# Applying Rough Sets to Information Tables Containing Probabilistic Values

Michinori Nakata[1] and Hiroshi Sakai[2]

[1] Faculty of Management and Information Science,
Josai International University
1 Gumyo, Togane, Chiba, 283-8555, Japan
nakatam@ieee.org
[2] Department of Mathematics and Computer Aided Sciences,
Faculty of Engineering, Kyushu Institute of Technology,
Tobata, Kitakyushu, 804-8550, Japan
sakai@mns.kyutech.ac.jp

**Abstract.** Rough sets are applied to information tables containing imprecise values that are expressed in a probability distribution. A family of weighted equivalence classes is obtained where each equivalence class is accompanied by the probability to which it is an actual one. By using the family of weighted equivalence classes, we derive lower and upper approximations. The lower and upper approximations coincide with ones obtained from methods of possible worlds. Therefore, the method of weighted equivalence classes is justified. In addition, this method is applied to missing values interpreted probabilistically. Using weighted equivalence classes correctly derives a lower approximation, even in the case where the method of Kryszkiewicz does not derive any lower approximation.

**Keywords:** Rough sets, Imprecise information, Probabilistic value, Weighted equivalence class, Lower and upper approximations.

## 1 Introduction

Rough sets play a significant role in the field of knowledge discovery and data mining since the first paper published by Pawlak [19]. The framework of rough sets is constructed under the premise that information tables consisting of precise information are obtained. However, there ubiquitously exists imprecise information in the real world [18]. Thus, it has been investigated to apply rough sets to information tables containing imprecise information represented by a missing value, an or-set, a possibility distribution, etc [2,4,5,10,11,13,14,15,20,21,22,23,25]. The methods are broadly separated into three ways.

The first method is one based on possible worlds [17,20,21,22]. In the method, possible tables, which consist of precise values, are obtained from an information table. Each possible table is dealt with by the traditional methods of applying rough sets to information tables containing precise information, and then the results from

V. Torra, Y. Narukawa, and Y. Yoshida (Eds.): MDAI 2007, LNAI 4617, pp. 282–294, 2007.

the possible tables are aggregated. In other words, the methods that are already established are applied to each possible table. Therefore, there is no doubt for correctness of the treatment. However, the method has difficulties for knowledge discovery at the level of a set of possible values, although it is suitable for finding knowledge at the level of possible values. This is because the number of possible tables exponentially increases as the number of imprecise attribute values increases.

The second method is to use assumptions on indiscernibility of missing values [2,5,8,10,11,25]. Under the assumptions, we can obtain a binary relation for indiscernibility between objects. To the binary relation, rough sets are applied by using a class of objects; for example, an indiscernible class. In the method, it is not clarified why the assumptions are valid to real data sets.

The third method directly deals with imprecise values, without using any assumptions for indiscernibility, under extending the traditional method of rough sets [13,14,15,16,25]. In the method, imprecise values are dealt with probabilistically or possibilistically and the traditional methods are probabilistically or possibilistically extended.[1] A binary relation for indiscernibility is constructed by calculating a degree for indiscernibility between objects. Indiscernible classes for each object are obtained from the binary relation for indiscernibility. The correctness criterion is that any extended method has to give the same results as the method of possible worlds [13]. This criterion is commonly used in the field of databases handling imprecise information [1,7,28].

Stefanowski and Tsoukiàs used implication operators to calculate an inclusion degree between indiscernible classes [25]. Nakata and Sakai have shown that the results in terms of implication operators do not satisfy the correctness criterion and has proposed the method that satisfies the correctness criterion [13,14,15]. However, the proposed method has difficulties for definability, because rough approximations are defined by constructing sets from singletons. Therefore, we propose a method using equivalence classes, called a method of weighted equivalence classes. In this paper, we show how weighted equivalence classes are applied to information tables containing imprecise values expressed in a probability distribution, called probabilistic values.[2]

In Section 2, we briefly address the traditional methods of applying rough sets to information tables containing precise information. In Section 3, methods of possible worlds are mentioned. In the methods, the extended set of possible tables is obtained from an information table containing imprecise values. The traditional methods of applying rough sets to precise information deal with each possible table, and then the results from possible tables are aggregated. In Section 4, methods of applying rough sets to information tables containing probabilistic values are described in terms of weighted equivalence classes. In Section 5, the method of weighted equivalence classes is applied to information tables containing missing values under probabilistic interpretation. Section 6 presents conclusions.

[1] Ziarko proposes methods of rough sets applying data tables where each data is accompanied by a probability [26,27].

[2] See the reference [16] for information tables containing possibilistic values expressed in a possibility distribution.

