UNIVERSITY OF WARSAW

SECURITY ON-DEMAND

Rough Sets in Data Mining & Databases: Foundations & Applications

Tutorial @ IJCRS 2018

Dominik Ślęzak & Arkadiusz Wojna

(Part I)

Rough Sets

- The theory of rough sets founded in early 80-ties by Professor Zdzisław Pawlak provides the means for handling incompleteness & uncertainty in data
- In the process of knowledge discovery, one can search for *decision reducts*, which are irreducible subsets of attributes that determine decision values
- Dependencies in data can be expressed in terms of, e.g., *discernibility* or *rough set approximations*
- There are also rough-set-inspired computational models, such as *rough clustering*, *rough SQL*, etc.

Rough Sets & Data

	Outlook	Temp.	Humid.	Wind	Sport?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cold	Normal	Weak	Yes
6	Rain	Cold	Normal	Strong	No
7	Overcast	Cold	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cold	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

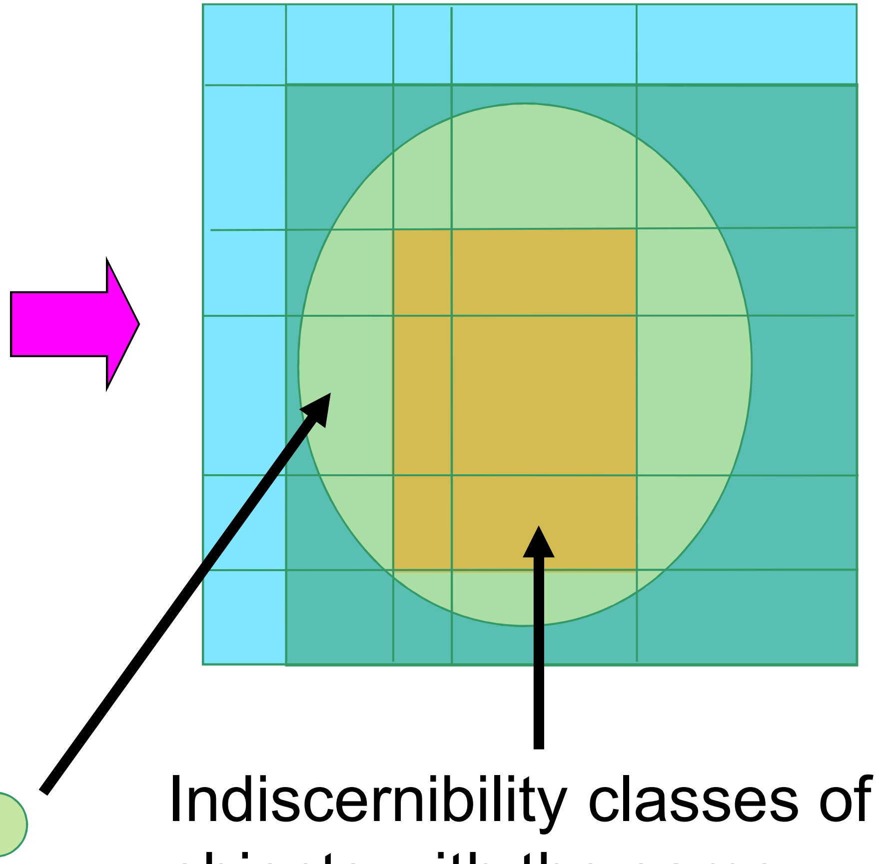
IF (H=Normal) AND (T=Mild) THEN (S=Yes)

It corresponds to a data block included in the positive region of the decision class "Yes"

Rules & Approximations

$$\mathbb{A} = (U, A \cup \{d\})$$

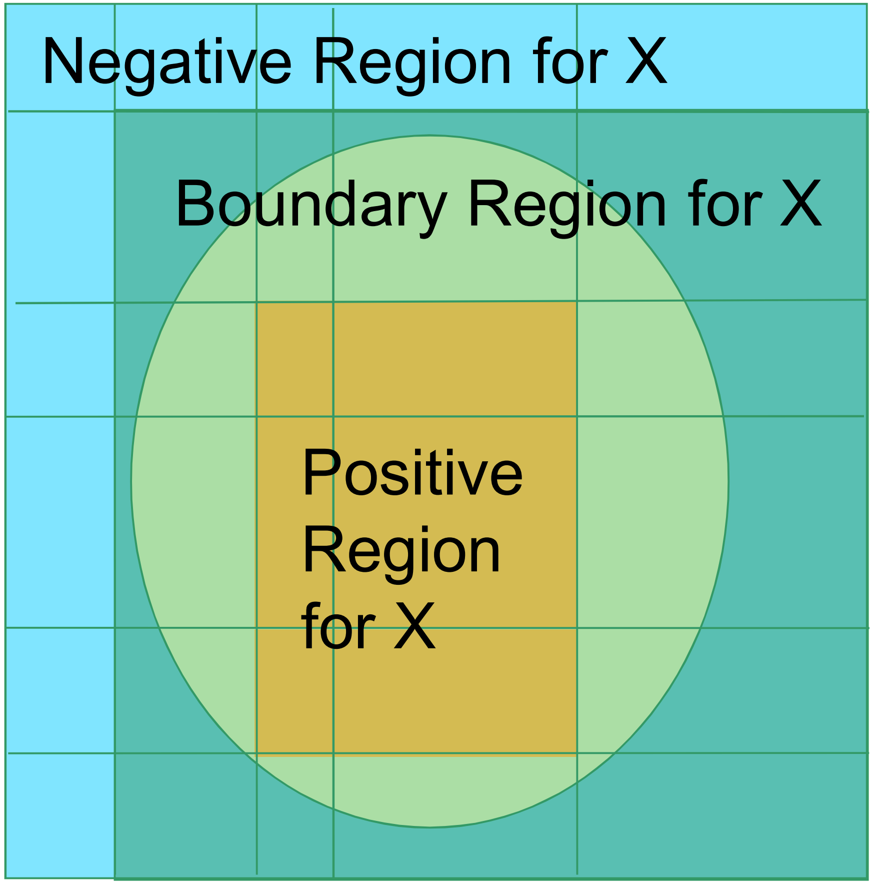
	Outlook	Temp.	Humid.	Wind	Sport?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cold	Normal	Weak	Yes
6	Rain	Cold	Normal	Strong	No
7	Overcast	Cold	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cold	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No



Sport? = Yes

Indiscernibility classes of objects with the same values of some attributes

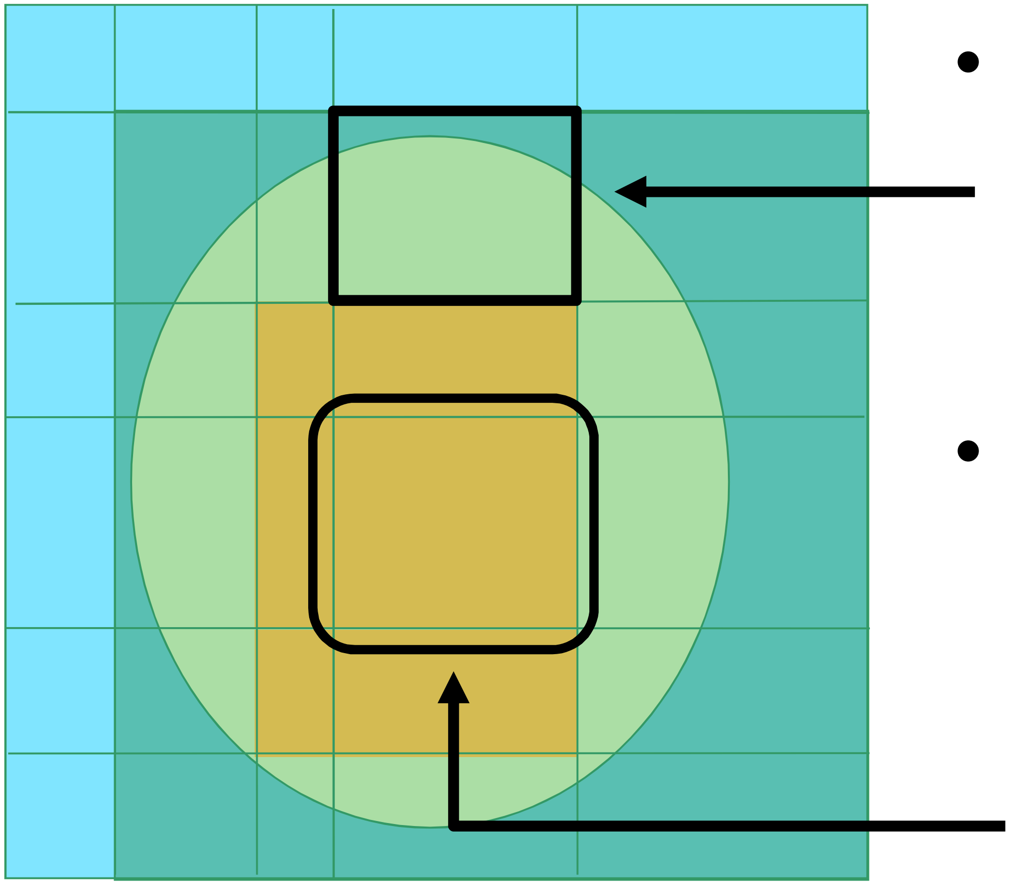
Approximations & Regions



Lower Approximation: Objects certainly in X (the exact rules for X)

Upper Approximation: Objects that may be in X (the rules, which don't exclude X)

Approximations - Extensions



- Indiscernibility classes can be almost in X (VPRS model)
- It does not need to be based on equivalences (DRSA, tolerance, covering models)

