Soft Comput (2016) 20:3153–3161
DOI 10.1007/s00500-016-2243-6

FOUNDATIONS

# Maximal limited similarity-based rough set model

Ahmed Hamed Attia[1] · Ahmed Sobhy Sherif[1] · Ghada Samy El-Tawel[1]

Published online: 30 June 2016

**Abstract** Non-symmetric similarity relation-based rough set model (NS-RSM) is viewed as mathematical tool to deal with the analysis of imprecise and uncertain information in incomplete information systems with "?" values. NS-RSM relies on the concept of non-symmetric similarity relation to group equivalent objects and generate knowledge granules that are then used to approximate the target set. However, NS-RSM results in unpromising approximation space when addressing inconsistent data sets that have lots of boundary objects. This is because objects in the same similarity classes are not necessarily similar to each other and may belong to different target classes. To enhance NS-RSM capability, we introduce the maximal limited similarity-based rough set model (MLS-RSM) which describes the maximal collection of indistinguishable objects that are limited tolerance to each other in similarity classes. This allows accurate computation to be done for the approximation space. Furthermore, approximation accuracy comparisons have been conducted among NS-RSM and MLS-RSM. The results demonstrate that MLS-RSM model outperforms NS-RSM and can approximate the target set more efficiently.

**Keywords** Incomplete information system · Non-symmetric similarity relation-based rough set · Limited tolerance relation

## 1 Introduction

Uncertainty in the data degrades the process of analyzing the data leading to uncertain conclusions (Gantayat et al. 2014). Classical rough set theory (RST) proposed by Pawlak (1982) made a great success in processing and analyzing complete information systems characterized by uncertainty (Qin et al. 2015). Reduct (Jensen et al. 2014; Jiang and Yu 2015) is the most important issue of RST which eliminates unnecessary attributes and creates a minimal sufficient subset of attributes for decision table. Approximation accuracy measure (Dai and Xu 2012) that measures the imprecision of approximation space is employed as heuristics measure to guide the reduct process. Consequently, for a minimal feasible subset of features, the approximation accuracy should be maximal. RST uses the lower and upper approximations that are defined using the indiscernibility relation to define the approximation space. The indiscernibility relation is considered equivalent relation because it is reflexive, symmetric, and transitive. However, the indiscernibility relation is a rigid relation (Huang et al. 2014) because it is based on the assumption that all object's values for every attribute are known. This assumption contrasts with several real-valued information systems situations where the information may be incomplete (DU and ZI 2014). This limits the applicability of RST in real-world applications where some of the attribute values are unknown. For RST to deal with incomplete information systems (IIS), two strategies are proposed. The first one is an indirect method that transforms the IIS into a complete one according to some rules such as

Communicated by A. Di Nola.

Ahmed Hamed Attia
ahmed_hamed@ci.suez.edu.eg

Ahmed Sobhy Sherif
ahmed_sobhy@ci.suez.edu.eg

Ghada Samy El-Tawel
ghada@ci.suez.edu.eg

[1] Computer Science Department, Faculty of Computers and Informatics, Suez Canal university, Ismailia, Egypt

probability statistical methods; this is called data preparation (Wang 2001; Grzymala-Busse and Hu 2005). However, this may cause loss in the original information leading to uncertain conclusions. The second one is a direct method that extends the basic concepts of RST under IIS by relaxing the requirement of indiscernibility relation of reflexivity, symmetry and transitivity (Kryszkiewicz 1998, 1999; Stefanowski and Tsoukiàs 1999, 2001; Skowron and Stepaniuk 1996; Wang 2002; Wang et al. 2008; Leung and Li 2003; Cheng et al. 2007; Yin and XiuyiJia 2006; Yang 2009; Yang and Yang 2012; Nguyen et al. 2013; Grzymala-Busse and Wang 1997; Liu and Shao 2014; Huang and Li 2014).

The unknown values were categorized by Grzymała-Busse (2004) as follows:

1. The unknown values are "lost ?", this unknown value means that it is an absent value and cannot be compared with any other values in the domain of the corresponding attribute.
2. The unknown values are "don't care *," this unknown value means that it is a missing value and can be compared with any other values in the domain of the corresponding attribute.