## 2 Rough Sets to Precise Information

A data set is represented as a table, called an information table, where each row represents an object and each column represents an attribute. The information table is pair $\mathcal{A} = (U, AT)$. $U$ is a non-empty finite set of objects called the universe. Concretely speaking, $U$ is the set of objects that comprise the information table. $AT$ is a non-empty finite set of attributes such that $\forall a \in AT \ : \ U \rightarrow V_a$. Set $V_a$ is called the domain of attribute $a$. In information table $T$ whose framework is set $AT$ of attributes, binary relation $IND(\Psi_A)$ for indiscernibility of objects in subset $\Psi \subseteq U$ on subset $A \subseteq AT$ of attributes is,

$$IND(\Psi_A) = \{(o, o') \in \Psi \times \Psi \mid \forall a \in A \ \ a(o) = a(o')\}, \tag{1}$$

where $a(o)$ and $a(o')$ denote values of attribute $a$ for objects $o$ and $o'$, respectively. This relation is called an indiscernibility relation. Obviously, $IND(\Psi_A)$ is an equivalence relation. From the indiscernibility relation, equivalence class $E(\Psi_A)_o (= \{o' \mid (o, o') \in IND(\Psi_A)\})$ containing object $o$ is obtained. This is also the set of objects that is indiscernible with object $o$, called the indiscernible class for object $o$. Finally, family $U/IND(\Psi_A)$ $(= \{E(\Psi_A)_o \mid o \in \Psi\})$ of equivalence classes is derived from the indiscernibility relation. All equivalence classes obtained from the indiscernibility relation do not intersect with each other. This means that the objects are uniquely partitioned.

Using equivalence classes, lower approximation $\underline{Apr}(\Phi_B, \Psi_A)$ and upper approximation $\overline{Apr}(\Phi_B, \Psi_A)$ of $\Phi/IND(\Phi_B)$ by $\Psi/IND(\Psi_A)$ are,

$$\underline{Apr}(\Phi_B, \Psi_A) = \{E(\Psi_A) \mid \exists E(\Phi_B) \ E(\Psi_A) \subseteq E(\Phi_B)\}, \tag{2}$$

$$\overline{Apr}(\Phi_B, \Psi_A) = \{E(\Psi_A) \mid \exists E(\Phi_B) \ E(\Psi_A) \cap E(\Phi_B) \neq \emptyset\}. \tag{3}$$

where $E(\Psi_A) \in \Psi/IND(\Psi_A)$ and $E(\Phi_B) \in \Phi/IND(\Phi_B)$ are equivalence classes for sets $\Psi$ and $\Phi$ of objects on sets A and B of attributes, respectively. These formulas are expressed in terms of a family of equivalence classes. When we express the approximations in terms of a set of objects, the following expressions are used:

$$\underline{apr}(\Phi_B, \Psi_A) = \{o \mid o \in E(\Psi_A) \wedge \exists E(\Phi_B) \ E(\Psi_A) \subseteq E(\Phi_B)\}, \tag{4}$$

$$\overline{apr}(\Phi_B, \Psi_A) = \{o \mid o \in E(\Psi_A) \wedge \exists E(\Phi_B) \ E(\Psi_A) \cap E(\Phi_B) \neq \emptyset\}. \tag{5}$$

## 3 Methods of Possible Worlds

In methods of possible worlds, the traditional ways addressed in the previous section are applied to each possible table, and then the results from possible tables are aggregated.

When probabilistic values expressed in a probability distribution is contained in information table $T$, we obtain extended set $rep(T)$ of possible tables,

$$rep(T) = \{(pt_1, \mu(pt_1)), \ldots, (pt_n, \mu(pt_n))\}, \tag{6}$$

where $\mu(pt_i)$ denotes the probability to which possible table $pt_i$ is the actual one and $n$ is equal to $\Pi_{i=1,m} l_i$, where the number of probabilistic values is $m$ and each of them is expressed in a probability distribution having $l_i (i = 1, m))$ elements. When possible table $pt_i$ is the table where probabilistic values in information table $T$ are replaced by $v_{i1}$, $v_{i2}$, ..., $v_{im}$,

$$\mu(pt_i) = \prod_{k=1,m} \pi(v_{ik}), \tag{7}$$

where $\prod$ denotes product and probability $\pi(v_{ik})$ of element $v_{ik}$ comes from probability distribution $\pi$ expressing the probabilistic value to which the element belongs.

Each possible table consists of precise values. Family $U/IND(\Psi_A)_{pt_i}$ of equivalence classes on set $A$ of attributes is obtained from each possible table $pt_i$. Possible table $pt_i$ is accompanied by probability $\mu(pt_i)$ to which it is the actual information table. Thus, the family of possible equivalence classes accompanied by a probability is obtained for each possible table, which is expressed by $(U/IND(\Psi_A)_{pt_i}, \mu(pt_i))$. When we express $(U/IND(\Psi_A)_{pt_i}, \mu(pt_i))$ in terms of equivalence classes,

$$(U/IND(\Psi_A)_{pt_i}, \mu(pt_i)) = \{(E(\Psi_A), \mu(pt_i)) \mid E(\Psi_A) \in U/IND(\Psi_A)_{pt_i}\}, \tag{8}$$

where equivalence class $E(\Psi_A)$ is a possible equivalence class on set $A$ of attributes and has probability $\mu(pt_i)$ to which it is one of actual equivalence classes. $U/IND(\Psi_A)$ is the union of $(U/IND(\Psi_A)_{pt_i}, \mu(pt_i))$,

$$U/IND(\Psi_A) = \cup_i (U/IND(\Psi_A)_{pt_i}, \mu(pt_i)). \tag{9}$$

Note that the summation of probabilities is taken in the union if there are the same elements accompanied by a probability. When we express family $U/IND(\Psi_A)$ in terms of equivalence classes,

$$U/IND(\Psi_A) = \{(E(\Psi_A), \kappa(E(\Psi_A) \in U/IND(\Psi_A))) \mid \kappa(E(\Psi_A) \in U/IND(\Psi_A)) > 0\}, \tag{10}$$

where probability $\kappa(E(\Psi_A) \in U/IND(\Psi_A))$ to which equivalence class $E(\Psi_A)$ is contained in $U/IND(\Psi_A)$ is,

$$\kappa(E(\Psi_A) \in U/IND(\Psi_A)) = \sum_{E(\Psi_A) \in U/IND(\Psi_A)_{pt_i}} \mu(pt_i). \tag{11}$$