Rough Sets in Data Mining & Databases: Foundations & Applications

The Idea of Reducts

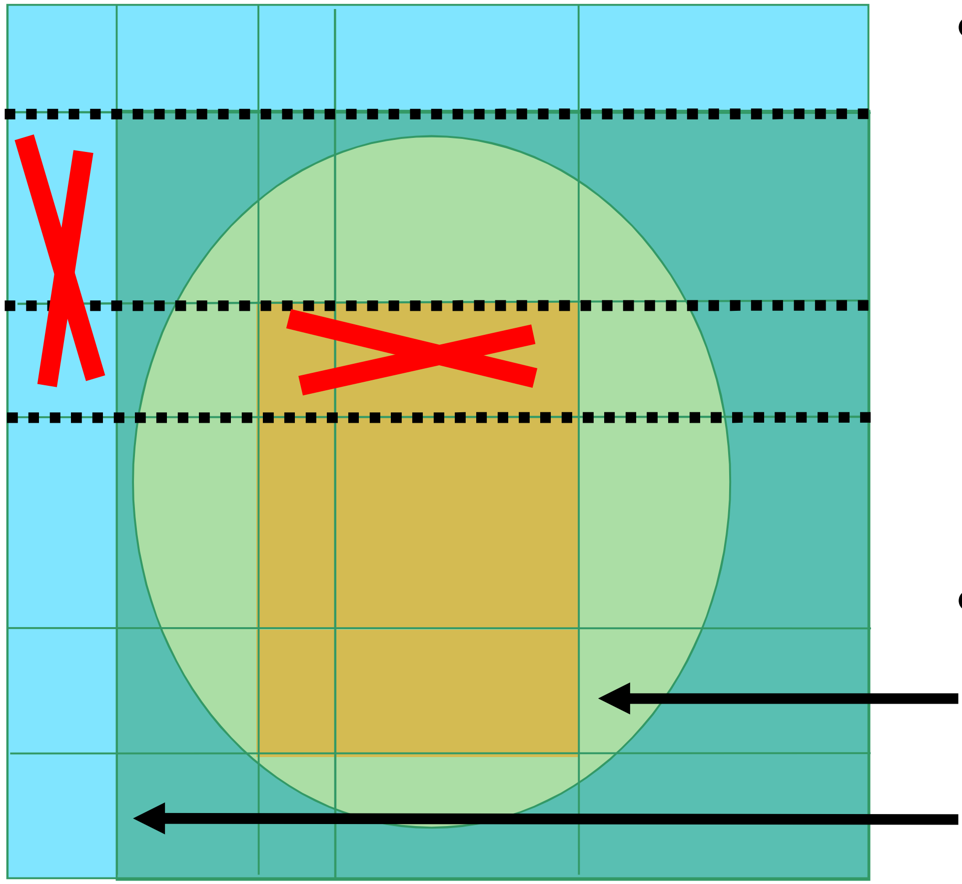
Attribute Selection & Reducts

- *Do we need all attributes?*
- *Do we need to store the entire data?*
- *Is it possible to avoid a costly test?*

Reducts are minimal subsets of attributes which contain a necessary portion of information from the set of all attributes. They are, however, (NP-)hard to find.

- Efficient and robust heuristics exist for reduct construction task
- Searching for reducts may be done efficiently with the use of, for example, evolutionary computation
- Overfitting can be avoided by considering several reducts, pruning rules and lessening constraints for keeping information

Attribute Selection & Approximations



- Approximations will (<u>or will not!!!</u>) change if we use a different subset of attributes to produce them
- Positive region generated by smaller subsets may decrease

Selection, Extraction, Reduction…

- Many tools for extracting possibly minimal amount of new features from original data
- For example, PCA provides new features as linear combinations of original features
- However, linear combinations still involve many original attributes in their definitions
- It would be better to start with rough set attribute reduction and then apply PCA

Illustration: Rules for {O,H,T,W}

- There are 14 rules supported in data
- However, the number of all possible combinations of conditions is 36
- We would not know how to classify some new cases with unseen combinations
- For instance:

    O=Sunny, T=Hot, H=Normal, W=Weak

Illustration: Rules for {O,H,W}

- O=Sunny & H=High & W=Weak => S=No
- O=Sunny & H=High & W=Strong => S=No
- O=Overcast & H=High & W=Weak => S=Yes
- O=Rain & H=High & W=Weak => S=Yes
- O=Rain & H=Normal & W=Weak => S=Yes
- O=Rain & H=Normal & W=Strong => S=No
- O=Overcast & H=Normal & W=Strong => S=Yes
- O=Sunny & H=Normal & W=Weak => S=Yes
- O=Sunny & H=Normal & W=Strong => S=Yes
- O=Overcast & H=High & W=Strong => S=Yes
- O=Overcast & H=Normal & W=Weak => S=Yes
- O=Rain & H=High & W=Strong => S=No

ALL COMBINATIONS !!!

Rough Sets in Data Mining & Databases: Foundations & Applications

The Idea of Discernibility

How to Discern Between Objects?

	Outlook	Temp.	Humid.	Wind	Sport?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cold	Normal	Weak	Yes
6	Rain	Cold	Normal	Strong	No
7	Overcast	Cold	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cold	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

{O,T,H} is not enough: it doesn't discern (5,6)

{T,H,W} is not enough: it doesn't discern (6,7)

{O,W} is not enough: it doesn't discern (8,9)

The only reducts are {O,T,W} and {O,H,W}. They discern all the pairs of objects with different decisions and cannot be further reduced.

Discernibility Matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1														
2														
3	O	O W												
4	O T	O T W												
5	O T H	O T H W												
6			O T H W	T H W	W									
7	O T H W	O T H				O								
8			O T	O	O T H		O T H W							
9	T H	T H W				O W		T H						
10	O T H	O T H W				T W		O H						
11	T H W	T H				O T		H W						
12	O T W	O T				O T H		O W						
13	O H	O H W				O T W		O T H						
14			O T W	W	T H W		O T H		O T H W	H W	O H	O	O T H W	

What Do We Want to Discern?

- Generalized decision function generated by subset of attributes $B \subseteq A$ labels each object $u \in U$ with a set of its *possible* decision values:

$$\partial_B(u) = \{ d(x) : \forall_{a \in B}\ a(u) = a(x) \}$$

- ∂-reduct is an irreducible subset $B \subseteq A$ such that:

$$\forall_{u \in U}\ \partial_B(u) = \partial_A(u)$$

  or **equivalently**:

$$\forall_{x,y \in U}\ \partial_A(x) \neq \partial_A(y) \rightarrow \exists_{a \in B}\ a(x) \neq a(y)$$

- $B \subseteq A$ is ∂-reduct, if and only if it is an irreducible subset of attributes such that a *multi-valued dependency* (MVD) $B \rightarrow\rightarrow \{d\}$ holds in data

Case Study: Survival Analysis

u	#	ttr	st_l	st_{cr}	loc	$\lvert [u]_C \rvert$	$\lvert [u]_C \cap def \rvert$	$\lvert [u]_C \cap unk \rvert$	$\lvert [u]_C \cap suc \rvert$
0	1	*only*	*T*3	*cN*1	*larynx*	25	15	4	6
4	1	*after*	*T*3	*cN*1	*larynx*	38	8	18	12
24	1	*radio*	*T*3	*cN*1	*larynx*	23	6	7	10
28	1	*after*	*T*3	*cN*0	*throat*	18	4	8	6
57	1	*after*	*T*4	*cN*1	*larynx*	32	12	14	6
91	1	*after*	*T*3	*cN*1	*throat*	35	5	16	14
152	1	*only*	*T*3	*cN*0	*larynx*	27	9	14	4
255	1	*after*	*T*3	*cN*0	*larynx*	15	2	6	7
493	1	*after*	*T*3	*cN*1	*other*	19	6	7	6
552	2	*after*	*T*4	*cN*2	*larynx*	14	6	3	5

In this case we operated with distributions of *rough membership functions* (data-derived probabilities):

$$\mu_d^C(u) = \left\langle \frac{\lvert [u]_C \cap def \rvert}{\lvert [u]_C \rvert}, \frac{\lvert [u]_C \cap unk \rvert}{\lvert [u]_C \rvert}, \frac{\lvert [u]_C \cap suc \rvert}{\lvert [u]_C \rvert} \right\rangle$$

It is not only about Discernibility Matrices…

- Selection Constraints:
  - Keep sufficiently good appro-ximations of decision classes
  - Discern between (almost) all pairs of objects with different decision values

- Optimization Goals:
  - Find subsets which induce minimum amount of rules
  - Find ensembles of subsets which work well together

- Algorithms & Structures:
  - Greedy, randomized, nature-inspired, Boolean, working on feature clusters
  - Discernibility matrices, sorting, hashing, MapReduce, FPGA, SQL-based scripts

Rough Sets in Data Mining & Databases: Foundations & Applications

Approximate Attribute Reduction

Fundamental Idea

- It is worth reducing irrelevant attributes and simplifying obtained decision rules
- Reduction (simplification) should not decrease the overall accuracy of rules
- In real-world situations, we may agree to ***slightly*** decrease the quality, if it leads to ***significantly*** simpler classification model

Approximate Discernibility

- For highly inconsistent data, we can focus on discerning only these pairs of objects that have *significantly* distant distributions of rough membership functions – then, after attribute reduction, new distributions will be *close* to the original distributions
- For numeric data, we can employ fuzzy discernibility (dissimilarity) and request that discernibility degrees for pairs of objects with different decision values do not decrease *significantly* after reduction

Dependency Functions / Criteria

- We can specify a function

  $M: P(A) \to \Re$

  measuring influence of A's subsets on d.
- $B \subseteq A$ is an (M,ε)-approximate reduct, if

  $M(B) / M(A) \geq 1-\varepsilon$

  and none of its proper subsets holds it.
- It is important for M to be monotonic

  $M(B) \geq M(C) \qquad C \subseteq B$

Rough-Set-Inspired Examples

- Cardinality of positive region induced by B
- Number of pairs of objects with different decision values that are discerned by B
- Measures based on cardinalities of generalized decision functions:
  - „∂-gini index": $\frac{1}{|U|}\sum_{u\in U}\frac{1}{|\partial_B(u)|}$
  - „∂-conditional entropy": $\frac{1}{|U|}\sum_{u\in U}\log(|\partial_B(u)|)$
  - „∂-Dempster-Shafer": $\frac{1}{|U|}\sum_{u\in U}\frac{1}{2^{|\partial_B(u)|-1}}$

o-GA for Approximate Reducts

- *Genetic part*, where each chromosome encodes a permutation of the attributes
- *Heuristic part*, where permutations are put into the following algorithm

REDORD algorithm:

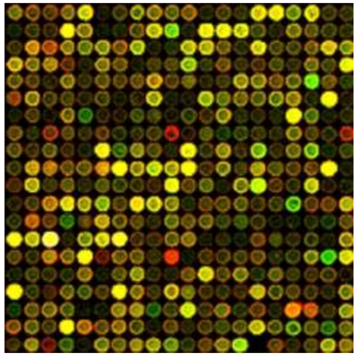
1. $\sigma:\{1,..,|A|\}\rightarrow\{1,..,|A|\}$, $B_\sigma = A$
2. For $i = 1$ to $|A|$ repeat 3 & 4
3. Let $B_\sigma \leftarrow B_\sigma \setminus \{a_{\sigma(i)}\}$
4. If not $B_\sigma \Rightarrow_\varepsilon d$ undo 3

