Sains Malaysiana 48(12)(2019): 2831–2839
http://dx.doi.org/10.17576/jsm-2019-4812-24

A Relative Tolerance Relation of Rough Set in Incomplete Information
(Perhubungan Toleransi Relatif Set Kasar dalam Maklumat tak Lengkap)

Rd Rohmat Saedudin*, Shahreen Kasim, Hairulnizam Mahdin, Mohd Farhan Md Fudzee,
Edi Sutoyo, Iwan Tri Riyadi Yanto, Rohayanti Hassan

ABSTRACT

University is an educational institution that has objectives to increase student retention and also to make sure students graduate on time. Student learning performance can be predicted using data mining techniques e.g. the application of finding essential association rules on student learning base on demographic data by the university in order to achieve these objectives. However, the complete data i.e. the dataset without missing values to generate interesting rules for the detection system, is the key requirement for any mining technique. Furthermore, it is problematic to capture complete information from the nature of student data, due to high computational time to scan the datasets. To overcome these problems, this paper introduces a relative tolerance relation of rough set (RTRS). The novelty of RTRS is that, unlike previous rough set approaches that use tolerance relation, non-symmetric similarity relation, and limited tolerance relation, it is based on limited tolerance relation by taking account into consideration the relatively precision between two objects and therefore this is the first work that uses relative~~ly~~ precision. Moreover, this paper presents the mathematical properties of the RTRS approach and compares the performance and the existing approaches by using real-world student dataset for classifying university's student performance. The results show that the proposed approach outperformed the existing approaches in terms of computational time and accuracy.

Keywords: Classification; educational data mining; incomplete information systems; rough set theory

ABSTRAK

Universiti adalah sebuah institusi pendidikan yang antara objektifnya adalah untuk meningkatkan penahanan pelajar dan juga untuk memastikan pelajar bergraduasi dalam jangka masa yang ditetapkan. Untuk mencapai objektif tersebut, pelajar perlulah memastikan prestasi pembelajaran sentiasa konsisten. Teknik perlombongan data boleh digunakan untuk meramal prestasi pembelajaran pelajar. Namun, isu data hilang atau data tidak lengkap membataskan keberkesanan teknik perlombongan data khasnya dalam mengenal pasti hubungan atribut pembelajaran pelajar dan atribut demografi pelajar. Isu menjadi lebih sukar apabila melibatkan data pelajar yang banyak. Maka, kertas ini mencadangkan teknik perhubungan toleransi relatif set kasar (RTRS) bagi mengatasi isu ini. Kelainan RTRS dalam kertas ini adalah dengan menggunakan ketepatan relatif antara dua objek atribut. Selain itu, kertas ini turut membentangkan formula matematik yang digunakan dalam RTRS. Seterusnya, prestasi cadangan teknik RTRS ini dibandingkan dengan teknik asal menggunakan set data pelajar universiti untuk mengelaskan prestasi pelajar tersebut. Hasil menunjukkan bahawa teknik RTRS yang dicadangkan mengatasi teknik sedia ada daripada segi masa komputer dan ketepatan.

Kata kunci: Pengelasan; perlombongan data pendidikan; sistem maklumat tidak lengkap; teori set kasar

Introduction

In university, students' performance is a great concern to the higher education where several factors may affect them. Detecting students from failure is a major problem and it has become very important for the higher education institution to get more understanding why so many students were failed to graduate on time. Higher education institution needs to have approximate prior knowledge of enrolled students to predict their performance in future academics. It helps them to identify promising students and also provides them an opportunity to pay attention and to improve those who would probably get lower grades, which would affect their graduation time. By evaluating students' performance, a strategic program can be planned during their period of studies in an institution well (Ibrahim & Rusli 2007), like arranging intensive guidance to improve students' academic performance. Moreover, the detection and prevention of student failure at university and early intervention make much more sense than remediation (Slavin et al. 1994). An effective way to detect student failure is the use of data mining techniques (Márquez-Vera et al. 2013). The computational process of discovering and extracting patterns in large datasets is Data Mining. It is a process that involves methods of applying data analysis and discovery algorithms, that under acceptable computational efficiency limitations, produce a particular enumeration of patterns (or models) over the data (Fayyad 1996). It also detects hidden knowledge and patterns which were previously unknown from large databases for easy and fast retrieval of data and information (Ogunde & Ajibade 2014). As an interdisciplinary field, data mining draws from statistical analysis, database systems, machine learning, pattern recognition, neural networks, and fuzzy systems (Dobrota et al. 2014).

In various fields like financial banking, medicine, manufacturing engineering, customer relationship management, web mining, geochemical and e-learning (Saedudin et al. 2018, 2017, 2016; Sutoyo et al. 2019, 2017; Yanto et al. 2018a, 2018b, 2016) can apply data mining concepts and methods. The new emerging technique of data mining is educational data mining that can be applied to the data related to the field of education (Romero & Ventura 2007), including students' performance evaluation. Machine learning (Chiroma et al. 2015; Kotsiantis et al. 2004; Pal 2012; Sutoyo et al. 2017; Yadav et al. 2012), association rule mining (Borkar & Rajeswari 2013), decision tree (Yadav & Pal 2012), attribute selection (Mohammed et al. 2016; Saedudin et al. 2017), and genetic algorithm (Minaei-Bidgoli et al. 2003) have been proposed in the domain of evaluating students' performance. However, all those methods do not consider the missing values and can only be applied if all the preference values are completely available. Even though in fact, in real world problems users are not able to provide all the preference values that are required, and then, we have to deal with incomplete information systems.

