INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Computational Complexity of Rough-Set-Based Feature Selection Algorithms

Dominik Ślęzak

U of Warsaw & Infobright Inc., Poland

slezak@{mimuw.edu.pl;infobright.com}

About the Talk

- Introductory notions
- Practical inspirations
- Towards scalability
- Further extensions

We discuss the rough-set-based approaches to data mining, paying a special attention to the notions of rough approximation, object discernibility and attribute reduction. We concentrate on the tasks of feature selection and feature subset selection. We study the computational complexity of the data processing and the model optimization problems, with respect to the amount of attributes and objects in the data. We outline the most popular heuristics that are used to analyze real-world data sets. Finally, we present some of the recent extensions of the rough-set-based methods aimed at learning robust classifier ensembles.

Introduction – Rough Sets

- Rough set theory proposed by Z. Pawlak in 82 is an approximate reasoning model
- In applications, it focuses on approximate knowledge derivable from data
- It provides good results in such domains as, e.g., Web analysis, finance, industry, multimedia, medicine, and bioinformatics

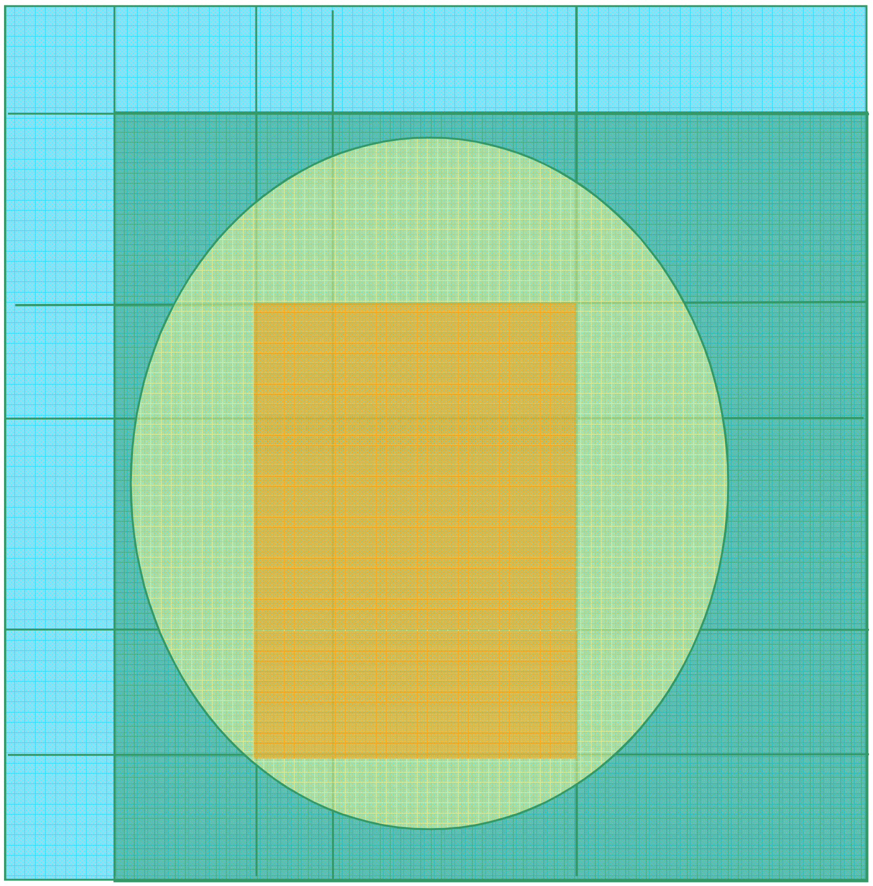
Introduction – Reduction

- Reducts: optimal attribute subsets, which approximate well enough the pre-defined target concepts or the whole data source
- Notion of reduct extended based on e.g.: Boolean reasoning, Bayesian reasoning, information theory, etc.
- Real-world data-based reduction algorithms based on e.g.: greedy heuristics and genetic algorithms

Attribute Reduction Criteria

- Find optimal subset of attributes providing (approximate) rules covering (almost) all the objects occurring in the available data
- Find optimal subset of attributes providing the rules approximating decisions at least (almost) as good as the full attribute set

Rough Approximations



Lower Approxi-mation: Objects certainly in X (the exact rules for X)

Upper Approxi-mation: Objects that may be in X (the rules which do not exclude X)

Reducts Preserving Positive Region

- Consider a system with r decision classes

  $X_0,\ldots,X_{r-1}$ (r is called a system's rank)
- By a B-positive region we mean the union of lower approximations of all the classes:

  $POS(B) = U_{k=0,..r-1}\ LOW_B(X_k)$
- We say that subset B of A is a reduct, if

  $POS(B) = POS(A)$

  and for any proper subset C of B there is

  $POS(C) \neq POS(A)$

Illustration

	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

- POS(O,T,H,W) is equal to U
- POS(O,H,W) is still equal to U
- POS(C), for any proper subset C of {O,H,W}, will decrease a lot

Rules Generated by {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

Rules Generated by {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 !!!

Reducts Preserving Discernibility

	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.

<table>
<tr><th></th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th><th>8</th><th>9</th><th>10</th><th>11</th><th>12</th><th>13</th><th>14</th></tr>
<tr><td>1</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>2</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>3</td><td>O</td><td>O W</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>4</td><td>O T</td><td>O T
W</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>5</td><td>O T
H</td><td>O T
H W</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>6</td><td>|||||</td><td>|||||</td><td>O T
H W</td><td>T H
W</td><td>W</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>7</td><td>O T
H W</td><td>O T
H</td><td>|||||</td><td>|||||</td><td>|||||</td><td>O</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>8</td><td>|||||</td><td>|||||</td><td>O T</td><td>O</td><td>O T
H</td><td>|||||</td><td>O T
H W</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>9</td><td>T H</td><td>T H
W</td><td>|||||</td><td>|||||</td><td>|||||</td><td>O W</td><td>|||||</td><td>T H</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>10</td><td>O T
H</td><td>O T
H W</td><td>|||||</td><td>|||||</td><td>|||||</td><td>T W</td><td>|||||</td><td>O H</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>11</td><td>T H
W</td><td>T H</td><td>|||||</td><td>|||||</td><td>|||||</td><td>O T</td><td>|||||</td><td>H W</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>12</td><td>O T
W</td><td>O T</td><td>|||||</td><td>|||||</td><td>|||||</td><td>O T
H</td><td>|||||</td><td>O W</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>13</td><td>O H</td><td>O H
W</td><td>|||||</td><td>|||||</td><td>|||||</td><td>O T
W</td><td>|||||</td><td>O T
H</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>14</td><td>|||||</td><td>|||||</td><td>O T
W</td><td>W</td><td>T H
W</td><td>|||||</td><td>O T
H</td><td>|||||</td><td>O T
H W</td><td>H W</td><td>O H</td><td>O</td><td>O T
H W</td><td>|||||</td></tr>
</table>

