Knowledge reduction in incomplete decision tables using Probabilistic Similarity-Based Rough set Model

Nguyen Do Van¹  Koichi Yamada²  Muneyuki Unehara³
Department of Management and Information Systems Science
Nagaoka University of Technology
Nagaoka, Japan
Email: ¹ngdovan@stn.nagaokaut.ac.jp ²yamada@kjs.nagaokaut.ac.jp ³unehara@kjs.nagaokaut.ac.jp

Abstract— This research discovers how the frequency of attributes values affects to definitions of similarity relations in incomplete information systems. Assuming that the domain of attribute values is defined, we apply the probability of values appearing in data tables in order to measure the self-information of similarity. This is defined as the uncertainty of similarity. Based on that, set approximation is defined by giving a threshold. The method to derive reducts and core is also introduced. Finally, the merit of this similarity is clarified and compared with other approaches.

Index Terms—Incomplete information systems, Rough Set, Set approximation, Reducts, Core.

I. INTRODUCTION

Rough set theory [1, 2] has been developed as a mean to analyze vague description of objects. Objects in information systems with precise values have been dealt by the theory. However, problems arise when the values of some objects are unknown. Therefore, it is necessary to develop a theory which enables classifications of objects even if there is only partial information available. Rough set models in [3, 4] deal with indiscernibility based on similarity relations. In such approaches, a missing value is considered as a special value that may take any possible value. In fact, there is an array of methods to handle the incomplete objects. In this research, we study probability of similarity and propose a method of handling missing value in incomplete information systems based on self-information of uncertainty in similarity measurement.

II. BASIC CONCEPT [3,4,5]

A. Incomplete information systems

An information system is defined as a pair \( I = (U, A) \), where \( U \) is a non-empty finite set of objects called the universe and \( A \) is a non-empty finite set of attributes such that \( f_a : U \rightarrow V_a \) for every \( a \in A \). The set \( V_a \) is called the value set of \( a \).

If \( U \) contains at least one object with an unknown (missing or null) value, then \( I \) is called an incomplete information system, otherwise complete. In incomplete information systems, objects may contain several unknown attribute values. Unknown values are denoted by special symbol “*” in incomplete information systems.

B. Indiscernibility relation

The relation \( \text{IND}(P) \), \( P \subseteq A \) denotes a binary relation between objects that are indiscernible in terms of values of attributes in \( P \).

\[
\text{IND}(P) = \{(x, y) \in U \times U \mid a \in P, f_a(x) = f_a(y) \} \tag{2.1}
\]

Let \( I_P(x) = \{ y \in U \mid (x, y) \in \text{IND}(P) \} \) is the set of objects which are indiscernible by \( P \) with \( x \), and is called equivalence class.

C. Similarity relations

A similarity relation \( \text{SIM}(P) \), \( P \subseteq A \) denotes a binary relation between objects that are possibly indiscernible in terms of values of attributes. In incomplete information systems, the similarity relation is defined by:

\[
\text{SIM}(P) = \{(x, y) \in U \times U \mid a \in P, f_a(x) = f_a(y) \text{ or } f_a(x) = \ast \text{ or } f_a(y) = \ast \} \tag{2.2}
\]

The relation is reflexive and symmetric, but not transitive.

Let \( S_P(x) = \{ y \in U \mid (x, y) \in \text{SIM}(P) \} \) is the set of objects which are similar to \( x \) in terms of \( P \) in the sense of the above similarity relation. \( S_P(x) \) is called equivalence class instead of \( I_P(x) \) in the rest of this paper. In the same way, we may say that \( y \) is equivalent to \( x \) in terms of \( P \) instead of saying that \( y \) is similar to \( x \), when \( y \in S_P(x) \).

D. Rough sets

Rough sets in incomplete information systems are defined in the same way as in complete information systems. Let \( U/\text{SIM}(P) \) denote a classification, which is a family set \( \{ S_P(x) \mid x \in U \} \). Note that \( U/\text{SIM}(P) \) is not a partition of \( U \), unlike \( U/\text{IND}(P) \).

E. Set approximations

Let \( X \subseteq U, P \subseteq A \). \( P_X \) is the lower approximation of \( X \) in terms of \( P \), if and only if

\[
P_X = \{ x \in U \mid S_P(x) \subseteq X \} = \{ x \in X \mid S_P(x) \subseteq X \} \tag{2.3}\]

\( \overline{P_X} \) is the upper approximation of \( X \) in terms of \( P \), if and only if

\[
\overline{P_X} = \{ x \in U \mid S_P(x) \cap X \neq \emptyset \} = \bigcup \{ S_P(x) \mid x \in X \}. \tag{2.4}\]
F. Decision tables

An incomplete decision table (DT) is an incomplete information system combined with a set of decision attributes. When there is only a decision attribute, it is denoted by DT = (U, A ∪ {d}), where d is a distinguished attribute called decision, d ∉ A, and * ∉ V_d, where V_d is the value set of decision d. All elements of A are called conditions.

