Tato prezentace je součástí wiki-prezentace Metoda GUHA, LISp-Miner a typové úlohy
Je dostupná z této adresy
Verse 17. 8. 2019

Typ úlohy: Podmíněná asociační pravidla, vysoká konfidence
Data: Hotel
Problém: Vyplývají z místa bydliště hosta nějaké typické parametry pobytu, případně i počasí? A to obecně i zvlášť pro rekreační a služební pobyty?

Jan Rauch
Katedra informačního a znalostního inženýrství
Vysoká škola ekonomická v Praze

© Jan Rauch

Asociační pravidla – příklad

Hotel.txt

HVek	HPohlavi	HMesto	HMesto_X	HMesto_Y	HStat	PPobytOd	PNoci	POsob	PTypPobytu	PCenaUbytovani	PCenaStrava	PCenaSleva	PCenaCelkem	DHodnoceni
21	žena	České Budějovice	14.4757883	48.9763169	ČR	31.5.2013	1	1	rekreační	1450.00	0.000	0.00	1450.00	spokojen 91 82 81 56
34	muž	Linec	14.2862742	48.3066489	Rakousko	2.8.2013	2	4	rekreační	11600.00	1440.000	200.00	12840.00	průměr 44 21 62 84
30	mu			48.2115631	Rakousko	5.6.2012	7	2	rekreační	16940.00	2100.000	200.00	18840.00	nespokojen 5 37 25 71
62	mu			283	50.2592108	Po								
35	že			48.2115631	Rakou									13 47
58	muž	České Budějovice		14.4757883	48.9763									
81	žena	Vídeň	16.3736767	48.2115631	Rakou									
22	žena	Drážďany	13.7397044	51.0497456										14 17
82	muž	Katovice	19.0241283	50.2592108	Po									
55	muž	Praha	14.4212806	50.0874967	ČR	14.12.2012	2	2	rekreační	4840.00	600.000	200.00	5240.00	spokojen 89 81 91 53
75	žena	Berlín	13.3908886	52.5176189	Německo	26.5.2012	1	4	rekreační	4840.00	0.000	200.00	4640.00	nespokojen 30 39 32 3
66	žena	Linec	14.2862742	48.3066489	Rakousko	23.2.2012	1	2	služební	2420.00	0.000	0.00	2420.00	spokojen 95 93 75 21
64	žena	Linec	14.2862742	48.3066489	Rakousko	16.3								52 30 59 52
35	muž	Košice	21.2543528	48.7160408	Slovensko	6.4.								52 56 36
32	muž	Mnichov	11.5836375	48.1364669	Německo	13.8.2								58 70 46
65	muž	Plzeň	13.3771556	49.7490406	ČR	13.3.2013	1							72
79	muž	Brno	16.6153758	49.1921808	ČR	3.5.2012	1	3	rekreační	3630.00	0.000	0.00	3630.00	průměr 59 43 48 59
28	žena	Drážďany	13.7397044	51.0497456	Německo	24.12.2012	1	2	služební	2420.00	0.000	0.00	2420.00	průměr 51 66 58 48
35	žena	Hamburg	10.0043528	53.5498325	Německo	5.1.2013	14	2	rekreační	40600.00	5040.000	200.00	45440.00	nespokojen 27 15 30 12
22	muž	Plzeň	13.3771556	49.7490406	ČR	9.11.2013	4	1	rekreační	5800.00	0.000	200.00	5600.00	spokojen 84 79 86 48
25	muž	Karlovy Vary	12.8690381	50.2311075	ČR	9.11.2013	7	2	rekreační	20300.00	0.000	600.00	19700.00	spokojen 88 74 94 54
20	žena	Hamburg	10.0043528	53.5498325	Německo	19.1.2013	14	4	rekreační	81200.00	10080.000	200.00	91080.00	průměr 59 23 46 96
30	žena	Linec 14				9.11.2013	2	4	rekreační	11600.00	0.000	200.00	11400.00	spokojen 88 82 84 61
45	žena	Karlovy				28.12.2012	2	2	rekreační	4840.00	600.000	200.00	5240.00	spokojen 81 99 87 23

Vyplývají z místa bydliště hosta nějaké typické parametry pobytu, případně i počasí?