<table>
<tr><th></th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th><th>8</th><th>9</th><th>10</th><th>11</th><th>12</th><th>13</th><th>14</th></tr>
<tr><td>1</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>2</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>3</td><td>O</td><td>O W</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>4</td><td>O T</td><td>O T
W</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>5</td><td>O T
H</td><td>O T
H W</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>6</td><td>|||||</td><td>|||||</td><td>O T
H W</td><td>T H
W</td><td>W</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>7</td><td>O T
H W</td><td>O T
H</td><td>|||||</td><td>|||||</td><td>|||||</td><td>O</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>8</td><td>|||||</td><td>|||||</td><td>O T</td><td>O</td><td>O T
H</td><td>|||||</td><td>O T
H W</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>9</td><td>T H</td><td>T H
W</td><td>|||||</td><td>|||||</td><td>|||||</td><td>O W</td><td>|||||</td><td>T H</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>10</td><td>O T
H</td><td>O T
H W</td><td>|||||</td><td>|||||</td><td>|||||</td><td>T W</td><td>|||||</td><td>O H</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>11</td><td>T H
W</td><td>T H</td><td>|||||</td><td>|||||</td><td>|||||</td><td>O T</td><td>|||||</td><td>H W</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>12</td><td>O T
W</td><td>O T</td><td>|||||</td><td>|||||</td><td>|||||</td><td>O T
H</td><td>|||||</td><td>O W</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>13</td><td>O H</td><td>O H
W</td><td>|||||</td><td>|||||</td><td>|||||</td><td>O T
W</td><td>|||||</td><td>O T
H</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td><td>|||||</td></tr>
<tr><td>14</td><td>|||||</td><td>|||||</td><td>O T
W</td><td>W</td><td>T H
W</td><td>|||||</td><td>O T
H</td><td>|||||</td><td>O T
H W</td><td>H W</td><td>O H</td><td>O</td><td>O T
H W</td><td>|||||</td></tr>
</table>

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

Matrices are not the Only Ones

- Given a discernibility matrix, we can:
  - search for all reducts (exponential) or
  - apply heuristics to find (sub-)optimal reducts (polynomial w.s.t. attributes – acceptable; but square w.s.t. no. of objects – not acceptable)
- But we can also base e.g. on data sorting:
  - A heuristic procedure chooses the subsets of attributes to be verified
  - A heuristic measure is calculated over the data sorted according to each given set of attributes

Hybrid Genetic Algorithms

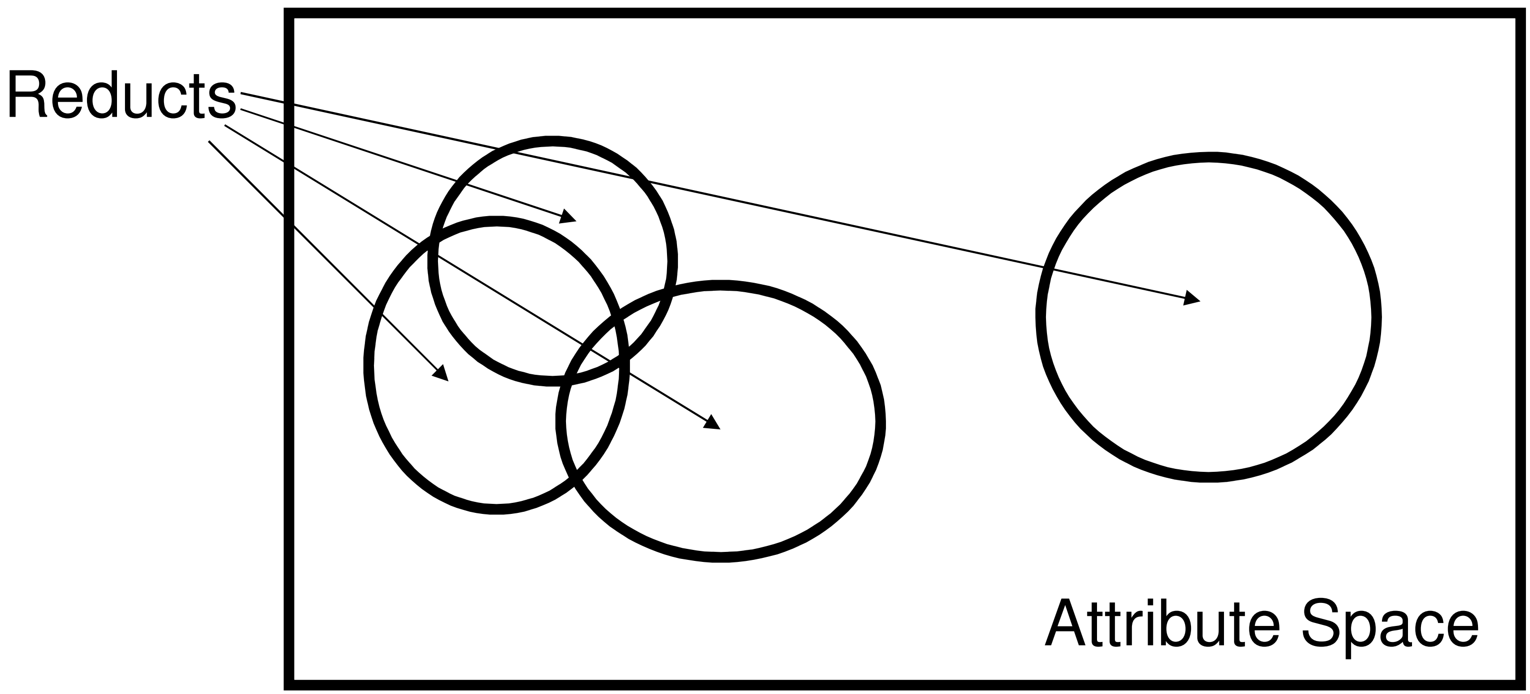
- *Genetic part*, where each chromosome encodes a permutation of attributes
- *Heuristic part*, where permutations are put into the following algorithm

REDORD algorithm:

1. For $\tau:\{1,..,|A|\}\rightarrow\{1,..,|A|\}$, let $B_\tau=A$;
2. For $i = 1$ to $|A|$ repeat steps 3 and 4;
3. Let $B_\tau \leftarrow B_\tau \setminus \{a_{\tau(i)}\}$;
4. If $POS(B_\tau) \neq POS(A)$ undo step 3

Reducts mapped by most permutations

- Those with least cardinality
- Those with least intersections with others
- A good basis for the classifier construction



Towards Approximate Reducts

- It is worth reducing irrelevant attributes and simplifying obtained decision rules
- Reduction (simplification) should not decrease the overall accuracy of rules, understood in terms of the rough set approximations of decision classes
- In real-life applications, we may agree to *slightly* decrease the quality, if it leads to significantly simpler classification models

Approximate Reducts

- We can specify a function

  $$M(d/\ ): P(A) \rightarrow \Re$$

  evaluating influence of attribute sets on d
- $B \subseteq A$ is an (M,ε)-approximate reduct, iff

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

  and none of its proper subsets holds it
- It is important for M to be somehow “good”

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

By a non-directed graph we understand a tuple $\mathbf{G} = (X, E)$, where X is the set of vertices and where the relation $E \subseteq X \times X$ is symmetric. Each element of E is represented as $e = \{l(e), r(e)\}$, where $l(e), r(e) \in X$ are called the vertices of e.

Definition 8.1. Let a non-directed $\mathbf{G} = (X, E)$ be given. We say that subset $Y \subseteq X$ *covers* $\mathbf{G}$ iff

$$\forall_{x \in X} \left(x \notin Y \Rightarrow \exists_{y \in Y} \left(\{x, y\} \in E \right) \right) \tag{83}$$

Definition 8.2. The *Minimal Graph Covering Problem* is the problem of finding minimal subset of vertices, which covers a given graph $\mathbf{G} = (X, E)$.

Theorem 8.1. *([2]) The Minimal Graph Covering Problem is NP-hard.*

The proof of Theorem 6.1 requires a generalization of the notion of a covering.

Definition 8.3. Let $\alpha \in (0, 1]$ and $\mathbf{G} = (X, E)$ be given. We say that subset $Y \subseteq X$ *α-covers* $\mathbf{G}$ iff

$$|Cov_{\mathbf{G}}(Y)| \,/\, |X| \geq \alpha \tag{84}$$

where

$$Cov_{\mathbf{G}}(Y) = Y \cup \{x \in X : \exists_{y \in Y} (\{x, y\} \in E)\} \tag{85}$$

is the set of vertices covered by Y in $\mathbf{G}$.

Definition 8.4. For any $\alpha \in (0, 1]$, the *Minimal Graph α-Covering Problem* is the problem of finding minimal subset of vertices, which is an α-covering for a given graph $\mathbf{G} = (X, E)$.

Theorem 8.2. *For any $\alpha \in (0, 1]$, the Minimal Graph α-Covering Problem is NP-hard.*

Examples of Quality Functions

- Disc(d/B) = Disc(B∪{d}) – Disc(B)

  where Disc(X)=

  $= |\{(u_1,u_2): a(u_1) \neq a(u_2) \text{ for some } a \in X\}|$
- Relative Gain R(d/B) =

$$\sum_{\textit{rules}\ \mathrm{r}\ \textit{induced by}\ \mathrm{B}} \left(\frac{\textit{number of objects recognizable by}\ \mathrm{r}}{\textit{number of objects in}\ \mathrm{U}} * \max_i \frac{\textit{probability of the}\ \mathrm{i}\textit{-th decision class induced by}\ \mathrm{r}}{\textit{prior probability of the}\ \mathrm{i}\textit{-th decision class}} \right)$$

- Conditional information entropy H(d/B)

o-GA for Approximate Reducts

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

(M,ε)-REDORD algorithm:

1. For $\tau:\{1,..,|A|\}\rightarrow\{1,..,|A|\}$, let $B_\tau=A$;
2. For $i = 1$ to $|A|$ repeat steps 3 and 4;
3. Let $B_\tau \leftarrow B_\tau \setminus \{a_{\tau(i)}\}$;
4. If $M(d/B_\tau) < (1-\varepsilon)M(d/A)$ undo step 3

evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Practical Inspirations

A sample of the gene expression data related to a selected type of soft tissue tumor

http://genome-www.stanford.edu/sarcoma/

A	a_1	a_2	a_3	a_4	a_5	d
x_1	-0.16	0.47	1.28	2.39	0.53	0.11
x_2	-0.97	-0.18	-0.32	-0.98	0.18	0.88
x_3	-0.23	0.44	1.32	2.91	-0.20	0.20
x_4	-1.45	0.66	-1.59	-1.01	-0.45	-0.99
x_5	-0.05	-0.63	-1.35	-1.01	-2.02	-0.59
x_6	1.06	1.31	0.82	1.32	2.21	-0.78
x_7	-1.99	2.36	-0.28	-1.32	0.13	-1.20

Decision Rules Revisited

- Suppose that we want to build a rule basing on object x and attribute set B
- We must verify whether, for each object y, the degree of closeness of d(y) to the rule's consequence d(x) is appropriately bounded by the degrees of closeness of a(y) to the rule's premises a(x), for a in B
- One can understand it as rule's stability

