Rough Set Analysis of Classification Data with Missing Values

Zastosowanie teorii zbiorów przybliżonych do analizy danych klasyfikacyjnych z brakującymi wartościami atrybutów

Marcin Szeląg

Institute of Computing Science, Poznań University of Technology

06.11.2019

Outline

Introduction

	outlook	temperature	humidity	windy	play golf	
	sunny	hot	high	weak	no	
	sunny	hot	high	strong	no	
	rain	cool	normal	strong	no	X_1
y	sunny	mild	high	weak	no	
	rain	mild	high	strong	no	
	overcast	hot	high	weak	yes	
	rain	mild	high	weak	yes	
	rain	cool	normal	weak	yes	
	overcast	cool	normal	strong	yes	
	sunny	cool	normal	weak	yes	X_2
	rain	mild	normal	weak	yes	
x	sunny	mild	high	weak	yes	
	overcast	mild	high	strong	yes	
	overcast	hot	normal	weak	yes	
	q_1	q_2	q_3	q_4	d	
	for all q in C, $q(x)=q(y) \Leftrightarrow y\,I\,x \Leftrightarrow y \in I(x)$				$d(x) \neq d(y)$	

Classical classification problem + MV

	outlook	temperature	humidity	windy	play golf	
	sunny	hot	high	weak	no	
	*	hot	high	*	no	
	rain	cool	normal	strong	no	X_1
y	sunny	*	high	weak	no	
	rain	mild	high	strong	no	
	overcast	hot	high	weak	yes	
	rain	mild	high	weak	yes	
	rain	cool	normal	weak	yes	
	overcast	*	normal	strong	yes	
	sunny	cool	normal	*	yes	X_2
	rain	mild	normal	weak	yes	
x	sunny	mild	*	weak	yes	
	overcast	mild	high	strong	yes	
	*	hot	normal	weak	yes	
	q_1	q_2	q_3	q_4	d	
	are y and x comparable?				$d(x) \neq d(y)$	

Ordinal classification with monotonicity constraints

	buying	maint	doors	persons	lug_boot	safety	class	
	vhigh	vhigh	2	2	small	med	unacc	X_1
	med	vhigh	3	more	small	med	unacc	
	vhigh	high	2	4	med	low	unacc	
	vhigh	high	2	4	big	low	unacc	
	med	low	2	4	big	low	unacc	
y	low	low	4	more	big	high	unacc	
	high	med	2	more	med	high	acc	X_2
	med	vhigh	3	more	med	med	acc	
	med	vhigh	3	more	med	high	acc	
	med	vhigh	3	more	big	med	acc	
	med	vhigh	3	more	big	high	acc	
	low	low	4	more	small	med	acc	
x	low	low	2	more	big	med	good	X_3
	low	low	4	more	small	high	good	
	low	low	4	more	big	med	good	
	med	med	4	more	med	high	vgood	X_4
	med	low	2	4	big	high	vgood	
	low	low	4	more	big	high	vgood	
	q_1	q_2	q_3	q_4	q_5	q_6	d	

$$\text{for all } q \in C, y \succeq_q x \Leftrightarrow \begin{cases} yDx \\ x \not D\, y \end{cases} \Leftrightarrow \begin{cases} y \in D^+(x) \\ x \in D^-(y) \end{cases} \quad d(y) \prec d(x)$$

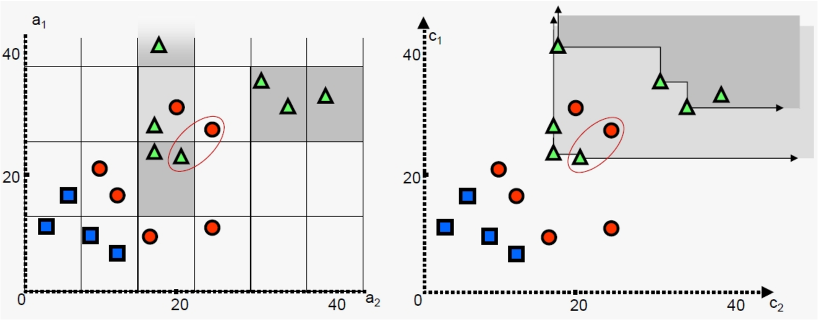
Ordinal classification with monotonicity constraints + MV

