Rough set analysis of classification data with missing values

Marcin Szeląg[1], Jerzy Błaszczyński[1], Roman Słowiński[1,2]

[1] Institute of Computing Science, Poznań University of Technology,
60-965 Poznań, Poland
{mszelag, jblaszczynski, rslowinski}@cs.put.poznan.pl
[2] Systems Research Institute, Polish Academy of Sciences,
01-447 Warsaw, Poland

Abstract. In this paper, we consider a rough set analysis of non-ordinal and ordinal classification data with missing attribute values. We show how this problem can be addressed by several variants of Indiscernibility-based Rough Set Approach (IRSA) and Dominance-based Rough Set Approach (DRSA). We propose some desirable properties that a rough set approach being able to handle missing attribute values should possess. Then, we analyze which of these properties are satisfied by the considered variants of IRSA and DRSA.

Keywords: Rough set, Indiscernibility-based Rough Set Approach, Dominance-based Rough Set Approach, Missing values

1 Introduction

In data mining concerning classification problems, it is quite common to have missing values for attributes describing objects [12]. To cope with the problem of missing values, several approaches have been proposed. The usual approach is to assume that some value(s) can represent correctly the missing one. Then, the missing values are replaced in some way by so-called representative values. In this case, the question is how to avoid data distortion [12].

Rough set approach to handling missing values avoids making changes in the data. The problem is addressed by a proper definition of the relation employed to form granules of knowledge.

In this work, we consider both Indiscernibility-based Rough Set Approach (IRSA), in which value sets of attributes describing objects are not supposed to be ordered, and Dominance-based Rough Set Approach (DRSA), which takes into account an order in the value sets of attributes, monotonically related with the order of decision classes. We focus on the following types of IRSA:

- classical rough set approach (CRSA) proposed by Pawlak [16],
- Variable Consistency Indiscernibility-based Rough Set Approach (VC-IRSA) proposed by Błaszczyński, Greco, Słowiński, and Szeląg [2, 3],

and on the following types of DRSA:

– classical Dominance-based Rough Set Approach (CDRSA) proposed by Greco, Matarazzo, and Słowiński [8, 9, 17],
– Variable Consistency Dominance-based Rough Set Approach (VC-DRSA) proposed by Błaszczyński, Greco, Słowiński, and Szeląg [2, 3].

Adaptations of the classical rough set model [16] to handling missing values, were presented in [6, 7, 10, 11, 14, 19]. Proposals of handling missing values in dominance-based rough set approaches were given in [1, 5, 6, 7, 13, 15, 20]. We review all these approaches and analyze their properties, refining and extending the research results presented in [1, 4].

The rest of this paper is structured as follows. Section 2 reminds basics of IRSA and DRSA. In Section 3, we present ways of handling missing values in IRSA and DRSA. We also propose a list of desirable properties that IRSA and DRSA adapted to handle missing values should possess. After characterizing variants of IRSA and DRSA coping with missing values, we discover non-dominated variants with respect to these properties. Section 4 concludes the paper.

2 Basics of IRSA and DRSA

Classification data analyzed by IRSA and DRSA concern a finite universe U of objects described by attributes from a finite set A. Moreover, A is divided into disjoint sets of condition attributes C and decision attributes Dec. The value set of $q \in C \cup Dec$ is denoted by V_q, $q(x) \in V_q$ denotes evaluation of object $x \in U$ on attribute q, and $V_C = \prod_{q=1}^{|C|} V_q$ is called C-evaluation space. For simplicity, we assume that $Dec = \{d\}$. Values of attribute d are class labels.

Decision attribute d makes a partition of set U into n disjoint sets of objects, called *decision classes*. We denote this partition by $\mathcal{X} = \{X_1, \ldots, X_n\}$.

2.1 Basics of IRSA

In IRSA, the value sets of attributes are not considered to be ordered, and thus *indiscernibility relation* is employed. Object y is considered to be indiscernible with object x (denoted by yIx) if and only if (iff) $q(y) = q(x)$ for each $q \in C$. Given an object $x \in U$,

$$I(x) = \{y \in U : yIx\} \tag{1}$$

denotes a set (granule) of objects indiscernible with referent x.

Given a non-ordinal classification problem, two objects $x, y \in U$ are said to be *inconsistent* with respect to (w.r.t.) indiscernibility relation, if they are indiscernible but they are assigned to different decision classes. In order to handle such inconsistency, one calculates lower approximations of considered classes.

CRSA In CRSA [16], *lower approximation* of class $X_i \in \mathcal{X}$ is defined as

$$\underline{X_i} = \{x \in U : I(x) \subseteq X_i\}, \tag{2}$$

and *upper approximation* of class $X_i \in \mathcal{X}$ is defined as

$$\overline{X_i} = \{x \in U : I(x) \cap X_i \neq \emptyset\}. \tag{3}$$

VC-IRSA In VC-IRSA [2, 3], *probabilistic lower approximation* of class $X_i \in \mathcal{X}$ is defined using an *object consistency measure*. We employ cost-type measure ϵ_{X_i}:

$$\epsilon_{X_i}(x) = \frac{|I(x) \cap \neg X_i|}{|\neg X_i|}, \tag{4}$$

where $\neg X_i = U \setminus X_i$. Then,

$$\underline{X_i} = \{x \in X_i : \epsilon_{X_i}(x) \leq \theta_{X_i}\}, \tag{5}$$

where threshold $\theta_{X_i} \in [0, 1]$. In the following, we will denote this version of VC-IRSA by ϵ-VC-IRSA.

