Data mining using Rough Sets

Alber Sánchez[1]
alber.ipia@inpe.br

[1]Instituto Nacional de Pesquisas Espaciais,
São José dos Campos, SP, Brazil

Referata Geoinformatica, 2015

Table of Contents

Rough Set Theory

- Who? Zdzisław Pawlak
- When? In the 80's
- What? Classificatory analysis of data tables.
- Why? To synthesize approximations of concepts from data.

Cutting to the chase - It allows to go from this...

Table 1 : Decision table

	Diploma	Experience	French	Reference	Decision
x_1	MCE	Low	No	Good	Stand By
x_2	MCE	Low	No	Neutral	Stand By
x_3	MBA	Low	No	Neutral	Rejected
x_4	MCE	Medium	No	Good	Rejected
x_5	MCE	Medium	No	Excellent	Accept
x_6	Msc	Medium	No	Excellent	Accept
x_7	Msc	High	Yes	Excellent	Accept
x_8	Msc	High	Yes	Excellent	Accept

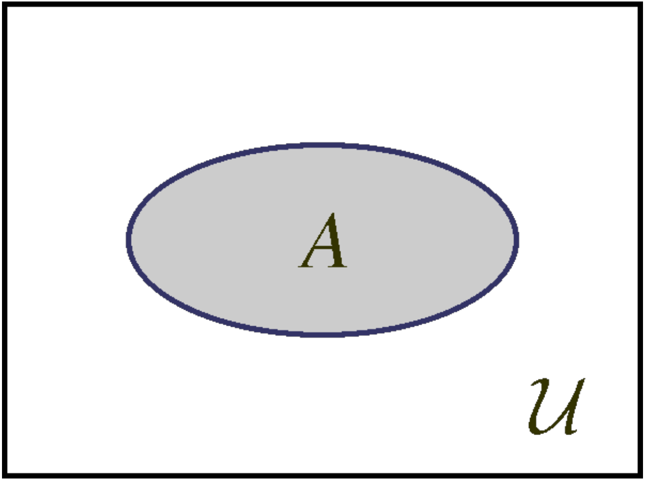
To this...

Table 2 : Core values of conditions

U	*Diploma*	*Experience*	*Reference*	*Decision*
x_1	MCE	Low	*	Stand By
x_2	MBA	*	*	Rejected
x_3	*	Medium	Good	Rejected
x_4	*	*	Excellent	Accept

Venn diagram

Figure 1 : A Set.



Taken from http://staff.www.ltu.se/~larserik/applmath/chap10en/

Information systems

Also known as tables

- $\mathcal{A} = (U, A)$
- U: non-empty finite set of objects
- A: non-empty finite set of attributes
- $a : U \to V_a, \quad \forall a \in A$
- V_a is the *value set of a*.

Table 3 : An information system

	Age	LEMS
x_1	16 - 30	50
x_2	16 - 30	0
x_3	31 - 45	1 - 25
x_4	31 - 45	1 - 25
x_5	46 - 60	26 - 49
x_6	16 - 30	26 - 49
x_7	46 - 60	26 - 49

Decision tables

Table 4 : An Decision Table

	Age	LEMS	Walk
x_1	16 - 30	50	Yes
x_2	16 - 30	0	No
x_3	**31 - 45**	**1 - 25**	**No**
x_4	**31 - 45**	**1 - 25**	**Yes**
x_5	46 - 60	26 - 49	No
x_6	16 - 30	26 - 49	Yes
x_7	46 - 60	26 - 49	No

Equivalence

- Equivalence relation
  - $R \subseteq X \times X$
  - Binary
  - Reflexive (xRx)
  - Symmetric ($xRy \iff yRx$)
  - Transitive ($xRy \wedge yRz \implies xRz$)
- Equivalence class
  - The EC of $x \in X$ consists all of $y \in X \quad | \quad xRy$

Indiscernibility relation

- $IND_A(B)$ is called the *B-Indiscernibility relation*
- Let $A = (U, A)$ be an Information System
- Then $\forall \quad B \subseteq A \quad \exists \quad IND_A(B)$
- Where $IND_A(B) = \{(x, x') \in U^2 \quad | \quad \forall a \in B \quad a(x) = a(x')\}$

