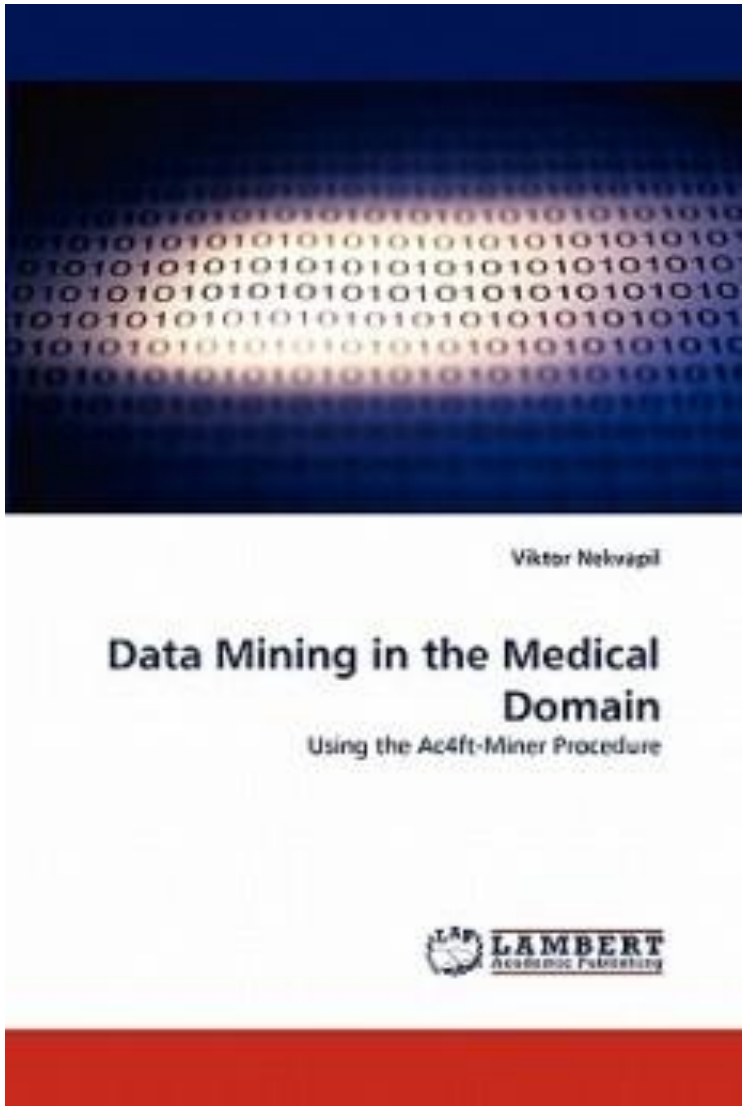
USING THE AC4FT-MINER PROCEDURE IN THE MEDICAL DOMAIN

Viktor Nekvapil

About the author

- VŠE: IT – 4IQ, 2. semestr
- Bachelor thesis:
  Using the Ac4ft-Miner procedure in the medical domain
- Supervisor: doc. Rauch
- Published



About the Ac4ft-Miner

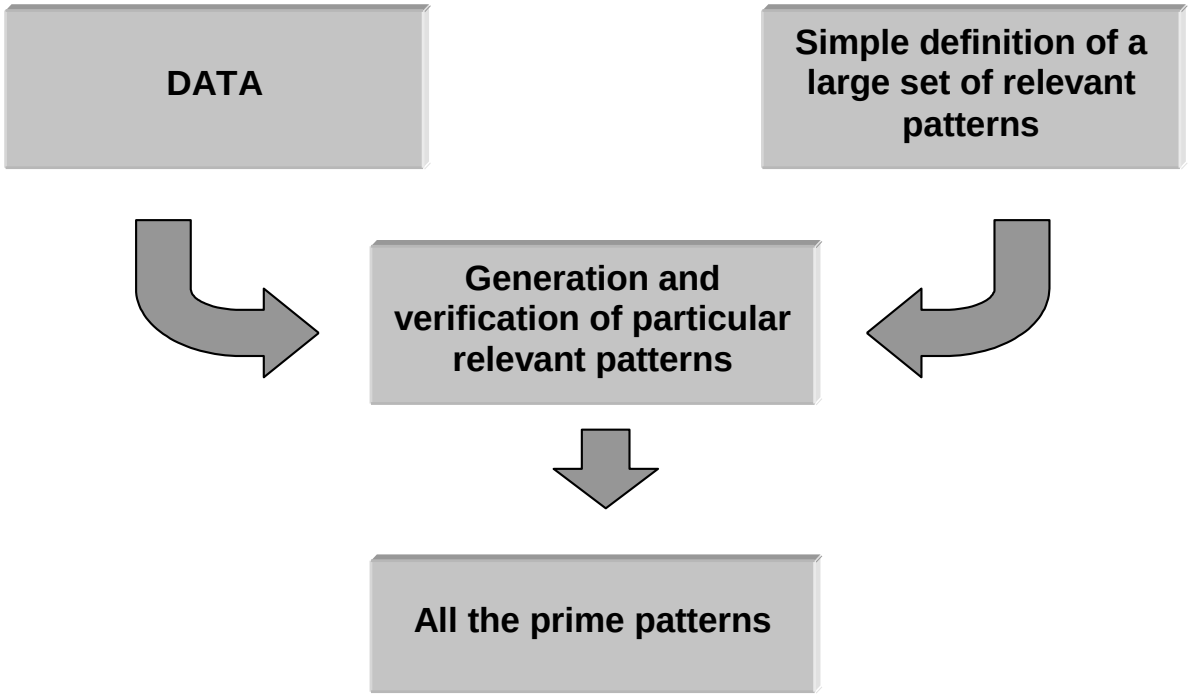
- *Mines for rules that express which actions should be performed to improve the defined state*
- Newest procedure of the LISp-Miner System
- Authors: Doc. RNDr. Jan Rauch, CSc., Ing. Milan Šimůnek, Ph.D.
- Implementation of the GUHA method
- Mines for G-action rules
- Expressed by two 4ft-association rules
- Inspired by action rules according to [Ras,Wieczorkowska, 2000]

Contents

Contents

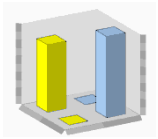
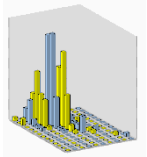
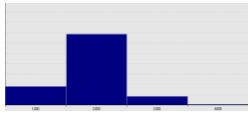
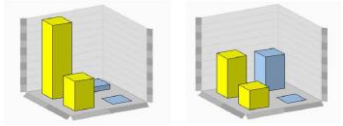
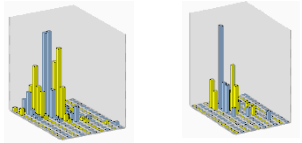
GUHA Method

- Offers all interesting patterns true in given data
- Method of exploratory data analysis
- Implemented by GUHA procedures



LISp-Miner, 7 GUHA procedures

http://lispminer.vse.cz

- 4ft-Miner 
- KL-Miner 
- CF-Miner 
- Ac4ft-Miner
- SD4ft-Miner 
- SDKL-Miner 
- SDCF-Miner 

Source: [Rauch, Šimůnek, c2011]

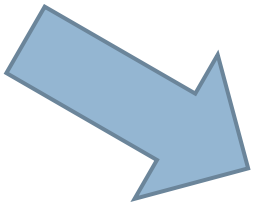
Data representation in LISp-Miner

- One database table formes data matrix $\mathcal{M}$
- Data matrix $\mathcal{M}$ ⇒ Boolean attributes(literals)

object i.e. row of $\mathcal{M}$	columns of $\mathcal{M}$ i.e. attributes				⇒ examples of literals	
	A_1	A_2	...	A_{50}	$A_1(1,2)$	$\neg A_{50}(6)$
o_1	1	4	...	4	T	T
o_2	4	3	...	6	F	F
o_3	2	6	...	7	T	T
⋮	⋮	⋮	⋱	⋮	⋮	⋮
o_n	3	1	...	36	F	T

Data representation in LISp-Miner (example)

