On Probability of Matching in Probability Based Rough Set Definitions

Do Van Nguyen*, Koichi Yamada**, and Muneyuki Unehara***
Department of Management and Information Systems Science
Nagaoka University of Technology
Nagaoka, Japan
E-mail: *ngdovan@stn.nagaokaut.ac.jp, **yamada@kjs.nagaokaut.ac.jp, ***unehara@kjs.nagaokaut.ac.jp

***Abstract*—The original rough set theory deals with precise and complete data, while real applications frequently contain imperfect information. A typical imperfect data studied in rough set research is the missing values. Though there are many ideas proposed to solve the issue in the literature, the paper adopts a probabilistic approach, because it can incorporate other types of imperfect data including imprecise and uncertain values in a single approach. The paper first discusses probabilities of attribute values assuming different type of attributes in real applications, and proposes a generalized method of probability of matching. It also discusses the case of continuous data as well as discrete one. The proposed probability of matching could be used for defining valued tolerance/similarity relations in rough set approaches.**

I. Introduction

Classical Rough set theory [23], [24] provides a mathematical tool to analyse databases under objects description. Objects characterized by some attributes may be indiscernible in view of the available information about them. The original rough sets approach presupposes that all objects in an information system have precise and complete attribute values. Problems arise when information systems contain imperfect data, which occasionally happens in the real world. Therefore, it is necessary to develop a theory which enables classifications of objects even if there is only partial information available.

Controversial rough set research mostly considers that imperfect data in information systems comes from missing values [7], [9], [10], [12], [13], [14], [26], [27], [29], [31]. An information system with missing values is called incomplete information system [13], [14]. In incomplete information systems, **Table I** for example [10], objects may contain several unknown attribute values. Unknown values are denoted by special symbol "*". In studies of rough set in incomplete information systems, some probabilistic solutions have been introduced based on the possibility of "missing value" [7], [19], [20], [26], [27]. Among them, some approaches [7], [26] suppose a priori assumption that there exists a uniform probability distribution on every attribute domain and compute valued tolerance (or similarity) classes based on the joint probability distribution. In this paper, we aim to generalize the method of determining the probability that two object may be tolerant of (similar to) each other on an attribute. The probability of matching will be defined based on the probability that two objects may take the same values on an attribute in the dataset.

TABLE I. An example of a dataset with missing values

Cases	Temperature	Headache	Nausea	Flu
x1	High	*	no	yes
x2	Very-high	yes	yes	yes
x3	*	no	no	no
x4	High	yes	yes	yes
x5	High	*	yes	no
x6	Normal	yes	no	no
x7	Normal	no	yes	no
x8	*	yes	*	yes

TABLE II. An information system with uncertainties

Employees	Deterministic Department	Stochastic Quality Bonus
Jon Smith	Toy	0.4[Great Yes] 0.4[Good Yes] 0.1[Fair Yes]
Fre Jones	Housewares	1.0[Good Yes]

Besides the missing values, there are many reasons why imperfect data are produced in datasets [16]. *Imprecision* is another type of possible imperfect data. Stored information is imprecise when it denotes a set of possible value and the real value is one of elements of this set. For example, John's age is between 30 and 35 or John's age is over 30.

One more possible type of imperfect data is *Uncertainty* [16]. Whereas the statement "John's age is either 31 or 32" is in the form of imprecision, the statement "John is probably 32" or "John's age is 32 with confidence 0.6" denotes uncertainty. Both imprecise and uncertain values can be represented by probabilistic data [4]. **Table II** [1] illustrates an information system with probabilistic information.

Hence, in information systems containing missing, imprecise and uncertain values, it is not appropriate to apply to the systems a method that can deal only with missing values. We must rely on a probabilistic approach that could be applied to these types of imperfect data. Therefore, it is important to define probability of matching as a step to define valued tolerance/similarity relations in information systems which contain both attributes with missing values and attributes with probabilistic data.

This paper is organized as follows; Section II re-introduces rough set theory as well as probabilistic relation definitions. Section III will suggest several methods to determine ***Probability of object attribute values*** in case of imprecise and missing

978-1-4799-0652-9/13 $31.00 © 2013 IEEE
DOI 10.1109/SMC.2013.82

values. Then, in section IV, we will generalize ***Probability of matching*** definitions. The probability of matching can be also extended for the case of continuous values in section V. Eventually, section VI briefly introduces an application of the proposed approach.