RELATED WORK

There have been many efforts in studying incomplete information systems. The simplest method to deal with incomplete information systems is to remove objects with unknown values (Bunting et al. 2002) or to replace missing values with most common values (Chmielewski et al. 1993). However, this approach certainly reduces the sample size of data. Another disadvantage to this approach is that the objects with missing values may be different than the objects without missing values (e.g. missing values that are non-random). Recently, the extension of the classical rough set theory based on tolerance relation (Kryszkiewicz 1999, 1998; Yang 2009; Zhou 2010; Zhou & Yang 2012), non-symmetric similarity relation (Stefanowski & Tsoukias 2001, 1999; Wu & Guo 2010), and limited tolerance relation (Wang 2002; Yang et al. 2011) have also been proposed and studied to cope with incomplete information systems. However, a tolerance relation approach leads to poor results in terms of approximation. Consequently, Stefanowski and Tsoukias (2001, 1999) introduced non-symmetric similarity relation to refining the results obtained using tolerance relation approach. However, Wang (2002) and Yang et al. (2011) proved that similarity relation will lose some information and proposed limited tolerance relation. Nevertheless, some information may be also lost because the limited tolerance relation does not consider the similarity precision between two objects. Nguyen et al. (2013) improved the tolerance relation by considering the probability matching between two objects. However, it needs to know the probability distribution of data in advance.

In order to overcome their drawbacks, in this paper, we propose a relative tolerance relation of the rough set (RTRS). The RTRS is based on limited tolerance relation by taking into consideration the relatively precision between two objects. The relative precision is defined when a threshold value is given. By adjusting the threshold, we are able to obtain better results as compared with limited tolerance relation. In summary, the contribution of this work is described as follows:

The relative precision of rough set is proposed with aims to modify the limited tolerance relation; A correctness proof and related algorithms of the proposed approach are presented; Comparative analysis and experiment results between the proposed approach with the existing baseline approach in terms of computational time and accuracy by using real-world student dataset for predicting university's student performance are elaborated; and the result found that the proposed approach outperforms as compared with the existing baseline approaches.

The rest of the paper is organized as follows. Next section will describes methods and the proposed method, the RTRS for handling incomplete information systems. Then, continues with describes the result, analysis, and discussion and compares them with existing approaches. Finally, the conclusion of this work is described in last section.

METHODS

In this section, the basic concepts of information systems and rough set theory will be explained. Afterward, the extensions of rough set in incomplete information systems namely tolerance relation (TR), non-symmetric similarity relation (NSSR), and limited tolerance relation (LTRS) are also explained.

TOLERANCE RELATION

Let give a complete information system equipped with decision $S = (U, A, V, f)$, where $A = C \cup \{d\}$, C is a set of condition attributes and d the decision attribute, such that $f: U \times A \rightarrow V$, for any $a \in A$, where V_a is called domain of an attribute a. In incomplete information systems $S^* = (U, A, V_*, f)$, for any subset $B \subseteq C$, the tolerance relation T is defined by the following definition.

Definition 1 Let $S^* = (U, A, V_*, f)$ be an incomplete information system. A tolerance relation T is defined as

$$\forall x,y \in U\ T(x, y) \Leftrightarrow \forall_{c_j \in B}(c_j(x) = c_j(y) \vee c_j(x) = * \vee c_j(y) = *).$$

Thus,

$$T = \{(x,y) x \in U \wedge y \in U \wedge \forall c_j (c_j \in B \rightarrow (c_j(x)$$
$$= c_j(\mathrm{y}) \vee c_j(x) = * \vee c_j(y) = *))\}$$

Obviously, T is reflexive and symmetric, but not transitive. From Definition 1, we describe the notion of tolerance class as follows.

Definition 2 Let $S^* = (U, A, V_*, f)$ be an incomplete information system. The tolerance class $I_B^T(x)$ of an object x with reference to an attribute set B is defined as $I_B^T(x) = \{y \mid y \in U \wedge T_B(x, y)\}$.

From Definition 2, we describe the notion of lower and upper approximations of tolerance class as follows.

Definition 3 Let $S^* = (U, A, V_*, f)$ be an incomplete information system. The lower approximation x_T^B and upper approximation x_B^T of an object set X with reference to attribute set B, respectively can be defined as follow:

$$x_B^T = \{x \mid x \in U \wedge I_B^T(x) \subseteq X\} \quad \text{and}$$

$$x_T^B = \{x \mid x \in U \wedge I_B^T(x) \cap X \neq \phi\}.$$

We can illustrate the this concepts with an incomplete information system from Wang (Yadav et al. 2012).

Example 1 Table 1 shows an incomplete information system, where $a_1, a_2, \ldots, a_{12}$ are the objects. The c_1, c_2, c_3, c_4 are condition attributes, where their domain values are $\{0,1,2,3\}$. The d is a decision attribute, where its domain values are $\{\beta, \Omega\}$, $\beta = \{a_1, a_2, a_4, a_7, a_{10}, a_{12}\}$ and $\Omega = \{a_3, a_5, a_6, a_8, a_9, a_{11}\}$.

From Table 1, we can easily obtain the results by analyzing it with the tolerance relation in Definition 1, as follows

$$I_C^T(a_1) = \{a_1, a_{11}, a_{12}\}, I_C^T(a_2) = \{a_2, a_3\}, I_C^T(a_3) = \{a_2, a_3\}, I_C^T(a_4) = \{a_4, a_5, a_{10}, a_{11}, a_{12}\}, I_C^T(a_5) = \{a_4, a_5, a_{10}, a_{11}, a_{12}\}, I_C^T(a_6) = \{a_6\}, I_C^T(a_7) = \{a_7, a_8, a_9, a_{11}, a_{12}\}, I_C^T(a_8) = \{a_7, a_8, a_{10}\}, I_C^T(a_9) = \{a_7, a_9, a_{11}, a_{12}\}, I_C^T(a_{10}) = \{a_4, a_5, a_8, a_{10}, a_{11}\}, I_C^T(a_{11}) = \{a_1, a_4, a_5, a_7, a_9, a_{10}, a_{11}, a_{12}\}, I_C^T(a_{12}) = \{a_1, a_4, a_5, a_7, a_9, a_{11}, a_{12}\},$$

and

$$\beta_C^T = \phi, \Omega_C^T = \{a_6\}, \beta_T^C = \{a_1, a_2, a_3, a_4, a_5, a_7, a_8, a_9, a_{10}, a_{11}, a_{12}\}, \Omega_T^C = U.$$

From the analysis, some objects that can be discerned intuitively cannot be classified, such as a_1 has complete information, but a_1 is not in the lower approximation of β. The reason is that the missing attribute values of a_{11} is considered similar to a_1. In the following sub-section, we discuss the non-symmetric similarity relation.

