{p}}, \tag{3}$$

where N_a denotes the number of attributes, x_{ia} denotes the a-th attribute of object x_i, and p denotes a positive integer, is a reasonable choice since it has been successfully applied to many areas and its mathematical properties have been well investigated. More generally, any type of dissimilarity measure can be used regardless of whether or not the triangular inequality is satisfied among objects.

Threshold of dissimilarity Th_{di} for object x_i is automatically determined based on the spatial density of objects. The procedure is summarized as follows.

1. Sort $d(x_i, x_j)$ in ascending order. For simplicity, we denote the sorted dissimilarity using the same representation $d(x_i, x_s)$, $1 \leq s \leq N$.
2. Generate a function $f(d)$ that represents the cumulative distribution of d. For a given dissimilarity d, function f returns the number of objects whose dissimilarity to x_i is smaller than d. Figure 1 shows an example. Function $f(d)$ can be generated by linearly interpolating $f(d(x_i, x_s)) = n$, where n corresponds to the index of x_s in the sorted dissimilarity list.
3. Obtain the smoothed second-order derivative of $f(d)$ as a convolution of $f(d)$ and the second-order derivative of Gaussian function as follows.

$$F''(d) = \int_{-\infty}^{\infty} f(u) \frac{-(d-u)}{\sigma^3 \sqrt{2\pi}} e^{-(d-u)^2/2\sigma^2} du, \tag{4}$$

where $f(d) = 1$ and $f(d) = N$ are used for $d < 0$ and $d > 1$ respectively. The smoothed first-order derivative $F'(d)$ of $f(d)$ represents spatial density

of objects because it represents increase or decrease velocity of the objects induced by the change of dissimilarity. Therefore, by calculating its further derivative as $F''(d)$, we find a sparse region between two dense regions. Figure 2 illustrates relationship between $f(d)$ and its smoothed derivatives. The most sparse point d^* should take a local minimum of the density where the following conditions are satisfied.

$$F''(d^* - \Delta d) < 0 \text{ and } F''(d^* + \Delta d) > 0. \tag{5}$$

Usually, there are some d^*s in $f(d)$ because $f(d)$ has multiple local minima. The value of σ in the above Gaussian function can be adjusted to eliminate meaningless small minima.

4. Choose the smallest d^* and object x_{j^*} whose dissimilarity is the closest to but not larger than d^*. Finally, the dissimilarity threshold Th_{d_i} is obtained as $Th_{d_i} = d(x_i, x_{j^*})$.

2.3 Refinement of Initial Equivalence Relations

Suppose we are interested in two objects, x_i and x_j. In indiscernibility-based classification, they are classified into different categories regardless of other relations, if there is at least one equivalence relation that has an ability to discern them. In other words, the two objects are classified into the same category only when all of the equivalence relations commonly regard them as indiscernible objects. This strict property is not acceptable in clustering because it will generate many meaningless small categories, especially when global associations between the equivalence relations are not taken into account. We consider that objects should be classified into the same category when most of, but not necessarily all of, the equivalence relations commonly regard the objects as indiscernible. In the second stage, we perform global optimization of initial equivalence relations so that they produce adequately coarse classification to the objects. The global similarity of objects is represented by a newly introduced measure, the *indiscernibility degree*. Our method takes a threshold value of the indiscernibility degree as an input and associates it with the user-defined granularity of the categories. Given the threshold value, we iteratively refine the initial equivalence relations in order to produce categories that meet the given level of granularity.

