On the Relations between Indiscernibility Degree and Cluster Number in Rough Clustering

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

Abstract

This paper discusses the relationships between indiscernibility degree and clustering results in rough clustering. We first examine the relationship between the threshold value of indiscernibility degree and resultant clusters. After that, we apply random disturbance to the perfect relations, and examine how the results change. The results imply the threshold-validity curve may have globally convex shape, and the best value may be selected according to the change of cluster numbers.

Keywords: Indiscernibility, Clustering, Rough Set

1 Introduction

This paper discusses the relationships between indiscernibility degree and clustering results in rough clustering[1]. Rough clustering is a novel grouping method that controls the granularity of discrimination by a measure called indiscernibility degree. The clustering results can be affected by two factors: (1) initial equivalence relations that form the basic partition of objects, (2) threshold value of indiscernibility degrees. In previous work, we have shown that (1) minor disturbance in the initial equivalence relations would be absorbed in refinement steps, and (2) there might be a range of indiscernibility degree that yield high cluster validity. However, these findings were all dependent on the method for determining initial equivalence relations - in this case density-based determination method.

In this work, we employ the perfect initial equivalence relations that are derived from class labels of objects so that we exclude the effect of the method for initial equivalence determination. We first examine the relationship between the threshold value of indiscernibility degree and resultant clusters. After that, we apply random disturbance to the perfect relations, and examine how the results change. The results imply the threshold-validity curve may have globally convex shape, and the best value may be selected according to the change of cluster numbers.

The remainder of this paper is organized as follows. Section 2 gives a brief explanation of rough clustering. Section 3 describes experimental results on artificial datasets, and Section 4 concludes the technical results.

2 Rough Clustering Overview

This section gives a brief overview of rough clustering, aka, indiscernibility-based clustering. This 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. The main benefits of this method is that (1) it can handle relative proximity, where no geometric measure such as centroids can not be defined, (2) it can take dissimilarity matrix as input and does not require any direct reference to the original data value.

There are two parameters that control the behavior of this clustering method: the threshold value $T_h$ for refinement of equivalence relations and the number $N_r$ of iteration of refinement. As shown in the experiments, $N_r$ can be determined automatically, because the equivalence relations will be stable after several cycles of refinement. The refinement process can be terminated when no candidates for refinement appear.

### 2.1 Assignment of Initial Equivalence Relations

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

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

where $P_i = \{x_j \mid d(x_i, x_j) \leq T_h d_i\}$, $\forall x_j \in U$. (2)

$d(x_i, x_j)$ denotes dissimilarity between objects $x_i$ and $x_j$, and $T_h d_i$ 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 $T_h d_i$, object $x_j$ is considered to be indiscernible to $x_i$. $U/R_i$ can be alternatively written as $U/R_i = \{\{x_i\}_R_i, \{x_i\}_R_i\}$, where $\{x_i\}_R_i \cap \{x_i\}_R_i = \phi$ and $\{x_i\}_R_i \cup \{x_i\}_R_i = U$ hold.

Methods for constructing initial equivalence relations, including the choice of dissimilarity measure, is arbitrary under the condition that it has the ability of performing binary classification of $U$. For example, one can simply use Euclidean distance and k-means with cluster number 2, if it is appropriate based on the property of the data. We have introduced a method for constructing initial equivalence relations based on the denseness of the objects in [1]: however, one may use another approach for this purpose.

### 2.2 Refinement of Initial Equivalence Relations

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. Rough clustering 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 $R = \{R_1, R_2, R_3, R_4, R_5\}$ is given as follows.

$$
U/R_1 = \{\{x_1, x_2, x_3\}, \{x_4, x_5\}\},
U/R_2 = \{\{x_1, x_2, x_3\}, \{x_4, x_5\}\},
U/R_3 = \{\{x_2, x_3, x_4\}, \{x_1, x_5\}\},
U/R_4 = \{\{x_1, x_2, x_3, x_4\}, \{x_5\}\},
U/R_5 = \{\{x_4, x_5\}, \{x_1, x_2, x_3\}\}.
$$

This example contains three types of equivalence relations: $R_1 (= R_2 = R_3)$, $R_3$ and $R_4$. Since each of them classifies $U$ slightly differently, classification of $U$ by the family of equivalence relations $R$, $U/R$, contains four very small, almost independent categories.

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

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.
From its definition, a large value, we consider that it represents excessively fine classification knowledge and refine it according to the following procedure (note that $R_0$ is rewritten as $R_1$ below for the purpose of generalization).

$$\gamma(x_i, x_j) = \frac{\sum_{k=1}^{[U]} \delta_k^{\text{indis}}(x_i, x_j)}{\sum_{k=1}^{[U]} \delta_k^{\text{indis}}(x_i, x_j) + \sum_{k=1}^{[U]} \delta_k^{\text{dis}}(x_i, x_j)},$$  \hspace{1cm} (5)

where

$$\delta_k^{\text{indis}}(x_i, x_j) = \begin{cases} 1, & \text{if } (x_i \in [x_k]_{R_k} \wedge x_j \in [x_k]_{R_k}) \\ 0, & \text{otherwise}. \end{cases}$$  \hspace{1cm} (6)

and

$$\delta_k^{\text{dis}}(x_i, x_j) = \begin{cases} 1, & \text{if } (x_i \in [x_k]_{R_k} \wedge x_j \notin [x_k]_{R_k}) \\ 0, & \text{otherwise}. \end{cases}$$  \hspace{1cm} (7)