In recent years, many researchers extended RST model by replacing the indiscernibility relation by another non-equivalent relation to process IIS directly. For example, Kryszkiewicz (1998, 1999) proposed the tolerance relation that is reflexive and symmetric but not necessarily transitive to deal with "*" values. In this relation, objects that have no values in common are considered indistinguishable. For example, with two objects $X = \{1, *, 3, *\}$ and $Y = \{*, 2, *, 4\}$ then according to the tolerance relation, object $X$ is considered indistinguishable from object $Y$ and will be gathered in the same tolerance class. Obviously, this is unreasonable case which limits the applicability of the tolerance relation. Stefanowski and Tsoukiàs (1999, 2001) investigated the similarity relation of Skowron and Stepaniuk (1996) to be reflexive and transitive but not necessarily symmetric to deal with "?" values. This relation separates two objects that are very similar to each other but with little loss in the information. At the same time, objects in the same similarity class are not necessarily similar to each other and may belong to different target classes. This excludes some objects from the lower approximation of the target set. For example, with four objects $X = \{a, ?, c, d, ?, d_1\}$, $Y = \{?, s, c, d, ?, d_1\}$, $Z = \{a, b, c, d, e, d_2\}$ and $W = \{a, s, c, d, f, d_1\}$ then according to the non-symmetric similarity relation $R^{-1}(X) = \{X, Z, W\}$. Obviously, object $Y$ is not included in the same similarity class of object $X$ in spite of the fact two attribute values are perceived as common, at the same time objects $X$, $Z$ and $W$ belong to the same similarity class in spite of the fact that objects $Z$ and $W$ are not similar, this lead to $R^{-1}(X) \nsubseteq d_1$ which prevents objects $X$ and $Z$ from being included in the lower approximation of $d_1$. This decreases the cardinality of the lower approximation leading to unpromising results with respect to approximation accuracy. Wang et al. (2008) recognized that the tolerance relation requirement is too weak as it regards two objects with no common values as indistinguishable, also, the requirement of the NS-RSM is too strict as it separates two objects that are very similar to each other but with little loss in the information. This makes the process too extreme. Consequently, Wang proposed the limited tolerance relation that tries to relax the requirements of both tolerance relation and NS-RSM. Limited tolerance relation is reflexive and symmetric but not necessarily transitive. However, limited tolerance relation has not differentiated the two types of unknown attribute values. Leung and Li (2003) proposed the maximal consistent block relation that is reflexive and symmetric but not necessarily transitive to deal with "*" values. Maximal consistent block relation describes the maximal collection of indistinguishable objects in the tolerance classes (Cheng et al. 2007). This relation achieves better approximation accuracy than that provided by tolerance relation, but it inherits the limitation of the tolerance relation where objects that have no values in common are considered indistinguishable. Yin and XiuyiJia (2006) proposed the constrained dis-symmetrical similarity relation that is reflexive and transitive but not necessarily symmetric to deal with "?" values. This relation extends the non-symmetrical similarity relation based on the limited tolerance relation. In this relation, two objects are considered to be similar due to high loss in the information of one of the two objects. For example, with two objects $A = \{3, 2, 1, 0, d_1\}$ and $B = \{?, 2, ?, ?, d_2\}$ then according to dis-symmetrical similarity relation, object $A$ is considered similar to object $B$ which prevents object $A$ from being included in the lower approximation of $d_1$ in spite of the fact it contains no "?" values. Yang (2009), Yang and Yang (2012) proposed the difference relation that is not necessarily reflexive, symmetric and transitive to deal with "?" values. The requirement of the difference relation is too strict as it separates two objects that are very dissimilar but with a slight bit of similarity. For example, with two objects $S = \{?, ?, 3, 4, 5\}$ and $D = \{?, 6, 7, 8, 5\}$ then according to the difference relation, object $S$ is not considered dissimilar to object $D$ because value 5 is common in both objects. This may not allow the difference relation to approximate the target set whenever all tuples are similar to each other, even so, with partial similarity. Nguyen et al. (2013) proposed a parametric relation that extends the non-symmetric similarity relation by computing the probability of matching for each attribute. This relation is reflexive and symmetric but not necessarily transitive. This relation needs in advance threshold ($\alpha$) to control the tolerance degree.

Słowiński et al. (2014) reported that up till now, NS-RSM is the only RST extension that correctly characterizes the target set as it computes the lower/upper approximations using two different relations. This is because of the fact that NS-RSM does not partition the data, instead it uses the similarity classes, but objects in the same similarity class are not necessarily similar to each other which leads to unpromising results with respect to approximation accuracy. Consequently, the aim of this paper is to enhance the capability of NS-RSM to provide promising results with respect to approximation accuracy which can be further used to provide promising results with respect to reduct. In this paper, we propose the maximal limited similarity-based rough set model (MLS-RSM) that is a modified version of NS-RSM able to provide promising approximation accuracy under IIS with "?" values. MLS-RSM finds the maximal limited consistent blocks of similar objects for each similarity class in NS-RSM. Maximal limited consistent blocks describes the maximal collection of indistinguishable objects that are limited tolerance to each other in the similarity classes. This ensures that objects in the same block are similar to each other leading to accurate computation to be done for the lower approximation. Furthermore, approximation accuracy comparisons have been conducted among NS-RSM and MLS-RSM. The results demonstrate that MLS-RSM model outperforms NS-RSM model.

The rest of this paper is organized as follows. Section 2 introduces some preliminaries on basic concepts and rough set theory. Some related RST extensions under IIS are reviewed in Sect. 3. In Sect. 4, we propose the maximal limited similarity-based rough set model (MLS-RSM) model. In Sect. 5, approximation accuracy comparisons among NS-RSM and MLS-RSM have been conducted. Section 6 concludes the paper.

## 2 Preliminaries

### 2.1 Basic concepts

For the convenience of discussion, some basic notions and relevant concepts are introduced at first.

**Definition 1** An information system is a quadruple IS $= (U, \mathrm{AT}, V, f)$, where, $U$ is a non-empty finite set of objects; $AT$ is a non-empty finite set of features; $V$ is the union of attribute domains, $V = \bigcup_{a \in A} V_{\mathrm{a}}$, where, $V_{\mathrm{a}}$ is the value set of attribute $a$; $f : U \times \mathrm{AT} \rightarrow V$ is the information function that assigns a particular value of the object attributes.

If there exist $x \in U$ such that $f_a(x)$ equals to an unknown value where $a \in \mathrm{AT}$, then the information system is said to be incomplete information system (IIS).

### 2.2 Rough set theory

The classical RST uses the indiscernibility relation that is given by $E(A) = \{(x, y) \in U^2 : \forall a \in A, f_a(x) = f_a(y)\}$, where $A \subseteq \mathrm{AT}$. The equivalence classes are given by $[x]_E = \{y \in U : xEy\}$. The lower/upper approximations of a set $X$ under RST are given respectively by, $\underline{\mathrm{Apr}}(X) = \bigcup\{[x]_E : [x]_E \subseteq X\}$, $\overline{\mathrm{Apr}}(X) = \bigcup\{[x]_E : [x]_E \bigcap X \neq \phi\}$. The pair $(\underline{\mathrm{Apr}}(X), \overline{\mathrm{Apr}}(X))$ is called the approximation space. The positive, negative and the boundary regions are defined respectively by, $\mathrm{POS}(X) = \underline{\mathrm{Apr}}(X)$, $\mathrm{NEG}(X) = U - \overline{\mathrm{Apr}}(X)$, $\mathrm{BND}(X) = \overline{\mathrm{Apr}}(X) - \underline{\mathrm{Apr}}(X)$.

### 2.3 Approximation accuracy measure

Approximation and reduction are two key issues in RST. The approximation refers to approximately describe a subset of the universe with respect to a given set of features. While the reduction is the process of finding a minimal subset of features that preserve the discriminatory power as the whole features using the approximation concept. Consequently, the finer the approximation, the finer the reduction. The approximation accuracy measure $\tau$ (Dai and Xu 2012) is uncertainty measure that measures the imprecision of rough approximation. This measure reflects RST ability to find the minimal subset of features that preserve the discriminatory power as the whole features. The form of approximation accuracy is given by

$$\tau = \frac{|\underline{\mathrm{Apr}}(X)|}{|\overline{\mathrm{Apr}}(X)|} \tag{1}$$

where $|\cdot|$ denotes cardinality. Clearly, the greater the approximation accuracy, the greater the characterizing power of the available features. Consequently, for a minimal subset of features, the approximation accuracy should be maximal.