Now let us assume $U = \{x_1, x_2, x_3, x_4, x_5\}$ and classifications of U by $\mathbf{R} = \{R_1, R_2, R_3, R_4, R_5\}$ is given as follows.

$$\begin{aligned}
U/R_1 &= \{\{x_1, x_2, x_3\}, \{x_4, x_5\}\}, \\
U/R_2 &= \{\{x_1, x_2, x_3\}, \{x_4, x_5\}\}, \\
U/R_3 &= \{\{x_2, x_3, x_4\}, \{x_1, x_5\}\}, \\
U/R_4 &= \{\{x_1, x_2, x_3, x_4\}, \{x_5\}\}, \\
U/R_5 &= \{\{x_4, x_5\}, \{x_1, x_2, x_3\}\}.
\end{aligned} \tag{6}$$

This example contains three types of equivalence relations: R_1 $(= R_2 = R_5)$, R_3 and R_4. Since each of them classifies U slightly differently, classification of U

by the family of equivalence relations $\mathbf{R}$, $U/\mathbf{R}$, contains four very small, almost independent categories.

$$U/\mathbf{R} = \{\{x_1\}, \{x_2, x_3\}, \{x_4\}, \{x_5\}\}. \tag{7}$$

In the following we present a method to reduce the variety of equivalence relations and to obtain coarser categories.

First, we define an *indiscernibility degree*, $\gamma(x_i, x_j)$, for two objects x_i and x_j as follows.

$$\gamma(x_i, x_j) = \frac{\sum_{k=1}^{|U|} \delta_k^{indis}(x_i, x_j)}{\sum_{k=1}^{|U|} \delta_k^{indis}(x_i, x_j) + \sum_{k=1}^{|U|} \delta_k^{dis}(x_i, x_j)}, \tag{8}$$

where

$$\delta_k^{indis}(x_i, x_j) = \begin{cases} 1, \text{ if } (x_i \in [x_k]_{R_k} \wedge x_j \in [x_k]_{R_k}) \\ 0, \text{ otherwise.} \end{cases} \tag{9}$$

and

$$\delta_k^{dis}(x_i, x_j) = \begin{cases} 1, \text{ if } (x_i \in [x_k]_{R_k} \wedge x_j \notin [x_k]_{R_k}) \text{ or} \\ \quad \text{if } (x_i \notin [x_k]_{R_k} \wedge x_j \in [x_k]_{R_k}) \\ 0, \text{ otherwise.} \end{cases} \tag{10}$$

Equation (9) shows that $\delta_k^{indis}(x_i, x_j)$ takes 1 only when the equivalence relation R_k regards both x_i and x_j as indiscernible objects, under the condition that both of them are in the same equivalence class as x_k. Equation (10) shows that $\delta_k^{dis}(x_i, x_j)$ takes 1 only when R_k regards x_i and x_j as discernible objects, under the condition that either of them is in the same class as x_k. By summing $\delta_k^{indis}(x_i, x_j)$ and $\delta_k^{dis}(x_i, x_j)$ for all $k(1 \leq k \leq |U|)$ as in Equation (8), we obtain the percentage of equivalence relations that regard x_i and x_j as indiscernible objects. Note that in Equation (9), we excluded the case when x_i and x_j are indiscernible but not in the same class as x_k. This is to exclude the case where R_k does not significantly put weight on discerning x_i and x_j. As mentioned in Section 2.2, P_k for R_k is determined by focusing on similar objects rather than dissimilar objects. This means that when both of x_i and x_j are highly dissimilar to x_k, their dissimilarity is not significant for x_k, when determining the dissimilarity threshold Th_{dk}. Thus we only count the number of equivalence relations that certainly evaluate the dissimilarity of x_i and x_j.

For example, the indiscernibility degree $\gamma(x_1, x_2)$ of objects x_1 and x_2 in the above case is calculated as follows.

$$\begin{aligned} \gamma(x_1, x_2) &= \frac{\sum_{k=1}^{5} \delta_k^{indis}(x_1, x_2)}{\sum_{k=1}^{5} \delta_k^{indis}(x_1, x_2) + \sum_{k=1}^{5} \delta_k^{dis}(x_1, x_2)} \\ &= \frac{1+1+0+1+0}{(1+1+0+1+0) + (0+0+1+0+0)} = \frac{3}{4}. \end{aligned} \tag{11}$$

