

Asociační pravidla I

prof. RNDr. Jan Rauch, CSc.

Katedra informačního a znalostního inženýrství

Asociační pravidla I

- Využívání metod data mining
- Asociační pravidla – příklad
- Podmíněná asociační pravidla – příklad
- Asociační pravidla – jiný příklad
- Asociační pravidla – přehledný popis
- Poznámka k seminárním pracím
- Rekapitulace
- Doporučení pro zadání 4ft-kvantifikátoru

S využitím zdrojů :

[1] Rauch J., Šimůnek M.: *Dobývání znalostí z databází, LISp-Miner a GUHA*.
1. vyd. Praha : Oeconomica, 2014. 462 s. ISBN 978-80-245-2033-9.

lispminer.vse.cz/wiki (autor M. Šimůnek)



Top 10 algorithms in data mining



Přihlásit se

Vše Mapy Videá Obrázky Zprávy Více Nastavení Nástroje

Přibližný počet výsledků: 1 770 000 (0,30 s)

Top 10 algorithms in data mining | SpringerLink

<https://link.springer.com/article/10.1007/s10115-007-0114-2> ▼ Přeložit tuto stránku

autor: X Wu - 2008 - Počet citací tohoto článku: 3610 - Související články

This paper presents the top 10 data mining algorithms identified by the IEEE International Conference on Data Mining (ICDM) in December 2006: C4.5, k-Means, SVM, Apriori, EM, PageRank, AdaBoost, kNN,...

[PDF] Top 10 algorithms in data mining - Department of Computer Science

www.cs.uvm.edu/~icdm/algorithms/10Algorithms-08.pdf ▼ Přeložit tuto stránku

autor: X Wu - 2007 - Počet citací tohoto článku: 3610 - Související články

4. 12. 2007 - Abstract This paper presents the top 10 data mining algorithms identified by the IEEE. International Conference on Data Mining (ICDM) in December 2006: C4.5, k-Means, SVM, Apriori, EM, PageRank, AdaBoost, kNN, Naive Bayes, and CART. These top 10 algorithms are among the most influential data ...

[PDF] Top 10 Algorithms in Data Mining Xindong Wu

home.etf.rs/~vm/os/dmsw/Top10DMAgorithms.pdf ▼ Přeložit tuto stránku

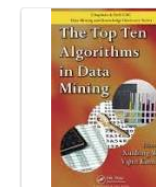
autor: X Wu - Počet citací tohoto článku: 3610 - Související články

2. Top 10 Algorithms in Data Mining: Xindong Wu and Vipin Kumar. "Top 10 Algorithms in Data Mining" by the IEEE ICDM Conference. 1. The 3-step identification process. 2. 18 identified candidates in 10 data mining topics. 3. The top 10 algorithms. 4. Follow-up actions ...

Top 10 algorithms in data mining - ACM Digital Library - Association ...

Nakupovat

Sponzorováno



Top Ten Algorithms in Data Mining (Wu Xindong)(Pevn...

2 485,00 Kč

ENbook.cz

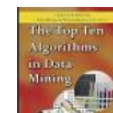
Z webu Google

Zobrazit výsledky pro

The Top Ten Algorithms in Data Mining (K...

Původní datum publikování: 2009

Redaktor: Vipin Kumar



Využívání metod data mining

<http://www.kdnuggets.com/2011/11/algorithms-for-analytics-data-mining.html>

(citováno 22. 2. 2018)

The latest KDNuggets Poll asked:
Which methods/algorithms did you use for data analysis in 2011?
The average number of algorithms per voter was 5.6.
The 10 most popular algorithms (by percent of voters who used that algorithm), are

Algorithm	Usage
Decision Trees/Rules (186)	59.8 %
Regression (180)	57.9 %
Clustering (163)	52.4 %
Statistics (descriptive) (149)	47.9 %
Visualization (119)	38.3 %
Time series/Sequence analysis (92)	29.6 %
Support Vector (SVM) (89)	28.6 %
Association rules (89)	28.6 %
Ensemble methods (88)	28.3 %
Text Mining (86)	27.7 %

Only 14% of voters used analytics in the cloud, Hadoop, EC2, etc in 2011.
Next table shows breakdown by employment type.

- KDNuggets™ News 18:n08, Feb 21: Neural network AI is simple - stop pretending you are a genius; Data Science at the command line
- Artificial Intelligence (AI) Conference, April 29 - May 2, 2018, NYC - KDNuggets Offer

KNIME Open for Innovation®
KNIME Spring Summit 2018
March 5 - 9 in Berlin, Germany
Use promo code **KDNUGGETS** for 10% off tickets.
[KNIME Spring Summit March 5-9, Berlin, Germany](#)
[Use code KDNUGGETS to save](#)

https://www.knime.com/about/events/knime-spring-summit-2018-berlin

Subscribe to KDNuggets News Follow

02_-_CF-Miner.zip

soupis fa 2017.xlsx

Zobrazit vše



Využívání metod data mining

<http://www.kdnuggets.com/2015/05/top-10-data-mining-algorithms-explained.html>

(citováno 6. 3. 2017)

The screenshot shows the KDnuggets website with the article 'Top 10 Data Mining Algorithms'. The page features a yellow header with the KDnuggets logo, social media icons for Twitter, Facebook, and LinkedIn, and a search bar. Below the header is a navigation menu with links for SOFTWARE, NEWS, Top stories, Opinions, Tutorials, JOBS, Companies, Courses, Datasets, EDUCATION, Certificates, Meetings, and Webinars. The main content area has a large purple circle with the text 'Top 10 Data Mining Algorithms'. Below this, it says 'Here are the algorithms:' followed by a list of ten algorithms. A red arrow points to '4. Apriori'. To the right of the list is a yellow box with text. Below the list, there is a 'Most Popular' section with a link to '17 More Must-Know Data Science Interview Questions and Answers'. The page also includes a sidebar with social media icons and a 'SAS Viya' advertisement.

Top 10 Data Mining Algorithms

Here are the algorithms:

- 1. C4.5
- 2. k-means
- 3. Support vector machines
- 4. Apriori
- 5. EM
- 6. PageRank
- 7. AdaBoost
- 8. kNN
- 9. Naive Bayes
- 10. CART

Consider every data problem solved. With analytics built for innovation. Explore SAS® Viya™ >

SAS Viya

Most Popular

1. **NEW** 17 More Must-Know Data Science Interview Questions and Answers

GUHA asociační pravidla jsou obecnější než ta poskytovaná Apriori

Subscribe to KDnuggets News

6. března 2017
pondělí

17:02
6.3.2017

Využívání metod data mining

<https://www.kdnuggets.com/2017/10/top-10-machine-learning-algorithms-beginners.html/>

Citováno 22. 2. 2018

Top 10 Machine Learning Algorithms for Beginners

2017 PLATINUM KDNuggets Blog

◀ Previous post Next post ▶

Like 2.1K Share 2.1K Tweet G+ Share 941

Tags: Adaboost, Algorithms, Apriori, Bagging, Boosting, Decision Trees, Ensemble methods, Explained, K-means, K-nearest neighbors, Linear Regression, Logistic Regression, Machine Learning, Naive Bayes, PCA, Top 10

Global network of experts. World of opportunities. SAS Business Knowledge Series. LEARN MORE. sas

GUHA asociační pravidla jsou obecnější než ta poskytovaná Apriori

Asociační pravidla – příklad

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	muž	Linec	14.2862742	48.2115631	Rakousko	5.6.2012	7	2	rekreační	16940.00	2100.000	200.00	18840.00	nespokojen 5 37 25 71
62	muž	Linec	14.2862742	48.2115631	Rakousko	5.6.2012	7	2	rekreační	16940.00	2100.000	200.00	18840.00	nespokojen 5 37 25 71
35	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
58	muž	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
81	žena	Videň	16.3736767	48.2115631	Rakousko	2.8.2013	2	4	rekreační	11600.00	1440.000	200.00	12840.00	průměr 44 21 62 84
22	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
82	muž	Katovice	19.0241283	50.2592108	Polsko	14.12.2012	2	2	rekreační	4840.00	600.000	200.00	5240.00	spokojen 89 81 91 53
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	16.3.2012	1	2	služební	2420.00	0.000	0.00	2420.00	spokojen 95 93 75 21
35	muž	Košice	21.2543528	48.7160408	Slovensko	6.4.2012	1	1	rekreační	1450.00	0.000	0.00	1450.00	spokojen 91 82 81 56
32	muž	Mnichov	11.5836375	48.1364669	Německo	13.8.2012	1	1	rekreační	1450.00	0.000	0.00	1450.00	spokojen 91 82 81 56
65	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
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	4	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	14.2862742	48.2115631	Rakousko	5.6.2012	7	2	rekreační	16940.00	2100.000	200.00	18840.00	nespokojen 5 37 25 71
45	žena	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

Hotel.txt

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

Bydliště ⇒? Pobyt, Meteo

Meteo.txt

MDatum	MTeplota	MOblaha
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
9.1.2012	-1.3	zataženo
10.1.2012	-1.3	srážky
11.1.2012	-1.3	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	Attribute	Used	DBCcolumn	Categories XCat	Sample categories
<ul style="list-style-type: none"> Root group of attrib Dotazník Host Bydliště Meteo Pobyt Cena Začátek Směnárna 	HCizinec_b	+	HStat	2	ne, ano
	HMesto	+	HMesto	28	Berlín, Bratislava, Brno, České Budějovice, C
	HMesto_m_hlavni		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	Attribute	Used	DBCcolumn	Categories XCat	Sample categories
<ul style="list-style-type: none"> Root group of attrib Dotazník Host Bydliště Meteo Pobyt Cena Začátek Směnárna 	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	Attribute	Used	DBCcolumn	Categories XCat	Sample categories
<ul style="list-style-type: none"> Root group of attrib Dotazník Host Bydliště Meteo Pobyt Cena Začátek Směnárna 	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, extri

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 => Pobyt... 030: Bydliště => Pobyt...