	buying	maint	doors	persons	lug_boot	safety	class	
	vhigh	vhigh	2	2	small	med	unacc	X_1
	med	*	3	more	small	med	unacc	
	vhigh	high	2	4	*	low	unacc	
	vhigh	high	2	4	big	low	unacc	
	med	low	2	4	big	low	unacc	
y	low	*	**4**	more	big	*	**unacc**	
	high	med	2	more	med	high	acc	X_2
	med	vhigh	3	more	*	med	acc	
	*	vhigh	3	more	med	high	acc	
	med	vhigh	*	more	big	med	acc	
	med	vhigh	3	more	big	high	acc	
	low	low	4	more	small	med	acc	
x	low	*	*	*	big	**med**	**good**	X_3
	low	low	4	more	small	high	good	
	low	low	4	*	big	med	good	
	med	med	4	more	med	high	vgood	X_4
	med	low	2	4	big	high	vgood	
	low	low	4	more	big	high	vgood	
	q_1	q_2	q_3	q_4	q_5	q_6	d	
	are y and x comparable?						$d(y) < d(x)$	

Research motivation

- Handle missing values of attributes in indiscernibility-based (classical) rough set approach (IRSA).
- Handle missing values of criteria in dominance-based rough set approach (DRSA).
- Introduce a set of desirable properties to compare different adaptations od IRSA and DRSA.
- Verify which properties are satisfied by which adaptation.
- Uncover approaches non-dominated w.r.t. desirable properties.
- Compare non-dominated approaches experimentally.

IRSA vs DRSA



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

$$\overline{X_i} = \bigcup_{x \in X_i} I(x)$$

approximation of classes

$$\underline{X_i^{\geq}} = \{x \in U : D^+(x) \subseteq X_i^{\geq}\}$$

$$\overline{X_i^{\geq}} = \bigcup_{x \in X_i^{\geq}} D^+(x)$$

approximation of unions of ordered classes

Adaptations of IRSA to Handle Missing Values

IRSA analysis of classification data with MV

Adaptations:

- redefinition of the indiscernibility relation $I \Rightarrow I_j$, IRSA-mv_j (j stands for the version id)
- as I_j may miss some properties, like symmetry or transitivity, we employ generalized definitions of rough approximations[a], where indiscernibility relation is only assumed to be reflexive:

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

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

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

[a] Słowiński, R., Vanderpooten, D.: A generalized definition of rough approximations based on similarity. IEEE Trans. Knowl. Data Eng. 12(2), 331–336 (2000)

IRSA-mv_1

- employs indiscernibility relation I_1[a] considered as a directional statement where a subject y is compared to a referent x which cannot have missing values:

$$yI_1x \Leftrightarrow \text{ for each } q \in C : \begin{cases} q(x) \neq * \\ q(y) = q(x) \text{ or } q(y) = * \end{cases}$$

	outlook	temperature	humidity	windy
y	sunny	*	high	*
x	sunny	hot	high	strong
	$y I_1 x$			

[a] 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)

- IRSA-$mv_{1.5}$[a] is an improvement over IRSA-mv_1
- $I_{1.5}$ – reflexive and transitive similarity relation; referent can have missing values:

$$yI_{1.5}x \Leftrightarrow q(y) = q(x) \text{ for each } q \in C \text{ such that } q(y) \neq *$$

	outlook	temperature	humidity	windy
y	sunny	*	high	*
x	sunny	hot	high	*
	$y\ I_{1.5}\ x$			

- this approach is treating missing values as "lost" ones!

[a] Stefanowski, J., Tsoukias, A.: Incomplete information tables and rough classification. Comput. Intell. 17(3), 545–566 (2001)

IRSA-mv_2

- IRSA-mv_2[a] employs a reflexive and symmetric tolerance relation:

$$yI_2x \Leftrightarrow \text{for each } q \in C \text{ there is } q(y) = q(x), \text{ or } q(y) = *, \text{ or } q(x) = *$$

	outlook	temperature	humidity	windy
y	sunny	*	high	*
x	sunny	hot	*	*
	$y\ I_2\ x$			

- this approach is treating missing values as "do not care" ones!