NON-SYMMETRIC SIMILARITY RELATION

The assumption of this approach is that the missing attribute values are not uncertain, but it is non-existing and it is not comparable to any other attribute values. An object x is considered to be similar to object y only if all their known attribute values are the same. Thus, one object may have a complete description than the other, the inverse relationship does not hold. The notion of a non-symmetric similarity relation is given in the following definition.

Definition 4 (Saedudin et al. 2018, 2017a) Let $S^* = (U, A, V_*, f)$ be an incomplete information system. A non-symmetric similarity relation S is defined as

$$\forall x, y \in U\ (S_B(x,y) \Leftrightarrow \forall_{c_j \in B}(c_j(x) = c_j(y) \vee c_j(x) = *)).$$

It is obvious that S is transitive and reflexive but not symmetric. From Definition 4, we can induce two similarity sets as given in Definitions 5 and 6.

Definition 5 Let $S^* = (U, A, V_*, f)$ be an incomplete information system and $B \subseteq A$.The set of objects similar to object x denoted by $\mathrm{Sim}_B(x)$ is defined as

$$\mathrm{Sim}_B(x) = \{y \mid y \in U \wedge S_B(y,x)\}$$

Definition 6 Let $S^* = (U, A, V_*, f)$ be an incomplete information system and $B \subseteq A$.The set of objects which x is similar to $\mathrm{Sim}_B^{-1}(x)$ is defined as

$$\mathrm{Sim}_B^{-1}(x) = \{y \mid y \in U \wedge S_B(x, \mathrm{y})\}.$$

Clearly, $\mathrm{Sim}_B(x)$ and $\mathrm{Sim}_B^{-1}(x)$ are two different sets. To clearly depict the two similarity sets as defined above, we illustrate through an example from Table 1.

Example 2 From Table 1, the results for non-symmetric similarity relations are as follows

$$\mathrm{Sim}_B^{-1}(a_5) = \{a_4, a_5\}, \mathrm{Sim}_B(a_5) = (a_4, a_5, a_{11}\}, \mathrm{Sim}_B^{-1}(a_6) = \{a_6\}, \mathrm{Sim}_B(a_6) = \{a_6\}, \mathrm{Sim}_B^{-1}(a_7) = \{a_7, a_9\}, \mathrm{Sim}_B(a_7) = \{a_7\}, \mathrm{Sim}_B^{-1}(a_8\} = \{a_8\}, \mathrm{Sim}_B(a_8) = \{a_8\}, \mathrm{Sim}_B^{-1}(a_9\} = \{a_9\}, \mathrm{Sim}_B(a_9) = \{a_7, a_9, a_{11}, a_{12}\}, \mathrm{Sim}_B^{-1}(a_{10}) = \{a_{10}\}, \mathrm{Sim}_B(a_{10}) = \{a_{10}\}, \mathrm{Sim}_B^{-1}(a_{11}\} = \{a_1, a_4, a_5, a_9, a_{11}, a_{12}\}, \mathrm{Sim}_B(a_{11}) = \{a_{11}\}, \mathrm{Sim}_B^{-1}(a_8\}(a_{12}) = \{a_1, a_9, a_{12}\}, \mathrm{Sim}_B(a_{12}) = \{a_{11}, a_{12}\},$$

and

$$B_C^T = \{a_1, a_{10}\}, B_T^C = \{a_1, a_2, a_3, a_4, a_5, a_7, a_{10}, a_{11}, a_{12}\},$$

$$\Omega_C^T = \{a_6, a_8, a_9\}, \Omega_T^C = \{a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{11}, a_{12}\}.$$

Further, from Definitions 5 and 6, the lower approximation and upper approximation of objects set X can be defined as follows.

Definition 7 Let $S^* = (U, A, V_*, f)$ be an incomplete information system and $B \subseteq A$. The lower-approximation

X_B^S and the upper-approximation X_S^B of an object set X with respect to an attribute set $B \subseteq A$ are, respectively, defined as

$$X_B^S = \{x \mid x \in \mathrm{U} \wedge \mathrm{Sim}_B^{-1}(x) \subseteq X\} \textit{ and}$$

$$X_S^B = \cup\{\mathrm{Sim}_B(x) \mid x \in X\}.$$

The approximations showed by the non-symmetric similarity relation are more informative than those resulted in tolerance relation. To clearly depict the non-symmetric similarity relation as defined above, we illustrate through an example from Table 1.

Example 3 From Table 1, the lower approximations of the set β and Ω include some objects which are intuitively expected to be classified into β and Ω, respectively. However, object a_1 and object a_{12} look alike but would not be similar in terms of non-symmetric similarity relation. In the following sub-section, we discuss the limited tolerance relation.

LIMITED TOLERANCE RELATION

In an information system, two objects may be distinct because of a little missing information. For example, two objects $a = \{x, *, y, z, w\}$ and $b = \{*, v, y, z, w\}$ are similar, but they do not satisfy the non-symmetric similarity relation. To avoid such a problem, Wang (2002) and Yang et al. (2011) developed a limited tolerance relation based on Definition 8 as follows.

Definition 8 (Kryszkiewicz 1998). Let $S^* = (U, A, V_*, f)$ be an incomplete information system, a subset $B \subseteq A$, and $P_B(x) = \{b \,|\, b \in B \wedge b(x) \neq *\}$. A binary relation L (limited tolerance relation) defined on U is given as $\forall_{x,y \in UxU}(L_B(x, y) \Leftrightarrow \forall_{b \in B}(b(x) = b(y) = *) \vee ((P_B(x) \cap P_B(y) \neq \phi) \wedge \forall_{b \in B}((b(x) \neq *) \wedge (b)y) \neq *) \rightarrow (b(x) = b(y))))$

Obviously, the limited tolerance relation is symmetric and reflexive but not transitive. In Definition 8, the condition that $(b(x) \neq *) \wedge (b(y) \neq *) \rightarrow (b(x) = b(y))$ is equivalent to $(b(x) = *) \vee (b(y) = *) \vee (b(x) = b(y))$. Thus, two objects that satisfy the tolerance relation but not limited tolerance relation are only those hold $P_B(x) \cap P_B(y) = \phi$.