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	QUANTIFIERS	SUCCEDENT
Host/Bydliště Con, 1 - 1 » HCizinec_b (subset), 1 - 1 B, pos » HMesto (subset), 1 - 1 B, pos » HStat (subset), 1 - 1 B, pos	PIM p= 0.800 BASE p= 50 Abs. $\Rightarrow ?$ Generation information Status: Solved, 4 run(s) Mode: Standard	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

Total length: 1 - 5

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 symmetric hypotheses: Yes Extensions minimal length check: Yes

Maximal number of hypotheses: 1000

Params Switch Validate Task Clone

Run Bkgrnd Run Grid Run Show Results

CONDITION

Default Con, 0 - 0

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

Ready

Start [Taskbar icons] 17:30

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(*)

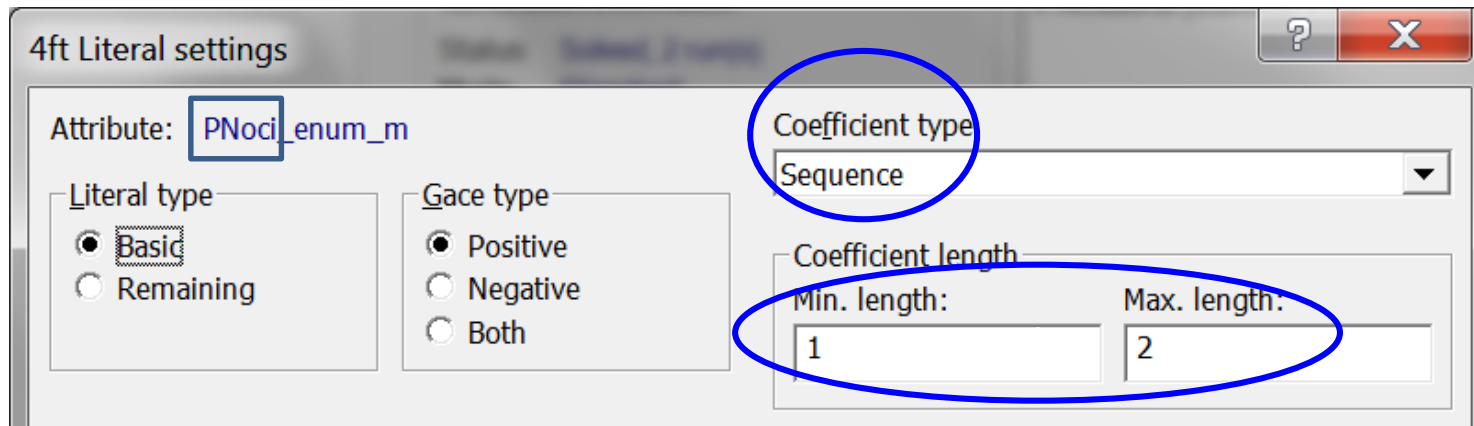
PNoci_enum_m(*) \wedge PDenTydne(*)
PNoci_enum_m(*) \wedge PDenTydne(*) \wedge Posobonoci_ef5 (*)
PNoci_enum_m(*) \wedge POsob(*) \wedge Posobonoci_ef5(*)
PDenTydne \wedge POsob(*) \wedge Posobonoci_ef5(*)

PNoci_enum_m(*) \wedge PDenTydne(*)
PNoci_enum_m(*) \wedge POsob(*)
PNoci_enum_m(*) \wedge Posobonoci_ef5(*)
PDenTydne \wedge POsob(*)
PDenTydne \wedge Posobonoci_ef5(*)
POsob(*) \wedge Posobonoci_ef5(*)

PNoci_enum_m(*) \wedge PDenTydne(*) \wedge POsob(*) \wedge Posobonoci (*)

Del Up Down

PNoci_enum_m(*)

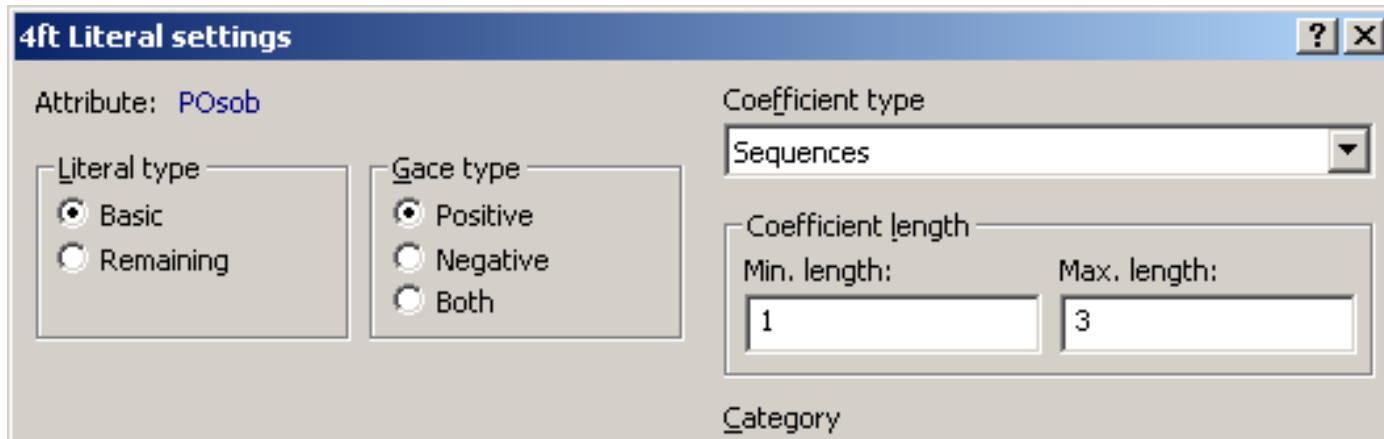


10 kategorií : 1, 2, ⟨3;6⟩, 7, ⟨8;13⟩, 14, ⟨15;20⟩, 21, ⟨22;27⟩, 28

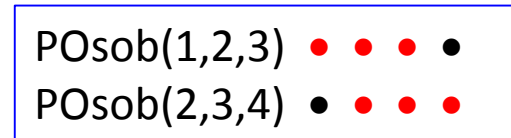
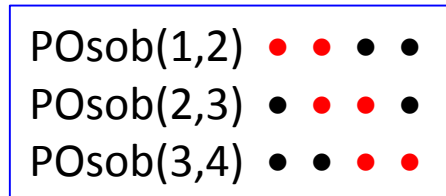
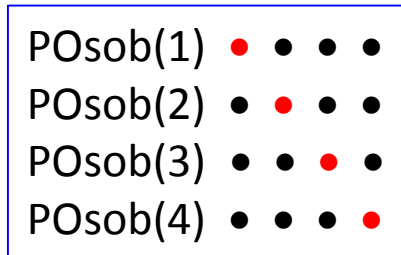
PNoci(1)	● ● ● ● ● ● ● ● ● ●
PNoci(2)	● ● ● ● ● ● ● ● ● ●
Pnoci(⟨3;6⟩)	● ● ● ● ● ● ● ● ● ●
PNoci(7)	● ● ● ● ● ● ● ● ● ●
PNoci(⟨8;13⟩)	● ● ● ● ● ● ● ● ● ●
PNoci(14)	● ● ● ● ● ● ● ● ● ●
PNoci(⟨15;20⟩)	● ● ● ● ● ● ● ● ● ●
PNoci(21)	● ● ● ● ● ● ● ● ● ●
PNoci(⟨22;27⟩)	● ● ● ● ● ● ● ● ● ●
PNoci(28)	● ● ● ●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

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



Bydliště \Rightarrow ? Pobyt, Meteo - zadání pro 4ft-Miner

The screenshot shows the LISp-Miner workspace for a task named "030: Bydliště => Pobyt, Meteo". The task parameters are:

- Name: 030: Bydliště => Pobyt, Meteo
- Taskgroup: 03: Typické pobyty podle bydliště hosta
- Task type: 4ft-Miner
- Data matrix: HotelPlusExterni

The interface displays the ANTECEDENT (A) and QUANTIFIERS (S) sections. The ANTECEDENT section lists:

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

The QUANTIFIERS section shows:

- PIM p= 0.800
- BASE p= 50 Abs.