$\Rightarrow_\varepsilon$ means the given attribute set determines *approximately* the decision **d**

Rough Sets in Data Mining & Databases: Foundations & Applications

High-Dimensional Data Sets

Case Study: Gene Expressions



	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Exp 6
Gene 1	-1.2	-2.1	-3	-1.5	1.8	2.9
Gene 2	2.7	0.2	-1.1	1.6	-2.2	-1.7
Gene 3	-2.5	1.5	-0.1	-1.1	-1	0.1
Gene 4	2.9	2.6	2.5	-2.3	-0.1	-2.3
Gene 5	0.1		2.6	2.2	2.7	-2.1
Gene 6	-2.9	-1.9	-2.4	-0.1	-1.9	2.9

- Thousands of genes-attributes to analyze
- Number of experiments-objects quite low
- Simple knowledge representation needed
  - Black-box approaches unacceptable
  - Standard discretization unacceptable
  - Rules too detailed for this level

	a	b	c	d
u1	3	7	3	0
u2	2	1	0	1
u3	4	0	6	1
u4	0	5	1	2

POS(a*,b*) = POS(a*,b*,c*)

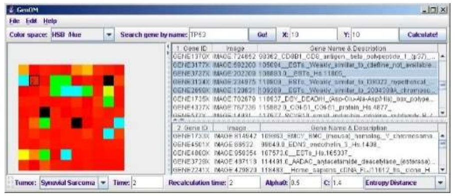
POS(a*) ⊂ POS(a*,b*,c*)

POS(b*) ⊂ POS(a*,b*,c*)

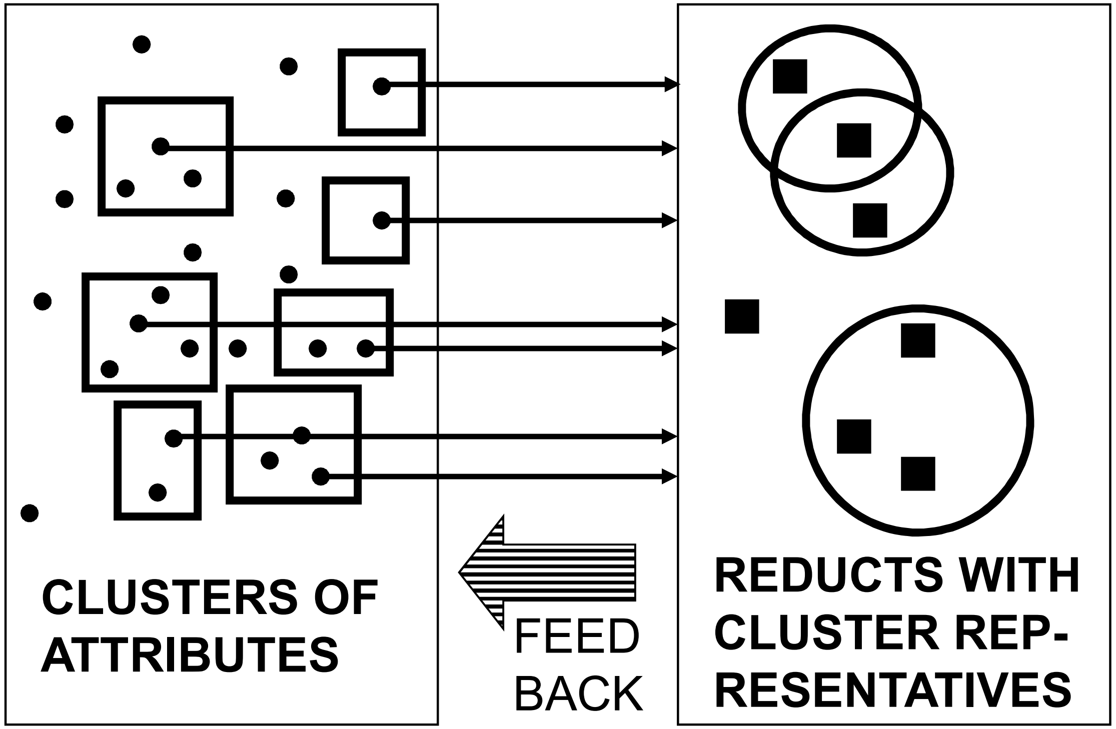
IF a≥3 AND b≥7 THEN d=0
IF a≥3 AND b<7 THEN d=1
IF a≥2 AND b<1 THEN d=1
IF a<2 AND b≥1 THEN d=2
IF a≥4 AND b≥0 THEN d=1
IF a≥0 AND b<5 THEN d=1

	a*	b*	c*	d*
(u1,u1)	1+	1+	1+	0
(u1,u2)	1–	1–	1–	1
(u1,u3)	1+	1–	1+	1
(u1,u4)	1–	1–	1–	2
(u2,u1)	2+	2+	2+	0
(u2,u2)	2+	2+	2+	1
(u2,u3)	2+	2–	2+	1
(u2,u4)	2–	2+	2+	2
(u3,u1)	3–	3+	3–	0
(u3,u2)	3–	3+	3–	1
(u3,u3)	3+	3+	3+	1
(u3,u4)	3–	3+	3–	2
(u4,u1)	4+	4+	4+	0
(u4,u2)	4+	4–	4–	1
(u4,u3)	4+	4–	4+	1
(u4,u4)	4+	4+	4+	2

Adaptive Clustering / Reduction



Grużdź, Ihnatowicz, Ślęzak: Interactive gene clustering – a case study of breast cancer microarray data. Inf. Systems Frontiers 8 (2006).



- Frequent occurrence of representatives in reducts yields splitting clusters
- Rare occurrence of pairs of close representatives yields merging clusters

Attribute Replaceability

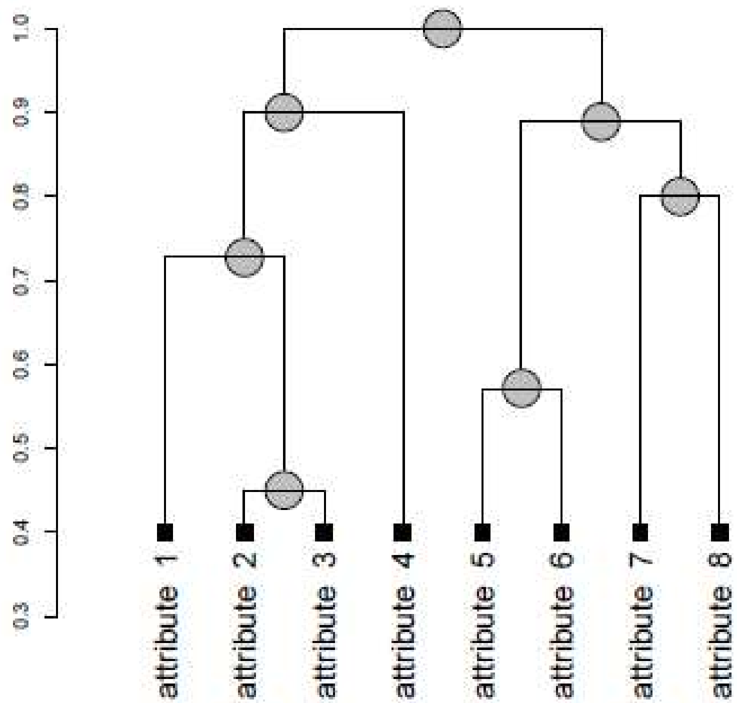
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1														
2														
3	O	O W												
4	O T	O T W												
5	O T H	O T H W												
6			O T H W	T H W	W									
7	O T H W	O T H				O								
8			O T	O	O T H		O T H W							
9	T H	T H W				O W		T H						
10	O T H	O T H W				T W		O H						
11	T H W	T H				O T		H W						
12	O T W	O T				O T H		O W						
13	O H	O H W				O T W		O T H						
14			O T W	W	T H W		O T H		O T H W	H W	O H	O	O T H W	

Attribute Clustering

Exemplary decision system $\mathbb{A} = (U, A \cup \{d\})$:

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	d
u_1	1	2	2	0	0	1	0	1	1
u_2	0	1	1	1	1	0	1	0	1
u_3	1	2	0	1	0	2	1	0	1
u_4	0	1	0	0	1	0	0	1	0
u_5	2	0	1	0	2	1	0	0	1
u_6	1	0	2	0	2	0	0	2	0
u_7	0	1	1	2	0	2	1	0	1
u_8	0	0	0	2	1	1	1	1	0
u_9	2	1	0	0	1	1	0	0	0

Hierarchical attribute clustering of $\mathbb{A}$:



$$direct(a,b) = 1 - \frac{|\{(u,u'): d(u) \neq d(u') \wedge a(u) \neq a(u') \wedge b(u) \neq b(u')\}|}{|\{(u,u'): d(u) \neq d(u') \wedge (a(u) \neq a(u') \vee b(u) \neq b(u'))\}|}$$

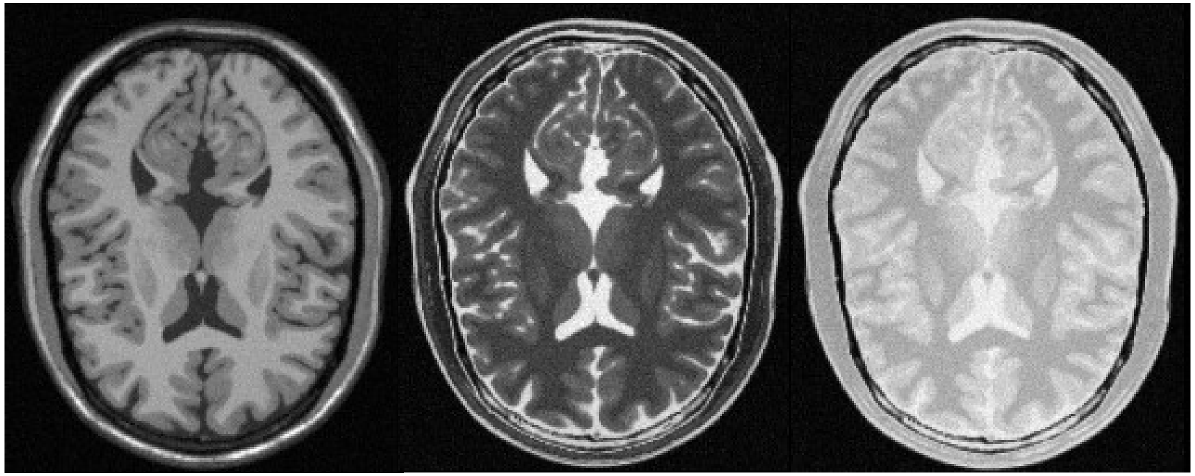
Rough Sets in Data Mining & Databases: Foundations & Applications