Indiscernibility relation example

Table 5 : A decision Table

	Age	LEMS	Walk
x_1	16 - 30	50	Yes
x_2	16 - 30	0	No
x_3	31 - 45	1 - 25	No
x_4	31 - 45	1 - 25	Yes
x_5	46 - 60	26 - 49	No
x_6	16 - 30	26 - 49	Yes
x_7	46 - 60	26 - 49	No

- $IND(\{Age\}) = \{\{x1, x2, x6\}, \{x3, x4\}, \{x5, x7\}\}$
- $IND(\{LEMS\}) = \{\{x1\}, \{x2\}, \{x3, x4\}, \{x5, x6, x7\}\}$
- $IND(\{Age, LEMS\}) = \{\{x1\}, \{x2\}, \{x3, x4\}, \{x5, x7\}, \{x6\}\}$
- The equivalence classes of the B-indiscernibility relation are denoted $[x]_B$

Set approximation

- The concept *walk* cannot be defined as a crisp set using *Age* and *LEMS* because of $\{x3, x4\}$
- However, we can approximate it using 3 sets.
  - Those objects which fulfil $Walk = Yes$
  - Those objects which fulfil $Walk = No$
  - The remaining objects

Set approximation 2

- Let $A = (U, A)$ be a IS
- Let $B \subseteq A$
- Let $X \subseteq U$

X can be approximated using only the information contained in B using 3 sets:

- *B-lower approximation of X*, $\underline{B}X = \{x \mid [x]_B \subseteq X\}$
- *B-upper approximation of X*, $\overline{B}X = \{x \mid [x]_B \cap X\}$
- *B-boundary region*, $BN_B = \overline{B}X - \underline{B}X$

Set approximation 3

On the basis of knowledge in B:

- Objects in $\underline{B}X$ can be with certainly classified as members of X
- Objects in $\overline{B}X$ can be only classified as possible members of X
- Objects we cannot decisively classify into X

Besides, there is the set *B-outside region of* X which is $U - \overline{B}X$

Rough Set definition

A set is said to be *rough* if the boundary region is non-empty.

Rough Set example

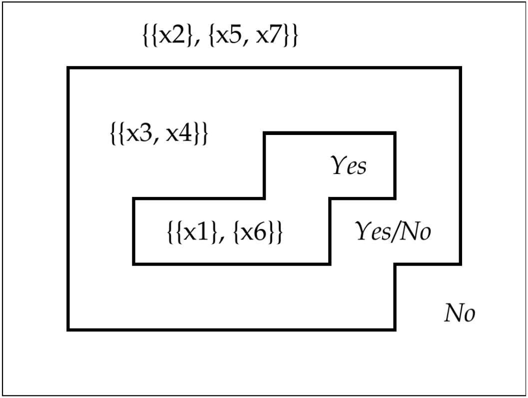
Table 6 : A decision Table

	Age	LEMS	Walk
x_1	16 - 30	50	Yes
x_2	16 - 30	0	No
x_3	31 - 45	1 - 25	No
x_4	31 - 45	1 - 25	Yes
x_5	46 - 60	26 - 49	No
x_6	16 - 30	26 - 49	Yes
x_7	46 - 60	26 - 49	No

- Let $W = \{x|Walk(x) = yes\}$, then:
  - $\underline{\underline{A}}W = \{x1, x6\}$
  - $\overline{A}W = \{x1, x3^{***}, x4, x6\}$
  - $BN_A(W) = \{x3, x4\}$
  - $U - \overline{A}W = \{x2, x5, x7\}$

Rough Set graphic example

Figure 2 : A rough set.