Objects	Attributes						
Patient	Sex	Age	Type of therapy	Success	Genetic predisposition	City	…
1	male	42	none	no	no	Prague	…
2	female	61	diet	yes	no	Čáslav	…
3	female	24	surgery	no	yes	Čáslav	…
4	male	54	medicaments	yes	no	Prague	…
…	…	…	…	…	…	…	…
632	female	57	medicaments	yes	no	Prague	…

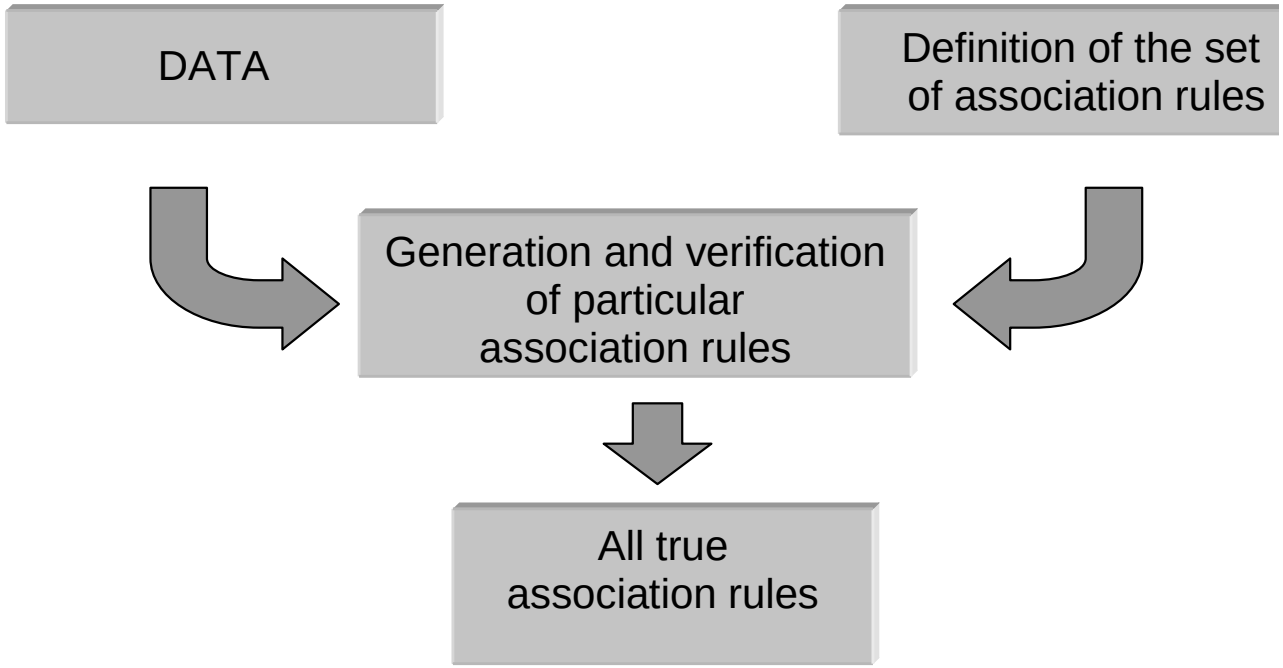


Objects	Basic Boolean attributes		Derived Boolean attributes		
Patient	Sex (male)	Type of therapy (surgery)	Sex (male) ∨ success (yes)	Genetic predisp. (no) ∧ Age ⟨50,60)	Sex (male) ∧ (Type of therapy (diet) ∨ Type of therapy (medicaments)) ∧ ¬ Age ⟨50,60)
1	true	false	true	false	false
2	false	false	true	false	true
3	false	true	false	false	false
4	true	false	true	true	false
…	…	…	…	…	…
632	false	false	true	true	false

Contents

GUHA method and Association rules – 4ft-Miner

- Mines for enhanced association rules = not just implication



Source: [Rauch, Šimůnek, c2011]

4ft-Miner procedure

	$\varphi \approx \psi$
φ	antecedent
ψ	succedent
$\approx$	4ft-quantifier

	ψ	$\neg\ \psi$
φ	a	b
$\neg\ \varphi$	c	d

object	Attributes				Boolean attributes			
	A_1	A_2	…	A_K	$A_1(6)$	$A_2(1,4)$	$A_2(1,4) \wedge A_K(2,7)$	…
o_1	6	4	…	2	1	1	1	…
o_2	9	3	…	5	0	0	0	…
…	…	…	…	…	…	…	…	…
o_n	4	1	…	3	0	1	0	…

Bit-string approach to mine association rules

- Apriori algorithm is not used

object i.e. row of $\mathcal{M}$	columns of $\mathcal{M}$ i.e. attributes				examples of literals	
	A_1	A_2	...	A_{50}	$A_1(1,2)$	$\neg A_{50}(6)$
o_1	1	4	...	4	T	T
o_2	4	3	...	6	F	F
o_3	2	6	...	7	T	T
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
o_n	3	1	...	36	F	T

Source: [Rauch, Šimůnek, 2005]

Bit-string approach to mine association rules (2)

- Attribute A_1 with 4 categories (1, 2, 3, 4)

row of $\mathcal{M}$	A_1	cards of categories of A_1			
		$A_1[1]$	$A_1[2]$	$A_1[3]$	$A_1[4]$
o_1	1	1	0	0	0
o_2	4	0	0	0	1
o_3	2	0	1	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
o_n	3	0	0	1	0

bit-wise Boolean operations

$\dot{\wedge}$, $\dot{\vee}$ $\dot{\neg}$

$$\mathcal{C}(\varphi \wedge \psi) = \mathcal{C}(\varphi) \;\dot{\wedge}\; \mathcal{C}(\psi)$$

$$\mathcal{C}(\varphi \vee \psi) = \mathcal{C}(\varphi) \;\dot{\vee}\; \mathcal{C}(\psi)$$

$$\mathcal{C}(\neg\varphi) = \dot{\neg}\, \mathcal{C}(\varphi)$$

$$\mathcal{C}(A_1(1,2)) = A_1[1] \;\dot{\vee}\; A_1[2]$$

Source: [Rauch, Šimůnek, 2005]

Bit-string approach to mine association rules (3)

4ft-table $4ft(\varphi, \psi, \mathcal{M})$ of φ and ψ on $\mathcal{M}$

$\mathcal{M}$	ψ	$\neg\psi$
φ	a	b
$\neg\varphi$	c	d

$$a = Count(\mathcal{C}(\varphi) \dot{\wedge} \mathcal{C}(\psi))$$
$$b = Count(\mathcal{C}(\varphi)) - a$$
$$c = Count(\mathcal{C}(\psi)) - a$$
$$d = n - a - b - c$$

$Count(\xi)$ =
= number of „1“ in ξ

Source: [Rauch, Šimůnek, 2005]

4ft-quantifiers

$\varphi \approx \psi$

	ψ	$\neg\ \psi$
φ	a	b
$\neg\ \varphi$	c	d

$\varphi \Rightarrow_{p,Base} \psi$ $\quad \frac{a}{a+b} \geq p \wedge a \geq Base$

$\varphi \Leftrightarrow_{p,Base} \psi$ $\quad \frac{a}{a+b+c} \geq p \wedge a \geq Base$

$\varphi \equiv_{p,Base} \psi$ $\quad \frac{a+d}{a+b+c+d} \geq p \wedge a \geq Base$

$\varphi \Rightarrow^{+}_{p,Base} \psi$ $\quad \frac{a}{a+b} \geq (1+p)\frac{a+c}{a+b+c+d} \wedge a \geq Base$

… and many other possibilities

Contents

Action rules

according to [Ras,Wieczorkowska, 2000]

- suggest a change in behaviour that can bring us an advantage
- Two sets of attributes: stable and flexible

Stable attributes

R: $(A_1 = \omega_1) \wedge \ldots \wedge (A_Q = \omega_Q) \wedge$

$(B_1, \alpha_1 \rightarrow \beta_1) \wedge \ldots \wedge (B_P, \alpha_P \rightarrow \beta_P) \Rightarrow (D, k_1 \rightarrow k_2)$

Flexible attributes **Decision**

Action rules – support and confidence

R: $(A_1 = \omega_1) \wedge \ldots \wedge (A_Q = \omega_Q) \wedge (B_1, \alpha_1 \rightarrow \beta_1) \wedge \ldots \wedge (B_P, \alpha_P \rightarrow \beta_P) \Rightarrow (D, k_1 \rightarrow k_2)$