The generation information indicates the task is solved in 4 runs using standard mode. The resulting rules are listed in the S section:

- » PNoci_enum_m (seq), 1 - 2
- » PDenTydne (subset), 1 - 1
- » POSob (seq), 1 - 3
- » POSobonoci_ef5 (seq), 1 - 2
- Meteo
- » MObloha (subset), 1 - 1
- » MTeploata_exp (subset), 1 - 1

A detailed view of the QUANTIFIERS section shows the parameters: PIM p= 0.800 and BASE p= 50 Abs.

The interface also includes a table for handling missing values and various control buttons like Params, Switch, Validate, Task Clone, Run, Bkgrnd Run, Grid Run, and Show Results.

QUANTIFIERS

PIM p= 0.800
BASE p= 50 Abs.

Matrice dat	S	$\neg S$
A	a	b
$\neg A$	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: 13.2016 22:28:21 Total time: 0h 0m 1s

Number of verifications: 200634

Number of hypotheses: 16 Mode: Standard Add group

Actual group of hypotheses: All hypotheses

Hypotheses in group: 16 Shown hypotheses: 16 Highlighted:

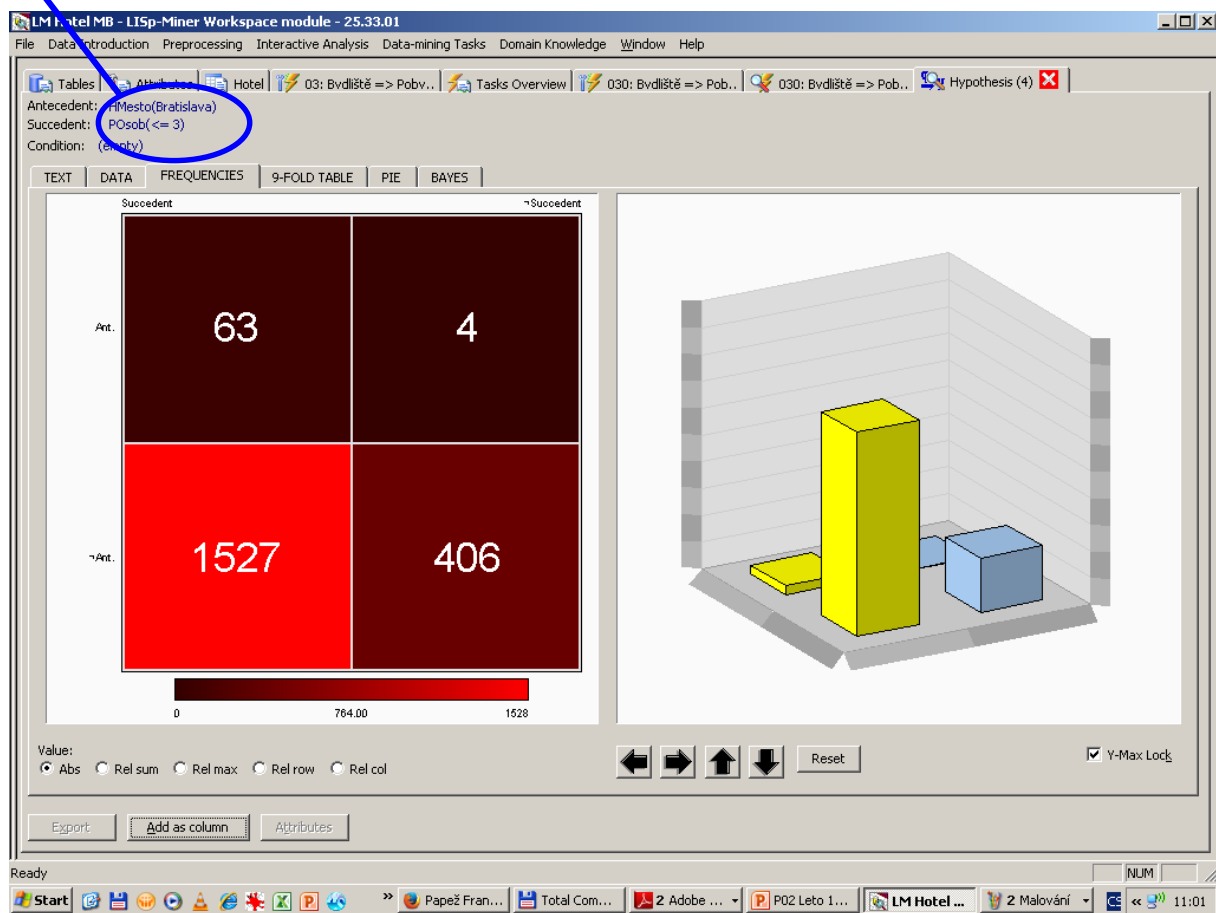
Nr.	Id	Conf	Hypothesis
1	2	0.940	HMesto(<i>Bratislava</i>) >÷< POsob(<=3)
2	4	0.896	HMesto(<i>Bratislava</i>) >÷< PNoci(<=2)
3	1	0.886	HMesto(<i>Berlín</i>) >÷< POsob(<=3)
4	15	0.873	HStat(<i>Polsko</i>) >÷< POsob(>=2)
5	16	0.862	HStat(<i>Slovensko</i>) >÷< POsob(<=3)
6	7	0.857	HMesto(<i>Drážďany</i>)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

Asociační pravidla I

Detailní výstup prvního pravidla

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

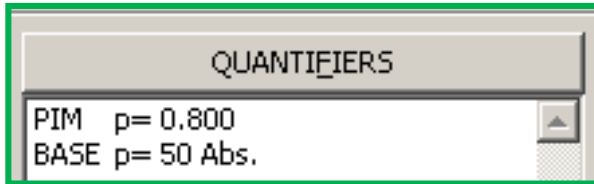
$HMesto(Bratislava) \Rightarrow_{0.8, 50} POsob(\leq 3)$



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	$\neg S$
A	a	b
$\neg A$	c	d

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

HotelplusExterni	POsob(≤ 3)	\neg POsob(≤ 3)
Mesto(Bratislava)	63	4
\neg 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)	\neg POsob(≤ 3)
HMesto(Bratislava)	63	4
\neg 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)

Asociační pravidla I

- Komentář ke kvízovým otázkám
- Využívání metod data mining
- Asociační pravidla – příklad
- Podmíněná asociační pravidla – příklad
- Asociační pravidla – jiný příklad, jiný 4ft-kvantifikátor
- Asociační pravidla – přehledný popis
- Poznámka k seminárním pracím
- Rekapitulace
- Doporučení pro zadání 4ft-kvantifikátoru

Podmíněná asociační pravidla – příklad

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

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	muž	Linec	14.2862742	48.2115631	Rakousko	5.6.2012	7	2	rekreační	16940.00	2100.000	200.00	18840.00	nespokojen 5 37 25 71
62	muž	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
35	žena	Linec	14.2862742	48.2115631	Rakousko	5.6.2012	7	2	rekreační	16940.00	2100.000	200.00	18840.00	nespokojen 5 37 25 71
58	muž	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
81	žena	Videň	16.3736767	48.2115631	Rakousko	5.6.2012	7	2	rekreační	16940.00	2100.000	200.00	18840.00	nespokojen 5 37 25 71
22	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
82	muž	Katovice	19.0241283	50.2592108	Polsko	2.8.2013	2	4	rekreační	11600.00	1440.000	200.00	12840.00	průměr 44 21 62 84
55	muž	Praha	14.4212806	50.0874967	ČR	14.11.2013	4	1	rekreační	5800.00	0.000	200.00	5600.00	spokojen 84 79 86 48
75	žena	Berlín	13.3908886	52.5176189	Německo	19.1.2013	14	4	rekreační	81200.00	10080.000	200.00	91080.00	průměr 59 23 46 96
66	žena	Linec	14.2862742	48.3066489	Rakousko	23.2.2012	1	2	služební	2420.00	0.000	0.00	2420.00	spokojen 91 82 81 56
64	žena	Linec	14.2862742	48.3066489	Rakousko	23.2.2012	1	2	služební	2420.00	0.000	0.00	2420.00	spokojen 91 82 81 56
35	muž	Košice	21.2543528	48.7160408	Slovensko	5.1.2013	14	2	rekreační	40600.00	5040.000	200.00	45440.00	nespokojen 27 15 30 12
32	muž	Mnichov	11.5836375	48.1364669	Německo	19.1.2013	14	4	rekreační	81200.00	10080.000	200.00	91080.00	průměr 59 23 46 96
65	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
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	14.2862742	48.2115631	Rakousko	5.6.2012	7	2	rekreační	16940.00	2100.000	200.00	18840.00	nespokojen 5 37 25 71
45	žena	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

Hotel.txt

Vyplývají z místa bydliště hosta nějaké typické parametry pobytu, případně i počasí? A to obecně i zvláště pro rekreační a služební pobyty.