[a] Kryszkiewicz, M.: Rough set approach to incomplete information systems. Inf. Sci. 112, 39–49 (1998)

IRSA-mv_3

- IRSA-mv_3 is introduced in the presented paper
- I_3 – reflexive and transitive similarity relation; referent can have missing values:

$$yI_3x \Leftrightarrow q(y) = q(x) \text{ for each } q \in C \text{ such that } q(x) \neq *$$

	outlook	temperature	humidity	windy
y	sunny	hot	high	*
x	sunny	*	high	*
	$y\,I_3\,x$			

Desirable Properties of IRSA Adapted to Handle Missing Values

Desirable properties of IRSA-mv_j (X – any class X_i)

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

Desirable properties of IRSA-mv_j

Table: 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**
M_1	**T**	**T**	**T**	**T**
M_2	**T**	**T**	**T**	**T**
MT	**T**	**T**	*F*	**T**
		non-dominated!	non-dominated!	

Adaptations of DRSA to Handle Missing Values

DRSA analysis of classification data with MV

Adaptations:

- redefinition of the dominance relations D, $\mathbb{D} \Rightarrow D_j$, $\mathbb{D}_j$, DRSA-mv_j (j stands for the version id)
- as D_j, $\mathbb{D}_j$ may miss some properties, like transitivity, we employ generalized definitions of rough approximations[a], where dominance relations are only assumed to be reflexive:

$$\mathbb{D}_j^+(x) = \{y \in U : x \, \mathbb{D}_j \, y\} \qquad \mathbb{D}_j^-(x) = \{y \in U : y \, \mathbb{D}_j \, x\}$$

$$\underline{X_i^{\geq}} = \{x \in U : \mathbb{D}_j^+(x) \subseteq X_i^{\geq}\} \qquad \underline{X_i^{\leq}} = \{x \in U : D_j^-(x) \subseteq X_i^{\leq}\}$$

$$\overline{X_i^{\geq}} = \bigcup_{x \in X_i^{\geq}} D_j^+(x) \qquad \overline{X_i^{\leq}} = \bigcup_{x \in X_i^{\leq}} \mathbb{D}_j^-(x)$$

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

- for D_1, $\reflectbox{D}_1$[a], referent x cannot have missing values:

$$yD_1x \Leftrightarrow \text{ for each } q \in C : \begin{cases} q(x) \neq * \\ y \succeq_q x \text{ or } q(y) = * \end{cases}$$

$$z \, \reflectbox{D}_1 \, x \Leftrightarrow \text{ for each } q \in C : \begin{cases} q(x) \neq * \\ x \succeq_q z \text{ or } q(z) = * \end{cases}$$

	buying	maint	doors	persons
y	med	*	4	4
x	med	high	3	4
z	med	high	*	2

[a] 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)

DRSA-$mv_{1.5}$

- DRSA-$mv_{1.5}$[a] is an improvement over DRSA-mv_1
- $D_{1.5}$, $D^{-1}_{1.5}$ – referent x can have missing values:

$$yD_{1.5}x \Leftrightarrow y \succeq_q x \text{ for each } q \in C \text{ such that } q(y) \neq *$$
$$z\, D^{-1}_{1.5}\, x \Leftrightarrow x \succeq_q z \text{ for each } q \in C \text{ such that } q(z) \neq *$$

	buying	maint	doors	persons
y	*	*	4	4
x	*	high	3	4
z	*	high	*	2

- this approach is treating missing values as "lost" ones!