Distance-Based Discernibility

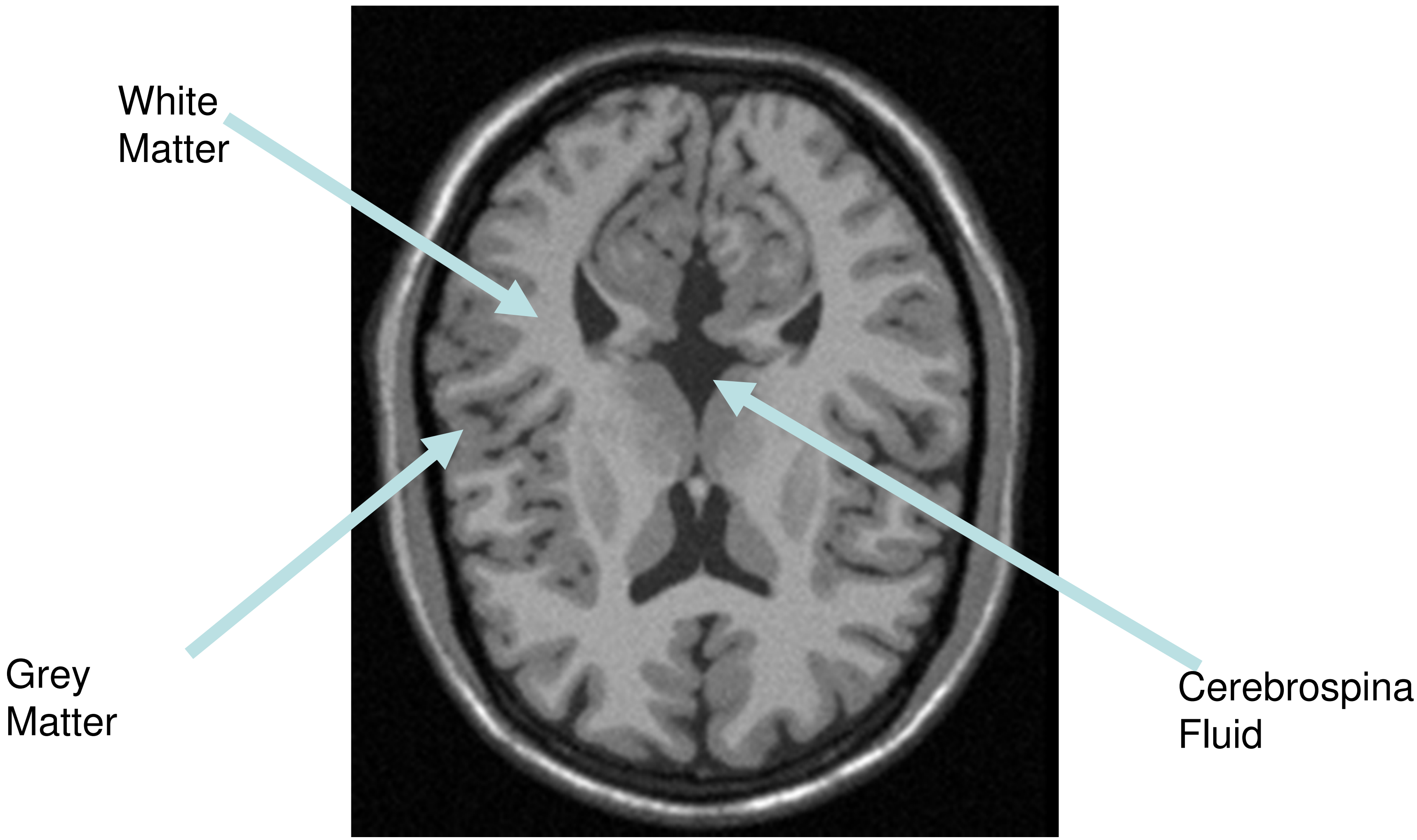
- Consider **c** = $(c_1,\ldots,c_m)$ as a vector of cuts over ranges of attributes in B = $(a_1,\ldots,a_m)$
- Then integral of the form

$$\mathrm{D(B)} = \iiint_{ranges} \mathrm{D(B_c)dc}$$

can be expressed using quantities

$$\mathrm{Dist(B)} = \sum_{x,y \in U} \prod_{a \in B} |a(x) - a(y)|$$

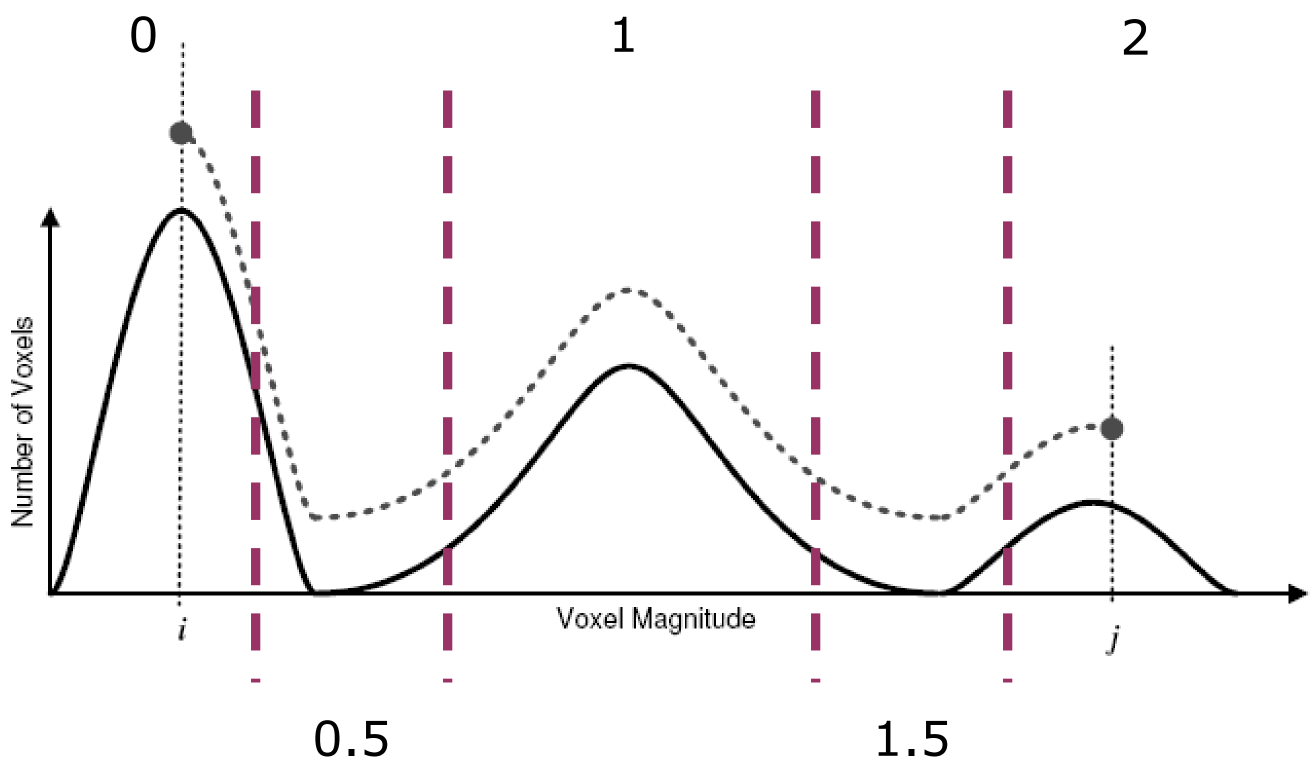
Approximate Reduction Case Study: MRI Segmentation



