Rough Set Extension Model of Incomplete Information System

Hou Yue
School of Electronic and Information Engineering, Lan Zhou Jiao Tong University
Lan Zhou 730070, P.R.China
email: houyue@mail.lzjtu.cn

Abstract
This paper investigated the incomplete information system in which situation of the more missing or absent unknown values. Based on the original characteristic relation, we proposed a new reflective relation which was controlled by the parameter alpha, beta. By analyzing illustrative examples and comparing to the original characteristic relation, this paper indicated the validity and practicability of the new-defined binary relation.

Keywords: rough set, incomplete information system, tolerance relation, similarity relation, characteristic relation

1. Introduction
Proposed by Poland mathematician Z. Pawlak in 1982, rough set theory is an emerging mathematical theory dealing with imprecise, uncertain and incomplete information [1, 2]. The main principle of rough set theory is to make approximate description to imprecise or uncertain knowledge utilizing the known repository [3]. Recently, rough set theory has achieved enormous success in the application of knowledge discovery.

The research object of classical rough set is the complete information system possessed with discrete attribute values, which means all the attribute values in the information system are known[4, 5]. However, the vast majority of information systems are incomplete in real problems. In general, concerning about unknown attribute values in incomplete information system there exist two explanations: 1) all the unknown attribute values exactly exist only to be missed, 2) all the unknown attribute values are considered to be lost but not be compared with. Kryszkiewicz has put forward the tolerance relation of the incomplete information system based on missing semantic and moreover make study on knowledge reduction [6, 7]. Stefanowski has proposed asymmetric similarity relation based on absent semantic [8]. In reality, however, the common situation always happened is that both missing semantic and absent semantic exist in incomplete information system at the same time, so using the above models would result difficulties. In order to use rough set to deal with the incomplete information system possessed with missing and lost unknown attributes, Grzymala-Busse has brought forward the characteristic relation, which is a generalized form combined with tolerance and similarity relation.

Yet, derived from characteristic relation, characteristic set has two unreasonable circumstances: 1) it is possible to make false judgment to classify two objects without any the same clear attributes into a set; 2) it is possible to separate two objects with large the same known attributes. This paper, aiming at the problems with more missing and absent values, has obtained more reasonable and realistic characteristic set under the analysis and discussion respectively based on new binary relation.

2 Basic Concepts
2.1. Incomplete Information System
An incomplete information system is a four-tuple class: \( S=\langle U, AT, V, f \rangle \). In this equation, \( U \) is a non-empty finite object set called the domain; \( AT \) is a non-empty finite attributes...
set; regarding $\forall a \in AT$, there is $a : U \rightarrow \mathcal{V}_a$. $\mathcal{V}_a$ is the range of attributes $a \in AT$ (including absent unknown attributes ? and missing unknown attributes *), $\mathcal{V}$ is the set of range of all attributes.

$$\mathcal{V} = \bigcup_{a \in AT} \mathcal{V}_a$$

(1)

Define $f$ as the information function, for $\forall a \in AT$, $\forall x \in U$, there is $f(x, a) \in \mathcal{V}_a$.

Table 1 is an incomplete information system analyzed in literature 6. Among the analysis, $U=\{1,2,3,4,5,6,7,8\}, AT=\{a,b,c\}=\{Temperature, Headache, Nausea\}, \mathcal{V}_a=\{Very_High, High, Normal\}$

<table>
<thead>
<tr>
<th></th>
<th>Temperature</th>
<th>Headache</th>
<th>Nausea</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>High</td>
<td>?</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Very-High</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>?</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>High</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>High</td>
<td>?</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>Normal</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Normal</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>8</td>
<td>*</td>
<td>Yes</td>
<td>*</td>
</tr>
</tbody>
</table>

### 2.2. Characteristic Relations

For the incomplete system shown in Table 1, no matter it is the tolerance relation or similarity relation, the domain can’t be classified because the tolerance relation only considers that all the unknown attributes are the missing type and the similarity relation is considered that all the unknown attributes are the absent type. Therefore, concerning that the incomplete system has both the missing and absent type, Grzymala-Busse has constructed the characteristic relation [9] as the following.

