SIMILARITY-BASED ROUGH SET APPROACH IN INCOMPLETE INFORMATION SYSTEM USING POSSIBLE EQUIVALENT VALUE-SET

Asma' Mustafa, Rabiei Mamat*, and Ahmad Shukri Mohd Nor

Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia

Emails: asma.mmustafa92@gmail.com, rab@umt.edu.my*, ashukri@umt.edu.my

ABSTRACT

Data analytics generally helps businesses or entities to make better and efficient decision making. But in the face of growing volume of data or information, it becomes challenging to achieve these goals. One of which is on classification of information with high accuracy. Furthermore, when the information is incomplete, definitely it is more challenging in order to classify the information with high accuracy. Although incomplete information is well discussed using rough set theory for data classification, such as based on tolerance and similarity relations, there are still issues on accuracy to evaluate data classification. The main objective of this paper is to introduce a new similarity approach with semantically justified based on possible equivalent value-set related to incomplete information systems. It is based on a classification of three semantics types of incomplete information i.e., "any value", "maybe value" and "not applicable value" for modelling similarity. Subsequently, the similarity precision between objects in incomplete information systems is considered. The comparative studies and simulation results between the proposed approach in terms of accuracy on synthetic data, four well-known classification datasets and one real marine dataset are presented. The proposed approach improves the accuracy up to two orders of magnitude and, thus verifying its data classification accuracy.

Keywords: Incomplete information. similarity relation. possible equivalent value-set. similarity precision

1.0 INTRODUCTION

The rough set theory proposed by Pawlak [1] was a successful and effective tool to deal with many real-life problems such as in human resource management [2], financial management [3], decision making [4,5,6] and etcetera. The theory has also been used in the study of intelligent systems characterized by data uncertainty or inconsistent data especially in rule extraction [7], data clustering [8, 9, 10], granular computing [11,12], data reduction [13, 14, 15, 16], pattern recognition [17, 18] and data classification [5, 6, 19]. The theory is based on indiscernibility relation or equivalence relation where two objects are equivalent to each other if both objects have the same values for attributes. Based on that relation, we can group the objects to obtain equivalence classes, which is the main concern. Besides, it has been proven to be an efficient mathematical tool compared with principal component analysis, neural networks and support vector machine methods [20, 21]. Unlike those methods, the rough set theory allows knowledge discovering process to be conducted automatically by the data themselves without any dependence on the prior knowledge [22].

However, the rough set theory can only be used to analyse information systems with exactly known information called a complete information system where all available objects in an information system have attribute values. A problem arises when some of the attribute values in information systems are unknown or missing that gives an imprecise answer to some queries and subsequently will result in poor decision making which sometimes happens in the real world. This information system is called incomplete information system. Because some attribute values are missing in incomplete information systems, and therefore, it is hard to process the incomplete information systems with the indiscernibility relation. In other words, the indiscernibility relation in a complete information system shows deficiency when dealing with incomplete information systems. Therefore, many researchers improved the indiscernibility relation in a complete information system to different kind of relations to deal with

incomplete information systems. There have been many efforts in studying incomplete information systems to modelling indiscernibility relation, including the works done by [22 – 28].

In general, there are two main approaches to handling incomplete information systems. One is the indirect approach, which transforms the incomplete information systems into complete information systems by some specific rules, i.e. probability statistical methods. However, it may change the original information on incomplete information systems [29]. The other is the direct approach, which extends the classical rough set theory by reducing the requirements of indiscernibility relation of reflexivity, symmetry and transitivity. For instance, Kryszkiewicz [26] introduced indiscernibility based on tolerance relation where the missing attribute value in an object can be replaced by any known possible values in that attribute. However, tolerance relation produced poor results in terms of accuracy of approximation [31]. Consequently, Stefanowski and Tsoukias [26] proposed a similarity relation to refining the results obtained from a tolerance relation approach. Wang [32] proved that similarity relation will lose some information and proposed limited tolerance relation. Nevertheless, the accuracy of the approximation is still outstanding and thus, needs to be improved.

Therefore, in this paper, we proposed an approach based on possible world semantic represent by Lipski [33] in order to improve the accuracy of approximation that can help in obtain accurate results especially in decision making process. In recent years, there have been studies that used semantic in handling missing attribute values in incomplete information system. Kryszkiewicz [30] considers incomplete information systems as "*do-not-care-value*" that can be replaced by any known values of an attribute. Grzymala-Busse [23] divided the incomplete information into two: "*do-not-care-value*" and "*lost-value*". For "*do-not-care* value", it can be replaced by any known values of an attribute which is similar to the concept proposed by Kryszkiewicz [30]. And for the "*lost-value*", it is inaccessible.

In this paper, we categorize the incomplete information systems into two categories: 1) "with values" (WV) and 2) "without values" (OV). The WV must exist; however, we do not know the value or we only know a range of certain values. The WV is slightly different compared to the Grzymala-Busse [23] approach, where a missing value to an attribute may be in a range of certain values. On the other hand, OV is the value that does not exist. For example, an attribute "salary" does not apply to housewife.

From these two categories, the similarity precision is considered to determine both objects are within a certain level of similarity. From these categories, the accuracy of approximation with and without similarity precision are presented. Comparative analysis and experimental result between the proposed approach and other baseline approaches in terms of accuracy of approximation are presented. We found that the proposed approaches with similarity precision are more favourable and better in terms of accuracy of approximation up to two orders of magnitude.

The rest of the paper is organized as follows: Section 2 discusses the theoretical background on information systems, tolerance relation and limited tolerance relation. In Section 3, two types of semantics of incomplete information systems were introduced. Experimental results by using the proposed approach are discussed in Section 4 and finally, Section 5 describes the conclusion of this work.

2.0 THEORETICAL BACKGROUND

This section reviews some basic concepts of information systems and thereafter the tolerance relation and limited tolerance relation

2.1 Information Systems

An information system is a 4-tuple, where $S = (O, AT, V, f)$, where $O = \{o_1, o_2, \cdots, o_{|O|}\}$ denotes a non-empty finite set of objects and $AT = \{a_1, a_2, \cdots, a_{|AT|}\}$ denotes a finite set of attributes/ dimensions, $V = \bigcup_{a \in AT} V_a$,where $V_a$is a value-set of attribute a,$f: O \times AT \rightarrow V$ is a function such that $f(o, a) \in V_a$for every $(o, a) \in O \times AT$, called information function [9]. $S = (O, AT, V, f)$ is called *complete information system (CIS)* if O in S contains all objects with known values, otherwise S is called *incomplete information system (IS)* if at least one object has an

unknown or missing value. In an incomplete information system, the unknown or missing value is denoted as "*". In this paper, the quadruple $IS^* = (O, AT, V \cup \{*\}, f)$ to denote an incomplete information system. From the notion of an information system above, we recall the notion of tolerance relation and limited tolerance relation approach for incomplete information systems in the following sub-section.

2.2 Tolerance Relation and Limited Tolerance Relation

Given a complete information system $S = (O, AT, V, f)$ where $AT = C \cup \{d\}$, *C* is a set of condition attributes and d is the decision attribute, such that $f: O \times AT \to V$, for any $a \in C$, where V_a is called domain of an attribute a. For any subset $B \subseteq C$, the tolerance relation *T* is defined as follows [32]:

Definition 2.1 *Let* $IS = (O, AT, V \cup \{*\}, f)$ *be an IS. A tolerance relation T is defined as*

$$\forall_{x,y \in O} T_B(x, y) \Leftrightarrow \forall_{a_j \in B}\big(a_j(x) = a_j(y) \vee a_j(x) = * \vee a_j(y) = *\big)$$

where known attribute values on attributes *x* and *y* are equal, i.e., $a(x) = a(y)$. Obviously, *T* is reflexive and symmetric, but not transitive. From Definition 2.1, we can describe the notion of tolerance class as follows.

Definition 2.2 *Let* $IS = (O, AT, V \cup \{*\}, f)$ *be an IS. The tolerance class* $I_B^T(x)$*of an object x with reference to an attribute set B is defined as* $I_B^T(x) = \{y | y \in O \wedge T_B(x, y)\}$.

From Definition 2.2, the notion of lower and upper approximations of tolerance class is defined as follows.

Definition 2.3 *Let* $IS = (O, AT, V \cup \{*\}, f)$ *be an IS. The lower approximation* x_B^T*and upper approximation* x_T^B*of an object set X with reference to attribute set B, respectively can be defined as follow* [32]*:*

$$x_B^T = \{x | x \in O \wedge I_B^T(x) \subseteq X\} \text{ and } x_T^B = \{x | x \in O \wedge I_B^T(x) \cap X \neq \varphi\}$$

Definition 2.4. *Let* $IS = (O, A, V \cup \{*\}, f)$ *be an IS, and* $B \subseteq A$*, the lower approximation is* $x_B^T = \{x | x \in O \wedge I_B^{LT}(x) \subseteq X\}$ *and the upper approximation is* $x_T^B = \{x | x \in O \wedge I_B^{LT}(x) \cap X \neq \varphi\}$ *of an object set X. The accuracy of approximation of an object set X with reference to attribute set B can be defined as:*

$$Accuracy = x_B^T / x_T^B \tag{1}$$

We can illustrate the tolerance class and accuracy of approximation with an *IS* for tolerance relation approach through an example below.

Example 1. (See [34]). Table 1 as follow is a description of scholarship-application attributes for a list of students, $S = \{s_i | i = 1, 2, \ldots, 10\}$ who apply for the scholarship. To explain the concepts, we assumed their decision is based on four criteria which are the ability to do analysis (C_1), studying BSc in Mathematics (C_2), the communication skills (C_3) and the ability to speak in Malay language (C_4). The score for C_1 and C_2 be given based on three different levels; 3=good, 2=moderate and 1=poor, while C_3 and C_4 are based on the other three different levels, 3= fluent, 2= moderate and 1= not-fluent.

Table 1: Description of scholarship-application attribute

Attribute Name	Description	Attribute Set Value
Students	Student objects	$\{s_1, s_2, \ldots, s_{10}\}$
C1	Ability to do analysis	{1,2,3}
C2	BSc in Mathematics	{1,2,3}
C3	Communication skills	{1,2,3}
C4	Ability to speak Malay language	{1,2,3}
Decision	Accept or reject the application	{accept, reject}

Table 2: An incomplete information table

Students	**C_1**	**C_2**	**C_3**	**C_4**	**Decision (*d*)**
s_1	3	3	3	*	*accept*
s_2	1	*	3	3	*accept*
s_3	*	*	1	3	*reject*
s_4	3	*	3	3	*accept*
s_5	3	*	3	3	*accept*
s_6	*	3	*	3	*reject*
s_7	1	3	3	*	*accept*
s_8	1	*	3	*	*accept*
s_9	*	3	*	*	*reject*
s_{10}	3	3	3	3	*accept*

Table 2 above is an incomplete information system where some attribute values that are unknown or missing are denoted as "*". The decision (*d*) has two different classes which are *accept* and *reject*. The object that has decision's class of *accept* = $\{s_1, s_2, s_4, s_5, s_7, s_8, s_{10}\}$ and *reject* = $\{s_3, s_6, s_9\}$. We will obtain the results by analyzing Table 2 with the tolerance class from Definition 2.2 as follows.