Bydliště ⇒? Pobyt, Meteo / PTypPobytu

Meteo.txt

MDatum	MTeplota	MOblaha
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
9.1.2012	-1.3	zataženo
10.1.2012	-1.3	srážky
11.1.2012	-1.3	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 ⇒? Pobyt, Meteo / PTypPobytu

Groups of attributes tree	Attribute	Used	DBCColumn	Categories XCat	Sample categories
<ul style="list-style-type: none"> Root group of attrib Dotazník Host Bydliště Meteo Pobyt Cena Začátek Směnárna 	HCizinec_b	+	HStat	2	ne, ano
	HMesto	+	HMesto	28	Berlín, Bratislava, Brno, České Budějovice, C
	HMesto_m_hlavni		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	Attribute	Used	DBCColumn	Categories XCat	Sample categories
<ul style="list-style-type: none"> Root group of attrib Dotazník Host Bydliště Meteo Pobyt Cena Začátek Směnárna 	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	Attribute	Used	DBCColumn	Categories XCat	Sample categories
<ul style="list-style-type: none"> Root group of attrib Dotazník Host Bydliště Meteo Pobyt Cena Začátek Směnárna 	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, extri

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 D..
 - 02: Range
 - 03: Bydliště => Pobyt, Meteo / PTypPobytu
 - Task Results
 - Task Settings
 - Hypothesis (2419)
 - 03: Bydliště => Pobyt, Meteo / PTypPobytu
 - Task Settings
- E. Domain knowledge
- W. Workspace

02: Dotaznik_ef3 x D.. | 03: Bydliště => Pobyt, Meteo / PTypPobytu | 03: Bydliště => Pobyt, Meteo / PTypPobytu | Hypothesis (2419) | 03: Bydliště => 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

ANTCEDENT	QUANTIFIERS	SUCCEEDENT
Host/Bydliště » HCizinec_b (subset), 1 - 1 » HMesto (subset), 1 - 1 » HStat (subset), 1 - 1	PIM p=0.900 BASE p=80 Abs.	Pobyt » PNoci_enum_m (seq), 1 - 2 » PNoci_exp (subset), 1 - 1 » PDenTydne (subset), 1 - 1 Meteo » MObloha (subset), 1 - 1

Generation information

Status: Solved, 2 run(s)

Mode: Standard

Total length: 1

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 symmetric hypotheses: Yes Extensions minimal length check: Yes

Maximal number of hypotheses: 1000

Params Switch Validate Task Clone

Run Bkgrnd Run Grid Run Show Results

CONDITION

Pobyt
» PTypPobytu (subset), 1 - 1

Total length: 0 - 5 {0 - 1}

Bydliště(*)

\Rightarrow ?

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

$PTypPobytu(*)$

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

Pobyt(*)

4ft Succedent Partial Redent Settings

Basic parameters

Name: Pobyt

Min. 2 Max. length: 2

Literals boolean operation type: Conjunction

Edit

Comment:

Literals Settings

Underlying attribute	Categories	X-cat	Coefficient type	Length	+/-	B/R	Class of equiv.
PNoci_enum_m	10	No	Sequence	1 - 2	pos	Basic	PNoci
PNoci_exp	5	No	Subset	1 - 1	pos	Basic	PNoci
PDenTydne	7	No	Subset	1 - 1	pos	Basic	

PNoci_enum_m(*) ^ PDenTydne(*)

PNoci_exp(*) ^ PDenTydne(*)

Literal Coefficient Eq. Class Add Del Up Down

Close

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	QUANTIFIERS	SUCCEEDENT									
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}	PIM p= 0.900 BASE p= 80 Abs. Generation information	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}									
<table border="1"> <thead> <tr> <th>Matrice dat/?</th> <th>S</th> <th>\negS</th> </tr> </thead> <tbody> <tr> <td>A</td> <td>a</td> <td>b</td> </tr> <tr> <td>\negA</td> <td>c</td> <td>d</td> </tr> </tbody> </table> $a / (a + b) \geq 0.9 \wedge a \geq 80$		Matrice dat/?	S	\neg S	A	a	b	\neg A	c	d	Total length: 1 - 5 {2 - 3}
Matrice dat/?	S	\neg S									
A	a	b									
\neg A	c	d									
Total length: 1		CONDITION									
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		Pobyt Con, 0 - 5 » PTypPobytu (subset), 1 - 1 B, pos									

Bydliště ⇒? 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 [Edit](#)
Data matrix: HotelPlusExterni
Task type: 4ft-Miner

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) ID: 5

Comment: Snížení p-Implikace na 0,7

Taskgroup: 03: Typické pobyty podle bydliště hosta

Task type: 4ft-Miner Data matrix: HotelPlusExterni

ANTECEDENT	QUANTIFIERS	SUCCEDENT
<p>Host/Bydliště Con, 1 - 1</p> <ul style="list-style-type: none"> » HCizinec_b (subset), 1 - 1 B, pos » HMesto (subset), 1 - 1 B, pos » HStat (subset), 1 - 1 B, pos 	<p>PIM p= 0.700</p> <p>BASE p= 80 Abs.</p>	<p>Pobyt Con, 2 - 2</p> <ul style="list-style-type: none"> » PNoci_enum_m (seq), 1 - 2 B, pos » PNoci_exp (subset), 1 - 1 B, pos » PDenTydne (subset), 1 - 1 B, pos <p>Meteo Con, 0 - 5</p> <ul style="list-style-type: none"> » MObloha (subset), 1 - 1 B, pos
<p>Total length: 1</p>		
<p>Task parameters</p> <p>Handling of missing values: Ignore X-categories</p> <p>Prime rule test for implications enabled: No</p> <p>Include succedent extensions of 100% implications: Yes</p> <p>Include extensions of coefficients with no change in the four-fold table: Yes</p> <p>Include extensions of cedents with no change in the four-fold table: Yes</p>		
<p>CONDITION</p> <p>Pobyt Con, 0 - 5</p> <ul style="list-style-type: none"> » PTypPobytu (subset), 1 - 1 B, pos 		

PIM: 0.9 \rightarrow 0.7

Matrice dat/?	S	\neg S
A	a	b
\neg A	c	d

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

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 [Edit]
Data matrix: HotelPlusExterni
Task type: 4ft-Miner

Task run
Start: 20.9.2015 14:42:45 Total time: 0h 0m 0s
Number of verifications: 3200
Number of hypotheses: 4 Mode: Standard [Add group] [Del group] [Edit group]

