퍼지 및 지능시스템학회 논문지 2001, Vol. 11. No. 5 pp. 441-447

Design of the Integrated Incomplete Information Processing System based on Rough Set

Gu-Beom Jeong*, Hwan-Mook Chung, Guk-Boh Kim and Kyung-Ok Park

*Sang Ju National University, Catholic University of Daegu, Daejin University and Samsung Electronics co., LTD

Abstract

In general, Rough Set theory is used for classification, inference, and decision analysis of incomplete data by using approximation space concepts in information system. Information system can include quantitative attribute values which have interval characteristics, or incomplete data such as multiple or unknown(missing) data. These incomplete data cause the inconsistency in information system and decrease the classification ability in system using Rough Sets. In this paper, we present various types of incomplete data which may occur in information system and propose INcomplete information Processing System(INiPS) which converts incomplete information system into complete information system in using Rough Sets.

Key Words : Rough Set, approximation space, incomplete information system

1. Introduction

Rough Set theory, introduced by Pawlak[6] and discussed in greater detail in [7, 8] is a technique for dealing with uncertainty and for identifying cause-effect relationships in information system. Rough Set theory is used for classification and inference of uncertain data in information system by using indiscernibility relation and approximation concepts. But if information system includes incomplete information such as quantitative attribute values, multiple attribute values, and null values, it will decrease the classification ability of Rough Set and cause an error in the result from the reasoning.

In this paper, we present various types of incomplete information which may occur in information system and propose ASPA (Approximation Space Partition Approach), ORE (Object Relation Entropy), ARE(Attribute Relation Entropy), and SAPA(Sub-Attribute Partition Approach) in order to process incomplete information. ASPA converts quantitative attribute values into qualitative attribute values. ORE and ARE substitutes similar attribute values for null attribute values and inconsistent attribute values. SAPA partitions multiple attribute values into unitary attribute values and then, extends multiple attribute values to sub-attribute in information system. And we design INcomplete information System Processor (INiSP) to use these various types of incomplete information processing methods efficiently.

NiSP is made up of RC(Rough Classifier), ASPM(Approximation Space Partition Module), EM(Entropy Module) and SAPM(Sub-Attribute Partition Module). RC, a main processing module of INiSP, performs interface function between users or each module, analyzes and manages incomplete information system. And ASPM, EM and SAPM convert incomplete information into complete information according to the control of RC.

접수일자 : 2001년 2월 5일
완료일자 : 2001년 7월 18일

2. Basic concepts of Rough Set and Incomplete Information

2.1 Rough Set

Rough Set involve the following :
U is the universe, which cannot be empty.
R is the indiscernibility relation, or equivalence relation.
A = (U, R), an ordered pair, is called approximation space.

$[x]_R$ denotes the equivalence class of R containing x, for any element x of U, elementary sets in A - the equivalence classes of R, definable set in A - any finite union of elementary sets in A. Therefore, for any given approximation space defined on some universe U and having an equivalence relation R imposed upon it, U is partitioned into equivalence classes called elementary sets which may be used to define other sets in A.

Given that $X \subseteq U$, X can be defined in terms of the definable sets in A by the following :

lower approximation of X in A is the set
$$R^*X = \{x \in U \mid [x]_R \subseteq X\},$$
upper approximation of X in A is the set
$$R_*X = \{x \in U \mid [x]_R \cap X \neq 0\}.$$

Another way to describe the set of approximations is as follows:

Given the upper and lower approximations R^*X and R_*X. The boundary region of X is $BN_R(X) = R^*X - R_*X$. X is called R-definable if and only if $R^*X = R_*X$. Otherwise, $R^*X \neq R_*X$ and X is rough with respect to R. $\alpha_A(X) = |R_*X| / |R^*X|$ called the accuracy of approximation, where $|x|$ denotes the cardinality of X. Obviously, $0 \leq \alpha A(X) \leq 1$.

2.2 Incomplete Information

If we can't precisely deal with incomplete information in application of real world as an expert system, we may not have an efficient system construction and depend upon the results of inference. According to this, studies for solving incomplete information have been continued in information system using Rough Set[2, 4, 5, 10, 11].