$I_C^T(s_1) = \{s_1, s_4, s_5, s_6, s_9, s_{10}\}$,$I_C^T(s_2) = \{s_2, s_6, s_7, s_8, s_9\}$, $I_C^T(s_3) = \{s_3, s_6, s_9\}$, $I_C^T(s_4) = \{s_1, s_4, s_5, s_6, s_9, s_{10}\}$, $I_C^T(s_5) = \{s_1, s_4, s_5, s_6, s_9, s_{10}\}$ $I_C^T(s_6) = \{s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8, s_9, s_{10}\}$,$I_C^T(s_7) = \{s_2, s_6, s_7, s_8, s_9\}$,$I_C^T(s_8) = \{s_2, s_6, s_7, s_8, s_9\}$,$I_C^T(s_9) = \{s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8, s_9, s_{10}\}$, and $I_C^T(s_{10}) = \{s_1, s_4, s_5, s_6, s_9, s_{10}\}$

$$\frac{O}{IND(d)} = \{\{s_1, s_2, s_4, s_5, s_7, s_8, s_{10}\}, \{s_3, s_6, s_9\}\}$$

Thus, we have the following values

$$accept_C^T = \varphi, reject_C^T = \{s_3\},$$
$$accept_T^C = \{s_1, s_2, s_4, s_5, s_6, s_7, s_8, s_9, s_{10}\}, reject_T^C = \{s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8, s_9, s_{10}\}.$$

From Eqn. (1), the accuracy can be computed as follow.

$accuracy_{accept} = \frac{|accept_C^T|}{|accept_T^C|} = \frac{0}{9} = 0$ and $accuracy_{reject} = \frac{|reject_C^T|}{|reject_T^C|} = \frac{1}{10} = 0.1$,

with the average accuracy = (0.0+0.10)/2 =0.05.

From the analysis above, the accuracy is very low and some objects that can be discerned intuitively cannot be classified, such as s_{10} has complete information, but s_{10} is not in the lower approximation of $accept_C^T$. The reason is that, the missing attribute values of s_9 is considered similar to s_{10}. Also, some objects are similar with respect to $O_{\{c_1,c_2,c_3,c_4\}}$ but they may not have the same ordinary value for any attribute. For example, from Table 2, two objects $s_2, s_9 \in O$,

$$O(s_2) =< c_1, 1 > \wedge < c_2, * > \wedge < c_3, 3 > \wedge < c_4, 3 >$$
$$O(s_9) =< c_1, * > \wedge < c_2, 3 > \wedge < c_3, * > \wedge < c_4, * >$$

are equivalent with respect to $O_{\{c_1,c_2,c_3,c_4\}}$, however they do not have the same ordinary attribute value. In order to overcome such problems, Wang [32] developed a limited tolerance relation based on the following definition.

Definition 2.5 (See [32]). *Let* $IS = (O, AT, V \cup \{*\}, f)$*be an incomplete information system, a subset* $B \subseteq AT$, *and*$P_B(x) = \{b | b \in B \wedge b(x) \neq *\}$. *A limited tolerance relation defined on O is given as*

$$\forall_{x,y \in OxO}(L_B(x,y) \Leftrightarrow \forall_{b \in B}(b(x) = b(y) = *) \vee \big((P_B(x) \cap P_B(y) \neq \varphi) \wedge \forall_{b \in B}\big((b(x) \neq *) \wedge (b(y) \neq *) \rightarrow \big(b(x) = b(y)\big)\big)\big)$$

Obviously, the limited tolerance relation is symmetric and reflexive but not transitive. In Definition 2.5, the condition that $(b(x) \neq *) \wedge (b(y) \neq *) \rightarrow \big(b(x) = b(y)\big)$ is equivalent to $(b(x) = *) \vee (b(y) = *) \vee \big(b(x) = b(y)\big)$. Thus, two objects that satisfy the tolerance relation but not limited tolerance relation are only those hold$P_B(x) \cap P_B(y) = \varphi$.

In other words, we can consider two objects are in limited tolerance relation if they fulfilled one of these two cases. The first case is that all attribute values for both objects are missing. The second case is there is at least one known attribute value of the two objects that are similar in corresponding to those attribute [31]. From Definition 2.5, the notion of the limited tolerance class can be expressed as follows:

Definition 2.6. *Let* $IS = (O, AT, V \cup \{*\}, f)$*be an incomplete information system and a subset*$B \subseteq AT$. *The limited tolerance class is defined as* $I_B^{LT}(x) = \{y \mid y \in O \wedge LT_B(x,y)\}$.

Table 2 illustrates the limited tolerance class from Definition 2.6 and its accuracy of approximation with an *IS* as follow:

$I_C^{LT}(s_1) = \{s_1, s_4, s_5, s_6, s_9, s_{10}\}$,$I_C^{LT}(s_2) = \{s_2, s_6, s_7, s_8, s_9\}$,$I_C^{LT}(s_3) = \{s_3, s_6, s_9\}$, $I_C^{LT}(s_4)\{s_1, s_4, s_5, s_6, s_9, s_{10}\}$,
$I_C^{LT}(s_5) = \{s_1, s_4, s_5, s_6, s_9, s_{10}\}$, $I_C^{LT}(s_6) = \{s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8, s_9, s_{10}\}$,$I_C^{LT}(s_7) = \{s_2, s_6, s_7, s_8, s_9\}$,
$I_C^{LT}(s_8) = \{s_2, s_6, s_7, s_8, s_9\}$,$I_C^{LT}(s_9) = \{s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8, s_9, s_{10}\}$, and$I_C^{LT}(s_{10}) = \{s_1, s_4, s_5, s_6, s_9, s_{10}\}$.

$$O/IND(d) = \big\{\{s_1, s_2, s_4, s_5, s_7, s_8, s_{10}\}, \{s_3, s_6, s_9\}\big\}.$$

Thus,

$$accept_C^{LT} = \{s_8\},\ reject_C^{LT} = \{s_3\},$$
$$accept_{LT}^{C} = \{s_1, s_2, s_4, s_5, s_6, s_7, s_8, s_9, s_{10}\},\ reject_{LT}^{C} = \{s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_9, s_{10}\}$$

From Eqn. (1), the

$$accuracy_{accept}^{LT} = \frac{|accept_C^{LT}|}{|accept_{LT}^{C}|} = \frac{1}{9} = 0.1111 \text{ and } accuracy_{reject}^{LT} = \frac{|reject_C^{LT}|}{|reject_{LT}^{C}|} = \frac{1}{9} = 0.1111,$$

with the average accuracy = (0.1111+0.1111)/2 =0.1111.

The terms accuracy of the approximation and the accuracy will be used interchangeably. From the analysis above, the limited tolerance relation improves the accuracy compared to tolerance relation approach. However, accuracy is still outstanding and thus need to be improved. Some objects that can be discerned intuitively still cannot be classified. For example, we have complete information about s_{10}nevertheless s_{10} is not in lower approximation of *Accept*. This is because of the missing attribute value s_6 is considered similar to s_{10}.

In the following section, we present the proposed similarity relation based on the possible equivalent value-set and similarity precision between objects.

3.0 POSSIBLE EQUIVALENT VALUE SET AND SIMILARITY PRECISION

3.1 Possible Equivalent Value-set

There have been many efforts in analysing incomplete information systems [22-23, 26-27, 33-34,36-38]. Lipski [33] presents a possible-world semantics to replace missing attribute values with a subset of values within a domain. And based on that semantics, unknown values in incomplete information systems can be represented by possible value-sets [35, 38]. In this paper, we categorize the incomplete information systems that represented possible value-sets into two categories: 1) "with values" (WV) and 2) "without values" (OV). The WV must exist; however, we do not know the value or we only know a range of certain values. On the other hand, OV are the values that do not exist. For example, the attribute "salary" does not apply to housewife. Considering the WV, we summarize the following two semantics.

(Y) "Any value" denoted by "": For $f_a(x) = *$, we can replace * with any value in V_a. For example, if $V_a = \{1,2,3\}$, then * can be interpreted as 1,2 or 3, and * can be only one of them.*

(M) "Maybe value" denoted by "λ: For $f_a(x) = \lambda$, the actual value is in a subset $M_a(x) \subseteq V_a$, where $|M_a(x)| > 1$. That is the actual value of λis one of the values in $M_a(x)$.

The semantic of the OV is as follows:

(N) "Not applicable value" denoted by ∞. For $f_a(x) = \infty$, we know that the value of an object $x \in O$on an attribute $a \in AT$does not exist.

Based on the different semantics of *IS* above, we define the valued set information function that replaces the *IS* with its possible values as follows:

If we denote $P_a(x)$ as the set of possible attribute values of object *x* with respect to an attribute *a*, then

- If $f_a(x) \in V_a$, then $P_a(x) = \{f_a(x)\}$,
- For "any value"; if $f_a(x) = *$, then $P_a(x) = V_a$, (2)
- For "may be value"; if $f_a(x) = \lambda$, then $P_a(x) = M_a \subseteq V_a$,
- For "not applicable value"; $f_a(x) = \infty$, then $P_a(x) = \infty$.

Considering the above semantics, we can replace the *IS* by its possible values. The possible values retain the original incomplete information. Based on the above semantics, we present a definition of incomplete information systems as follows:

Definition 3.1 *An incomplete information systems is a tuple IS = (O, AT, V', F), where O is a finite nonempty set of objects, AT is a finite nonempty set of attributes, $V' = V \cup \{*\} \cup \{\lambda\} \cup \{\infty\}$, where $V = \cup_{a \in AT} V_a$, $V_a$is the domain of attribute a. The *,λand ∞are special symbols of any value, maybe value and not applicable value, respectively. $F = \{f_a | a \in AT\}$, $f_a: O \to V'$ is an information function, such that*

$$f_a(x) \in V_a', x \in O, V_a' = V_a \cup \{*\} \cup \{\lambda\} \cup \{\infty\}.$$

3.2 Similarity Relation

In this sub-section, we review a concept of similarity relation and similarity class in rough set which will be used in our work on incomplete information systems.

Definition 3.2 *Given an incomplete information system, IS = (O, AT, V', F), as in Definition 3.1, and $B \subseteq AT$, a similarity relation with possible equivalent value-set, Sim, is defined as*;

$$\forall x, y \in O, Sim_B(x, y) \Leftrightarrow \forall a \in B, (f_a(x) = f_a(y) \vee P_a(x) \cap P_a(y) \neq \varphi)$$

and the similarity class for the above similarity relation can be defined as follows:

Definition 3.3. *Given an incomplete information system IS = (O,AT,V',F), and similarity relation as in definition 3.2, the similarity class, $I_B^{Sim}(x)$ of an object x, with reference to set B, is defines as;*

$$I_B^{Sim}(x) = \{y | Sim_B(x,y) \in O\}.$$

The computation of the similarity class for incomplete information table with different semantics will be illustrated through example 2 by using information in Table 1.