Let us explain this example with the calculation of the numerator (1+1+0+1+0). The first value 1 is for $\delta_1^{indis}(x_1, x_2) as shown$. Since x_1 and x_2 are in the same class of R_1 and obviously, they are in the same class to x_1, $\delta_1^{indis}(x_1, x_2) = 1$

Table 1. Degree γ for objects in Eq. (6).

	x_1	x_2	x_3	x_4	x_5
x_1	3/3	3/4	3/4	1/5	0/4
x_2		4/4	4/4	2/5	0/5
x_3			4/4	2/5	0/5
x_4				3/3	1/3
x_5					1/1

Table 2. Degree γ after the first refinement.

	x_1	x_2	x_3	x_4	x_5
x_1	3/3	3/4	3/4	2/4	1/5
x_2		4/4	4/4	3/4	0/5
x_3			4/4	3/4	0/5
x_4				3/3	1/5
x_5					1/1

Table 3. Degree γ after the second refinement.

	x_1	x_2	x_3	x_4	x_5
x_1	4/4	4/4	4/4	4/4	0/5
x_2		4/4	4/4	4/4	0/5
x_3			4/4	4/4	0/5
x_4				4/4	0/5
x_5					1/1

holds. The second value is for $\delta_2^{indis}(x_1, x_2)$, and analogously, it becomes 1. The third value is for $\delta_3^{indis}(x_1, x_2)$. Since x_1 and x_2 are in the different classes of R_3, it becomes 0. The fourth value is for $\delta_4^{indis}(x_1, x_2)$ and it obviously, becomes 1. The last value is for $\delta_5^{indis}(x_1, x_2)$. Although x_1 and x_2 are in the same class of R_5, their class is different to that of x_5. Thus $\delta_5^{indis}(x_1, x_2)$ returns 0.

Indiscernibility degrees for all of the other pairs in U are tabulated in Table 1. Note that the indiscernibility degree of object x_i to itself, $\gamma(x_i, x_i)$, will always be 1.

From its definition, a larger $\gamma(x_i, x_j)$ represents that x_i and x_j are commonly regarded as indiscernible objects by the large number of the equivalence relations. Therefore, if an equivalence relation R_l discerns the objects that have high γ value, we consider that it represents excessively fine classification knowledge and refine it according to the following procedure (note that R_l is rewritten as R_i below for the purpose of generalization).

Let $R_i \in \mathbf{R}$ be an initial equivalence relation on U. A refined equivalence relation $R_i' \in \mathbf{R}'$ of R_i is defined as

$$U/R_i' = \{P_i',\ U - P_i'\}, \tag{12}$$

where P_i' denotes a set of objects represented by

$$P_i' = \{x_j | \gamma(x_i, x_j) \geq T_h\}, \quad \forall x_j \in U. \tag{13}$$

and T_h denotes the lower threshold value of the indiscernibility degree above, in which x_i and x_j are regarded as indiscernible objects. It represents that when $\gamma(x_i, x_j)$ is larger than T_h, R_i is modified to include x_j into the class of x_i.

Suppose we are given $Th = 3/5$ for the case in Equation (6). For R_1 we obtain the refined relation R_1' as

$$U/R_1' = \{\{x_1, x_2, x_3\}, \{x_4, x_5\}\}, \tag{14}$$

because, according to Table 1, $\gamma(x_1, x_1) = 1 \geq T_h = 3/5$, $\gamma(x_1, x_2) = 3/4 \geq 3/5$, $\gamma(x_1, x_3) = 3/4 \geq 3/5$, $\gamma(x_1, x_4) = 1/5 \leq 3/5$, $\gamma(x_1, x_5) = 0/5 \leq 3/5$ hold. In the same way, the rest of the refined equivalence relations are obtained as follows.