Various types of incomplete information which may occur in information system are as follows:

(1) discretization of quantitative attributes,
(2) imprecise descriptors,
(3) unknown(missing) descriptors,
(4) multiple descriptors.

Incompleteness of type (1) is an essential issue in information processing for the Rough Set analysis. Attributes creating the information system are divided, in general, into qualitative and quantitative ones. An original domain of a quantitative attribute is usually a subset of a interval while the domain of a qualitative attribute is a finite set of qualitative terms, usually of a low cardinality. In practice, values of quantitative attributes are rarely directly used in the Rough Set analysis. Instead, prior to the analysis, they are interpreted in qualitative terms, e.g. low, medium, high etc. So, the original domain(interval) is divided into few subintervals corresponding to the qualitative terms. Bounds of these subintervals are established according to norms, conventions, traditions existing in the field of a given application. It must be noticed, however, that such a definition is more or less arbitrary and may influence the results of the Rough Set analysis.

Incompleteness of type (2) appears when instead of a precise value of a quantitative attribute for a given object (i.e. a single descriptor) a subinterval of possible values is known. For instance, the statement "the temperature is between 35℃ and 40℃" may result from an imprecise measurement of the attribute value.

Incompleteness of type (3) refers to an unknown (missing) value of the descriptor for pair [*object, attribute*], so-called null value.

Incompleteness of type (4) occurs when instead of a single value of a descriptor for pair [*object, attribute*], a finite set of attribute values is known(i.e. the object is described by a multiple descriptor).

3. Incomplete Information Processing System

3.1 Incomplete Information Process

In this section, we propose incomplete information processing approach that include ASPA, ORE, ARE, and SAPA. These methods are based upon Rough Set theory for constructing complete information system in Fig. 1.

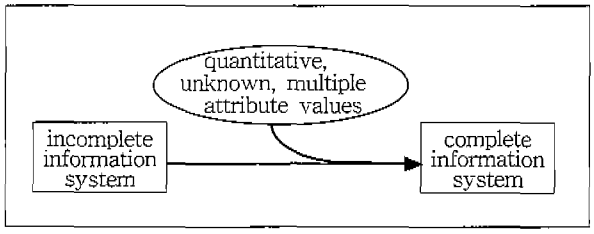

Fig. 1. efficient incomplete information process

3.1.1 Approximation Space Partition Approach (ASPA)

ASPA is a method which converts quantitative attribute values into qualitative attribute values in order to remove incompleteness occurring when the value of [*object, attribute*] is quantitative. For this, we partition quantitative attribute values in information system into sub-intervals included in approximation space in Rough Set. Each subinterval is substituted for 'low, medium, high', etc. corresponding to the qualitative linguistic terms and we can adapt range symbols to each linguistic term.

Incompleteness of quantitative attribute values, included in an approximation space in Rough Set, exists in the duplicated boundary region among classes. This boundary region becomes indiscernibility region and the bound of $BR[P]$ in Fig. 2.

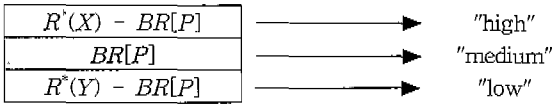

Fig. 2. approximation space partition for quantitative attribute values

A. Boundary Region Decision

ASPA generates the maximum or minimum attribute values of objects included in the upper approximations, and then decides the duplicated boundary region among classes. And the decision methods are performed by the following definitions:

[Definition 1] Consider the minimum or maximum values, included in the upper approximations $R^*(X)$ and $R^*(Y)$ of Set X and Y, as $LV(X)$, $HV(X)$, $LV(Y)$ and $HV(Y)$. Given that the boundary region among classes exists, the maximum or minimum values of $BR[X]$ and $BR[Y]$, the boundary region of $R^*(X)$ and $R^*(Y)$, defines as follows:

$BR[X]_{min} = \min\{x \mid LV(X) \le BR[x_i] \le BR[X]_{max},\ X \neq 0\}$,
$BR[X]_{max} = \max\{x \mid BR[X]_{min} \le BR[x_i] \le HV(X),\ X \neq 0\}$,
$BR[Y]_{min} = \min\{x \mid LV(Y) \le BR[x_i] \le BR[Y]_{max},\ Y \neq 0\}$,
$BR[Y]_{max} = \max\{x \mid BR[Y]_{min} \le BR[x_i] \le HV(Y),\ Y \neq 0\}$. (1)

[Definition 2] Boundary region $BR[P]$ defines by the following [Definition 1].