Table 3**:** An incomplete information table with different semantics

Students	**C_1**	**C_2**	**C_3**	**C_4**	**Decision (*d*)**
s_1	3	3	3	λ_{c4}^{s1}	*accept*
s_2	1	λ_{c2}^{s2}	3	3	*accept*
s_3	λ_{c1}^{s3}	∞	1	3	*reject*
s_4	3	λ_{c2}^{s4}	3	3	*accept*
s_5	3	λ_{c2}^{s5}	3	3	*accept*
s_6	λ_{c1}^{s6}	3	*	3	*reject*
s_7	1	3	3	λ_{c4}^{s7}	*accept*
s_8	1	λ_{c2}^{s8}	3	λ_{c4}^{s8}	*accept*
s_9	∞	3	*	λ_{c4}^{s9}	*reject*
s_{10}	3	3	3	3	*accept*

Example 2. Table 3 illustrates the computation of the similarity class for incomplete information table with different semantics which the domain of each attribute, $V_{c_i} = \{1,2,3\}$,for *i*=1,2,3,4. Based on our discussion in Section 3.1, Table 3 can be equivalently interpreted by a value-set table as shown in Table 4 where we simply use a value to represent a singleton set for a known value. The possible value of $f_a(x) = *$ is {1,2,3}, and let the possible values for$\lambda_{c4}^{s1} = \lambda_{c2}^{s2} = \lambda_{c1}^{s3} = \lambda_{c2}^{s4} = \lambda_{c4}^{s7} = \lambda_{c4}^{s9} = \{1,2\}$, and $\lambda_{c2}^{s5} = \lambda_{c1}^{s6} = \lambda_{c2}^{s8} = \lambda_{c4}^{s8} = \{2,3\}$, respectively.

Table 4**:** Equivalent value-set table from Table 3

Students	**C_1**	**C_2**	**C_3**	**C_4**	**Decision (*d*)**
s_1	3	3	3	1,2	*accept*
s_2	1	1,2	3	3	*accept*
s_3	1,2	∞	1	3	*reject*
s_4	3	1,2	3	3	*accept*
s_5	3	2,3	3	3	*accept*
s_6	2,3	3	1,2,3	3	*reject*
s_7	1	3	3	1,2	*accept*
s_8	1	2,3	3	2,3	*accept*
s_9	∞	3	1,2,3	1,2	*reject*
s_{10}	3	3	3	3	*accept*

The similarity class based on Definition 3.3 be given as;

$I_C^S(s_1) = \{s_1\}$,$I_C^S(s_2) = \{s_2, s_8\}$,$I_C^S(s_3) = \{s_3\}$,$I_C^{Sim}(s_4) = \{s_4, s_5\}$,$I_C^{Sim}(s_5) = \{s_4, s_5, s_{10}\}$,$I_C^{Sim}(s_6) = \{s_5, s_6, s_{10}\}$,$I_C^{Sim}(s_7) = \{s_7, s_8\}$,$I_C^{Sim}(s_8) = \{s_2, s_7, s_8\}$,$I_C^{Sim}(s_9) = \{s_9\}$,$I_C^{Sim}(s_{10}) = \{s_5, s_6, s_{10}\}$, and

$$\frac{O}{IND(d)} = \{\{s_1, s_2, s_4, s_5, s_7, s_8, s_{10}\}, \{s_3, s_6, s_9\}\}.$$

Thus, we have the following values

$$accept_C^{Sim} = \{s_1, s_2, s_4, s_5, s_7, s_8\}, reject_C^{Sim} = \{s_3, s_9\},$$
$$accept_{Sim}^{C} = \{s_1, s_2, s_4, s_5, s_6, s_7, s_8, s_{10}\}, reject_{Sim}^{C} = \{s_3, s_6, s_9, s_{10}\}.$$

The accuracy is computed as follow.

$$accuracy_{accept} = \frac{|accept_C^{Sim}|}{|accept_{Sim}^C|} = \frac{6}{8} = 0.75 \text{ and } accuracy_{reject} = \frac{|reject_C^{Sim}|}{|reject_{Sim}^C|} = \frac{2}{4} = 0.5,$$

with the average accuracy = (0.75+0.50)/2 =0.6250.

From the above analysis, we found that the similarity relation with possible equivalent value-set improves the accuracy compared to the limited tolerance relation approach. However, accuracy needs to be improved.

In the following sub-section, we propose a similarity precision that will be used thereafter for the similarity relation between objects x and $y$on incomplete information systems with possible equivalent value-set of different semantics to improve the accuracy.

3.3 Similarity Precision

Given an incomplete information system *IS* = (*O,AT,V',F*), and similarity relation as in Definition 3.2, the *similarity precision* for each attribute between objects *x* and y with reference to set $B \subseteq AT$, is defines as follows:

Definition 3.4 *Let* $P_B(x) = \{b | b \in B \wedge b(x) \neq \infty\}$, *the similarity precision,*$Sim_{prec}(x, y)$ *between objects x and y on attribute b is defined as*:

$$Sim_{prec}^b(x,y) = \begin{cases} \frac{|P_b(x) \cap P_b(y)|}{|P_b(x) \times P_b(y)|}, & P_b(x) \cap P_b(y) \neq \varphi \\ 0, & P_b(x) \neq P_b(y) \end{cases} \quad (3)$$

where $|\cdot|$ *represent the cardinality of the set and* $\times$ *denotes the Cartesian product.*

Definition 3.5 *Given an incomplete information system IS = (O, AT, V', F) and* $B \subseteq AT$,*the similarity precision of objects on attributes*$b \in B$*is given by a mapping*

$$Sim_{prec}^b: OxO \rightarrow [0,1].$$

This can be illustrated through example below.

Example 3. Two objects s_7 and s_8 from Table 4 are considered. The *similarity precision* between objects s_7 and s_8 for attributes C_2 and C_4 are:

$$Sim_{prec}^{C_2}(s_7, s_8) = \frac{|P_{C2}(s_7) \cap P_{C2}(s_8)|}{|P_{C2}(s_7) x P_{C2}(s_8)|} = \frac{|\{3\} \cap \{2,3\}|}{|\{3\} x \{2,3\}|} = \frac{|\{3\}|}{|\{3,2\},\{3,3\}|} = \frac{1}{2} = 0.5 \text{ and}$$

$$Sim_{prec}^{C_4}(s_7, s_8) = \frac{|P_{C4}(s_7) \cap P_{C4}(s_8)|}{|P_{C4}(s_7) x P_{C4}(s_8)|} = \frac{|\{1,2\} \cap \{2,3\}|}{|\{1,2\} x \{2,3\}|} = \frac{|\{2\}|}{|\{1,2\},\{1,3\},\{2,2\},\{2,3\}|} = \frac{1}{4} = 0.25.$$

From Definition 3.4, Eq. (3) measures the similarity precision of two objects with respect to a single attribute. Therefore, the similarity precision between objects x and y for a set of attributes is defined as:

$$Sim_{prec}(x,y) = \frac{\sum_{b \in B} Sim_{prec}^b(x,y)}{|B|}, B \subseteq AT. \quad (4)$$

Thus, from the above example, the *similarity precision* between objects s_7 and s_8 is;

$$Sim_{prec}(s_7, s_8) = \frac{\sum_{b \in B} Sim_{prec}^b(s_7, s_8)}{|B|} = \frac{1 + 0.5 + 1 + 0.25}{4} = 0.69$$

From Eqn. (4), it is clear that the $0 \leq Sim_{prec}(x, y) \leq 1$. This can be explained by the following proposition.

Proposition 1. The similarity precision has the following properties:

(i) $0 \leq Sim_{prec}(x, y) \leq 1$;
(ii) $Sim_{prec}(x, y) = 1$ when $P_b(x), P_b(y) \in V_b$, $P_b(x) = P_b(y)$ from Eqn. (3),
(iii) $Sim_{prec}(x, y) = Sim_{prec}(y, x)$.

These properties satisfy the Definition 3.5 and Definition of similarity measures as in [17, 18].

From Definition 3.2, we proposed a new similarity relation with possible equivalent value-set and similarity precision as follows;

Definition 3.6. *Given an incomplete information system, IS = (O, AT, V', F), as in Definition 3.1, and* $B \subseteq AT$, *a similarity relation with possible value-set and similarity precision, LSim, is defined as;*

$$\forall x, y \in O, (x, y) \in LSim_B \Leftrightarrow \forall a \in B, Sim_{prec}(x, y) \geq \delta$$

where $\delta \in (0,1]$ *is a threshold value.*

Since $\delta \in (0,1]$, and $0 \leq Sim_{prec}(x, y) \leq 1$which implies that $P_a(x) \cap P_a(y) \notin \varphi$ hold. To clearly depict the similarity relation with possible equivalent value-set and similarity precision as defined above, we illustrate through example as follows;

Example 4. From Table 4, two objects s_7 and s_8 are similar with $\delta \geq 0.65$. However, both objects are not similar if we set $\delta \geq 0.75$, i.e.$(s_7, s_8) \notin LSim(s_7, s_8)$. That is, the two objects are not similar if the value of similarity precision does not hold the threshold value, δ.

In the following sub-section, two propositions for similarity relations with possible equivalent value-set and similarity precision are presented.

Proposition 2. *For* $x, y \in O$,$LSim_{prec}(x, y)$ *is reflexive, symmetric, but not transitive.*

Proof:

- From Eq. (3), we can get $Sim^b_{prec}(x, x) = 1, \forall b \in B \subseteq AT$. Therefore the relation is reflexive.
- For $x, y \in O$,$(x, y) \in LSim_{prec}(x, y)$, we can obtain $Sim^b_{prec}(x, y) \geq \delta$,$\forall b \in B \subseteq AT$. We have

  $Sim^b_{prec}(x, y) = \frac{|P_b(x) \cap P_b(y)|}{|P_b(x) \times P_b(y)|} = \frac{|P_b(y) \cap P_b(x)|}{|P_b(y) \times P_b(x)|} = Sim^b_{prec}(y, x) \geq \delta$

  Thus, the relation is symmetric.
- Suppose that $\delta \geq 0.55$, based on Table 4, we can compute $Sim^b_{prec}(s_4, s_5) \geq 0.81$ and $Sim^b_{prec}(s_5, s_6) \geq 0.58$. Hence, $(s_4, s_5), (s_5, s_6) \in LSim_{prec=.55}(x, y)$. However, $Sim^b_{prec}(s_4, s_6) = 0 \notin LSim_{prec=.55}(x, y)$. Therefore, the relation is not transitive.

Definition 3.7. *Given an incomplete information system, IS = (O, AT, V', F) as in Definition 3.6. The similarity class is defined as;*

$$I_B^{LSim}(x) = \{y | y \in O \wedge Sim_{prec}(x, y)\}$$

To clearly depict the similarity class as defined above, we illustrate through an example below based on equivalent value-set table from Table 4.