To obtain lower and upper approximations, the traditional methods addressed in the previous section are applied to possible tables. Let $\underline{Apr}(\Phi_B, \Psi_A)_{pt_i}$ and $\overline{Apr}(\Phi_B, \Psi_A)_{pt_i}$ denote the lower and upper approximations of $U/IND(\Phi_B)_{pt_i}$ by $U/IND(\Psi_A)_{pt_i}$ in possible table $pt_i$ having probability $\mu(pt_i)$. $\underline{Apr}(\Phi_B, \Psi_A)_{pt_i}$ and $\overline{Apr}(\Phi_B, \Psi_A)_{pt_i}$ are accompanied by probability $\mu(pt_i)$, which is expressed by $(\underline{Apr}(\Phi_B, \Psi_A)_{pt_i}, \mu(pt_i))$ and $(\overline{Apr}(\Phi_B, \Psi_A)_{pt_i}, \mu(pt_i))$. $\underline{Apr}(\Phi_B, \Psi_A)$ and $\overline{Apr}$

$(\Phi_B, \Psi_A)$ are the union of $(\underline{Apr}(\Phi_B, \Psi_A)_{pt_i}, \mu(pt_i))$ and $(\overline{Apr}(\Phi_B, \Psi_A)_{pt_i}, \mu(pt_i))$, respectively.

$$\underline{Apr}(\Phi_B, \Psi_A) = \cup_i(\underline{Apr}(\Phi_B, \Psi_A)_{pt_i}, \mu(pt_i)), \tag{12}$$

$$\overline{Apr}(\Phi_B, \Psi_A) = \cup_i(\overline{Apr}(\Phi_B, \Psi_A)_{pt_i}, \mu(pt_i)). \tag{13}$$

When we express approximations in terms of equivalence classes,

$$\underline{Apr}(\Phi_B, \Psi_A) = \{(E(\Psi_A), \kappa(E(\Psi_A) \in \underline{Apr}(\Phi_B, \Psi_A))) \mid \kappa(E(\Psi_A) \in \underline{Apr}(\Phi_B, \Psi_A)) > 0\}, \tag{14}$$

$$\overline{Apr}(\Phi_B, \Psi_A) = \{(E(\Psi_A), \kappa(E(\Psi_A) \in \overline{Apr}(\Phi_B, \Psi_A))) \mid \kappa(E(\Psi_A) \in \overline{Apr}(\Phi_B, \Psi_A)) > 0\}, \tag{15}$$

where probabilities $\kappa(E(\Psi_A) \in \underline{Apr}(\Phi_B, \Psi_A))$ and $\kappa(E(\Psi_A) \in \overline{Apr}(\Phi_B, \Psi_A))$ to which equivalence class $E(\Psi_A)$ is contained in $\underline{Apr}(\Phi_B, \Psi_A)$ and $\overline{Apr}(\Phi_B, \Psi_A)$ are,

$$\kappa(E(\Psi_A) \in \underline{Apr}(\Phi_B, \Psi_A)) = \sum_{E(\Psi_A) \in \underline{Apr}(\Phi_B, \Psi_A))_{pt_i}} \mu(pt_i), \tag{16}$$

$$\kappa(E(\Psi_A) \in \overline{Apr}(\Phi_B, \Psi_A)) = \sum_{E(\Psi_A) \in \overline{Apr}(\Phi_B, \Psi_A))_{pt_i}} \mu(pt_i). \tag{17}$$

These formulas show that the summation of the probabilities of possible tables where equivalence class $E(\Psi_A)$ is contained in rough approximations is equal to the probability for equivalence class $E(\Psi_A)$.

**Proposition 1**
When $(E(\Psi_A), \kappa(E(\Psi_A) \in \underline{Apr}(\Phi_B, \Psi_A)))$ is an element of $\underline{Apr}(\Phi_B, \Psi_A)$ in information table $T$, there exists set $\underline{PT}$ of possible tables where for all $pt \in \underline{PT}$ $\underline{Apr}(\Phi_B, \Psi_A)_{pt}$ contains $E(\Psi_A)$ and $\sum_{pt \in PT} \mu(pt)$ is equal to $\kappa(E(\Psi_A) \in \underline{Apr}(\Phi_B, \Psi_A))$.

**Proposition 2**
When $(E(\Psi_A), \kappa(E(\Psi_A) \in \overline{Apr}(\Phi_B, \Psi_A)))$ is an element of $\overline{Apr}(\Phi_B, \Psi_A)$ in information table $T$, there exists set $\overline{PT}$ of possible tables where for all $pt \in \overline{PT}$ $\overline{Apr}(\Phi_B, \Psi_A)_{pt}$ contains $E(\Psi_A)$ and $\sum_{pt \in PT} \mu(pt)$ is equal to $\kappa(E(\Psi_A) \in \overline{Apr}(\Phi_B, \Psi_A))$.