## 3 Some RST extensions under IIS

In this section, we review some related RST extensions under IIS with their issues.

### 3.1 Non-symmetric similarity relatio-based rough set model (NS-RSM)

Stefanowski and Tsoukiàs (1999, 2001) redefined the similarity relation of Skowron and Stepaniuk (1996) to deal with the "?" values as Definition 2.

**Definition 2** Given $IIS$ in which $A \subseteq \mathrm{AT}$, the non-symmetric similarity relation is given by

$$\forall_{x,y} S(x, y) \Longleftrightarrow \forall_{a \in A} f_a(x) \neq ?,\ f_a(x) = f_a(y) \tag{2}$$

**Table 1** Small illustrative example

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ | $x_9$ | $x_{10}$ | $x_{11}$ | $x_{12}$ | $x_{13}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $a_1$ | 3 | 2 | 2 | ? | ? | 2 | 3 | 2 | 3 | 1 | 3 | 3 | 3 |
| $a_2$ | 2 | 3 | 3 | 2 | 2 | 3 | ? | 3 | 2 | ? | 2 | 2 | 2 |
| $a_3$ | 1 | 2 | 2 | ? | ? | 2 | ? | 2 | 1 | ? | ? | 1 | 2 |
| $a_4$ | ? | ? | 0 | 1 | 1 | 1 | 3 | 2 | 3 | ? | ? | 1 | 3 |
| $d$ | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 2 | 1 | 2 |

For each object, two similarity classes are defined as follows:

$$R(x) = \{y \in U : S(y, x)\} \tag{3}$$

$$R^{-1}(x) = \{y \in U : S(x, y)\} \tag{4}$$

where, $R(x)$ represents the set of objects similar to $x$ and $R^{-1}(x)$ represents the set of objects to which $x$ is similar.

The lower/upper approximations of set $X$ under NS-RSM are given respectively as follows:

$$\underline{R}_A^{-1}(X) = \{x \in U : R^{-1}(x) \subseteq X\} \tag{5}$$

$$\overline{R}_A(X) = \bigcup \{R(x) : x \in X\} \tag{6}$$

Słowiński et al. (2014) stated that the discrimination task between objects using indiscernibility relation is difficult due to the imprecision of data describing the objects. This situation is considered perfectly using non-symmetric similarity relation.

In the following we will illustrate NS-RSM with a small IIS shown in Table 1.

*Example 1* Consider the IIS presented in Table 1, where $U = \{x_1, x_2, ..., x_{13}\}$, $A = \{a_1, a_2, a_3, a_4\}$ with 25 % lost ("?") values and $d$ is the decision attribute that determines a partition on the universe such that $U/d = \{d_1, d_2\} = \{\{x \in U : f_d(x) = 1\}, \{x \in U : f_d(x) = 2\}\}$.

Thus,

$$\begin{array}{ll}
R^{-1}(x_1) = \{x_1, x_9, x_{12}\} & R(x_1) = \{x_1, x_{11}\} \\
R^{-1}(x_2) = \{x_2, x_3, x_6, x_8\} & R(x_2) = \{x_2\} \\
R^{-1}(x_3) = \{x_3\} & R(x_3) = \{x_2, x_3\} \\
R^{-1}(x_4) = \{x_4, x_5, x_{12}\} & R(x_4) = \{x_4, x_5\} \\
R^{-1}(x_5) = \{x_4, x_5, x_{12}\} & R(x_5) = \{x_4, x_5\} \\
R^{-1}(x_6) = \{x_6\} & R(x_6) = \{x_2, x_6\} \\
R^{-1}(x_7) = \{x_7, x_9, x_{13}\} & R(x_7) = \{x_7\} \\
R^{-1}(x_8) = \{x_8\} & R(x_8) = \{x_2, x_8\} \\
R^{-1}(x_9) = \{x_9\} & R(x_9) = \{x_1, x_7, x_9, x_{11}\} \\
R^{-1}(x_{10}) = \{x_{10}\} & R(x_{10}) = \{x_{10}\} \\
R^{-1}(x_{11}) = \{x_1, x_9, x_{11}, x_{12}, x_{13}\} & R(x_{11}) = \{x_{11}\} \\
R^{-1}(x_{12}) = \{x_{12}\} & R(x_{12}) = \{x_1, x_4, x_5, x_{11}, x_{12}\} \\
R^{-1}(x_{13}) = \{x_{13}\} & R(x_{13}) = \{x_7, x_{11}, x_{13}\}
\end{array}$$

Obviously, NS-RSM separates objects that are very similar to each other but with little loss in the information, for example, in $R^{-1}(x_3)$ object $x_2$ is not considered similar to object $x_3$ in spite of the fact three of their attribute values are common. At the same time, objects are included in the same similarity class in spite of the fact they are not similar to each other. For example, in $R^{-1}(x_1) = \{x_1, x_9, x_{12}\}$ object $x_9$ and object $x_{12}$ are not similar, this leads to $R^{-1}(x_1) \nsubseteq d_1$ which prevents object $x_1$ from being included in the lower approximation of $d_1$. In $R^{-1}(x_2) = \{x_2, x_3, x_6, x_8\}$ objects $x_3$, $x_6$ and $x_8$ are not similar, this leads to $R^{-1}(x_2) \nsubseteq d_1$ which prevents object $x_2$ from being included in the lower approximation of $d_1$. This shrinks the lower approximation leading to unpromising results with respect to approximation accuracy.

The lower/upper approximations are as follows:

$$\underline{R}_A^{-1}(d_1) = \{x_6, x_{10}, x_{12}\}$$

$$\overline{R}_A(d_1) = \{x_1, x_2, x_4, x_5, x_6, x_7, x_{10}, x_{11}, x_{12}\}$$

$$\underline{R}_A^{-1}(d_2) = \{x_3, x_8, x_9, x_{13}\}$$

$$\overline{R}_A(d_2) = \{x_1, x_2, x_3, x_4, x_5, x_7, x_8, x_9, x_{11}, x_{13}\}$$

Consequently, $\tau(d_1) = \frac{3}{9} = \frac{1}{3}$ and $\tau(d_2) = \frac{4}{10} = \frac{2}{5}$.