Definition 1: Assuming $S$ is an incomplete information system, for $\forall A \subseteq AT$, the expression of characteristic relation decided by $A$ is $K(A)$,

$$K(A) = \{(x, y) \in U^2 : \forall a \in A \land f(x, a) \neq ? \}$$

(2)

$$f(x, a) = f(y, a) \vee f(x, a) = * \vee f(y, a) = *$$

(3)

Definition 2: Assuming $S$ is an incomplete information system, and $A \subseteq AT$, so for $\forall X \subseteq U$, the lower-approximation and upper-approximate set of $X$ based on characteristic relation ($A$) are respectively regarded as $\mathcal{A}_K(X), \overline{\mathcal{A}_K(X)}$ and:

$$\mathcal{A}_K(X) = \{x \in U : K_A(x) \subseteq X\}$$

(4)

$$\overline{\mathcal{A}_K(X)} = \{x \in U : K_A(x) \neq \emptyset\}$$

(5)

In the equation, $K_A(x) = \{y \in U : (x, y) \in K(A)\}$.

In the incomplete information system, characteristic relation not only can deal with missing type, but also can process absent unknown attributes incomplete information system. If all the unknown attributes are considered as missing type, the characteristic relation $K_A$ degenerates to tolerance relation [6, 7]; from the other view, if all the unknown attributes in the incomplete system are considered as lost type, the characteristic relation $K_A$ degenerates to
asymmetrical similarity relation [9]. Therefore, the characteristic relation has preserved the relevant characters of the tolerance and also has preserved the relevant characters of asymmetrical similarity relation.

3. New Binary Relation
For the characteristic relation $K(A)$, because it has inherited the relevant characters of tolerance and similarity relation, it would make objects without any the same known attributes classify into a set or separate majority objects with known same attributes easily [10]. For instance, in Table 1 there is $(1,8) \in K(A), (4,5) \notin K(A)$. Hence, aiming at incomplete information system has relatively more absent and missing values, this paper has constructed a kind of new binary relation with parameters respectively.

3.1. Binary Relation with more Missing Values
Definition 3 [11]: Assuming $S$ is an incomplete information system with more missing values, regarding $\forall A \subseteq AT$, new binary relation decided by $A$ is represented as $R_{\alpha,\beta}^A$ and:

$$R_{\alpha,\beta}^A = \{(x,y) \in U^2 : (\forall b \in B (f(x,b) = f(y,b) \vee f(x,b) = * \vee f(y,b) = *)) \land$$

$$\left( |B|/|N_\alpha(x) \geq \alpha \right) \land \left( |C|/|B| \geq \beta \right) \right\}$$

(6)

With: $B = N_\alpha(x) \cap N_\beta(y)$
$C = M_\alpha(x) \cap M_\beta(y) \cap M_\alpha(y) \cap N_\beta(y)$
$N_\alpha(x) = \{a \in A : f(x,a) \neq ? \}$
$M_\alpha(x) = \{a \in A : f(x,a) \neq * \}$

$|X|$ represents the cardinal number of set $X$, $\alpha \in [0,1]$, $\beta \in [0,1]$.

3.2. Binary Relation with more Absent Values
Definition 4: Assuming $S$ is an incomplete information system with more absent values, regarding $\forall A \subseteq AT$, new binary relation decided by $A$ is represented as $K_{\alpha,\beta}^A$ and:

$$K_{\alpha,\beta}^A = \{(x,y) \in U^2 : (\forall b \in B (f(x,b) = f(y,b) \vee f(x,b) = * \vee f(y,b) = *)) \land$$

$$\left( |B|/|N_\alpha(x) \geq \alpha \right) \land \left( |C|/|B| \geq \beta \right) \right\}$$

(7)

With $B = M_\alpha(x) \cap M_\beta(y)$
$C = M_\alpha(x) \cap M_\beta(y) \cap M_\alpha(y) \cap N_\beta(y)$
$N_\alpha(x) = \{a \in A : f(x,a) \neq ? \}$
$M_\alpha(x) = \{a \in A : f(x,a) \neq * \}$

$|X|$ represents the cardinal number of set $X$, $\alpha \in [0,1]$, $\beta \in [0,1]$.

Binary relation $R_{\alpha,\beta}^A, K_{\alpha,\beta}^A$ in definition 4, 5 has introduced two parameters $\alpha$ and $\beta$ which only satisfy reflexive. Setting parameter $\alpha$ is to prevent the separation of two objects with enormous the same known attributes because of the existence of unknown attribute “?”. Setting parameter $\beta$ is to prevent two objects without or with only little the same known attributes are classified into a set because of the existence of unknown attribute “*”.

3.3. Algorithm Process
Algorithm of improved characteristic relation is shown as the following:

**Step 1:** Given an incomplete information system involving both missing and absent attributes.
Step 2: Through observation, the incomplete information system is divided into two situations:

If the system has more missing values, the characteristic relation $R^{\alpha, \beta}(A)$ is calculated according to Equation (4).