Decision Table $\mathrm{A}=(U,A \cup \{d\})$

- Records in U correspond to the voxels
- Columns in A correspond to the voxels' features extracted from images (we are describing possible features further)
- Decision d corresponds to the voxels' tissue types taken from the phantom image created by the experts

Histogram Attributes



Another Case Study: Rough Set Approach to 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	$only$	$T3$	$cN1$	$larynx$	25	15	4	6
4	1	$after$	$T3$	$cN1$	$larynx$	38	8	18	12
24	1	$radio$	$T3$	$cN1$	$larynx$	23	6	7	10
28	1	$after$	$T3$	$cN0$	$throat$	18	4	8	6
57	1	$after$	$T4$	$cN1$	$larynx$	32	12	14	6
91	1	$after$	$T3$	$cN1$	$throat$	35	5	16	14
152	1	$only$	$T3$	$cN0$	$larynx$	27	9	14	4
255	1	$after$	$T3$	$cN0$	$larynx$	15	2	6	7
493	1	$after$	$T3$	$cN1$	$other$	19	6	7	6
552	2	$after$	$T4$	$cN2$	$larynx$	14	6	3	5

Rough Memberships

- For each u∈U we can calculate rough membership distribution of the form

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

- During the reduction process, we want to discern between only these object pairs, which induce rough memership distributions far enough to each other

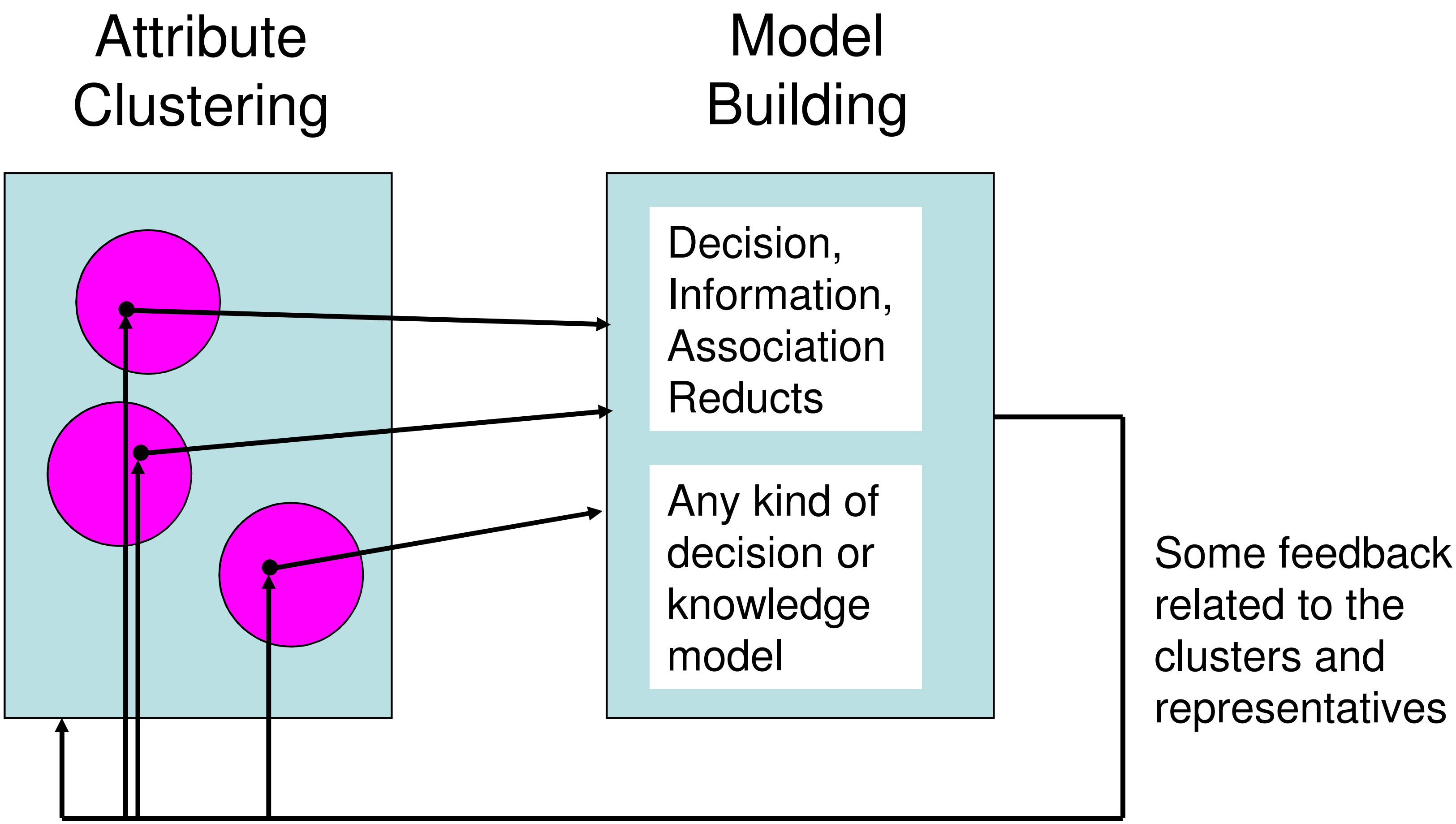
Compound Decision Values

- Distributions of the types of recurrences
- Kaplan-Meier plots (various distances)
- Prognostic indexes of the Cox model
- Pairs of all the above kinds of values calculated for two kinds of operations

- For MRI – Tissue distributions resulting from the fuzzy phantoms (especially in case of Partial Volume Effect analysis)