Example 5. From Table 4, let $Sim_{prec}(x, y) > 0.75$, we have the similarity class as follows.

$$I_C^{LSim.75}(s_1) = \{s_1\}, I_C^{LSim.75}(s_2) = \{s_2\}, I_C^{LSim.75}(s_3) = \{s_3\}, I_C^{LSim.75}(s_4) = \{s_4, s_5\}, I_C^{LSim.75}(s_5) = \{s_4, s_5, s_{10}\}, I_C^{LSim.75}(s_6) = \{s_6\}, I_C^{LSim.75}(s_7) = \{s_7\}, I_C^{LSim.75}(s_8) = \{s_8\}, I_C^{LSim.75}(s_9) = \{s_9\}, I_C^{LSim.75}(s_{10}) = \{s_5, s_{10}\}, \text{ and}$$

$$\frac{O}{IND(d)} = \{\{s_1, s_2, s_4, s_5, s_7, s_8, s_{10}\}, \{s_3, s_6, s_9\}\}.$$

Thus, we have the following values

$$accept_C^{LSim_{.75}} = \{s_1, s_2, s_4, s_5, s_7, s_8, s_{10}\}, reject_C^{LSim_{.75}} = \{s_3, s_6, s_9\},$$
$$accept_{LSim_{.75}}^C = \{s_1, s_2, s_4, s_5, s_7, s_8, s_{10}\}, reject_{LSim_{.75}}^C = \{s_3, s_6, s_9\}.$$

Therefore, the accuracy can be computed as follow.

$$accuracy_{accept}^{LSim_{.75}} = \frac{|accept_C^{LSim_{.75}}|}{|accept_{LSim_{.75}}^C|} = \frac{7}{7} = 1.0 \text{ and } accuracy_{reject}^{LSim_{.75}} = \frac{|reject_C^{LSim_{.75}}|}{|reject_{LSim_{.75}}^C|} = \frac{3}{3} = 1.0,$$

with the average accuracy = 1.0.

From the above analysis, the result of the proposed approach is more precise and flexible compared to the previous approaches, where objects that can be discerned intuitively can be divided into different classes.

The similarity precision for similarity relation with possible equivalent value-set can be presented using similarity precision matrix to easily compute the accuracy of approximation.

Definition 3.8 *Given an incomplete information system, IS = (O, AT, V', F) and* $B \subseteq AT$. *Suppose that* $X = \{x_1, x_2, \dots, x_n\}$*and the similarity precision matrix is defined by*

$$MSM_{nxm}^B = \begin{pmatrix} Sim_{prec}(x_1, y_1) & Sim_{prec}(x_1, y_2) & \dots & Sim_{prec}(x_1, y_m) \\ Sim_{prec}(x_2, y_1) & Sim_{prec}(x_2, y_2) & \dots & Sim_{prec}(x_2, y_m) \\ \vdots & \square & \square & \square \vdots \square & \square & \square \ddots \square & \square & \square & \vdots \\ Sim_{prec}(x_n, y_1) & Sim_{prec}(x_n, y_2) & \dots & Sim_{prec}(x_n, y_m) \end{pmatrix}$$

From the Table 1, we have

$$MSM_{10x10}^B = \begin{array}{c} \square \\ s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \\ s_7 \\ s_8 \\ s_9 \\ s_{10} \end{array} \begin{array}{c} \begin{array}{cccccccccc} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 & s_8 & s_9 & s_{10} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & .71 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & .81 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & .81 & 1 & .58 & 0 & 0 & 0 & .88 \\ 0 & 0 & 0 & 0 & .58 & 1 & 0 & 0 & 0 & .71 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & .71 & 0 & 0 \\ 0 & .71 & 0 & 0 & 0 & 0 & .71 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & .88 & .71 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

For example, from the above similarity precision matrix, the

$$Sim_{prec}(s_4, s_5) = \frac{\sum_{b \in B} Sim_{prec}^b(s_4, s_5)}{|B|} = \frac{1+0.25+1+1}{4} = 0.81,$$
$$Sim_{prec}(s_5, s_6) = \frac{\sum_{b \in B} Sim_{prec}^b(s_5, s_6)}{|B|} = \frac{0.5 + 0.5 + 0.333 + 1}{4} = 0.58.$$

From the above similarity precision matrix, we can easily obtain similarity classes for different similarity precision. For example, the

$$Sim_{prec}(x, y) \geq 0.55: I_C^{LSim}(s_2) = \{s_2, s_8\}, I_C^{LSim}(s_5) = \{s_4, s_5, s_6, s_{10}\}, I_C^{LSim}(s_{10}) = \{s_5, s_6, s_{10}\},$$
$$Sim_{prec}(x, y) \geq 0.65: I_C^{LSim}(s_2) = \{s_2, s_8\}, I_C^{LSim}(s_5) = \{s_4, s_5, s_{10}\}, I_C^{LSim}(s_{10}) = \{s_5, s_6, s_{10}\},$$
$$Sim_{prec}(x, y) \geq 0.75: I_C^{LSim}(s_2) = \{s_2\}, I_C^{LSim}(s_5) = \{s_4, s_5, s_{10}\}, I_C^{LSim}(s_{10}) = \{s_5, s_{10}\},$$

that will give different values of accuracy. This can be illustrated by the following example when $Sim_{prec}(x, y) \geq 0.65$. From the above matrix, we obtain,

$$I_C^{LSim.65}(s_1) = \{s_1\}, I_C^{LSim.65}(s_2) = \{s_2, s_8\}, I_C^{LSim.65}(s_3) = \{s_3\}, I_C^{LSim.65}(s_4) = \{s_4, s_5\}, I_C^{LSim.65}(s_5) = \{s_4, s_5, s_{10}\}, I_C^{LSim.65}(s_6) = \{s_6, s_{10}\}, I_C^{LSim.65}(s_7) = \{s_7, s_8\}, I_C^{LSim.65}(s_8) = \{s_2, s_7, s_8\}, I_C^{LSim.65}(s_9) = \{s_9\}, I_C^{LSim.65}(s_{10}) = \{s_5, s_6, s_{10}\}, \text{ and}$$

$$\frac{O}{IND(d)} = \{\{s_1, s_2, s_4, s_5, s_7, s_8, s_{10}\}, \{s_3, s_6, s_9\}\}.$$

Thus,

$$accept_C^{LSim.65} = \{s_1, s_2, s_4, s_5, s_7, s_8, s_{10}\},$$
$$reject_C^{LSim.65} = \{s_3, s_9\},$$
$$accept_{LSim.65}^C = \{s_1, s_2, s_4, s_5, s_6, s_7, s_8, s_{10}\},$$
$$reject_{LSim.65}^C = \{s_3, s_6, s_9, s_{10}\}.$$

The

$$accuracy_{accept}^{LSim.65} = \frac{\left|accept_C^{LSim.65}\right|}{\left|accept_{LSim.65}^C\right|} = \frac{7}{8} = 0.875 \text{ and}$$
$$accuracy_{reject}^{LSim.65} = \frac{\left|reject_C^{LSim.65}\right|}{\left|reject_{LSim.65}^C\right|} = \frac{2}{4} = 0.5,$$

with the average accuracy = (0.875+0.5)/2 = 0.6875.

From Definition 3.6, the following proposition can be described.

Proposition 3. Given an incomplete information system, IS = (*O,AT,V',F*), $B \subseteq AT$ and $x, y \in O$.
If $0 \leq Sim_{prec(1)}(x,y) < Sim_{prec(2)}(x,y) \leq 1$, then $I_C^{LSim(2)} \subseteq I_C^{LSim(1)}$.

Proof.
For $\forall c \in I_B^{LS_2}(x)$,we have $\delta_B(x,y) \geq Sim_{prec(2)}(x,y)$. Since $Sim_{prec(2)}(x,y) > Sim_{prec(1)}(x,y)$, then $\delta_B(x,y) \geq Sim_{prec(1)}(x,y)$, that is $\forall c \in I_B^{LS_1}(x)$ which implies$Sim_{prec(1)}(x,y) = Sim_{prec(2)}(x,y)$. On the other hand, if $\delta_B(x,y) \geq Sim_{prec(1)}(x,y)$, then it does not necessarily $\delta_B(x,y) \geq Sim_{prec(2)}(x,y)$. Hence $I_C^{LSim(2)} \subseteq I_C^{LSim(1)}$.

The following example will illustrate the proposition 3 based on Table 4.

Example 6. From Table 4, we have $I_C^{LSim(1)}(s_2) = \{s_2, s_8\}$for $Sim_{prec}(s_2, s_8) > 0.65$. However, for $Sim_{prec}(s_2, s_8) > 0.75$, we have $I_C^{LSim(2)}(s_2) = \{s_2\}$ and thus,$I_C^{LSim(2)}(s_2) \subseteq I_C^{LSim(1)}(s_2)$.

4.0 SIMULATION RESULTS

In this section, the experimental results illustrate the improvement of the accuracy of the proposed approach. In this study, we will compare the proposed method with limited tolerance relation approach (since limited tolerance relation approach is better compared to tolerance relation). Four different datasets were obtained from UCI Machine Learning Repository [36] and a real marine dataset from [37] are considered for simulations. The accuracy of the approximation is calculated from the similarity class matrix by using the algorithm as shown below.

4.1 Algorithms

Algorithm: The accuracy from incomplete information tables with possible equivalent value-set and similarity precision

Input: An incomplete information table *IS*=(*O,AT* U{d},*V* ,*f*), *d* is decision attribute;

For i=1 to |O|; O number of objects;
For j=1 to |B| number of attributes;
For k=1 to |Bj|, Bj value-domain in attribute j;
Insert $g(i,j,k)$: g is attribute-value
If $g(i,j,k) \neq \infty$ and $g(i,j,k) \neq 0$then
$f(i,j) = \cup \{g(i,j,k\}$
Insert equivalent class induced by decision attribute (d): $O/IND(d)$
Insert similarity precision threshold value $0 < \delta \leq 1$;
Output: Accuracy of approximation
Begin
For i=1 to |O|; //O number of objects;
For j=1 to |B|;// number of attributes;
Get $Sim^{b}_{prec}(x_i, x_j) = \frac{|P_b(x_i) \cap P_b(x_j)|}{|P_b(x_i) \times P_b(x_j)|}$; //where $P_b(x_i) = f(i,b)$,
// similarity precision between objects $x_i, x_j \in O$ for attribute $b \in B$
Get $Sim_{prec}(x_i, x_j) = \frac{Sim^{b}_{prec}(x_i, x_j)}{|B|}$
Get similarity class matrix MSM^{B}_{OxO} as in Definition 3.8
From similarity class matrix, get similarity classes based on similarity precision
Get equivalent class of lower approximation and upper approximation for each $IND(d)$
Get the accuracy and, subsequently, get the average accuracy for each $IND(d)$
end

The following are several datasets for our simulations based on the above algorithm.

4.2 Four UCI Dataset

The description of the datasets is presented in Table 5. The Soybean and Mammographic dataset is incomplete dataset while the Monk, Tic-tac-toe and Car are complete datasets. About 10% of the known attribute value was randomly removed from the complete datasets to create incomplete datasets. Then, the missing attribute value of the dataset will be replaced by possible equivalent value-set. Examples of the incomplete datasets for Soybean is as in Table 6. The same procedure is also applied to the other incomplete datasets. Originally, the five datasets contain many objects. However, only several objects were considered as shown in Table 5 for simulation purposes.