In other words, two objects are in limited tolerance relation if there are in one of the two cases. Firstly, is that all attribute values of the two objects are missing. Secondly, is where at least an attribute having an ordinary value for both objects and the two objects have the same value for those attributes. The notion of limited tolerance class is given as follow:

Definition 9 Let $S^* = (U, A, V_*, f)$ be an incomplete information system and a subset $B \subseteq A$. The limited tolerance class is defined as $I_B^L(x) = \{y \mid y \in U \wedge L_B(x, y)\}$.

To clearly depict the limited tolerance class as defined above, we illustrate through an example from Table 1.

Example 4 Analyzing Table 1, with limited tolerance relation, we can get the following results of limited tolerance classes.

$$I_C^L(a_1) = \{a_1, a_{11}, a_{12}\}, I_C^L(a_2) = \{a_2, a_3\}, I_C^L(a_3) = \{a_2, a_3\}, I_C^L(a_4) = \{a_4, a_5, a_{11}, a_{12}\}, I_C^L(a_5) = \{a_4, a_5, a_{11}, a_{12}\}, I_C^L(a_6) = \{a_6\}, I_C^L(a_7) = \{a_7, a_9, a_{12}\}, I_C^L(a_8) = \{a_8\}, I_C^L(a_9) = \{a_7, a_9, a_{11}, a_{12}\}, I_C^L(a_{10}) = \{a_{10}\}, I_C^L(a_{11}) = \{a_1, a_4, a_5, a_9, a_{11}, a_{12}\}, I_C^L(a_{12}) = \{a_1, a_4, a_5, a_7, a_9, a_{11}, a_{12}\},$$

and

$$\beta_C^T = \{a_{10}\}, \beta_T^C = \{a_1, a_2, a_3, a_4, a_5, a_7, a_9, a_{10}, a_{11}, a_{12}\},$$

$$\Omega_C^T = \{a_6, a_8\}, \Omega_T^C = \{a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{11}, a_{12}\}.$$

From these results, there is no known equal attribute value for the object a_{10} and a_{11}. There are indiscernible in the tolerance relation but discernible in the limited tolerance relation. Also, most attribute values (c_1, c_2, c_3) are equal to each other for objects a_9 and a_{12}. There are discernible in the non-symmetric similarity relation but indiscernible in the limited tolerance relation. However, a_1 has complete information, but a_1 is not in the lower approximation of β.

From Definition 9, the notions of lower approximation and the upper approximation of an object x based on the limited tolerance class are given in the following definition.

Definition 10 The lower approximation and the upper approximation of an object x based on the limited tolerance class $I_B^L(x)$ are respectively defined as

$$D_L^B = \{x \mid x \in U \wedge I_B^L(x) \cap D \neq \phi\} \text{ and}$$

$$D_B^L = \{x \mid x \in U \wedge I_B^L(\mathrm{x}) \subseteq D\}.$$

From non-symmetric similarity and limited tolerance relations, in the following section, we present the proposed new limited tolerance relation approach.

TABLE 1. An incomplete information tables

	C_1	C_2	C_3	C_4	d
a_1	3	2	1	0	β
U/A	C_1	C_2	C_3	C_4	d
a_2	2	3	2	0	β
a_3	2	3	2	0	Ω
a_4	*	2	*	1	β
a_5	*	2	*	1	Ω
a_6	2	3	2	1	Ω
a_7	3	*	*	3	β
a_8	*	0	0	*	Ω
a_9	3	2	1	3	Ω
a_{10}	1	*	*	*	β
a_{11}	*	2	*	*	Ω
a_{12}	3	2	1	*	β

PROPOSED RELATIVE TOLERANCE RELATION OF ROUGH SET

In this section, we introduce the concept of relatively precision between objects x and y in order to determine that two objects are tolerant.

RELATIVELY PRECISION

Let given an incomplete information system $S^* = (U, A, V_*, f)$, where $A = C \cup \{\mathrm{d}\}$, C is a set of condition attributes and d the decision attribute, such that $f: U \times A \rightarrow V_*$. For any $a \in A$, where V_a is called the domain of an attribute a and a subset $B \subseteq C$, the relatively precision is defined as follows.

Definition 11 Let $P_B(x) = \{b \mid b \in B \wedge b(x) \neq *\}$, the relatively precision δ, is defined as

$$\delta(x, y) = \delta(x, y) = \frac{|P_B(x) \cap P_B(y)|}{|P_B(x) \cup P_B(y)|},$$

where $|\bullet|$ represents the cardinality of the set.

From Definition 11, it is clear that $0 < \alpha(x, y) \leq 1$. From Definition 11, the relatively precision limited tolerance relation with relatively precision is given in Definition 12 as follow:

Definition 12 Let $S^* = (U, A, V_*, f)$ be an incomplete information system. The relatively precision limited tolerance relation L δ is defined as follows

$\forall_{x,y \in U \times U}(L\delta_B(x,y) \Leftrightarrow \forall_{b \in B}(b(x) = b(y) = *) \vee ((\alpha(x, y)) \geq \delta) \wedge$

$\forall_{b \in B}(((b(x) \neq *) \wedge (b(y) \neq *)) \rightarrow (b(x) = b(y)))$ where $\delta \in (0,1]$ is a threshold value.

Since $\delta \in (0,1]$, then $0 < \alpha(x,y) \leq 1$ which implies that $P_B(x) \cap P_B(y) \neq \phi$ holds, but not vice versa if the certain threshold value of the similarity is given.

To clearly depict the limited tolerance class as defined above, we illustrate through an example as follows.

