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

Typ úlohy: 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

Podmíněná asociační pravidla – příklad

http://lispminer.vse.cz/wiki/doku.php?id=lmdemo:hotel2015:task:ft

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			7	48.2115631	Rakou												
58	muž	České Budějovice		14.4757883	48.9763													
81	žena	Vídeň	16.3736767	48.2115631	Rakou													
22	žena	Drážďany	13.7397044	51.0497456														
82	muž	Katovice	19.0241283	50.2592108	Po													
55	muž	Praha	14.4212806	50.0874967	ČR	14												
75	žena	Berlín	13.3908886	52.5176189	Něm													
66	žena	Linec	14.2862742	48.3066489	Rakousko	23.2.2012	1	2	služební	2420.00	0.000	0.00	2420.00	spokojen	93	93	73	21
64	žena	Linec	14.2862742	48.3066489	Rakousko													
35	muž	Košice	21.2543528	48.7160408	Slovensk													
32	muž	Mnichov	11.5836375	48.1364669	Německo													
65	muž	Plzeň	13.3771556	49.7490406	ČR	13.3.												
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? A to obecně i zvlášť pro rekreační a služební pobyty?

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

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 / PTypPobytu

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 $\Rightarrow^{?}$ Pobyt, Meteo / PTypPobytu

LM Hotel MB - LISp-Miner Workspace module - 25.35.00

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

Tab Tree Hide

- A. Data Introduction
  - Tables
- B. Data Preprocessing
- C. Interactive Analysis
- D. Data-mining Tasks
  - Overview
  - 02: Dotaznik_ef3 x Dot
  - 02: Range
  - 03: Bydliště => Pobyt,
    - Task Results
    - Task Settings
    - Hypothesis (2419)
  - 03: Bydliště => Pobyt,
    - Task Settings
- E. Domain knowledge
- W. Workspace

02: Dotaznik ef3 x D.. | 03: Bydliště => Poby.. | 03: Bydliště => Poby.. | Hypothesis (2419) | 03: Bydliště => Poby..

Data-mining Task basic parameters

Name: 03: Bydliště => Pobyt, Meteo / PTypPobytu ID: 4

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

Bydliště(*)

Total length: 1

QUANTIFIERS

PIM p= 0.900

BASE p= 80 Abs.

$\Rightarrow^{?}$

Generation information

Status: Solved, 2 run(s)

Mode: Standard

SUCCEDENT

Pobyt Con, 2 - 2

» PNoci_enum_m (seq), 1 - 2 B, pos

» PNoci_exp (subset), 1 - 1 B, pos

» PDenTydne (subset), 1 - 1 B, pos

Meteo Con, 0 - 5

» MObloha (subset), 1 - 1 B, pos

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

Total length: 1 - 5 {2 - 3}

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

Pobyt Con, 0 - 5

» PTypPobytu (subset), 1 - 1 B, pos

PTypPobytu(*)