II. Probabilistic relation in rough set approach

Studies in [7], [19], [20], [26], [27] introduce probabilistic relations as follows: first, a probability distribution is defined on the domain of each attribute, and the probability that a pair of objects are tolerant of (similar to) each other on the attribute is determined. Then, the degree that two objects are tolerant of (similar to) each other on a set of attributes is calculated, for example, using the joint probabilities. This section will summarize the rough set approach as well as some concepts in probabilistic relation definitions. The problems of the former probability-based rough set approaches are also addressed in this section.

An information system in rough set study 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 $a \in A$ [23], [24]. The non-empty discrete value set V_a is called the domain of a. The original rough set theory deals with complete information systems in which $\forall x \in U$, $a \in A$, $f_a(x)$ is a precise value.

The relation $EQU_P(x, y)$, $P \subseteq A$ denotes a binary relation between objects that are equivalent in terms of values of attributes in P [23]. The equivalence relation is reflexive, symmetric, and transitive. Let $E_P(x) = \{y \in U | EQU_P(y, x)\}$ be the set of all objects that are equivalent to x by P, and is called equivalence class.

Any information system of the form $I = (U, A \cup \{d\})$ is called decision table where $d \notin A$ is called decision and the elements of A are called conditions. We assume $V_d = \{d_1, ..., d_r\}$. The decision d determines a partition of the universe U, where $C_k = \{x \in U | f_d(x) = d_k\}$, $1 \leq k \leq r$. The set C_k is called the decision class or concept on U.

In this paper, $f_a(x) = *$ denotes attribute value of x is missing on a. However, we assume that any attribute domain does not contain the special symbol "*" representing the missing value and that the real value could be found in the domain.

Now, for an information system, in which some attribute values of objects are missing and/or associated with probabilistic data, we define probabilities of attribute values. For a discrete attribute, ***Probability of object attribute value*** denoted by $Pr_a(f_a(x) = v)$ represents the probability that $x \in U$ takes the value $v \in V_a$ on attribute $a \in A$. Several methods to estimate the probabilities of object attribute values will be discussed in the next section.

Based on the probability of object attribute value, ***Probability of matching*** between two objects $x, y \in U$ on attribute $a \in A$ denoted by $\theta_a(x, y)$ defines the probability that object x takes the same value as object y on attribute a. In [7], [26], [27], it is supposed that there is an uniform probability distribution on an attribute, and the probability of matching is defined as $\theta_a(x, y) = Pr_a(f_a(x) = v_k)) \times Pr_a(f_a(y) = v_k) = 1/|V_a|^2$ where v_k is a value in the domain of "a". The definition is clearly inadequate when we suppose the attribute values of both "x" and "y" are missing on "a". The definition of probability of matching is discussed and calculated in several cases in section III.

From the probability of matching between two objects, we can induce the degree that $x, y \in U$ are tolerant of (similar to) each other on a set of attributes $P \subseteq A$, which is denoted by R_P. The degree of tolerance/similarity can be defined as the probability that two objects have the same values on all attributes in set P [7], [26], [27] and calculated by joint probability $R_P = \prod_{a \in P} \theta_a(x, y)$. Other methods of tolerance (similarity) degree definitions can be found in [19], [20].

III. Probability of object attributes values

In some studies with missing values [7], [26], [27], the probability of object attribute values is determined by uniform probability distribution. In this kind of attribute, assuming that the set of possible values on this attribute is discrete we do not know any information about probability distribution of attribute values. Hence, we have to make the hypothesis that all values have the same probability to be associated to an element of data set. Consider an attribute $a \in A$ and its domain $V_a = \{v_1, v_2, ..., v_m\}$, given an object $x \in U$, if $f_a(x) = *$ the probability $Pr_a(f_a(x) = v_i) = 1/|V_a|$ for any $v_i \in V_a$. However, depending on their characteristics, there are possibly numerous kinds of distribution on attribute domains. This section will discuss some cases where they are not uniform probability distributions.

A. Probability of attribute values in case of imprecision

In case of imprecision, the value of object attribute is described by a set of possible value. Thus, we are able to suppose that the probabilities of every values in the set of choice are equal, such that $Pr_a(f_a(x) = v) = \frac{1}{|T_a(x)|}$ if $v \in T_a(x)$. Actually, *precise* value and *missing* value can be considered as two extreme kinds of imprecision [16]. A value is precise when the set of possibilities is a singleton. In this case $|T_a(x)| = 1$. Missing value with equal probability could be regarded as imprecise information where the set of possible values encompasses the entire attribute domain, such that $T_a(x) = V_a$. In this research, we distinguish missing and imprecise values. Probability of attribute value in case of missing value can be determined based on various kind of distribution.