### 3.2 Limited tolerance relation

Wang et al. (2008) recognized that the requirement of NS-RSM is too strict as it separates two objects that are very similar to each other but with little loss in the information. This makes the process too extreme. Consequently, Wang proposed the limited tolerance relation that tries to relax the requirements of NS-RSM as Definition 3.

**Definition 3** Given $IIS$ in which $A \subseteq \text{AT}$, the limited tolerance relation is given by

$$\begin{aligned}
\forall_{x,y \in U \times U}(\text{LT}_A(x, y) \Longleftrightarrow & \forall_{a \in A}(f_a(x) = f_a(y) = \text{unknown}) \\
& \vee ((P_A(x) \bigcap P_A(y) \neq \phi) \\
& \wedge \forall_{a \in A}(f_a(x) \neq \text{unknown} \\
& \wedge f_a(y) \neq \text{unknown}) \\
& \rightarrow (f_a(x) = f_a(y))))
\end{aligned} \tag{7}$$

where $P_A(x) = \{a \in A : f_a(x)\, is\, known\ value\}$ and the unknown values can be interpreted as * or ? values.

The limited tolerance classes of $x$ is denoted by $I_A(x)$ where

$$I_A(x) = \{y : y \in U \wedge \text{LT}_A(x, y)\} \tag{8}$$

The lower and upper approximations of set $X$ under limited tolerance relation are given, respectively, as follows:

$$\underline{I}_A(X) = \{x : x \in U \wedge I_A(x) \subseteq X\} \tag{9}$$

$$\overline{I}_A(X) = \{x : x \in U \wedge I_A(x) \bigcap X \neq \phi\} \tag{10}$$

**Continue with Example** 1. we illustrate Definition 3. Thus,

$$\begin{array}{ll}
I_A(x_1) = \{x_1, x_4, x_5, x_7, x_9, & \\
\quad x_{11}, x_{12}\} & I_A(x_8) = \{x_2, x_8\} \\
I_A(x_2) = \{x_2, x_3, x_6, x_8\} & I_A(x_9) = \{x_1, x_7, x_9, x_{11}\} \\
I_A(x_3) = \{x_2, x_3\} & I_A(x_{10}) = \{x_{10}\} \\
I_A(x_4) = \{x_1, x_4, x_5, x_{11}, x_{12}\} & I_A(x_{11}) = \{x_1, x_4, x_5, x_7, x_9, \\
 & \quad x_{11}, x_{12}, x_{13}\} \\
I_A(x_5) = \{x_1, x_4, x_5, x_{11}, x_{12}\} & I_A(x_{12}) = \{x_1, x_4, x_5, x_{11}, x_{12}\} \\
I_A(x_6) = \{x_2, x_6\} & I_A(x_{13}) = \{x_7, x_{11}, x_{13}\} \\
I_A(x_7) = \{x_1, x_7, x_9, x_{11}, x_{13}\} &
\end{array}$$

Obviously, limited tolerance relation relaxed the requirement of NS-RSM, for example in $\mathrm{LT}_A(x_3) = \{x_2, x_3\}$, object $x_2$ is considered indistinguishable from object $x_3$ which is not the case in $R^{-1}(x_3)$. However, limited tolerance relation failed to relax the requirement of gathering non-similar objects in the same limited tolerance class, for example, in $I_A(x_2) = \{x_2, x_3, x_6, x_8\}$ objects $x_3$, $x_6$ and $x_8$ are not similar to each other.

The lower/upper approximations are as follows:

$$\underline{I}_A(d_1) = \{x_2, x_6, x_{10}\},$$
$$\overline{I}_A(d_1) = U,$$
$$\underline{I}_A(d_2) = \phi,$$
$$\overline{I}_A(d_2) = U - x_{10}.$$

## 4 Maximal limited similarity-based rough set model (MLS-RSM)

As we discussed in Sect. 3.1, objects in the same similarity classes are not necessarily similar to each other and may belong to different target classes which increase uncertainty in the data. This excludes some objects from the lower approximation of the target set in spite of the fact they could be classified in the lower approximation leading to unpromising results with respect to approximation accuracy. In order to overcome this problem, we propose the maximal limited similarity-based rough set model (MLS-RSM) which finds the maximal limited consistent blocks of similar objects for each similarity class $(R(x), R^{-1}(x))$. Maximal limited consistent blocks describe the maximal collection of indistinguishable objects that are limited tolerance to each other in similarity classes. This reduces the uncertainty from the data allowing accurate computation to be done for the lower approximation leading to promising results with respect to approximation accuracy.

Similar to NS-RSM, for each object, two similarity classes are defined as Definitions 4 and 5, respectively.

**Definition 4** Given $IIS$ and $R(x)$ be set of objects similar to $x$ in which $A \subseteq \mathrm{AT}$ and $X \subseteq R(x)$, we say that $X$ is limited consistent block of objects similar to $x$ with respect to $A$ if $(x, y) \in S(y, x)$ and $\forall y, z \in X, (y, z) \in \mathrm{LT}_A(y, z)$. We say that $X$ is maximal limited consistent block of objects similar to $x$ if there does not exist $Y \subseteq R(x)$ where $X \subset Y$ and $Y$ is limited consistent with respect to $A$. We denote the maximal limited consistent blocks of all $R(x)$ as $\zeta^{\mathrm{LT}}_{R_A}$ and the maximal limited consistent blocks of objects similar to $x$ with respect to $A$ as $\zeta^{\mathrm{LT}}_{R_A(x)}$.

**Definition 5** Given $IIS$ and $R^{-1}(x)$ be set of objects to which $x$ is similar in which $A \subseteq \mathrm{AT}$ and $X \subseteq R^{-1}(x)$, we say that $X$ is limited consistent block of objects to which $x$ is similar with respect to $A$ if $(x, y) \in S(x, y)$ and $\forall y, z \in X, (y, z) \in \mathrm{LT}_A(y, z)$. We say that $X$ is maximal limited consistent block of objects to which $x$ is similar if there does not exist $Y \subseteq R^{-1}(x)$ where $X \subset Y$ and $Y$ is limited consistent with respect to $A$. We denote the maximal limited consistent blocks of all $R^{-1}$ as $\zeta^{\mathrm{LT}}_{R^{-1}_A}$ and the maximal limited consistent blocks of objects to which $x$ is similar with respect to $A$ as $\zeta^{\mathrm{LT}}_{R^{-1}_A(x)}$.