In [3], we introduced some *monotonicity properties* required from an object consistency measure. For IRSA, relevant properties are: $(m1)$ – monotonicity w.r.t. growing set of attributes, and $(m2)$ – monotonicity w.r.t. growing set of objects (class). As proved in [3], ϵ_{X_i} has both property $(m1)$ and property $(m2)$.

2.2 Basics of DRSA

In DRSA, it is supposed that value sets of condition attributes, as well as decision classes, are ordered. Then, it is often meaningful to consider *monotonicity constraints* (*monotonic relationships*) between ordered class labels and values of attributes expressed on ordinal or cardinal (numerical) scales [8, 9, 17]. In order to make a meaningful representation of classification decisions, one has to consider the *dominance relation* D in the C-evaluation space. Let us denote by $\succeq_q$ the *weak preference relation* over U confined to single attribute $q \in C$:

$$y \succeq_q x \Leftrightarrow \begin{cases} q(y) \text{ is not missing,} \\ q(x) \text{ is not missing,} \\ q(y) \text{ is at least as good as } q(x). \end{cases} \tag{6}$$

Then, classically (i.e., when there are no missing attribute values), given $x, y \in U$, object y is said to *dominate* object x, denoted by yDx, iff $y \succeq_q x$ for each $q \in C$. Moreover, y is said to *be dominated* by x, denoted by $y \,\mathit{D}\!\!\!\!/\; x$, iff $x \succeq_q y$ for each $q \in C$. Let us observe that, classically, yDx iff $x \,\mathit{D}\!\!\!\!/\; y$.

Dominance relations D and $\mathit{D}\!\!\!\!/$ are partial preorders, i.e., they are reflexive, transitive, and not necessarily complete. For any object $x \in U$, two types of dominance cones can be defined in the C-evaluation space. Positive dominance cone with the origin in x w.r.t. relation D:

$$D^+(x) = \{y \in U : yDx\}, \tag{7}$$

and negative dominance cone with the origin in x w.r.t. relation D:

$$D^-(x) = \{y \in U : xDy\}. \tag{8}$$

In DRSA, if $1 \leq i < j \leq n$, then class X_i is considered to be worse than X_j. Moreover, rough approximations concern unions of classes: upward unions $X_i^{\geq} = \bigcup_{t \geq i} X_t$, and downward unions $X_i^{\leq} = \bigcup_{t \leq i} X_t$, where $i = 1, \ldots, n$.

CDRSA In CDRSA [8, 9, 17], *lower approximations* of unions of classes $X_i^{\geq}$, $X_i^{\leq}$, $i = 1, \ldots, n$, are defined using strict inclusion relation:

$$\underline{X_i^{\geq}} = \{x \in U : D^+(x) \subseteq X_i^{\geq}\}, \quad \underline{X_i^{\leq}} = \{x \in U : D^-(x) \subseteq X_i^{\leq}\}. \tag{9}$$

Moreover, *upper approximations* of unions of classes $X_i^{\geq}$, $X_i^{\leq}$ are defined as

$$\overline{X_i^{\geq}} = \{x \in U : D^-(x) \cap X_i^{\geq} \neq \emptyset\}, \quad \overline{X_i^{\leq}} = \{x \in U : D^+(x) \cap X_i^{\leq} \neq \emptyset\}. \tag{10}$$

VC-DRSA Definition (9) appears to be too restrictive in practical applications. This explains the interest in VC-DRSA [2, 3] which is a probabilistic extension of CDRSA. We use *object consistency measures* $\epsilon_{X_i^{\geq}} : U \rightarrow [0, 1]$, $\epsilon_{X_i^{\leq}} : U \rightarrow [0, 1]$, introduced in [2, 3]:

$$\epsilon_{X_i^{\geq}}(x) = \frac{|D^+(x) \cap \neg X_i^{\geq}|}{|\neg X_i^{\geq}|}, \quad \epsilon_{X_i^{\leq}}(x) = \frac{|D^-(x) \cap \neg X_i^{\leq}|}{|\neg X_i^{\leq}|}. \tag{11}$$

Then, *probabilistic lower approximations* of $X_i^{\geq}$, $X_i^{\leq}$, $i = 1, \ldots, n$, are defined as

$$\underline{X_i^{\geq}} = \{x \in X_i^{\geq} : \epsilon_{X_i^{\geq}}(x) \leq \theta_{X_i^{\geq}}\}, \quad \underline{X_i^{\leq}} = \{x \in X_i^{\leq} : \epsilon_{X_i^{\leq}}(x) \leq \theta_{X_i^{\leq}}\}, \tag{12}$$

where $\theta_{X_i^{\geq}}, \theta_{X_i^{\leq}} \in [0, 1)$. In the following, we will denote this version of VC-DRSA by ϵ-VC-DRSA.

As proved in [3], $\epsilon_{X_i^{\geq}}$, $\epsilon_{X_i^{\leq}}$ have monotonicity properties $(m1)$, $(m2)$, and $(m4)$ (monotonicity w.r.t. dominance relation), sufficient in practical applications.

3 Different Ways of Handling Missing Values in IRSA and DRSA

In the following, a missing attribute value is denoted by $*$. We assume that each object $x \in U$ has at least one known value, i.e., for each $x \in U$ there exists $q \in C$ such that $q(x) \neq *$. Moreover, we use symbol X to denote an approximated set of objects. In IRSA, X denotes a single decision class $X_i \in \mathcal{X}$. In DRSA, X denotes a union of decision classes $X_i^{\geq}$ or $X_i^{\leq}$, $i \in \{1, \ldots, n\}$.