Total length: 0 - 5 {0 - 1}

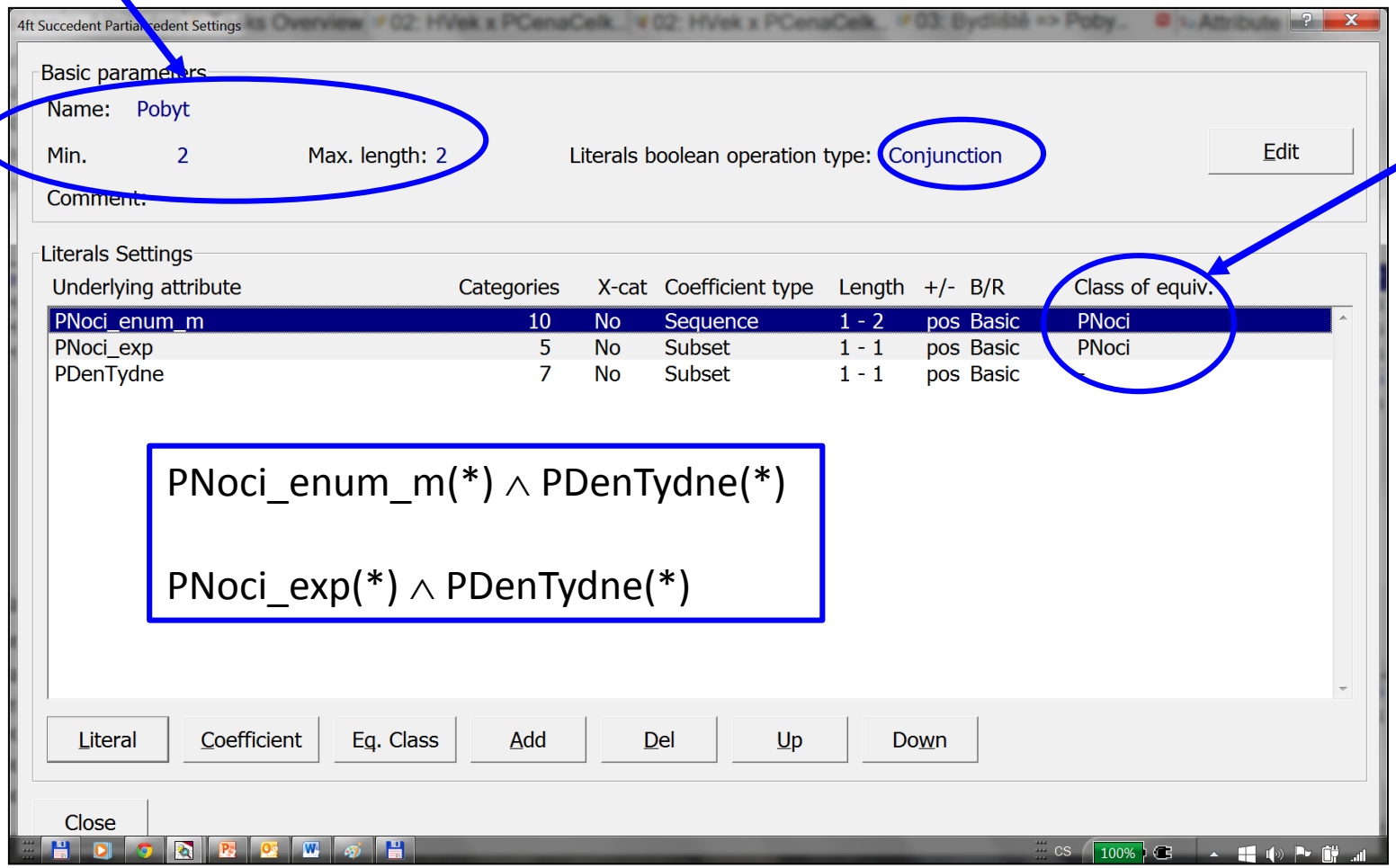
Params | Switch | Validate | Task Clone

Run | Bkgrnd Run | Grid Run | Show Results

Ready

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

Pobyt(*)



PNoci_enum_m(*) $\wedge$ PDenTydne(*)

PNoci_exp(*) $\wedge$ PDenTydne(*)

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

Data-mining Task basic parameters

Name: 03: Bydliště => Pobyt, Meteo / PTypPobytu — ID: 4

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

{A}

Total length: 1

QUANTIFIERS

PIM p= 0.900
BASE p= 80 Abs.

Generation information

Matice dat/?	S	¬S
A	a	b
¬A	c	d

$$a / (a + b) \geq 0.9 \wedge a \geq 80$$

SUCCEDENT

Pobyt	Con, 2 - 2
» PNoci_enum_m (seq), 1 - 2	B, pos
» PNoci_exp (subset), 1 - 1	B, pos
» PDenTydne (subset), 1 - 1	B, pos
Meteo	Con, 0 - 5
» MObloha (subset), 1 - 1	B, pos

{S}

Total length: 1 - 5 {2 - 3}

CONDITION

Pobyt	Con, 0 - 5
» PTypPobytu (subset), 1 - 1	B, pos

Task parameters

Handling of missing values: Ignore X-ca

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

Bydliště $\Rightarrow^{?}$ Pobyt, Meteo / PTypPobytu - Výstup

Tables | Attributes | Tasks Overview | HotelPlusExterni | 03: Bydliště => Poby.. | 03: Bydliště => Poby..

Task: 03: Bydliště => Pobyt, Meteo / PTypPobytu
Comment: -
Taskgroup: 03: Typické pobyty podle bydliště hosta
Data matrix: HotelPlusExterni
Task type: 4ft-Miner

Edit

Task run
Start: 20.9.2015 14:38:57 Total time: 0h 0m 0s
Number of verifications: 3200
Number of hypotheses: 0 Mode: Standard

Show all | Show not in group
Show hypotheses just from group:

Add group | Del group | Edit group

Actual group of hypotheses: All hypotheses
Hypotheses in group: 0 Shown hypotheses: 0 Highlighted: 0

Nr.	Id	Conf	Hypothesis

Modifikace PIM (p-implikace)

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

Data-mining Task basic parameters

Name: 03: Bydliště => Pobyt, Meteo / PTypPobytu (01)

Comment: Snížení p-Implikace na 0,7

Taskgroup: 03: Typické pobyty podle bydliště hosta

Task type: 4ft-Miner

Data matrix: HotelPlusExterni

ID: 5

Edit

PIM: $0.9 \rightarrow 0.7$

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= 0.700

BASE p= 80 Abs.