$$U/R_2' = \{\{x_1, x_2, x_3\}, \{x_4, x_5\}\},$$

$$
\begin{aligned}
U/R'_3 &= \{\{x_1, x_2, x_3\}, \{x_4, x_5\}\}, \\
U/R'_4 &= \{\{x_4\}, \{x_1, x_2, x_3, x_5\}\}, \\
U/R'_5 &= \{\{x_5\}, \{x_1, x_2, x_3, x_4\}\}.
\end{aligned} \tag{15}
$$

Then we obtain classification of U by the refined family of equivalence relations $\mathbf{R}'$ as follows.

$$
U/\mathbf{R}' = \{\{x_1, x_2, x_3\}, \{x_4\}, \{x_5\}\}. \tag{16}
$$

In the above example, R_3, R_4 and R_5 are modified so that they include similar objects into the equivalence class of x_3, x_4 and x_5, respectively. Three types of the equivalence relations remain, however, the categories become coarser than those in Equation (7) by the refinement.

2.4 Iterative Refinement of Equivalence Relations

It should be noted that the state of the indiscernibility degrees could also be changed after refinement of the equivalence relations, since the degrees are recalculated using the refined family of equivalence relations $\mathbf{R}'$.

Suppose we are given another threshold value $T_h = 2/5$ for the case in Equation (6). According to Table 1, we obtain $\mathbf{R}'$ after the first refinement, as follows.

$$
\begin{aligned}
U/R'_1 &= \{\{x_1, x_2, x_3\}, \{x_4, x_5\}\}, \\
U/R'_2 &= \{\{x_1, x_2, x_3, x_4\}, \{x_5\}\}, \\
U/R'_3 &= \{\{x_1, x_2, x_3, x_4\}, \{x_5\}\}, \\
U/R'_4 &= \{\{x_2, x_3, x_4\}, \{x_1, x_5\}\}, \\
U/R'_5 &= \{\{x_5\}, \{x_1, x_2, x_3, x_4\}\}.
\end{aligned} \tag{17}
$$

Hence

$$
U/\mathbf{R}' = \{\{x_1\}, \{x_2, x_3\}, \{x_4\}, \{x_5\}\}. \tag{18}
$$

The categories in $U/\mathbf{R}'$ are exactly the same as those in Equation (7). However, the state of the indiscernibility degrees are not the same because the equivalence relations in $\mathbf{R}'$ are different from those in $\mathbf{R}$. Table 2 summarizes the indiscernibility degrees, recalculated using $\mathbf{R}'$. In Table 2, it can be observed that the indiscernibility degrees of some pairs of objects, for example $\gamma(x_1, x_4)$, increased after the refinement, and now they exceed the threshold $th = 2/5$. Thus we perform refinement of equivalence relations again using the same T_h and the recalculated γ. Then we obtain

$$
\begin{aligned}
U/R'_1 &= \{\{x_1, x_2, x_3, x_4\}, \{x_5\}\}, \\
U/R'_2 &= \{\{x_1, x_2, x_3, x_4\}, \{x_5\}\}, \\
U/R'_3 &= \{\{x_1, x_2, x_3, x_4\}, \{x_5\}\}, \\
U/R'_4 &= \{\{x_1, x_2, x_3, x_4\}, \{x_5\}\}, \\
U/R'_5 &= \{\{x_5\}, \{x_1, x_2, x_3, x_4\}\}.
\end{aligned} \tag{19}
$$

Hence

$$U/\mathbf{R}' = \{\{x_1, x_2, x_3, x_4\}, \{x_5\}\}. \tag{20}$$

After the second refinement, the number of the equivalence relations in $\mathbf{R}'$ are reduced from 3 to 2, and the number of categories are also reduced from 4 to 2. We further update the state of the indiscernibility degrees according to the equivalence relations after the second refinement. The results are shown in Table 3. Since no new pairs, whose indiscernibility degree exceeds the given threshold appear, refinement process may be halted and the stable categories may be obtained, as in Equation (20).