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

- DRSA-mv_2[a] employs relations D_2, $\mathit{Ɑ}_2$:

$$yD_2x \Leftrightarrow \forall q \in C \text{ there is: } y \succeq_q x, \text{ or } q(y) = *, \text{ or } q(x) = *$$

$$z\, \mathit{Ɑ}_2\, x \Leftrightarrow \forall q \in C \text{ there is: } x \succeq_q z, \text{ or } q(z) = *, \text{ or } q(x) = *$$

	buying	maint	doors	persons
y	med	*	4	*
x	*	*	3	4
z	med	*	*	2

- this approach is treating missing values as "do not care" ones!

[a] 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)

DRSA-$mv_{2.5}$

- in DRSA-$mv_{2.5}$[a], dominance relations $D_{2.5}$, $D^{\!\!\!/}_{2.5}$ are defined similarly as in DRSA-mv_2
- just an additional condition:

$$\frac{|\text{set of ``common'' attributes for } x \text{ and } y|}{|C|} \geq \lambda,$$

where: "common" attributes = attributes for which both x and y have a non-missing value; $\lambda \in [0, 1]$

[a]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, 18-20, Nov. 2007, Nanjing, China

- DRSA-mv_3[a] employs relations D_3, $\reflectbox{$D$}_3$:

$$yD_3x \Leftrightarrow y \succeq_q x \text{ for each } q \in C \text{ such that } q(x) \neq *$$
$$z\,\reflectbox{D}_3\,x \Leftrightarrow x \succeq_q z \text{ for each } q \in C \text{ such that } q(x) \neq *$$

	buying	maint	doors	persons
y	*	high	4	4
x	*	*	3	4
z	*	high	3	2

[a] Błaszczyński, J., Słowiński, R., Szeląg, M.: Induction of ordinal classification rules from incomplete data. In: Yao et al. (eds.) RSCTC 2012. LNCS (LNAI), vol. 7413, pp. 56–65, Springer, Heidelberg (2012)

- DRSA-mv_4[a] employs relations D_4, $\not D_4$:

$$yD_4x \Leftrightarrow \forall q \in C : y \succeq_q x, \text{ or } q(x) = *, \text{ or } q(x) = \inf(V_q)$$
$$z \not D_4 x \Leftrightarrow \forall q \in C : x \succeq_q z, \text{ or } q(z) = *, \text{ or } q(z) = \inf(V_q),$$

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

	buying	maint	doors	persons
y	*	high	4	4
x	*	*	3	4
z	*	vhigh	*	2

[a] Dembczyński, K., Greco, S., Słowiński, R.: Rough set approach to multiple criteria classification with imprecise evaluations and assignments. Eur. J. Oper. Res. 198(2), 626–636 (2009)

- DRSA-mv_5[a] employs relations D_5, $\mathit{\Pi}_5$:

$$yD_5x \Leftrightarrow \forall q \in C : y \succeq_q x, \text{ or } q(y) = *, \text{ or } q(y) = \sup(V_q)$$
$$z\,\mathit{\Pi}_5\,x \Leftrightarrow \forall q \in C : x \succeq_q z, \text{ or } q(x) = *, \text{ or } q(x) = \sup(V_q),$$

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

	buying	maint	doors	persons
y	*	*	4	4
x	*	low	3	4
z	*	*	3	2

[a] Dembczyński, K., Greco, S., Słowiński, R.: Rough set approach to multiple criteria classification with imprecise evaluations and assignments. Eur. J. Oper. Res. 198(2), 626–636 (2009)

- DRSA-mv_6[a], employs relations D_6, $\reflectbox{$D$}_6$:

$$D_6 = \{(y,x) \in U \times U : \widetilde{D}(y,x) \geq \alpha\} \cup \{(x,x) : x \in U\}$$
$$\reflectbox{D}_6 = \{(z,x) \in U \times U : \widetilde{\reflectbox{D}}(z,x) \geq \alpha\} \cup \{(x,x) : x \in U\},$$

where fuzzy relation $\widetilde{D}$ ($\widetilde{\reflectbox{$D$}}$) is such that $\widetilde{D}(y,x)$ (resp. $\widetilde{\reflectbox{$D$}}(z,x)$) reflects a probability of yDx (resp. $z \, \reflectbox{$D$} \, x$), for $y,x,z \in U$; threshold $\alpha \in [0,1]$