- n: the total number of objects in the database
- CPL(R): the number of objects matching $(\omega_1,\ldots,\omega_Q,\alpha_1,\ldots,\alpha_P, k_1)$
- CPR(R): the number of objects matching $(\omega_1,\ldots,\omega_Q,\beta_1,\ldots, \beta_P, k_2)$
- CVL(R): the number of objects matching $(\omega_1,\ldots,\omega_Q,\alpha_1,\ldots,\alpha_P)$
- CVR(R): the number of objects matching $(\omega_1,\ldots,\omega_Q,\beta_1,\ldots, \beta_P)$
- LeftSup(R) = $\frac{CPL(R)}{n}$ RightSup(R) = $\frac{CPR(R)}{n}$
- Sup(R) = LeftSup(R) = $\frac{CPL(R)}{n}$
- Conf(R) = $\frac{CPL(R)}{CVL(R)} * \frac{CPR(R)}{CVR(R)}$

Action rules - example

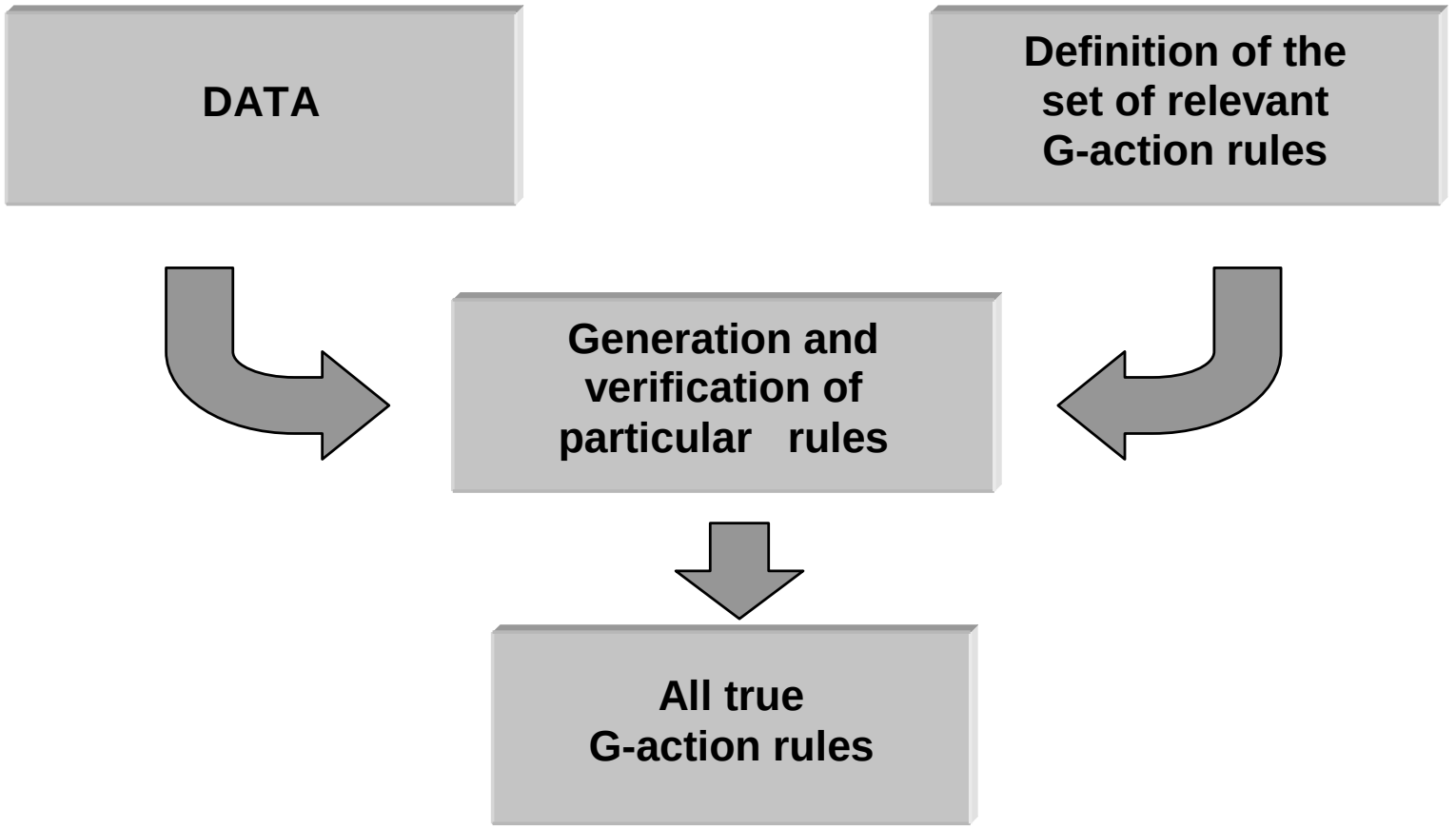
R: Sex (female) $\wedge$ Age $\langle 50,60)$ $\wedge$ Type of therapy (diet$\rightarrow$medicaments) $\Rightarrow$ Success (no$\rightarrow$yes)

- n = 632 — There are 632 patients in a database
- CPL(R) = 42 — (female, $\langle 50; 60)$, diet, no)
- CPR(R) = 74 — (female, $\langle 50; 60)$, medicaments, yes)
- CVL(R) = 90 — (female, $\langle 50; 60)$, diet)
- CVR(R) = 105 — (female, $\langle 50; 60)$, medicaments)

- Sup(R) = LeftSup(R) = $\frac{CPL(R)}{n} = \frac{42}{632} = 0.07$
- Conf(R) = $\frac{CPL(R)}{CVL(R)} * \frac{CPR(R)}{CVR(R)} = \frac{42}{90} * \frac{74}{105} = 0.33$

Contents

GUHA method and G-action rules – Ac4ft-Miner



Source: [Rauch, Šimůnek, 2009]

G-action rules

$$\varphi_{St} \wedge \Phi_{Chg} \approx^* \psi_{St} \wedge \Psi_{Chg}$$

M	Stable attributes			Flexible attributes		
object	A_1	…	A_Q	B_1	…	B_P
o_1	6	…	4	2	…	9
o_2	9	…	8	3	…	7
…	…	…	…	…	…	…
o_n	4	…	3	5	…	6

φ_{St} the stable antecedent (or antecedent stable part)

Φ_{Chg} the change of antecedent (or antecedent flexible part)

ψ_{St} the stable succedent (or succedent stable part),

Ψ_{Chg} the change of succedent (or succedent flexible part)

$\approx^*$ Ac4ft-quantifier

G-action rule

$$\varphi_{St} \wedge \Phi_{Chg} \approx^{*} \psi_{St} \wedge \Psi_{Chg}$$

Rule describing the initial state

$$R_I:\ \varphi_{St} \wedge I(\Phi_{Chg}) \approx_I \psi_{St} \wedge I(\Psi_{Chg})$$

$$R_I:\ \varphi_I \approx_I \psi_I$$

	ψ_I	$\neg\ \psi_I$
φ_I	a_I	b_I
$\neg\ \varphi_I$	c_I	d_I

Rule describing the final state

$$R_F:\ \varphi_{St} \wedge F(\Phi_{Chg}) \approx_F \psi_{St} \wedge F(\Psi_{Chg})$$

$$R_F:\ \varphi_F \approx_F \psi_F$$

	ψ_F	$\neg\ \psi_F$
φ_F	a_F	b_F
$\neg\ \varphi_F$	c_F	d_F

Ac4ft-quantifiers

$$\Rightarrow^{F > I}_{q, B_I, B_F} \quad \frac{a_F}{a_F + b_F} - \frac{a_I}{a_I + b_I} \geq q \wedge a_I \geq B_I \wedge a_F \geq B_F$$

$$\Rightarrow^{I > F}_{q, B_I, B_F} \quad \frac{a_I}{a_I + b_I} - \frac{a_F}{a_F + b_F} \geq q \wedge a_I \geq B_I \wedge a_F \geq B_F,$$

$$\Rightarrow^{I \lesseqgtr F}_{q, B_I, B_F} \quad \left| \frac{a_I}{a_I + b_I} - \frac{a_F}{a_F + b_F} \right| \geq q \wedge a_I \geq B_I \wedge a_F \geq B_F$$

where $0 < q \leq 1, B_I > 0$ and $B_F > 0$.