3.1 Adaptations of IRSA to handle missing values

Handling of missing attribute values requires a proper adaptation of IRSA by redefinition of the indiscernibility relation I. Once we fix this definition, we can proceed by calculating rough approximations of decision classes, and then inducing decision rules from data structured in the rough set way.

The approaches resulting from different definitions of the indiscernibility relation are denoted by CRSA-mv_j and ϵ-VC-IRSA-mv_j, and the respective indiscernibility relations are denoted by I_j, where j stands for the version id. When these approaches are described jointly, we use denotation IRSA-mv_j.

It is important to underline that due to missing values, considered indiscernibility relation I_j may miss some properties, like symmetry or transitivity. For this reason, in the following, we employ generalized definitions of rough approximations proposed in [18], where indiscernibility relation is only assumed to be reflexive (so it may be not symmetric and/or not transitive). According to [18],

$$I_j^{-1}(x) = \{y \in U : xI_jy\} \tag{13}$$

denotes the set (granule) of objects with which x is indiscernible (to which x is similar). Then, in CRSA-mv_j, *generalized lower approximation* of class $X_i \in \mathcal{X}$ is defined as

$$\underline{X_i} = \{x \in U : I_j^{-1}(x) \subseteq X_i\}. \tag{14}$$

Generalized upper approximation of class $X_i \in \mathcal{X}$ is defined as

$$\overline{X_i} = \bigcup_{x \in X_i} I_j(x). \tag{15}$$

Let us remark that if I_j is symmetric, then $I_j^{-1}(x) = I_j(x)$, and then, definitions (14) and (2) are equivalent [18].

Analogously, ϵ-VC-IRSA is adjusted to the case of I_j, possibly being not symmetric, by redefining object consistency measure ϵ_{X_i}, given by (4), in the following way:

$$\epsilon_{X_i}(x) = \frac{|I_j^{-1}(x) \cap \neg X_i|}{|\neg X_i|}. \tag{16}$$

IRSA-mv_1 employs the indiscernibility relation defined in [6, 7], which we denote by I_1. This relation is considered as a directional statement where a subject is compared to a referent which cannot have missing values. Subject y is considered to be indiscernible with referent x iff for each $q \in C$, $q(x) \neq *$, and either $q(y) = q(x)$ or $q(y) = *$. Thus, it is not true that xI_1x when object $x \in U$ has some missing attribute values (i.e., I_1 is, in general, not reflexive). Nevertheless, it is still interesting to see consequences of adapting IRSA by using relation I_1.

Note that in [6, 7], lower approximation of class X_i was not defined using (14), and moreover, some properties considered in these papers (like rough inclusion or complementarity), were defined with respect to subset U_C of the universe U, where U_C is composed of all objects from U which have no missing value. Thus, we have to verify if these properties hold also for U.

IRSA-$mv_{1.5}$ [19] can be considered as an improvement over IRSA-mv_1. It defines a reflexive and transitive similarity relation without imposing that a referent cannot have missing values. In this approach, subject y is considered to be indiscernible with referent x iff $q(y) = q(x)$ for each $q \in C$ such that $q(y) \neq *$. Let us remark that this approach is treating missing values as "lost" ones (see, e.g., [10, 11]).

IRSA-mv_2 [6, 7, 14, 19] employs a reflexive and symmetric tolerance relation. In this approach, subject y is considered to be indiscernible with referent x iff for each $q \in C$ there is $q(y) = q(x)$, or $q(y) = *$, or $q(x) = *$. Note that this approach is treating missing values as "do not care" ones (see, e.g., [10, 11]).

IRSA-mv_3 is a new approach which is an indiscernibility-based counterpart of DRSA-mv_3 proposed in [1]. In this approach, subject y is considered to be indiscernible with referent x iff $q(y) = q(x)$ for each $q \in C$ such that $q(x) \neq *$.

3.2 Desirable properties of IRSA adapted to handle missing values

We consider the following desirable properties of IRSA-mv_j, $j = 1, 1.5, 2, 3$:

1. Property S (reflecting symmetry of indiscernibility relation): IRSA-mv_j has property S iff $yI_jx \Leftrightarrow xI_jy$, for any $x, y \in U$.
2. Property R (reflecting reflexivity of indiscernibility relation): IRSA-mv_j has property R iff xI_jx, for any $x \in U$.
3. Property T (reflecting transitivity of indiscernibility relation): IRSA-mv_j has property T iff $yI_jx \wedge xI_jz \Rightarrow yI_jz$, for any $x, y, z \in U$.
4. Property B (robustness): given $x \in U$, let $C^x = \{q \in C : q(x) \neq *\}$; IRSA-$mv_j$ has property B iff for each $x \in \underline{X}$, $I_j^{-1\prime}(x) \cap \neg X \subseteq I_j^{-1}(x) \cap \neg X$, where $I_j^{-1\prime}(x)$ is a set of objects such that in C^x-evaluation space, object x is indiscernible with them.
5. Property P (reflecting precisiation of data): IRSA-mv_j has property P iff the lower approximation of any $X \subseteq U$ does not shrink when any missing attribute value is replaced by some non-missing value.
6. Property RI (rough inclusion): IRSA-mv_j has property RI iff $\underline{X} \subseteq X \subseteq \overline{X}$, for any $X \subseteq U$.
7. Property C (complementarity): IRSA-mv_j has property C iff $\underline{X} = U \setminus \overline{\neg X}$, for any $X \subseteq U$.
8. Property M_1 (monotonicity w.r.t. growing set of attributes): IRSA-mv_j has property M_1 iff the lower approximation of any $X \subseteq U$ does not shrink when set P is extended by new attributes.
9. Property M_2 (monotonicity w.r.t. growing set of objects): IRSA-mv_j has property M_2 iff the lower approximation of any $X \subseteq U$ does not shrink when this set is augmented by new objects.
10. Property MT (transitivity of membership to lower approximation): IRSA-mv_j has property MT iff for any $X \subseteq U$ and for any $x, y \in U$ it is true that $x \in \underline{X} \wedge y \in X \wedge xI_jy \Rightarrow y \in \underline{X}$.