$BR[P] = \{x \mid BR[P]_{min} \le x_k \le BR[P]_{max}\}$. (2)

B. Subinterval Decision

Subintervals $BR[P]$, $RS(X)$ and $RS(Y)$, partitioned according to the upper approximations of Set X and Y, are divided into the following two cases, that is the case of the boundary region $BR[P] = 0$ and the case of the boundary region $BR[P] \neq 0$, and then decided.

If $BR[P] = 0$ then (3)
$RS(X) = \{x \mid LV(X) < x_k \le HV(X)\}$,
$RS(Y) = \{x \mid LV(Y) \le x_k < HV(Y)\}$.

If $BR[P] \neq 0$ then (4)
$BR[Y]_{min} \le BR[X]_{min} \le BR[Y]_{max} < BR[X]_{max}$
$\Rightarrow BR[P] = \{x \mid BR[X]_{min} \le x_k \le BR[Y]_{max}\}$,
$RS(X) = \{x \mid BR[Y]_{max} < x_k \le HV(X)\}$,
$RS(Y) = \{x \mid LV(Y) \le x_k < BR[X]_{min}\}$,
$BR[X]_{min} \le BR[Y]_{min} \le BR[X]_{max} \le BR[Y]_{max}$
$\Rightarrow BR[P] = \{x \mid BR[Y]_{min} \le x_k \le BR[X]_{max}\}$,
$RS(X) = \{x \mid LV(X) \le x_k < BR[Y]_{min}\}$,
$RS(Y) = \{x \mid BR[X]_{max} < x_k \le HV(Y)\}$,
$BR[X]_{min} \le BR[Y]_{min} \le BR[X]_{max} \le BR[Y]_{max}$
$\Rightarrow BR[P] = \{x \mid BR[X]_{min} \le x_k \le BR[X]_{max}\}$,
$RS(X) = \{x \mid LV(Y) \le x_k < BR[X]_{min}\}$,
$RS(Y) = \{x \mid BR[X]_{max} < x_k \le HV(Y)\}$,
$BR[X]_{min} \le BR[Y]_{min} \le BR[Y]_{max} < BR[X]_{max}$
$\Rightarrow BR[P] = \{x \mid BR[Y]_{min} \le x_k \le BR[Y]_{max}\}$,
$RS(X) = \{x \mid LV(X) \le x_k < BR[Y]_{min}\}$,
$RS(Y) = \{x \mid BR[Y]_{max} < x_k \le HV(Y)\}$.

If we give range symbols to subintervals $RS(X)$, $BR[P]$, and $RS(Y)$ partitioned according to this procedure, conversion as qualitative attribute values for quantitative attribute values is completed.

3.1.2 Object Relation Entropy and Attribute Relation Entropy

A. Object Relation Entropy(ORE)

ORE is used when an unknown(missing) value or condition attribute value of a decision attribute is identical, but a decision attribute value differs.

[Definition 3] ORE Eo(xt) for object x∈U defines as follows:

$$Eo(x_t) = -\sum_j (\rho_j(R))\ [P_i \log_2(P_i)]$$
where $1 \le i \le n$, $1 \le j \le m$ and $1 \le t \le l$. (5)

Class i is the equivalence class of a condition attribute j which denotes condition attributes C_1, C_2, ⋯, C_m, and t is the object of class containing the specified attribute value. The term $\rho_j(R)$ is the roughness for the condition attribute j, and $P_i \log_2(P_i)$is the probability of the equivalence class i for the condition attribute j. Given that ci is the number of elements in class i and C_m is the number of whole elements in all equivalence classes, the probability P_i that the specified element may exist in class i is $P_i = c_i / C_m$.

■ In case that the decision attribute value $a_d(x_t)$ of (x_t, d) is null, or condition attribute values are identical but decision attribute values differ, the substituted decision attribute values, by using formula (5), are decided as follows:

① We get $E_O(x_t^1) \sim E_O(x_t^{l-t})$ by adapting possible attribute values (except null or identical attribute values in $a_d(x_1) \sim a_d(x_{l-t})$) to $a_d(x_t)$ one by one. In this case, $E_O(x_t^1) \sim E_O(x_t^{1-t})$ becomes ORE calculated in terms of decision attribute values applicable to null values.