B. On attributes with pre-defined probability distribution of values

On some attributes, it may be possible to assume that there exists a pre-defined probability distribution among attribute values. One example is a game of four people playing with dice. Their scores can be calculated based on the sum of two dice thrown for each of them. **Table III** shows their scores. In this table, as we can see that the score of Tom is unknown due to some reasons. However, the probability of each value for Tom's score can be identified by probability distribution for the sum of two dice. The probability that Tom's score is 7, for example, is 1/6. On the other hand, the probability that his score equals to 11 is 1/18. Hence, the probability of object attribute value can be assigned the probability function, such

TABLE III. THE SCORES OF A DICE GAME.

Players	Score
Terry	3
David	6
Tom	*
Anna	8

TABLE IV. PROBABILITY OF ATTRIBUTE VALUES

Attributes	Values	Probability
Temperature	very-hig	0.17
Temperature	high	0.50
Temperature	normal	0.33
Headache	yes	0.67
Headache	no	0.33
Nausea	yes	0.57
Nausea	no	0.43

that $Pr_a(f_a(x) = v) = \lambda_a(v)$ if $f_a(x) = *$, where "v" is an element of V_a, $\lambda_a(v)$ is the probability mass function on "a".

C. Method of the frequency of attribute value

We first study the method in [19]. The approach is based on the notion of "The most common method". This is a method of handling missing value summarized by Grzymala-Busse [8], [10], in which, missing values are replaced by the most common value of the attribute. The method of handling missing attribute values is implemented, e.g., in well-known machine learning algorithm CN2 [6].

Suppose the value domains are known. First, we define the probability that each value of the attribute appears based on the frequency of the available value in dataset. The probability that a value $v \in V_a$ appears as a value of a certain object is define by:

$$\rho_a(v) = \begin{cases} \dfrac{|V_a(v)|}{|U - V_a(*)|} & \text{if } V_a(*) \subset U, \\ \dfrac{1}{|V_a|} & \text{otherwise.} \end{cases} \quad (1)$$

where $V_a(v)$ and $V_a(*)$ are sets of objects whose attribute value is "v" and the set of objects whose value on "a" is missing, respectively. The symbol "$\subset$" denotes a proper subset. As seen in the equation, the probability $\rho_a(v)$, $v \in V_a$ is defined by the ratio of the value "v" among objects whose values are not missing. If $V_a(*) = U$, that is, values of attribute "a" are missing in all objects, the equal probability distribution is given. The value of $\rho_a(v)$ is greater than zero if there is at least an object such that $f_a(x) = v$. Since it could be zero for many values if the size of U is small, the size of U should be large enough when using the approach.

The probabilities of attribute values are illustrated in the **Table IV**. From this table, we can see that the value "high" of "Temperature" occurs more frequently than the other values. The most frequent values of "Headache" and "Nausea" happen to be "yes".

Now, we define the probability of object attribute values by the frequency of values in a dataset. Formally, in incomplete information system $I = (U, A)$, an attribute $a \in A$ and its domain $V_a = \{v_1, v_2, ...v_m\}$, $\rho_a(v_i)$ denotes the frequency of each value $v_i \in V_a$ in the dataset. Given an object $x \in U$, if $f_a(x) = *$ the probability $Pr_a(f_a(x) = v_i) = \rho_a(v_i)$ for any $v_i \in V_a$.

TABLE V. PROBABILITY OF ATTRIBUTE VALUES RELATED TO CONCEPTS

Attributes	Values	Probability in concepts	
		Flu=Yes	*Flu=No*
Temperature	very-hig	0.33	0.00
Temperature	high	0.67	0.33
Temperature	normal	0.00	0.67
Headache	yes	1.00	0.33
Headache	no	0.00	0.67
Nausea	yes	0.67	0.50
Nausea	no	0.33	0.50

D. Method of the frequency of attribute value related to concepts

This method is first discussed in [20] and is an extension of the previous method. Observing some systems, we recognized that attribute values might relate to some concepts. Supposed the value domains are known, the probability that a value $v \in V_a$ appears as a value of objects contained in a concept $X \subseteq U$ is defined as follows:

$$\rho_a(v)_X = \begin{cases} \dfrac{|V_a(v)_X|}{|X - V_a(*)_X|} & \text{if } V_a(*)_X \subset X, \\ \dfrac{1}{|V_a|} & \text{otherwise.} \end{cases} \quad (2)$$

where $V_a(v)_X$ and $V_a(*)_X$ are sets of objects in concept X whose attribute value is "v" and the set of objects whose value on "a" is missing, respectively.

From **Table V**, we can see that flu relates to high and very-high temperature, headache and nausea. Meanwhile non-flu corresponds to the cases of low temperature, no headache.