These relations ($\zeta^{\mathrm{LT}}_{R_A}$ and $\zeta^{\mathrm{LT}}_{R^{-1}_A}$) are reflexive but not necessarily transitive and symmetric.

The lower/upper approximations of set $X$ under MLS-RSM are given respectively as follows:

$$\underline{\zeta^{\mathrm{LT}}_{R^{-1}_A}}(X) = \bigcup \left\{Y \in \zeta^{\mathrm{LT}}_{R^{-1}}(A) : Y \subseteq X\right\} \tag{11}$$

$$\overline{\zeta^{\mathrm{LT}}_{R_A}}(X) = \bigcup \left\{Z \in \zeta^{\mathrm{LT}}_{R(x)}(A) : x \in X\right\} \tag{12}$$

Similar to the classical RST, the positive, negative and the boundary regions under MLS-RSM are given, respectively, as follows:

$$\mathrm{POS}(X) = \underline{\zeta^{\mathrm{LT}}_{R^{-1}_A}}(X), \tag{13}$$

$$\mathrm{NEG}(X) = U - \overline{\zeta^{\mathrm{LT}}_{R_A}}(X), \tag{14}$$

$$\mathrm{BND}(X) = \overline{\zeta^{\mathrm{LT}}_{R_A}}(X) - \underline{\zeta^{\mathrm{LT}}_{R^{-1}_A}}(X). \tag{15}$$

### 4.1 Properties of MLS-RSM

**Property 1** *Any similarity class $R^{-1}(x)$ of attributes subset $A$ can be represented as the union of maximal limited consistent blocks included in it. In other words,*

$$R^{-1}(x) = \bigcup \left\{ Y \in \zeta^{\mathrm{LT}}_{R_A^{-1}} : Y \subseteq R^{-1}(x) \right\}$$
$$= \bigcup \left\{ Y \in \zeta^{\mathrm{LT}}_{R_A^{-1}(x)} \right\}.$$

*Proof* Suppose that $R^{-1}(x_1) = \{x_1, x_2, x_3\}$, then one of the following is true with respect to MLS-RSM model.

1. $\zeta^{\mathrm{LT}}_{R_A^{-1}(x)} = \{Y_1 = \{x_1, x_2, x_3\}\}$
2. $\zeta^{\mathrm{LT}}_{R_A^{-1}(x)} = \{Y_1 = \{x_1, x_2\}, Y_2 = \{x_1, x_3\}\}$

Obviously, for both cases it is clear that $R^{-1}(x_1)$ is the union of the blocks $Y_s$ found in $\zeta^{\mathrm{LT}}_{R_A^{-1}(x)}$, consequently, we can say that the similarity class $R^{-1}(x_1)$ is the union of the blocks $Y$ found in $\zeta^{\mathrm{LT}}_{R_A^{-1}(x)}$.

For instance, continue with example 1, we have $\zeta^{\mathrm{LT}}_{R_A^{-1}(x_1)} = \{Y_1 = \{x_1, x_9\}, Y_2 = \{x_1, x_{12}\}\}$ and $\bigcup\{Y_1, Y_2\} = \{x_1, x_9, x_{12}\} = R^{-1}(x_1)$. □

**Property 2** *Any similarity class $R(x)$ of attributes subset $A$ can be represented as the union of maximal limited consistent blocks included in it. In other words,*

$$R(x) = \bigcup \left\{ Z \in \zeta^{\mathrm{LT}}_{R_A} : Z \subseteq R(x) \right\} = \bigcup \left\{ Z \in \zeta^{\mathrm{LT}}_{R_A(x)} \right\}.$$

*Proof* please refer to the proof of Property 1. □

**Property 3** *Given $IIS$ in which $A, B \subseteq AT$ and $X \subseteq U$, then*

1. $\underline{\zeta^{\mathrm{LT}}_{R_A^{-1}}}(X) \subseteq X \subseteq \overline{\zeta^{\mathrm{LT}}_{R_A}}(X)$.
2. *If* $A \subseteq B \implies \overline{\zeta^{\mathrm{LT}}_{R_B}}(X) \subseteq \overline{\zeta^{LT}_{R_A}}(X)$. *But* $\underline{\zeta^{\mathrm{LT}}_{R_A^{-1}}}(X) \subseteq \underline{\zeta^{\mathrm{LT}}_{R_B^{-1}}}(X)$ *does not hold.*

**Theorem 1** *Given $IIS$, the lower approximation of $X \subseteq U$ obtained using MLS-RSM model is a refinement of the one obtained using NS-RSM.*

*Proof* We have to clarify that the lower approximation of NS-RSM obtained using Eq. (5) is subset of lower approximation of MLS-RSM obtained using Eq. (11).

Suppose that $x \in \underline{R}_A^{-1}(X)$ obtained using Eq. (5), then $R^{-1}(x) \subseteq X$ holds. By property 1, $R^{-1}(x) = \bigcup\{Y \in \zeta^{\mathrm{LT}}_{R_A^{-1}} : Y \subseteq R^{-1}(x)\} = \bigcup\{Y \in \zeta^{\mathrm{LT}}_{R_A^{-1}(x)}\}$. Then there exist $Y \in \zeta^{\mathrm{LT}}_{R_A^{-1}(x)}$ that contain $x$ where $Y \subseteq X$. Therefore, $x \in \underline{\zeta^{\mathrm{LT}}_{R_A^{-1}}}(X)$ obtained using Eq. (11). This means that $\forall x \in \underline{R}_A^{-1}(X)$, $x \in \underline{\zeta^{\mathrm{LT}}_{R_A^{-1}}}(X)$. The inverse is not necessarily true. Consequently, the lower approximation of $X$ obtained using MLS-RSM is at least equal to the lower approximation of $X$ obtained using NS-RSM. □