G. Generalized decision

The function \( \hat{d}_p : U \rightarrow F(V_d) \) where \( P \subseteq A \), and \( F(V_d) \) is the power set of \( V_d \), is defined as
\[
\hat{d}_p(x) = \{ i \mid i = f_d(y), y \in S_p(x) \},
\]
(2.5)
where \( f_d(y) \) is the decision value of an object \( y \).
\( \hat{d}_p \) will be called generalized decision in DT. \( \hat{d}_p \) determines which decision classes the object \( x \) may be classified to based on the available information on \( x \).

H. Reduct and core

A subset of conditional attributes \( RED \subseteq A \) is a reduct if the equivalence classes induced by \( RED \) are the same as the equivalence classes induced by all attributes in set \( A \) and no attribute can be removed from set \( RED \) without changing the equivalence classes.

In decision tables, reduction of knowledge that preserves generalizations for all objects is lossless from decision making standpoint. Reduct of a decision table is defined as follows.

A reduct \( P \) of decision table DT as minimal subset of \( A \), such that \( \hat{d}_P(x) \) of the reduct \( P \) is equal to \( \hat{d}_A(x) \) for any \( x \in U \). Formally, a set \( P \subseteq A \) is a reduct of decision table if and only if
\[
\hat{d}_P(x) = \hat{d}_A(x), \forall Q \subset P, \hat{d}_Q(x) \neq \hat{d}_A(x).
\]
(2.6)
Core is a set of all indispensable attributes, which appear in all reducts. Actually, core is the intersection of all reducts. The core possibly is empty. In such case, there is no indispensable attribute.

III. PROBABILISTIC SIMILARITY

First, we define minimum probability that each value of an attribute appears based on the frequency in the data set. Then, the minimum probability that two objects have the same values is calculated in order to measure the uncertainty of similarity.

The probability that a value \( j \in V_a \), \( j \neq * \) appears as a value of a certain object is between \( \frac{|V_a(j)|}{|U|} \) and \( \frac{|V_a(j)| + |V_a(*)|}{|U|} \), where \( V_a(j) \) and \( V_a(*) \) are sets of objects whose value of attribute "a" is "j" and "*", respectively. If \( f_a(x) \neq * \), that is, the attribute value of an object \( x \) is not missing, the probability that \( f_a(x) \) appears is between \( \frac{|V_a(x)|}{|U|} \) and \( \frac{|V_a(x)| + |V_a(*)|}{|U|} \).

Now, let us define probabilities \( \rho_a(x) \) and \( \rho_a(oba(j)) \), which are the minimum probability that an attribute value \( f_a(x) \neq * \) appears, and the minimum probability that value of attribute "a" is "j", respectively. \( oba(j) \) is an object whose value of attribute "a" is "j".

Definition 3.1. Let \( I = (U, A) \) be an incomplete information system, \( a \in A \). The minimum probability that each possible value \( j \in V_a \) appears is given by
\[
\rho_a(oba(j)) = \frac{|V_a(j)| + 1}{|U|}.
\]
(3.1)
Then, the minimum probability that the value \( f_a(x) \neq * \) appears can be defined:
\[
\rho_a(x) = \frac{|V_a(x)|}{|U|}, \text{ if } f_a(x) \neq *.
\]
(3.2)

Now, we define probability that objects \( x \) and \( y \) are similar to each other on attribute "a" as the minimum estimation of probability that the two objects have the same value of "a".

Definition 3.2. Let \( I = (U, A) \) be an incomplete information system. Given \( a \in A \) and \( x, y \in U \), probability of similarity between \( x \) and \( y \) denoted by \( \theta_a(x, y) \) is defined as the minimum estimation of probability that \( x \) and \( y \) take the same value on \( a \), and is given by the following equation:
\[
\begin{array}{ll}
1, & \text{if } f_a(x) = f_a(y) \neq *, \\
0, & \text{if } f_a(x) \neq *, f_a(y) \neq *, f_a(x) \neq f_a(y), \\
\rho_a(x), & \text{if } f_a(x) \neq *, f_a(y) = *, \\
\rho_a(y), & \text{if } f_a(x) = *, f_a(y) \neq *, \\
\sum_{j \in V_a} \{\rho_a(oba(j))\}^2, & \text{if } f_a(x) = f_a(y) = *.
\end{array}
\]
(3.3)

In the case where two values is certain (not missing), if two values are equal, the probability of similarity is exactly 1. On the other hand, if two values are different from each other, the probability of similarity is 0. When there is a missing value, it cannot define they are the same or not. In this case, we have to employ the probability.