As shown in this example, refinement of the equivalence relations may change the indiscernibility degree of objects. Thus we iterate the refinement process using the same T_h until the categories become stable. Note that each refinement process is performed using the previously 'refined' set of equivalence relations.

3 Experimental Results

We applied the proposed method to some artificial numerical datasets and evaluated its clustering ability. Note that we used numerical data, but clustered them without using any type of geometric measures.

3.1 Effects of Iterative Refinement

We first examined the effects of refinement of the initial equivalence relations. A two-dimensional numerical dataset was artificially created using Neyman-Scott method [2]. The number of seed points was set to 5. Each of the five clusters contained approximately 100 objects, and a total of 491 objects were included in the data. We evaluated validity of the clustering result based on the following measure:

$$\text{Validity } v_{\mathbf{R}}(C) = \min\left(\frac{|X_{\mathbf{R}} \cap C|}{|X_{\mathbf{R}}|}, \frac{|X_{\mathbf{R}} \cap C|}{|C|}\right),$$

where $X_{\mathbf{R}}$ and C denote the clusters obtained by the proposed method and the expected clusters, respectively. The threshold value for refinement T_h was set to 0.2, meaning that if two objects were commonly regarded as indiscernible by 20% of objects in the data, all the equivalence relations were modified to regard them as indiscernible objects.

Without refinement, the method produced 461 small clusters. Validity of the result was 0.011, which was the smallest value assigned to this dataset. This was because the small size of the clusters produced very low coverage, namely, amount of overlap between the generated clusters and their corresponding expected clusters was very small compared with the size of the expected clusters.

By performing refinement one time, the number of clusters was reduced to 429, improving validity to 0.013. As the refinement proceeds, the small clusters merged as shown in Figures 3 and 4. Validity of the results continued to increase. Finally, clusters became stable at the 6th refinement, where 10 clusters were

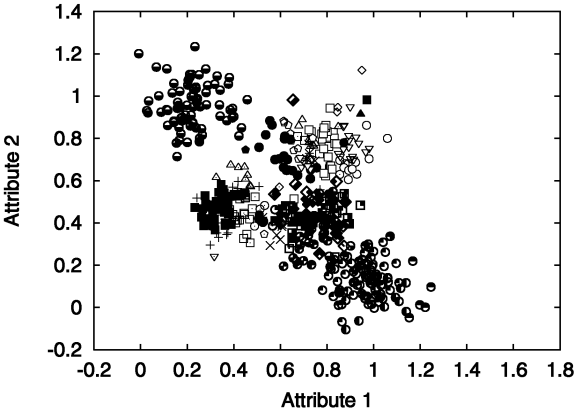

Fig. 3. Clusters after 4th refinement.

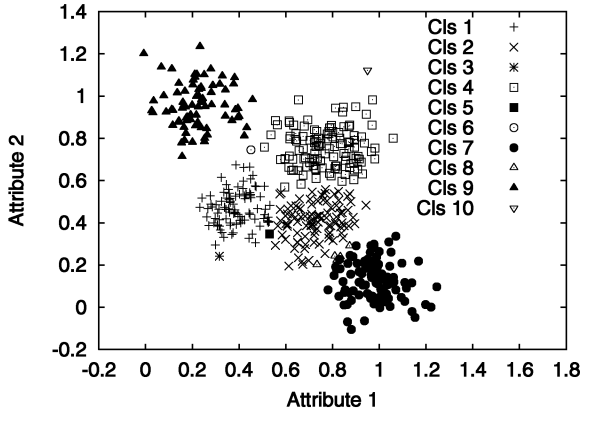

Fig. 4. Clusters after 6th refinement.

formed as shown in Figure 4. Validity of the clusters was 0.927. One can observe that a few small clusters, for example, clusters 5 and 6, were formed between the large clusters. These objects were classified into independent clusters because of the competition of the large clusters containing almost the same populations. Aside from this, the results revealed that the proposed method automatically produced good clusters that have high correspondence to the original ones.