Like the previous method, it is possible to define the probability of object attribute values by the frequency of values in a dataset. Formally, in incomplete information system $I = (U, A)$, an attribute $a \in A$ and its domain $V_a = \{v_1, v_2, ...v_m\}$, $\rho_a(v_i)_X$ denotes the frequency of each value $v_i \in V_a$ in concept X. Given an object $x \in X$, if $f_a(x) = *$ the probability $Pr_a(f_a(x) = v_i) = \rho_a(v_i)_X$ for any $v_i \in V_a$.

IV. PROBABILITY OF MATCHING

This section will re-define the degree that two object have the same value on an attribute if at least one of the two objects has the missing, imprecise or uncertain value on the attribute.

***Definition** 4.1:* Given an information system $I = (U, A)$, on an attribute $a \in A$ with its domain V_a, the probability that the value of "x" matching with the value of "y" on "a" is given by:

$$\theta_a(x, y) = \sum_{v_i \in V_a} Pr_a(f_a(x) = v_i | f_a(y) = v_i) Pr_a(f_a(y) = v_i) \quad (3)$$

when $x \neq y$. Otherwise $\theta_a(x, y) = \theta_a(x, x) = 1$. $Pr_a(f_a(x) = v_i | f_a(y) = v_i)$ denotes the conditional probability of $f_a(x) = v_i$ given $f_a(y) = v_i$.

It is not common to see that the occurrence of $f_a(y) = v_i$ affect the probability of $f_a(x) = v_i$. Hence, in the rest of the paper, to define probability of matching, we assume that two events $f_a(x) = v_i$ and $f_a(y) = v_j$, $x, y \in U$, $a \in A$ are independent with each other for any $v_i, v_j \in V_a$.

A. Probability of matching on attribute with probabilistic values

Let $I = (U, A)$ be an information system, an probabilistic associated attribute $a \in A$. In case of uncertainty, for each object $x \in U$, the probability $Pr_a(f_a(x) = v)$ for any possible value $v \in V_a$ is given in the dataset. Otherwise, for imprecision, the probability of attribute value can be calculated by the method shown in section III. Formally, let $T_a(x) = \{v \in V_a | Pr_a(f_a(x) = v) > 0\}$, the probability of matching between two objects on "a" is given by:

$$\theta_a(x,y) = \begin{cases} \sum\limits_{v_i \in T_a(x) \cap T_a(y)} Pr_a(f_a(x) = v_i) Pr_a(f_a(y) = v_i) \\ \qquad\qquad \text{if } T_a(x) \cap T_a(y) \neq \emptyset, \\ 0 \qquad\qquad \text{if } T_a(x) \cap T_a(y) = \emptyset, \end{cases} \tag{4}$$

when $x \neq y$. Otherwise $\theta_a(x, y) = \theta_a(x, x) = 1$.

In case of probabilistic associated attribute, the probability that two objects have the same value for an attribute is given by the joint probability of the attribute value. Thus, to calculate the probability of matching between two objects, the sum of the probabilities should be taken for the set of values so that these values and their probability would be available for both two objects.

B. Probability of matching on attribute with missing values

In incomplete information system $I = (U, A)$, on a attribute $a \in A$ with its domain V_a , the probability of matching between two objects on "a" is given as follows:

$$\theta_a(x,y) = \begin{cases} Pr_a(f_a(y) = f_a(x)) \quad \text{if } f_a(x) \neq *, f_a(y) = *, \\ Pr_a(f_a(x) = f_a(y)) \quad \text{if } f_a(x) = *, f_a(y) \neq *, \\ \sum\limits_{v_i \in V_a} Pr_a(f_a(x) = v_i) Pr_a(f_a(y) = v_i) \\ \qquad\qquad \text{if } f_a(x) = *, f_a(y) = *, \end{cases} \tag{5}$$

when $x \neq y$. Otherwise $\theta_a(x, y) = \theta_a(x, x) = 1$.

If one of the two objects has a certain value, $f_a(x)$ for example, the probability value that $f_a(y) = f_a(x)$ is given by $Pr_a(f_a(y) = f_a(x))$. If both of them are missing, for each value of the domain, the probability that they are equal is given by the joint probability of attribute value. For the whole domain, the sum of such joint probability should be taken.