Deployment of Rough Set Methods

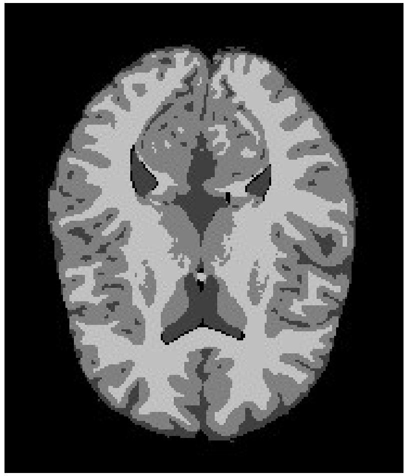
Case Study: MRI Segmentation

The source of conditional attributes | Decision

T1 T2 PD | Phantom



+



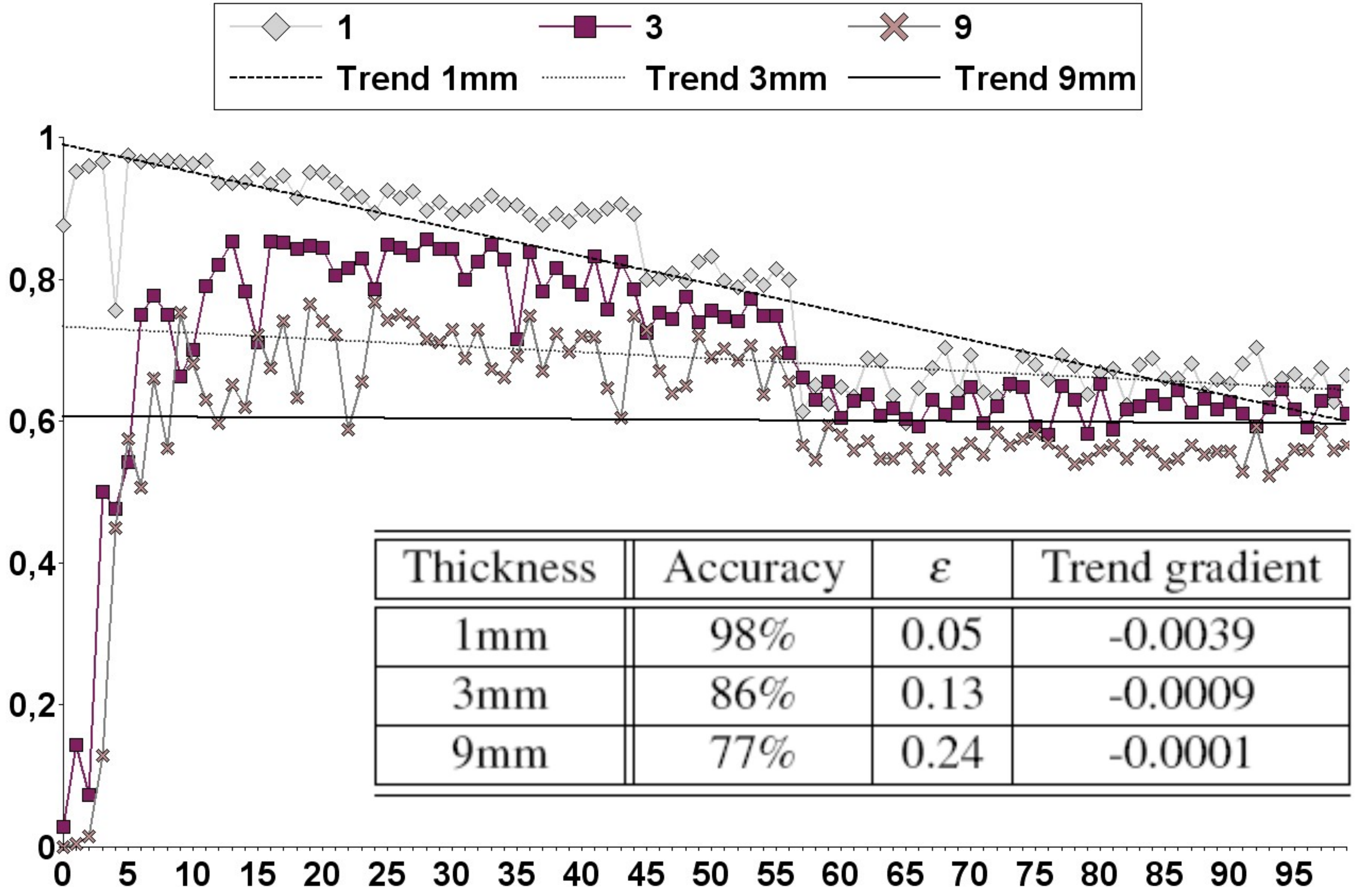
(relaxation time 1) (relaxation time 2) (proton density) **(tissue type)**

Preparing Decision Table (U, A ∪ {d})

- Records in U correspond to the voxels
- Columns in A correspond to the voxels' **<u>attributes extracted from the images</u>**
- Decision d corresponds to the voxels' tissue types taken from the phantom

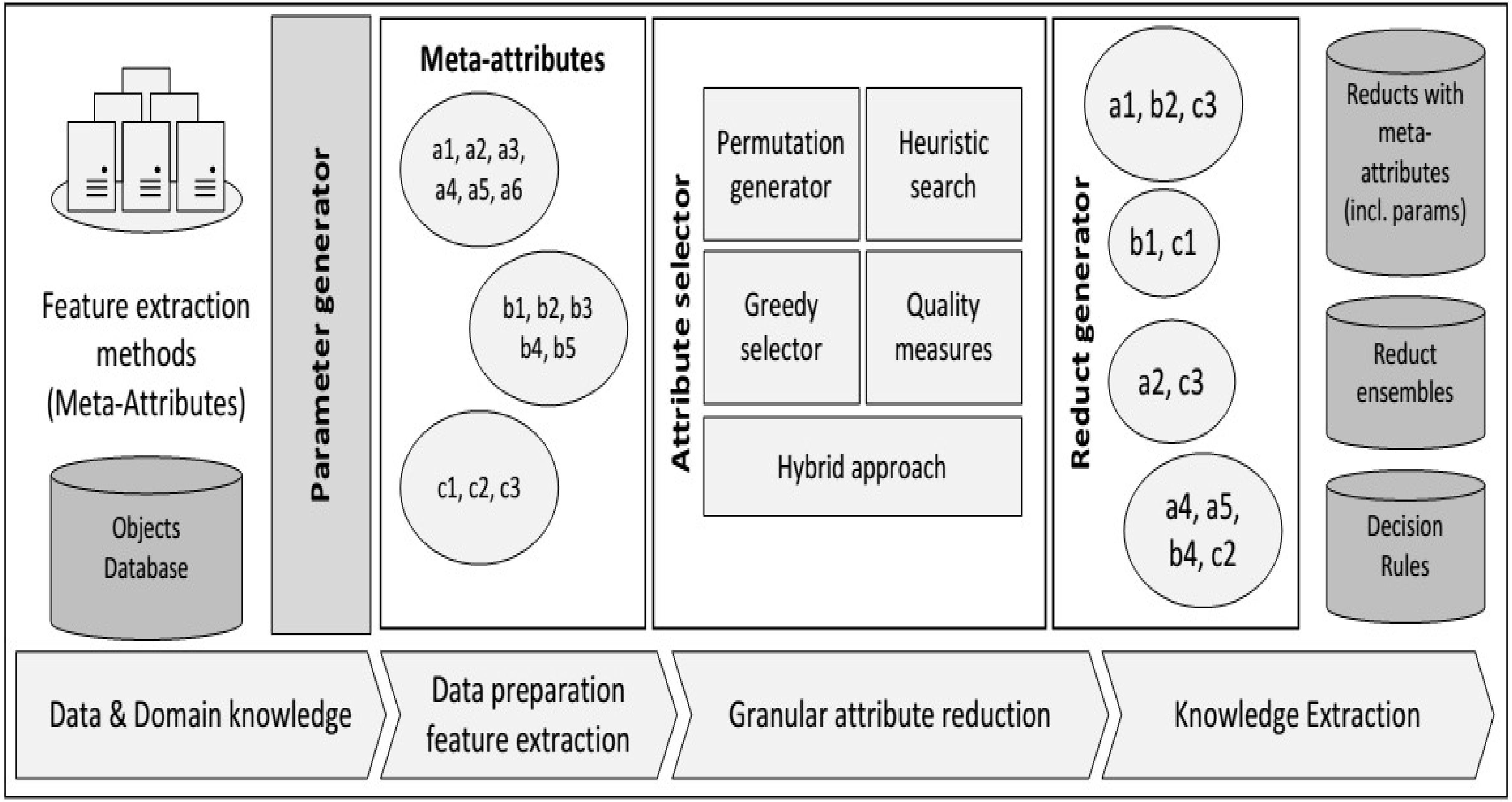
	edge_T1	edge_T2	edge_PD	hcMag_T1_3	hcMag_T2_3	hcMag_PD_3	hcNbr_T1_3	hcNbr_T2_3	hcNbr_PD_3	hcMag_T1_5	hcMag_T2_5	hcMag_PD_5	hcNbr_T1_5	hcNbr_T2_5	hcNbr_PD_5	somMag_T1	somMag_T2	somMag_PD	somNbr_T1	somNbr_T2	somNbr_PD	mask	decision
voxel(80;18)	0	0	1	2	2	1	2	2	1	1	2	1	1	2	1	1	2	1	1	2	1	1	WM
voxel(81;18)	0	0	1	2	2	1	2	2	1	1	2	1	1	2	1	1	3	1	1	3	1	1	WM
voxel(82;18)	0	1	1	2	2	1	2	2	1	1	2	1	1	2	1	1	3	2	1	3	1	1	WM
voxel(83;18)	0	1	1	2	2	1	2	2	1	1	2	1	1	2	1	1	3	1	1	3	1	1	WM
voxel(114;23)	1	0	1	2	2	2	2	2	2	1	2	2	1	2	2	1	3	3	1	3	3	1	WM
voxel(115;23)	1	1	1	2	2	2	2	2	2	1	2	2	1	2	2	1	3	3	1	3	3	1	WM
voxel(116;23)	1	1	1	2	2	2	2	1	1	1	2	2	1	1	1	1	3	2	1	1	1	1	WM
voxel(62;24)	1	1	1	2	2	1	2	2	2	1	2	1	1	2	2	1	3	2	1	2	3	1	WM
voxel(63;24)	1	0	1	2	2	2	2	2	2	1	2	2	1	2	2	2	3	3	1	2	3	1	WM
voxel(64;24)	1	1	1	3	2	2	2	2	2	1	2	2	1	2	2	2	3	3	2	2	3	1	GM
voxel(65;24)	1	1	0	3	2	2	3	1	2	1	2	2	1	1	2	2	2	3	2	1	3	1	GM
voxel(66;24)	1	1	1	3	1	2	3	1	2	2	1	2	1	1	2	2	2	2	2	1	2	1	GM
voxel(67;24)	1	0	1	3	1	2	3	1	2	2	1	2	2	1	2	3	1	2	3	1	2	1	CSF