Rough Set properties

1. $\underline{B}(X) \subseteq X \subseteq \overline{B}(X)$
2. $\underline{B}(\emptyset) = \overline{B}(\emptyset), \quad \underline{B}(U) = \overline{B}(U) = U$
3. $\overline{B}(X \cup Y) = \overline{B}(X) \cup \overline{B}(Y)$
4. $\underline{B}(X \cap Y) = \underline{B}(X) \cap \underline{B}(Y)$
5. $X \subseteq Y$ implies $\underline{B}(X) \subseteq \underline{B}(Y)$ and $\overline{B}(X) \subseteq \overline{B}(Y)$
6. $\underline{B}(X \cup Y) \supseteq \underline{B}(X) \cup \underline{B}(Y)$
7. $\overline{B}(X \cap Y) \subseteq \overline{B}(X) \cap \overline{B}(Y)$
8. $\underline{B}(-X) = -\overline{B}(X)$
9. $\overline{B}(-X) = -\underline{B}(X)$
10. $\underline{B}(\underline{B}(X)) = \overline{B}(\underline{B}(X)) = \underline{B}(X)$
11. $\overline{B}(\overline{B}(X)) = \underline{B}(\overline{B}(X)) = \overline{B}(X)$

Where $-X$ denotes $U - X$

Rough Set classification

- X is *roughly B-definable*, iff $\underline{B}(X) \neq \emptyset$ and $\overline{B}(X) \neq U$
- X is *internally B-indefinable*, iff $\underline{B}(X) = \emptyset$ and $\overline{B}(X) \neq U$
- X is *externally B-indefinable*, iff $\underline{B}(X) \neq \emptyset$ and $\overline{B}(X) = U$
- X is *totally B-indefinable*, iff $\underline{B}(X) = \emptyset$ and $\overline{B}(X) = U$

Accuracy of approximation

$$\alpha_B(X) = \frac{|\underline{B}(X)|}{|\overline{B}(X)|}, \text{ where } |X| \text{ is the cardinality of } X \neq \emptyset$$

- $0 \leq \alpha_B(X) \leq 1$
- if $\alpha_B(X) = 1$, X is *crisp* with respect to B
- If $\alpha_B(X) < 1$, X is *rough* with respect to B

Quality of approximation

$$\gamma_B(X) = \frac{|\underline{B}(X)|}{|U|}\text{, where } |X| \text{ is the cardinality of } X \neq \emptyset$$

It express the percentage of possible correct decisions when classifying objects employing the knowledge B

Table of Contents

Reducts

Let $A = (U, A)$

A reduct of A is a minimal set of attributes $B \subseteq A$ such that $IND_A(B) = IND_A(A)$

A reduct is a minimal set of attributes from A that preserves the partitioning of the universe, and hence, the ability to perform classifications as the whole attribute set A does.

Reduct example

$$A = (U, \{Diploma, Experience, French, Reference\})$$

Table 7 : An unreduced decision table

	Diploma	**Experience**	French	**Reference**	Decision
x_1	MBA	Medium	Yes	Excellent	Accept
x_2	MBA	Low	Yes	Neutral	Reject
x_3	MCE	Low	Yes	Good	Reject
x_4	Msc	High	Yes	Neutral	Accept
x_5	Msc	Medium	Yes	Neutral	Reject
x_6	Msc	High	Yes	Excellent	Accept
x_7	MBA	High	No	Good	Accept
x_8	MCE	Low	No	Excellent	Reject

Discernibility matrix and function

DM is a symmetric *nxn* matrix which entries are:
$c_{ij} = \{a \in A | a(x_i) \neq a(x_j)\}$ for $i, j = 1, ..., n$

DF f_A is a Boolean function of m Boolean variables $a_1^*, ..., a_m^*$ (corresponding to attributes $a_1, ..., a_m$) defined as below, where $c_{ij}^* = \{a^* | a \in c_{ij}\}$

$$f_A(a_1^*, ..., a_m^*) = \bigwedge\{\bigvee c_{ij}^* | 1 \leq j \leq i \leq n, c_{ij} \neq \emptyset\}$$

The set of all prime implicants of f_A determines the set of all reducts of A[1]

---

[1] An implicant of a Boolean function f is any conjuntion of literals (variables or their negations) such that if the values of that literals are true under an arbitrary valuation v of variables then thge value of the function f under v is also true. A rpime implicant is a minimal implicant. Here we are interested in implicants of monotone Boolean functions only i.e. functions constructed without negation.

k-relative discernibility function & reducts

Resulting from constructing a Boolean function by restricting the conjuntion to only run over column k in the discernibility matrix (instead of all the columns).

The set of all prime implicants of this function determines the set of all *k-relative reducts* of A. These reducts reveal the minimum amount of information needed to discern $x_k \in U$ (or more pecisely $[x_k] \subseteq U$) from all other objects.