In (5), if two objects $x, y \in U$ contain missing value on an attribute, depending on the characteristics of the attribute, the probabilities of attribute value of the two objects may and may not be the same for each value. For example, in incomplete information systems with pre-defined value probability distribution $\lambda_a(v)$ of attribute "a", the probability of matching between two separate objects $x, y \in U$ is $\theta_a(x, y) = \sum_{v_i \in V_a} \{\lambda_a(v_i)\}^2$ if $f_a(x) = *, f_a(y) = *$. On the other hand, when the probability of attribute is defined based on the *Method of the frequency of attribute value related to concepts*, the probability of matching between $x \in X$,$y \in Y$ is $\theta_a(x, y) = \sum_{v_i \in V_a} \rho_a(v_i)_X \rho_a(v_i)_Y$ if $f_a(x) = *, f_a(y) = *$.

V. Discussion for continuous values

In information system coming with continuous value, keeping the consistency of information systems, continuous attributes have to be transformed into discrete ones. The solution is the discretization of these attributes into ranges where each interval is mapped to a discrete value [2], [3], [21], [30]. In general, the target of such studies is to find the minimum interval without weakening the discernibility in the dataset.

On continuous attributes containing imperfect data, the indiscernibility relation is not available at all. There exists a way to deal with them using rough set technique. First discretizing the continuous data to discrete data [5], and then finding the attribute reduction using methods proposed in [9], [11], [12], [13], [14], [26], [27], [29], [31].

Let $I = (U, A)$ be an information system. Any pair (a, c), where continuous attribute $a \in A$, $c \in \mathbb{R}$, will be called a cut on V_a. For $a \in A$, any set of cuts $\{(a, c_1^a), (a, c_2^a), ..., (a, c_k^a)\}$ on $V_a = [v_{min}^a, v_{max}^a) \subset \mathbb{R}$ defines a partition $V_a' = \{[c_0^a, c_1^a), [c_1^a, c_2^a), ...[c_k^a, c_{k+1}^a)\}$ where $v_{min}^a = c_0^a < c_1^a < ... < c_k^a < c_{k+1}^a = v_{max}^a)$, and $V_a = [c_0^a, c_1^a) \cup [c_1^a, c_2^a) \cup ... \cup [c_k^a, c_{k+1}^a)$. Therefore, any set of cuts defines a new attribute domain V_a' on "a" and the equivalence between two object on "a" [21] is defined as

$$EQU_{\{a\}}(x, y) \Leftrightarrow \big(\text{iff } f_a(x), f_a(y) \in [c_i^a, c_{i+1}^a)\big) \tag{6}$$

On attributes associated with continuous values, two objects are equivalent if their attribute values fall in the same interval. If there is a missing, uncertain or imprecise value , the equivalence relation cannot be determined. We have to define the degree of tolerance (similar) instead. For continuous values, the probability that an attribute value of $x \in U$ on attribute $a \in A$ falls into an interval, say $[c_1, c_2) \subseteq V_a$, is given by $Pr_a(c_1 \leq f_a(x) < c_2)$. From the probability of object attribute value, we are able to define a valued tolerance/similarity relation.

A. Continuous attributes with uncertainty and imprecision

In case of uncertainty, the possible values and their probability is provided in the data set. Let $T_a(x) = \cup\{v \in V_a | Pr_a(f_a(x) = v) > 0\}$, we define the probability that continuous value falls in interval $[c_1, c_2)$ as the following equation:

$$Pr_a(c_1 \leq f_a(x) < c_2) = \sum_{\substack{v \in T_a(x) \\ c_1 \leq v < c_2}} Pr_a(f_a(x) = v) \tag{7}$$

In case of imprecision, the possible value of object attribute value is often described by a set of possible ranges. The probability that continuous values of a range falls in an interval could be defined as how large the interval cover the range.

Formally, let $[v_1, v_2]$, $v_1, v_2 \in \mathbb{R}$, $v_2 > v_1$ denotes a possible range of an imprecise value. The probability that a continuous value v in the range falls in interval $[c_1, c_2]$ can be determined as follows:

$$Pr_a(c_1 \leq v \leq c_2) = \frac{\min(c_2, v_2) - \max(c_1, v_1)}{v_2 - v_1} \tag{8}$$

Let $T_a(x) = \{[v_{1,1}, v_{1,2}], [v_{2,1}, v_{2,2}], ..., [v_{t,1}, v_{t,2}]\}$ be the set of possible range, which is described in the data set for object x with imprecision, we define the probability that continuous value falls in interval $[c_1, c_2)$ as follows:

$$\begin{aligned} &Pr_a(c_1 \leq f_a(x) < c_2) \\ &\quad = \lim_{\varepsilon \to 0} \sum_{i=1}^{t} \frac{\min(c_2 - \varepsilon, v_{i,2}) - \max(c_1, v_{i,1})}{v_{i,2} - v_{i,1}} \\ &\quad = \sum_{i=1}^{t} \frac{\min(c_2, v_{i,2}) - \max(c_1, v_{i,1})}{v_{i,2} - v_{i,1}} \end{aligned} \tag{9}$$