Bydliště $\Rightarrow^{?}$ Pobyt, Meteo

Meteo.txt

MDatum	MTeplota	MObloha
4.1.2012	-6.3	slunečno
5.1.2012	-6.6	zataženo
6.1.2012	6.1	srážky
7.1.2012	1.6	srážky
8.1.2012	-1.3	srážky
	7	zataženo
		srážky
		zataženo
12.1.2012	-3.1	srážky
13.1.2012	-8.1	zataženo
14.1.2012	-10.7	srážky
15.1.2012	-5.5	zataženo
16.1.2012	2.3	zataženo
17.1.2012	-1.9	zataženo
18.1.2012	-8.6	zataženo

Bydliště $\Rightarrow^{?}$ Pobyt, Meteo

Groups of attributes tree

- Root group of attrib
  - Dotazník
  - Host
    - Bydliště
  - Meteo
  - Pobyt
    - Cena
    - Začátek
  - Směnárna

Attribute	Used	DBColumn	Categories	XCat	Sample categories
HCizinec_b	+	HStat	2		ne, ano
HMesto	+	HMesto	28		Berlín, Bratislava, Brno, České Budějovice, D
HMesto_m_hlavní		HMesto	5		Berlín, Bratislava, Praha, Varšava, Vídeň
HStat	+	HStat	5		ČR, Německo, Polsko, Rakousko, Slovensko
HStat_m_bezČR		HStat	4		Německo, Polsko, Rakousko, Slovensko

Groups of attributes tree

- Root group of attrib
  - Dotazník
  - Host
    - Bydliště
  - Meteo
  - Pobyt
    - Cena
    - Začátek
  - Směnárna

Attribute	Used	DBColumn	Categories	XCat	Sample categories
PNoci_enum_m	+	PNoci	10		1, 2, <3;6>, 7, <8;13>, 14, <15;20>, 21,
PNoci_exp	+	PNoci	5		1, 2, 7/14/21, 28, ostatni
POsob		POsob	4		1, 2, 3, 4
POsobonoci_ef5		POsobonoci	5		nejnižší, nižší, průměr, vyšší, nejvyšší
PPresSobotniNoc		PPresSobotniNoc	2		ne, ano
PTurnus		PTurnus	2		ne, ano
PTypPobytu		PTypPobytu	2		rekreační, služební

Groups of attributes tree

- Root group of attrib
  - Dotazník
  - Host
    - Bydliště
  - Meteo
  - Pobyt
    - Cena
    - Začátek
  - Směnárna

Attribute	Used	DBColumn	Categories	XCat	Sample categories
MObloha	+	MObloha	3		slunečno, srážky, zataženo
MTeplota_ed5		MTeplota	10		<-17.5;-12.5), <-12.5;-7.5), <-7.5;-2.5),
MTeplota_exp		MTeplota	5		extrémní mrazy, zima, neutrální, teplo, extr

Bydliště $\approx^?$ Pobyt, Meteo - zadání pro 4ft-Miner

LM Hotel MB - LISp-Miner Workspace module - 25.33.01

File Data Introduction Preprocessing Interactive Analysis Data-mining Tasks Domain Knowledge Window Help

032: Bydliště + Host.. | 032: Bydliště + Host.. | Hypothesis (271) | 030: Bydliště => Pob.. | 030: Bydliště => Pob..

Data-mining Task basic parameters

Name: 030: Bydliště => Pobyt, Meteo — ID: 2

Comment: -

Taskgroup: 03: Typické pobyty podle bydliště hosta

Task type: 4ft-Miner — Data matrix: HotelPlusExterni — Edit

ANTECEDENT

Host/Bydliště	Con, 1 - 1
» HCizinec_b (subset), 1 - 1	B, pos
» HMesto (subset), 1 - 1	B, pos
» HStat (subset), 1 - 1	B, pos

Total length: 1

QUANTIFIERS

PIM p= 0.800
BASE p= 50 Abs.