**Theorem 2** *Given $IIS$, the upper approximation of $X \subseteq U$ obtained using MLS-RSM relation is equal to the one obtained using NS-RSM.*

*Proof* The upper approximation of NS-RSM is given by $\overline{R}_A(X) = \bigcup\{R(x) : x \in X\}$. From property 2, $R(x) = \bigcup\{Z \in \zeta^{\mathrm{LT}}_{R_A} : Z \subseteq R(x)\} = \bigcup\{Z \in \zeta^{\mathrm{LT}}_{R_A(x)}\}$ for any $x \in U$. So, $\overline{R}_A(X) = \bigcup\{\bigcup\{Z \in \zeta^{\mathrm{LT}}_{R_A(x)}\} : x \in X\} = \bigcup\{Z \in \zeta^{\mathrm{LT}}_{R_A(x)} : x \in X\} = \overline{\zeta^{\mathrm{LT}}_{R_A}}(X) =$ upper approximation of MLS-RSM. □

We know that, in RST, approximation accuracy for a given target set $X \subseteq U$ is defined as the cardinality of the lower approximation of $X$ divided by the cardinality of the upper approximation of $X$.

Theorem 1 means that a better approximation accuracy for a given target set $X \subseteq U$ can be obtained using MLS-RSM than NS-RSM. The following example is an illustration

**Continue with Example** 1. we illustrate Definitions 4 and 5.

Thus,

$$\zeta^{\mathrm{LT}}_{R_A(x_1)} = \{Z_1 = \{x_1, x_{11}\}\}$$
$$\zeta^{\mathrm{LT}}_{R_A^{-1}(x_1)} = \{Y_1 = \{x_1, x_9\}, Y_2 = \{x_1, x_{12}\}\}$$
$$\zeta^{\mathrm{LT}}_{R_A(x_2)} = \{Z_2 = \{x_2\}\}$$
$$\zeta^{\mathrm{LT}}_{R_A^{-1}(x_2)} = \{Y_3 = \{x_2, x_3\}, Y_4 = \{x_2, x_6\}, \quad Y_5 = \{x_2, x_8\}\}$$
$$\zeta^{\mathrm{LT}}_{R_A(x_3)} = \{Z_3 = \{x_2, x_3\}\} \quad \zeta^{\mathrm{LT}}_{R_A^{-1}(x_3)} = \{Y_6 = \{x_3\}\}$$
$$\zeta^{\mathrm{LT}}_{R_A(x_4)} = \{Z_4 = \{x_4, x_5\}\} \quad \zeta^{\mathrm{LT}}_{R_A^{-1}(x_4)} = \{Y_7 = \{x_4, x_5\, x_{12}\}\}$$
$$\zeta^{\mathrm{LT}}_{R_A(x_5)} = \{Z_4 = \{x_4, x_5\}\} \quad \zeta^{\mathrm{LT}}_{R_A^{-1}(x_5)} = \{Y_7 = \{x_4, x_5, x_{12}\}\}$$
$$\zeta^{\mathrm{LT}}_{R_A(x_6)} = \{Z_5 = \{x_2, x_6\}\} \quad \zeta^{\mathrm{LT}}_{R_A^{-1}(x_6)} = \{Y_8 = \{x_6\}\}$$
$$\zeta^{\mathrm{LT}}_{R_A(x_7)} = \{Z_6 = \{x_7\}\} \quad \zeta^{\mathrm{LT}}_{R_A^{-1}(x_7)} = \{Y_9 = \{x_7, x_9\}, Y_{10} = \{x_7, x_{13}\}\}$$
$$\zeta^{\mathrm{LT}}_{R_A(x_8)} = \{Z_7 = \{x_2, x_8\}\} \quad \zeta^{\mathrm{LT}}_{R_A^{-1}(x_8)} = \{Y_{11} = \{x_8\}\}$$
$$\zeta^{\mathrm{LT}}_{R_A(x_9)} = \{Z_8 = \{x_1, x_7, x_9, x_{11}\}\} \quad \zeta^{\mathrm{LT}}_{R_A^{-1}(x_9)} = \{Y_{12} = \{x_9\}\}$$
$$\zeta^{\mathrm{LT}}_{R_A(x_{10})} = \{Z_9 = \{x_{10}\}\} \quad \zeta^{\mathrm{LT}}_{R_A^{-1}(x_{10})} = \{Y_{13} = \{x_{10}\}\}$$
$$\zeta^{\mathrm{LT}}_{R_A(x_{11})} = \{Z_{10} = \{x_{11}\}\}$$
$$\zeta^{\mathrm{LT}}_{R_A^{-1}(x_{11})} = \{Y_{14} = \{x_1, x_9, x_{11}\}, \quad Y_{15} = \{x_1, x_{11}, x_{12}\}, Y_{16} = \{x_{11}, x_{13}\}\}$$
$$\zeta^{\mathrm{LT}}_{R_A(x_{12})} = \{Z_{11} = \{x_1, x_4, x_5, x_{11}, x_{12}\}\} \quad \zeta^{\mathrm{LT}}_{R_A^{-1}(x_{12})} = \{Y_{17} = \{x_{12}\}\}$$
$$\zeta^{\mathrm{LT}}_{R_A(x_{13})} = \{Z_{12} = \{x_7, x_{11}, x_{13}\}\} \quad \zeta^{\mathrm{LT}}_{R_A^{-1}(x_{13})} = \{Y_{18} = \{x_{13}\}\}$$

**Table 2** Data sets description

| Data set | Percent of "?" values (%) | # of instances | # of attributes |
|---|---|---|---|
| Mammographic | 3 | 961 | 6 |
| Hepatitis | 6 | 155 | 18 |
| Congress | 5.6 | 435 | 16 |

**Table 3** Approximation space comparison

| Data set | Model | $\lvert\underline{\mathrm{Apr}}_A(d_1)\rvert$ | $\lvert\overline{\mathrm{Apr}}_A(d_1)\rvert$ | $\tau(d_1)(\%)$ | $\lvert\underline{\mathrm{Apr}}_A(d_2)\rvert$ | $\lvert\overline{\mathrm{Apr}}_A(d_2)\rvert$ | $\tau(d_2)(\%)$ |
|---|---|---|---|---|---|---|---|
| Mammographic | NS-RSM | 413 | 606 | 68 | 355 | 548 | 64 |
| | MLS-RSM | 421 | 606 | 70 | 363 | 548 | 67 |
| Hepatitis | NS-RSM | 92 | 100 | 92 | 71 | 79 | 89 |
| | MLS-RSM | 92 | 100 | 92 | 78 | 79 | 99 |
| Congress | NS-RSM | 260 | 270 | 96 | 160 | 170 | 94 |
| | MLS-RSM | 266 | 270 | 98.5 | 168 | 170 | 98.8 |