When the lower and upper approximations are expressed in terms of a set of objects,

$$\underline{apr}(\Phi_B, \Psi_A) = \{(o, \kappa(o \in \underline{apr}(\Phi_B, \Psi_A))) \mid \kappa(o \in \underline{apr}(\Phi_B, \Psi_A)) > 0\}, \tag{18}$$

$$\overline{apr}(\Phi_B, \Psi_A) = \{(o, \kappa(o \in \overline{apr}(\Phi_B, \Psi_A))) \mid \kappa(o \in \overline{apr}(\Phi_B, \Psi_A)) > 0\}, \tag{19}$$

and

$$\kappa(o \in \underline{apr}(\Phi_B, \Psi_A)) = \sum_{E(\Psi_A) \ni o} \kappa(E(\Psi_A) \in \underline{Apr}(\Phi_B, \Psi_A)), \tag{20}$$

$$\kappa(o \in \overline{apr}(\Phi_B, \Psi_A)) = \sum_{E(\Psi_A) \ni o} \kappa(E(\Psi_A) \in \overline{Apr}(\Phi_B, \Psi_A). \tag{21}$$

We adopt results from methods of possible worlds as a correctness criterion of extended methods of applying rough sets to imprecise information. This is commonly used in the field of databases handling imprecise information [1,7,28].

**Correctness criterion**
*Results obtained from applying an extended method to an information table containing imprecise values are the same as ones obtained from applying the corresponding traditional method to every possible table derived from that information table and aggregating the results created in the possible tables.*

## 4 Rough Sets to Information Tables Containing Probabilistic Values

When object $o$ takes imprecise values for attributes, we calculate the degree to which the attribute values are the same as another object $o'$. The degree is the indiscernibility degree of object $o$ and $o'$ on the attributes. In this case, a binary relation for indiscernibility on set $A$ of attributes is,

$$IND(\Psi_A) = \{((o, o'), \kappa(A(o) = A(o'))) \mid (\kappa(A(o) = A(o')) \neq 0) \wedge (o \neq o')\} \cup \{((o, o), 1)\}, \tag{22}$$

where $\kappa(A(o) = A(o'))$ denotes the indiscernibility degree of objects $o$ and $o'$ on set $A$ of attributes and is equal to $\kappa((o, o') \in IND(\Psi_A))$,

$$\kappa(A(o) = A(o')) = \bigotimes_{a \in A} \kappa(a(o) = a(o')), \tag{23}$$

where operator $\bigotimes$ depends on properties of imprecise attribute values. When the imprecise attribute values are expressed in a probability distribution, the operator is product denoted by $\prod$.

From binary relation $IND(\Psi_A)$ for indiscernibility, family $U/IND(\Psi_A)$ of weighted equivalence classes is obtained via indiscernible sets. Among the elements of $IND(\Psi_A)$, set $S_A(o)$ of objects that are paired with object $o$, called the indiscernible set on set $A$ of attributes for object $o$, is,

$$S_A(o) = \{o' \mid \kappa((o, o') \in IND(\Psi_A)) > 0\}. \tag{24}$$

$S_A(o)$ is the greatest possible equivalence class among possible equivalence classes containing objects $o$, when objects $o$ has a precise value on all attributes in set $A$

of attributes. Let $PS_A(o)$ denote the power set of $S_A(o)$. From $PS_A(o)$, family $Can(U/IND(\Psi_A)_o)$ of candidates for possible equivalence classes containing object $o$ is obtained,

$$Can(U/IND(\Psi_A)_o) = \{E(\Psi_A) \mid E(\Psi_A) \in PS_A(o) \wedge o \in E(\Psi_A)\}. \quad (25)$$

Whole family $Can(U/IND(\Psi_A))$ of candidates for possible equivalence classes is,

$$Can(U/IND(\Psi_A)) = \cup_o Can(U/IND(\Psi_A)_o). \quad (26)$$

Probability $\kappa(E(\Psi_A) \in U/IND(\Psi_A))$ to which candidate $E(\Psi_A) \in Can(U/IND(\Psi_A))$ is an actual one is,

$$\kappa(E(\Psi_A) \in U/IND(\Psi_A)) = \kappa(\wedge_{o \in E(\Psi_A) \ and \ o' \in E(\Psi_A)}(A(o) = A(o')) \wedge_{o \in E(\Psi_A) \ and \ o' \notin E(\Psi_A)}(A(o) \neq A(o'))), \quad (27)$$

where $o \neq o'$, $\kappa(f)$ is the probability to which formula $f$ is satisfied, and $\kappa(f) = 1$ when there exists no $f$. When set $\Psi$ of objects contains $k$ objects and equivalence class $E(\Psi_A)$ consists of $l$ objects,

$$\kappa(E(\Psi_A) \in U/IND(\Psi_A)) = \sum_{(u, v_1, \cdots, v_{k-l})} \Big( \prod_{o \in E(\Psi_A)} \pi_{A(o)}(u) \times \prod_{o_i \notin E(\Psi_A)} (\pi_{A(o_1)}(v_1), \pi_{A(o_2)}(v_2), \ldots, \pi_{A(o_{k-l})}(v_{k-l}))\Big), \quad (28)$$

where

$$\pi_{A(o)}(u) = \prod_{j=1,m} \pi_{a_j(o)}(u_j), \quad (29)$$

$$\pi_{A(o_i)}(v_i) = \prod_{j=1,m} \pi_{a_j(o_i)}(v_{ij}), \quad (30)$$

where two values $u$ and $v_i$ are different and are expressed in $(u_1, \cdots, u_m)$ and $(v_{i1}, \cdots, v_{im})$ on set $A(= \{a_1, a_2, \ldots, a_m\})$ of attributes, respectively. Finally, family $U/IND(\Psi_A)$ of weighted equivalence classes is,

$$U/IND(\Psi_A) = \{(E(\Psi_A), \kappa(E(\Psi_A) \in U/IND(\Psi_A))) \mid \kappa(E(\Psi_A) \in U/IND(\Psi_A)) > 0\}. \quad (31)$$