② we substitute a null value or imprecision value of $a_d(x_t)$ for a decision attribute value applied to $E_O(x) = \min\{E_O(x_t^1), E_O(x_t^2), \cdots, E_O(x_t^{l-t})\}$.

B. Attribute Relation Entropy (ARE)

ARE, which revised formula (5), is used when a condition attribute value is null, or priority of multiple attribute value is calculated.

[Definition 4] ARE $E_A(x_t)$ for the object $x \in U$ of Rough Set is defined as follows:

$$E_A(x_t) = (\rho_j(R))\ [P_i \log_2(P_i)]$$
where $1 \le i \le n$, $1 \le j \le m$ and $1 \le t \le l$. (6)

■ In case that a condition attribute value $a_k(x_t)$ of (x_t, A_k) is null value, the substituted attribute value, by using formula (6), is decided as follows:

① we get $E_A(x_t^1) \sim E_A(x_t^{l-t})$ by adapting the possible values of the condition attribute Ak (except null or identical attribute values in $a_k(x_1) \sim a_k(x^{l-t})$) to $a_k(x_t)$ one by one. In this case, $E_A(x_t^1) \sim E_A(x_t^{l-t})$ becomes ARE calculated in terms of condition attribute values applicable to null values.

② we substitute a null value of $a_k(x_t)$ for a condition attribute value applied to $E_A(x) = \min\{E_A(x_t^1), E_A(x_t^2), \cdots, E_A(x_t^{l-t})\}$.

3.1.3 Subattribute Partition Approach

SAPA is a method for converting multiple attribute values in information system into sub-attributes of Single attribute values, and then extends multiple attribute values to the attribute item in information system.

■ Consider a condition attribute $C \subseteq A$ and object $x_l \in U(l = 1, \cdots, k;\ k$ is the number of whole objects existing in information system) in information system $IS = \{U, A, V\}$ Given that [object, attribute], which denotes a condition attribute $p_i \in C$ $(1 \le i \le n;\ n$ is the number

of whole condition attributes in information system), is $MDes(x_l, p_l) = \{v_l : v_l \in V_i\}$, information system IS is generalized as follows:

$$S_{MDes} = \{ U, A, V, MDes(x_l, p_l) \}. \quad (7)$$

[Definition 5] Single Sub-attribute $Des(x_l, p_l^j)$, subdivided from an attribute value vi of a multiple condition attribute $MDes(x_l, p_l)$, defines by the following:

$$Des(x_l, p_l^j) = v_l^j, \text{ for } j = 1, \ldots, m. \quad (8)$$

Where m denotes the maximum number of multiple condition attribute values in (x_l, p_l), and j is the sequence of dividing multiple condition attribute values into sub-attributes.

■ In case that the condition attribute value of (x_l, p_l) is the multiple attribute value, it is dealt as follows:

① In case that multiple attribute values of (x_l, p_l) are quantitative or null, we remove incomplete attribute values by using an ASPA or ARE in advance.
② As a result of analyzing multiple attribute values of $MDes(x_l, p_l)$, we partition attributes which have the higher relation into sub-attribute unit without calculating priority when multiple attribute values are not divided by different attributes precisely or have the higher relation among attribute values.
③ For the multiple attribute value v_l of (x_l, p_l), we calculate priority in terms of ARE and decide priority of multiple attribute values, beginning with the low entropy of attribute values in order.
④ we subdivide the multiple attribute value v_l^j, whose priority is decided, priority, into sub-attribute according to formula (8), provided we adapt from the sub-attribute values which have the high priority in turn.
⑤ Given that v_l^j is null, we deal with 'don't care'.

3.2 Design of the INcomplete information Processing System(INiSP)

After all classifying information in information system or creating inference rules, incompleteness still remains with the results as long as we don't remove incompleteness of information itself. In this section we propose INiPS, using ASPA, ORE, ARE and SAPA presented in above section, to remove various incompleteness occurring in information system using Rough Set.