Accuracy & Approximation Degree



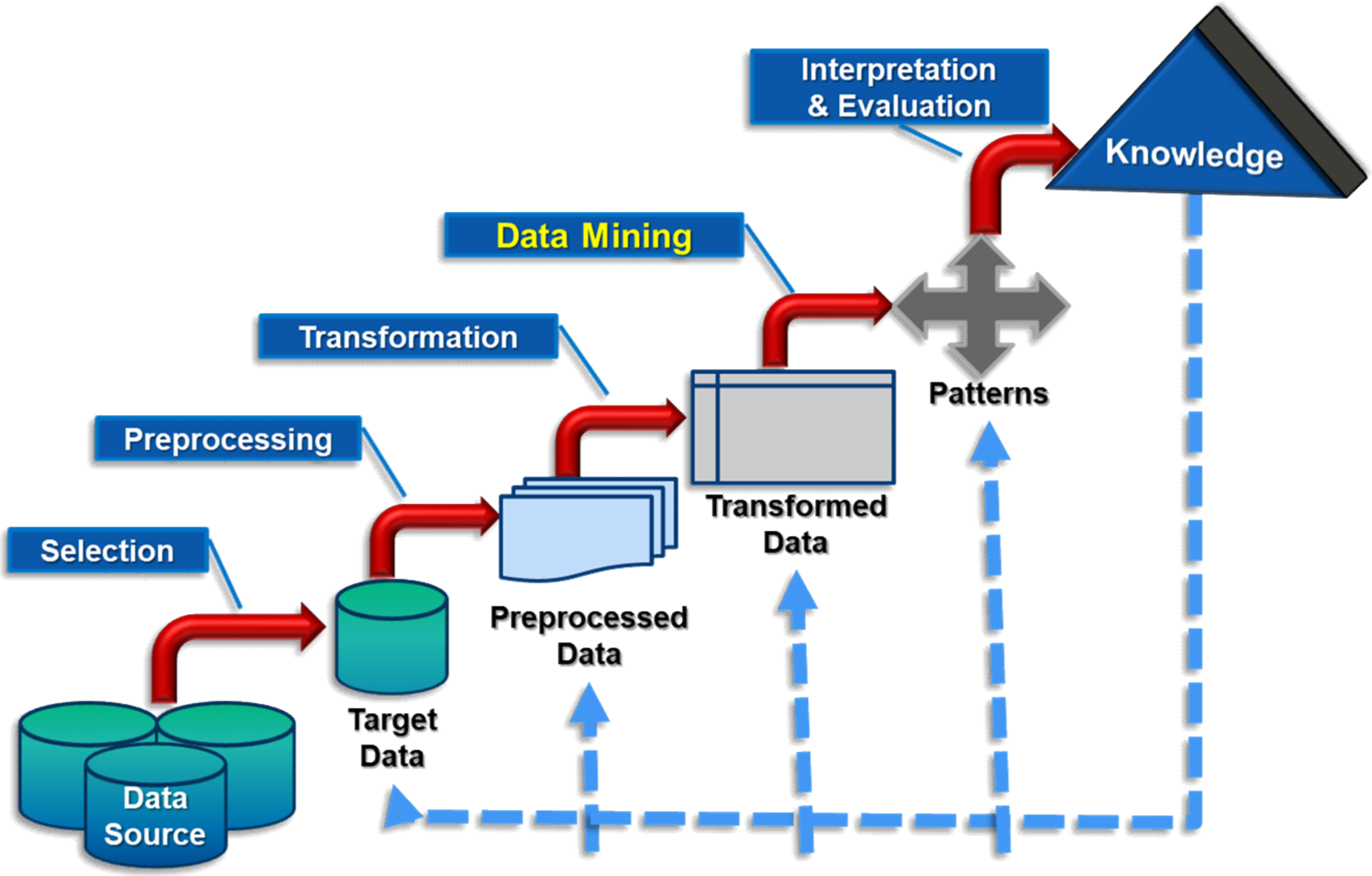
Thickness	Accuracy	ε	Trend gradient
1mm	98%	0.05	-0.0039
3mm	86%	0.13	-0.0009
9mm	77%	0.24	-0.0001

„Granular" Attribute Selection



S. Widz: Ensembles of Approximate Decision Reducts in Classification Problems. PhD Thesis, Polish Academy of Sciences 2017

Rough Sets in KDD

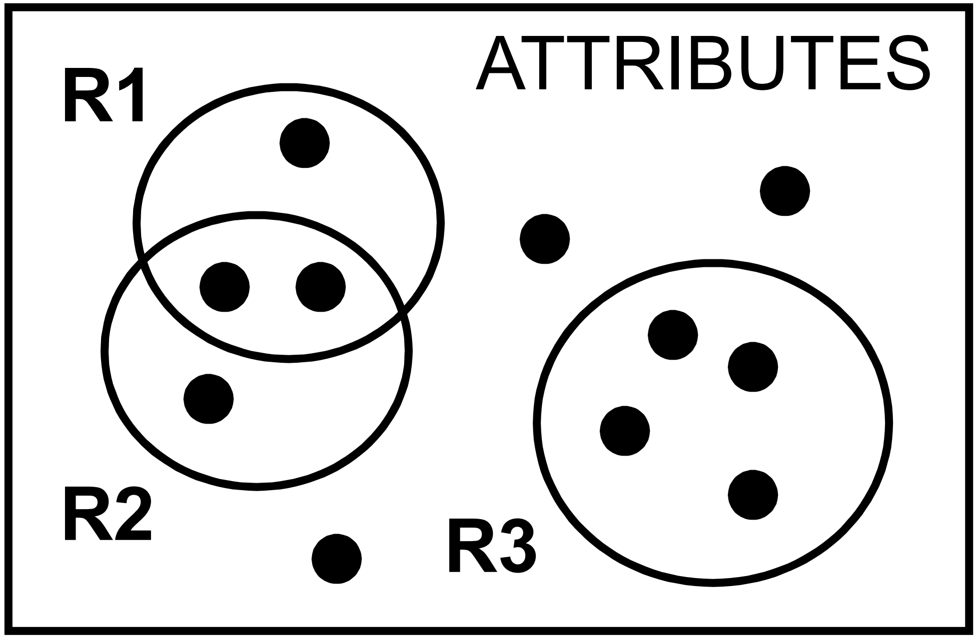


Rough Sets in Data Mining & Databases: Foundations & Applications

Ensembles of Reducts

“Good” Ensembles of Reducts

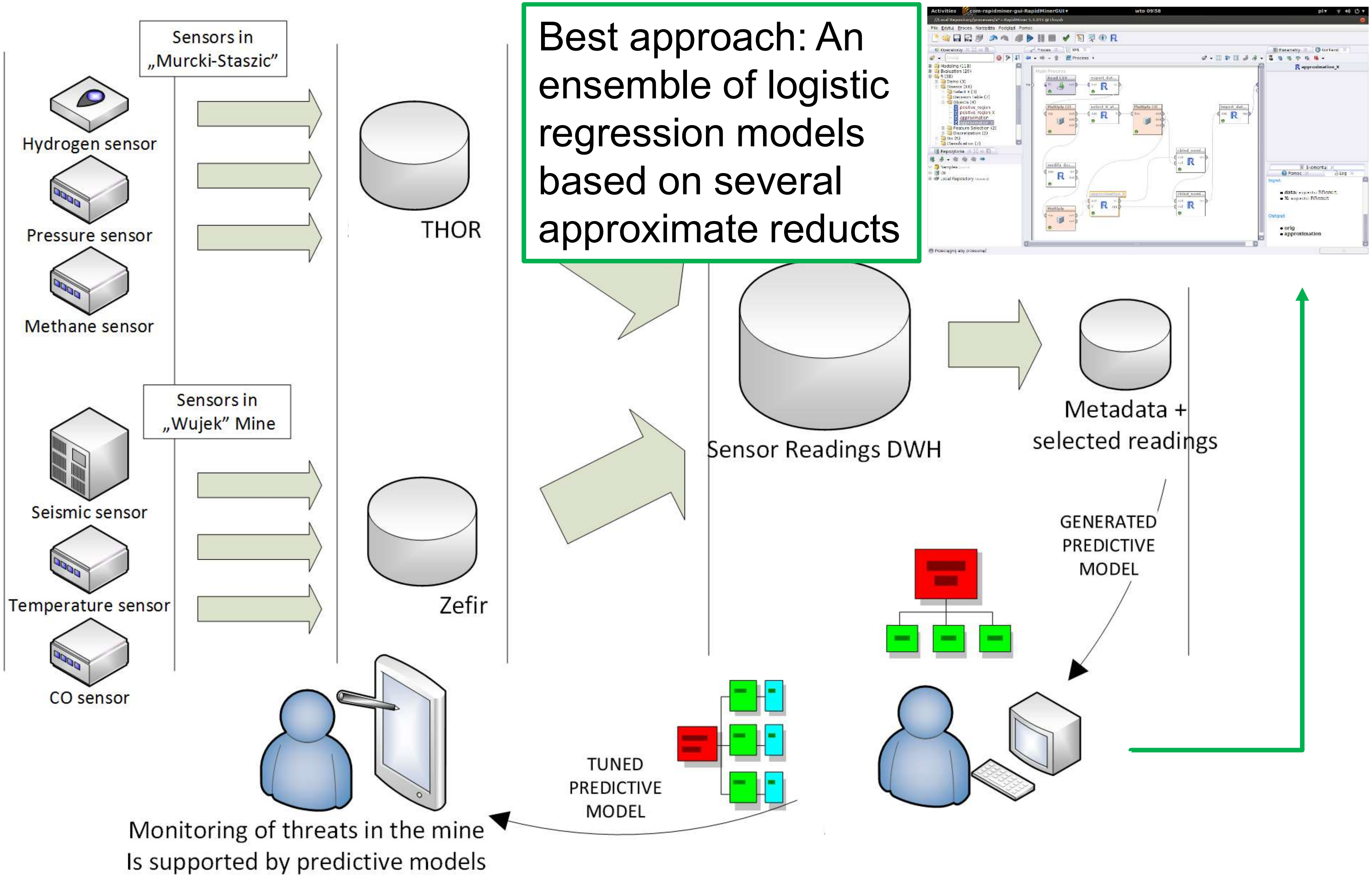
- Reducts with minimal cardinalities (or rules)
- Reducts with minimal pairwise intersections
- Reducts that „cooperate” in decision making