Table 5: Descriptions of datasets

Dataset	Number of attributes	Number of objects	Number of classes
Soybean	7	20	2
Monk	6	25	2
Tic-tac-toe	5	30	2
Car	6	40	4
Mammographic	5	40	2

4.3 Implementation of technique on soybean dataset

Table 6 is consists of incomplete table of 20 different soybeans. Let soybean = {1,2,…,20} be the objects, a set of attributes, C={stem,seed-size,shriveling, roots,mycelium}, with stem={abnorm, norm}, seed-size={norm, it-norm}, shriveling={present, absent}, roots={normal,galls-cysts,rotted}, mycelium={present,absent}, and the decision(d) ={present,absent}.

Table 6: Incomplete table of 20 different soybeans

Soybean	Stem	Seed-size	Shrivelling	Roots	Mycelium	Decision
1	Abnorm	Norm	Absent	Normal	Absent	Present
2	Norm	Norm	Absent	Normal	Absent	Present
3	Abnorm	Norm	*	Normal	Absent	Present
4	Abnorm	Norm	Absent	Galls-cysts	Present	Present
5	*	It-Norm	Absent	Normal	Absent	Present
6	Norm	Norm	*	Normal	*	Present
7	Abnorm	Norm	Absent	Normal	Absent	Present
8	Abnorm	*	Absent	Galls-cysts	Absent	Present
9	*	Norm	Absent	Normal	Absent	Present
10	Abnorm	Norm	Absent	Normal	Absent	Present
11	Abnorm	It-Norm	*	Galls-cysts	*	Present
12	Norm	Norm	*	Galls-cysts	*	Present
13	Norm	It-Norm	*	Normal	Absent	Present
14	Norm	It-Norm	*	Galls-cysts	*	Absent
15	Abnorm	It-Norm	*	Galls-cysts	*	Absent
16	Norm	It-Norm	*	Galls-cysts	Present	Absent
17	Norm	*	*	*	*	Absent
18	*	*	*	Rotted	*	Present
19	Abnorm	Norm	*	Rotted	Present	Absent
20	Abnorm	*	*	Rotted	*	Present

From the table, the limited tolerance class is obtained as:

$$I_C^{LT}(1) = \{1,3,7,9,10\}, I_C^{LT}(2) = \{2,6,9,17\}, I_C^{LT}(3) = \{1,3\},\ I_C^{LT}(4) = \{4\}, I_C^{LT}(5) = \{5,13\}, I_C^{LT}(6) = \{2,6,9,17\}, I_C^{LT}(7) = \{1,7,9,10\}, I_C^{LT}(8) = \{8,11\}, I_C^{LT}(9) = \{1,2,6,7,9,10\}, I_C^{LT}(10) = \{1,7,9,10\}, I_C^{LT}(11) = \{11,15\}, I_C^{LT}(12) = \{12,17\}, I_C^{LT}(13) = \{5,13,17\},\ I_C^{LT}(14) = \{14,16,17\}, I_C^{LT}(15) = \{11,15\}, I_C^{LT}(16) = \{14,16,17\},\ I_C^{LT}(17) = \{2,6,12,13,14,16,17\}, I_C^{LT}(18) = \{18,19,20\}, I_C^{LT}(19) = \{18,19,20\} \text{and } I_C^{LT}(20) = \{18,19,20\}$$

$$O/IND(d) = \{\{1,2,3,4,5,6,7,8,9,10,11,12,13,18,20\}, \{14,15,16,17,19\}\}$$

Thus,

$$present_C^{LT} = \{1,3,4,5,7,8,9,10\},$$
$$absent_C^{LT} = \{14,16\},$$
$$present_{LT}^{C} = \{1,2,3,4,5,6,7,8,9,10,11,12,13,15,17,18,19,20\},$$
$$absent_{LT}^{C} = \{11,12,13,14,15,16,17,18,19,20\}.$$

Therefore, the

$$accuracy_{present} = \frac{|present_C^{LT}|}{|present_{LT}^{C}|} = \frac{8}{18} = 0.4444 \text{ , and}$$
$$accuracy_{absent} = \frac{|absent_C^{LT}|}{|absent_{LT}^{C}|} = \frac{2}{10} = 0.20,$$

with the average accuracy = (0.4444+0.20)/2 =0.3131.

The incomplete information system from Table 6 can be replaced with equivalent value-set as given in Table 7

Table 7: Equivalent value set replacement

Soybean	Stem	Seed-size	Shriveling	Roots	Mycelium	Decision
1	Abnorm	Norm	Absent	Normal	Absent	Present
2	Norm	Norm	Absent	Normal	Absent	Present
3	Abnorm	Norm	Absent,Present	Normal	Absent	Present
4	Abnorm	Norm	Absent	Galls-cysts	Present	Present
5	Abnorm, Norm	It-Norm	Absent	Normal	Absent	Present
6	Norm	Norm	Absent,Present	Normal	Absent,Present	Present
7	Abnorm	Norm	Absent	Normal	Absent	Present
8	Abnorm	Norm, It-Norm	Absent	Galls-cysts	Absent	Present
9	Abnorm, Norm	Norm	Absent	Normal	Absent	Present
10	Abnorm	Norm	Absent	Normal	Absent	Present
11	Abnorm	It-Norm	Absent,Present	Galls-cysts	Absent,Present	Present
12	Norm	Norm	Absent,Present	Galls-cysts	Absent,Present	Present
13	Norm	It-Norm	Absent,Present	Normal	Absent	Present
14	Norm	It-Norm	Absent,Present	Galls-cysts	Absent,Present	Absent
15	Abnorm	It-Norm	Absent,Present	Galls-cysts	Absent,Present	Absent
16	Norm	It-Norm	Absent,Present	Galls-cysts	Present	Absent
17	Norm	∞	Absent,Present	Normal, Galls-cysts	Absent,Present	Absent
18	Abnorm, Norm	Norm, It-Norm	Absent,Present	Rotted	Absent,Present	Present
19	Abnorm	Norm	Absent,Present	Rotted	Present	Absent
20	Abnorm	Norm, It-Norm	Absent,Present	Rotted	Absent,Present	Present

The similarity class from Table 7 is obtained as follows:

$$I_C^{Sim}(1) = \{1,3,7,9,10\}, I_C^{Sim}(2) = \{2,6,9\}, I_C^{Sim}(3) = \{1,3,7,9,10\}, I_C^{Sim}(4) = \{4\}, I_C^{Sim}(5) = \{5,13\} I_C^{Sim}(6) = \{2,6,9\}, I_C^{SIM}(7) = \{1,3,7,9,10\}, I_C^{Sim}(8) = \{8,11,15\}, I_C^{Sim}(9) = \{1,2,3,6,7,9,10\}, I_C^{Sim}(10) = \{1,3,7,9,10\}, I_C^{Sim}(11) = \{8,11,15\}, I_C^{Sim}(12) = \{12\}, I_C^{Sim}(13) = \{5,13\}, I_C^{Sim}(14) = \{14,16\}, I_C^{Sim}(15) = \{8,11,15\}, I_C^{Sim}(16) = \{14,16\}, I_C^{Sim}(17) = \{17\}, I_C^{Sim}(18) = \{18,19,20\}, I_C^{Sim}(19) = \{18,19,20\}$$ and $$I_C^{Sim}(20) = \{18,19,20\}$$

$$O/IND(d) = \{\{1,2,3,4,5,6,7,8,9,10,11,12,13,18,20\}, \{14,15,16,17,19\}\}$$

Thus,

$$present_C^{Sim} = \{1,2,3,4,5,6,7,8,9,10,12,13\},$$
$$absent_C^{Sim} = \{14,16,17\},$$
$$present_{Sim}^{C} = \{1,2,3,4,5,6,7,8,9,10,11,12,13,15,18,19,20\},$$
$$absent_{Sim}^{C} = \{11,14,15,16,17,18,19,20\}.$$

Therefore, the

$$accuracy_{present} = \frac{|present_C^{Sim}|}{|present_{Sim}^{C}|} = \frac{14}{17} = 0.823, \text{ and}$$
$$accuracy_{absent} = \frac{|absent_C^{Sim}|}{|absent_{Sim}^{C}|} = \frac{3}{8} = 0.375,$$

with the average accuracy=(0.823+0.375)/2=0.599

Let s_1, s_2,…, s_{20}, be the Soybean objects from the above Soybean data set. The similarity precision matrix from Table 7 can be presented as:

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}	s_{11}	s_{12}	s_{13}	s_{14}	s_{15}	s_{16}	s_{17}	s_{18}	s_{19}	s_{20}
s_1	1	0	.9	0	0	0	1	0	.9	1	0	0	0	0	0	0	0	0	0	0
s_2	0	1	0	0	0	.8	0	0	.9	0	0	0	0	0	0	0	0	0	0	0
s_3	.9	0	1	0	0	0	.9	0	.8	.9	0	0	0	0	0	0	0	0	0	0
s_4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s_5	0	0	0	0	1	0	0	0	0	0	0	0	.8	0	0	0	0	0	0	0
s_6	0	.8	0	0	0	1	0	0	.7	0	0	0	0	0	0	0	0	0	0	0
s_7	1	0	.9	0	0	0	1	0	.9	1	0	0	0	0	0	0	0	0	0	0
s_8	0	0	0	0	0	0	0	1	0	0	.7	0	0	0	.7	0	0	0	0	0
s_9	.9	.9	.8	0	0	.7	.9	0	1	.9	1	0	0	0	0	0	0	0	0	0
s_{10}	1	0	.9	0	0	0	1	0	.9	1	0	0	0	0	0	0	0	0	0	0
s_{11}	0	0	0	0	0	0	0	.7	0	0	1	0	0	0	.8	0	0	0	0	0
s_{12}	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
s_{13}	0	0	0	0	.8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
s_{14}	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	.8	0	0	0	0
s_{15}	0	0	0	0	0	0	0	.7	0	0	.8	0	0	0	1	0	0	0	0	0
s_{16}	0	0	0	0	0	0	0	0	0	0	0	0	0	.8	0	1	0	0	0	0
s_{17}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
s_{18}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	.6	.6
s_{19}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.6	1	.7

By considering the similarity precision, say $Sim_{prec}(x, y) \geq 0.75$, the similarity class for the above similarity precision matrix is given as;

$I_C^{LSim}(1) = \{1,3,7,9,10\}$,$I_C^{LSim}(2) = \{2,6,9\}$,$I_C^{LSim}(3) = \{1,3,7,9,10\}$, $I_C^{Sim}(4) = \{4\}$,$I_C^{Sim}(5) = \{5,13\}$,
$I_C^{LSim}(6) = \{2,6\}$,$I_C^{LSim}(7) = \{1,3,7,9,10\}$,$I_C^{LSim}(8) = \{8\}$,$I_C^{LSim}(9) = \{1,2,3,7,9,10\}$,$I_C^{LSim}(10) = \{1,3,7,9,10\}$,
$I_C^{LSim}(11) = \{11,15\}$,$I_C^{LSim}(12) = \{12\}$,$I_C^{LSim}(13) = \{5,13\}$, $I_{C}^{LSim}(14) = \{14,16\}$ $I_C^{LSim}(15) = \{11,15\}$,
$I_C^{LSim}(16) = \{14,16\}$,$I_C^{LSim}(17) = \{17\}$,$I_C^{LSim}(18) = \{18\}$,$I_C^{LSim}(19) = \{19\}$and $I_C^{LSim}(20) = \{20\}$

$$O/IND(d) = \{\{1,2,3,4,5,6,7,8,9,10,11,12,13,18,20\},\{14,15,16,17,19\}\}$$

Thus,

$present_C^{LSim} = \{1,2,3,4,5,6,7,8,9,10,12,13,18,20\}$, $absent_C^{LSim} = \{14,16,17,19\}$,
$present_{LSim}^{C} = \{1,2,3,4,5,6,7,8,9,10,11,12,13,15,18,20\}$, $absent_{LSim}^{C} = \{11,14,15,16,17,19\}$.