… and many other possibilities

	ψ_I	$\neg\ \psi_I$
φ_I	a_I	b_I
$\neg\ \varphi_I$	c_I	d_I

	ψ_F	$\neg\ \psi_F$
φ_F	a_F	b_F
$\neg\ \varphi_F$	c_F	d_F

G-action rule – example

R: Sex (female) $\wedge$ Age $\langle 50; 60)$ $\wedge$ Type of therapy (diet $\rightarrow$ medicaments) $\Rightarrow_{\substack{F > I \\ 0.386, 17, 25}}$ Success (yes)

Initial rule

R_I: Sex (female) $\wedge$ Age $\langle 50; 60)$ $\wedge$ Type of therapy (diet) $\Rightarrow_{0.607,17}$ Success (yes)

	Success (yes)	¬ Success (yes)
Sex (female) $\wedge$ Age $\langle 50; 60)$ $\wedge$ Type of therapy (diet)	17	11
¬ (Sex (female) $\wedge$ Age $\langle 50; 60)$ $\wedge$ Type of therapy (diet))	205	398

$$p = \frac{17}{17+11} = 0.607 \qquad B = 17$$

G-action rule – example (2)

R: Sex (female) $\wedge$ Age $\langle 50; 60)$ $\wedge$ Type of therapy (diet $\rightarrow$ medicaments) $\Rightarrow_{0.386,\,17,\,25}^{F > I}$ Success (yes)

Final rule

$R_{F:}$ Sex (female) $\wedge$ Age $\langle 50; 60)$ $\wedge$ Type of therapy (medicaments) $\Rightarrow_{0.893,25}$ Success (yes)

	Success (yes)	¬ Success (yes)
Sex (female) $\wedge$ Age $\langle 50; 60)$ $\wedge$ Type of therapy (medicaments)	25	3
¬ (Sex (female) $\wedge$ Age $\langle 50; 60)$ $\wedge$ Type of therapy (medicaments))	267	382

$$p = \frac{25}{25+3} = 0.893 \qquad B = 25$$

G-action rule - interpretation

R: Sex (female) $\wedge$ Age $\langle 50; 60)$ $\wedge$ Type of therapy (diet $\rightarrow$ medicaments) $\Rightarrow_{0.386,\,17,\,25}^{F > I}$ Success (yes)

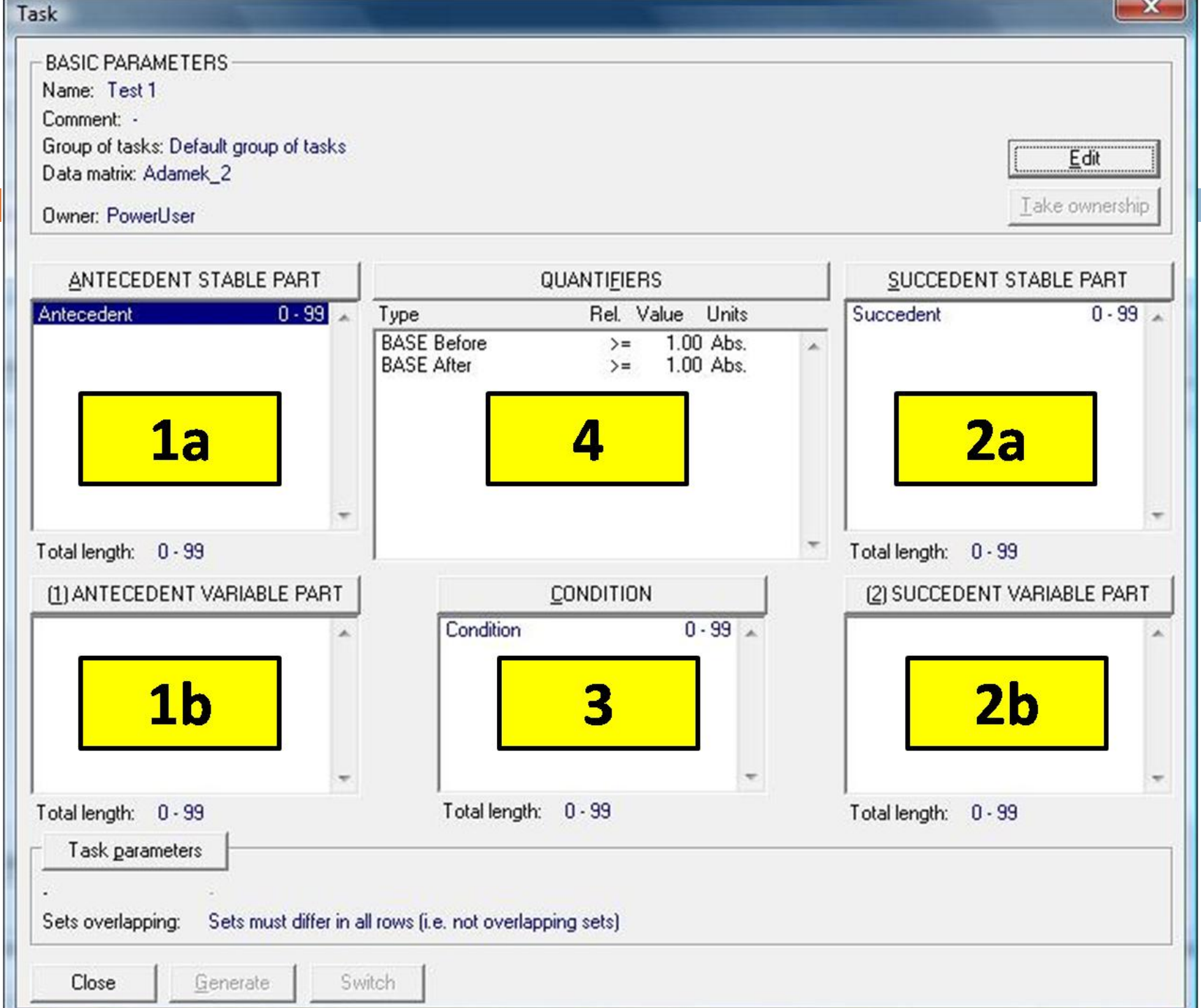
$q = p_F - p_I = 0.893 - 0.607 = 0.386$

„If the therapy is changed from diet to medicaments among female patients between the ages of 50 and 60, the probability of successful treatment increases by 38.6 percentage points.“

Contents

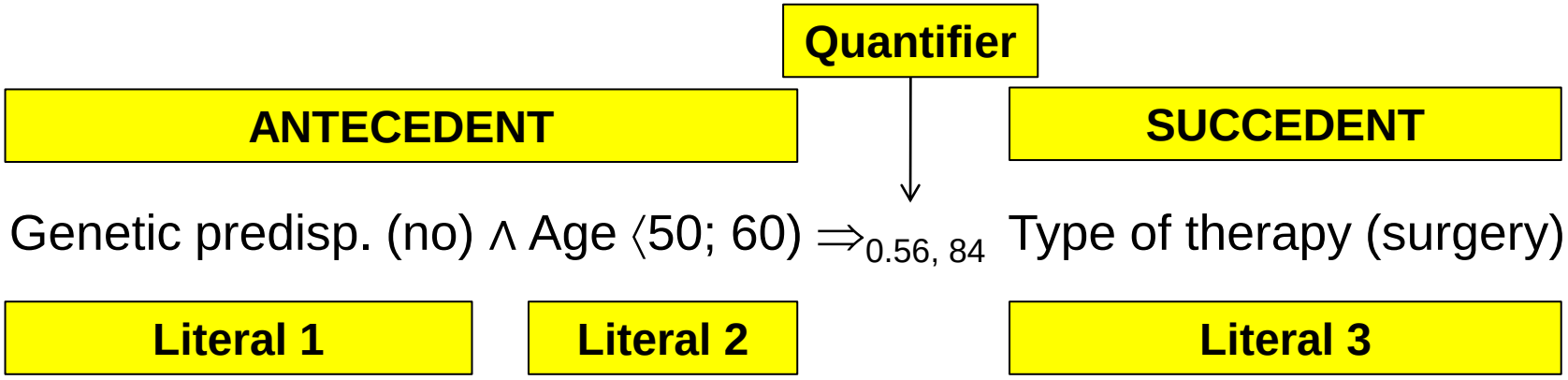
Input in Ac4ft-Miner

- Data matrix
- Definition of a set of relevant rules
  1. Entering a set of relevant antecedents
     a. Antecedent stable part
     b. Antecedent variable part
  2. Entering a set of relevant succedents
     a. Succedent stable part
     b. Succedent variable part
  3. Entering a set of relevant conditions
  4. Entering an Ac4ft-quantifier