Now we define the probability of matching for objects containing uncertainty and imprecision. Let $I = (U, A)$ be an information system, an attribute $a \in A$ with uncertainty or imprecision. The probability of attribute value can be determined by (7) and (9). Note that in case of precise value, say $f_a(x) = v \in V_a$, $Pr_a(c_i^a \leq f_a(x) < c_{i+1}^a) = 1$ if $v \in [c_i^a, c_{i+1}^a)$. Let $T'_a(x) = \{c_i^a | Pr_a(c_i^a \leq f_a(x) < c_{i+1}^a) > 0\}$, we define the probability of matching as follows:

$$\theta_a(x,y) = \begin{cases} \sum\limits_{c_i^a \in T'_a(x) \cap T'_a(y)} Pr_a(c_i^a \leq f_a(x) < c_{i+1}^a) Pr_a(c_i^a \leq f_a(y) < c_{i+1}^a) \\ \qquad\qquad \text{if } T'_a(x) \cap T'_a(y) \neq \emptyset, \\ 0 \qquad\qquad \text{if } T'_a(x) \cap T'_a(y) = \emptyset, \end{cases} \tag{10}$$

when $x \neq y$. Otherwise $\theta_a(x, y) = \theta_a(x, x) = 1$.

B. Continuous attributes with missing values

When a random variable takes values from a continuous range, in some cases, we have to do experiments to estimate probability distribution of the data. In other cases, the data is already described by a known probability distribution such as Gaussian, Laplace, Gamma distribution [15], [22]. In information system $I = (U, A)$, suppose a probability function $\lambda_a(v)$, $v \in V_a$ (that is called probability density function for continuous values). The probability that continuous value falls in interval $[c_1, c_2]$ is determined by an integral:

$$Pr_a(c_1 \leq f_a(x) \leq c_2) = \int_{c_1}^{c_2} \lambda_a(v) d(v) \tag{11}$$

For the interval $[c_1, c_2) \subseteq V_a$, the probability is

$$Pr_a(c_1 \leq f_a(x) < c_2) = \lim_{\varepsilon \to 0} \int_{c_1}^{c_2 - \varepsilon} \lambda_a(v) d(v) \tag{12}$$

In incomplete information system, when the attribute values of the two objects are present, the relation between these objects can be determined by (6). If at least one of two object attribute values is missing, probability of matching can be defined as the following equation:

$$\theta_a(x,y) = \begin{cases} Pr_a(c_i^a \leq f_a(y) < c_{i+1}^a) \\ \qquad \text{if } f_a(x) \in [c_i^a, c_{i+1}^a), f_a(y) = *, \\ Pr_a(v_i^a \leq f_a(x) < c_{i+1}^a) \\ \qquad \text{if } f_a(y) \in [c_i^a, c_{i+1}^a), f_a(x) = *, \\ \sum\limits_{c_i = c_0^a}^{c_k^a} Pr_a(c_i^a \leq f_a(x) < c_{i+1}^a) Pr_a(c_i^a \leq f_a(y) < c_{i+1}^a) \\ \qquad \text{if } f_a(x) = *, f_a(y) = *, \end{cases} \tag{13}$$

when $x \neq y$. Otherwise $\theta_a(x, y) = \theta_a(x, x) = 1$.

Hence, with a pre-defined probability distributed function $\lambda_a(v)$, the probability of matching between two separate object is:

$$\theta_a(x,y) = \begin{cases} \lim\limits_{\varepsilon \to 0} \int\limits_{c_i^a}^{c_{i+1}^a - \varepsilon} \lambda_a(v) d(v) \text{ if } f_a(x) \in [c_i^a, c_{i+1}^a), f_a(y) = *, \\ \lim\limits_{\varepsilon \to 0} \int\limits_{c_i^a}^{c_{i+1}^a - \varepsilon} \lambda_a(v) d(v) \text{ if } f_a(y) \in [c_i^a, c_{i+1}^a), f_a(x) = *, \\ \lim\limits_{\varepsilon \to 0} \int\limits_{v_{min}}^{v_{max} - \varepsilon} (\lambda_a(v))^2 d(v) \qquad \text{if } f_a(x) = *, f_a(y) = *, \end{cases} \tag{14}$$

Equation (10) and (13) show that in information systems containing continuous attributes with missing, uncertain or imprecise values, it is possible to use probability of matching for defining probabilistic based tolerance (similarity) relations. On the other hand, in such kind of information systems, we may be able to define a distance function such as $distance_a(f_a(x), f_a(y)) = 1 - \theta_a(x, y)$ for defining similarity relation based on distance [28].