Therefore, the

$$accuracy_{present} = \frac{|present_C^{LSim}|}{|present_{LSim}^{C}|} = \frac{14}{16} = 0.875, \text{ and}$$
$$accuracy_{absent} = \frac{|absent_C^{LSim}|}{|absent_{LSim}^{C}|} = \frac{4}{6} = 0.667,$$

with the average accuracy = (0.875+0.667)/2 =0.771.

4.4 Comparative Analysis from Four Datasets

In this section, we compare the proposed similarity relation with possible equivalent value-set with limited tolerance relation approach based on accuracy. The accuracy of each approach is calculated based on roughness accuracy as in Definition 2.4, i.e. lower approximation over upper approximation. Hence, it is adopted from the standard rough set accuracy of the approximation. The results of the comparison between the two approaches are presented in Table 8.

Table 8: The accuracy of each approach

Dataset	Limited tolerance relation (LTR)	Similarity relation with possible equivalent value-set (SR-PEVS)	Improvement
Soybean	0.3131	0.5993	91.41%
Monk	0.5458	0.8607	57.70%
Tic-tac-toe	0.4885	0.7122	45.79%
Car	0.3657	0.5570	52.17%
Mammographic	0.1576	0.2368	50.25%

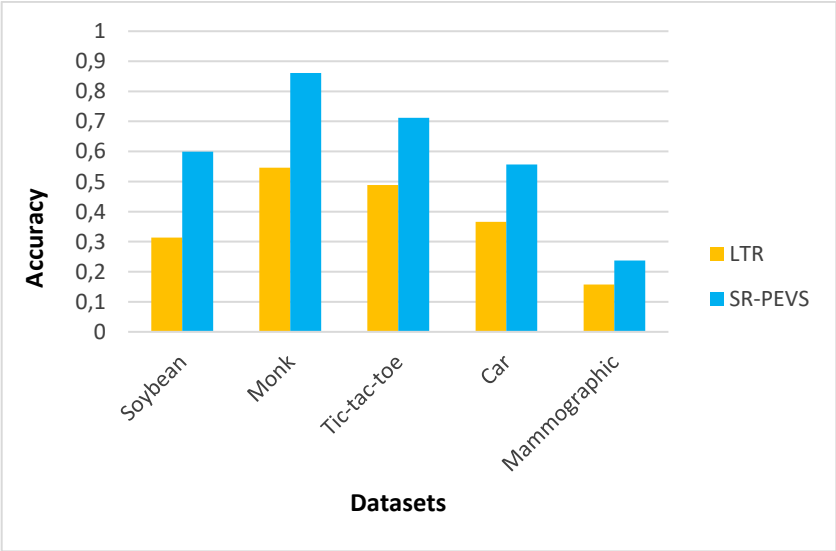

Fig. 1: Comparison of accuracy of approximation between limited tolerance relation and similarity relation with possible equivalent value-set.

We further compare similarity relation with possible equivalent value-set with and without similarity precision. The results as shown in Table 9.

Table 9**:** The accuracy of each approach

Dataset	Similarity relation with possible equivalent value-set (SR-PEVS)	Similarity relation with possible equivalent value-set and with similarity precision (≥75%) (SR-PEVS-SP)	Improvement
Soybean	0.5993	0.7709	28.63%
Monk	0.8607	1.0000	16.18%
Tic-tac-toe	0.7122	0.9354	31.34%
Car	0.5570	0.7325	31.51%
Mammographic	0.2368	0.2762	16.64%

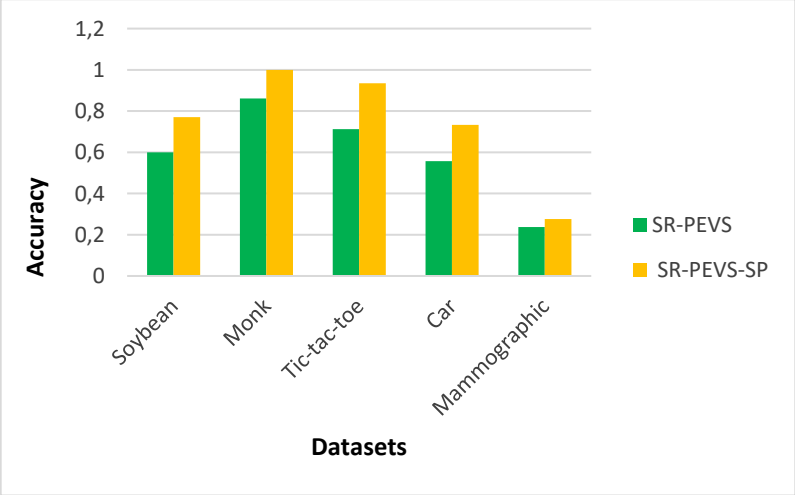

Fig. 2: Comparison of accuracy of approximation between similarity relation with possible equivalent value-set with and without similarity precision

From Table 8 and Fig.1, the results show that the proposed approach achieved better accuracy as compared to the limited tolerance relation approach. Consequently, the proposed approach with similarity precision (≥75%) significantly improves the accuracy of approximation as shown in Table 7 and Fig. 2. For example, the accuracy for Soybean dataset using limited tolerance relation is 0.3131 while by using similarity relation with possible equivalent value-set is 0.5990 which improves more than 91%. Subsequently, the accuracy for Soybean dataset using similarity relation with possible equivalent value-set and similarity precision (≥75%) is 0.7709 which improves more than 146% as compared to the limited tolerance relation approach. The accuracy for Monk dataset using limited tolerance relation is 0.5458 while by using similarity relation with possible equivalent value-set is 0.8607 which improves more than 57%. Subsequently, the accuracy for Monk dataset using similarity relation with possible equivalent value-set and similarity precision (≥75%) is 1.000 which improves more than 83.22% as compared to the limited tolerance relation approach.

4.5 Real Marine Dataset

The Marine dataset was obtained from [37]. It contained 544 objects and 7 attributes. The attributes are Region ID, Number of Sea Fish (SF), Number of Sea Aquaculture (SA), Number of Coral Reefs (CR), Mangrove Area (MA), Coastal Forest Area (CF), Plankton Abundance (PA) and Potential Fish Yield (PF). A sample of 30 objects from 544 objects will be considered in this study as shown in Table 9. Table 8 below shows the description of each attribute in the dataset.

Table 9**:** Description of Marine dataset

Attribute name	**Description**	**Attribute set value**
ID	Region ID	{1,2,3,…,544}
SF	Number of Sea Fish (tons)	{1,2,3,4,5}
SA	Number of Sea Aquaculture (tons)	{1,2,3,4,5}
CR	Number of Coral Reefs (km^2)	{1,2,3,4,5}
MA	Mangrove Area (km^2)	{1,2,3,4,5}
CF	Coastal Forest Area (km^2)	{1,2,3,4,5}
PA	Plankton Abundance (cell/L)	{1,2,3,4,5}
PF	Potential Fish Yield	{1,2,3}

Table 10 30 objects of Marine dataset with 7 attributes

ID	SF	SA	CR	MA	CF	PA	PF (Dec)		ID	SF	SA	CR	MA	CF	PA	PF (Dec)
1	2	1	1	2	2	1	1		16	2	3	1	2	1	3	2
2	*	*	1	3	1	*	1		17	3	1	1	1	2	2	2
3	2	1	2	1	1	1	1		18	2	1	2	1	3	1	2
4	1	1	2	3	1	1	1		19	3	1	1	2	2	1	2
5	1	1	2	1	1	1	1		20	1	3	2	1	1	2	2
6	1	1	3	1	1	1	1		21	2	2	*	1	2	1	2
7	1	1	1	1	1	1	1		22	3	1	2	4	1	*	2
8	1	1	1	3	1	1	1		23	2	1	1	4	1	1	2
9	1	1	2	1	1	1	1		24	4	*	3	*	*	1	2
10	2	1	2	1	1	*	1		25	4	1	3	1	1	1	2
11	1	1	2	*	1	1	1		26	3	1	1	1	2	2	2
12	*	*	1	1	2	*	1		27	3	1	1	2	4	1	3
13	1	1	3	1	1	1	1		28	3	1	1	2	4	1	3
14	2	*	1	2	1	3	1		29	4	2	3	3	4	1	3
15	2	2	*	1	*	*	1		30	4	2	3	3	4	1	3

4.5.1 Comparative analysis from Marine dataset

The Marine dataset was obtained from Saedudin et al. (2018). It contained 544 objects and 7 attributes. The attributes are Region ID, Number of Sea Fish (SF), Number of Sea Aquaculture (SA), Number of Coral Reefs (CR), Mangrove Area (MA), Coastal Forest Area (CF), Plankton Abundance (PA) and Potential Fish Yield (PF). A sample of 30 objects from 544 objects will be considered in this study as shown in Table 9. Table 8 below shows the description of each attribute in the dataset.