Consequently, $\zeta^{\mathrm{LT}}_{R_A} = \{Z_1 = \{x_1, x_{11}\}, Z_2 = \{x_2\}, Z_3 = \{x_2, x_3\}, Z_4 = \{x_4, x_5\}, Z_5 = \{x_2, x_6\}, Z_6 = \{x_7\}, Z_7 = \{x_2, x_8\}, Z_8 = \{x_1, x_7, x_9, x_{11}\}, Z_9 = \{x_{10}\}, Z_{10} = \{x_{11}\}, Z_{11} = \{x_1, x_4, x_5, x_{11}, x_{12}\}, Z_{12} = \{x_7, x_{11}, x_{13}\}\}$ and $\zeta^{\mathrm{LT}}_{R_A^{-1}} = \{Y_1 = \{x_1, x_9\}, Y_2 = \{x_1, x_{12}\}, Y_3 = \{x_2, x_3\}, Y_4 = \{x_2, x_6\}, Y_5 = \{x_2, x_8\}, Y_6 = \{x_3\}, Y_7 = \{x_4, x_5, x_{12}\}, Y_8 = \{x_6\}, Y_9 = \{x_7, x_9\}, Y_{10} = \{x_7, x_{13}\}, Y_{11} = \{x_8\}, Y_{12} = \{x_9\}, Y_{13} = \{x_{10}\}, Y_{14} = \{x_1, x_9, x_{11}\}, Y_{15} = \{x_1, x_{11}, x_{12}\}, Y_{16} = \{x_{11}, x_{13}\}, Y_{17} = \{x_{12}\}, Y_{18} = \{x_{13}\}\}$.

Noting $\zeta^{\mathrm{LT}}_{R_A^{-1}(x_1)} = \{Y_1 = \{x_1, x_9\}, Y_2 = \{x_1, x_{12}\}\}$ object $x_9$ and object $x_{12}$ are not included in the same block as NS-RSM does. This leads to $Y_2 \subseteq d_1$ which allows object $x_1$ to be included in the lower approximation of $d_1$. In $\zeta^{\mathrm{LT}}_{R_A^{-1}(x_2)} = \{Y_3 = \{x_2, x_3\}, Y_4 = \{x_2, x_6\}, Y_5 = \{x_2, x_8\}\}$ objects $x_3$, $x_6$ and $x_8$ are not in the same block. This leads to $Y_4 \subseteq d_1$ which allows object $x_2$ to be included in the lower approximation of $d_1$.

In $\zeta^{\mathrm{LT}}_{R_A^{-1}(x_{11})} = \{Y_{14} = \{x_1, x_9, x_{11}\}, Y_{15} = \{x_1, x_{11}, x_{12}\}, Y_{16} = \{x_{11}, x_{13}\}\}$ objects $x_1$, $x_{12}$ and $x_{13}$ are not included in the same block. This leads to $Y_{16} \subseteq d_2$ which allows object $x_{11}$ to be included in the lower approximation of $d_2$.

The lower/upper approximations are as follows:

$$\underline{\zeta^{\mathrm{LT}}_{R_A^{-1}}}(d_1) = \{x_1, x_2, x_6, x_{10}, x_{12}\}$$

$$\overline{\zeta^{\mathrm{LT}}_{R_A}}(d_1) = \{x_1, x_2, x_4, x_5, x_6, x_7, x_{10}, x_{11}, x_{12}\}$$

$$\underline{\zeta^{\mathrm{LT}}_{R_A^{-1}}}(d_2) = \{x_3, x_8, x_9, x_{11}, x_{13}\}$$

$$\overline{\zeta^{\mathrm{LT}}_{R_A}}(d_2) = \{x_1, x_2, x_3, x_4, x_5, x_7, x_8, x_9, x_{11}, x_{13}\}$$

Consequently, $\tau(d_1) = \frac{5}{9}$ and $\tau(d_2) = \frac{5}{10} = \frac{1}{2}$. The results are more informative than NS-RSM model. This is because MLS-RSM finds the maximal limited consistent blocks of indiscernible objects and now objects in the same block are similar to each other.

It is worth noting that $\underline{R}_A^{-1}(d_1) \subseteq \underline{\zeta^{\mathrm{LT}}_{R_A^{-1}}}(d_1)$, $\underline{R}_A^{-1}(d_2) \subseteq \underline{\zeta^{\mathrm{LT}}_{R_A^{-1}}}(d_2)$, $\overline{\zeta^{\mathrm{LT}}_{R_A}}(d_1) = \overline{R}_A(d_1)$ and $\overline{\zeta^{\mathrm{LT}}_{R_A}}(d_2) = \overline{R}_A(d_2)$ which verifies the validity of MLS-RSM.

## 5 Experimental evaluation

### 5.1 Experimental setup

The proposed approach (MLS-RSM) and the related one (NS-RSM) are implemented using MATLAB R12a on PC with windows 8, Intel(R) Core(TM) i7 CPU 2.4 GHZ and 6GB memory, to verify the validity of MLS-RSM. Our experiments employ three publicly accessible data sets; mammographic (Team 2015), hepatitis and congress (UCI 2015). The datasets are outlined in Table 2. The objective of the experiment is to compare the approximation space located by MLS-RSM and NS-RSM.

### 5.2 Approximation space comparison

We obtain the numbers of objects in the lower and upper approximations and the approximation accuracy in terms of MLS-RSM and NS-RSM as Table 3.

By Table 3, it is easy to note that using MLS-RSM, the number of objects in the lower approximations of MLS-RSM model is equal or greater than those in NS-RSM lower approximations. The number of objects in the upper approximations of MLS-RSM model is equal to those in NS-RSM upper approximation. This shrinks the boundary region leading to promising approximation accuracy which verifies the validity of MLS-RSM model.