Example 5 Two objects $x = \{1,*,*,2,*,*,0,0\}$ and $y = \{*,*,*,2,0,0,*,*\}$ are tolerant if it is based on limited tolerance relation., i.e., $P(x) \cap P(y) = \{2\}$ or $\alpha(x, y) = 0.125$. From the value of $\alpha(x, y)$, we believed that both objects are tolerant is too loose. Moreover, if we set $\delta = 0.4$, then $(x, y) \notin L\delta$. That is the two objects are not tolerant if the relatively precision does not hold the threshold value.

Now, we define the extended tolerance relation by using relatively precision with a threshold.

Definition 13 Let $S^* = (U, A, V_*, f)$ be an incomplete information system, a subset $B \subseteq C$, and a threshold δ. The relatively precision limited tolerance relation is defined as follows

$$L\delta_B(x,y) \Leftrightarrow \alpha_B(x, y) \geq \delta.$$

It is easy to observe that these relation is reflexive and symmetric but not necessarily transitive. To clearly depict the limited tolerance with similarity precision as defined, we illustrate through an example from Table 1.

Example 6 From Table 1, two objects a_1 and a_{11} are not tolerant if $\delta = 0.4$. However, two objects a_1 and a_{12} are tolerant due to $\alpha(a_1, a_{12}) \geq 0.4$.

In the following sub-section, we present two properties of our proposed relatively precision limited tolerance relation and their correctness proofs.

PROPERTIES AND CORRECTNESS PROOFS

Proposition 1 Let given an incomplete information system $S^* = (U, A, V_*, f)$, a subset $B \subseteq C$ and $x \in U$. If $\delta > 0$, then

a. For any x and y, $L\delta_B(x, y) \Rightarrow L_B(x, y)$

b. $L\delta_B(x, y) \Leftarrow L_B(x, y)$ except for the case when $P_B(x) \cap P_B(y) = \phi$.

Proof

a. When $\delta > 0$, then $L\delta_B(x, y) \Leftrightarrow \alpha_B(x, y) > 0$
$\Leftrightarrow P_B(x) \cap P_B(y) \neq \phi$
$\wedge \forall a \in P_B(\mathrm{x}) \cap P_B(y).\, f_a(x) = f_a(y)$
$\Rightarrow L_B(x, y)$

b. It is clear that $L_B(x, y) \Rightarrow L\delta_B$ except the case when $P_B(x) \cap P_B(y) = \phi$.

Definition 14 Let $S^* = (U, A, V_*, f)$ be an incomplete information system and $B \subseteq A$. The limited tolerance class is defined as $I_B^{L\delta} = \{y \mid y \in U \wedge L\delta_B(x,y)\}$.

To clearly depict the limited tolerance class as defined, we illustrate through an example from Table 1.

Example 7 From Table 1, and let $\delta > 0.5$, we have the tolerance classes as follows

$I_C^{L\delta}(a_1) = \{a_1, a_{12}\}$, $I_C^{L\delta}(a_2) = \{a_2, a_3\}$, $I_C^{L\delta}(a_3) = \{a_2, a_3\}$, $I_C^{L\delta}(a_4) = \{a_4, a_5\}$, $I_C^{L\delta}(a_5) = (a_4, a_5)$, $I_C^{L\delta}(a_6) = \{a_6\}$, $I_C^{L\delta}(a_7) = \{a_7, a_9\}$, $I_C^{L\delta}(a_8) = \{a_8\}$, $I_C^{L\delta}(a_9) = \{a_7, a_9, a_{12}\}$, $I_C^{L\delta}(a_{10}) = \{a_{10}\}$, $I_C^{L\delta}(a_{11}) = \{a_{11}\}$, $I_C^{L\delta}(a_{12}) = \{a_1, a_9, a_{12}\}$,

and

$\beta_C^{L\delta} = \{a_1, a_{10}, a_{12}\}$, $\beta_{L\delta}^C = \{a_1, a_2, a_3, a_4, a_5, a_7, a_9, a_{10}, a_{11}, a_{12}\}$, $\Omega_C^{L\delta} = \{a_6, a_8, a_{11}\}$, $\Omega_{L\delta}^C = \{a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{11}, a_{12}\}$.

From the analysis, the results of the proposed approach are more flexible and precise as compared to the tolerance relation, non-symmetric similarity relation, and limited tolerance relation, where in this case a_1, a_5 and a_{11} are divided into a different class. We also found that, $\beta_{L\delta}^C \subseteq \beta_L^C$, $\Omega_C^{L\delta} \supseteq \Omega_C^L$, and $\Omega_{L\delta}^C \subseteq \Omega_L^C$.

Definition 15 Let given an incomplete information system $S^* = (U, A, V_*, f)$. The lower approximation and the upper approximation of an object x based on the limited tolerance class $I_B^{L\delta}(x)$ denoted as $D_{L\delta}^B(x)$ and $D_B^{L\delta}(x)$ respectively, are defined as

$$D_B^{L\delta} = \{x \mid x \in U \wedge I_B^{L\delta}(x) \subseteq D\} \textit{ and}$$

$$D_{L\delta}^B = \{x \mid x \in U \wedge I_B^{L\delta}(x) \cap D \neq \varphi\}.$$

From Definition 12, we can generalize Proposition 1 as described in the following proposition.

Proposition 2. Let $S^* = (U, A, V_*, f)$ be an incomplete information system, a subset $B \subseteq A$ and $x \in U$. If $0 \leq \delta_1 < \delta_2 \leq 1$, then

$$I_B^{L\delta_2} \subseteq I_B^{L\delta_1}$$

Proof

For every $a \in I_B^{L\delta_2}(x)$, we have $\alpha_B(x,y) \geq \delta_2$. Since $\delta_2 > \delta_1$, then $\alpha_B(x, y) \geq \delta_1$, that is $\forall a \in I_B^{L\delta_1}(x)$ which implies $I_B^{L\delta_2}(x) = I_B^{L\delta_1}(x)$. However, if $\alpha_B(x, y) \geq \delta_1$ then it does not necessarily $\alpha_B(x, y) \geq \delta_2$. Hence $I_B^{L\delta_2} \subseteq I_B^{L\delta_1}$.