Table 9**:** Description of Marine dataset

Attribute name	**Description**	**Attribute set value**
ID	Region ID	{1,2,3,…,544}
SF	Number of Sea Fish (tons)	{1,2,3,4,5}
SA	Number of Sea Aquaculture (tons)	{1,2,3,4,5}
CR	Number of Coral Reefs (km^2)	{1,2,3,4,5}
MA	Mangrove Area (km^2)	{1,2,3,4,5}
CF	Coastal Forest Area (km^2)	{1,2,3,4,5}
PA	Plankton Abundance (cell/L)	{1,2,3,4,5}
PF	Potential Fish Yield	{1,2,3}

Table 10**:** 30 objects of Marine dataset with 7 attributes

ID	SF	SA	CR	MA	CF	PA	PF (Dec)		ID	SF	SA	CR	MA	CF	PA	PF (Dec)
1	2	1	1	2	2	1	1		16	2	3	1	2	1	3	2
2	*	*	1	3	1	*	1		17	3	1	1	1	2	2	2
3	2	1	2	1	1	1	1		18	2	1	2	1	3	1	2
4	1	1	2	3	1	1	1		19	3	1	1	2	2	1	2
5	1	1	2	1	1	1	1		20	1	3	2	1	1	2	2
6	1	1	3	1	1	1	1		21	2	2	*	1	2	1	2
7	1	1	1	1	1	1	1		22	3	1	2	4	1	*	2
8	1	1	1	3	1	1	1		23	2	1	1	4	1	1	2
9	1	1	2	1	1	1	1		24	4	*	3	*	*	1	2
10	2	1	2	1	1	*	1		25	4	1	3	1	1	1	2
11	1	1	2	*	1	1	1		26	3	1	1	1	2	2	2
12	*	*	1	1	2	*	1		27	3	1	1	2	4	1	3
13	1	1	3	1	1	1	1		28	3	1	1	2	4	1	3

14	2	*	1	2	1	3	1		29	4	2	3	3	4	1	3
15	2	2	*	1	*	*	1		30	4	2	3	3	4	1	3

Let Table 10 be transformed into its equivalent value-set as shown in Table 11. We will obtain the similarity class with possible equivalent value-set and its accuracy as

$$I_C^{sim}(1) = \{1\}, I_C^{sim}(2) = \{2,8\}, I_C^{sim}(3) = \{3,10\}, I_C^{sim}(4) = \{4,11\}, I_C^{sim}(5) = \{5,9\}, I_C^{sim}(6) = \{6,13\}, I_C^{sim}(7) = \{7\}, I_C^{sim}(8) = \{2,8\}, I_C^{sim}(9) = \{5,9\}, I_C^{sim}(10) = \{3,10\}, I_C^{sim}(11) = \{4,11\}, I_C^{sim}(12) = \{12,15\}, I_C^{sim}(13) = \{6,13\}, I_C^{sim}(14) = \{14\}, I_C^{sim}(15) = \{12,15,21\}, I_C^{sim}(16) = \{16\}, I_C^{sim}(17) = \{17,26\}, I_C^{sim}(18) = \{18\}, I_C^{sim}(19) = \{19\}, I_C^{sim}(20) = \{20\}, I_C^{sim}(21) = \{15,21\}, I_C^{sim}(22) = \{22\}, I_C^{sim}(23) = \{23\}, I_C^{sim}(24) = \{24,29,30\}, I_C^{sim}(25) = \{25\}, I_C^{sim}(26) = \{17,26\}, I_C^{sim}(27) = \{27,28\}, I_C^{sim}(28) = \{27,28\}, I_C^{sim}(29) = \{24,29,30\}$$ and $$I_C^{sim}(30) = \{24,29,30\}$$

$$O/IND(Dec) = \{\{1,2,3,\dots,15\}, \{16,17,18,\dots,26\}, \{27,28,29,30\}\}$$

where

$$Dec1_C^{sim} = \{1,2,3,4,5,6,7,8,9,10,11,12,13,14\},$$
$$Dec1_{sim}^C = \{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,21\},$$
$$Dec2_C^{sim} = \{16,17,18,19,20,22,23,25,26\},$$
$$Dec2_{sim}^C = \{15,16,17,18,19,20,21,22,23,24,25,26,29,30\},$$
$$Dec3_C^{sim} = \{27,28\}, \text{ and}$$
$$Dec3_{sim}^C = \{24,27,28,29,30\}$$

From Definition 2.6, the,

$$accuracy_{Dec1} = \frac{\left|Dec1_C^{sim}\right|}{\left|Dec1_{sim}^C\right|} = \frac{14}{16} = 0.8750,$$
$$accuracy_{Dec2} = \frac{\left|Dec2_C^{sim}\right|}{\left|Dec2_{sim}^C\right|} = \frac{9}{14} = 0.6429, \text{ and}$$
$$accuracy_{Dec3} = \frac{\left|Dec3_C^{sim}\right|}{\left|Dec3_{sim}^C\right|} = \frac{2}{5} = 0.4000,$$

with the average accuracy = (0.8750+0.6429+0.4000)/3 =0.6393.

The similarity precision matrix for marine data set cannot be presented in this paper due to large number of objects that is difficult to portray in the paper. However, the similarity classes for different similarity precision are given as follows;

By considering the similarity precision, say $Sim_{prec}(x,y) \geq 0.65$, the similarity class with similarity precision for Table 10 is calculated as;

$$I_C^{LSim_{\cdot 65}}(1) = \{1\}, I_C^{LSim_{65}}(2) = \{2,8\}, I_C^{LSim_{65}}(3) = \{3,10\}, I_C^{LSim_{65}}(4) = \{4,11\}, I_C^{LSim_{65}}(5) = \{5,9\}, I_C^{LSim_{65}}(6) = \{6,13\}, I_C^{LSim_{65}}(7) = \{7\}, I_C^{LSim_{65}}(8) = \{2,8\}, I_C^{LSim_{65}}(9) = \{5,9\}, I_C^{LSim_{65}}(10) = \{3,10\}, I_C^{LSim_{65}}(11) = \{4,11\}, I_C^{LSim_{65}}(12) = \{12\}, I_C^{LSim_{65}}(13) = \{6,13\}, I_C^{LSim_{65}}(14) = \{14\}, I_C^{LSim_{65}}(15) = \{15,21\}, I_C^{LSim_{65}}(16) = \{16\}, I_C^{LSim_{65}}(17) = \{17,26\}, I_C^{LSim_{65}}(18) = \{18\}, I_C^{LSim_{65}}(19) = \{19\}, I_C^{LSim_{65}}(20) = \{20\}, I_C^{LSim_{65}}(21) = \{15,21\}, I_C^{LSim_{65}}(22) = \{22\}, I_C^{LSim_{65}}(23) = \{23\}, I_C^{LSim_{65}}(24) = \{24,29,30\}, I_C^{LSim_{65}}(25) = \{25\}, I_C^{LSim_{65}}(26) = \{17,26\},$$

$$I_C^{LSim_{65}}(27) = \{27,28\}, I_C^{LSim_{65}}(28) = \{27,28\}, I_C^{LSim_{65}}(29) = \{24,29,30\}$$ and $$I_C^{LSim_{65}}(30) = \{24,29,30\}$$

$$\frac{O}{IND(Dec)} = \{\{1,2,3,\dots,15\}, \{16,17,18,\dots,26\}, \{27,28,29,30\}\}$$

Thus,

$$Dec1_C^{LSim_{65}} = \{1,2,3,4,5,6,7,8,9,10,11,12,13,14\},$$
$$Dec1_{LSim_{65}}^C = \{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,21\},$$

$$Dec2_C^{LSim_{65}} = \{16,17,18,19,20,21,22,23,24,25,26\},$$
$$Dec2_{LSim_{65}}^{C} = \{15,16,17,18,19,20,21,22,23,24,25,26,29,30\},$$
$$Dec3_C^{LSim_{65}} = \{27,28\}, \text{ and}$$
$$Dec3_{LSim_{65}}^{C} = \{27,28,29,30\}.$$

From Definition 2.6, the

$$accuracy_{Dec1_{65}} = \frac{\left|Dec1_C^{LSim_{65}}\right|}{\left|Dec1_{LSim_{65}}^{C}\right|} = \frac{14}{16} = 0.8750,$$
$$accuracy_{Dec2_{65}} = \frac{\left|Dec2_C^{LSim_{65}}\right|}{\left|Dec2_{LSim_{65}}^{C}\right|} = \frac{8}{13} = 0.6429, \text{ and}$$
$$accuracy_{Dec3_{65}} = \frac{\left|Dec3_C^{LSim_{65}}\right|}{\left|Dec3_{LSim_{65}}^{C}\right|} = \frac{2}{4} = 0.5,$$

with the average accuracy = (0.8750+0.6429+0.5)/3 =0. 6726.

For $Sim_{prec}(x, y) \geq 0.75$, from Table 10, we will obtain the similarity class with similarity precision as;

$I_C^{LSim_{.75}}(1) = \{1\}$,$I_C^{LSim_{.75}}(2) = \{2\}$,$I_C^{LSim}(3) = \{3,10\}$, $I_C^{LSim}(4) = \{4,11\}$, $I_C^{LSim}(5) = \{5,9\}$,$I_C^{LSim}(6) = \{6,13\}$, $I_C^{LSim}(7) = \{7\}$, $I_C^{LSim}(8) = \{8\}$,$I_C^{LSim}(9) = \{5,9\}$,$I_C^{LSim}(10) = \{3,10\}$, $I_C^{LSim}(11) = \{4,11\}$, $I_C^{LSim}(12) = \{12\}$,$I_C^{LSim}(13) = \{6,13\}$,$I_C^{LSim}(14) = \{14\}$,$I_C^{LSim}(15) = \{15\}$,$I_C^{LSim}(16) = \{16\}$,$I_C^{LSim}(17) = \{17,26\}$,$I_C^{Lsim}(18) = \{18\}$,$I_C^{LSim}(19) = \{19\}$,$I_C^{LSim}(20) = \{20\}$, $I_C^{LSim}(21) = \{21\}$, $I_C^{LSim}(22) = \{22\}$,$I_C^{LSim}(23) = \{23\}$, $I_C^{LSim}(24) = \{24\}$,$I_C^{LSim}(25) = \{25\}$,$I_C^{LSim}(26) = \{17,26\}$, $I_C^{LSim}(27) = \{27,28\}$, $I_C^{LSim}(28) = \{27,28\}$,$I_C^{LSim}(29) = \{29,30\}$and $I_C^{LSim}(30) = \{29,30\}$

where

$$\frac{O}{IND(Dec)} = \{\{1,2,3,\ldots,15\},\{16,17,18,\ldots,26\},\{27,28,29,30\}\}$$

Thus,

$$Dec1_C^{sim} = \{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15\},$$
$$Dec1_{sim}^{C} = \{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15\},$$
$$Dec2_C^{sim} = \{16,17,18,19,20,21,22,23,24,25,26\},$$
$$Dec2_{sim}^{C} = \{16,17,18,19,20,21,22,23,24,25,26\},$$
$$Dec3_C^{sim} = \{27,28,29,30\},\ Dec3_{sim}^{C} = \{27,28,29,30\}.$$

From Definition 2.6, the

$$accuracy_{Dec1} = \frac{\left|Dec1_C^{sim}\right|}{\left|Dec1_{sim}^{C}\right|} = \frac{15}{15} = 1.0000,$$
$$accuracy_{Dec2} = \frac{\left|Dec2_C^{sim}\right|}{\left|Dec2_{sim}^{C}\right|} = \frac{11}{11} = 1.0000, \text{ and}$$

$$accuracy_{Dec3} = \frac{\left|Dec3_C^{sim}\right|}{\left|Dec3_{sim}^{C}\right|} = \frac{4}{4} = 1.0000,$$

with the average accuracy = (1.0000+1.0000+1.0000)/3 = 1.0000.

Table 12**:** The accuracy of each approach

Dataset	Limited tolerance relation (LTR)	Similarity relation with possible equivalent value-set and (SR-PEVS)	Improvement
Marine	0.5230	0.6393	22.24%

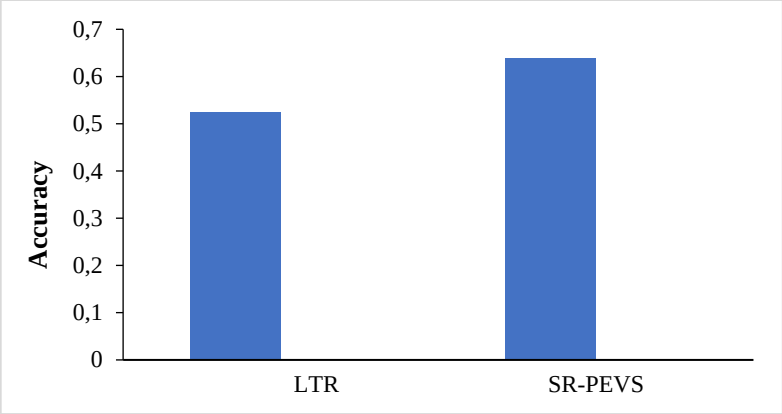

Fig. 3: Comparison of accuracy using different approaches for Marine dataset

Next, we compare similarity relation with possible equivalent value-set with different values of similarity precision. The results are shown in Table 12.