TABLE VI. KANSEI INFORMATION TABLE FOR MOBILE PHONE DESIGN

U	a_1	a_2	a_3	a_4	D
x1	2	1	0	0.8[metal],0.2[plastic]	deluxe
x2	{1,2}	1	{0,1}	1.0[metal]	deluxe
x3	2	{0,1}	1	0.7[metal],0.3[plastic]	deluxe
x4	0	2	{1,2}	0.1[metal],0.8[plastic]	cute
x5	1	0	2	0.2[metal],0.8[plastic]	cute
x6	1	{0,1}	1	1.0[plastic]	sporty
x7	1	0	{0,1}	0.1[metal],0.9[plastic]	sporty
x8	2	0	0	1.0[plastic]	sporty

VI. APPLICATION IN KANSEI ENGINEERING

In actual application, there are many situation in which, we have to describe examples in imprecise as well as uncertain representation rather than singleton values such as human feeling and impression used in Kansei engineering [17], [18]. Data of a WEB-based form feature extraction system for mobile phone design [25], [32] is an example of imperfect information. Let the product feature set be $A = \{a_1, a_2, a_3, a_4\}$, which denotes body shape, body ratio, bottom shape and material, respectively, and Kansei adjectives set be $V_d = \{deluxe, cute, sporty\}$. Each kind of body shape, body ratio, conner shape can be shown by a picture (see [25] for more detail). Considered material types are plastic and metal [32]. Then numerous interviewees were invited to give their felling

about a desired product. Using numbers to represent the type of features, the Kansei data is illustrated in **Table VI**. Now, we want to extract Kansei knowledge from the data collected using rough set theory. It probably is impracticable to employ a simple traditional solution in this situation. However, working with probabilistic method may help us define a rough set model. Consequently, Kansei knowledge acquisition methods [25], [32] within the proposed approach can be used in this kind of information.

VII. Conclusion

In this paper, the method ***Probability of Matching*** is introduced. The approach can be used in any study about valued tolerance/similarity relations in information systems containing imperfect data. The method allows defining probabilistic rough set models with both discrete and continuous values. Furthermore, it avoids the inadequateness of the former studies.

To calculate probability of matching, the paper also suggests several ways to determine ***Probability of object attribute values***. In some cases, the probability distributions of attribute values are provided. In the other cases, we have to use equal probability or statistic method such as method of the frequency of attribute values to define the probability of attribute values. It depends on the characteristic of each information system.

References

[1] Babara D., Garcia-Monila H. and Porter D., "The Management of Probabilistic Data," IEEE Transactions on Knoweldge and Data Engineering, vol.4, no.5, 1992.

[2] Beynon M. J., "Stability of continuous value discretisation: an application within rough set theory," International Journal of Approximate Reasoning 35, pp. 29-53, Elsevier Inc, 2004.

[3] de Carvalho M.A., de Moraes C.H.V., Lambert-Torres G., Borges da Silva L.E., Aoki A.R., Vivaldi A., "Transforming ContinuousAttributes using GA for Applications of Rough Set Theory to Control Centers," Intelligent System Application to Power Systems (ISAP), pp. 1-8 , 2011.

[4] Cavallo R. and Pittarelli M., "The theory of probabilistic databases," Proceeding of the 13th Very Large Database, Brighton, 1987.

[5] Chmielewski M. R., Grzymala-Busse J. W., "Global Discretization of Continuous Attributes as Preprocessing for Machine Learning," International Journal of Approximate Resoning 15, pp.319-331, Elsevier Inc 1996.

[6] Clark P. and Niblett T., "The CN2 Induction Algorithm," Machine Learning, vol.3, no.4, pp.261-283, 1989.

[7] Feng Y., Li W., Lv Z. and Ma X., "Probabilistic approximation under incomplete information systems," IFIP International Federation for Information Proceeding, vol. 228, Intelligent Information processing III, eds. Shi Z., Shimohara K., Feng D., Springer, Boston pp. 71-80, 2006.

[8] Grzymala-Busse J. W. and Hu M., "A comparison of several approaches to missing attribute values in data mining," RSCTC 2000, LNAI 2005, pp.378-385, Springer-Verlag, 2001.

[9] Grzymala-Busse J. W., "Characteristic relations for incomplete data: A generalization of the indiscernibility relation," In: Proceedings of the RSCTC2004, Fourth International Conference on Rough Sets and Current Trends in Computing, Uppsala,Sweden, pp.244253, 2004.