Table of Contents

Example

Table 8 : Decision table

	Diploma	Experience	French	Reference	Decision
x_1	MCE	Low	No	Good	Stand By
x_2	MCE	Low	No	Neutral	Stand By
x_3	MBA	Low	No	Neutral	Rejected
x_4	MCE	Medium	No	Good	Rejected
x_5	MCE	Medium	No	Excellent	Accept
x_6	Msc	Medium	No	Excellent	Accept
x_7	Msc	High	Yes	Excellent	Accept
x_8	Msc	High	Yes	Excellent	Accept

Encode values

- Diploma
  - 0 → MBA
  - 1 → MCE
  - 2 → Msc
- Experience
  - 0 → Low
  - 1 → Medium
  - 2 → High
- French
  - 0 → No
  - 2 → Yes
- Reference
  - 0 → Neutral
  - 1 → Good
  - 2 → Excellent
- Decision
  - 0 → Rejected
  - 1 → Stand By
  - 2 → Accept

Table 9 : Encoded decision table

U	a	b	c	d	e
x_1	1	0	0	1	1
x_2	1	0	0	0	1
x_3	0	0	0	0	0
x_4	1	1	0	1	0
x_5	1	1	0	2	2
x_6	2	1	0	2	2
x_7	2	2	2	2	2
x_8	2	2	2	2	2

Compute indiscernibility relation

Table 10 : Encoded decision table

U	a	b	c	d	e
x_1	1	0	0	1	1
x_2	1	0	0	0	1
x_3	0	0	0	0	0
x_4	1	1	0	1	0
x_5	1	1	0	2	2
x_6	2	1	0	2	2
x_7	2	2	2	2	2

- $IND\{a\} = \{\{x_1, x_2, x_4, x_5\}, \{x_3\}, \{\{x_6, \{x_7\}\}$
- $IND\{a, b, c\} = \{\{x_1, x_2\}, \{x_3\}, \{x_4, x_5\}, \{x_6\}, \{x_7\}\}$
- (...)
- $IND\{a, b, d\} = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5\}, \{x_6\}, \{x_7\}\}$
- $IND\{a, b, c, d\} = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5\}, \{x_6\}, \{x_7\}\}$
- Attribute c is superfluous because $IND\{a, b, d\} = IND\{a, b, c, d\}$

Compute core values of conditions

Table 11 : Reduced decision table

U	a	b	d	e
x_1	**1**	**0**	**1**	**1**
x_2	**1**	**0**	0	**1**
x_3	0	**0**	0	0
x_4	**1**	1	**1**	0
x_5	**1**	1	2	2
x_6	2	1	2	2
x_7	2	2	2	2

For rule x_1:

- $F = \{[x_1]_a, [x_1]_b, [x_1]_d\}$
- $F = \{\{x_1, x_2, x_4, x_5\}, \{x_1, x_2, x_3\}, \{x_1, x_4\}\}$

Consider that

- $[x_1]_{a,b,d} = [x_1]_a \cap [x_1]_b \cap [x_1]_d = \{x_1\}$
- $[x_1]_e = \{x_1, x_2\}$

Find a smaller relation being a subset of $[x_1]_e$

- $[x_1]_b \cap [x_1]_d = \{x_1\} \subseteq [x_1]_e$
- $[x_1]_a \cap [x_1]_d = \{x_1, x_4\}$
- $[x_1]_a \cap [x_1]_b = \{x_1, x_2\} \subseteq [x_1]_e$

So, $b(x_1) = 0$ is a *core value* because it is present in $[x_1]_b \cap [x_1]_d$ and $[x_1]_a \cap [x_1]_b$, both are subsets of $[x_1]_e$

Result of computing the core values of conditions

Table 12 : Core values of conditions

U	a	b	d	e
x_1	-	0	-	1
x_2	1	-	-	1
x_3	0	-	-	0
x_4	-	1	1	0
x_5	-	-	2	2
x_6	-	-	-	2
x_7	-	-	-	2

Compute value reducts

For rule x_1:
$F = \{[x_1]_a, [x_1]_b, [x_1]_d\} = \{\{x_1, x_2, x_4, x_5\}, \{x_1, x_2, x_3\}, \{x_1, x_4\}\}$