To clearly depict the property of generalized limited tolerance class in Proposition 2, we illustrate through an example from Table 1.

Example 7 From Table 1, we have $I_C^{L\delta_1}(a_1) = I_C^{L\delta_2}(a_1) = \{a_1, a_{11}, a_{12}\}$ for $\delta_1 = 0.25$. However, for $\delta_2 = 0.5$, we have $I_C^{L\delta_2}(a_1) = \{a_1, a_{12}\}$ and thus $I_C^{L\delta_2}(a_1) \neq I_C^{L\delta_1}(a_1)$.

From Definition 15 and Proposition 2, we have the following property of the lower approximation and the upper approximation.

Proposition 3 Let given an incomplete information system $S^* = (U, A, V_*, f)$, a subset $B \subseteq A$ and $x \in U$. If $0 \leq \delta 1 < \delta 2 \leq 1$, then and.

Proof

Firstly, $\forall x \in D_B^L$, then $I_L^B(x) \subseteq D$ holds. For $0 < \delta \leq 1$, we have $I_{L\delta}^B(x) \subseteq I_L^B(x)$. Thus, $I_{L\delta}^B(x) \subseteq D$, i.e., $x \in D_B^{L\delta}$. Hence $D_B^{L\delta} \supseteq D_B^L$ holds.

Secondly, since $I_{L\delta}^B(x) \cap D \neq \phi$ for $\forall x \in D_{L\delta}^B$, and also we have $I_{L\delta}^B(x) \subseteq I_L^B(x)$, then $I_L^B(x) \cap D \neq \phi$, i.e., $x \in D_L^B$. Thus, $D_{L\delta}^B \subseteq D_L^B$ holds.

Therefore, from Proposition 3, we conclude that the proposed relative tolerance relation with relatively precision is an improved approach of limited tolerance relation in incomplete information systems.

Results, Analysis, and Discussion

In this section, we compare the proposed Relative Tolerance Relation of Rough Set (RTRS) with the existing baseline approaches i.e. Tolerance Relation (TR), Limited Tolerance Relation (LTR), and Non-Symmetric Similarity Relation (NSSR) approaches based on accuracy in terms of flexibility.

A real-world dataset that contains incomplete missing values is used. This dataset was obtained from the Directorate of Information Systems (SISFO), Telkom University. It contains 1250 instances and 8 categorical attributes. The attributes are a Student ID, National exam score (NES), university entrance exam score (UEES), 1st GPA, 2nd GPA, 3rd GPA, 4th GPA and probability of graduating on time, respectively. Here, irrelevant attributes such as name, gender, student residential address have been removed. The occurrence of missing values might be due to several possibilities, such as the student was on leave, the unofficial results of GPA, the student is not enrolled in certain semester. The probability field represents the probability of graduating on time. The description of each attribute of the dataset is shown in Table 2 as follows:

The values of GPA are in the form of letter representation of their actual numeric score (4.0 scale). The conversion of the actual score of GPAs to a letter representation based on a standard that is implemented by Telkom University is as shown in Table 3.

The sample of 10 out of 1250 of student data that are used as a dataset in this paper is shown in Table 4 as follows:

We will first recall the notion of accuracy. The accuracy in term of is defined as follows (Table 5):

TABLE 2. Description of dataset attributes

Attribute name	Description	Attribute set value
ID	ID of student	$\{1, 2, 3, \ldots, 1250\}$
NES	Letter representation of national exam score (NES)	$\{A, AB, B, BC, C, D, E\}$
UEES	Letter representation of university entrance exam score (UEES)	$\{A, AB, B, BC, C, D, E\}$
1st GPA	Letter representation of GPA of the student in first semester	$\{A, AB, B, BC, C, D, E\}$
2nd GPA	Letter representation of GPA of the student in second semester	$\{A, AB, B, BC, C, D, E\}$
3rd GPA	Letter representation of GPA of the student in third semester	$\{A, AB, B, BC, C, D, E\}$
4th GPA	Letter representation of GPA of the student in fourth semester	$\{A, AB, B, BC, C, D, E\}$
Probability	Probability of graduated on time	$\{0\%, 25\%, 50\%, 75\%, 100\%\}$

TABLE 3. Conversion of GPA

Range of GPA	GPA Letter	Category
3.51-4.00	A	Excellent
3.01-3.50	AB	Very Good
2.51-3.00	B	Good
2.01-2.50	BC	Fair
1.51-2.00	C	Satisfactory
1.10-1.50	D	Passing
0.00-1.00	E	Poor

TABLE 4. Sample of dataset

ID	1st	2nd	3rd	4th	5th	6th	Performance
1	B	B	AB	B	*	*	75%
2	B	BC	BC	AB	AB	A	75%
3	B	BC	BC	*	BC	B	50%
4	B	C	C	D	B	*	50%
5	D	C	E	C	*	*	0%
6	B	B	AB	C	AB	C	50%
7	D	C	C	BC	D	BC	25%
8	BC	B	B	B	*	*	75%
9	AB	B	AB	AB	AB	A	100%
10	A	A	AB	B	*	*	100%

TABLE 5. The accuracy measurement of each approach

Approaches	Accuracy measurement	Description
Tolerance realtion	$\frac{x_B^T = \left\{x \mid x \in U \wedge I_B^T(x) \subseteq X\right\}}{x_T^B = \left\{x \mid x \in U \wedge I_B^T(x) \cap X \neq \varphi\right\}}$	Definition 3
Non-symmetric similarity relation	$\frac{X_B^S = \left\{x \mid x \in U \wedge \mathrm{Sim}_B^{-1}(x) \subseteq X\right\}}{X_S^B = \bigcup\left\{\mathrm{Sim}_B(x) \mid x \in X\right\}}$	Definition 7
Limite tolerance relation	$\frac{D_L^B = \left\{x \mid x \in U \wedge I_B^L(x) \cap D \neq \varphi\right\}}{D_B^L = \left\{x \mid x \in U \wedge I_B^L(x) \subseteq D\right\}}$	Definition 10
Relative tolerance relation of rough set	$\frac{D_B^{L\delta} = \left\{x \mid x \in U \wedge I_B^{L\delta}(x) \subseteq D\right\}}{D_{L\delta}^B = \left\{x \mid x \in U \wedge I_B^{L\delta}(x) \cap D \neq \varphi\right\}}$	Definition15

In the experimentation, the proposed approach and other three baseline approaches are implemented in MATLAB version 8.3.0.532 (R2014a). They are executed sequentially on a processor Intel Core i5-6200U Processor 2.30 GHz CPUs. The total main memory is 4GB and the operating system is Windows 10. The computation results comparing all four (4) techniques in terms of accuracy are shown in Figure 1.