**Proposition 3**
When $(E(\Psi_A), \kappa(E(\Psi_A) \in U/IND(\Psi_A)))$ is an element of $U/IND(\Psi_A)$ in information table $T$, there exists set $PT$ of possible tables where for all $pt \in PT$ $U/IND(\Psi_A)_{pt}$ contains $E(\Psi_A)$ and $\sum_{pt \in PT} \mu(pt)$ is equal to $\kappa(E(\Psi_A) \in U/IND(\Psi_A))$.

**Proposition 4**
$U/IND(\Psi_A)$ in an information table is equal to the union of the families of possible equivalence classes accompanied by a probability, where each family of possible equivalence classes is obtained from a possible table created from the information table.

**Proposition 5**
For any object $o$,

$$\sum_{E(\Psi_A)\ni o} \kappa(E(\Psi_A) \in U/IND(\Psi_A)) = 1. \tag{32}$$

Using families of weighted equivalence classes, we can obtain lower approximation $\underline{Apr}(\Phi_B, \Psi_A)$ and upper approximation $\overline{Apr}(\Phi_B, \Psi_A)$ of $U/IND(\Phi_B)$ by $U/IND(\Psi_A)$. For the lower approximation,

$$\underline{Apr}(\Phi_B, \Psi_A) = \{(E(\Psi_A), \kappa(E(\Psi_A) \in \underline{Apr}(\Phi_B, \Psi_A))) \mid \kappa(E(\Psi_A) \in \underline{Apr}(\Phi_B, \Psi_A)) > 0\}, \tag{33}$$

$$\kappa(E(\Psi_A) \in \underline{Apr}(\Phi_B, \Psi_A)) = \sum_{E(\Phi_B)} (\kappa(E(\Psi_A) \subseteq E(\Phi_B)) \times \kappa(E(\Psi_A) \in U/IND(\Psi_A)) \times \kappa(E(\Phi_B) \in U/IND(\Phi_B))), \tag{34}$$

where

$$\kappa(E(\Psi_A) \subseteq E(\Phi_B)) = \begin{cases} 1 \text{ if } E(\Psi_A) \subseteq E(\Phi_B), \\ 0 \text{ otherwise.} \end{cases} \tag{35}$$

**Proposition 6**
If $(E(\Psi_A), \kappa(E(\Psi_A) \in \underline{Apr}(\Phi_B, \Psi_A)))$ in information table $T$ is an element of $\underline{Apr}(\Phi_B, \Psi_A)$, there exists set $\underline{PT}$ of possible tables where for all $pt \in \underline{PT}$ $\overline{Apr}(\Phi_B, \Psi_A)_{pt}$ contains $E(\Psi_A)$ and $\sum_{pt\in PT} \mu(pt)$ is equal to $\kappa(E(\Psi_A) \in \underline{Apr}(\Phi_B, \Psi_A))$.

For the upper approximation,

$$\overline{Apr}(\Phi_B, \Psi_A) = \{(E(\Psi_A), \kappa(o \in \overline{Apr}(\Phi_B, \Psi_A))) \mid \kappa(E(\Psi_A) \in \overline{Apr}(\Phi_B, \Psi_A)) > 0\}, \tag{36}$$

$$\kappa(E(\Psi_A) \in \overline{Apr}(\Phi_B, \Psi_A)) = \kappa(E(\Psi_A) \cap \Phi_B \neq \emptyset) \times \kappa(E(\Psi_A) \in U/IND(\Psi_A)), \tag{37}$$

where

$$\kappa(E(\Psi_A) \cap \Phi_B \neq \emptyset) = \begin{cases} 1 \text{ if } E(\Psi_A) \cap \Phi_B \neq \emptyset, \\ 0 \text{ otherwise.} \end{cases} \tag{38}$$

From this formula, the upper approximation is trivial when $\Phi_B = U_B$; namely, $\overline{Apr}(U_B, \Psi_A) = U/IND(\Psi_A)$.

**Proposition 7**
If $(E(\Psi_A), \kappa(E(\Psi_A) \in \overline{Apr}(\Phi_B, \Psi_A)))$ in information table $T$ is an element of $\overline{Apr}(\Phi_B, \Psi_A)$, there exists set $\overline{PT}$ of possible tables where for all $pt \in \overline{PT}$ $\overline{Apr}(\Phi_B, \Psi_A)_{pt}$ contains $E(\Psi_A)$ and $\sum_{pt \in PT} \mu(pt)$ is equal to $\kappa(E(\Psi_A) \in \overline{Apr}(\Phi_B, \Psi_A))$.

For expressions in terms of a set of objects, the same expressions as in Section 3 are used.

**Proposition 8**
The lower and upper approximations that are obtained by the method of weighted equivalence classes coincide with ones obtained by the method of possible worlds.

## 5 Information Tables Containing Missing Values

We apply the method of weighted equivalence classes to information tables containing missing values. We briefly compare the method where Kryszkiewicz uses indiscernible classes with the method of weighted equivalence classes.