Consequently, we can say that MLS-RSM model can approximate the target set more accurately than NS-RSM in IIS with "?" values. This is a perfect indicator that MLS model can improve the feature subset selection technique in IIS with "?" values.

## 6 Conclusion

Uncertainty in the data degrades the process of analyzing the data. Non-symmetric similarity relation-based rough set model (NS-RSM) is viewed as a mathematical tool to deal with uncertainty in incomplete information systems with "?" values. Unfortunately, NS-RSM results in unpromising approximation space when addressing inconsistent data sets that have lots of boundary objects. This is because objects in the same similarity classes are not necessarily similar to each other and may belong to different target classes. To enhance NS-RSM capability, we introduce the maximal limited similarity relation-based rough set model (MLS-RSM) which describes the maximal collection of indistinguishable objects that are limited tolerance to each other in similarity classes. This allows accurate computation to be done for the approximation space.Theoretical analysis shows that MLS-RSM can approximate the target set more efficiently and more accurately than NS-RSM. The experimental results show that MLS-RSM can deal with the analysis of imprecise and uncertain information in IIS. In the future work, MLS-RSM can be used to improve the reduct issue.

**Acknowledgments** The authors would like to thank the editors and the anonymous reviewers for their valuable comments and suggestions which helped immensely in improving the quality of the paper.

**Funding** This study is not funded by any organization.

**Compliance with ethical standards**

**Conflict of interest** The three authors declare that they have no conflict of interest.

**Ethical approval** This article does not contain any studies with human or animal participants performed by any of the authors.

## References

Cheng YS, Zhang YS, Hu XG, Zhang YZ (2007) Uncertainty measure of knowledge and rough set based on maximal consistency block technique. In: Proceedings of the sixth international conference on machine learning and cybernetics, Hong Kong, pp 3069–3074

Dai J, Xu Q (2012) Approximations and uncertainty measures in incomplete information systems. Inf Sci 198:62–80

DU WF, ZI T (2014) Properties of lower and upper approximation operators under various kinds of relations. In: International conference on mechanics and civil engineering, pp 472–477

Gantayat SS, Misra A, Panda BS (2014) A study of incomplete data—a review. In: Advances in intelligent systems and computing, proceedings of the international conference on frontiers of intelligent computing: theory and applications (FICTA), vol 247, pp 401–408

Grzymała-Busse JW (2004) Characteristic relations for incomplete data: a generalization of the indiscernibility relation. In: Lecture notes in computer science. Proceeding of the fourth international conference on rough sets and current trends in computing, vol 3066, pp 244–253

Grzymala-Busse JW, Hu M (2005) A comparison of several approaches to missing attribute values in data mining. In: Lecture notes in computer science, proceedings of the second international conference on rough sets and current trends in computing. Springer, Berlin, pp 378–385

Grzymala-Busse JW, Wang A (1997) Modified algorithms lem1 and lem2 for rule induction from data with missing attribute values. In: Machine learning and cybernetics communications in computer and information science, proceeding of the fifth international workshop on rough sets and soft computing at the third joint conference on information sciences, pp 69–72

Huang J, Guan Y, Shen J, Wang H (2014) Rough approximations in tolerance rough set models. In: 11th International conference on fuzzy systems and knowledge discovery, pp 61–66

Huang S, Li M (2014) Limited and variable precision rough set model. J Inf Comput Sci 11:3493–3501

Jensen R, Tusonb A, Shena Q (2014) Finding rough and fuzzy-rough set reducts with sat. Inf Sci 225:100–120

Jiang Y, Yu Y (2015) Minimal attribute reduction with rough set based on compactness discernibility information tree. Soft Comput 20:2233–2243

Kryszkiewicz M (1998) Rough set approach to incomplete information system. Inf Sci 112:39–49

Kryszkiewicz M (1999) Rules in incomplete information systems. Inf Sci 113:271–292

Leung Y, Li D (2003) Maximal consistent block technique for rule acquisition in incomplete information systems. Inf Sci 153:85–106

Liu X, Shao M (2014) Approachs to computing maximal consistent block. Machine Learning and Cybernetics Communications in Computer and Information Science 481:264–274

Nguyen DV, Yamada K, Unehara M (2013) Extended tolerance relation to define a new rough set model in incomplete information systems. Adv Fuzzy Syst 13:1–10

Pawlak Z (1982) Rough sets. Int J Inf Comput Sci 11:341–356

Qin B, Xia G, Yan K (2015) Similarity of binary relations based on rough set theory and topology: an application for topological structures of matroids. Soft Comput 20:853–861

Skowron A, Stepaniuk J (1996) Tolerance approximation space. Fundam Inform 27:245–253

Słowiński R, Greco S, Matarazzo B (2014) Rough-set-based decision support. In: Burke EK, Kendall G (eds) Search methodologies. Springer, pp 475–527

Stefanowski J, Tsoukiàs A (1999) On the extension of rough sets under incomplete information. In: Lecture notes in computer science, 7th international workshop on new directions in rough sets, data mining, and granular-soft computing, vol 1711, pp 73–81

Stefanowski J, Tsoukiàs A (2001) Incomplete information tables and rough classification. Comput Intell 17:545–566

Team W (2015) Knowledge extraction based on evolutionary learning. http://sci2s.ugr.es/keel/missing.php

UCI (2015) Uci machine learning repository. https://archive.ics.uci.edu/ml/index.html

Wang G (2001) Rough set theory and knowledge acquisition. Xi'an Jiaotong University Press, Xi'an

Wang G (2002) Extension of rough set under incomplete information systems*. Fuzzy systems, 2002. FUZZ-IEEE'02. In: Proceedings of the IEEE international conference 2, pp 1098–1103

Wang G, Guan L, Hu F (2008) Rough set extensions in incomplete information systems. Front Electr Electron Eng China 3:399–405

Yang X (2009) Difference relation-based rough set and negative rules in incomplete information system. Int J Uncertain Fuzziness Knowl Based Syst 18:649–665

Yang X, Yang J (2012) Expansions of rough sets in incomplete information systems. Science Press, Beijing

Yin X, XiuyiJia Shang L (2006) A new extension model of rough sets under incomplete information, vol 4062. Springer, Berlin