We need to find all subfamilies $G \subseteq F | \bigcap G \subseteq [x_1]_e = \{x_1, x_2\}$

- $[x_1]_b \cap [x_1]_d = \{x_1, x_2, x_3\} \cap \{x_1, x_4\} = \{x_1\} \subseteq [x_1]_e$
- $[x_1]_a \cap [x_1]_d = \{x_1, x_2, x_4, x_5\} \cap \{x_1, x_4\} = \{x_1, x_4\}$
- $[x_1]_a \cap [x_1]_b = \{x_1, x_2, x_4, x_5\} \cap \{x_1, x_2, x_3\} = \{x_1, x_2\} \subseteq [x_1]_e$

So, only $[x_1]_b \cap [x_1]_d$ and $[x_1]_a \cap [x_1]_b$ are reducts of the family F

Results of computing value reducts

Table 13 : Core values of conditions

U	a	b	d	e
x_1	1	0	*	1
x_1'	*	0	1	1
x_2	1	0	*	1
x_2'	1	*	0	1
x_3	0	*	*	0
x_4	*	1	1	0
x_5	*	*	2	2
x_6	*	*	2	2
x_6'	2	*	*	2
x_7	*	*	2	2
x_7'	*	2	*	2
x_7''	2	*	*	2

Many possible solutions

Table 14 : Core values of conditions

U	a	b	d	e
x_1	1	0	*	1
x_2	1	*	0	1
x_3	0	*	*	0
x_4	*	1	1	0
x_5	*	*	2	2
x_6	*	*	2	2
x_7	2	*	*	2

Table 15 : Core values of conditions

U	a	b	d	e
x_1	1	0	*	1
x_2	1	0	*	1
x_3	0	*	*	0
x_4	*	1	1	0
x_5	*	*	2	2
x_6	*	*	2	2
x_7	*	*	2	2

Minimal solution

After removing duplicates and re-numbering

Table 16 : Core values of conditions

U	a	b	d	e
x_1	1	0	*	1
x_2	0	*	*	0
x_3	*	1	1	0
x_4	*	*	2	2

Minimal solution decoded

Table 17 : Core values of conditions

U	*Diploma*	*Experience*	*Reference*	*Decision*
x_1	MCE	Low	*	Stand By
x_2	MBA	*	*	Rejected
x_3	*	Medium	Good	Rejected
x_4	*	*	Excellent	Accept

Table of Contents

Applications

- Data mining
- AI

Sense, plan, act

Figure 3 : Sense, plan, act cycle.

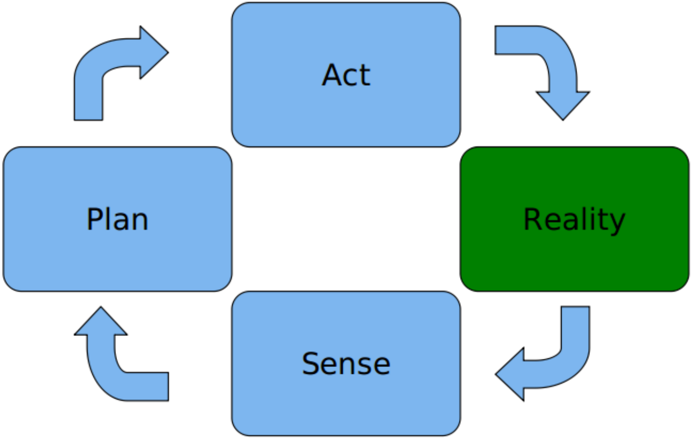


Table of Contents

Software

- R packages *RoughSetKnowledgeReduction* and *RoughSets*
- RSES - Rough Set Exploration System
  `http://logic.mimuw.edu.pl/~rses/start.html`
- Infobright Community Edition `http://www.infobright.org`

Table of Contents

References I

Pawlak, Zdzisław.
Rough Sets: Theoretical Aspects of Reasoning about Data.
Springer Netherlands, 1991.

Komorowski, Jan et al.
Rough sets: A tutorial.
Rough fuzzy hybridization: A new trend in decision-making, pages 3–98, 1999.