Comparing to the list of desirable properties introduced in [4], we propose new property B which postulates that an object x, belonging to the lower approximation of class X_i when considering all condition attributes, should also belong to this approximation when considering only these attributes, for which evaluation of x is not missing. Moreover, we modify definition of property MT to reflect definition of generalized lower approximation given by (14) (for CRSA-mv_j), and by (5), (16) (for ϵ-VC-IRSA-mv_j).

The properties of IRSA-mv_j, $j = 1, 1.5, 2, 3$, are summarized in Table 1, where **T** and F denote presence and absence of a given property, respectively. Moreover, in case of two symbols $\cdot/\cdot$, the first (resp. the second) one concerns only CRSA (resp. only ϵ-VC-IRSA).

Table 1. Properties of IRSA-mv_j, $j = 1, 1.5, 2, 3$

Property / Approach	IRSA-mv_1	IRSA-$mv_{1.5}$	IRSA-mv_2	IRSA-mv_3
S	F	F	**T**	F
R	F	**T**	**T**	**T**
T	**T**	**T**	F	**T**
B	F	**T**	**T**	F
P	F	F	**T**	F
RI	F	**T**	**T**	**T**
C	F/**T**	**T**	**T**	**T**
M_1	**T**	**T**	**T**	**T**
M_2	**T**	**T**	**T**	**T**
MT	**T**	**T**	F	**T**

According to Table 1, IRSA-$mv_{1.5}$ and IRSA-mv_3 dominate IRSA-mv_1, which has the least number of desirable properties; IRSA-mv_3 is dominated by IRSA-$mv_{1.5}$. Thus, taking into account the considered properties, we can conclude that there are two non-dominated approaches: IRSA-$mv_{1.5}$ and IRSA-mv_2.

3.3 Adaptations of DRSA to handle missing values

Handling of missing attribute values requires a proper adaptation of DRSA by redefinition of the dominance relations D and $\mathcal{I}$. Once we fix these definitions, we can proceed by calculating rough approximations of unions of decision classes, and then inducing decision rules from data structured in the rough set way.

In this sub-section, we review several ways of adapting DRSA to missing values known from the literature, and we propose some new adaptations. All of them are based on specific definitions of dominance relations.

The approaches, resulting from different definitions of the dominance relations, are denoted by CDRSA-mv_j and ϵ-VC-DRSA-mv_j, and the respective dominance relations are denoted by D_j and $\mathcal{I}_j$, where j stands for the version id. When these approaches are described jointly, we use denotation DRSA-mv_j.

It is important to underline that due to missing values, an approach employing dominance relation D_j may miss some properties, like transitivity. Moreover, it may be the case that yD_jx while not $x\,\mathcal{I}_j\,y$ (lack of a specific kind of symmetry). For this reason, in the following, we employ generalized definitions of rough approximations formulated in [20], related to generalized definitions of rough approximations proposed for IRSA in [18]. These generalized definitions are valid for the case when considered relations D_j and $\reflectbox{$D$}_j$ are reflexive (regardless of their being transitive or satisfying $yD_jx \Leftrightarrow x\, \reflectbox{D}_j\, y$).

According to [20], for any object $x \in U$, apart from dominance cones $D_j^+(x)$ and $D_j^-(x)$, two more types of dominance cones in the C-evaluation space should be considered. Positive dominance cone with the origin in x w.r.t. relation $\reflectbox{$D$}_j$:

$$\reflectbox{D}_j^+(x) = \{y \in U : x\, \reflectbox{D}_j\, y\}, \tag{17}$$

and negative dominance cone with the origin in x w.r.t. relation $\reflectbox{$D$}_j$:

$$\reflectbox{D}_j^-(x) = \{y \in U : y\, \reflectbox{D}_j\, x\}. \tag{18}$$

Let us observe that, when the description of objects has no missing values, $\reflectbox{$D$}_j^+(x) = D_j^+(x)$ and $\reflectbox{$D$}_j^-(x) = D_j^-(x)$. Then, according to [20], in CDRSA-mv_j:

– *generalized lower approximation* of $X_i^{\geq}$, $i \in \{1, \ldots, n\}$, is defined as

$$\underline{X_i^{\geq}} = \{x \in U : \reflectbox{D}_j^+(x) \subseteq X_i^{\geq}\}, \tag{19}$$

where $\reflectbox{$D$}_j^+(x)$ is read as "the set of objects that x is dominated by";

– *generalized upper approximation* of $X_i^{\geq}$, $i \in \{1, \ldots, n\}$, is defined as

$$\overline{X_i^{\geq}} = \{x \in U : D_j^-(x) \cap X_i^{\geq} \neq \emptyset\}, \tag{20}$$

where $D_j^-(x)$ is read as "the set of objects that x dominates";

– *generalized lower approximation* of $X_i^{\leq}$, $i \in \{1, \ldots, n\}$, is defined as

$$\underline{X_i^{\leq}} = \{x \in U : D_j^-(x) \subseteq X_i^{\leq}\}, \tag{21}$$

where $D_j^-(x)$ is read as "the set of objects that x dominates";

– *generalized upper approximation* of $X_i^{\leq}$, $i \in \{1, \ldots, n\}$, is defined as

$$\overline{X_i^{\leq}} = \{x \in U : \reflectbox{D}_j^+(x) \cap X_i^{\leq} \neq \emptyset\}, \tag{22}$$

where $\reflectbox{$D$}_j^+(x)$ is read as "the set of objects that x is dominated by".