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

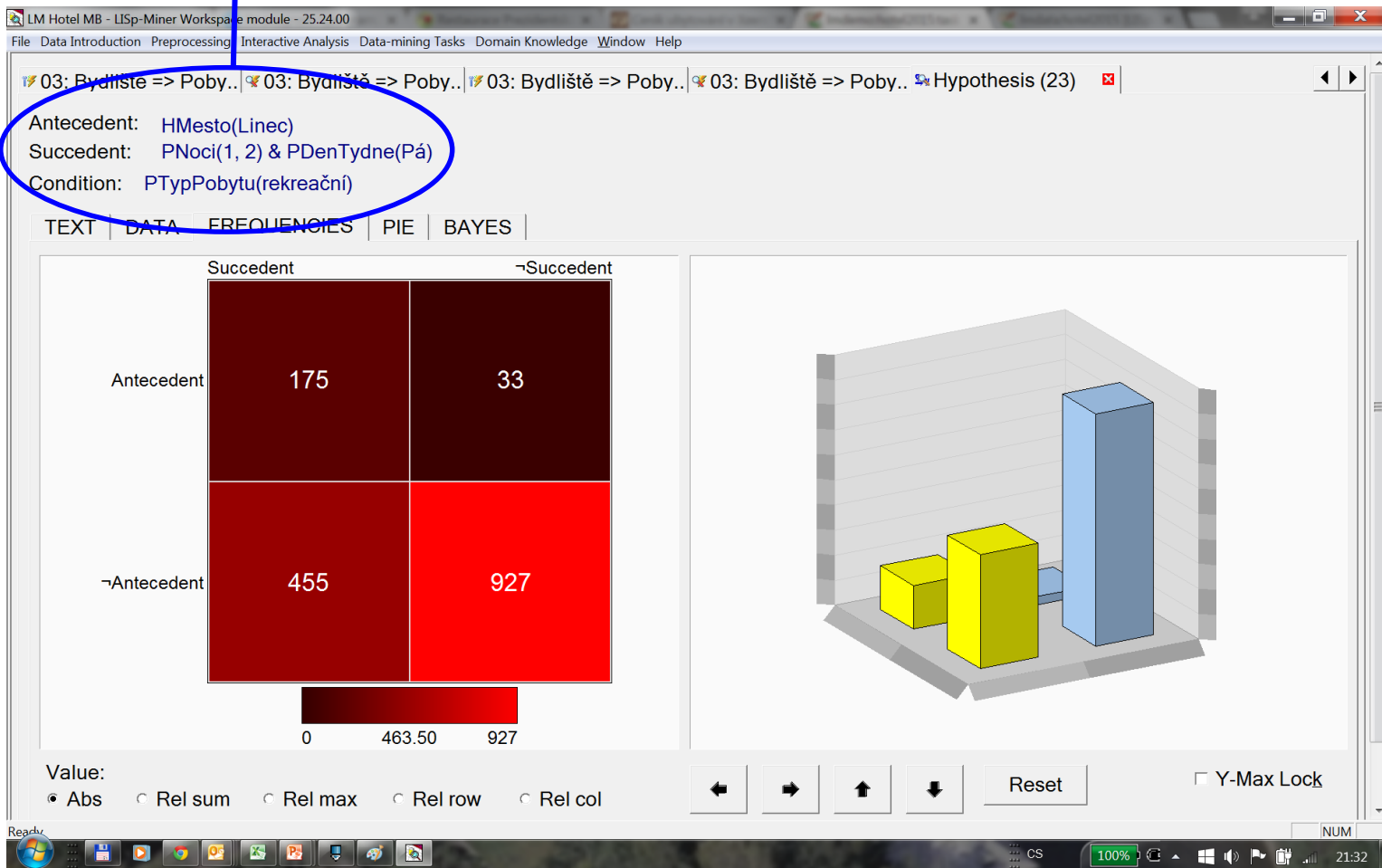
Nr.	Id	Conf	Hypothesis
1	1	0.841	HMesto(<i>Linec</i>) >+< PNoci(<=2) & PDenTydne(<i>Pá</i>) / PTypPobytu(<i>rekreační</i>)
2	3	0.752	HMesto(<i>České Budějovice</i>) >+< PNoci(<=2) & PDenTydne(<i>Pá</i>) / PTypPobytu(<i>rekreační</i>)
3	2	0.721	HMesto(<i>Linec</i>) >+< PNoci(<=2) & PDenTydne(<i>Pá</i>)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

[Detail] [Goto ID] [Copy] [Remove] [Filter] [Syntax Filter] [BK Filter] [BK Survey] [Sorting] [Export]

Ready NUM 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)	$P_{\text{noci}(1,2)} \wedge P_{\text{DenPobytu}(\text{Pá})}$	$\neg(P_{\text{noci}(1,2)} \wedge P_{\text{DenPobytu}(\text{Pá})})$
HMesto(Linec)	175	33
\neg HMesto(Linec)	455	927

○ $175 / (175 + 33) = 0.84$

○ **Pravidlo**

$HMesto(Linec) \Rightarrow_{0.84,175} P_{\text{noci}(1,2)} \wedge P_{\text{DenTydne}(\text{Pá})} / P_{\text{TypPobytu}(\text{rekreační})}$

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.

Kvízy – následuje kvízová otázka

- V průběhu většiny přednášek bude několik kvízových otázek.
- Odpovědi se vkládají přes ISIS pomocí vašeho notebooku nebo chytrého telefonu.
- Kvízové otázky souvisí s přednášenou látkou a jejich správným zodpovězením prokazujete, že rozumíte přednášené látce.
- K výsledkům kvízů bude přihlédnuto v závěrečné klasifikaci.
- Obdobné otázky budou součástí „ostrých“ testů .
- Pro odpověď na kvíz nepoužívejte žádný software.

Otázka 2

Předpokládejte, že je dána čtyřpolní tabulka

Matice dat	V	$\neg V$
U	b	y
$\neg U$	x	a

kde U a V jsou booleovské atributy – sloupce matice dat.

Který vzorec vyjadřuje, že současně platí obě následující tvrzení a) i b)?

- a) nejméně 91% řádků splňujících U splňuje i V
- b) nejméně 263 řádků splňuje U i V .

Asociační pravidla I

- Komentář ke kvízovým otázkám
- Využívání metod data mining
- Asociační pravidla – příklad
- Podmíněná asociační pravidla – příklad
- **Asociační pravidla – jiný příklad, jiný 4ft-kvantifikátor**
- Asociační pravidla – přehledný popis
- Poznámka k seminárním pracím
- Rekapitulace
- Doporučení pro zadání 4ft-kvantifikátoru

Jiný 4ft-kvantifikátor – příklad

euromise.vse.cz/challenge2004/data/entry/

Homepage | People | Projects

Data STULONG

Projects > Discovery Challenge 2004

Analytická otázka:
Existuje skupina pacientů definovaná pomocí osobních údajů, spotřeby alkoholu, cukru, kávy a čaje taková, že nejméně 50 % pacientů ze skupiny má riziko obezity a zároveň počet pacientů ze skupiny s rizikem obezity je minimálně 30?

Table 1: Groups of the attributes in

Groups of attri	
identification data	2
social characteristics	6
physical activity	4
smoking	3
drinking of alcohol	9
sugar, coffee, tea	3
personal anamnesis	18
questionnaire A ₂	3
physical examination	8
biochemical examination	3
risk faktors	5

? 4ft: Osobní, Alkohol, CKČ $\Rightarrow_{0.5,30}$ Obezrisk(ano)

Print page PDF version

Projects > Discovery Challenge 2004 > Data set > Entry

Mail to: webmaster

Osobní údaje

Basic parameters

Name: Osobní údaje

Min. length: 0

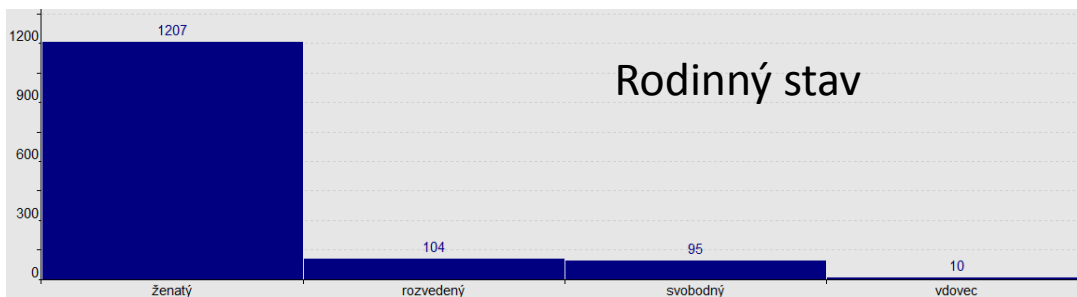
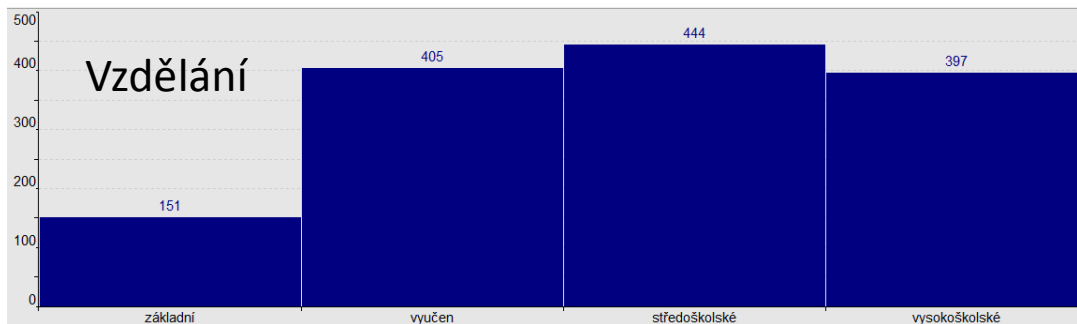
Max. length: 2

Literals boolean operation type: Conjunction

Comment: -

Literals Settings

Underlying attribute	Categories	X-cat	Coefficient type	Length	+/-	B/R
Vzdělání	4	Yes	Subsets	1 - 1	pos	Basic
Rodinný stav	4	Yes	Subsets	1 - 1	pos	Basic