Towards Scalability

Computing with Attribute Sets



Attribute Replaceability

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

Attribute Replaceability

- Discernibility approach corresponds to

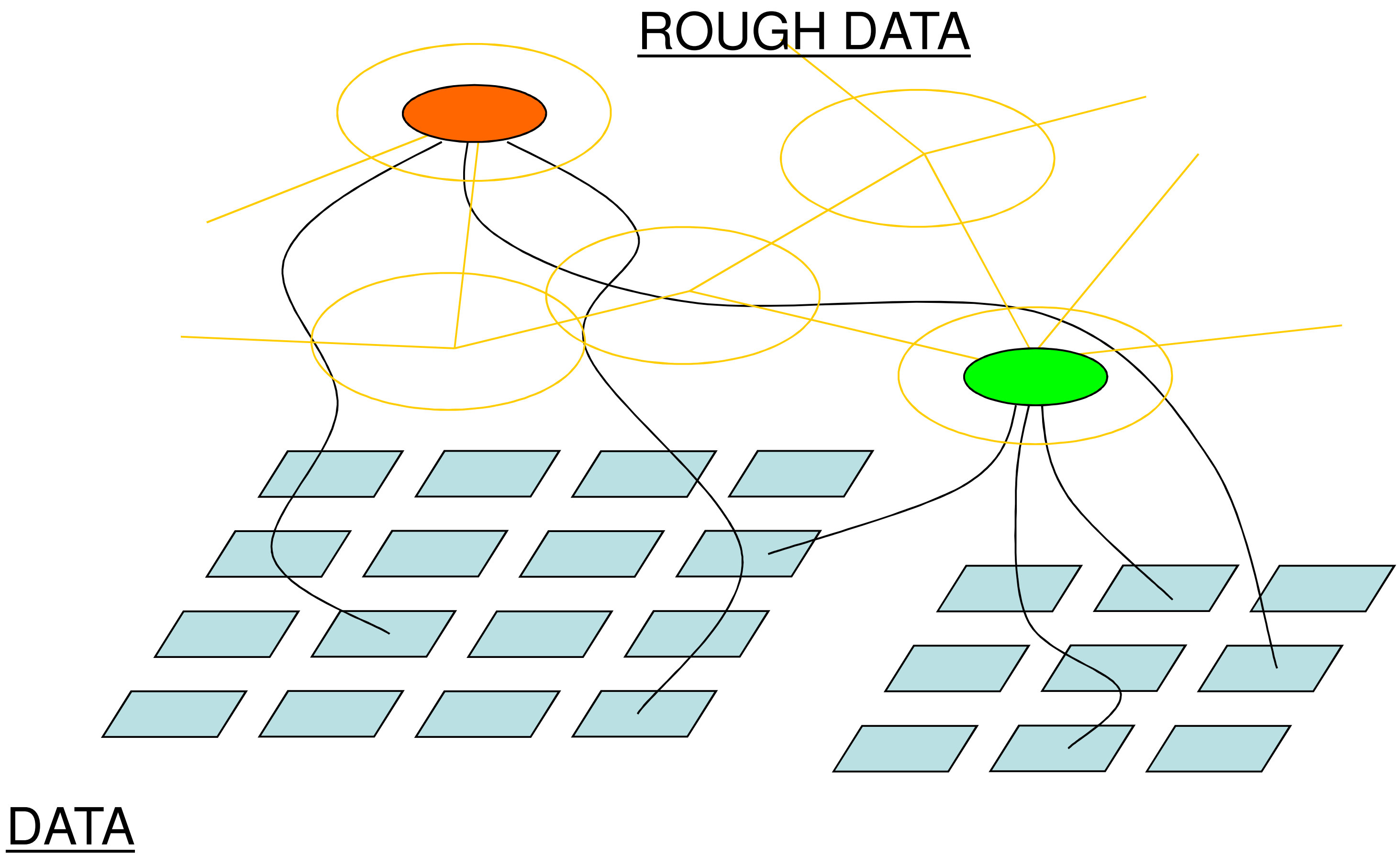
  $$\mathrm{Disc}(B) = |\{(u_1,u_2): \exists_{a\in B}\ a(u_1)\neq a(u_2)\}|$$

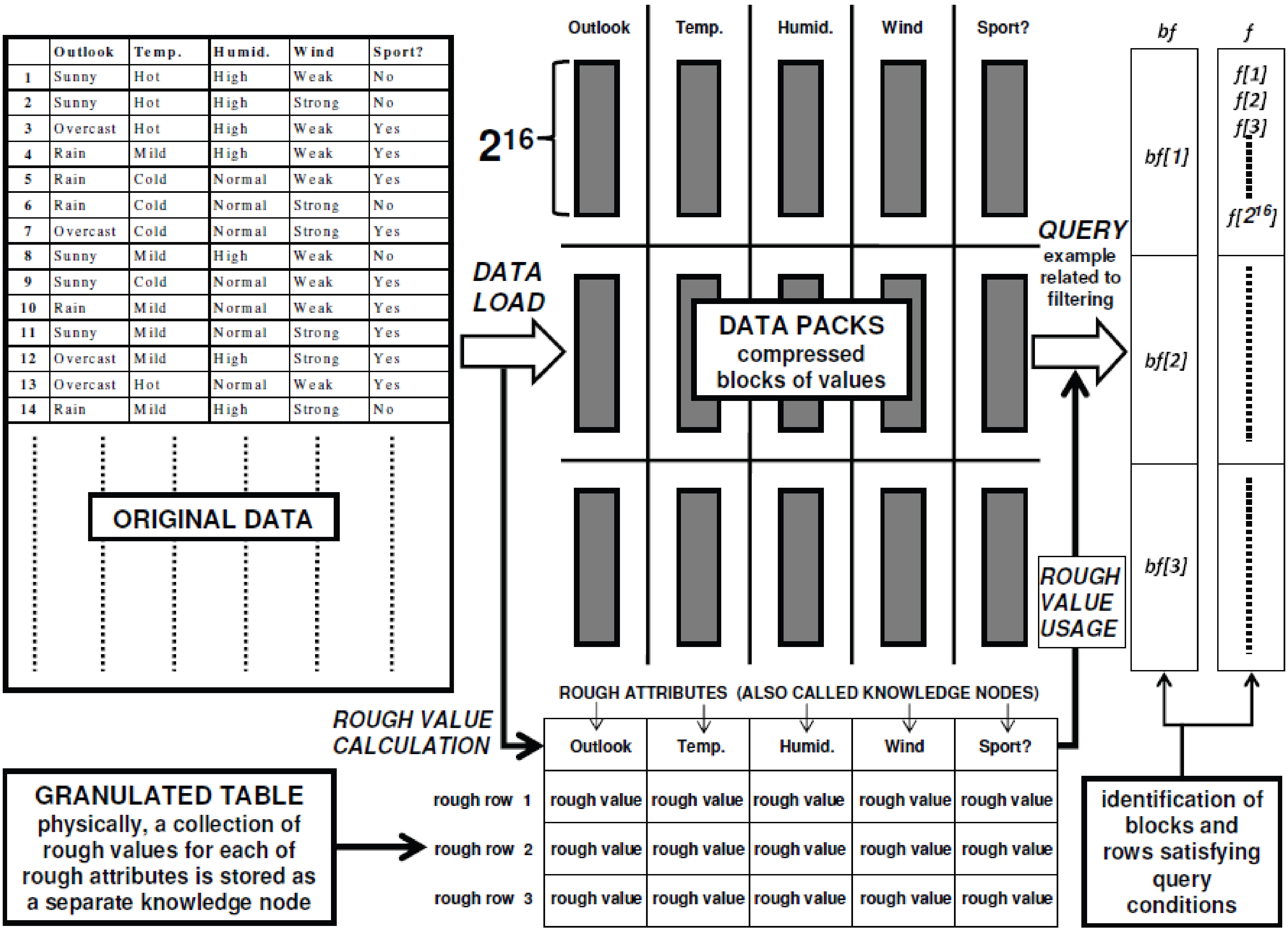
- Analysis can be based e.g. on distances

  $$\mathrm{Disc}(a/b)+\mathrm{Disc}(b/a)$$

- It can be also more sensitive with respect to interactions with the rest of attributes