$\Rightarrow ?$

Generation information

Status: Solved, 4 run(s)
Mode: Standard

SUCCEDENT

Pobyt	Con, 1 - 4
» PNoci_enum_m (seq), 1 - 2	B, pos
» PDenTydne (subset), 1 - 1	B, pos
» POsob (seq), 1 - 3	B, pos
» POsobonoci_ef5 (seq), 1 - 2	B, pos
Meteo	Con, 0 - 2
» MObloha (subset), 1 - 1	B, pos
» MTeplota_exp (subset), 1 - 1	B, pos

Total length: 1 - 5

Bydliště(*)

Pobyt(*) $\wedge$ Meteo(*)

Task parameters

Handling of missing values: Ignore X-categories
Prime rule test for implications enabled: No
Include succedent extensions of 100% implications: Yes
Include extensions of coefficients with no change in the four-fold table: Yes
Include extensions of cedents with no change in the four-fold table: Yes
Include 'worse' antecedent extensions (for implications and AAD/BAD): Yes
Include both symetric hypotheses: Yes — Extensions minimal length check: Yes
Maximal number of hypotheses: 1000

CONDITION

Default	Con, 0 - 0

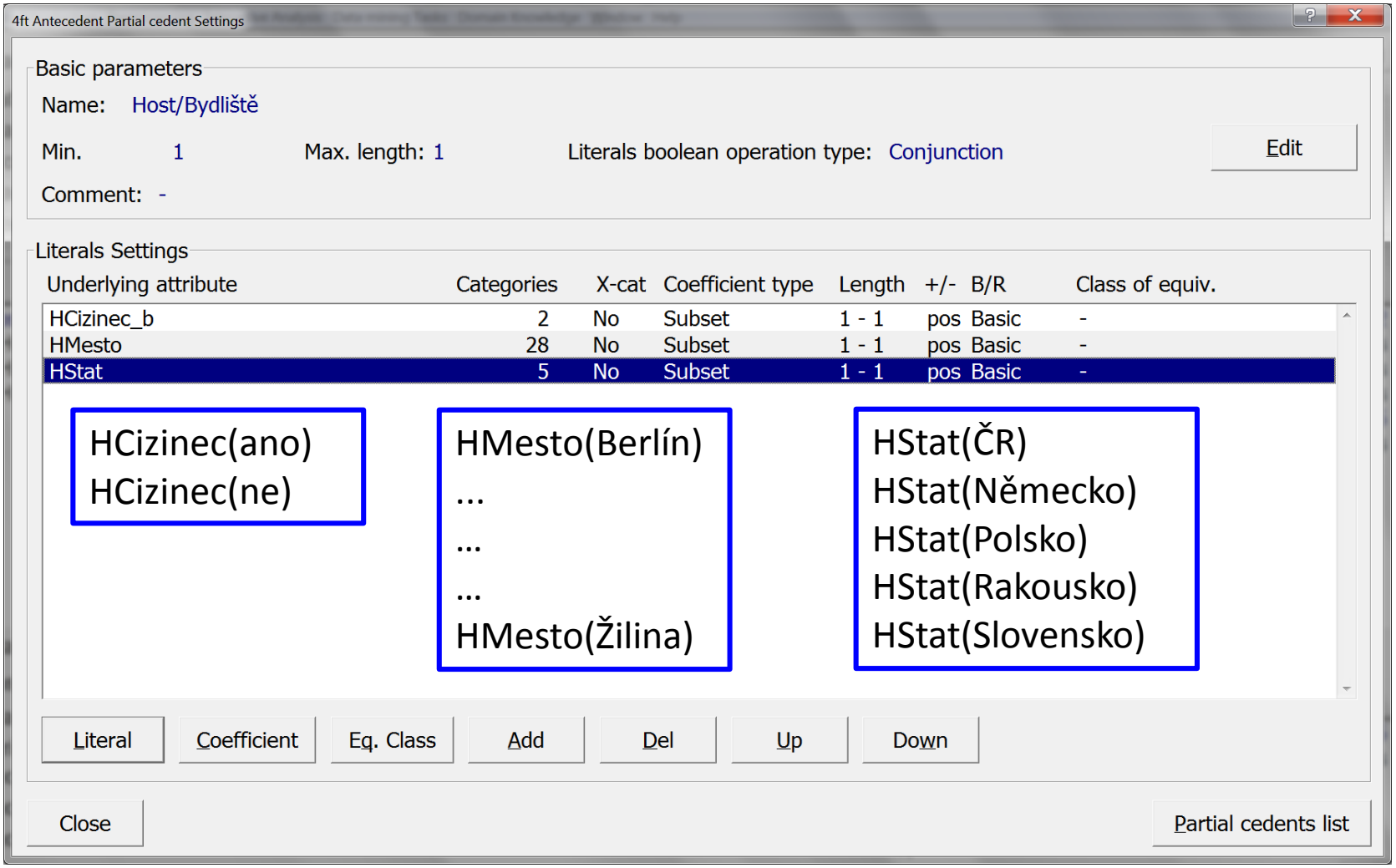
Params | Switch | Validate | Task Clone

Run | Bkgrnd Run | Grid Run | Show Results

Jedno z mnoha možných zadání pro řešení dané analytické otázky!!