Spotřeba alkoholu

Basic parameters

Name: Alkohol

Min. length: 0

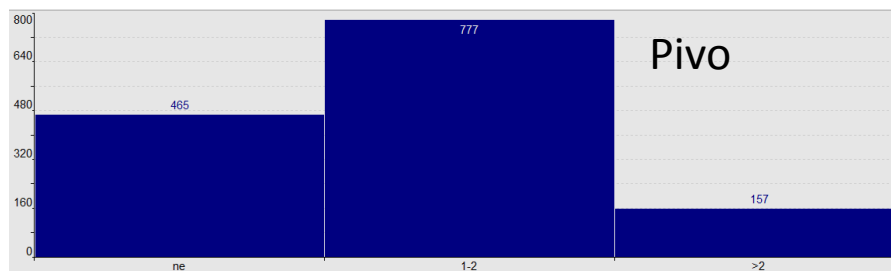
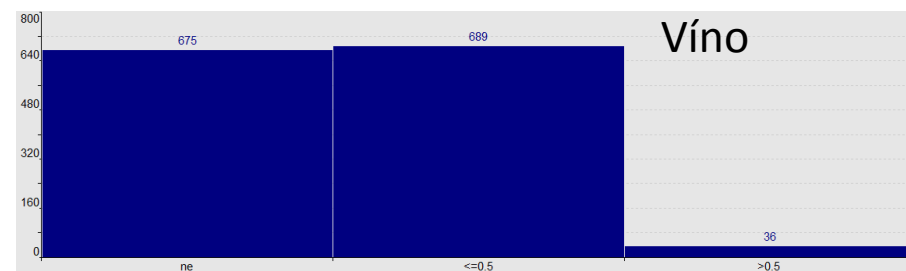
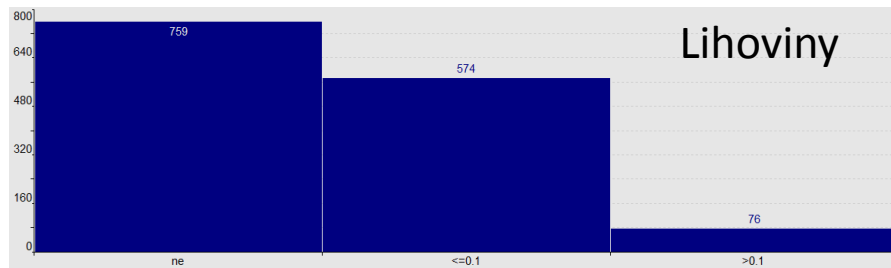
Max. length: 3

Literals boolean operation type: Conjunction

Comment: -

Literals Settings

Underlying attribute	Categories	X-cat	Coefficient type	Length	+/-	B/R
Lihoviny	3	Yes	Subsets	1 - 1	pos	Basic
Pivo	3	Yes	Subsets	1 - 1	pos	Basic
Víno	3	Yes	Subsets	1 - 1	pos	Basic



CKČ – Cukr, Káva, Čaj

Basic parameters

Name: Cukr, káva, čaj

Min. length: 0

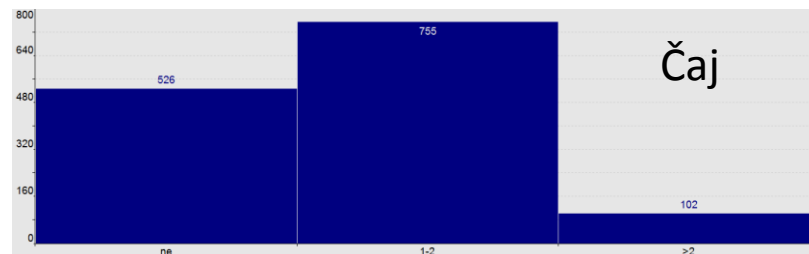
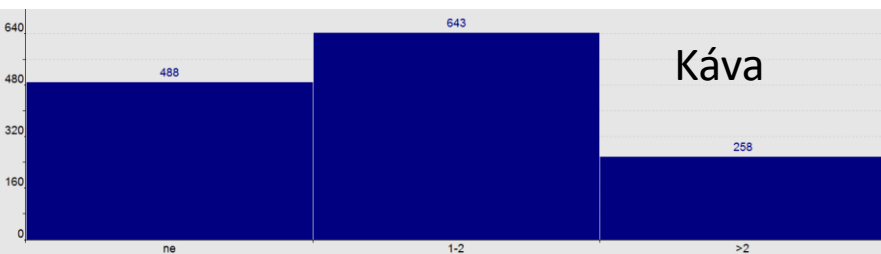
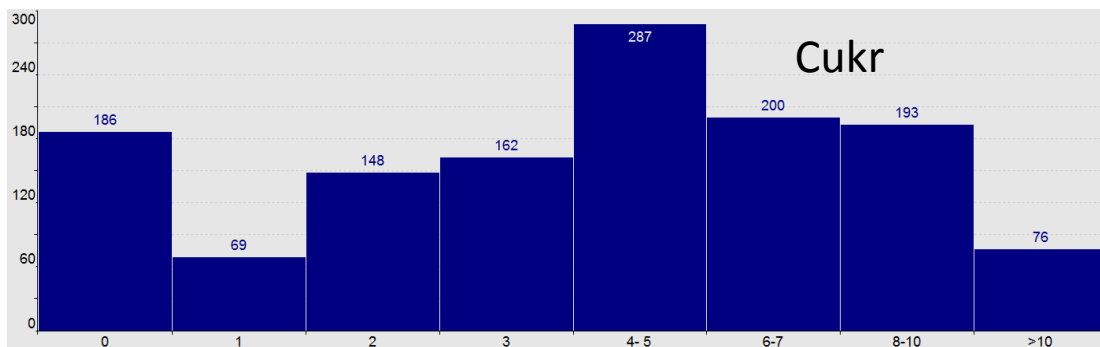
Max. length: 3

Literals boolean operation type: Conjunction

Comment: -

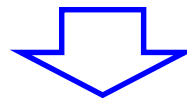
Literals Settings

Underlying attribute	Categories	X-cat	Coefficient type	Length	+/-	B/R
Cukr	8	Yes	Sequences	1 - 3	pos	Basic
Káva	3	Yes	Sequences	1 - 2	pos	Basic
Čaj	3	Yes	Sequences	1 - 2	pos	Basic



? 4ft: Osobní, Alkohol, CKČ $\Rightarrow_{0.5,30}$ Obezrisk(ano)

ANTECEDENT	QUANTIFIERS	SUCCEDENT
Osobní údaje » Vzdělání (seq), 1 - 2 » Rodinný stav (subset), 1 - 1 Alkohol » Lihoviny (subset), 1 - 1 » Pivo (subset), 1 - 1 » Víno (subset), 1 - 1 Cukr, káva, čaj » Cukr (seq), 1 - 3 » Káva (seq), 1 - 2 » Čaj (seq), 1 - 2	Con, 0 - 2 B, pos B, pos Con, 0 - 3 B, pos B, pos B, pos Con, 0 - 3 B, pos B, pos B, pos	Default Partial Cedent » Obezrisk(ano) Con, 1 - 1 B, pos
	<div style="border: 2px solid blue; padding: 2px;"> PIM p= 0.500 BASE p= 30 Abs. </div> Generation information Status: Not generated, 11 run(s) Mode: -	



Task run			
Start:	2.3.2016 16:31:31	Total time:	0h 0m 1s
Number of verifications:	44545	Mode:	Standard
Number of hypotheses:	0	Add group	
Actual group of hypotheses:	All hypotheses		
Hypotheses in group:	0	Shown hypotheses:	0
		Highlighted:	0
Nr.	Id	Conf	Hypothesis

Změna analytické otázky

Analytická otázka:

Existuje skupina pacientů definovaná pomocí osobních údajů, spotřeby alkoholu, cukru, kávy a čaje taková, že **nejméně 50 % pacientů z této skupiny má riziko obezity** a zároveň počet pacientů ze skupiny s rizikem obezity je minimálně 30?



Analytická otázka:

Existuje skupina pacientů definovaná pomocí osobních údajů, spotřeby alkoholu, cukru, kávy a čaje taková, že **relativní četnost pacientů s rizikem obezity je v této skupině o 50 % větší než v celém souboru** a zároveň počet pacientů ze skupiny s rizikem obezity je minimálně 30?

? 4ft: Osobní, Alkohol, CKČ $\Rightarrow_{0.5,30}$ Obezrisk(ano)



? 4ft: Osobní, Alkohol, CKČ $\sim^+_{0.5,30}$ Obezrisk(ano)

Změna 4ft-kvantifikátoru $\Rightarrow_{0.5,30}$ na $\sim^+_{0.5,30}$

Matrice dat	S	$\neg S$
A	a	b
$\neg A$	c	d

$$\Rightarrow_{0.5,30} \frac{a}{a+b} \geq 0.5 \wedge a \geq 30$$



$$\sim^+_{0.5,30} \frac{a}{a+b} \geq (1+0.5) \frac{a+c}{a+b+c+d} \wedge a \geq 30$$

4ft-kvantifikátor $\sim^+_{0.5,30}$

Matice dat	S	$\neg S$
A	<i>a</i>	<i>b</i>
$\neg A$	<i>c</i>	<i>d</i>

$$\underbrace{\frac{a}{a+b}}_{\text{Relativní četnost S pokud platí A}} \geq \underbrace{(1+0.5)}_{\text{je nejméně o 50\% větší než}} \underbrace{\frac{a+c}{a+b+c+d}}_{\text{relativní četnost S v celé matici}} \wedge a \geq 30$$