The structure of INiPS is shown in Fig. 3 that is an unitary incomplete information processing system which converts incomplete information system into complete information system. Main functions of INiPS components are as follows:

① RC(Rough Classifier) is a main processing module of INiPS, performs interface function between users or each module and manages information system. Likewise, it analyzes types of attribute and incomplete in incomplete information system and classifies an approximation space.
② ASPM(Approximation Space Partition Module) converts, by using ASPA, quantitative attribute values in incomplete information system into qualitative attribute values.
③ EM(Entropy Module) substitutes null attribute values or inconsistency attribute values in incomplete information system by using ORE and ARE for similar attribute values and calculate priority of multiple attribute values.
④ SAPM(Sub-Attribute Partition Module) extends a decision table in information system to sub-attribute according to priority of multiple attribute values calculated in EM.

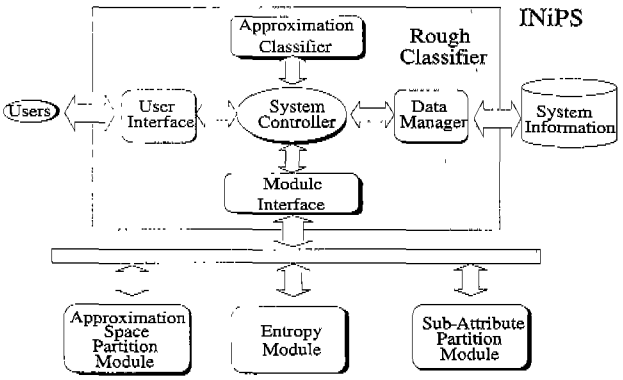

Fig. 3. Structure of INiPS

4. Application Cases

In this section, we study, through application cases, about the processing procedure and its results of RC, ASPS, EM, or SAPM module of INiPS, proposed as incomplete information processing methods, and induce inference rules.

4.1 Cases of incomplete information system

In application cases about incomplete information system processing methods proposed in this paper, we reconstruct and use a car relation table used in Hu[12], as in car efficiency comparison in Table 1.

4.2 Incomplete information processing of INiPS

Incomplete information processing of Table 1 is in the same line with the procedure of INiPS.

4.2.1. Incomplete information system analysis

According to RC, we analyze information of Table 1 and its results are as follows:

① Object number: 16, $\{x_1, x_2, \cdots, x_{16}\}$.
② Attribute:
- condition attribute = {a, b, c, d, e, f, g}.
- decision attribute = {m}.

Table 1. Car efficiency comparison

U	a	b	c	d	e	f	g	m
x_1	0	0	1,6	1,130	2	1	0	1
x_2	0	0	1,6	1,297	2	1	1	1
x_3	0	0	1	1,100	1	1	1	1
x_4	0	1	0	790	0	2	1	2
x_5	0	1	0	1,068	0	1	1	1
x_6	0	1	0,6	1,187	1	1	0	1
x_7	1	0	0	Λ	0	2	1	2
x_8	1	0	1	698	1	1	1	2
x_9	0	0	1	1,210	1	1	1	1
x_{10}	1	1	0	1,023	0	1	1	2
x_{11}	0	0	1,t	950	2	1	1	1
x_{12}	0	0	1,t	Λ	2	1	0	1
x_{13}	1	0	0	1,039	1	2	1	1
x_{14}	1	0	0	1,039	1	2	1	2
x_{15}	0	0	0	950	1	2	1	2
x_{16}	0	0	0,6	1,000	1	2	0	Λ

1. Maker (a: A→'0', B→'1'),
2. fuel system (b: ECU→'0', EFI→'1'),
3. engine (c: small→'0', medium→'1', 6 cylinder→'6', turbocharge→'t'),
4. weight (d),
5. power (e: low→'0', medium→'1', high→'2'),
6. compression ratio (f: low→'0', medium→'1', high→'2'),
7. transmitter (g: auto→'0', manual→'1'),
8. mileage (m: medium→'1', high→'2').

③ Attribute containing incomplete information : {c}, {d}, {m}.
- attribute {c} : multiple values = {'0', '1', '6', 't'}.
- attribute {d} : quantitative values = {698(min), 1,297(max)}, null values = {d_7, d_{12}}.
- attribute {m} : inconsistency values = {m_{13}, m_{14}}, null value = {m_{16}}.