Ready

Bydliště(*)


4ft Antecedent Partial cedent Settings
Basic parameters
Name: Host/Bydliště
Min. 1
Max. length: 1
Literals boolean operation type: Conjunction
Edit
Comment: -
Literals Settings
Underlying attribute Categories X-cat Coefficient type Length +/- B/R Class of equiv.
HCizinec_b 2 No Subset 1 - 1 pos Basic -
HMesto 28 No Subset 1 - 1 pos Basic -
HStat 5 No Subset 1 - 1 pos Basic -
HCizinec(ano)
HCizinec(ne)
HMesto(Berlín)
...
...
...
HMesto(Žilina)
HStat(ČR)
HStat(Německo)
HStat(Polsko)
HStat(Rakousko)
HStat(Slovensko)
Literal
Coefficient
Eq. Class
Add
Del
Up
Down
Close
Partial cedents list

Pobyt(*)

4ft Succedent Partial cedent Settings

Basic parameters

Name: Pobyt

Min. length: 1 Max. length: 4 Literals boolean operation type: Conjunction

Comment: -

Edit

Literals Settings

Underlying attribute	Categories	X-cat	Coefficient type	Length	+/-	B/R	Class of equiv.
PNoci_enum_m	10	No	Sequences	1 - 2	pos	Basic	-
PDenTydne	7	No	Subsets	1 - 1	pos	Basic	-
POsob	4	No	Sequences	1 - 3	pos	Basic	-
POsobonoci_ef5	5	No	Sequences	1 - 2	pos	Basic	-

PNoci_enum_m(*)
PDenTydne(*)
POsob(*)
POsobonoci_ef5(*)

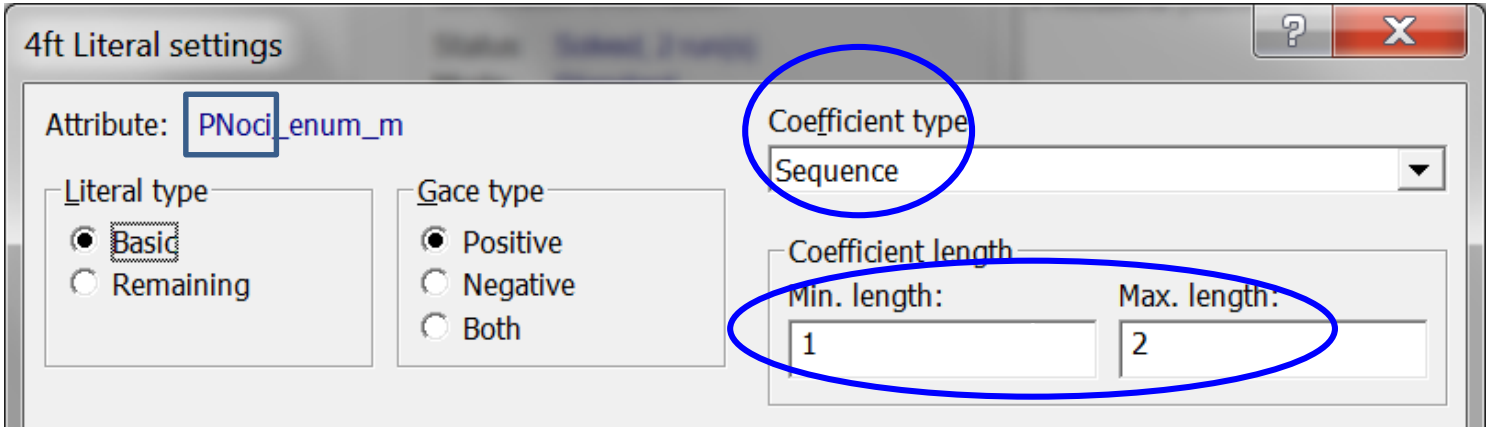
PNoci_enum_m(*) ∧ PDenTydne(*) ∧ POsob(*)
PNoci_enum_m(*) ∧ PDenTydne(*) ∧ Posobonoci_ef5 (*)
PNoci_enum_m(*)∧ POsob(*) ∧ Posobonoci_ef5(*)
PDenTydne ∧ POsob(*)∧ POsobonoci_ef5(*)