Relativní četnost
S pokud platí A

je nejméně o 50% větší než

relativní četnost S v celé matici

? 4ft: Osobní, Alkohol, CKČ $\sim^+_{0.5,30}$ Obezrisk(ano)

ANTECEDENT		QUANTIFIERS	SUCCEDENT	
Osobní údaje	Con, 0 - 2	AAD p= 0.500 BASE p= 30 Abs.	Default Partial Cedent	Con, 1 - 1
» Vzdělání (seq), 1 - 2	B, pos		» Obezrisk(ano)	B, pos
» Rodinný stav (subset), 1 - 1	B, pos	Generation information Status: Solved, 11 run(s) Mode: Standard		
Alkohol	Con, 0 - 3			
» Lihoviny (subset), 1 - 1	B, pos			
» Pivo (subset), 1 - 1	B, pos			
» Víno (subset), 1 - 1	B, pos			
Cukr, káva, čaj	Con, 0 - 3			
» Cukr (seq), 1 - 3	B, pos			
» Káva (seq), 1 - 2	B, pos			
» Čaj (seq), 1 - 2	B, pos			

? 4ft: Osobní, Alkohol, CKČ $\sim^+_{0.5,30}$ Obezrisk(ano) výstup

Task run

Start: 2.3.2016 16:29:12

Total time: 0h 0m 3s

Number of verifications: 44545

Number of hypotheses: 235

Mode: Standard

Add group

Del group

Edit group

Actual group of hypotheses: All hypotheses

Hypotheses in group: 235

Shown hypotheses: 235

Highlighted: 0

Nr.	Id	AvgDf	Hypothesis
1	102	1.250	Vzdělání(\leq vyučen) & Rodinný stav (ženatý) & Víno(ne) & Cukr(0,1,2) >+< Obezrisk(ano)
2	80	1.124	Vzdělání(\leq vyučen) & Víno(ne) & Cukr(0,1,2) & Káva(\leq 1-2) >+< Obezrisk(ano)
3	79	1.106	Vzdělání(\leq vyučen) & Víno(ne) & Cukr(0,1,2) >+< Obezrisk(ano)
4	81	1.080	Vzdělání(\leq vyučen) & Víno(ne) & Cukr(0,1,2) & Káva(\leq 1-2) & Čaj(\leq 1-2) >+< Obezrisk(ano)
5	82	1.057	Vzdělání(\leq vyučen) & Víno(ne) & Cukr(0,1,2) & Čaj(\leq 1-2) >+< Obezrisk(ano)
6	88	1.028	Vzdělání(\leq vyučen) & Rodinný stav (ženatý) & Cukr(0,1) & Káva(\leq 1-2) >+< Obezrisk(ano)
7	170	1.020	Vzdělání(vyučen, středoškolské) & Rodinný stav (ženatý) & Cukr(0,1) & Čaj(\geq 1-2) >+< Obezrisk(ano)
8	87	1.017	Vzdělání(\leq vyučen) & Rodinný stav (ženatý) & Cukr(0,1) >+< Obezrisk(ano)
9	223	0.920	Rodinný stav (ženatý) & Víno(ne) & Cukr(0,1) & Káva(\leq 1-2) >+< Obezrisk(ano)
10	125	0.916	Vzdělání(vyučen) & Rodinný stav (ženatý) & Cukr(0,1,2) >+< Obezrisk(ano)
11	112	0.883	Vzdělání(vyučen) & Cukr(1,2) >+< Obezrisk(ano)

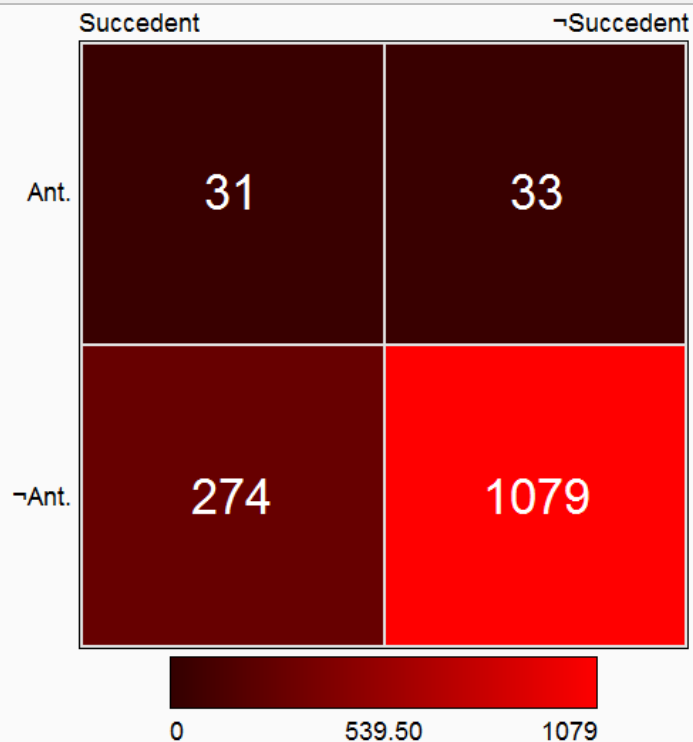
Detail nejsilnějšího pravidla

Antecedent: Vzdělání(základní, vyučen) & Rodinný stav (ženaný) & Víno(ne) & Cukr(0, 1, 2)

Succedent: Obezisk(ano)

Condition: (empty)

TEXT | DATA | FREQUENCIES | 9-FOLD TABLE | PIE | BAYES



$$\frac{31}{31 + 33} = (1 + 1.25) \frac{31 + 274}{31 + 33 + 274 + 1079}$$

Pokud platí Ant, tak je relativní četnost rizika obezity o 125 % vyšší, než je relativní četnost rizika obezity mezi všemi pacienty.

$\Rightarrow_{0.5,30} a \sim^+_{0.5,30}$ - poznámky

Matice dat	S	$\neg S$
A	a	b
$\neg A$	c	d

$$\Rightarrow_{0.5,30} \frac{a}{a+b} \geq 0.5 \wedge a \geq 30$$

$$\sim^+_{0.5,30} \frac{a}{a+b} \geq (1+0.5) \frac{a+c}{a+b+c+d} \wedge a \geq 30$$

- $A \Rightarrow_{0.5,30} S / P$ platí právě když $A \wedge P \Rightarrow_{0.5,30} S$
- Není pravda, že $A \sim^+_{0.5,30} S / P$ platí právě když $A \wedge P \sim^+_{0.5,30} S$
- $\Rightarrow_{0.5,30}$ se týká pouze prvního řádku čtyřpolní tabulky
- $\sim^+_{0.5,30}$ se týká celé čtyřpolní tabulky

Kvízy – následuje kvízová otázka

- V průběhu většiny přednášek bude několik kvízových otázek.
- Odpovědi se vkládají přes ISIS pomocí vašeho notebooku nebo chytrého telefonu.
- Kvízové otázky souvisí s přednášenou látkou a jejich správným zodpovězením prokazujete, že rozumíte přednášené látce.
- K výsledkům kvízů bude přihlédnuto v závěrečné klasifikaci.
- Obdobné otázky budou součástí „ostrých“ testů .
- Pro odpověď na kvíz nepoužívejte žádný software.

Otázka 3

Předpokládejte, že je dána čtyřpolní tabulka:

KLIENT	Úvěr(splácí)	\neg Úvěr(splácí)
Vzdělání(základní) \wedge Stav(svobodný)	t	u
\neg (Vzdělání(základní) \wedge Stav(svobodný))	v	w

Který vzorec vyjadřuje následující tvrzení:

Relativní četnost klientů splňujících Úvěr(splácí) mezi klienty splňujícími $Vzdělání(základní) \wedge Stav(svobodný)$ je nejméně o 75% vyšší, než relativní četnost klientů splňujících Úvěr(splácí) v celé matici KLIENT a zároveň nejméně 100 klientů splňuje $Vzdělání(základní) \wedge Stav(svobodný)$ i Úvěr(splácí).