When missing values are contained in information table $T$, Kryszkiewicz defines binary relation $IND(U_A)$ for indiscernibility between objects on set $A$ of attributes as follows [8,10]:

$$IND(U_A) = \{(o, o') \in U \times U \mid \forall a \in A, a(o) = a(o') \vee a(o) = * \vee a(o') = *\}, \quad (39)$$

where $*$ denotes a missing value and $U$ is used in place of $\Psi$ when $\Psi$ is equal to universe $U$. From this definition, an object having missing values for all attributes on set $A$ of attributes is indiscernible with any object. This corresponds to "do not care" semantics of missing values addressed by Grzymala-Busse [4,5]. By using indiscernible classes obtained from $IND(U_A)$, Kryszkiewicz expresses lower and upper approximations of set $\Phi \subseteq U$ of objects:

$$\underline{apr}(\Phi, U_A) = \{o \in U \mid S_A(o) \subseteq \Phi\}, \quad (40)$$
$$\overline{apr}(\Phi, U_A) = \{o \in U \mid S_A(o) \cap \Phi \neq \emptyset\}, \quad (41)$$

where $S_A(o) (= \{o' \mid (o, o') \in IND(U_A)\})$ denotes the indiscernible class for object $o$.

When we use the method of weighted equivalence classes, a missing value in an attribute is probabilistically interpreted. In the missing value, every element in the domain of the attribute has the same probability to which the element is the actual value. In other words, a missing value in attribute $a$ is equal to the probabilistic value expressed in the uniform probability distribution where every element over the domain has the same probability $1/|V_a|$. When attribute value $a(o)$ of object $o$ is a missing value,

$$\kappa(a(o) = a(o')) = \sum_{u,v \in V_a} (\mu_=(u, v) \times \pi_{a(o)}(u) \times \pi_{a(o')}(v)) = 1/|V_a|,$$

where $\pi_{a(o)}(u)$ and $\pi_{a(o')}(u)$ denote probability distributions expressing attribute values $a(o)$ and $a(o')$,[3] respectively, and,

$$\mu_{=}(u,v) = \begin{cases} 1 \text{ if } u = v, \\ 0 \text{ otherwise.} \end{cases}$$

This shows that the indiscernibility degree of an object taking a missing value on attribute $a$ with the other objects is equal to $1/|V_a|$; namely, the object is indiscernible with all objects with probability $1/|V_a|$. We express lower and upper approximations in terms of weighted equivalence classes, as is shown in the previous section. Differences between the method of Kryszkiewicz and the method of weighted equivalence classes are clarified in the following simple example:

**Example**
We suppose that information table $T$ is obtained:

$T$

| $O$ | $a_1$ | $a_2$ | $a_3$ |
|---|---|---|---|
| 1 | $x$ | 1 | $a$ |
| 2 | $x$ | 1 | $a$ |
| 3 | $x$ | 1 | $a$ |
| 4 | $x$ | 1 | $a$ |
| 5 | $*$ | 2 | $b$ |

The mark $O$ denotes the object identity. Domains $V_{a_1}$, $V_{a_2}$, and $V_{a_3}$ of attributes $a_1$, $a_2$, and $a_3$ are $\{x, y\}$, $\{1, 2\}$, and $\{a, b\}$, respectively.

First, we apply the method of Kryszkiewicz to information table $T$. For indiscernible classes of each object on attribute $a_1$,

$$S_{a_1}(o_1) = S_{a_1}(o_2) = S_{a_1}(o_3) = S_{a_1}(o_4) = S_{a_1}(o_5) = \{o_1, o_2, o_3, o_4, o_5\}.$$

We suppose that $\Phi = \{o_1, o_2, o_3, o_4\}$ for simplicity. For the lower approximation, using formula (40), because of $\{o_1, o_2, o_3, o_4, o_5\} \not\subseteq \{o_1, o_2, o_3, o_4\}$,

$$\underline{apr}(\Phi, U_{a_1}) = \emptyset$$

This shows that we do not obtain any information for the lower approximation.[4] This is true for different expressions [4,6,12] proposed by several authors. For the upper approximation, using formula (41), because of $\{o_1, o_2, o_3, o_4, o_5\} \cap \{o_1, o_2, o_3, o_4\} \neq \emptyset$,

$$\overline{apr}(\Phi, U_{a_1}) = \{o_1, o_2, o_3, o_4, o_5\}.$$

Second, we use the method of weighted equivalence classes. Missing value $*$ in information table $T$ is expressed in probability distribution $\{(x, 1/2), (y, 1/2)\}_p$. Using formulas (24) – (31),

[3] When $a(o')$ is a precise value; for example, $a(o') = x$, probability distribution $\pi_{a(o')}$ is expressed in $\{(x, 1)\}_p$, where subscript $p$ denotes a probability distribution.