Note that when yD_jx implies $x\, \reflectbox{$D$}_j\, y$, and vice versa (presence of a specific kind of symmetry), then:

– the lower approximation of a union of classes $X_i^{\geq}$ defined by (19) is identical to the lower approximation of the same union defined by (9);
– the upper approximation of a union of classes $X_i^{\leq}$ defined by (22) is identical to the upper approximation of the same union defined by (10).

Analogously, ϵ-VC-DRSA is generalized by redefining object consistency measures $\epsilon_{X_i^{\geq}}$, $\epsilon_{X_i^{\leq}}$, given by (11), in the following way:

$$\epsilon_{X_i^{\geq}}(x) = \frac{|\reflectbox{D}_j^+(x) \cap \neg X_i^{\geq}|}{|\neg X_i^{\geq}|}, \qquad \epsilon_{X_i^{\leq}}(x) = \frac{|D_j^-(x) \cap \neg X_i^{\leq}|}{|\neg X_i^{\leq}|}. \tag{23}$$

DRSA-mv_1 employs two dominance relations defined in [6, 7], which we denote by D_1 and $\reflectbox{$D$}_1$. These relations are considered as directional statements where subject y is compared to referent x which cannot have missing values:

- subject y *dominates* referent x (denoted by yD_1x) iff for each $q \in C$, $q(x) \neq *$, and either $y \succeq_q x$ or $q(y) = *$;
- subject y *is dominated by* referent x (denoted by $y \reflectbox{$D$}_1 x$) iff for each $q \in C$, $q(x) \neq *$, and either $x \succeq_q y$ or $q(y) = *$.

In view of the above definitions of D_1 and $\reflectbox{$D$}_1$, neither xD_1x nor $x \reflectbox{$D$}_1 x$ (i.e., D_1, $\reflectbox{$D$}_1$ are not reflexive), in general. Nevertheless, it is still interesting to see consequences of adapting DRSA to handle missing values by using relations D_1 and $\reflectbox{$D$}_1$. Note that in [6, 7], lower approximations of unions of classes $X_i^{\geq}$ and $X_i^{\leq}$ were not defined using (19) and (21), and moreover, some properties considered in these papers (like rough inclusion or complementarity), were defined with respect to $U_C \subseteq U$, where U_C is composed of all objects from U which have no missing value. Thus, we have to verify if these properties hold also for U.

DRSA-$mv_{1.5}$ [20] can be considered as an improvement over DRSA-mv_1. In this approach, the authors propose two relations (called in [20] *similarity dominance relations*), which we denote by $D_{1.5}$ and $\reflectbox{$D$}_{1.5}$:

- subject y *dominates* referent x (denoted by $yD_{1.5}x$) iff $y \succeq_q x$ for each $q \in C$ such that $q(y) \neq *$;
- subject y *is dominated by* referent x (denoted by $y \reflectbox{$D$}_{1.5} x$) iff $x \succeq_q y$ for each $q \in C$ such that $q(y) \neq *$.

Taking into account the semantics of missing values considered in [10, 11], it can be said that DRSA-$mv_{1.5}$ treats missing values as “lost” values.

DRSA-mv_2 was first proposed in [6, 7], and extended in [5] to handle imprecise evaluations on attributes and imprecise assignments to decision classes, both modeled by intervals. When considering missing values only, each object is assigned to a single class, and each missing attribute value corresponds to an interval spanning over entire value set of this attribute. This implies the following definitions of so-called *possible dominance relations*, denoted by D_2 and $\reflectbox{$D$}_2$:

- subject y *dominates* referent x (denoted by yD_2x) iff for each $q \in C$, $y \succeq_q x$, or $q(y) = *$, or $q(x) = *$;
- subject y *is dominated by* referent x (denoted by $y \reflectbox{$D$}_2 x$) iff for each $q \in C$, $x \succeq_q y$, or $q(y) = *$, or $q(x) = *$.

Taking into account the semantics of missing values considered in [10, 11], it can be said that DRSA-mv_2 treats missing values as “do not care” values.

In DRSA-$mv_{2.5}$ [13], two dominance relations (called in [13] *generalized extended dominance relations*) are defined as in DRSA-mv_2, only with additional condition that the ratio of the number of "common" attributes (i.e., attributes for which both x and y have simultaneously a non-missing value) and the number of all attributes in set C is not less than a given user-defined threshold $\lambda \in [0,1]$. We denote these relations by $D_{2.5}$ and $\reflectbox{$D$}_{2.5}$. The additional condition was introduced to restrict the dominance relations used in DRSA-mv_2 to pairs of objects that have at least one, or more, "common" attribute(s).

In DRSA-mv_3 [1], we employ dominance relations D_3 and $\reflectbox{$D$}_3$, defined as:

- subject y *dominates* referent x (denoted by yD_3x) iff $y \succeq_q x$ for each $q \in C$ such that $q(x) \neq *$;
- subject y *is dominated by* referent x (denoted by $y \reflectbox{$D$}_3\, x$) iff $x \succeq_q y$ for each $q \in C$ such that $q(x) \neq *$.