[10] Grzymala-Busse J. W. and Grzymala-Busse W. J., "Handling missing attribute values," in The Data Mining and Knowledge Discovery Handbook, eds Olded Maimon and Lior Rokach, Springer-Verlag, pp.37-57, 2005.

[11] Grzymala-Busse J. W., "Incomplete data and generalization of indiscernibility relation, definability, and approximations," In: Proceedings of the RSFDGrC2005, Tenth International Conference on Rough Sets, Fuzzy Sets, data Mining, and Granular Computing, Springer-Verlag, pp.244253, 2005.

[12] Grzymala-Busse J.W., "A Rough Set approach to Data with Missing Attribute values," In: Proceedings of the RSKT 2006, LNAI 4062, Springer-Verlag, pp.5867, 2006.

[13] Kryszkiewicz M., "Rough set approach to incomplete information system," Information Sciences, vol.112, issues 1-4, pp.39-49, 1998.

[14] Kryszkiewicz M., "Rules in incomplete information systems," Information Sciences, vol.113, issues 3-4, pp.271-292, 1999.

[15] Miller S. L., Childers D., "Probability and Random Process With Application to Signal Processing and Communications," Elsevier Inc, 2004.

[16] Motro A., "Uncertainty management information systems: From needs to solutions", Kluwer Acad. Publisher, Chapter 2, 1997.

[17] Nagamachi M., "KANSEI Engineering: a new ergonomics consumer-oriented technology for product development," Int. J. Ind. Des. 15 (1), pp.311, 1995.

[18] Nagamachi M., "KANSEI Engineering as a powerful consumer-oriented technology for product development," Appl. Ergonom. 33 (3) pp.289294, 2002.

[19] Nguyen D. V., Yamada K., Unehara M., "Knowledge reduction in incomplete decision tables using Probabilistic Similarity-Based Rough set Model," 12thInternational Symposium on Advanced Intelligent Systems (ISIS 2011), pp.147-150, Suwon, Korea, Sep. 29, 2011.

[20] Nguyen D. V., Yamada K., Unehara M., "Rough Set Model Based on Parameterized Probabilistic Similarity Relation in Incomplete Decision Tables," SCIS-ISIS2012, Kobe, Japan,Nov, 2012.

[21] Nguyen. H. S, "Discretization Problem for Rough Sets Methods,"Rough Sets and Current Trends in Computing, Lecture Notes in Computer Science, Vol. 1424, pp 545-552, Springer-Verlag, 1998.

[22] Papoulis A., Pillai S. U., "Probability, random variables, and stochastic process," Forth edition, McGraw-Hill, 2002.

[23] Pawlak Z., "Rough Sets," International Journal of Computer and Information Sciences, vol.11, pp.341-356, 1982.

[24] Pawlak Z., "Rough Sets. Theoretical Aspects of Reasoning about Data," Kluwer Acad. Publisher, 1991.

[25] Shi F.,Sun S., Xu J., "Employing rough sets and association rule mining in KANSEI knowledge extraction," Information Sciences 196, pp.118128, 2012.

[26] Stefanowski J. and Tsoukias A., "On the extension ofrough sets under incomplete information," Lecture Notes in ArtificialIntelligence 1711, pp.73-81, 1999.

[27] Stefanowski J. and Tsoukias A., "Incomplete Information Tables and Rough Classification," Computational Intelligence 17th, pp.545-566, 2001.

[28] Stepaniuk J., "Approximation spaces, reducts and representatives," In: Lech Polkowski , J. Kacprzyk , A. Skowron, Rough Sets in Knowledge Discovery 2: Applications, Case Studies, and Software Systems, chapter 6, pp.109-126, Physica-Verlag, 1998

[29] Wang G. Y., "Extension of Rough Set under Incomplete Information-Systems". FUZZ-IEEE 2002, the 2002 IEEE International Conference on Fuzzy Systems, pp.1098-1103, 2002.

[30] Xin G., Xiao Y., You H., "Discretization of continous interval-valued attributes in rough set theory and its application," Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, HongKong, 19-22 August 2007.

[31] Yang X., Song X., Hu X., "Generalisation of rough set for rule induction in incomplete system," International Journal of Granular Computing, Rough Sets and Intelligent Systems, vol.2, number 1/2011, pp.37-50, 2011.

[32] Zhai L.-Y., Khoo L.-P., Zhong Z.-W. "A rough set based decision support approach to improving consumer affective satisfaction in product design," International Journal of Industrial Ergonomics, 39 (2) , pp. 295-302, 2009.