Following is an example of an incomplete information.

<table>
<thead>
<tr>
<th>U</th>
<th>a1</th>
<th>a2</th>
<th>a3</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>*</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>x2</td>
<td>1</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>x3</td>
<td>2</td>
<td>1</td>
<td>*</td>
</tr>
<tr>
<td>x4</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>x5</td>
<td>*</td>
<td>*</td>
<td>2</td>
</tr>
<tr>
<td>x6</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Table I

An incomplete information system.

<table>
<thead>
<tr>
<th>U</th>
<th>a1</th>
<th>a2</th>
<th>a3</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>NG</td>
<td>2/6</td>
<td>4/6</td>
</tr>
<tr>
<td>x2</td>
<td>1/6</td>
<td>NG</td>
<td>NG</td>
</tr>
<tr>
<td>x3</td>
<td>1/6</td>
<td>1/6</td>
<td>NG</td>
</tr>
<tr>
<td>x4</td>
<td>2/6</td>
<td>1/6</td>
<td>4/6</td>
</tr>
<tr>
<td>x5</td>
<td>NG</td>
<td>NG</td>
<td>4/6</td>
</tr>
<tr>
<td>x6</td>
<td>2/6</td>
<td>2/6</td>
<td>4/6</td>
</tr>
</tbody>
</table>

Table II

Appearing probabilities \( \rho_a(x) \).

NG = Not given
Consider the case where \( a = a_1, x = x_1 \) and \( y = x_2 \) in Table 1. The minimum probability that the value \( f_{a_1}(x_1) \) is the same as \( f_{a_1}(x_2) = 1 \) is \( \rho_{a_1}(x_2) = 1/6 \). Then, the minimum probability that \( f_{a_1}(x_1) = f_{a_1}(x_3) \) is obtained as follows;

- Probability of \( f_{a_1}(x_1) = f_{a_1}(x_3) = 1 \) is \( 1/6 \).
- Probability of \( f_{a_1}(x_1) = f_{a_1}(x_3) = 2 \) is \( 2/36 = 1/18 \).
- Probability of \( f_{a_1}(x_1) = f_{a_1}(x_3) = 3 \) is \( 3/36 = 1/12 \).

So, the probability that \( f_{a_1}(x_1) = f_{a_1}(x_3) \) is \( 1/36 + 1/36 + 4/36 = 6/36 \).

From the probability of matching, we can measure the uncertainty of similarity information in order to define the similarity between objects.

**Definition 3.3.** Let \( I = (U, A) \) be an incomplete information system, \( P \subseteq A \). Given \( x, y \in U \), the self-information of similarity measurement is denoted by

\[
\varphi_p(x, y) = - \sum_{a \in P} \log_2 \theta_a(x, y).
\]

(3.4)

Obviously \( \varphi_p(x, x) = 0 \), and \( \varphi_p(x, y) = \varphi_p(y, x) \).

By the definition, the amount of self-information contained in a probabilistic similarity depends only on the probability of that similarity: the smaller its probability, the larger the self-information associated with receiving the information in that the similarity indeed occurred.

**Definition 3.4.** Similarity relations based on probability: The similarity relation in incomplete information is modified as \( \text{SIM}_\alpha(P) \). \( P \subseteq A \) and can be defined as follows:

\[
\text{SIM}_\alpha(P) = \{(x, y) \in U \times U \mid \varphi_p(x, y) \leq \alpha\}.
\]

(3.5)

The given threshold \( \alpha \geq 0 \) is different in each system. This depends on the characteristic of attribute values of objects.

Let \( \text{SIM}_\alpha(P) = \{x \in U \mid (x, y) \in \text{SIM}_\alpha(P)\} \) is the set of objects which are similar with \( x \) on \( P \).

From table 2, it can be achieved that \( \varphi_p(x_1, x_2) = 4.75489 \); \( \varphi_p(x_1, x_3) = 0 \); \( \varphi_p(x_1, x_4) = 4.16993 \); \( \varphi_p(x_1, x_5) = 1.58496 \). Thus, given \( \alpha = 2 \), \( \text{SIM}_\alpha(P) = \{x \in U \mid (x, y) \in \text{SIM}_\alpha(P)\} \).