3.2 Capability of Handling Relative Proximity

In order to validate the method's capability of handling relative proximity, we performed clustering experiments with another dataset. The data was originally generated on the two-dimensional Euclidean space likewise the previous dataset; however, in this case we randomly modified distances between data points in

Table 4. Comparison of the clustering results

Mutation Ratio[%]	0	10	20	30	40	50
AL-AHC	0.990	0.688 ± 0.011	0.670 ± 0.011	0.660 ± 0.011	0.633 ± 0.013	0.633 ± 0.018
CL-AHC	0.990	0.874 ± 0.076	0.792 ± 0.093	0.760 ± 0.095	0.707 ± 0.098	0.729 ± 0.082
Our method	0.981	0.980 ± 0.002	0.979 ± 0.003	0.980 ± 0.003	0.977 ± 0.003	0.966 ± 0.040

order to make the induced proximity matrix not fully satisfy the triangular inequality.

The dataset was prepared as follows. First, we created a two-dimensional data set by using the Neyman-Scott method [2]. The number of seed points was set to three, and a total of 310 points were included in the dataset. Next, we calculated the Euclidean distances between the data points and constructed a 310 × 310 proximity matrix. Then we randomly selected some elements of the proximity matrix and mutated them to zero. The ratio of elements to be mutated was set to 10%, 20%, 30%, 40%, and 50%. For each of these mutation ratio, we created 10 proximity matrices in order to include enough randomness. Consequently, we obtained a total of 50 proximity matrices.

We took each of the proximity matrices as an input and performed clustering of the dataset. Parameters used in the proposed method were manually determined to $\sigma = 15.0$ and $T_h = 0.3$. Additionally, we employed average-linkage and complete-linkage agglomerative hierarchical clustering methods (for short, AL-AHC and CL-AHC respectively) [3] for the purpose of comparison. Note that we partly disregarded the original data values and took the mutated proximity matrix as input of the clustering methods. Therefore, we did not employ clustering methods that require direct access to the data value.

We evaluated validity of the clustering results using the same measures as in the previous case. Table 4 shows the comparison results. The first row of the table represents the ratio of mutation. For example, 30 represents 30% of the elements in the proximity matrix were mutated to zero. The next three rows contain the validity obtained by AL-AHC, CL-AHC and the proposed method, respectively. Except for the cases in zero mutation ratio, validity is represented in the form of 'mean ± standard deviation', summarized from the 10 randomly mutated proximity matrices.

Without any mutation, the proximity matrix exactly corresponded to the one obtained by using the Euclidean distance. Therefore, both of AL-AHC and CL-AHC could produce high validity over 0.99. The proposed method also produced the high validity over 0.98. However, when mutation had occurred, the validity of clusters obtained by AL-AHC and CL-AHC largely reduced to 0.688 and 0.874, respectively. They kept decreasing moderately following the increase of mutation. The primary reason for inducing decrease of the validity was considered as follows. When the distance between two objects was forced to be mutated into zero, it brought a kind of local warp to the proximity of the objects. Thus the two objects could become candidates of the first linkage. If the two objects were originally belonged to the different clusters, these clusters were merged at

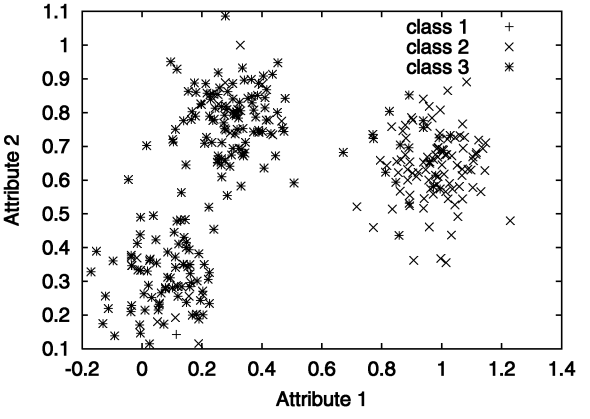

Fig. 5. Clustering results by AL-AHC. Ratio of mutation was 40%. Linkage was terminated when three clusters were formed.