④ Equivalence class of decision attribute {m} : X1, X2.
- X1 = {$x_1, x_2, x_3, x_5, x_6, x_9, x_{11}, x_{12}, x_{13}$}.
- X2 = {$x_4, x_7, x_8, x_{10}, x_{14}, x_{15}$}.

4.2.2. Incomplete information processing

A. Processing of quantitative attribute values : ASPM

In decision attribute {m}, the indiscernibility relation IND/d of weight {d} and attribute values of the domain of the upper approximations R^*_dX1 and R^*_dX2 are classified as follows:

X1 = {$x_1, x_2, x_3, x_5, x_6, x_9, x_{11}, x_{12}, x_{13}$},
X2 = {$x_4, x_7, x_8, x_{10}, x_{14}, x_{15}$},
$IND/$(d, m) = {{$x_1, x_2, x_3, x_5, x_6, x_9, x_{11}, x_{13}$}, {$x_4, x_8, x_{10}, x_{14}, x_{15}$}},
R^*_dX1 = {950, 1068, 1100, 1130, 1187, 1210, 1297, 1039},
R^*_dX2 = {698, 790, 950, 1023, 1039}.

The boundary region BR[d] and subinterval RS_d(X1), RS_d(X2) and range symbols for attribute values of weight {d} are as follows :

$1{,}039 < RS_d(X1) \le 1{,}297$; $RS_d(X1)$ = '2',
$950 \le BR[d] \le 1{,}039$; $BR[d]$ = '1',
$698 \le RS_d(X2) < 950$; $RS_d(X2)$ = '0'.

■ According to performance results of ASPM, the results of updating a quantitative attribute value of attribute {d} to a qualitative attribute value are given in Table 2.

B. Processing of null attribute values : EM

We adapt attribute values, substitutive for null attribute values(denoted as Λ) of weight attribute {d} in Table 1, to ARE of EM, calculate, and then decide the attribute value which has the minimum entropy.

$\{d_7\}$ = '0' ⇒ $E_{A0}(X2)$ = 0.244,
$\{d_7\}$ = '1' ⇒ $E_{A1}(X2)$ = 0.349,
$\{d_7\}$ = '2' ⇒ $E_{A2}(X2)$ = 0.401,
$\{d_{12}\}$ = '0' ⇒ $E_{A0}(X1)$ = 0.282,
$\{d_{12}\}$ = '1' ⇒ $E_{A1}(X1)$ = 0.2,
$\{d_{12}\}$ = '2' ⇒ $E_{A2}(X1)$ = 0.144.

■ As a result of calculating ARE, the attribute value of $\{b_7\}$ is '0' because of $E_{A0}(X2) < E_{A1}(X2) < E_{A2}(X2)$ and the attribute value of $\{b_{12}\}$ is '2' because of $E_{A2}(X1) < E_{A1}(X1) < E_{A0}(X1)$. According to performance results of EM, null attribute values of attribute {d_7, d_{12}} are updated to {'0', '2'}, respectively. The results are given in Table 2.

C. Processing of multiple attribute values : SAPM

Multiple attribute values of the engine attribute {c} are subdivided into the following single vector $Des(x_i, p_i^j) = c_i^j$.

$\{c_1^1\}$='1', $\{c_1^2\}$='6', $\{c_2^1\}$='1', $\{c_2^2\}$='6',
$\{c_3^1\}$='1', $\{c_4^1\}$='0', $\{c_5^1\}$='1', ···

Table 2. Complete information system for car efficiency comparison

U	a	b	c^1	c^2	d	e	f	g	m
x_1	0	0	1	6	2	2	1	0	1
x_2	0	0	1	6	2	2	1	1	1
x_3	0	0	1	×	2	1	1	1	1
x_4	0	1	0	×	0	0	2	1	2
x_5	0	1	0	×	2	0	1	1	1
x_6	0	1	0	6	2	1	1	0	1
x_7	1	0	0	×	0	0	2	1	2
x_8	1	0	1	×	0	1	1	1	2
x_9	0	0	1	×	2	1	1	1	1
x_{10}	1	1	0	×	1	0	1	1	2
x_{11}	0	0	1	t	1	2	1	1	1
x_{12}	0	0	1	t	2	2	1	0	1
x_{13}	1	0	0	×	1	1	2	1	2
x_{14}	1	0	0	×	1	1	2	1	2
x_{15}	0	0	0	×	1	1	2	1	2
x_{16}	0	0	0	6	1	1	2	0	1