[4] Stefanowski and Tsoukiàs points out that the method of Kryszkiewicz using "do not care" semantics creates quite poor results [24]. To handle the problem, other assumptions for indiscernibility of missing values are proposed [2,24].

$$U/IND(U_{a_1}) = \{(\{o_1, o_2, o_3, o_4\}, 1/2), (\{o_1, o_2, o_3, o_4, o_5\}, 1/2)\},$$

Applying formulas (33) – (38),

$$\underline{\underline{Apr}}(\Phi, U_{a_1}) = \{(\{o_1, o_2, o_3, o_4\}, 1/2)\},$$
$$\overline{\overline{Apr}}(\Phi, U_{a_1}) = \{(\{o_1, o_2, o_3, o_4\}, 1/2), \{(\{o_1, o_2, o_3, o_4, o_5\}, 1/2)\}.$$

Using formulas (18) – (21),

$$\underline{apr}(\Phi, U_{a_1}) = \{(o_1, 1/2), (o_2, 1/2), (o_3, 1/2), (o_4, 1/2)\},$$
$$\overline{apr}(\Phi, U_{a_1}) = \{(o_1, 1), (o_2, 1), (o_3, 1), (o_4, 1), (o_5, 1/2)\}.$$

Last, we show results by the method of possible worlds. Extended set $rep(T)$ of possible tables is,

$$rep(T) = \{(pt_1, 1/2), (pt_2, 1/2)\}_p.$$

$pt_1$

| $O$ | $a_1$ | $a_2$ | $a_3$ |
|---|---|---|---|
| 1 | $x$ | 1 | $a$ |
| 2 | $x$ | 1 | $a$ |
| 3 | $x$ | 1 | $a$ |
| 4 | $x$ | 1 | $a$ |
| 5 | $x$ | 2 | $b$ |

$pt_2$

| $O$ | $a_1$ | $a_2$ | $a_3$ |
|---|---|---|---|
| 1 | $x$ | 1 | $a$ |
| 2 | $x$ | 1 | $a$ |
| 3 | $x$ | 1 | $a$ |
| 4 | $x$ | 1 | $a$ |
| 5 | $y$ | 2 | $b$ |

For families of equivalence classes of possible tables,

$$(U/IND(U_{a_1}), 1/2)_{pt_1} = \{(\{o_1, o_2, o_3, o_4, o_5\}, 1/2)\},$$
$$(U/IND(U_{a_1}), 1/2)_{pt_2} = \{(\{o_1, o_2, o_3, o_4\}, 1/2), (\{o_5\}, 1/2)\},$$

For lower and upper approximations of each possible table,

$$\underline{\underline{Apr}}(\Phi, U_{a_1})_{pt_1} = \emptyset,$$
$$\overline{\overline{Apr}}(\Phi, U_{a_1})_{pt_1} = \{(\{o_1, o_2, o_3, o_4, o_5\}, 1/2)\},$$
$$\underline{\underline{Apr}}(\Phi, U_{a_1})_{pt_2} = \{(\{o_1, o_2, o_3, o_4\}, 1/2)\}.$$
$$\overline{\overline{Apr}}(\Phi, U_{a_1})_{pt_2} = \{(\{o_1, o_2, o_3, o_4\}, 1/2)\}.$$

Finally, using formulas (12) – (17) and (18) – (21),

$$\underline{\underline{Apr}}(\Phi, U_{a_1}) = \{(\{o_1, o_2, o_3, o_4\}, 1/2)\},$$
$$\overline{\overline{Apr}}(\Phi, U_{a_1}) = \{(\{o_1, o_2, o_3, o_4\}, 1/2), (\{o_1, o_2, o_3, o_4, o_5\}, 1/2)\},$$
$$\underline{\underline{apr}}(\Phi, U_{a_1}) = \{(o_1, 1/2), (o_2, 1/2), (o_3, 1/2), (o_4, 1/2)\},$$
$$\overline{\overline{apr}}(\Phi, U_{a_1}) = \{(o_1, 1), (o_2, 1), (o_3, 1), (o_4, 1), (o_5, 1/2)\}.$$

Indeed, the results obtained from the method of weighted equivalence classes coincide with ones from the method of possible worlds.

This simple example shows that we obtain correct results for the lower approximation when weighted equivalence classes are used. On the other hand, we cannot obtain any information for the lower approximation by the existence of only the missing value in the method of Kryszkiewicz where indiscernible classes are used.

## 6 Conclusions

We have proposed a method, where weighted equivalence classes are used, to deal with imprecise information expressed in a probability distribution. The lower and upper approximations by the method of weighted equivalence classes coincide with ones by the method of possible worlds. In other words, this method satisfies the correctness criterion that is used in the field of incomplete databases. This is justification of the method of weighted equivalence classes.

We have applied the method of weighted equivalence classes to information tables containing missing values under probabilistic interpretation. We obtain correct results for rough approximations when weighted equivalence classes are used, even if we do not obtain any results for the lower approximation when the method of Kryszkiewicz is used.

**Acknowledgment.** This research has been partially supported by the Grant-in-Aid for Scientific Research (C), Japan Society for the Promotion of Science, No. 18500214.