Task
BASIC PARAMETERS
Name: Test 1
Comment: -
Group of tasks: Default group of tasks
Data matrix: Adamek_2
Owner: PowerUser
Edit
Take ownership
ANTECEDENT STABLE PART
Antecedent 0 - 99
1a
Total length: 0 - 99
QUANTIFIERS
Type Rel. Value Units
BASE Before >= 1.00 Abs.
BASE After >= 1.00 Abs.
4
SUCCEDENT STABLE PART
Succedent 0 - 99
2a
Total length: 0 - 99
(1) ANTECEDENT VARIABLE PART
1b
Total length: 0 - 99
CONDITION
Condition 0 - 99
3
Total length: 0 - 99
(2) SUCCEDENT VARIABLE PART
2b
Total length: 0 - 99
Task parameters
Sets overlapping: Sets must differ in all rows (i.e. not overlapping sets)
Close
Generate
Switch

Composition of a rule

- Antecedent or succedent or condition = cedent
- Cedent = conjunction or disjunction of literals
- Literal = basic Boolean attribute or negation of basic Boolean attribute



Entering a set of relevant literals

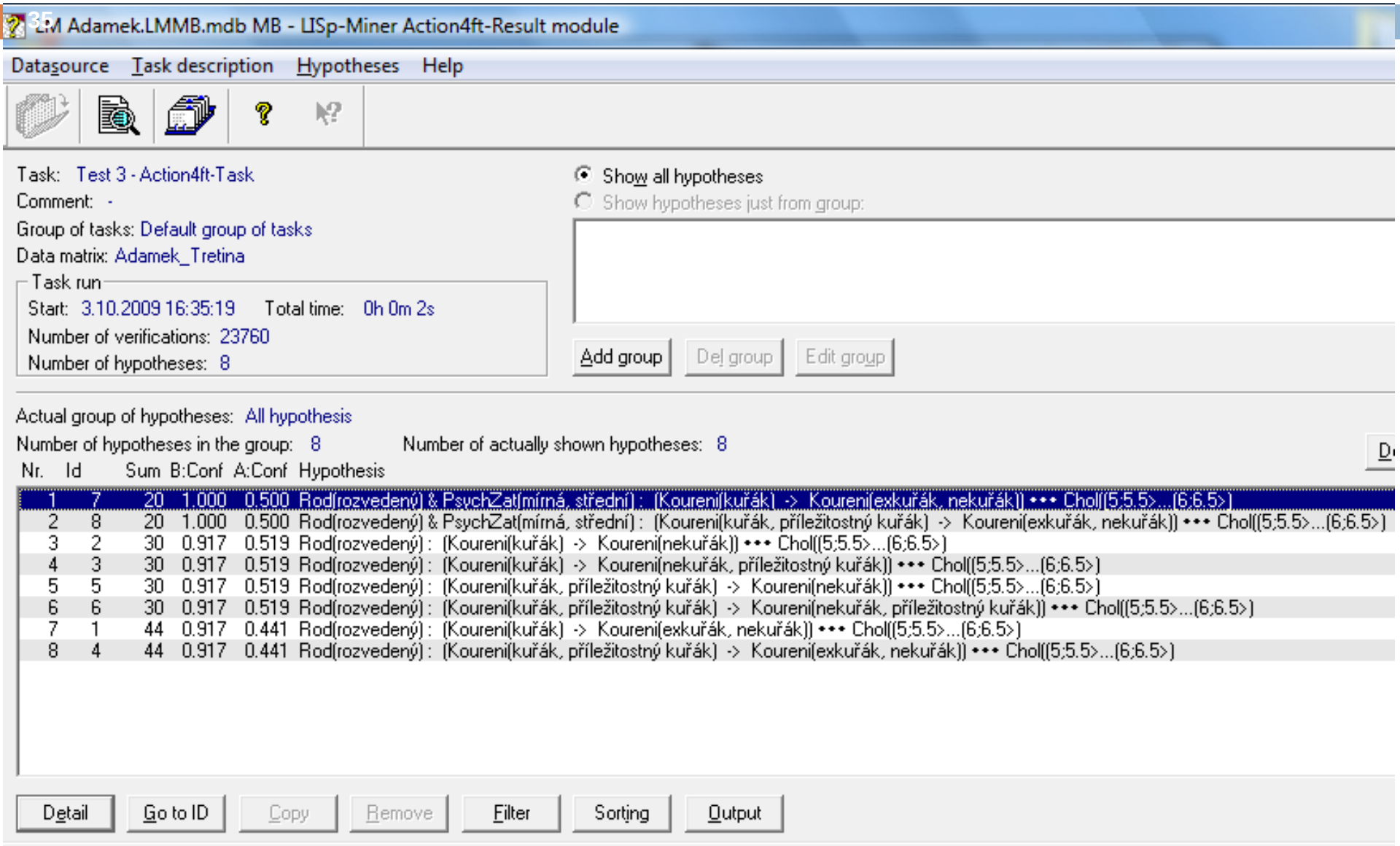
For each attribute we should define:

1. Minimum and maximum length of a literal.
   - Number of categories each literal has
   - Type of therapy (diet, medicaments)…..length = 2
2. The type of coefficient – subsets, intervals, cyclical intervals, left cuts, right cuts, cuts, one particular value
3. One of the following options:
   - Generate only positive literals – no literals with negation are created
   - Generate only negative literals – only literals with negation are created
   - Generate both positive and negative literals

Type of coefficients of literals

- Subsets: Type of therapy {diet, medicaments, surgery, none}, length: min 1, max 2
  - Tot(diet), Tot(medicaments), Tot(surgery), Tot(none), Tot(diet, medicaments), Tot (diet, surgery), Tot (diet, none), Tot (medicaments, surgery), Tot (medicaments, none), Tot(surgery, none)
- Intervals: Age {⟨20; 30), ⟨30; 40), ⟨40; 50), ⟨50; 60), ⟨60; 70)}, length: min 2, max 3
  - Age [⟨20; 30), ⟨30; 40)], Age [⟨30; 40), ⟨40; 50)], Age [⟨40; 50), ⟨50; 60)], Age [⟨50; 60), ⟨60; 70)], Age [⟨20; 30), ⟨30; 40), ⟨40; 50)], Age [⟨30; 40), ⟨40; 50), ⟨50; 60)], Age [⟨40; 50), ⟨50; 60), ⟨60; 70)]
  - Age ⟨20; 40), Age ⟨30; 50), Age ⟨40; 60), Age ⟨50; 70), Age ⟨20; 50), Age ⟨30; 60), Age ⟨40; 70)
- One particular value: Type of therapy {diet, medicaments, surgery, none} containing only surgery:
  - Type of therapy (surgery)
- Cyclical intervals, Left cuts, Right cuts, Cuts

Output in Ac4ft-Miner



Contents

Case study

- Aim
  - testing
  - formulation of analytical questions
  - interpretation (graphics)
- Data
  - Real medical data set ADAMEK
  - Cardiological patients
  - 1395 rows (patients)
  - 37 columns (properties of patients)
- Entering a set of relevant rules – rules containing risk factors of atherosclerosis

Risk factors of atherosclerosis

- Sex – male
- Age over 45 for the male and 55 for the female
- Positive family case history
- Smoking
- Lack of physical activity
- Diabetes mellitus
- Arterial hypertension – blood pressure over 140/90 mm Hg
- Hypercholesterolemia – cholesterol over 5 mmol/l
- HDL cholesterol – less than 1.1 mmol/l (men), less than 1.3 mmol/l (women)
- LDL cholesterol – over 3 mmol/l
- Hypertriglyceridemia – triglycerides over 2 mmol/l
- Overweight – BMI over 25
- Waist circumference – over 94 cm with men and over 80 cm with women
- Blood sugar – over 6 mmol/l
- Uric acid