<u>Challenge</u>:

How to modify the existing attribute reduction methods to search for such „good” ensembles

Case Study: Coal Mine Monitoring



Example of Optimization Goal

- *Ensembles of reducts should all together contain relatively many attributes but with small amount of attributes that they share*
- Good for ensembles of classifiers – diversity improves predictive performance
- And for information representation – more complete knowledge about dependencies
- And for domain experts – lower risk of a complete removal of important attributes

Approximate ∂-reducts that „cooperate"

- Irreducible subsets of attributes B and C such that:

$$\forall_{u \in U}\ \partial_B(u) \cap \partial_C(u) = \partial_A(u)$$

- Each subset can lose some ∂-information but the same ∂-information cannot be lost by both of them

a1	a2	a3	a4	a5	d
No	No	No	No	No	green
No	No	Yes	No	Yes	green
No	No	Yes	No	No	red
No	Yes	No	Yes	No	red
No	Yes	No	No	No	blue
Yes	No	Yes	No	Yes	blue

IF a1 = No AND a2 = Yes AND a3 = No THEN d = blue OR d = red

IF a3 = No AND a4 = No AND a5 = No THEN d = blue OR d = green

Definition (Decision bireduct)

Let $\mathbb{A} = (U, A \cup \{d\})$ be a decision system. A pair (B, X), where $B \subseteq A$ and $X \subseteq U$, is called a decision bireduct, if and only if B discerns all pairs $i, j \in X$ where $d(i) \neq d(j)$, and the following properties hold:

1. There is no $C \subsetneq B$ such that C discerns all pairs $i, j \in X$ where $d(i) \neq d(j)$;
2. There is no $Y \supsetneq X$ such that B discerns all pairs $i, j \in Y$ where $d(i) \neq d(j)$.

Some intuition

A decision bireduct (B, X) can be regarded as an inexact functional dependence linking the subset of attributes B with the decision d in a degree X, denoted by $B \Rrightarrow_X d$. The objects in $U \setminus X$ can be treated as the outliers. The objects in X can be used to learn a classifier based on B from data.

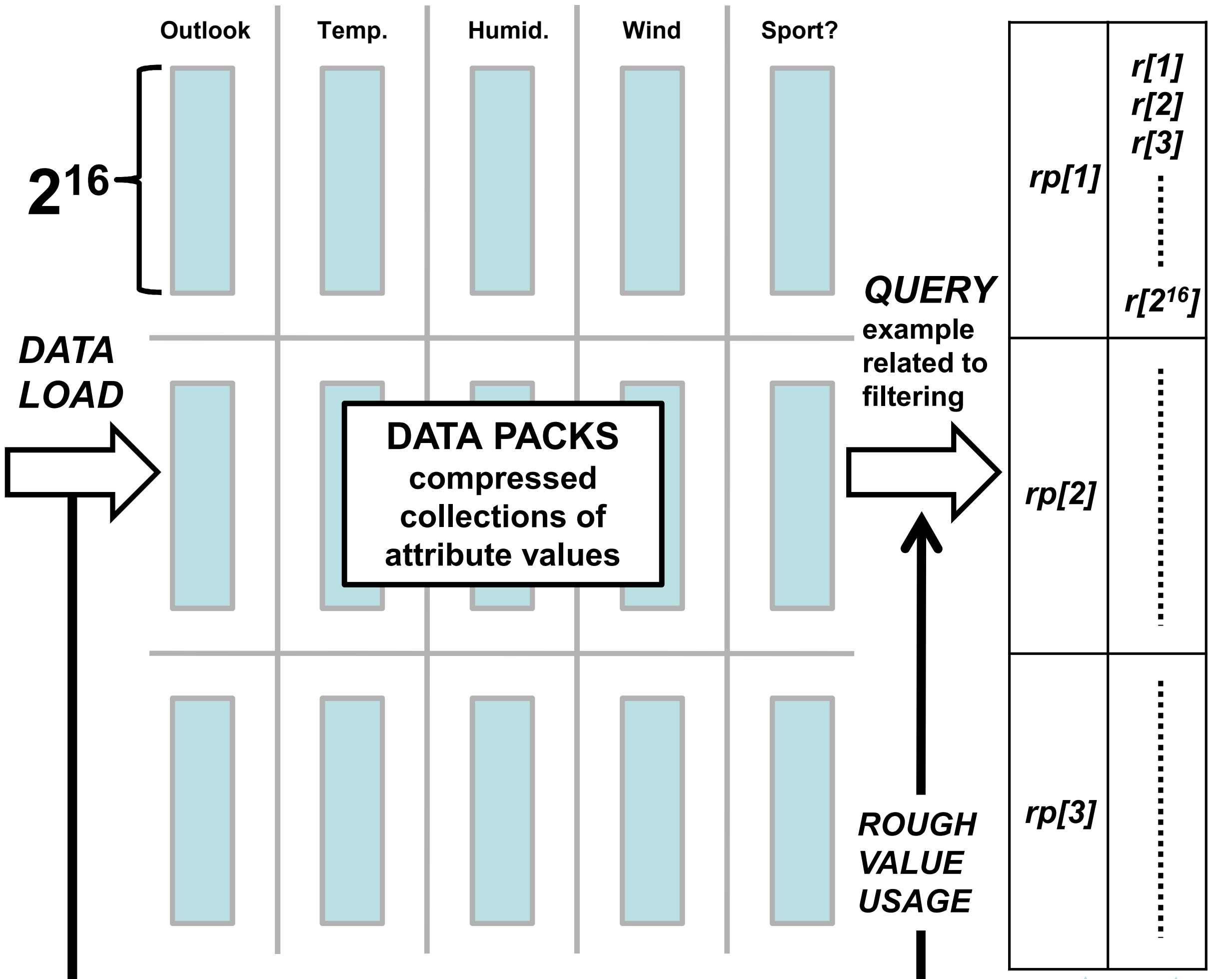
Rough Sets in Data Mining & Databases: Foundations & Applications

Industry Software Case Study 1

Analytical Database Engine – Infobright (2005-2017)

	Outlook	Temp.	Humid.	Wind	Sport?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cold	Normal	Weak	Yes
6	Rain	Cold	Normal	Strong	No
7	Overcast	Cold	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cold	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

ORIGINAL DATA



ROUGH VALUE CALCULATION

ROUGH ATTRIBUTES

	Outlook	Temp.	Humid.	Wind	Sport?
row pack 1	rough value	rough value	rough value	rough value	rough value
row pack 2	rough value	rough value	rough value	rough value	rough value
row pack 3	rough value	rough value	rough value	rough value	rough value

GRANULATED TABLE
a collection of rough values for each of rough attributes is stored as a separate knowledge node

identification of row packs and rows which satisfy query conditions

OUR SOLUTIONS / INFORMATION TECHNOLOGY SOLUTIONS

Scalable Big Data Analytics

Overview

Polystar

Ignite's Infobright DB-Architecture Overview

Ignite's Infobright DB powers applications to perform interactive, complex queries resulting in better, faster business decisions. It is a high performance, scalable solution for storing and analyzing large volumes of machine-generated data at a lower cost and significantly less administrative effort than other database solutions.

High Performance Data Analytics for Better, Faster Business Decisions at a Low Cost

Powered by our innovative Knowledge Grid architecture, Infobright DB is easy to implement and manage — helping you get the answers your business users need at a price you can afford.

- **High Performance:** Sub second response times for complex ad-hoc queries
- **Scalable:** Load terabytes of data per hour and scale to petabytes of data
- **Low Cost High ROI:** No need for complex hardware and storage infrastructure
- **Load and Go:** Infobright DB doesn't require data partitioning, tuning or index creation — just load and go with your existing schemas

SELECT MAX(A) FROM T WHERE B > 15

T (~350K rows) | **B > 15**

Pack A1 Min = 3 Max = 25	Pack B1 Min = 10 Max = 30		S
Pack A2 Min = 1 Max = 15	Pack B2 Min = 10 Max = 20		S
Pack A3 Min = 18 Max = 22	Pack B3 Min = 5 Max = 50		S
Pack A4 Min = 2 Max = 10	Pack B4 Min = 20 Max = 40		R
Pack A5 Min = 7 Max = 26	Pack B5 Min = 5 Max = 10		I
Pack A6 Min = 1 Max = 8	Pack B6 Min = 10 Max = 20		S

- **I**: Irrelevant Granules (*Negative Region*)
- **S**: Suspect Granules (*Boundary Region*)
- **R**: Relevant Granules (*Positive Region*)
- **E**: Exact Computation (necessary, if the final query result cannot be obtained only from the statistical snapshots)

SELECT MAX(A) FROM T WHERE B > 15;

T (~350K rows)		B > 15		MAX(A) ≥ 18		MAX(A) ≥ X	
Pack A1 Min = 3 Max = 25	Pack B1 Min = 10 Max = 30		S	S	S	E	E
Pack A2 Min = 1 Max = 15	Pack B2 Min = 10 Max = 20		S	I	I	I	I
Pack A3 Min = 18 Max = 22	Pack B3 Min = 5 Max = 50		S	S	S	I ⇔ X ≥ 22	I ⇔ X ≥ 22
Pack A4 Min = 2 Max = 10	Pack B4 Min = 20 Max = 40		R	I	I	I	I
Pack A5 Min = 7 Max = 26	Pack B5 Min = 5 Max = 10		I	I	I	I	I
Pack A6 Min = 1 Max = 8	Pack B6 Min = 10 Max = 20		S	I	I	I	I

[18,25] → [18,Y], Y∈[22,25], after accessing A1 & B1

More About Generalized Decisions

- Decision values can take form of numbers, long strings and so on. In such cases, a generalized decision should be rather a kind of description:

$$\partial_B^*(u) = description(\partial_B(u))$$

- Description functions should allow to test whether a given decision value does <u>not</u> occur for a given set of objects (e.g: decision interval, Bloom filter).
- We should also expect monotonicity with respect to an imprecision function (e.g.: interval length):

$$imprecision(\partial_B^*(u)) \geq imprecision(\partial_A^*(u))$$

Rough Sets in Data Mining & Databases: Foundations & Applications

Industry Software Case Study 2

New Query Execution Process

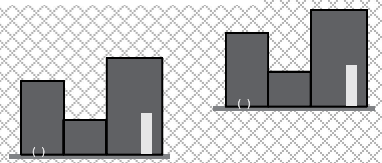
Database

Petabytes

Traditional Query Execution:
- long time to do computations
- lots of disk/memory/processing resources required
- hard to manage in data lake / data cloud environments