PNoci_enum_m(*) ∧ PDenTydne(*) ∧ POsob(*) ∧ Posobonoci (*)

PNoci_enum_m(*) ∧ PDenTydne(*)
PNoci_enum_m(*) ∧ POsob(*)
PNoci_enum_m(*) ∧ POsobonoci_ef5(*)
PDenTydne ∧ POsob(*)
PDenTydne ∧ POsobonoci_ef5(*)
POsob(*) ∧ POsobonoci_ef5(*)

Del Up Down

PNoci_enum_m(*)

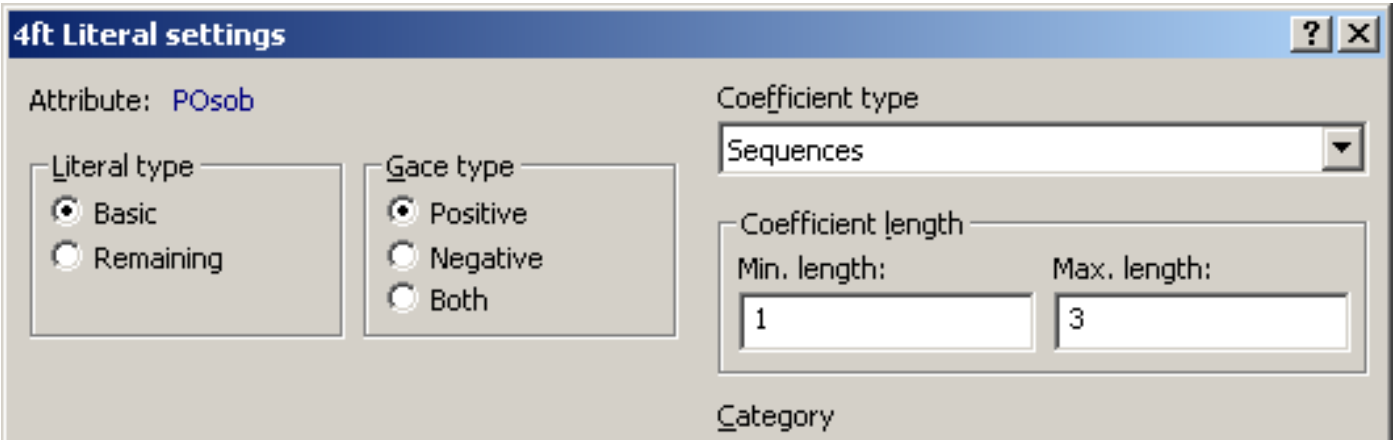


10 kategorií : 1, 2, ⟨3;6⟩, 7, ⟨8;13⟩, 14, ⟨15;20⟩, 21, ⟨22;27⟩, 28

PNoci(1)	●	●	●	●	●	●	●	●	●	●
PNoci(2)	●	●	●	●	●	●	●	●	●	●
Pnoci(⟨3;6⟩)	●	●	●	●	●	●	●	●	●	●
PNoci(7)	●	●	●	●	●	●	●	●	●	●
PNoci(⟨8;13⟩)	●	●	●	●	●	●	●	●	●	●
PNoci(14)	●	●	●	●	●	●	●	●	●	●
PNoci(⟨15;20⟩)	●	●	●	●	●	●	●	●	●	●
PNoci(21)	●	●	●	●	●	●	●	●	●	●
PNoci(⟨22;27⟩)	●	●	●	●	●	●	●	●	●	●
PNoci(28)	●	●	●	●	●	●	●	●	●	●

PNoci(1,2)	●	●	●	●	●	●	●	●	●	●
PNoci(2,⟨3;6⟩)	●	●	●	●	●	●	●	●	●	●
PNoci(⟨3;6⟩,7)	●	●	●	●	●	●	●	●	●	●
PNoci(7,⟨8;13⟩)	●	●	●	●	●	●	●	●	●	●
PNoci(⟨8;13⟩,14)	●	●	●	●	●	●	●	●	●	●
PNoci(14,⟨15;20⟩)	●	●	●	●	●	●	●	●	●	●
PNoci(⟨15;20⟩,21)	●	●	●	●	●	●	●	●	●	●
PNoci(21, ⟨22;27⟩)	●	●	●	●	●	●	●	●	●	●
PNoci(⟨22;27⟩,28)	●	●	●	●	●	●	●	●	●	●