Sex, age and positive family case history – stable attributes
Other attributes – stable or flexible

Grouping of attributes

Group			Attribute				
No	**Abb.**	**Name**	**No**	**Name**	**Name of column in database**	**Type**	**Categories**
1	OSB (3)	Personal data	1	Age	Vek	R	
			2	Sex	Pohl	N	woman / man
			3	Outpatient department	M	N	Čáslav / Prague
2	SCA (5)	Social history	4	Marital status	Rod	N	divorced / single / widow(er)/ married
			5	Lives alone	Sam	N	yes / no
			6	Education	Vzd	O	primary / secondary / university
			7	Mental load	PsychZat	O	none / low / medium / high
			8	Smoking	Koureni	N	Non-smoker/ ex-smoker/ occasional smoker / smoker
3	AKT (2)	Activities	9	Workload	FyzZat	N	unemployed / none/ low / medium / high
			10	Physical activity	TelAkt	O	none / low / medium / high
4	MRY (7)	Personal measurements	11	Weight	Hmotnost	R	
			12	Height	Vyska	R	
			13	Waist circumference	Pas	R	
			14	Hip circumference	Boky	R	
			15	BMI	BMI	R	
			16	WHR	WHR	R	
			2	Sex	Pohl	N	woman / man

Grouping of attributes (2)

Group			Attribute				
No	**Abb.**	**Name**	**No**	**Name**	**Name of column in database**	**Type**	**Categories**
5	RFK (4)	Risk factors	17	Hyperlipoproteinemia	Hlp	N	yes / no
			18	Diabetes mellitus	DM	N	yes / no
			19	Hypertension	HT	N	yes / no
			20	Positive family case history	Ra_f	N	yes / no
6	OBT (5)	Complaints	21	Breathlessness	Dusnost	O	none / due to activity/ while inactive
			22	Chest pain	Bolesthrud	O	none / due to activity/ while inactive
			23	Palpitation	Palpitace	N	yes / no
			24	Swelling	Otoky	N	yes / no
			25	Claudication	Klaudikace	N	yes / no
7	KTL (2)	Blood pressure	26	Systolic	ST	R	
			28	Diastolic	DT	R	
8	EKG (3)	EKG	29	Pulse rate	TepFrek	R	
			30	PQ_int	PQ_int	R	
			31	QRS_int	QRS_int	R	
9	CHL (4)	Laboratory Cholesterol	32	Total Cholesterol	Chol	R	
			33	HDL Cholesterol	HDL	R	
			34	LDL Cholesterol	LDL	R	
			35	Triacyglycerols	Tgl	R	
10	GKM (2)	Laboratory GKM	36	Blood sugar	Glyk	R	
			37	Uric acid	Kmoc	R	

Three-level formulation of analytical questions

1. „Intuitive formulation“
2. „More formal formulation“
3. Input in Ac4ft-Miner

Antecedent stable part – green

Antecedent variable part – red

Succedent stable part – blue

Condition – orange

„Intuitive formulation“

- For which properties of patients (group 1) does the change of other properties (group 2) cause a change in the incidence of other properties (group 3)?
- For which personal measurements does a change of activities cause a change in the incidence of risk factors?

„More formal formulation“

- For which combinations of properties from group 1 does the relative frequency of combinations of properties from group 3 change by at least q by changing the combinations of properties from group 2.

„More formal formulation“ (2)

- For which combinations of BMI and/or Waist circumference and/or Hip circumference does the relative frequency of Hyperlipoproteinemy and/or Hypertension change by at least 0.3 by changing Physical activity at work and/or Physical activity *(and the number of objects satisfying both initial antecedent and initial succedent is at least 30 and the number of objects satisfying both final antecedent and final succedent is at least 30)*.

„Formal definition“

- *The formal definition consists of the definition of the antecedent stable part (group 1), the antecedent variable part (group 2), the succedent stable part (group 3), and the definition of quantifier - the quantifier Founded implication* $\Rightarrow^{I \lesseqgtr F}_{q, B_I, B_F}$, which is defined as

$$\left| \frac{a_I}{a_I + b_I} - \frac{a_F}{a_F + b_F} \right| \geq q \wedge a_I \geq B_I \wedge a_F \geq B_F$$

where $0 < q \leq 1$, > 0 and > 0.

Analytical tasks

Analytical task	Main idea
1	Too strict formal definition
2	A detailed demonstration of the rule founded and selected
3	Use of condition, all the rules presented in detail, various ways of presenting rules
4	Loose definition but no rules found
5	Loose definition, many insignificant rules found, long duration of the procedure

Analytical tasks

Analytical task	Main idea
1	Too strict formal definition
2	A detailed demonstration of the rule founded and selected
3	Use of condition, all the rules presented in detail, various ways of presenting rules
4	Loose definition but no rules found
5	Loose definition, many insignificant rules found, long duration of the procedure

ANALYTICAL QUESTION 1

Intuitive formulation

For which personal measurements does a change of activities cause a change in the incidence of risk factors?

More formal formulation

For which combinations of BMI and/or Waist circumference and/or Hip circumference does the relative frequency of Hyperlipoproteinemy and/or Diabetes mellitus and/or Hypertension and/or Positive family case history change by at least 0.4 by changing Physical activity at work and/or Physical activity *(and the number of objects satisfying both initial antecedent and initial succedent is at least 40 and the number of objects satisfying both final antecedent and final succedent is at least 40).*

INPUT

	Antecedent	Succedent
Stable part	BMI *(int. per 1, int. max length 6)* Waist *(int. per 5 cm, int. max length 3)* Hip *(int. per 5cm, int. max length 3)*	Hyperlipoproteinemy *(subset max length 1)* Diabetes mellitus *(subset max length 1)* Hypertension *(subset max length 1)* Positive family case history *(subset max length 1)*
Variable part	Physical activity at work *(subset max length 1)* Physical activity *(int. max length 2)*	
Quantifier	$\Rightarrow_{q, B_I, B_F}^{I \lesseqgtr F}$	$\left\lvert \frac{a_I}{a_I + b_I} - \frac{a_F}{a_F + b_F} \right\rvert \geq 0.4 \wedge a_I \geq 40 \wedge a_F \geq 40$

Output analytical task 1

OUTPUT			
Number of hypotheses (rules) found:	0	Number of verifications:	18,951,486
Duration at PC with 2 GHz and 895 MB RAM		0h 34m 45s	
Interpretation			
There are no combinations of BMI and/or Waist circumference and/or Hip circumference for which the relative frequency of Hyperlipoproteinemy and/or Diabetes mellitus and/or Hypertension and/or Positive family case history change by at least 0.4 by changing Physical activity at work and/or Physical activity.			
COMMENT			
Both the absolute difference of the quantifier values, B_I, and also B_F were defined relatively high. This may be due to not finding any rules. It is possible to try to lower q, B_I and B_F and see if it brings any results (see Analytical question 2).			

Analytical tasks

Analytical task	Main idea
1	Too strict formal definition
2	A detailed demonstration of the rule founded and selected
3	Use of condition, all the rules presented in detail, various ways of presenting rules
4	Loose definition but no rules found
5	Loose definition, many insignificant rules found, long duration of the procedure

ANALYTICAL QUESTION 2

Intuitive formulation

For which personal measurements causes a change of activities a change in the incidence of risk factors?