Equation (6) shows that $\delta_k^{\text{indis}}(x_i, x_j)$ 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 (7) shows that $\delta_k^{\text{dis}}(x_i, x_j)$ 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 $\delta_k^{\text{indis}}(x_i, x_j)$ and $\delta_k^{\text{dis}}(x_i, x_j)$ for all $k(1 \leq k \leq [U])$ as in Equation (5), we obtain the percentage of equivalence relations that regard $x_i$ and $x_j$ as indiscernible objects. Note that in Equation (6), 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 2.1, $P_k$ 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 $T_h$. Thus we only count the number of equivalence relations that certainly evaluate the dissimilarity of $x_i$ and $x_j$.

From its definition, a large $\gamma(x_i, x_j)$ 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 $\gamma$ value, we consider that it represents excessively fine classification knowledge and refine it according to the following procedure (note that $R_0$ is rewritten as $R_1$ below for the purpose of generalization).

Let $R_t \in R$ be an initial equivalence relation on $U$. A refined equivalence relation $R'_t \in R'$ of $R_t$ is defined as

$$U/R'_t = \{P'_i \mid U - P'_i\},$$  \hspace{1cm} (8)

where $P'_i$ denotes a set of objects represented by

$$P'_i = \{x_j | \gamma(x_i, x_j) \geq T_h \}, \ \forall x_j \in U.$$  \hspace{1cm} (9)

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 $T_h$, $R_t$ is modified to include $x_j$ into the class of $x_i$.

2.3 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 $R'$. 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

3.1 Perfect equivalence relations

Our aim is to analyze the relationships between the threshold value $T_h$ of indiscernibility degree $\gamma$ and cluster numbers, while minimizing the influence of methods for determining initial equivalence relations in step 1. We prepared equivalence relations called perfect equivalence relations, which can classify the data into correct groups. Taking them as initial equivalence relations, we performed step 2 of the rough clustering several times by changing $T_h$ and observed the change of resultant clusters. We also performed clustering experiments on randomly disturbed perfect equivalence relations.
A perfect equivalence relation $R_i$ for object $x_i$ is denoted as follows.

$$U/R_i = \{P_i, U - P_i\}, \quad \text{(10)}$$

where

$$P_i = \{x_j | c[x_i] = c[x_j]\}, \quad \forall x_j \in U. \quad \text{(11)}$$

where $c[x_i]$ denotes the class label of $x_i$ assigned when creating the dataset. Obviously, the types of perfect equivalence relations in $R$ are equal to the number of classes in the dataset, because if objects $x_i$ and $x_j$ belong to the same class, $R_i$ and $R_j$ become identical.

### 3.2 Datasets

We artificially created a total of four numerical datasets named c3-1, c3-2, c5-1, and c5-2 shown in Table 1. Datasets c3-1 and c3-2 contain three clusters, and c5-1 and c5-2 contain five clusters respectively. The number of data points in each cluster was controlled to be substantially different and imbalanced, because the balanced data may induce special effect of $T_h$ on a specific range. The data points were generated based on a two-dimensional normal distribution for easy visualization; however, note that the geometric distribution of data points is not significant in this experiment because we used only their class labels for creating the perfect equivalence relations.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>cls 1</th>
<th>cls 2</th>
<th>cls 3</th>
<th>cls 4</th>
<th>cls 5</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>c3-1</td>
<td>52</td>
<td>40</td>
<td>93</td>
<td>-</td>
<td>-</td>
<td>185</td>
</tr>
<tr>
<td>c3-2</td>
<td>224</td>
<td>31</td>
<td>177</td>
<td>-</td>
<td>-</td>
<td>432</td>
</tr>
<tr>
<td>c5-1</td>
<td>52</td>
<td>171</td>
<td>148</td>
<td>215</td>
<td>55</td>
<td>641</td>
</tr>
<tr>
<td>c5-2</td>
<td>64</td>
<td>164</td>
<td>126</td>
<td>58</td>
<td>155</td>
<td>567</td>
</tr>
</tbody>
</table>

### 3.3 Procedures

The following procedure was applied to each dataset.

1. Form perfect initial equivalence relations: according to Eqs. (10) and (11), assign a perfect initial equivalence relation for each $x_i \in U$.

2. Disturb the initial relations: Select one of the following disturbance operation randomly at each time and apply it to initial relations. This process is repeated $\text{card}(P_i) \times \rho$ times, where $\rho$ denotes disturbance ratio (from 0.0 to 1.0, interval 0.2).

   - **Delete**: Randomly select one element in $P_i$ and remove it from $P_i$.
   - **Add**: Randomly select one element from $U$ and add it to $P_i$.
   - **Replace**: Randomly select one element from $P_i$ and replace it with randomly selected element in $U$.