Query Result

Summary Database



Gigabytes

Querying on Data Summaries:
- orders of magnitude faster (*original operations replaced by fast summary transformations*)
- far less resources consumed
- original data remaining in-place

Approximate Query Result (*accurate enough from business perspective*)

Practical Use Cases

Use Case	Improvements
Intrusion Detection	faster analytics ⇨ ⇨ improved reaction time ⇨ improved customer retention
Digital Advertising	richer sources of analytics ⇨ ⇨ improved quality of customer profiles ⇨ increased click-thru customer revenue
Sensor-based Monitoring of Industry Processes	faster/deeper machine learning ⇨ ⇨ improved risk prediction efficiency ⇨ lower cost of incorrect predictions

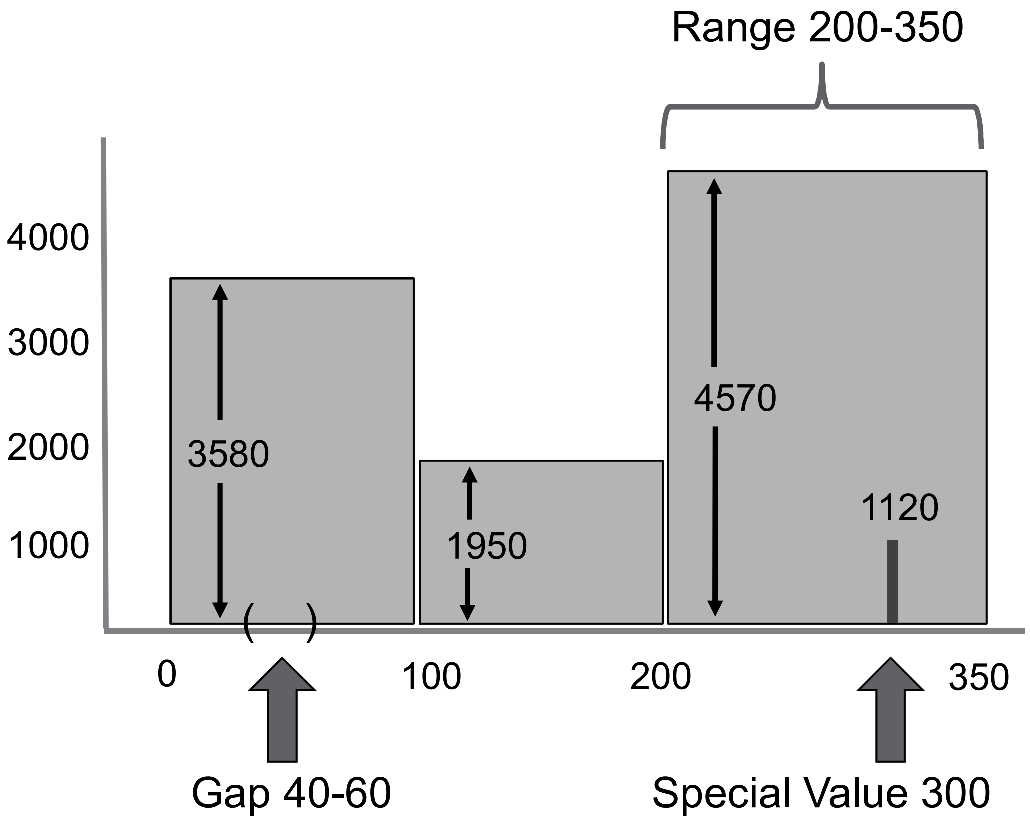
One of the current deployments of the considered new engine assumes working with 30-day periods, wherein there are over 10 billions of new data rows coming every day and ad-hoc analytical queries are required to execute in 2 seconds.

SECURITY ON-DEMAND

Single-Column Summaries

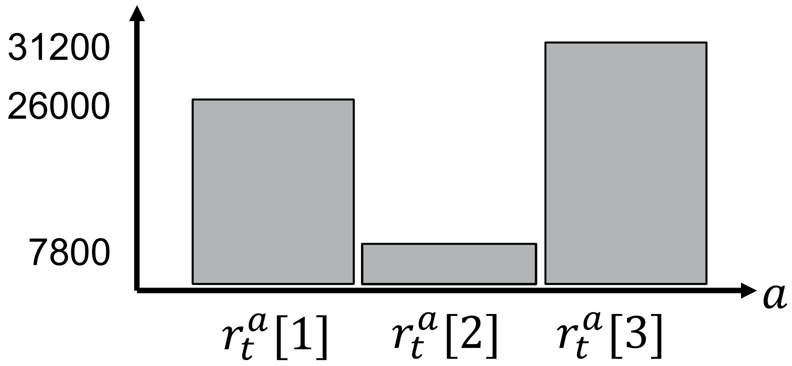
Examples of captured knowledge:

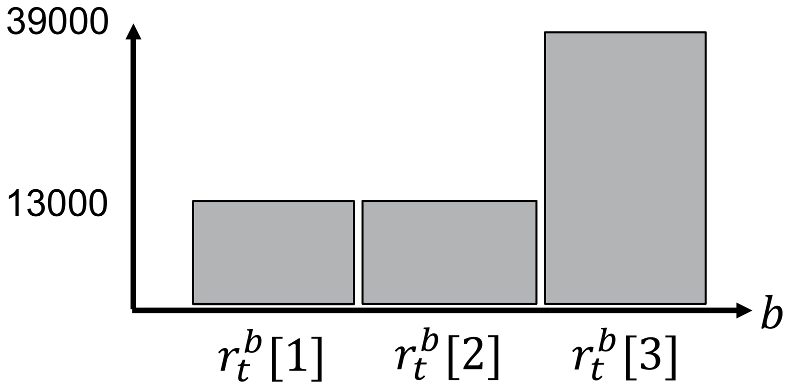
- Value 300 occurred 1120 times
- There were 4570 occurrences of values between 200 and 350 (including value 300)
- There were no occurrences of values between 40 and 60
- Values 0, 40, 60, 100, 200, 350 occurred at least once



On-load selection of borders between histogram bars resembles the tasks of discretization deeply considered in the theory of rough sets.

Two-Column Summaries





$p_t(r_t^a[1]) = \frac{26000}{6500} = \frac{2}{5}$ $\quad p_t(r_t^a[2]) = \frac{7800}{65000} = \frac{3}{25}$ $\quad p_t(r_t^a[3]) = \frac{31200}{65000} = \frac{12}{25}$

$p_t\left(r_t^b[1]\right) = \frac{1300}{65000} = \frac{1}{5}$ $\quad p_t\left(r_t^b[2]\right) = \frac{1300}{65000} = \frac{1}{5}$ $\quad p_t\left(r_t^b[3]\right) = \frac{3900}{65000} = \frac{3}{5}$

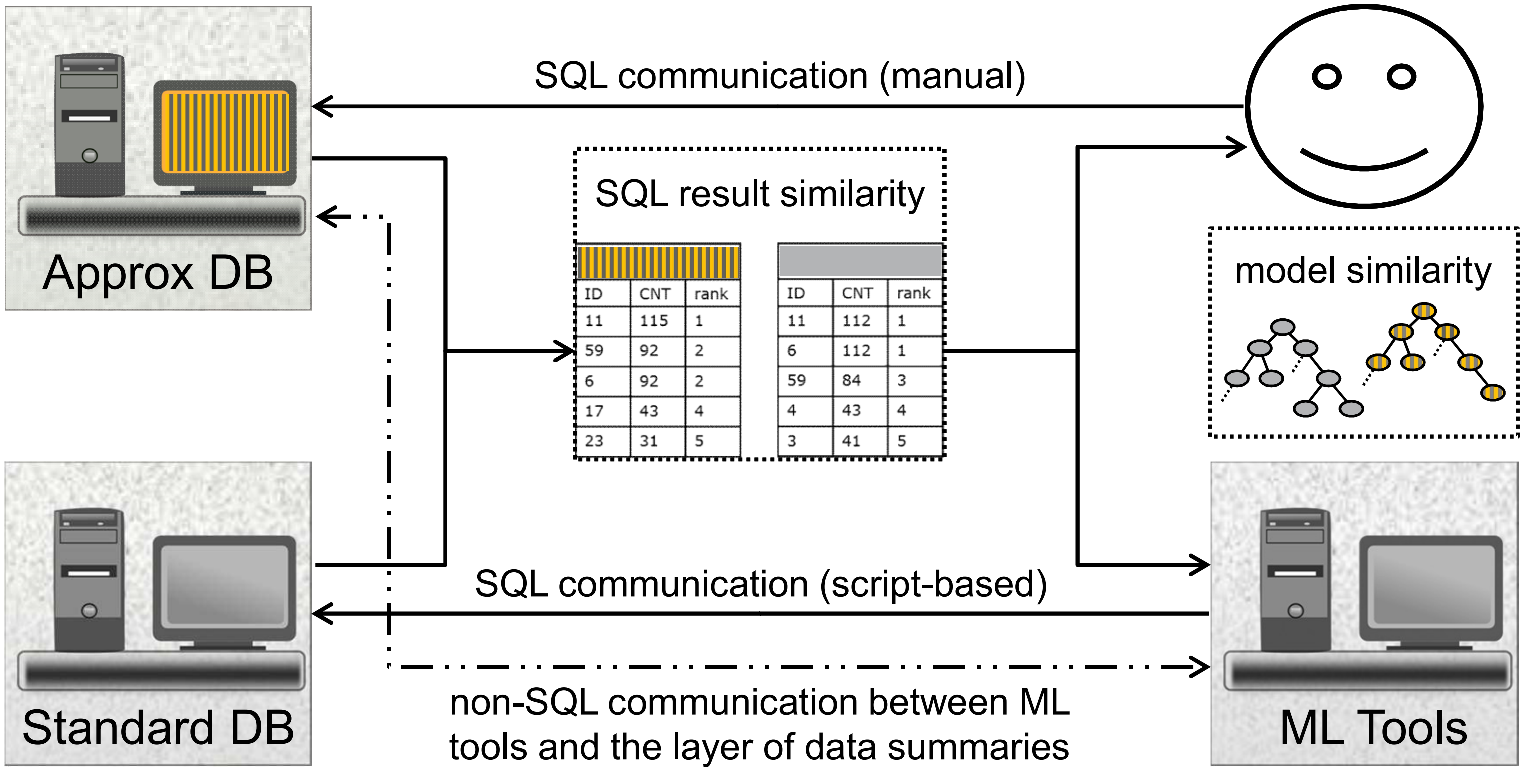
$$p_t\left(r_t^a[1], r_t^b[3]\right) = \frac{20800}{65000} = \frac{8}{25} \implies$$

$$\tau_t\left(r_t^a[1], r_t^b[3]\right) = \frac{8/25}{2/5 \cdot 3/5} = \frac{4}{3}$$

$$\tau_t(a,b) = \frac{1 - p_t\left(r_t^a[1], r_t^b[3]\right)}{1 - p_t\left(r_t^a[1]\right) \cdot p_t\left(r_t^b[3]\right)} = \frac{1 - 8/25}{1 - 2/5 \cdot 3/5} = \frac{17}{19}$$

	$r_t^a[1]$	$r_t^a[2]$	$r_t^a[3]$
$r_t^b[1]$	$\tau_t(a,b)$	$\tau_t(a,b)$	$\tau_t(a,b)$
$r_t^b[2]$	$\tau_t(a,b)$	$\tau_t(a,b)$	$\tau_t(a,b)$
$r_t^b[3]$	4/3	$\tau_t(a,b)$	$\tau_t(a,b)$

How Accurate Calculations Do We Need in Knowledge Discovery?