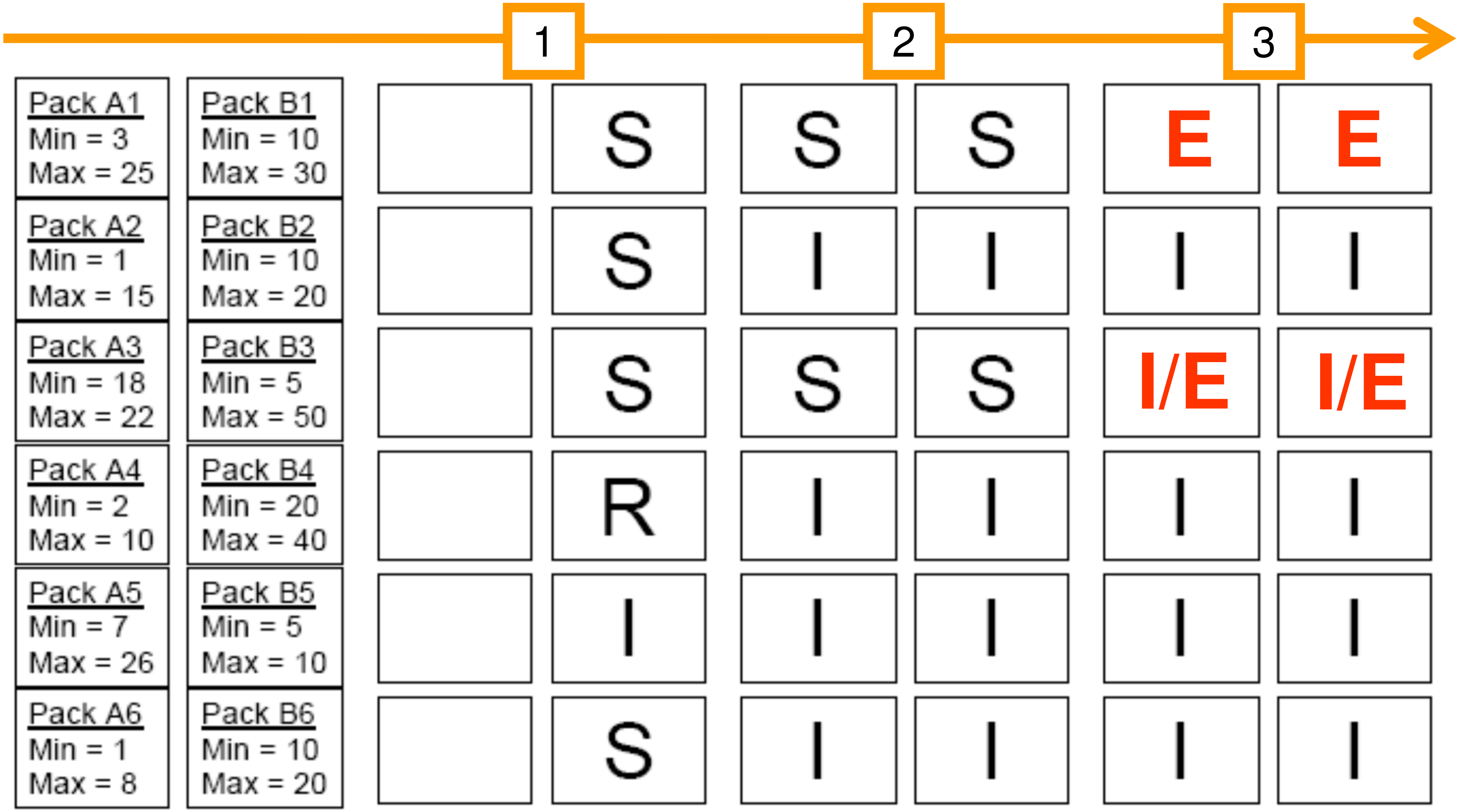
Computing with Object Sets





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
DATA LOAD
2^{16}
Outlook
Temp.
Humid.
Wind
Sport?
DATA PACKS
compressed
blocks of values
QUERY
example
related to
filtering
bf
f
bf[1]
bf[2]
bf[3]
f[1]
f[2]
f[3]
$f[2^{16}]$
ROUGH VALUE USAGE
ROUGH VALUE CALCULATION
ROUGH ATTRIBUTES (ALSO CALLED KNOWLEDGE NODES)
Outlook Temp. Humid. Wind Sport?
rough row 1 rough value rough value rough value rough value rough value
rough row 2 rough value rough value rough value rough value rough value
rough row 3 rough value rough value rough value rough value rough value
GRANULATED TABLE
physically, a collection of
rough values for each of
rough attributes is stored as
a separate knowledge node
identification of
blocks and
rows satisfying
query
conditions

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



I/S/R denotes irrelevant/suspect/relevant; E – exact computation (decompression)

Further Extensions

Association Reducts

- Association reduct (**C**,**D**) is supposed to represent <u>strong</u> dependency of **D** on **C**
- Association reduct is supposed to be:
  - Non-Extendible: impossible to add attributes to **D** without losing <u>strong</u> dependency on **C**
  - Irreducible: impossible to remove attributes from **C** and keep <u>strong</u> determination of **D**
- Association reduct is <u>most informative</u> if card(**C**) is smallest comparing to card(**D**)

Illustration

	a	b	c	d	e	f
u1	1	1	1	1	1	1
u2	0	0	0	1	1	1
u3	1	0	1	1	0	1
u4	0	1	0	0	0	0
u5	1	0	0	0	0	1
u6	1	1	1	1	1	0
u7	0	1	1	0	1	2

abc $\Rightarrow$ de is:

- <u>non-extendable</u>
  - not abc $\Rightarrow$ def
- <u>irreducible</u>
  - not ab $\Rightarrow$ de
  - not ac $\Rightarrow$ de
  - not bc $\Rightarrow$ de

How many reducts?