3. Clustering: Apply the iterative refinement process of rough clustering to the disturbed initial equivalence relations and obtain clusters. This process is repeated by changing $T_h$ (from 0 to 1.0, interval 0.05). For each $T_h$, calculate the validity of clustering result according to the following measure.

   $$v_R(C) = \min\left(\frac{|X_R \cap C|}{|X_R|}, \frac{|X_R \cap C|}{|C|}\right),$$

   where $X_R$ and $C$ denote the obtained clusters and original classes, respectively.

### 3.4 Results and discussions

Figures 1-4 show the results on the four datasets respectively. Each of the figures consists of two sub-figures: Th-Number of clusters curves (left) and Th-Cluster Validity curves (right). The horizontal axis corresponds to the threshold value $T_h$ of indiscernibility degree $\gamma$. The vertical axis corresponds to the number of clusters or cluster validity for the left or right figure, respectively. Each figure contains six curves indexed by “$\rho = x$”, which corresponds to the ratio of disturbance of the perfect initial equivalence relations described previously.

Let us first see the global characteristics the curves. At $T_h = 0$, every equivalence relation was modified to include all objects. This means that, regardless of the characteristics of initial equivalence relations, all objects would
be grouped into the same cluster. Therefore the number of clusters was always 1 at $T_h = 0$. The cluster validity took a constant value which was dependent only to the class distribution of the dataset (around 0.5 or 0.3 for the datasets used here).

When $\rho = 0$, initial equivalence relations were identical to the perfect relations since no disturbance was applied. In this case the indiscernibility degrees were 0 for all pairs of objects belonging to different clusters, and 1 for those belonging to the same cluster. Therefore, correct clusters of validity=1 were formed for all values of $T_h > 0$ without any refinement.

If $\rho > 0$, situations become close to those of real-world datasets. The variety of initial equivalence relations drastically increase because of disturbance. Even a small difference of equivalence relations results in producing fine clusters due to the increase of total discrimination ability. Hence, without refinement of the relations, excessively large number of fine clusters would be produced. Let us first see the case of dataset c3-1 in Figure 1. For large values of $T_h > 0.8$, only a few equivalence relations satisfied the condition for refinement in Eq. (9). As most of the relations remained unchanged, the number of induced clusters kept high value - almost equal to the number of objects in the dataset. When $T_h$ became smaller, the number of equivalence relations to be refined increased. The refinement made classification coarser and made the number of clusters smaller, inducing the increase of cluster validity. The level of $T_h$ for starting this improvement was higher if $\rho$ was smaller, because at smaller $\rho$ initial equivalence relations were only slightly and
locally modified from the perfect equivalence. Therefore, the indiscernibility degree of each object pairs kept high value, while the types of equivalence relations are quite large. As $\rho$ becomes larger, more severe and global disturbance could occur. Since it induced the decrease of average level of the indiscernibility, the values of $T_h$ should be smaller to do the necessary refinement.

For $0.5 > T_h > 0.1$, the number of clusters kept 3 with the highest validity of 1. In this range, the method could produce the correct cluster assignment with the help of iterative refinement of the disturbed initial equivalence relations. The range became narrow as $\rho$ became small. For example, when $\rho = 0.6$ the range was about $2.7 > T_h > 1.8$ and when $\rho = 1.0$, there was no range of $T_h$ that could generate the correct cluster assignment. If there exists too much disturbance, the level of indiscernibility degrees for objects that should belong to the same cluster would be close to those of objects that belong to different clusters. Hence it would be difficult to form correct clusters, especially for small clusters.

For small values of $T_h < 0.1$, the number of clusters decreased to 1, followed by the decrease of cluster validity. In this range, too coarse cluster was obtained due to too much refinement of equivalence relations. Let us denote by $\min_{\gamma}$ the minimum value of indiscernibility degrees. Actually, for $Th < \min_{\gamma}$, the results are identical with the case of $T_h = 0$, due to the discrete property of indiscernibility degree.

The above characteristics were commonly observed for all the other datasets used in this experiment. It demonstrated that, by changing the threshold value of indiscernibility de-
gree, we could control the roughness of classification knowledge, namely, granularity of the data.

Furthermore, an interesting feature about the number of clusters was observed on all datasets. Around $T_h = 0.1 - 0.2$, there existed a short spike at the left end of the range for yielding the correct number of clusters. Although it could disappears on extremely disturbed cases, the convex features of the curve may be used for determining the best range of $T_h$ semi-automatically.

4 Conclusions

In this work, we have empirically investigated the characteristics of indiscernibility degree in rough clustering. By the use of perfect equivalence relations, we could observe more basic relationships between the threshold value of indiscernibility degree and resultant clusters, without the effect of methods for determining initial equivalence relations. The result demonstrated that the threshold parameter might be associated with roughness of knowledge, which also controls the granularity of dataset. Additionally, although it still requires exploratory approach, the convex shape of th-nc curve suggested the possibility of guiding appropriate range of the thresholds. It remains as a future work to investigate the reason why these spikes occur.

Acknowledgements

This work was supported in part by the Grant-in-Aid for Scientific Research (B) (No. 17300048) by the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References