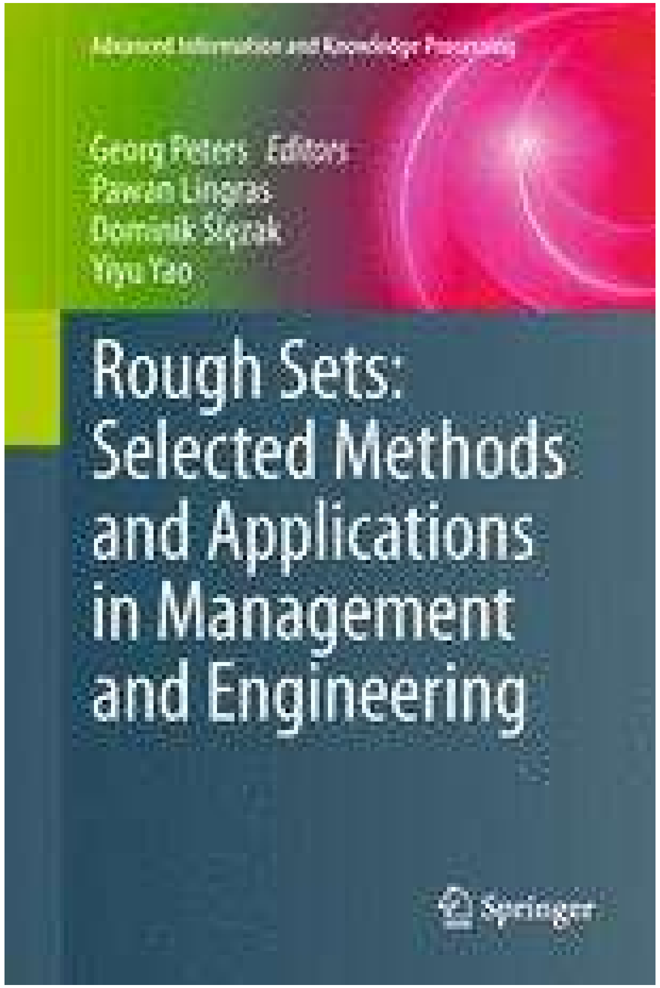


Rough Sets in Data Mining & Databases: Foundations & Applications

Additional Remarks & Materials

Lots of Other Things to Talk About

- Good background for approximate reasoning, knowledge representation, agent communication, etc.
- Powerful methods for hierarchical learning!
- Extending computational models: rough clustering, rough neurons, soft trees…
- Applications: Web and text analysis, finance, multimedia, biomedicine…

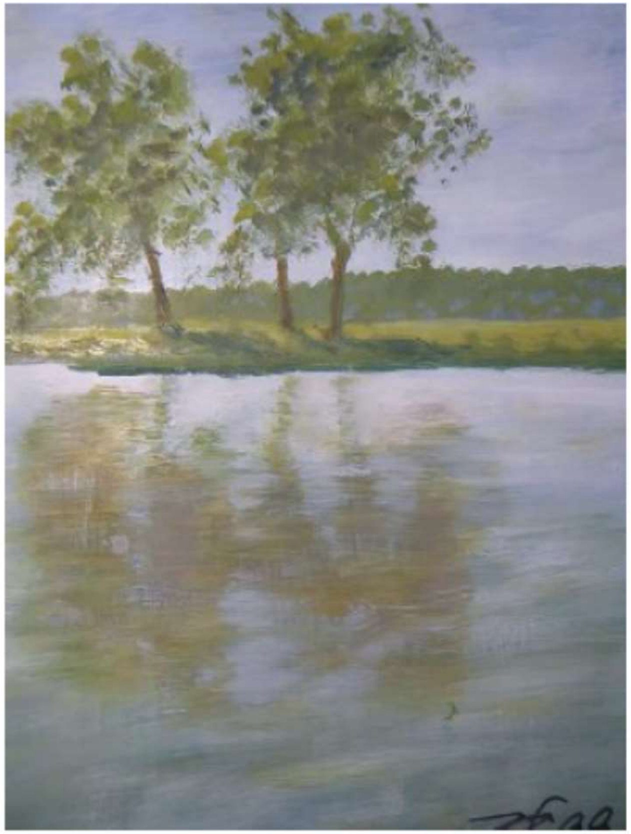


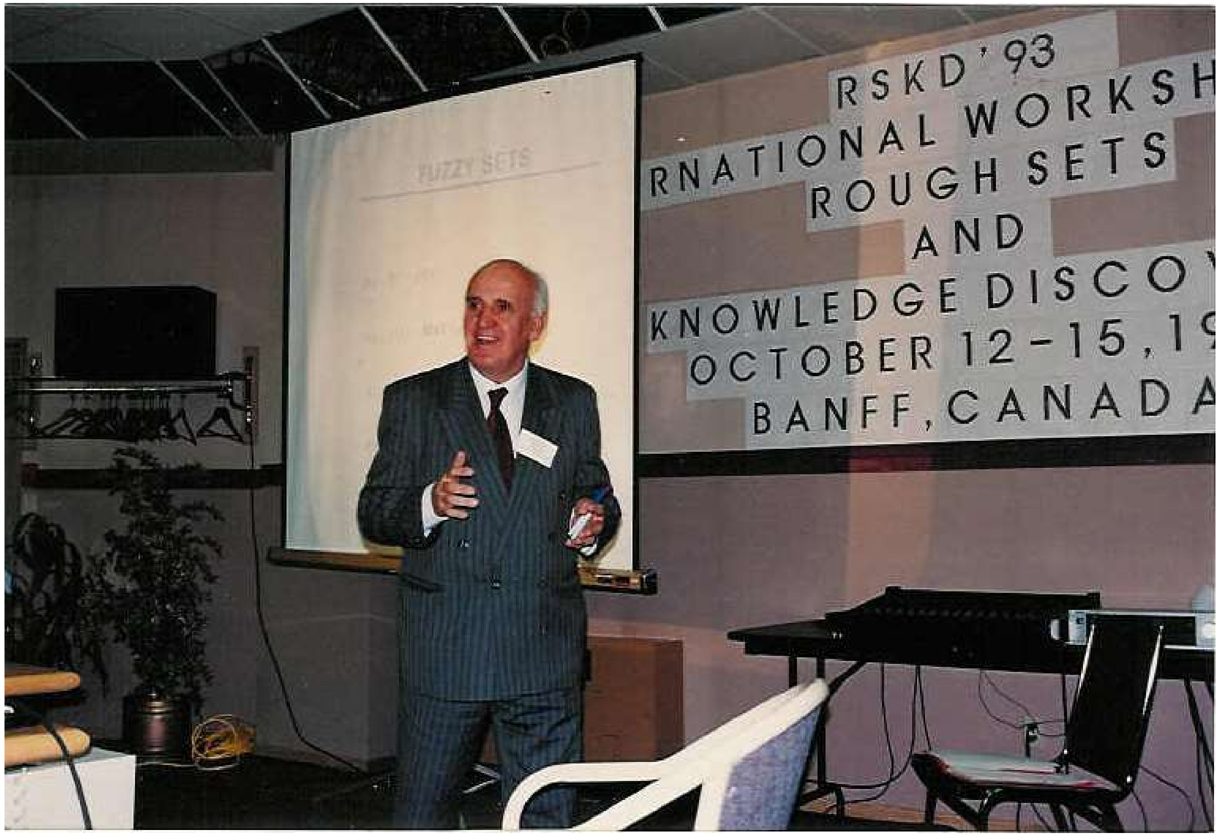
Literature & Useful Links

- Three papers by Z. Pawlak and A. Skowron published in Information Sciences in 2007
- Materials from plenary panel at FedCSIS 2016: https://www.fedcsis.org/2016/plenary_panel
- Materials from Rough Set Summer Schools: http://www.roughsets.org/roughsets/guides/
- Thousands of rough-set-related papers gathered at: http://rsds.univ.rzeszow.pl/

- L.S. Riza et al.: Implementing Algorithms of Rough Set Theory and Fuzzy Rough Set Theory in the R Package „RoughSets”. Inf. Sci. 287: 68-89 (2014)
- S. Stawicki et al.: Decision Bireducts and Decision Reducts - A Comparison. Int. J. Approx. Reasoning 84: 75-109 (2017)
- A. Janusz and D. Ślęzak: Rough Set Methods for Attribute Clustering and Selection. Applied Artificial Intelligence 28(3): 220-242 (2014)
- A. Janusz et al.: Predicting Seismic Events in Coal Mines Based on Underground Sensor Measurements. Eng. Appl. of AI 64: 83-94 (2017)
- D. Ślęzak et al.: Two Database Related Interpretations of Rough Approximations: Data Organization and Query Execution. Fundam. Inform. 127(1-4): 445-459 (2013)
- D. Ślęzak et al.: A New Approximate Query Engine Based on Intelligent Capture and Fast Transformations of Granulated Data Summaries. J. Intell. Inf. Syst. (2017) [Open Access]

Picture of Professor Zdzisław Pawlak





taken from the slides prepared
by Professor Andrzej Skowron

End of Part I

slezak@mimuw.edu.pl

arek.wojna@securityondemand.com