**Property 3.1.** Let \( I = (U, A) \) be an incomplete information system, \( P \subseteq A \). Given \( x, y, z \in U \), if \( \varphi_p(x, y) < \varphi_p(x, z) \), then \( y \) is more similar with \( x \) than \( z \).

**Property 3.2.** Let \( I = (U, A) \) be an incomplete information system, \( P \subseteq A \). \( \forall x, y \in U \), \( \varphi_p(x, y) \leq \varphi_p(x, x) \).

**Property 3.3.** Let \( I = (U, A) \) be an incomplete information system, \( P \subseteq A \). \( \forall x, y \in U \), \( \varphi_p(x, y) \leq \varphi_p(x, x) \).

IV. PROBABILISTIC SIMILARITY-BASED ROUGH SET MODEL

Rough set in incomplete information systems defined in section 2 is defined newly.

Let \( X \subseteq U, P \subseteq A \), \( \text{app}_p^\alpha X \) is PS-lower approximation of \( X \), if and only if

\[
\text{app}_p^\alpha X = \{x \in U \mid S_p^\alpha(x) \subseteq X\} = \{x \in X \mid S_p^\alpha(x) \subseteq X\}. \tag{4.1}
\]

---

**Properties:** Let \( I = (U, A) \) be an incomplete information system, \( X, Y \subseteq U \) and \( P, Q \subseteq A \). The properties of the original rough set \([1,2]\) are also hold for this rough set model. Besides that, there are some new properties of the new model:

1. If \( \alpha \leq \beta \Rightarrow \text{app}_p^\alpha X \supseteq \text{app}_p^\beta X \)
2. If \( \alpha \leq \beta \Rightarrow \text{app}_p^\alpha X \subseteq \text{app}_p^\beta X \)
3. If \( \alpha \leq \beta \Rightarrow \text{BN}_p^\alpha X \subseteq \text{BN}_p^\beta X \)
4. If \( \alpha \leq \beta \Rightarrow \text{AC}_P^\alpha X \geq \text{AC}_P^\beta X \)
5. \( \lim_{\alpha \to \infty} \text{app}_p^\alpha X = U \)
6. \( \lim_{\alpha \to \infty} \text{app}_p^\alpha X = \phi \)

V. REDUCT AND CORE

In Probabilistic Similarity Based rough set, we have to use all attributes to define the similarity. Thus, the original method to induce reduct and core is not applicable. Following is a method to derive reduction in our approach.

**Definition 5.1.** The similarity between two objects \( x, y \in U \) in terms of \( P \subseteq A \) is defined as

\[
\text{SB}_p^\alpha(x, y) = \begin{cases} 1 & \text{if } \varphi_p(x, y) \leq \alpha \\ -1 & \text{if } \varphi_p(x, y) > \alpha \end{cases}. \tag{5.1}
\]

**Definition 5.2.** The comparison function between two partitions induced by similarity relation in terms attribute sets \( P, Q \subseteq A \) in an incomplete information system is defined as

\[
\varphi^\alpha(P, Q) = \begin{cases} -1 & \text{if } \exists \exists(x, y) \in U \times U, \varphi^\alpha(P, Q) = -1 \\ 1 & \text{otherwise} \end{cases}. \tag{5.2}
\]

If \( \varphi^\alpha(P, Q) = 1 \), the two similarity relations in terms of \( P, Q \) make the same partitions.

**Definition 5.3.** The comparison function between two partitions induced by similarity relation in terms attribute sets \( P, Q \subseteq A \) in an incomplete decision table is defined as

\[
\varphi^\alpha(P, Q) = \begin{cases} -1 & \text{if } \exists \exists(x, y) \in U \times \Gamma^x, \varphi^\alpha(P, Q) = -1 \\ 1 & \text{otherwise} \end{cases}. \tag{5.3}
\]

where \( \Gamma^x = \{x \in U, d(z) \neq d(x)\} \)

**Proposition 5.4.** The attribute \( a \in A \) is indispensable in \( A \) if
and only if \( \sigma^\alpha (A - \{a\}, A) = -1 \) .

(5.4)

This proposition is applied to both incomplete information systems and in incomplete decision tables. \( \sigma^\alpha (A - \{a\}, A) = -1 \) means if \( a \) is removed from \( A \), the partition made by similarity relation in terms of \( A - \{a\} \) is different from the partition based on \( A \). Hence, \( a \) is indispensable in \( A \).