POsob(*)



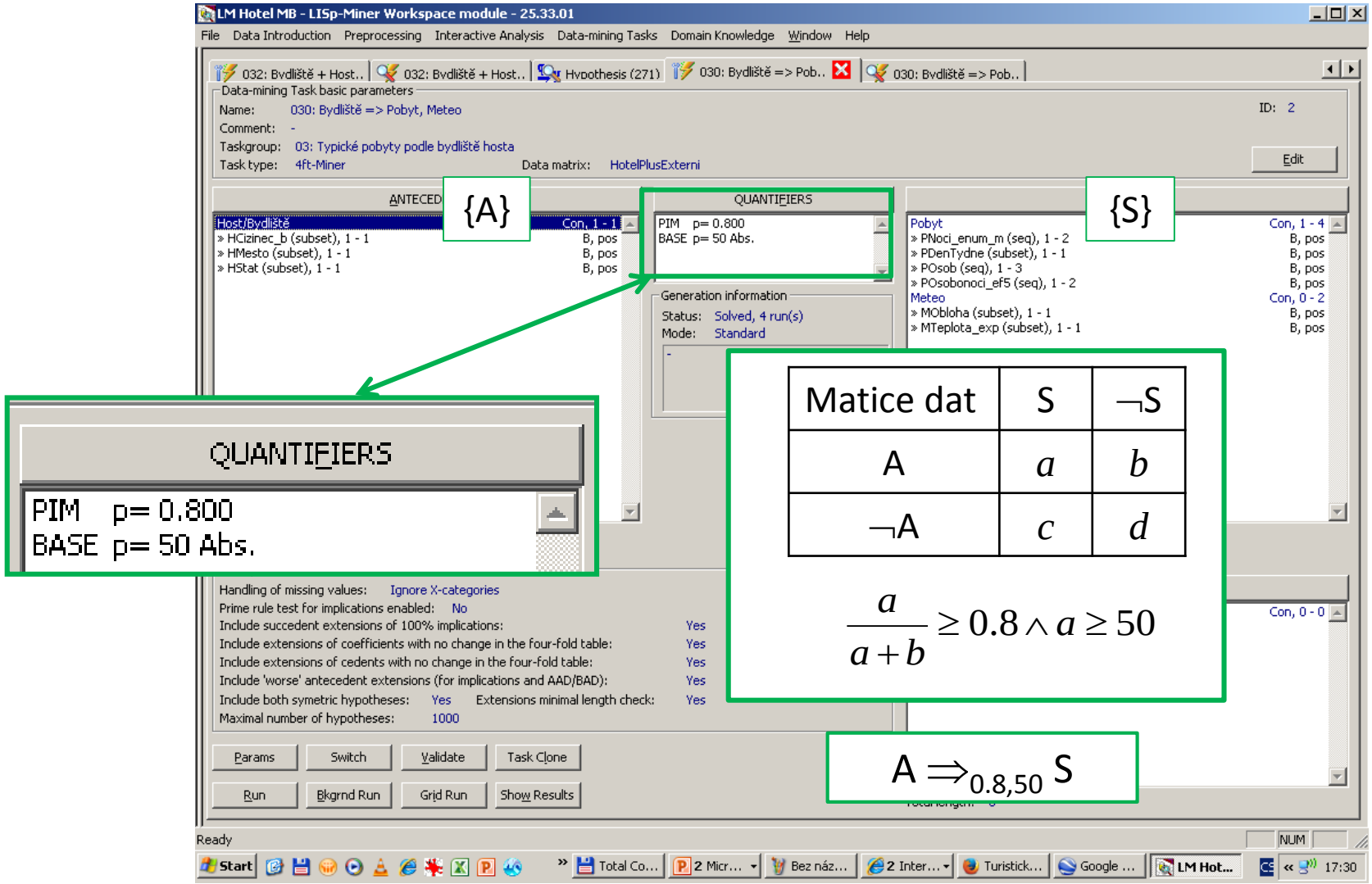
4 kategorie : 1, 2, 3, 4

POsob(1) ● ● ● ●
POsob(2) ● ● ● ●
POsob(3) ● ● ● ●
POsob(4) ● ● ● ●

POsob(1,2) ● ● ● ●
POsob(2,3) ● ● ● ●
POsob(3,4) ● ● ● ●

POsob(1,2,3) ● ● ● ●
POsob(2,3,4) ● ● ● ●

Bydliště $\Rightarrow^{?}$ Pobyt, Meteo - zadání pro 4ft-Miner



Matice dat	S	¬S
A	a	b
¬A	c	d

$$\frac{a}{a+b} \geq 0.8 \wedge a \geq 50$$