More formal formulation

For which combinations of BMI and/or Waist circumference and/or Hip circumference does the relative frequency of Hyperlipoproteinemy and/or Diabetes mellitus and/or Hypertension and/or Positive family case history change by at least 0.3 by changing Physical activity at work and/or Physical activity *(and the number of objects satisfying both initial antecedent and initial succedent is at least 30 and the number of objects satisfying both final antecedent and final succedent is at least 30).*

INPUT

	Antecedent	Succedent
Stable part	BMI *(int. per 1, int. max length 6)* Waist *(int. per 5 cm, int. max length 3)* Hip *(int. per 5cm, int. max length 3)*	Hyperlipoproteinemy *(subset max length 1)* Diabetes mellitus *(subset max length 1)* Hypertension *(subset max length 1)* Positive family case history *(subset max length 1)*
Variable part	Physical activity at work *(subset max length 1)* Physical activity *(int. max length 2)*	
Quantifier	$\Rightarrow_{q, B_I, B_F}^{I \lesseqgtr F}$	$\left\lvert \frac{a_I}{a_I + b_I} - \frac{a_F}{a_F + b_F} \right\rvert \geq 0.3 \wedge a_I \geq 30 \wedge a_F \geq 30$

OUTPUT

Number of hypotheses (rules) found:	80	Number of verifications:	46,648,080
Duration at PC with 2 GHz and 895 MB RAM:	1h 21m 54s		

One of the founded rules:

Waist $\langle 70; 85\rangle \wedge$ Hip $\langle 95; 110\rangle \wedge$ Physical activity at work (low $\rightarrow$ unemployed) $\wedge$ Physical activity (low, medium $\rightarrow$ none, low) $\Rightarrow_{0.302,43,31}$ Hypertension (no)

Initial rule

Waist $\langle 70; 85\rangle \wedge$ Hip $\langle 95; 110\rangle \wedge$ Physical activity at work (low) $\wedge$ Physical activity (low, medium) $\Rightarrow_{0.935,43}$ Hypertension (no)

	Hypertension (no)	¬ Hypertension (no)
Waist $\langle 70; 85\rangle \wedge$ Hip $\langle 95; 110\rangle \wedge$ Physical activity at work (low) $\wedge$ Physical activity (low, medium)	43	3
¬ (Waist $\langle 70; 85\rangle \wedge$ Hip $\langle 95; 110\rangle \wedge$ Physical activity at work (low) $\wedge$ Physical activity (low, medium))	908	441

Confidence $= \frac{a}{a+b} = \frac{43}{43+3} = 0.935$

Interpretation of initial rule

There are 43 patients with waist circumference between 70 and 85 centimetres, hip circumference between 95 and 110 centimetres, with low physical activity at work and low or medium physical activity, who represent 93.5 % of all patients with waist circumference between 70 and 85 centimetres, Hip circumference between 95 and 110 centimetres, with low physical activity at work and low or medium physical activity, who do not have hypertension.

Final rule		
Waist⟨70; 85⟩ ∧ Hip ⟨95; 110⟩ ∧ Physical activity at work (unemployed) ∧ Physical activity (none, low) $\Rightarrow_{0.633,31}$ Hypertension (no)		
	Hypertension (no)	¬ Hypertension (no)
Waist⟨70; 85⟩ ∧ Hip ⟨95; 110⟩ ∧ Physical activity at work (unemployed) ∧ Physical activity (none, low)	31	18
¬ (Waist⟨70; 85⟩ ∧ Hip ⟨95; 110⟩ ∧ Physical activity at work (unemployed) ∧ Physical activity (none, low))	920	426
Confidence = $\frac{a}{a+b} = \frac{31}{31+18} = 0.633$		

Interpretation of final rule

There are 31 patients with waist circumference between 70 and 85 centimetres, hip circumference between 95 and 110 centimetres, unemployed, with none or low physical activity, who represent 63.3 % of all patients with waist circumference between 70 and 85 centimetres, hip circumference between 95 and 110 centimetres, unemployed, with none or low physical activity, who do not have hypertension.

Interpretation of the whole action rule

If the physical activity at work is changed from low to unemployed and physical activity from low or medium to none or medium among the patients with waist circumference between 70 and 85 centimetres and hip circumference between 95 and 110 centimetres, the incidence of patients not having hypertension decreases by 30.2 percentage points.

Analytical tasks

Analytical task	Main idea
1	Too strict formal definition
2	A detailed demonstration of the rule founded and selected
3	Use of condition, all the rules presented in detail, various ways of presenting rules
4	Loose definition but no rules found
5	Loose definition, many insignificant rules found, long duration of the procedure

ANALYTICAL QUESTION 3

Intuitive formulation

For which personal measurements causes a change of activities a change in the incidence of risk factors in male patients?

More formal formulation

In case of male patients, for which combinations of BMI and/or Waist circumference and/or Hip circumference does the relative frequency of Hyperlipoproteinemy and/or Hypertension change by at least 0.3 by changing Physical activity at work and/or Physical activity *(and the number of objects satisfying both initial antecedent and initial succedent is at least 30 and the number of objects satisfying both final antecedent and final succedent is at least 30).*

INPUT

	Antecedent	Succedent
Stable part	BMI *(int. per 1, int. max length 6)* Waist *(int. per 5 cm, int. max length 3)* Hip *(int. per 5cm, int. max length 3)*	Hyperlipoproteinemy *(subset max length 1)* Hypertension *(subset max length 1)*
Variable part	Physical activity at work *(subset max length 1)* Physical activity *(int. max length 2)*	
Condition	Sex *(one category-(male))*	
Quantifier	$\Rightarrow^{I \lesseqgtr F}_{q, B_I, B_F}$	$\left\lvert \frac{a_I}{a_I + b_I} - \frac{a_F}{a_F + b_F} \right\rvert \geq 0.3 \wedge a_I \geq 30 \wedge a_F \geq 30$

Output analytical task 3

OUTPUT			
Number of hypotheses (rules) found:	6	Number of verifications:	598,064
Duration at PC with 2 GHz and 895 MB RAM:		0h 1m 14s	
Note: In this analytical question, different approaches of presenting founded rules are introduced. All the rules founded are presented. The text in orange indicates the condition.			

ACTION RULE 1
BMI (20; 26⟩ & Hip ⟨90; 105) ∗ Physical activity at work (low) &Physical activity (low, medium) → Physical activity (none) Hyperlipoproteinemy (no) 86.8% → Hyperlipoproteinemy (no) 54.4% / Sex (male)
Interpretation of the action rule
In case of male patients with BMI between 20 and 26 and with hip circumference between 90 cm and 105 cm, **IF** we change low physical activity at work and low or medium physical activity **TO** none physical activity, probability of not having hyperlipoproteinemy decreases **FROM** 86.8 % **TO** 54.4 %.

ACTION RULE 2

BMI (23; 28⟩ ∧ Hip ⟨75; 90) ∧ Physical activity (medium → none) $\Rightarrow_{0.311,30,30}$ Hyperlipoproteinemy (no) / Sex (male)

Interpretation of the action rule

In case of male patients with BMI between 23 and 28 and with hip circumference between 75 cm and 90 cm, **IF** we change physical activity **FROM** medium **TO** none, the incidence of not having hyperlipoproteinemy decreases **BY** 31.1 percentage points.

Initial rule

BMI (23; 28⟩ ∧ Hip ⟨75; 90) ∧ Physical activity (medium) $\Rightarrow_{0.811,30,30}$ Hyperlipoproteinemy (no) / Sex (male)

	Hyperlipoproteinemy (no)	**NOT** Hyperlipoproteinemy (no)
BMI (23; 28⟩ ∧ Hip ⟨75; 90) ∧ Physical activity (medium)	30	7
NOT **[** BMI (23; 28⟩ ∧ Hip ⟨75; 90) ∧ Physical activity (medium) **]**	305	241

Interpretation of initial rule

There are 30 male patients with BMI between 23 and 28, hip circumference between 75 and 90 centimetres with medium physical activity, who represent 81.1 % of all male patients with BMI between 23 and 28, hip circumference between 75 and 90 centimetres with medium physical activity,

Final rule

BMI (23; 28⟩ ∧ Hip ⟨75; 90) ∧ Physical activity (none) $\Rightarrow_{0.500,30}$ Hyperlipoproteinemy (no) / Sex (male)

	Hyperlipoproteinemy (no)	**NOT** Hyperlipoproteinemy (no)

Output analytical task 3, rule 3

ACTION RULE 3
BMI (22; 28⟩ ∧ Hip (75; 90⟩ ∧ Physical activity (medium → none) $\Rightarrow_{0.300,32,32}$ Hyperlipoproteinemy (no) / Sex (male)
Interpretation of the action rule:
In case of male patients with BMI between 22 and 28 and with hip circumference between 75 cm and 90 cm, **IF** we change physical activity **FROM** medium **TO** none, the incidence of not having hyperlipoproteinemy decreases **BY** 30 percentage points.