**Definition 5.5.** The core of \( A \) is the set of all indispensable attributes

\[
\text{core}(A) = \{ a \in A \mid \sigma^\alpha (A - \{a\}, A) = -1 \}. 
\]

(5.5)

**Proposition 5.6.** A subset \( P \subseteq A \) is a reduct of \( A \) for threshold \( \alpha \) if and only if

1. \( \sigma^\alpha (P, A) = 1 \).
2. \( \sigma^\alpha (P - \{a\}, P) = -1 \) for every \( a \in P \).

(5.6) (5.7)

VI. DISCUSSION

Numerous methods to handle missing value in incomplete information system have been discussed and tested. In such methods, several ones [6,7,8,9] concentrate on probability of matching values.

In [6,7], a “Method of possible worlds” is proposed. Missing values are replaced by all possible values in the set to derive a family of possible complete information systems. Based on that, approximation set is induced as the union of all approximation of derived information systems. Therefore, to induce equivalence classes for an incomplete information system, in which there are \( n \) missing values in an attribute \( a \), we have to derive possible equivalence classes of \( \forall A \)\(^n \) information systems.

In [8], the authors applied probability estimation of the unavailable induced from the available attributes to depict the suited degree between objects. The probability of matching between of \( x \) and \( y \) on attribute \( a_j \) is denoted as \( p(a_j) \). In equivalence relation \( p(a_j) = 1 \). On the other hand, if any attribute of the two objects is missing, \( p(a_j) \) is relative to the cardinality of all the suited attributes \( \forall a_j \mid f_{a_j}(x) = f_{a_j}(y) \neq \emptyset \). Then an expected function to define similarity relation is expressed by multiplying the probability of matching on all attribute. In this approach, only the cardinality of attributes, in which attribute values of objects are certain, controls the expected function.

In [9], similarity relation is defined based on equivalence degree

\[
\phi_p(x, y) = \prod_{a \in P} p(f_a(x) = f_a(y))
\]

where \( p(f_a(x) = f_a(y)) \) is the probability of \( f_a(x) = f_a(y) \). It means objects are similar to each other by some degree. Two objects are similar if equivalence degree is larger than 0. However, the probability \( p(f_a(x) = f_a(y)) \) has not been defined yet.

In the probabilistic approaches discussed above, all of them ignore the frequencies of attributes values. In our approach, the frequency of attributes values pays a vital attention to the relationship between objects. In some kind of attribute, one value may dominate in data tables. For example, when the age of students in the first year of undergraduate is valued, the missing values are likely “18 years old” which is the most common values. Another example is shown in the table below:

**TABLE III**

<table>
<thead>
<tr>
<th>Car</th>
<th>Size</th>
<th>Max-speed</th>
<th>Mileage</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Full</td>
<td>High</td>
<td>Low</td>
<td>Excel</td>
</tr>
<tr>
<td>2</td>
<td>Full</td>
<td>High</td>
<td>Mid</td>
<td>Good</td>
</tr>
<tr>
<td>3</td>
<td>Full</td>
<td>High</td>
<td>Mid</td>
<td>Good</td>
</tr>
<tr>
<td>4</td>
<td>*</td>
<td>High</td>
<td>*</td>
<td>Good</td>
</tr>
<tr>
<td>5</td>
<td>Compact</td>
<td>High</td>
<td>*</td>
<td>Good</td>
</tr>
<tr>
<td>6</td>
<td>Full</td>
<td>Low</td>
<td>Low</td>
<td>Good</td>
</tr>
<tr>
<td>7</td>
<td>Full</td>
<td>*</td>
<td>Low</td>
<td>Good</td>
</tr>
<tr>
<td>8</td>
<td>Full</td>
<td>High</td>
<td>Low</td>
<td>Good</td>
</tr>
<tr>
<td>9</td>
<td>Compact</td>
<td>Low</td>
<td>High</td>
<td>Poor</td>
</tr>
<tr>
<td>10</td>
<td>*</td>
<td>Low</td>
<td>High</td>
<td>Poor</td>
</tr>
</tbody>
</table>

In table 3, the set of observed objects is \( X_{\text{Good}} \) (Objects with \( d=\text{Good} \)). The size of Car 4 is missing. Due to the frequency of “Full” appearing in the set “Size”, the probability of matching between missing value and “Full” is larger than that between missing value and “Compact”.

Furthermore, self-information is a measure of the information content associated with the outcome of random variable. This assists us to measure the uncertainty of similarity. The given threshold allows us to decide which level of uncertainty is acceptable with the system.

Besides that, from the self-information, we can define a family of set approximation, which may connect to the fuzzy-set. This leads to a deeper research on fuzzy rough set.

**REFERENCES**