DRSA-mv_4 uses the concept of a *lower-end dominance relation* introduced in [5]. Resulting dominance relations D_4 and $\reflectbox{$D$}_4$ are defined as:

- subject y *dominates* referent x (denoted by yD_4x) iff for each $q \in C$, $y \succeq_q x$, or $q(x) = *$, or $q(x) = \inf(V_q)$;
- subject y *is dominated by* referent x (denoted by $y \reflectbox{$D$}_4\, x$) iff for each $q \in C$, $x \succeq_q y$, or $q(y) = *$, or $q(y) = \inf(V_q)$,

where $\inf(V_q)$ denotes the worst value in V_q (if no such value exists, $\inf(V_q) = -\infty$).

DRSA-mv_5 uses the concept of an *upper-end dominance relation* introduced in [5]. Resulting dominance relations D_5 and $\reflectbox{$D$}_5$ are defined as:

- subject y *dominates* referent x (denoted by yD_5x) iff for each $q \in C$, $y \succeq_q x$, or $q(y) = *$, or $q(y) = \sup(V_q)$;
- subject y *is dominated by* referent x (denoted by $y \reflectbox{$D$}_5\, x$) iff for each $q \in C$, $x \succeq_q y$, or $q(x) = *$, or $q(x) = \sup(V_q)$,

where $\sup(V_q)$ denotes the best value in V_q (if there is no such value, $\sup(V_q) = \infty$).

In DRSA-mv_6 [15], the authors define so-called *new extended dominance relation*, which we denote by D_6. It is an α-cut of fuzzy dominance relation $\widetilde{D}$, such that $\widetilde{D}(y,x)$ reflects a possibility of yDx, for $y,x \in U$. Threshold $\alpha \in [0,1]$ is a parameter estimated using decision-theoretic rough set model. This approach assumes that the value set of each attribute is discrete. Relation $\widetilde{D}$ is defined as

$$\widetilde{D}(y,x) = \prod_{q \in C} \widetilde{\succeq}_q(y,x), \tag{24}$$

where *fuzzy weak preference relation* over U confined to single attribute $q \in C$

$$\widetilde{\succeq}_q(y,x) = \begin{cases} 0, & \text{if } q(y) \neq *, q(x) \neq *, \text{not } y \succeq_q x \\ 1, & \text{if } q(y) \neq *, q(x) \neq *, y \succeq_q x \\ \frac{|\{v : v \in V_q, v \text{ is not worse than } q(x)\}|}{|V_q|}, & \text{if } q(y) = *, q(x) \neq * \\ \frac{|\{v : v \in V_q, q(y) \text{ is not worse than } v\}|}{|V_q|}, & \text{if } q(y) \neq *, q(x) = * \\ \frac{1}{2} + \frac{1}{2|V_q|}, & \text{if } q(y) = *, q(x) = * \end{cases}. \tag{25}$$

Then,

$$D_6 = \{(y,x) \in U \times U : \widetilde{D}(y,x) \geq \alpha\} \cup \{(x,x) : x \in U\}, \tag{26}$$

where threshold $\alpha \in [0,1]$. Moreover, once can define dominance relation $\reflectbox{D}_6$ as

$$\reflectbox{D}_6 = \{(y,x) \in U \times U : \widetilde{\reflectbox{D}}(y,x) \geq \alpha\} \cup \{(x,x) : x \in U\}, \tag{27}$$

where fuzzy dominance relation $\widetilde{\reflectbox{D}}$, reflecting for a pair $(y,x) \in U \times U$ the possibility of $y \, \reflectbox{D} \, x$, is defined as

$$\widetilde{\reflectbox{D}}(y,x) = \prod_{q \in C} \widetilde{\succeq_q}(x,y). \tag{28}$$

3.4 Desirable properties of DRSA adapted to handle missing values

We consider the following desirable properties of DRSA-mv_j, where $j = 1, 1.5, 2, 2.5, 3, \ldots, 6$:

1. Property S (reflecting a specific kind of symmetry): DRSA-mv_j has property S iff $yD_jx \Leftrightarrow x \, \reflectbox{D}_j \, y$, for any $x, y \in U$.
2. Property R (reflecting reflexivity of dominance relations): DRSA-mv_j has property R iff xD_jx and $x \, \reflectbox{D}_j \, x$, for any $x \in U$.
3. Property T (reflecting transitivity of dominance relations): DRSA-mv_j has property T iff $yD_jx \wedge xD_jz \Rightarrow yD_jz$, and $y \, \reflectbox{D}_j \, x \wedge x \, \reflectbox{D}_j \, z \Rightarrow y \, \reflectbox{D}_j \, z$, for any $x, y, z \in U$.
4. Property B (robustness): let $C^x = \{q \in C : q(x) \neq *\}$; DRSA-$mv_j$ has property B iff the following two conditions hold simultaneously:
   - for each $x \in \underline{X_i^{\geq}}$, $\reflectbox{D}_j^{+\prime}(x) \cap \neg X_i^{\geq} \subseteq \reflectbox{D}_j^{+}(x) \cap \neg X_i^{\geq}$, where $\reflectbox{D}_j^{+\prime}(x)$ is a positive dominance cone with the origin in x w.r.t. relation $\reflectbox{D}_j$, defined in the C^x-evaluation space,
   - for each $x \in \underline{X_i^{\leq}}$, $D_j^{-\prime}(x) \cap \neg X_i^{\leq} \subseteq D_j^{-}(x) \cap \neg X_i^{\leq}$, where $D_j^{-\prime}(x)$ is a negative dominance cone with the origin in x w.r.t. relation D_j, defined in the C^x-evaluation space.
5. Property P (reflecting precisiation of data): DRSA-mv_j has property P iff the lower approximation of any $X \subseteq U$ does not shrink when any missing attribute value is replaced by some non-missing value.
6. Property RI (rough inclusion): DRSA-mv_j has property RI iff $\underline{X} \subseteq X \subseteq \overline{X}$, for any $X \subseteq U$.
7. Property C (complementarity): DRSA-mv_j has property C iff $\underline{X} = U \setminus \overline{\neg X}$, for any $X \subseteq U$.
8. Property M_1 (monotonicity w.r.t. growing set of attributes): DRSA-mv_j has property M_1 iff the lower approximation any $X \subseteq U$ does not shrink when set P is extended by new attributes.
9. Property M_2 (monotonicity w.r.t. growing union of classes): DRSA-mv_j has property M_2 iff for any $X \subseteq U$, the lower approximation of X does not shrink when this set is augmented by new objects.