If the system has more absent values, the characteristic relation $K^{\alpha, \beta}(A)$ is calculated according to Equation (5).

Step 3: The improved characteristic relation between any two objects among all sample objects are obtained through the set to the threshold $\alpha$ and $\beta$.

The algorithm process chart is demonstrated in Figure 1:

![Algorithm Process Diagram](image)

4. Case Analysis

The incomplete information system with more missing values shown in Table 2, $U=\{O_i\}_{1 \leq i \leq 12}$, $AT=\{a,b,c,d\}$ is the set of all attributes.

According to the characteristic relation demonstrated in Definition 2, following characteristic sets can be obtained: $K_{AT}(O_1) = \{O_1, O_{11}, O_{12}\}$, $K_{AT}(O_2) = \{O_2, O_3\}$, $K_{AT}(O_3) = \{O_2, O_3\}$, $K_{AT}(O_4) = \{O_4, O_{11}, O_{12}\}$, $K_{AT}(O_5) = \{O_5\}$, $K_{AT}(O_6) = \{O_6\}$, $K_{AT}(O_7) = \{O_7, O_8, O_9, O_{11}, O_{12}\}$, $K_{AT}(O_8) = \{O_8\}$, $K_{AT}(O_9) = \{O_9, O_{11}, O_{12}\}$, $K_{AT}(O_{10}) = \{O_4, O_8, O_{10}, O_{11}\}$, $K_{AT}(O_{11}) = \{O_1, O_4, O_9, O_{11}, O_{12}\}$, $K_{AT}(O_{12}) = \{O_1, O_4, O_9, O_{11}, O_{12}\}$.
According to the characteristic relation demonstrated in definition 3, assuming $\alpha = \beta = 0.5$, results can be obtained:

$\mathit{R}_{\alpha,\beta}^{AT}(O_1) = \{O_1, O_{12}\}$, $\mathit{R}_{\alpha,\beta}^{AT}(O_2) = \{O_2, O_3\}$, $\mathit{R}_{\alpha,\beta}^{AT}(O_3) = \{O_2, O_3\}$, $\mathit{R}_{\alpha,\beta}^{AT}(O_4) = \{O_4\}$, $\mathit{R}_{\alpha,\beta}^{AT}(O_5) = \{O_5\}$, $\mathit{R}_{\alpha,\beta}^{AT}(O_6) = \{O_6\}$, $\mathit{R}_{\alpha,\beta}^{AT}(O_7) = \{O_7, O_9\}$, $\mathit{R}_{\alpha,\beta}^{AT}(O_8) = \{O_8\}$, $\mathit{R}_{\alpha,\beta}^{AT}(O_9) = \{O_7, O_{10}, O_{12}\}$, $\mathit{R}_{\alpha,\beta}^{AT}(O_{10}) = \{O_{10}\}$, $\mathit{R}_{\alpha,\beta}^{AT}(O_{11}) = \{O_{11}, O_{12}\}$, $\mathit{R}_{\alpha,\beta}^{AT}(O_{12}) = \{O_1, O_9, O_{12}\}$.

Clearly, $\mathit{R}_{\alpha,\beta}^{AT}(x)$ is better than $\mathit{K}_{\mathit{AT}}(x)$. For instance, object $O_5$ and $O_8$ have not a known equal attribute, but $O_8 \in \mathit{K}_{\mathit{AT}}(O_8)$, while $O_8 \notin \mathit{R}_{\alpha,\beta}^{AT}(O_8)$, the similar situation also has occurred at $O_5$ and $O_{11}$, $O_4$ and $O_{12}$. On the other side, $O_7$ and $O_9$ have greater possibility in similarity, but $O_7 \notin \mathit{K}_{\mathit{AT}}(O_9)$, while $O_7 \in \mathit{R}_{\alpha,\beta}^{AT}(O_9)$.

### Table 2. Incomplete Information System with more Missing Values

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O_1$</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$O_2$</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>$O_3$</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>$O_4$</td>
<td>*</td>
<td>2</td>
<td>*</td>
<td>1</td>
</tr>
<tr>
<td>$O_5$</td>
<td>*</td>
<td>?</td>
<td>*</td>
<td>1</td>
</tr>
<tr>
<td>$O_6$</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>$O_7$</td>
<td>3</td>
<td>?</td>
<td>*</td>
<td>3</td>
</tr>
<tr>
<td>$O_8$</td>
<td>*</td>
<td>0</td>
<td>0</td>
<td>*</td>
</tr>
<tr>
<td>$O_9$</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>$O_{10}$</td>
<td>1</td>
<td>*</td>
<td>*</td>
<td>?</td>
</tr>
<tr>
<td>$O_{11}$</td>
<td>*</td>
<td>2</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>$O_{12}$</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>*</td>
</tr>
</tbody>
</table>

The incomplete information system with more absent values shown in Table 3, $U=\{O_i\}_{1 \leq i \leq 12}$, $\mathit{AT}=(a,b,c,d)$ is the set of all attributes.