- details $\rightarrow$ proceedings

---

[a] 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., Greco, S., Lingras, P., Wang, G., Skowron, A. (eds.) RSKT 2010. LNCS (LNAI), vol. 6401, pp. 175–180. Springer, Heidelberg (2010)

Desirable Properties of DRSA Adapted to Handle Missing Values

1. S (reflecting "symmetry"): $yD_jx \Leftrightarrow x\,\text{Ɑ}_j\,y$, for any $x, y \in U$
2. R (reflexivity): xD_jx and $x\,\text{Ɑ}_j\,x$, for any $x \in U$
3. T (transitivity): $yD_jx \wedge xD_jz \Rightarrow yD_jz$, and $y\,\text{Ɑ}_j\,x \wedge x\,\text{Ɑ}_j\,z \Rightarrow y\,\text{Ɑ}_j\,z$, for any $x, y, z \in U$
4. B (robustness): ($C^x = \{q \in C : q(x) \neq *\}$)
   - for each $x \in \underline{X_i^{\geq}}$, $\text{Ɑ}_j^{+\prime}(x) \cap \neg X_i^{\geq} \subseteq \text{Ɑ}_j^{+}(x) \cap \neg X_i^{\geq}$, where $\text{Ɑ}_j^{+\prime}(x)$ is a positive dominance cone with the origin in x w.r.t. relation Ɑ_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. P (reflecting precisiation of data): $\underline{X}$ does not shrink when any missing attribute value becomes known, $\forall X \subseteq U$
6. RI (rough inclusion): $\underline{X} \subseteq X \subseteq \overline{X}$, $\forall X \subseteq U$
7. C (complementarity): $\underline{X} = U \setminus \overline{\neg X}$, $\forall X \subseteq U$

Desirable properties of DRSA-mv_j (X – any union $X_i^{\geq}/X_i^{\leq}$)

8. M_1 (monotonicity w.r.t. growing set of attributes): $\underline{X}$ does not shrink when set C is extended by new attributes, $\forall X \subseteq U$
9. M_2 (monotonicity w.r.t. growing union of classes): $\underline{X}$ does not shrink when set X is augmented by new objects, $\forall X \subseteq U$
10. M_3 (monotonicity w.r.t. super-union of classes): $\forall X_i^{\geq}, X_k^{\geq}$, with $1 \leq i < k \leq n$, there is $\underline{X_i^{\geq}} \supseteq \underline{X_k^{\geq}}$, and, moreover, $\forall X_i^{\leq}, X_k^{\leq}$, with $1 \leq i < k \leq n$, there is $\underline{X_i^{\leq}} \subseteq \underline{X_k^{\leq}}$
11. M_4 (monotonicity w.r.t. dominance relation):
    - for any $X_i^{\geq} \subseteq U$, with $i \in \{1, \ldots, n\}$, and for any $x, y \in U$ such that $x \, \reflectbox{$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 $(x \in \underline{X_i^{\leq}} \wedge y \in X_i^{\leq} \Rightarrow y \in \underline{X_i^{\leq}})$

Desirable properties of DRSA-mv_j

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

Prop/Appr	-mv_1	-$mv_{1.5}$	-mv_2	-$mv_{2.5}$	-mv_3	-mv_4	-mv_5	-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	T	T	T	T	T	T	T
M_4	T	T	F	F	T	T	T	F
		n-d!	n-d!			n-d!	n-d!	

Conclusions

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)
- We proposed some desirable properties for IRSA and DRSA that a rough set approach capable of dealing with missing attribute values should possess
- We uncovered approaches non-dominated with respect to these desirable properties:
  - 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.

References

- M. Szeląg, J. Błaszczyński, R. Słowiński, Rough Set Analysis of Classification Data with Missing Values. [In]: L. Polkowski et al. (Eds.): Rough Sets, International Joint Conference, IJCRS 2017, Olsztyn, Poland, July 3–7, 2017, Proceedings, Part I. LNAI, vol. 10313, Springer, 2017, pp. 552–565 $\Rightarrow$ manuscript.