Therefore, the attribute {c} is extended to the sub-attribute

$\{c^1, c^2\}$. A possible combination for multiple attribute values, and objects included in each sub-attribute are as follows:

$\{c^1\}=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_{10},x_{11},x_{12},x_{13},x_{14},x_{15},x_{16}\}$
$\{c^2\}=\{x_1,x_2,x_6,x_{11},x_{12}\}$.

In attribute values of the sub-attribute $\{c^1\}$, '×' is given in the case of c_i^j = null and the results are given in Table 2.

D. Processing of inconsistency attribute values or null attribute values of decision attributes: EM

$\{m_{13}\}$ = '2' ⇒ $E_{O2}(x_{13})$ = 3.007,
$\{m_{14}\}$ = '1' ⇒ $E_{O1}(x_{14})$ = 3.129,
$\{m_{16}\}$ = '1' ⇒ $E_{O1}(x_{16})$ = 2.989,
$\{m_{16}\}$ = '2' ⇒ $E_{O2}(x_{16})$ = 3.491.

■ As a result of calculating ORE, the attribute value of $\{m_{13}\}$ becomes '2' because of $E_{O2}(x_{13}) < E_{O1}(x_{14})$ and the attribute value of $\{m_{16}\}$ becomes '1' because of $E_{O1}(x_{16}) < E_{O2}(x_{16})$. As a result of performing EM module for removing incomplete information of decision attribute {m} like this, attribute values of attribute $\{m_{13}, m_{16}\}$ are updated to {'2', '1'} as in Table 2, respectively.

4.3 Derivation of inference rules

We derive inference rules using Rough Set from complete information system to evaluate processing results of INiPS. For this, we derive Table 3 from the reduction of unnecessary condition attributes and condition attribute values of Table 2 by the discernibility matrix.

We derive the following optimum inference rules by removing duplicated lines or unnecessary decision rules in Table 3.

Table 3. Reduction result of duplicated lines and unnecessary decision rules

U	a	c^2	d	c	g	m
x_1	×	6	×	×	×	
x_2,x_3,x_5,x_6,x_9	×	×	2	×	×	
x_{11}	×	×	1	2	×	1
x_{12}	×	t	×	×	×	
x_{16}	×	×	1	×	0	
x_4,x_7,x_8	×	×	0	×	×	
x_{10}	×	×	1	0	×	2
x_{13},x_{14},x_{15}	×	×	1	1	×	

× : don't care

if {(engine='6 cylinder')} or
{(1039 <weight≤1297)} or
{(950≤weight≤1039)} and (output='high')} or
{(engine='turbo')} or
{(950≤weight≤1039) and (gearbox='automatic')}
then (mileage efficiency = 'medium')
if {(698≤weight <950)} or
{(950≤weight≤1039)} and (output='low')} or
{(950≤weight≤1039)} and (output='medium')}
then {mileage = 'high'}.

5. Conclusion

The existing incomplete information process methods, presented for dealing with the only specified incomplete information occurring from condition attributes in information system, had their limits which couldn't manage various incomplete information, and incomplete information occurring from decision attributes as well. We needed, therefore, rather efficient methods for removing incomplete information in information system using Rough Set. For that reason, this paper summarizes types of incomplete information existing in information system and proposes ASPA, ORE, ARE, and SAPA that will remove different types of incomplete information in information system. Also, this paper proposes INiPS that integrates these methods to convert various incomplete information into complete information. INiPS integrates ASPA, ORE, ARE, and SAPA to unitary system in order to perform preprocessor function of an inference rule generator, and then converts incomplete system into complete information system efficiently.

Application cases show that the functions and characteristics of this INiPS are true, and establish justification and efficiency of INiPS by deriving optimum inference rules from complete information system managed in INiPS. But we have to experiment in the field of a real application, where various incomplete information exist, and need to generalize INiPS by supplementing functions for problems in the result from it so that INiPS may become rather incomplete information processing system. And studies for incorporating INiPS with the inference engine using Rough Set have to be continued to utilize INiPS in the real information processing system efficiently.