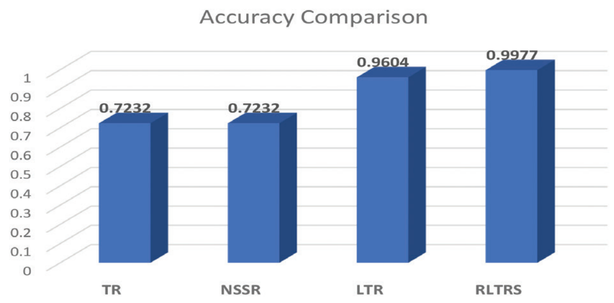

FIGURE 1. Accuracy comparison of each technique

To handle incomplete information systems, many approaches have been proposed but those approaches have never been tested in real datasets. In this study, the researcher proposed an RTRS method based on rough set theory which to cope with incomplete information systems. Furthermore, the researcher has tested the proposed approach and compared with the existing approaches in real dataset, incomplete data in student dataset. Based on the experiments that have been done, the proposed method has successfully handled the problem of incomplete information systems in real data. Based on Figure 1, the results have shown that the proposed approach outperforms as compared to the existing approaches in terms of flexibility and precision of accuracy, which achieved 99.7% of accuracy. At this stage of the research, we show the relative tolerance relation of rough set can be used to classify incomplete student dataset.

Conclusion

A common phenomenon in real-world problems is missing values data. Knowing how to handle missing values is important since the data insights or the performance of the predictive model could be impacted if the missing values are not appropriately handled. To overcome these problems, this paper introduces a relative tolerance relation of rough set (RTRS). The proposed approach is based on limited tolerance relation by taking account into consideration the relatively precision between two objects and therefore this is the first work that uses relatively precision. The results show that the proposed approach outperforms as compared with the existing approaches in terms of flexibility and precision of accuracy. The result shows relative tolerance relation of rough set can be used to classify incomplete student dataset. The results may potentially contribute to give insight/knowledge of the incomplete dataset. By knowing the knowledge from the dataset, during their period of studies in an institution, a strategic program can be planned to prevent student failure and early intervention make much more sense than remediation.

ACKNOWLEDGEMENTS

Authors would like to thanks Universiti Tun Hussein Onn Malaysia for research grant Vote Number U558.

REFERENCES

Borkar, S. & Rajeswari, K. 2013. Predicting students academic performance using education data mining. *IJCSMC International Journal of Computer Science and Mobile Computing* 2(7): 273-279.

Bunting, B.P., Adamson, G. & Mulhall, P.K. 2002. A Monte Carlo examination of an MTMM model with planned incomplete data structures. *Structural Equation Modeling* 9(3): 369-389.

Chiroma, H., Abdulkareem, S., Muaz, S.A., Abubakar, A.I., Sutoyo, E., Mungad, M., Younes, Saadi., Eka, Novita, Sari. & Herawan, T. 2015. An intelligent modeling of oil consumption. *Advances in Intelligent Systems and Computing* 320: 557-568.

Chmielewski, M.R., Grzymala-Busse, J.W., Peterson, N.W. & Than, S. 1993. The rule induction system LERS-a version for personal computers. *Foundations of Computing and Decision Sciences* 18(3-4): 181-212.

Dobrota, M., Bulajić, M. & Radojičić, Z. 2014. Data mining models for prediction of customers' satisfaction: The CART analysis. In *Innovative Management and Firm Performance,* edited by Jakšić, M.L., Rakočević, S.B. & Martić, M. London: Palgrave Macmillan. pp. 401-421.

Fayyad, U.M. 1996. Data mining and knowledge discovery: Making sense out of data. *IEEE Expert: Intelligent Systems and Their Applications* 11(5): 20-25.

Ibrahim, Z. & Rusli, D. 2007. Predicting students' academic performance: Comparing artificial neural network, decision tree and linear regression. *21st Annual SAS Malaysia Forum, 5th September*.

Kotsiantis, S., Pierrakeas, C. & Pintelas, P. 2004. Predicting students'performance in distance learning using machine learning techniques. *Applied Artificial Intelligence* 18(5): 411-426.

Kryszkiewicz, M. 1999. Rules in incomplete information systems. *Information Sciences* 113(3): 271-292.

Kryszkiewicz, M. 1998. Rough set approach to incomplete information systems. *Information Sciences* 112(1): 39-49.

Márquez-Vera, C., Cano, A., Romero, C. & Ventura, S. 2013. Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied Intelligence* 38(3): 315-330.

Minaei-Bidgoli, B., Kashy, D.A., Kortemeyer, G. & Punch, W.F. 2003. Predicting student performance: An application of data mining methods with an educational web-based system. *Proceedings-Frontiers in Education Conference 2003* 1: 13-18.

Mohammed, M.A.T., Mohd, W.M.W., Arshah, R.A., Mungad, M., Sutoyo, E. & Chiroma, H. 2016. Analysis of parameterization value reduction of soft sets and its algorithm. *International Journal of Software Engineering and Computer Systems* 2(1): 51-57.

Ogunde, A.O. & Ajibade, D.A. 2014. A data mining system for predicting university students' graduation grades using ID3 decision tree algorithm. *Journal of Computer Science and Information Technology* 2(1): 21-46.