$$A \Rightarrow_{0.8,50} S$$

Bydliště $\approx^?$ Pobyt, Meteo - 4ft-Miner, výstup

Task run

Start: 1.3.2016 22:28:21 Total time: 0h 0m 1s

Number of verifications: 200634

Number of hypotheses: 16 Mode: Standard

Add grou

Actual group of hypotheses: All hypotheses

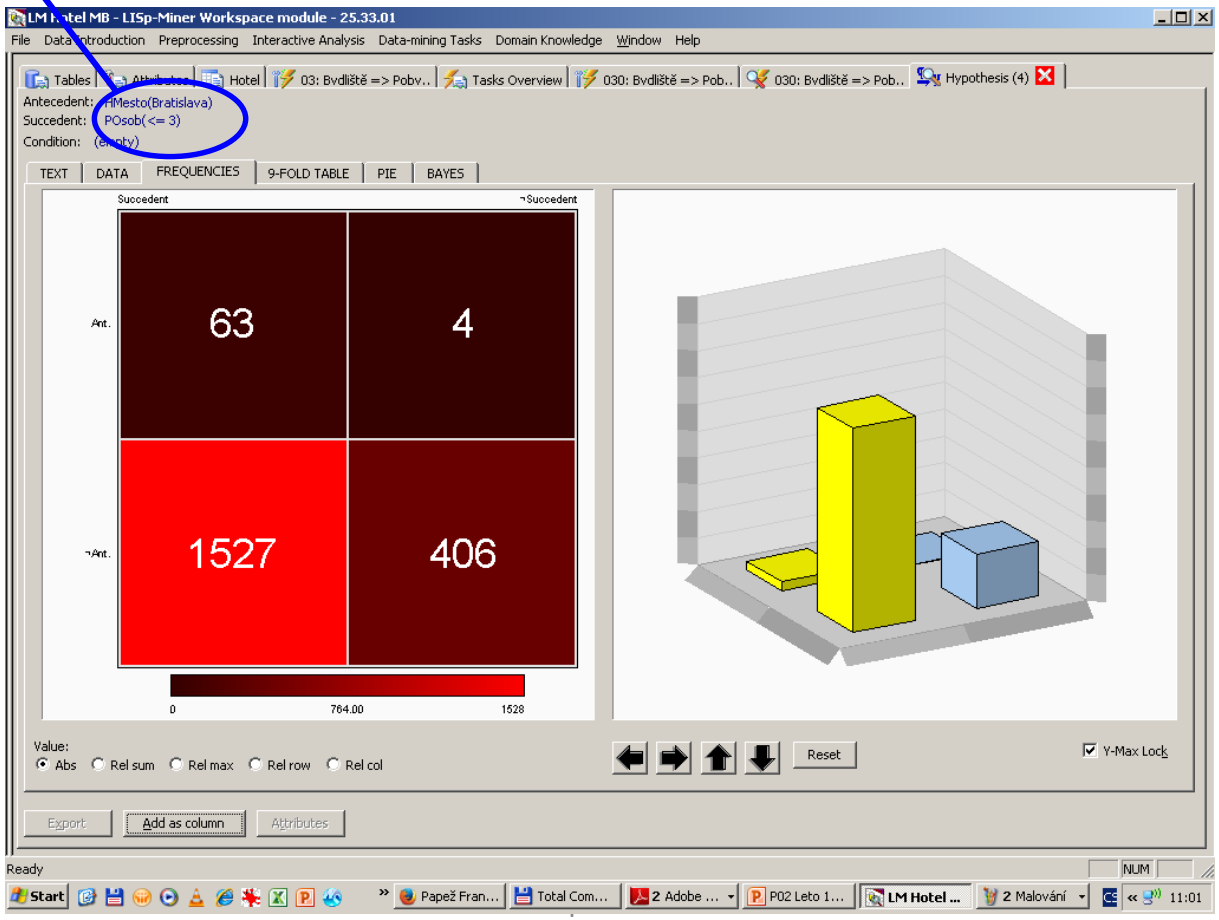
Hypotheses in group: 16 Shown hypotheses: 16 Highlighted:

Nr.	Id	Conf	Hypothesis
1	2	0.940	**HMesto**(*Bratislava*) >÷< **POsob**(<=*3*)
2	4	0.896	**HMesto**(*Bratislava*) >÷< **PNoci**(<=*2*)
3	1	0.886	**HMesto**(*Berlín*) >÷< **POsob**(<=*3*)
4	15	0.873	**HStat**(*Polsko*) >÷< **POsob**(>=*2*)
5	16	0.862	**HStat**(*Slovensko*) >÷< **POsob**(<=*3*)
6	7	0.857	**HMesto**(*Drážďany*) >÷< **POsob**(<=*3*)
7	5	0.849	**HMesto**(*Brno*) >÷< **POsob**(<=*3*)
8	14	0.846	**HStat**(*Německo*) >÷< **POsob**(<=*3*)
9	13	0.841	**HMesto**(*Vídeň*) >÷< **POsob**(<=*3*)
10	3	0.836	**HMesto**(*Bratislava*) >÷< **PNoci**(<=*2*) & **POsob**(<=*3*)
11	6	0.829	**HMesto**(*České Budějovice*) >÷< **POsob**(<=*3*)
12	10	0.823	**HMesto**(*Mnichov*) >÷< **POsob**(<=*3*)
13	12	0.823	**HMesto**(*Praha*) >÷< **POsob**(<=*3*)
14	11	0.823	**HMesto**(*Nitra*) >÷< **POsob**(>=*2*)
15	8	0.808	**HMesto**(*Jihlava*) >÷< **POsob**(>=*2*)
16	9	0.807	**HMesto**(*Linec*) >÷< **PNoci**(<=*2*)

Detailní výstup prvního pravidla

Antecedent: HMesto(Bratislava)
Succedent: POsob(<= 3)

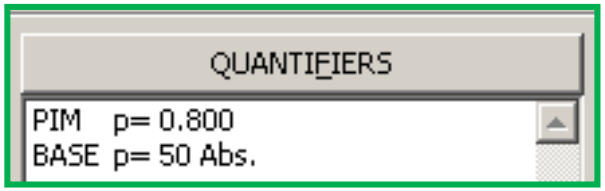
$$\text{HMesto(Bratislava)} \Rightarrow_{0.8,\ 50} \text{POsob}(\leq 3)$$



Detailní výstup prvního pravidla – komentář

HMesto(Bratislava) $\Rightarrow_{0.8,\ 50}$ POsob($\leq$ 3)

je v HotelplusExterni pravdivé protože splňuje zadané podmínky



Matice dat	S	¬S
A	a	b
¬A	c	d

$$\frac{a}{a+b} \geq 0.8 \wedge a \geq 50$$

HotelplusExterni	POsob(≤3)	¬POsob(≤ 3)
Mesto(Bratislava)	63	4
¬HMesto(Bratislava)	1 527	406

$$\frac{a}{a+b} = \frac{63}{63+4} = 0.94 \geq 0.8 \wedge 63 \geq 50$$

Detailní výstup prvního pravidla – komentář

Vzhledem ke konkrétním hodnotám je pravdivé i pravidlo

HMesto(Bratislava) $\Rightarrow_{0.94,\ 63}$ **POsob(≤ 3)**

HotelplusExterni	POsob(≤3)	¬POsob(≤ 3)
HMesto(Bratislava)	63	4
¬HMesto(Bratislava)	1 527	406

$$\frac{63}{63+4} = 0.94$$

Pravidlo **HMesto(Bratislava)** $\Rightarrow_{0.94,\ 63}$ **POsob(≤ 3)** říká

- (nejméně) 94% řádků splňujících HMesto(Bratislava) splňuje i POsob(≤3)
- (nejméně) 63 řádků splňuje HMesto(Bratislava) i POsob(≤3)