Matice dat/?	S	¬S
A	a	b
¬A	c	d

$$a / (a + b) \geq 0.7 \wedge a \geq 80$$

SUCCEDENT

Pobyt	Con, 2 - 2
» PNoci_enum_m (seq), 1 - 2	B, pos
» PNoci_exp (subset), 1 - 1	B, pos
» PDenTydne (subset), 1 - 1	B, pos
Meteo	Con, 0 - 5
» MObloha (subset), 1 - 1	B, pos

Total length: 1 - 5 {2 - 3}

CONDITION

Pobyt	Con, 0 - 5
» PTypPobytu (subset), 1 - 1	B, pos

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

Výstup po modifikaci

LM Hotel MB - LISp-Miner Workspace module - 25.24.00

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

HotelPlusExterni | 03: Bydliště => Poby.. | 03: Bydliště => Poby.. | 03: Bydliště => Poby.. | 03: Bydliště => Poby..

Task: 03: Bydliště => Pobyt, Meteo / PTypPobytu (01)
Comment: Snížení p-Implikace na 0,7
Taskgroup: 03: Typické pobyty podle bydliště hosta
Data matrix: HotelPlusExterni
Task type: 4ft-Miner

Edit

Task run
Start: 20.9.2015 14:42:45 Total time: 0h 0m 0s
Number of verifications: 3200
Number of hypotheses: 4 Mode: Standard

Show all | Show not in group | Highlight
Show hypotheses just from group:

Add group | Del group | Edit group

Actual group of hypotheses: All hypotheses
Hypotheses in group: 4 Shown hypotheses: 4 Highlighted: 0

Delete hypotheses

Nr.	Id	Conf	Hypothesis
1	1	0.841	HMesto(*Linec*) >÷< PNoci(<=*2*) & PDenTydne(*Pá*) / PTypPobytu(*rekreační*)
2	3	0.752	HMesto(*České Budějovice*) >÷< PNoci(<=*2*) & PDenTydne(*Pá*) / PTypPobytu(*rekreační*)
3	2	0.721	HMesto(*Linec*) >÷< PNoci(<=*2*) & PDenTydne(*Pá*) & MObloha(*slunečno*) / PTypPobytu(*rekreační*)
4	4	0.720	HMesto(*České Budějovice*) >÷< PNoci(<=*2*) & PDenTydne(*Pá*) & MObloha(*slunečno*) / PTypPobytu(*rekreační*)

Detail | Goto ID | Copy | Remove | Filter | Syntax Filter | BK Filter | BK Survey | Sorting | Export

Ready NUM

CS 100% 21:29

Detailní výstup prvního pravidla

Antecedent: HMesto(Linec)
Succedent: PNoci(1, 2) & PDenTydne(Pá)
Condition: PTypPobytu(rekreační)

Detailní výstup prvního pravidla – komentář (1)

Antecedent: HMesto(Linec)
Succedent: PNoci(1, 2) & PDenTydne(Pá)
Condition: PTypPobytu(rekreační)

- Jedná se o podmíněné asociační pravidlo

  HMesto(Linec) $\Rightarrow_{0.84,175}$ PNoci(1,2) $\wedge$ PDenTydne(Pá) / PTypPobytu(rekreační)
- Vyhodnocuje se na matici dat HotelPlusExterni / TypPobytu(rekreační)
- Řádky matice HotelPlusExterni splňující TypPobytu(rekreační)
- 175 + 33 + 455 + 927 = 1 590 (ne 2 000 = počet řádků v HotelplusExterni)

HotelplusExterni / TypPobytu(rekreační)	Pnoci(1,2) $\wedge$ PDenPobytu(Pá)	$\neg$(Pnoci(1,2) $\wedge$ PDenPobytu(Pá))
HMesto(Linec)	175	33
$\neg$HMesto(Linec)	455	927

Detailní výstup prvního pravidla – komentář (2)

HotelplusExterni / TypPobytu(rekreační)	Pnoci(1,2) $\wedge$ PDenPobytu(Pá)	$\neg$(Pnoci(1,2) $\wedge$ PDenPobytu(Pá))
HMesto(Linec)	175	33
$\neg$HMesto(Linec)	455	927

- 175 / (175 + 33) = 0.84
- **Pravidlo**
  HMesto(Linec) $\Rightarrow_{0.84,175}$ PNoci(1,2) $\wedge$ PDenTydne(Pá) / PTypPobytu(rekreační)
  říká:
  Pokud bereme v úvahu pouze pobyty rekreačního typu, tak platí
  - pokud je host z Lince, tak v 84 % případů se jedná o pobyt na jednu nebo dvě noci, který začíná v pátek
  - pobytů hostů z Lince na jednu nebo dvě noci začínajících v pátek je 175.