Pal, S. 2012. Mining educational data to reduce dropout rates of engineering students. *International Journal of Information Engineering and Electronic Business* 4(2): 1. Doi: 10.5815/ijieeb.2012.02.01.

Romero, C. & Ventura, S. 2007. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications* 33(1): 135-146.

Saedudin, R.R., Kasim, S., Mahdin, H., Sutoyo, E., Riyadi Yanto, I.T., Hassan, R. & Ismail, M.A. 2018. A relative tolerance relation of rough set (RTRS) for potential fish yields in Indonesia. *Journal of Coastal Research: Special Issue 82 - Coastal Ecosystem Responses to Human and Climatic Changes throughout Asia.* pp. 84-92.

Saedudin, R.R., Sutoyo, E., Kasim, S., Mahdin, H. & Yanto, I.T.R. 2017a. A comparative analysis of rough sets for incomplete information system in student dataset. *International Journal on Advanced Science, Engineering and Information Technology* 7(6): 2078-2084.

Saedudin, R.R., Sutoyo, E., Kasim, S., Mahdin, H. & Yanto, I.T.R. 2017b. Attribute selection on student performance dataset using maximum dependency attribute. *Electrical, Electronics and Information Engineering (ICEEIE), 2017 5th International Conference*. pp. 176-179.

Saedudin, R.R., Kasim, S.B., Mahdin, H. & Hasibuan, M.A. 2016. Soft set approach for clustering graduated dataset. *International Conference on Soft Computing and Data Mining*. pp. 631-637.

Slavin, R.E., Karweit, N.L. & Wasik, B.A. 1994. *Preventing Early School Failure: Research, Policy, and Practice*. Boston: Allyn & Bacon.

Stefanowski, J. & Tsoukias, A. 2001. Incomplete information tables and rough classification. *Computational Intelligence* 17(3): 545-566.

Stefanowski, J. & Tsoukiàs, A. 1999. On the extension of rough sets under incomplete information. *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing*. pp. 73-81.

Sutoyo E., Yanto, I.T.R., Saadi, Y., Chiroma, H., Hamid, S. & Herawan, T. 2019. A framework for clustering of web users transaction based on soft set theory. In *Proceedings of the International Conference on Data Engineering 2015 (DaEng-2015). Lecture Notes in Electrical Engineering*, edited by Abawajy, J., Othman, M., Ghazali, R., Deris, M., Mahdin H. & Herawan T. Singapore: Springer. 520: 307-314.

Sutoyo, E., Yanto, I.T.R., Saedudin, R.R. & Herawan, T. 2017a. A soft set-based co-occurrence for clustering web user transactions. *Telkomnika (Telecommunication Computing Electronics and Control)* 15(3): 1344-1353.

Sutoyo, E., Saedudin, R.R., Yanto, I.T.R. & Apriani, A. 2017b. Application of adaptive neuro-fuzzy inference system and chicken swarm optimization for classifying river water quality. *Electrical, Electronics and Information Engineering (ICEEIE), 2017 5th International Conference*. pp. 118-122.

Van Nguyen, D., Yamada, K. & Unehara, M. 2013. Extended tolerance relation to define a new rough set model in incomplete information systems. *Advances in Fuzzy Systems* 2013: 37209.

Wang, G. 2002. Extension of rough set under incomplete information systems. *Proceedings of the 2002 IEEE International Conference* 2: 1098-1103.

Wu, Y. & Guo, Q. 2010. An extension model of rough set in incomplete information system. *Future Computer and Communication (ICFCC), 2010 2nd International Conference* 2: 434-438.

Yadav, S.K., Bharadwaj, B. & Pal, S. 2012. Mining education data to predict student's retention: A comparative study. *International Journal of Computer Science and Information Security* 10(2): 113-117.

Yadav, S.K. & Pal, S. 2012. Data mining: A prediction for performance improvement of engineering students using classification. *World of Computer Science and Information Technology Journal WCSIT* 2(2): 51-56.

Yang, X. 2009. An improved model of rough sets on incomplete information systems. *Management of e-Commerce and e-Government, 2009. ICMECG'09. International Conference*. pp. 193-196.

Yang, X., Song, X. & Hu, X. 2011. Generalisation of rough set for rule induction in incomplete system. *International Journal of Granular Computing, Rough Sets and Intelligent Systems* 2(1): 37-50.

Yanto, I.T.R., Saedudin, R.R., Hartama, D. & Herawan, T. 2016. Clustering based on classification quality (CCQ). *International Conference on Soft Computing and Data Mining*. pp. 327-335.

Yanto, I.T.R., Saedudin, R.R., Lashari, S.A. & Haviluddin. 2018a. A numerical classification technique based on fuzzy soft set using hamming distance. *International Conference on Soft Computing and Data Mining*. pp. 252-260.

Yanto, I.T.R., Sutoyo, E., Apriani, A. & Verdiansyah, O. 2018b. Fuzzy soft set for rock igneous clasification. *2018 International Symposium on Advanced Intelligent Informatics (SAIN)*. pp. 199-203.

Zhou, J. & Yang, X. 2012. Rough set model based on hybrid tolerance relation. *International Conference on Rough Sets and Knowledge Technology*. pp. 28-33.

Zhou, Q. 2010. Research on tolerance-based rough set models. *System Science, Engineering Design and Manufacturing Informatization (ICSEM), 2010 International Conference* 2: 137-139.

Rd Rohmat Saedudin* & Edi Sutoyo
School of Industrial Engineering
Telkom University
40257, Bandung, West Java
Indonesia

Shahreen Kasim, Hairulnizam Mahdin &
Mohd Farhan Md Fudzee
Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia
86400 Batu Pahat, Johor Darul Takzim
Malaysia

Iwan Tri Riyadi Yanto
Department of Information Systems
Universitas Ahmad Dahlan
55161, Yogyakarta
Indonesia

Rohayanti Hassan
Faculty of Computing
Universiti Teknologi Malaysia
81310 Skudai, Johor Darul Takzim
Malaysia

*Corresponding author; email: rdrohmat@telkomuniversity.ac.id

Received: 21 February 2019
Accepted: 25 December 2019