## References

1. Abiteboul, S., Hull, R., Vianu, V.: Foundations of Databases. Addison-Wesley, London, UK (1995)
2. Greco, S., Matarazzo, B., Slowinski, R.: Handling Missing Values in Rough Set Analysis of Multi-attribute and Multi-criteria Decision Problem. In: Zhong, N., Skowron, A., Ohsuga, S. (eds.) New Directions in Rough Sets, Data Mining, and Granular-Soft Computing. LNCS (LNAI), vol. 1711, pp. 146–157. Springer, Heidelberg (1999)
3. Grzymala-Busse, J.W.: On the Unknown Attribute Values in Learning from Examples. In: Raś, Z.W., Zemankova, M. (eds.) ISMIS 1991. LNCS(LNAI), vol. 542, pp. 368–377. Springer, Heidelberg (1991)
4. Grzymala-Busse, J.W: Data with Missing Attribute Values: Generalization of Indiscernibility Relation and Rule Induction, Transactions on Rough Sets I, 78–95 (2004)
5. Grzymala-Busse, J.W.: Characteristic Relations for Incomplete Data: A Generalization of the Indiscernibility Relation, Transactions on Rough Sets IV, 58–68 (2005)
6. Guan, Y.-Y., Wang, H.-K.: Set-valued Information Systems. Information Sciences 176, 2507–2525 (2006)
7. Imielinski, T., Lipski, W.: Incomplete Information in Relational Databases. Journal of the ACM 31(4), 761–791 (1984)
8. Kryszkiewicz, M.: Rough Set Approach to Incomplete Information Systems. Information Sciences 112, 39–49 (1998)
9. Kryszkiewicz, M.: Properties of Incomplete Information Systems in the framework of Rough Sets. In: Polkowski, L., Skowron, A. (eds.) Rough Set in Knowledge Discovery 1: Methodology and Applications, Studies in Fuzziness and Soft Computing, vol. 18, pp. 422–450. Physica Verlag, Heidelberg (1998)
10. Kryszkiewicz, M.: Rules in Incomplete Information Systems. Information Sciences 113, 271–292 (1999)
11. Latkowski, R.: On Decomposition for Incomplete Data. Fundamenta Informaticae 54, 1–16 (2003)

12. Leung, Y., Li, D.: Maximum Consistent Techniques for Rule Acquisition in Incomplete Information Systems. Information Sciences 153, 85–106 (2003)
13. Nakata, N., Sakai, H.: Rough-set-based approaches to data containing incomplete information: possibility-based cases, pp. 234–241. IOS Press, Amsterdam, Trento, Italy (2005)
14. Nakata, N., Sakai, H.: Checking Whether or Not Rough-Set-Based Methods to Incomplete Data Satisfy a Correctness Criterion. In: Torra, V., Narukawa, Y., Miyamoto, S. (eds.) MDAI 2005. LNCS (LNAI), vol. 3558, pp. 227–239. Springer, Heidelberg (2005)
15. Nakata, N., Sakai, H.: Rough Sets Handling Missing Values Probabilistically Interpreted. In: Ślęzak, D., Wang, G., Szczuka, M., Düntsch, I., Yao, Y. (eds.) RSFDGrC 2005. LNCS (LNAI), vol. 3641, pp. 325–334. Springer, Heidelberg (2005)
16. Nakata, N., Sakai, H.: Lower and Upper Approximations in Data Tables Containing Possibilistic Information, Transactions on Rough Sets VII, 170–189 (2007)
17. Orłowska, E., Pawlak, Z.: Representation of Nondeterministic Information. Theoretical Computer Science 29, 313–324 (1984)
18. Parsons, S.: Current Approaches to Handling Imperfect Information in Data and Knowledge Bases. IEEE Transactions on Knowledge and Data Engineering 8(3), 353–372 (1996)
19. Pawlak, Z.: Rough Sets. International Journal of Computer and Information Sciences 11, 341–356 (1982)
20. Sakai, H.: Effective Procedures for Handling Possible Equivalence Relation in Nondeterministic Information Systems. Fundamenta Informaticae 48, 343–362 (2001)
21. Sakai, H., Nakata, M.: An Application of Discernibility Functions to Generating Minimal Rules in Non-deterministic Information Systems. Journal of Advanced Computational Intelligence and Intelligent Informatics 10, 695–702 (2006)
22. Sakai, H., Okuma, A.: Basic Algorithms and Tools for Rough Non-deterministic Information Systems, Transactions on Rough Sets I, 209–231 (2004)
23. Słowiński, R., Stefanowski, J.: Rough Classification in Incomplete Information Systems. Mathematical and Computer Modelling 12(10/11), 1347–1357 (1989)
24. Stefanowski, J., Tsoukiàs, A.: On the Extension of Rough Sets under Incomplete Information. In: Zhong, N., Skowron, A., Ohsuga, S. (eds.) New Directions in Rough Sets, Data Mining, and Granular-Soft Computing. LNCS (LNAI), vol. 1711, pp. 212–219. Springer, Heidelberg (1999)
25. Stefanowski, J., Tsoukiàs, A.: Incomplete Information Tables and Rough Classification. Computational Intelligence 17(3), 545–566 (2001)
26. Ziarko, W.: Probabilistic Rough Sets. In: Ślęzak, D., Wang, G., Szczuka, M., Düntsch, I., Yao, Y. (eds.) RSFDGrC 2005. LNCS (LNAI), vol. 3641, pp. 283–293. Springer, Heidelberg (2005)
27. Ziarko, W.: Stochastic Approach to Rough Set Theory. In: Greco, S., Hata, Y., Hirano, S., Inuiguchi, M., Miyamoto, S., Nguyen, H.S., Słowiński, R. (eds.) RSCTC 2006. LNCS (LNAI), vol. 4259, pp. 38–48. Springer, Heidelberg (2006)
28. Zimányi, E., Pirotte, A.: Imperfect Information in Relational Databases. In: Motro, A., Smets, P. (eds.) Uncertainty Management in Information Systems: From Needs to Solutions, pp. 35–87. Kluwer Academic Publishers, Boston, MA (1997)