10. Property M_3 (monotonicity w.r.t. super-union of classes): DRSA-mv_j has property M_3 iff given any two upward unions of classes $X_i^{\geq}, X_k^{\geq}$, with $1 \leq i < k \leq n$, there is $\underline{X_i^{\geq}} \supseteq \underline{X_k^{\geq}}$, and, moreover, given any two downward unions of classes $X_i^{\leq}, X_k^{\leq}$, with $1 \leq i < k \leq n$, there is $\underline{X_i^{\leq}} \subseteq \underline{X_k^{\leq}}$.
11. Property M_4 (monotonicity w.r.t. dominance relation): DRSA-mv_j has property M_4 iff the following two conditions hold simultaneously:
    - for any $X_i^{\geq} \subseteq U$, with $i \in \{1, \ldots, n\}$, and for any $x, y \in U$ such that $x \, D_j \, y$, it is true that $(x \in \underline{X_i^{\geq}} \wedge y \in X_i^{\geq} \Rightarrow y \in \underline{X_i^{\geq}})$;
    - for any $X_i^{\leq} \subseteq U$, with $i \in \{1, \ldots, n\}$, and for any $x, y \in U$ such that $x D_j y$, it is true that $(x \in \underline{X_i^{\leq}} \wedge y \in X_i^{\leq} \Rightarrow y \in \underline{X_i^{\leq}})$.

Comparing to the list of desirable properties introduced in [1], we propose new property B which postulates that an object x, belonging to the lower approximation of any union of classes when considering all condition attributes, should also belong to this approximation when considering only these attributes, for which evaluation of x is not missing. Moreover, we modify definition of property M_4 to reflect definitions of generalized lower approximations.

Note that there is a correspondence between the above properties M_1, M_2, M_3, and M_4, and monotonicity properties (m1), (m2), (m3), and (m4), introduced in [3]. However, in VC-DRSA-mv_j, it may happen that for some $k \in \{1, \ldots, 4\}$, (mk) is satisfied while M_k is not satisfied.

The properties of DRSA-mv_j, $j = 1, 1.5, 2, 2.5, 3, \ldots, 6$, are summarized in Table 2, where **T** and F denote presence and absence of a given property, respectively. Moreover, in case of two symbols $\cdot/\cdot$, the first one reflects only CDRSA-mv_j while the second one reflects only ϵ-VC-DRSA-mv_j.

Table 2. Properties of DRSA-mv_j, $j = 1, 1.5, 2, 2.5, 3, \ldots, 6$

Prop. / Approach	DRSA-mv_1	DRSA-$mv_{1.5}$	DRSA-mv_2	DRSA-$mv_{2.5}$	DRSA-mv_3	DRSA-mv_4	DRSA-mv_5	DRSA-mv_6
S	F	F	**T**	**T**	F	**T**	**T**	**T**
R	F	**T**	**T**	F	**T**	**T**	**T**	**T**
T	**T**	**T**	F	F	**T**	**T**	**T**	F
B	F	**T**	**T**	**T**	F	F	F	F
P	F	F	**T**	F	F	F	F	F
RI	F	**T**	**T**	F	**T**	**T**	**T**	**T**
C	**T**	**T**	**T**	**T**	**T**	**T**	**T**	**T**
M_1	**T**	**T**	**T**	F	**T**	**T**	**T**	**T**
M_2	**T**	**T**	**T**	**T**	**T**	**T**	**T**	**T**
M_3	**T**/F	**T**/F	**T**/F	**T**/F	**T**/F	**T**/F	**T**/F	**T**/F
M_4	**T**	**T**	F	F	**T**	**T**	**T**	F

According to Table 2, DRSA-$mv_{2.5}$ is the least attractive due to lack of many important properties (R, T, P, RI, M_1, and M_4). DRSA-mv_1 is dominated by: DRSA-$mv_{1.5}$, DRSA-mv_3, DRSA-mv_4, and DRSA-mv_5. DRSA-mv_3 is dominated by: DRSA-$mv_{1.5}$, DRSA-mv_4, and DRSA-mv_5. DRSA-mv_6 is dominated by: DRSA-mv_2, DRSA-mv_4, and DRSA-mv_5. The only non-dominated approaches are DRSA-$mv_{1.5}$, DRSA-mv_2, DRSA-mv_4, and DRSA-mv_5.