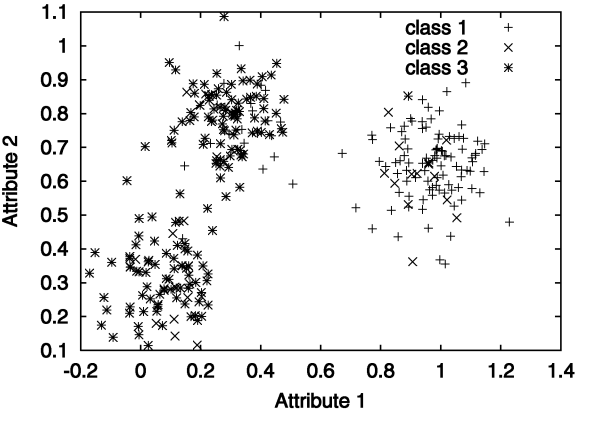

Fig. 6. Clustering results by CL-AHC. Ratio of mutation was 40%. Linkage was terminated when three clusters were formed.

an early stage of the merging process. Since both of AL-AHC and CL-AHC do not allow inverse of the cluster hierarchy, these clusters would never be separated. Consequently, inappropriately bridged clusters were obtained as shown in Figures 5 and 6.

On the contrary, the proposed method produced high validity even when the mutation ratio approached to 50%. In this method, effects of a mutation was very limited. The two concerning objects would consider themselves as indiscernible objects, however, the majority of other objects never change their classification. Although the categories obtained by the initial equivalence relations could be distorted, they could be globally adjusted through iterative refinement of the equivalence relations. Consequently, good clusters were obtained as shown in

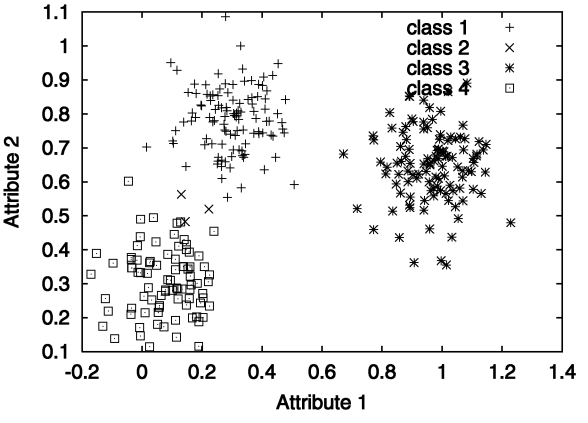

Fig. 7. Clustering results by the proposed method. Ratio of mutation was 40%. Iteration terminated at the fourth cycle.

Figure 7. This demonstrates the capability of the method for handling locally distorted proximity matrix that do not satisfy the triangular inequality.

4 Conclusions

In this paper, we have presented an indiscernibility-based clustering method, which clusters objects according to their relative proximity. Experimental results from the artificially created numerical datasets demonstrated that this method could produce good clusters even when the proximity of the objects did satisfy the triangular inequality. Future work include reduction of the computational complexity of the method and empirical evaluation of its clustering ability on large and complex real-life databases.

Acknowledgment

This work was supported in part by the Grant-in-Aid for Scientific Research on Priority Area (B)(No.759) by the Ministry of Education, Culture, Science and Technology of Japan.

References

1. P. Berkhin (2002): Survey of Clustering Data Mining Techniques. Accrue Software Research Paper. URL: http://www.accrue.com/products/researchpapers.html.
2. J. Neyman and E. L. Scott (1958): "Statistical Approach to Problems of Cosmology," *Journal of the Royal Statistical Society*, Series B20: 1–43.
3. B. S. Everitt, S. Landau, and M. Leese (2001): Cluster Analysis Fourth Edition. Arnold Publishers.