References

[1] Abe, S. and Lan, M. S., "A Method for Fuzzy Rules Extraction Directly from Numerical Data and Its Application to Pattern Classification," *IEEE Trans. on Fuzzy Systems*, vol. 3, No. 1, pp. 18-28, Feb. 1995.

[2] Beaubouef, T., Petry, F. E., and Arora, G., "Information-theoretic measures of uncertainty for Rough Sets and rough relational databases," *Information Science*, vol. 109, No. 1/4, pp. 185-195, 1998.

[3] Klir, G. J. and Folger, T. A., *Fuzzy Sets, Uncertainty, and Information*, Prentice Hall P T R, 1988.

[4] Kryszkiewicz, M., "Rules in incomplete information systems," *Information Science*, vol. 113, No. 3-4, pp. 271-292, 1999.

[5] Kryszkiewicz, M., "Rough Set approach to incom-

plete information systems," *Information Science*, vol. 112, No. 1/4, pp. 39-49, 1998.
[6] Pawlak, Z., "Rough Sets," *International Journal of Information and Computer Sciences*, vol. 11, No. 5, pp. 341-356, 1982.
[7] Pawlak, Z., *Rough Sets - Theoretical Aspects of Reasoning about Data*, Kluwer, 1991.
[8] Pawlak, Z., "Rough Set Theory and Its Applications to Data Analysis," Cybernetics and Systems : An International Journal, pp. 661-688, 1998.
[9] Skowron, A. and Rauszer, C., "The discernibility matrices and functions in information systems," in : Słowiński, R.(Ed.), *Intelligent Decision Support : Handbook of Applications and Advances of Rough Sets Theory*, Kluwer Academic Publisher, Dordrecht, pp. 331-362, 1992.
[10] Słowiński, R. and Stefanowski, J., "Handling Various Types of Uncertainty in the Rough Set Approach," in : W. Ziarko(ed.), *Rough Sets Fuzzy Sets and Knowledge Discovery(RSKD '93)*, Springer, Berlin, 1994.
[11] Słowiński, R. and Stefanowski, J., "Rough classification in incomplete information systems," *Mathematical and Compute. Modelling*, vol. 12, No. 10/11, pp. 1347- 1357, 1989.
[12] Hu, X., Cercone, N., and Han, J., "An Attribute-Oriented Rough Set Approach for Knowledge Discovery in Databases," in : W. Ziarko(ed.), *Rough Sets Fuzzy Sets and Knowledge Discovery(RSKD '93)*, Springer, Berlin, 1994..
[13] Zhang, K., "IRI : A Quantitative Approach to Inference Analysis in Relational Databases," *Proceedings of the IFIP WG 11.3 Eleventh Annual Working Conference on Database Security*, pp. 214-221, 1997.
[14] G. B. Jeong, D. Y. Kim and H. M. Chung, "A Study on the Processing of Imprecision Data by Rough Sets," *Proceedings of KFIS Spring Conference '98*, vol. 8, No. 1, pp. 11-15, 1998.
[15] G. B. Jeong and H. M. Chung, "A Study on the Processing of Incomplete Data in Information Systems Using Rough Set," *Journal of Korea Fuzzy Logic and Intelligent Systems Society*, vol. 9, No. 3, 1999.

저 자 소 개

정구범(Gu-Beom Jeong)
1999년 : 대구 가톨릭대학교 박사
1997년~현재 : 상주대학교 컴퓨터공학부 교수

관심분야 : 인공지능, 전자상거래

정환묵(Hwan-Mook Chung)
제 10권 제 5호 참조

김국보(Guk-Boh Kim)
제 7권 제 2호 참조
1997년 : 대구 가톨릭대학교 박사
1993년~현재 : 대진대학교 컴퓨터공학부 교수

관심분야 : 인공지능, 시스템 공학

박경옥(Kyung-Ok Park)
1998년 : 연세대학교 박사
1999년~현재 : 삼성전자(주) 경영혁신팀 기획그룹 과장

관심분야 : 통계학, 퍼지집합