### Table 3. Incomplete Information System with more Absent Values

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O_1$</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$O_2$</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>$O_3$</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>$O_4$</td>
<td>?</td>
<td>2</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>$O_5$</td>
<td>*</td>
<td>*</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>$O_6$</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>$O_7$</td>
<td>3</td>
<td>*</td>
<td>?</td>
<td>3</td>
</tr>
<tr>
<td>$O_8$</td>
<td>?</td>
<td>0</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>$O_9$</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>$O_{10}$</td>
<td>1</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>$O_{11}$</td>
<td>?</td>
<td>2</td>
<td>1</td>
<td>*</td>
</tr>
<tr>
<td>$O_{12}$</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>?</td>
</tr>
</tbody>
</table>
According to the characteristic relation demonstrated in Definition 2, following characteristic set can be obtained:

\[ K_{AT}(O_1) = \{O_1\}, \quad K_{AT}(O_2) = \{O_2, O_3\}, \]
\[ K_{AT}(O_3) = \{O_2, O_3\}, \quad K_{AT}(O_4) = \{O_4, O_5, O_{11}\}, \quad K_{AT}(O_5) = \{O_5, O_6\}, \quad K_{AT}(O_6) = \{O_6\}, \]
\[ K_{AT}(O_7) = \{O_7, O_9\}, \quad K_{AT}(O_8) = \{O_8\}, \quad K_{AT}(O_{10}) = \{O_{10}\}, \]
\[ K_{AT}(O_{11}) = \{O_{11}\}, \quad K_{AT}(O_{12}) = \{O_1, O_2, O_{12}\}. \]

According to the characteristic relation demonstrated in Definition 3, assuming \( \alpha = \beta = 0.5 \), results can be obtained:

\[ K_{AT}(x) = \{O_1, O_{12}\}, \quad K_{AT}(O_2) = \{O_2, O_3\}, \quad K_{AT}(O_3) = \{O_2, O_3\}, \]
\[ K_{AT}(O_4) = \{O_4\}, \quad K_{AT}(O_5) = \{O_5\}, \quad K_{AT}(O_6) = \{O_6\}, \quad K_{AT}(O_7) = \{O_7\}, \]
\[ K_{AT}(O_8) = \{O_8\}, \quad K_{AT}(O_9) = \{O_9, O_{12}\}, \quad K_{AT}(O_{10}) = \{O_{10}\}, \quad K_{AT}(O_{11}) = \{O_{11}, O_{12}\}, \]
\[ K_{AT}(O_{12}) = \{O_1, O_2, O_{11}, O_{12}\}. \]

Seen from the results, we can find \( \text{K}_{AT}(x) \) is better than \( \text{K}_{AT}(x) \). For instance, object \( O_5 \) and \( O_6 \) have only one known equal attribute, but \( O_6 \in \text{K}_{AT}(O_5) \). \( O_1 \) and \( O_2 \) have greater possibility in similarity, but \( O_1 \notin \text{K}_{AT}(O) \), while \( O_2 \in \text{K}_{AT}(O) \). \( O_7 \) and \( O_8 \) have only two same attributes which cannot be judged whether the two attributes can be distinguished or not intuitively, but it considered to be indistinguishable under \( \alpha = \beta = 0.7 \). Therefore, by setting \( \alpha \) and \( \beta \), the new characteristic relation has better solved the unreasonable condition caused by classification of characteristic relation and also conformed the intuitive feeling dealing with data.

5. Conclusion

Application of rough set in the incomplete information system has become a study hot spot recently. In order to use rough set to deal with the incomplete information system possessed with missing and lost unknown attributes at the same time, Grzymala-Busse has proposed the concept of characteristic relation and characteristic set. However for different information system, according to the characteristic relation, it is possible that the classification would divide two objects into a set or separate two objects with large known same attributes. Therefore, aiming at two incomplete information system with more missing and absent values, this paper has proposed a new binary relation with two parameters. New constructed rough set model is better than general rough set model as long as setting \( \alpha \) and \( \beta \) properly.

Acknowledgements

This work was supported by the Youth Science Fund Project of Lanzhou Jiaotong University (No. 20130006).

References