4 Conclusions

We considered different ways of dealing with missing attribute values in ordinal and non-ordinal classification data when analyzed using Indiscernibility-based Rough Set Approach (IRSA) or Dominance-based Rough Set Approach (DRSA). Moreover, we proposed some desirable properties for IRSA and DRSA that a rough set approach capable of dealing with missing attribute values should possess. We analyzed which of these properties are satisfied by the considered rough set approaches resulting from different definitions of indiscernibility or dominance relations, suitable for the case of missing values. Based on this analysis, we uncovered some non-dominated, with respect to desirable properties, indiscernibility-based and dominance-based rough set approaches. These are:

- in IRSA: IRSA-$mv_{1.5}$ and IRSA-mv_2,
- in DRSA: DRSA-$mv_{1.5}$, DRSA-mv_2, DRSA-mv_4, and DRSA-mv_5.

Our future work will focus on experimental comparison of non-dominated variants uncovered in this paper. One of them, called DRSA-mv_2, was already compared with respect to classification performance against some other ordinal and non-ordinal classifiers. The results reported in [1] show that DRSA-mv_2-based rule classifier performs better than other well known methods like: Naive Bayes, SVM, Ripper, or C4.5 when the share of missing values in a data set is below 20%.

Acknowledgment The first author acknowledges financial support from the Poznań University of Technology, grant no. 09/91/DSMK/0609.

Bibliography

[1] Błaszczyński, J., Słowiński, R., Szeląg, M.: Induction of ordinal classification rules from incomplete data. In: Yao, J., et al. (eds.) RSCTC 2012. LNAI, vol. 7413, pp. 56–65. Springer (2012)
[2] Błaszczyński, J., Greco, S., Słowiński, R., Szeląg, M.: Monotonic variable consistency rough set approaches. In: Yao, J., et al. (eds.) RSKT 2007. LNAI, vol. 4481, pp. 126–133. Springer-Verlag, Berlin Heidelberg (2007)
[3] Błaszczyński, J., Greco, S., Słowiński, R., Szeląg, M.: Monotonic variable consistency rough set approaches. International Journal of Approximate Reasoning 50(7), 979–999 (2009)
[4] Błaszczyński, J., Słowiński, R., Szeląg, M.: Rough set approach to classification of incomplete data. Research Report RA-22/2013, Poznań University of Technology (2013)
[5] Dembczyński, K., Greco, S., Słowiński, R.: Rough set approach to multiple criteria classification with imprecise evaluations and assignments. European Journal of Operational Research 198(2), 626–636 (2009)

[6] Greco, S., Matarazzo, B., Słowiński, R.: Handling missing values in rough set analysis of multi-attribute and multi-criteria decision problems. In: Zhong, N., et al. (eds.) RSFDGrC'99. LNCS, vol. 1711, pp. 146–157. Springer, Berlin (1999)

[7] Greco, S., Matarazzo, B., Słowiński, R.: Dealing with missing data in rough set analysis of multi-attribute and multi-criteria decision problems. In: Zanakis, S., et al. (eds.) Decision Making: Recent Developments and Worldwide Applications, pp. 295–316. Kluwer, Dordrecht (2000)

[8] Greco, S., Matarazzo, B., Słowiński, R.: Rough sets theory for multicriteria decision analysis. European J. of Operational Research 129(1), 1–47 (2001)

[9] Greco, S., Matarazzo, B., Słowiński, R.: Granular computing for reasoning about ordered data: the dominance-based rough set approach. In: Pedrycz, W., et al. (eds.) Handbook of Granular Computing, chap. 15. Wiley, Chichester (2008)

[10] Grzymala-Busse, J.W., Hu, M.: A comaprison of several approaches in missing attribute values in data mining. In: Ziarko, W., Yao, Y. (eds.) RSCTC 2000. LNAI, vol. 2005, pp. 378–385. Springer, Berlin (2001)

[11] Grzymala-Busse, J.W.: Mining incomplete data – a rough set approach. In: Yao, J.T., et al. (eds.) RSKT 2011. LNCS, vol. 6954, pp. 1–7. Springer, Berlin (2011)

[12] Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, Berlin (2009)

[13] Hu, M.L., Liu, S.F.: A rough analysis method of multi-attribute decision making for handling decision system with incomplete information. In: Proceedings of 2007 IEEE International Conference on Grey Systems and Intelligent Services, November 18-20, 2007, Nanjing, China (2007)

[14] Kryszkiewicz, M.: Rough set approach to incomplete information systems. Information Sciences 112, 39–49 (1998)

[15] Liang, D., Yang, S.X., Jiang, C., Zheng, X., Liu, D.: A new extended dominance relation approach based on probabilistic rough set theory. In: Yu, J., et al. (eds.) RSKT 2010. LNAI, vol. 6401, pp. 175–180. Springer (2010)

[16] Pawlak, Z.: Rough Sets. Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht (1991)

[17] Słowiński, R., Greco, S., Matarazzo, B.: Rough set methodology for decision aiding. In: Kacprzyk, J., Pedrycz, W. (eds.) Handbook of Computational Intelligence, chap. 22, pp. 349–370. Springer, Berlin (2015)

[18] Słowiński, R., Vanderpooten, D.: A generalized definition of rough approximations based on similarity. IEEE Transactions on Knowledge and Data Engineering 12(2), 331–336 (2000)

[19] Stefanowski, J., Tsoukias, A.: Incomplete information tables and rough classification. Computational Intelligence 17(3), 545–566 (2001)

[20] Yang, X., Yang, J., Wu, C., Yu, D.: Dominance-based rough set approach and knowledge reductions in incomplete ordered information system. Information Sciences 178(4), 1219–1234 (2008)