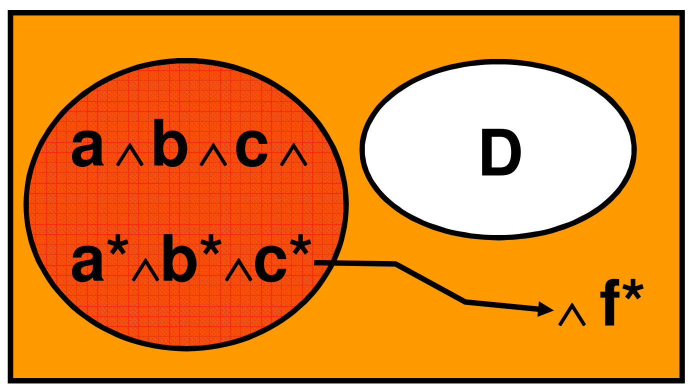
	a	**b**	**c**	**d**	**e**	**f**
u1	1	1	1	1	1	1
u2	0	0	0	1	1	1
u3	1	0	1	1	0	1
u4	0	1	0	0	0	0
u5	1	0	0	0	0	1
u6	1	1	1	1	1	0
u7	0	1	1	0	1	2

- abc $\Rightarrow$ de
- abdf $\Rightarrow$ ce
- abf $\Rightarrow$ e
- ace $\Rightarrow$ bd
- acf $\Rightarrow$ d
- ade $\Rightarrow$ bc
- adf $\Rightarrow$ c
- aef $\Rightarrow$ b
- bcd $\Rightarrow$ ae
- bde $\Rightarrow$ ac
- bef $\Rightarrow$ a
- cdf $\Rightarrow$ a
- cef $\Rightarrow$ abd

Boolean Representation

- We build formula α with *prime implicants* corresponding to the association reducts
- We use two types of Boolean variables:
  - **a** is truth iff attribute **a** belongs to **C**, in (**C**,**D**)
  - **a*** is truth iff attribute **a** does not belong to **D**
- We want association reducts (**C**,**D**) to look like: $\Lambda_{\mathbf{a}\in \mathbf{C}}\, \mathbf{a} \wedge \Lambda_{\mathbf{a}\notin \mathbf{D}}\, \mathbf{a^*}$

  (elements of **C** count twice!)



Most Interesting Reducts

- Given association reduct (C,D), we evaluate it with the value F(|C|,|D|)
- Function F: $N \times N \rightarrow R$ should hold:

  IF $n1 < n2$ THEN $F(n1,m) > F(n2,m)$

  IF $m1 < m2$ THEN $F(n,m1) < F(n,m2)$
- F(|C|,|D|) is maximized subject to # from the space of approximation parameters
- Such maximization problem is NP-hard

We proceed analogously to [9]. We reduce the Minimal Dominating Set Problem (MDSP) to $F\Theta$ARP. MDSP, widely known as NP-hard, is defined by INPUT as undirected graph $\mathcal{G} = (A, E)$, and OUTPUT as the smallest $B \subseteq A$ such that $Cov_{\mathcal{G}}(B) = A$, where $Cov_{\mathcal{G}}(B) = B \cup \{a \in A : \exists_{b \in B}(a, b) \in E\}$. To reduce MDSP to $F\Theta$ARP, we construct information system $\mathbb{A}_{\mathcal{G}} = (U_{\mathcal{G}}, A_{\mathcal{G}})$, $U_{\mathcal{G}} = \{u_1, \ldots, u_n, o_1, \ldots, o_n, u_*\}$, $A_{\mathcal{G}} = \{a_1, \ldots, a_n, a_*\}$, $n = |A|$, as follows:

$$\begin{array}{llll} a_i(u_j) = 1 \Leftrightarrow i = j \vee (i, j) \in E & & a_i(u_j) = 0 & \text{otherwise} \\ a_i(o_j) = 1 \Leftrightarrow i = j & & a_i(o_j) = 2 & \text{otherwise} \\ a_i(u_*) = 0\,, \quad a_*(u_j) = 0 & & a_*(o_j) = 0\,, & a_*(u_*) = 1 \end{array} \tag{10}$$

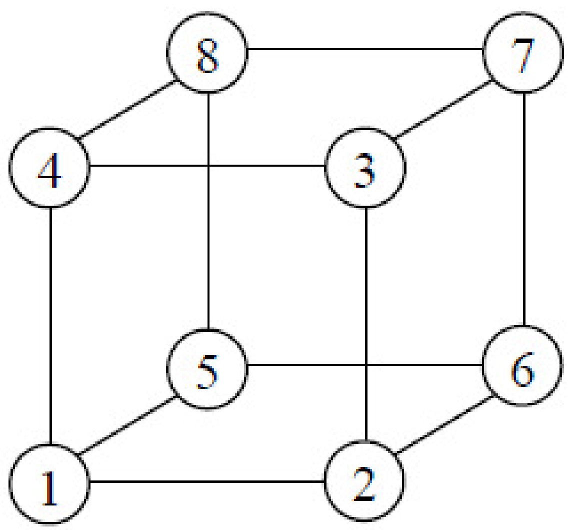


Fig. 1. $\mathcal{G} = (A, E)$ with 8 nodes and $\mathbb{A}_{\mathcal{G}} = (U_{\mathcal{G}}, A_{\mathcal{G}})$ constructed using (10).

$U_{\mathcal{G}}$	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_*
u_1	1	1	0	1	1	0	0	0	0
u_2	1	1	1	0	0	1	0	0	0
u_3	0	1	1	1	0	0	1	0	0
u_4	1	0	1	1	0	0	0	1	0
u_5	1	0	0	0	1	1	0	1	0
u_6	0	1	0	0	1	1	1	0	0
u_7	0	0	1	0	0	1	1	1	0
u_8	0	0	0	1	1	0	1	1	0
o_1	1	2	2	2	2	2	2	2	0
$\vdots$									$\vdots$
o_8	2	2	2	2	2	2	2	1	0
u_*	0	0	0	0	0	0	0	0	1

Summary

- Interesting tasks of rough-set-based feature subset selection are NP-hard
- Complexity should refer also to searching for optimal ensembles of feature subsets
- Complexity relates also to such tasks as creating features, comptuting measures...
- We should consider rough set extensions also from the computational point of view

evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

THANK YOU!!!

Dominik Ślęzak
U of Warsaw & Infobright Inc., Poland
slezak@{mimuw.edu.pl;infobright.com}