Table 13**:** The accuracy of the proposed approach with different values of similarity precision

Dataset	Similarity relation with possible equivalent value-set and with similarity precision (≥65%) (SR-PEVS-SP)	Similarity relation with possible equivalent value-set and with similarity precision (≥75%) (SR-PEVS-SP)	Improvement
Marine	0.6726	1.0000	48.68%

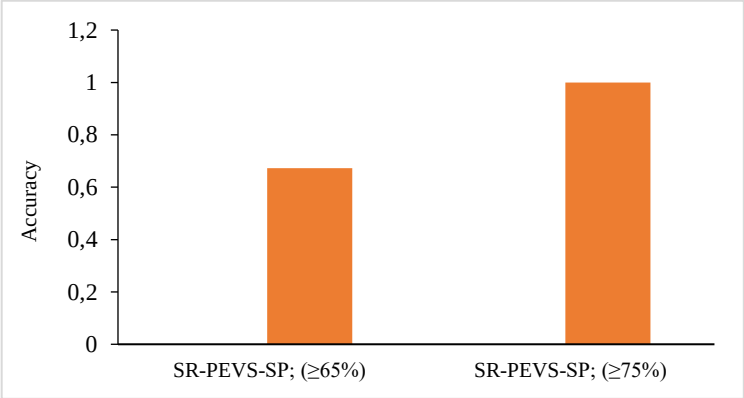

Fig. 4: Comparison of accuracy using different values of similarity precision for Marine dataset

Results from Table 12 and Fig. 3 show that the similarity precision with possible equivalent value-set achieved better accuracy as compared to the limited tolerance relation approach to 22.24%. Results in Table 13 and Fig. 4 show that by using the proposed approach with similarity precision (≥75%) achieved higher accuracy as compared to proposed approach with similarity precision (≥65%) up to 48%. It means similarity precision plays a big role in achieving higher classification accuracy.

5.0 CONCLUSION

Classical rough set theory such as tolerance relation and limited tolerance relation can be used to handle incomplete information systems. However, these approaches unable to produce promising results in terms of accuracy of approximation. Thus, to overcome the limitation, we proposed a new approach based on similarity relation with semantically justified using possible equivalent value-set. It is based on a classification of three semantics types of incomplete information (i.e., "any value", "may be value" and "not applicable value") for modelling similarity. The advantage of using the presented approach is there are more similar objects within the same indiscernibility classes that leads to greater value of lower approximation. Consequently, the similarity precision is considered to improve the accuracy of similarity relation with possible equivalent value-set. From the simulation results by using different and real dataset respectively, we are able to obtain better value for accuracy of approximation as compared to limited tolerance relation up to two orders of magnitude. Thus, the new approach is more flexible and precise as compared to the limited tolerance relation. This paper also discusses the mathematical properties of the proposed new approach.

ACKNOWLEDGMENT

This research is supported by Research Management Center , Universiti Malaysia Terengganu

REFERENCES

[1] Z. Pawlak, "Rough sets", *International Journal of Computer and Information Science*, Vol. 11, No. 5, Oct 1982, pp. 341–356.

[2] S. Imai, C. W. Lin, J. Watada, and G. H. Tzeng, "Knowledge acquisition in human resource management based on rough sets", in *Portland International Conference on Management of Engineering & Technology (PICMET)*, Cape Town, South Africa, July 2008, pp. 969–974.

[3] F. H. Chen, D. J. Chi, and C. Y. Kuo, "Using rough set theory and decision trees to diagnose enterprise distress - Consideration of corporate governance variables", in *International Conference on Intelligent Computing*, Vol. 8589, No. 55, 2014, pp. 199–211.

[4] Q. Yang, P. A. Du, Y. Wang, and B. Liang, "A rough set approach for determining weights of decision makers in group decision making," *PLoS One*, Vol. 12, No. 2, 2017, pp. 1–16.

[5] L. Sun, W. Wang, J. Xu, and S. Zhang, "Improved LLE and neighborhood rough sets-based gene selection using Lebesgue measure for cancer classification on gene expression data", *Journal of Intelligent & Fuzzy Systems*, Vol. 37, No. 4, 2019, pp. 1–12.

[6] L. Sun, X. Zhang, Y. Qian, J. Xu, and S. Zhang, "Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification", *Information Science*, Vol. 502, 2019, pp. 18–41.

[7] Z. Lu and Z. Qin, "Rule extraction from incomplete decision system based on novel dominance relation", in *4th International Conference on Intelligent Networks and Intelligent Systems*, China, November 2011, pp. 149–152.

[8] I. T. R. Yanto, P. Vitasari, T. Herawan, and M. M. Deris, "Applying variable precision rough set model for clustering student suffering study's anxiety", *Expert Systems with Applications*, Vol. 39, No. 1, 2012, pp. 452–459.

[9] T. Herawan, M. M. Deris, and J. H. Abawajy, "A rough set approach for selecting clustering attribute" , *Knowledge-Based System*, Vol. 23, No. 3, 2010, pp. 220–231, 2010.

[10] D. Parmar, T. Wu, and J. Blackhurst, "MMR: An algorithm for clustering categorical data using Rough Set Theory", *Data & Knowledge Engineering*, Vol. 63, No. 3, 2007, pp. 879–893.

[11] A. Skowron and P. Wasilewski, "Toward interactive Rough-Granular Computing", *Control and Cybernertics*, Vol. 40, No. 2, 2011, pp. 213–235.

[12] A. Skowron, J. Stepaniuk, and R. Swiniarski, "Approximation spaces in rough-granular computing" , *Fundamenta Informaticae*, Vol. 100, No. 1–4, 2010, pp. 141–157, 2010.

[13] M. M. Deris, N. Senan, Z. Abdullah, R. Mamat, and B. Handaga, "Dimensional Reduction using conditional entropy for incomplete information systems", in *International Conference on Parallel Computing Technologies*, 2019, pp. 263–272.

[14] L. Ke, Z. Feng, and Z. Ren, "An efficient ant colony optimization approach to attribute reduction in rough set theory", *Pattern Recognition Letters*, Vol. 29, No. 9, 2008, pp. 1351–1357.

[15] Q. Hu, D. Yu, and Z. Xie, "Information-preserving hybrid data reduction based on fuzzy-rough techniques", *Pattern Recognition Letters*, Vol. 27, No. 5, 2006, pp. 414–423.

[16] X. Jia, L. Shang, B. Zhou, and Y. Yao, "Generalized attribute reduct in rough set theory", *Knowledge-Based System.*, Vol. 91, 2016, pp. 204–218.

[17] L. Hedjazi, J. Aguilar-Martin, and M. V. Le Lann, "Similarity-margin based feature selection for symbolic interval data", *Pattern Recognition Letters*, Vol. 32, No. 4, 2011, pp. 578–585.

[18] Y. Li and Z. F. Wu, "Fuzzy feature selection based on min-max learning rule and extension matrix", *Pattern Recognition Letters*, Vol. 41, No. 1, 2008, pp. 217–226, 2008.

[19] D. Kim, "Data classification based on tolerant rough set", *Pattern Recognition Letters*, Vol. 34, No. 8, 2001, pp. 1613–1624.

[20] S. Trabelsi, Z. Elouedi, and P. Lingras, "Classification systems based on rough sets under the belief function framework", *International Journal of Approximate Reasoning*, Vol. 52, No. 9, 2011, pp. 1409–1432.

[21] K. Kaneiwa, "A rough set approach to multiple dataset analysis", *Applied Soft Computing*, Vol. 11, No. 2, 2011, pp. 2538–2547.

[22] H. Chongzao, "A novel approach of rough conditional entropy-based attribute selection for incomplete decision system", *Mathematical Problems in Engineering*, 2014.

[23] J. W. Grzymala-Busse, "Rough set strategies to data with missing attribute values", *Foundations and Novel Approaches in Data Mining*, 2003, pp. 56–63.

[24] N. F. Md Nasir, N. Ibrahim, M. M. Deris, and M. Z. Saringat, "Test case and requirement selection using rough set theory and conditional entropy", in *International Conference on Computational Intelligence in Information System*, 2018, pp. 61–71.

[25] B. Qin, F. Zeng, and K. Yan, "Knowledge structures in a tolerance knowledge base and their uncertainty measures", *Knowledge-Based System*, Vol. 151, 2018, pp. 198–215.

[26] J. Stefanowski and A. Tsoukiàs, "Incomplete information tables and rough classification", *Computational Intelligence*, Vol. 17, No. 3, 2001, pp. 545–566.

[27] J. W. Grzymala-Busse, "Characteristic relations for incomplete data: A generalization of the indiscernibility relation", in *International Conference on Rough Sets and Current Trends in Computing*, Vol. 3066, 2004, pp. 244–253.

[28] Y. Y. Guan and H. K. Wang, "Set-valued information systems", *Information Sciences.*, Vol. 176, No. 17, 2006, pp. 2507–2525.

[29] G. Wang, L. Guan, and F. Hu, "Rough set extensions in incomplete information systems," *Frontier of Electrical and Electron. Engineering in. China*, Vol. 3, No. 4, 2008, pp. 399–405.

[30] M. Kryszkiewicz, "Rough set approach to incomplete information systems", *Information Science.*, Vol. 112, 1998, pp. 39–49.

[31] D. Van Nguyen, K. Yamada, and M. Unehara, "Extended tolerance relation to define a new rough set model in incomplete information systems", *Advances in Fuzzy Systems*, Vol. 2013, 2013, pp. 1–11.

[32] G. Wang, "Extension of rough set under incomplete information systems," in *IEEE International Conference*

on Fuzzy Systems, USA, August 2002, pp. 1098–1103.

[33] W. Lipski, "On Databases with Incomplete Information", *Journal of the Association for Computing Machinery*, Vol. 28, No. 1, 1981, pp.41-70.

[34] M. M. Deris, M. A. Hamid, N. Ibrahim, R. Efendi, and I. T. Riyadi Yanto, "Data Reduction using Similarity Class and Enhanced Tolerance Relation for Complete and Incomplete Information Systems", in *10th International Conference on Information and Communication Systems (ICICS)*, Jordan, August 2019, pp. 134–139.

[35] J. Luo, H. Fujita, Y. Yao, and K. Qin, "On modeling similarity and three-way decision under incomplete information in rough set theory", *Knowledge-Based System*, Vol. 191, 2020, pp. 105251.

[36] UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets.php

[37] R. R. Saedudin, S. Kasim, H. Mahdin et al, "A Relative Tolerance Relation of Rough Set (RTRS) for Potential Fish Yields in Indonesia", *Journal of Coastal Research*, 2018.