ACTION RULE 4		
	IN PATIENTS WITH	
Antecedent stable part	**BMI (20; 26) & Hip ⟨90; 105)**	
	WHO ARE	
Condition	**Sex (male)**	
	IF WE CHANGE	
Ant. variable part - from	**Physical activity (none)**	
Proposed change	TO	
Ant .variable part - to	**Physical activity at work (low) & Physical activity (low, medium)**	
	FOLLOWING PROPERTY	
Succedent stable part	**Hyperlipoproteinemy (no)**	
	WILL INCREASE IN ITS INCIDENCE BY	
Difference of quantifier values	**0.324**	
	MULTIPLE 100 PERCENTAGE POINTS	

ACTION RULE 5
BMI (23; 28⟩ ∧ Hip (75; 90⟩ ∧ Physical activity (none → medium) $\Rightarrow_{0.311,30,30}$ Hyperlipoproteinemy (no) / Sex (male)
Interpretation of the action rule:
In case of male patients with BMI between 23 and 28 and with hip circumference between 75 cm and 90 cm, **IF** we change physical activity **FROM** none **TO** medium, the incidence of not having hyperlipoproteinemy increases **BY** 31.1 percentage points.

ACTION RULE 6
BMI (22; 28⟩ ∧ Hip (75; 90⟩ ∧ Physical activity (none → medium) $\Rightarrow_{0.300,32,32}$ Hyperlipoproteinemy (no) / Sex (male)
Interpretation of the action rule:
In case of male patients with BMI between 22 and 28 and with hip circumference between 75 cm and 90 cm, **IF** we change physical activity **FROM** none **TO** medium, the incidence of not having hyperlipoproteinemy increases **BY** 30 percentage points.

COMMENT
In this analytical question, the condition was used. Only the columns containing the attribute Sex (male) were included in the four-fold tables.
The "opposite" rules were founded. Rules 1 and 4; 2 and 5; 3 and 6 are "opposite". One rule from the pair suggests an action from the state X to the state Y while the succedent decreases in its incidence; the other rule from the pair suggests an action from the state Y to the state X while the succedent increases in its incidence. Both antecedent and succedent of the "opposite" rules are the same.

Analytical tasks

Analytical task	Main idea
1	Too strict formal definition
2	A detailed demonstration of the rule founded and selected
3	Use of condition, all the rules presented in detail, various ways of presenting rules
4	Loose definition but no rules found
5	Loose definition, many insignificant rules found, long duration of the procedure

ANALYTICAL QUESTION 4

Intuitive formulation

For which age groups and/or which sex and/or which weight of patients does a change of BMI cause the cholesterol to change?

More formal formulation

For which combinations of Age and/or Sex and/or Weight does the relative frequency of Cholesterol change by at least 0.2 by changing BMI *(and the number of objects satisfying both initial antecedent and initial succedent is at least 10 and the number of objects satisfying both final antecedent and final succedent is at least 10)*.

INPUT

	Antecedent	Succedent
Stable part	Age *(int. per 5 years, int. max length 4)* Sex *(subset max length 1)* Weight *(int. per 5kg, int. max length 4)*	Total cholesterol *(int. per 0.5, int. max length 3)*
Variable part	BMI *(int. per 1, int. max length 6)*	
Quantifier	$\Rightarrow_{q, B_I, B_F}^{I \leq F}$	$\left\lvert \frac{a_I}{a_I + b_I} - \frac{a_F}{a_F + b_F} \right\rvert \geq 0.2 \wedge a_I \geq 10 \wedge a_F \geq 10$

Output analytical task 4

OUTPUT			
Number of hypotheses (rules) found:	0	Number of verifications:	0
Duration at PC with 2 GHz and 895 MB RAM		7h 33m 3s	
Interpretation			
There are no combinations of Age and/or Sex and/or Weight for which the relative frequency of Cholesterol changes at least by 0.2 as a consequence of changing BMI.			
COMMENT			
Although the q, B_I and B_F were defined relatively low, there were no rules found. The process of finding rules was relatively long due to many categories of attributes (BMI has 18 categories, Age 11 categories, Weight 12 categories, Cholesterol 9 categories) and also due to the maximum length of intervals of each attribute. There were, therefore, many possible combinations which have to be created and checked.			

Analytical tasks

Analytical task	Main idea
1	Too strict formal definition
2	A detailed demonstration of the rule founded and selected
3	Use of condition, all the rules presented in detail, various ways of presenting rules
4	Loose definition but no rules found
5	Loose definition, many insignificant rules found, long duration of the procedure

ANALYTICAL QUESTION 5

Intuitive formulation

For which risk factors does a change in cholesterol cause a change in the incidence of complaints?

More formal formulation

For which combinations of Hyperlipoproteinemy and/or Diabetes mellitus and/or Hypertension and/or Positive family case history does the relative frequency of Breathlessness and/or Claudication and/or Palpitation change by at least 0.2 by changing Total cholesterol and/or HDL cholesterol and/or LDL cholesterol and/or Triacylglycerols *(and the number of objects satisfying both initial antecedent and initial succedent is at least 10 and the number of objects satisfying both final antecedent and final succedent is at least 10)*.

INPUT

	Antecedent	Succedent
Stable part	Hyperlipoproteinemy *(one category-(yes))* Diabetes mellitus *(one category-(yes))* Hypertension *(one category-(yes))* Positive family case history *(one category-(yes))*	Breathlessness *(one category-(due to the activity))* Breathlessness *(one category-(while inactive))* Claudication *(one category-(yes))* Palpitation *(one category-(yes))*
Variable part	Total cholesterol *(int. per 0.5, int. max length 2)* HDL cholesterol *(int. per 0.5, int. max length 2)* LDL cholesterol *(int. per 0.5, int. max length 2)* Triacylglycerols *(int. per 0.5, int. max length 2)*	
Quantifier	$\Rightarrow_{q, B_I, B_F}^{I \lesssim F}$	$\left\lvert \frac{a_I}{a_I + b_I} - \frac{a_F}{a_F + b_F} \right\rvert \geq 0.2 \wedge a_I \geq 10 \wedge a_F \geq 10$

Output analytical task 5

OUTPUT			
Number of hypotheses (rules) found:	437	Number of verifications:	742,569,534
Duration at PC with 2 GHz and 895 MB RAM:		22h 50m 31s Interrupted	
Interpretation:			
There were 437 rules found. They are, however, not very significant – the highest confidence of initial rule is only 0.318.			
COMMENT			
In the input of the antecedent stable part and succedent stable part, in the definition of literals the option "one category" was used. This is because we want to find rules containing risk factors and complaints (we do not want to find rules containing, for example, Palpitation (no)). The generation and verification was interrupted after nearly 23 hours. This example shows that by defining the q, B_I and B_2 too low, we can obtain insignificant rules (they are true for very low number of patients), and it also demonstrates that the run of the Ac4ft-Miner is very long in this case.			

Contents

Conclusions

- Ac4ft-Miner is a very complex tool
  - Many possibilities of defining input
- Problems with defining analytical questions
  - It proved useful to group attributes
  - Defining too specific questions – usually no results
  - Defining too general questions – too many rules found

Future work

- Create sophisticated methodology of use
  - How to formulate analytical questions?
  - How to interpret results?
  - How to tune parameters of the procedure?
- Test more Ac4ft-quantifiers

Bibliography

- Nekvapil, V. 2010. *Data Mining in the Medical Domain: Using the Ac4ft-Miner Procedure*. LAP LAMBERT Academic Publishing GmbH & Co. KG, Saarbrucken, 2010. ISBN 978-3-8383-9818-1
- Ras, Z.; Wieczorkowska, A. 2000. *Action-Rules: How to Increase Profit of a Com*pany In: D.A. Zighed, J. Komorowski, and J. Zytkow (Eds.): PKDD 2000, LNAI 1910, Berlin: Springer Verlag, pp. 587-592.
- Rauch, J.; Šimůnek, M. 2005. GUHA Method and Granular Computing. In: HU, X et al. (eds.). Proceedings of IEEE conference Granular Computing, Beijing: IEEE Computer Society.
- Rauch, J.; Šimůnek, M. 2009. *Action Rules and the GUHA Method: Preliminary Considerations and Results*. Praha 14.09.2009 – 17.09.2009. In: Foundations of Intelligent Systems. Berlin: Springer Verlag, pp. 76–87.
- Rauch, J.; Šimůnek, M. c2011. GUHA Method and the LISp-Miner System [PowerPoint presentation]. [cited 2011 April 20]