Asociační pravidla I

- Komentář ke kvízovým otázkám
- Využívání metod data mining
- Asociační pravidla – příklad
- Podmíněná asociační pravidla – příklad
- Asociační pravidla – jiný příklad
- **Asociační pravidla – přehledný popis**
- Poznámka k seminárním pracím
- Rekapitulace
- Doporučení pro zadání 4ft-kvantifikátoru

Asociační pravidla – přehledný popis

- Matice dat
 - Booleovský atribut
 - Asociační pravidlo a podmíněné asociační pravidlo
 - Procedury ASSOC a 4ft-Miner
 - Asociační pravidla a neúplná informace
 - Asociační pravidla, nákupní košík, algoritmus apriori a jejich vztah k proceduře ASSOC -
- viz 1. a 2. přednášku
- viz dále
- viz další přednášky

Asociační pravidlo a podmíněné asociční pravidlo

- Formální zápis
- Čtyřpolní tabulka
- Míry zajímavosti asocičního pravidla
- 4ft-kvantifikátory
- Asociační pravidlo je pravdivé v matici dat
- Příklady 4ft-kvantifikátorů
- Podmíněné asociční pravidlo je pravdivé v matici dat

Formální zápis

Asociační pravidlo:

$$\varphi \approx \psi$$

Podmíněné associační pravidlo:

$$\varphi \approx \psi / \chi$$

- φ , ψ a χ jsou booleovské atributy
 - φ se nazývá antecedent
 - ψ se nazývá sukcedent (konsekvent)
 - χ se nazývá podmínka
- \approx je 4ft-kvantifikátor, vyjadřuje vztah φ a ψ
- φ , ψ a χ se souhrnně nazývají cedenty
- obecnější definice než ta zavedená v souvislosti s analýzou nákupního košíku

Čtyřpolní tabulka

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

\mathcal{M} - matice dat

φ, ψ - booleovské atributy

- Čtyřpolní tabulka φ a ψ v matici \mathcal{M} se značí $4ft(\varphi, \psi, \mathcal{M})$
- Platí $4ft(\varphi, \psi, \mathcal{M}) = \langle a, b, c, d \rangle$ kde
 - a - počet řádků splňujících φ i ψ
 - b - počet řádků splňujících φ a nespňujících ψ
 - c - počet řádků nespňujících φ a splňujících ψ
 - d - počet řádků nespňujících φ ani ψ

Míry zajímavosti asociačního pravidla

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

\mathcal{M} - matice dat

φ, ψ - booleovské atributy

- Spolehlivost $\frac{a}{a+b}$
- Podpora $\frac{a}{a+b+c+d}$
- Jaccardova míra $\frac{a}{a+b+c}$
- Přesnost $\frac{a+d}{a+b+c+d}$
- Lift $\frac{a(a+b+c+d)}{(a+b)(a+c)}$
- AA-míra $\frac{a(a+b+c+d)}{(a+b)(a+c)} - 1$
- BA-míra $1 - \frac{a(a+b+c+d)}{(a+b)(a+c)}$
- Další viz [1]

Spolehlivost asociačního pravidla $\varphi \approx \psi$

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

Spolehlivost: $\frac{a}{a+b}$

- Další názvy: konfidence, p-implikace
- Relativní četnost ψ pokud platí φ
- $100 \frac{a}{a+b}$ = procento řádků splňujících ψ z počtu řádků splňujících φ

Podpora asociačního pravidla $\varphi \approx \psi$

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

Podpora: $\frac{a}{a + b + c + d}$

- Relativní četnost řádků splňujících φ i ψ v celé matici
- $100 \frac{a}{a + b + c + d}$ = procento řádků splňujících ψ i φ

Jaccardova míra asociačního pravidla $\varphi \approx \psi$ (1)

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

Jaccardova míra: $\frac{a}{a + b + c}$

- Další název: dvojitá p-implikace
- Relativní četnost řádků splňujících φ i ψ mezi řádky splňujícími φ nebo ψ
- $100 \frac{a}{a + b + c}$ = procento řádků splňujících ψ i φ z řádků splňujících φ nebo ψ
- $\frac{a}{a + b + c}$ = konfidence pravidla $\varphi \vee \psi \approx \varphi \wedge \psi$, viz následující slide

Jaccardova míra asociačního pravidla $\varphi \approx \psi$ (2)

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

Jaccardova míra: $\frac{a}{a+b+c}$

$$\varphi \vee \psi \approx \varphi \wedge \psi$$

\mathcal{M}	$\varphi \wedge \psi$	$\neg(\varphi \wedge \psi) = \neg\varphi \vee \neg\psi$
$\varphi \vee \psi$	a	$b+c$
$\neg(\varphi \vee \psi) = \neg\varphi \wedge \neg\psi$	0	d

Spolehlivost: $\frac{a}{a+b+c}$

Přesnost asociačního pravidla $\varphi \approx \psi$

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

Přesnost:

$$\frac{a + d}{a + b + c + d}$$

- Další název: p-ekvivalence
- Relativní četnost řádků pro které mají φ i ψ stejnou hodnotu (pravda nebo nepravda)
- $100 \frac{a + d}{a + b + c + d}$ = procento řádků matice, pro které má φ i ψ stejnou hodnotu

Lift asociačního pravidla $\varphi \approx \psi$

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

$$\text{Lift : } \frac{a(a+b+c+d)}{(a+b)(a+c)}$$

$$\circ \quad \frac{a(a+b+c+d)}{(a+b)(a+c)} = \frac{\frac{a}{a+b}}{\frac{a+c}{a+b+c+d}} = \frac{\text{relativní četnost } \psi \text{ pokud platí } \varphi}{\text{relativní četnost } \psi \text{ v celé matici dat}}$$

- Pokud lift > 1, pak platnost φ zvyšuje relativní četnost ψ
- Pokud lift < 1, pak platnost φ snižuje relativní četnost ψ
- Pokud lift = 1, pak platnost φ nemá vliv na relativní četnost ψ

AA míra asociačního pravidla $\varphi \approx \psi$

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

$$\text{AA-míra : } \frac{a(a+b+c+d)}{(a+b)(a+c)} - 1$$

- AA-míra = lift – 1, neboli

$$\text{AA-míra} = \frac{\text{relativní četnost } \psi \text{ pokud platí } \varphi}{\text{relativní četnost } \psi \text{ v celé matici dat}} - 1, \text{ tedy}$$

relativní četnost ψ pokud platí φ = (AA-míra + 1) relativní četnost ψ v celé matici dat

- Pokud lift > 1, pak 100*AA-míra udává, o kolik procent vzroste relativní četnost ψ pokud platí φ oproti četnosti ψ v celé matici dat

BA míra asociačního pravidla $\varphi \approx \psi$

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

$$\text{BA-míra : } 1 - \frac{a(a+b+c+d)}{(a+b)(a+c)}$$

- BA-míra = 1 – lift, neboli

$$\text{BA-míra} = 1 - \frac{\text{relativní četnost } \psi \text{ pokud platí } \varphi}{\text{relativní četnost } \psi \text{ v celé matici dat}}, \text{ tedy}$$

$$\text{relativní četnost } \psi \text{ pokud platí } \varphi = (1 - \text{BA-míra}) \text{ relativní četnost } \psi \text{ v celé matici dat}$$

- Pokud lift < 1, pak 100*BA-míra udává, o kolik procent klesne relativní četnost ψ pokud platí φ oproti četnosti ψ v celé matici dat

4ft-kvantifikátory

- Symbol \approx , součást asociačního pravidla – výrazu $\varphi \approx \psi$, definuje vztah φ a ψ
- Každému 4ft-kvantifikátoru je přiřazena podmínka týkající se čtyřpolních tabulek $\langle a,b,c,d \rangle$
- Různé typy podmínek:
 - *míra zajímavosti* $\varphi \approx \psi \geq \text{parametr}$
 - testy hypotéz
 - jednoduché podmínky na frekvence a,b,c,d
- Podmínku přiřazenou \approx chápeme jako $\{0,1\}$ -hodnotovou funkci $F_{\approx}(a,b,c,d)$
- Většinou píšeme pouze $\approx(a,b,c,d)$ místo $F_{\approx}(a,b,c,d)$
- Funkce $F_{\approx}(a,b,c,d)$ se nazývá asociovaná funkce 4ft-kvantifikátoru, používá se pro definici pravdivosti asociačního pravidla

Asociační pravidlo je pravdivé v matici dat

- Asociační pravidlo $\varphi \approx \psi$ je pravdivé v matici dat \mathcal{M} pokud platí $F_{\approx}(a,b,c,d) = 1$ kde $\langle a,b,c,d \rangle = 4ft(\varphi,\psi,\mathcal{M})$

$4ft(\varphi,\psi,\mathcal{M}) =$

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

- Asociační pravidlo $\varphi \approx \psi$ je pravdivé v matici dat \mathcal{M} pokud je v této matici dat pro $4ft(\varphi,\psi,\mathcal{M})$ splněna podmínka přiřazená 4ft-kvantifikátoru \approx .

Podmíněné asociační pravidlo je pravdivé v matici dat

- Asociační pravidlo $\varphi \approx \psi / \chi$ je pravdivé v matici dat \mathcal{M} , pokud je asociační pravidlo $\varphi \approx \psi$ pravdivé v matici dat \mathcal{M} / χ
- Matice \mathcal{M} / χ vznikne z matice dat \mathcal{M} vynecháním všech řádků které nesplňují χ
- $\varphi \approx \psi / \chi$ je pravdivé v matici dat \mathcal{M} pokud platí $F_{\approx}(a,b,c,d) = 1$ kde $\langle a,b,c,d \rangle = 4ft(\varphi,\psi,\mathcal{M} / \chi)$

$$4ft(\varphi,\psi,\mathcal{M} / \chi) = \begin{array}{c|c|c} \mathcal{M} / \chi & \psi & \neg\psi \\ \hline \varphi & a & b \\ \hline \neg\varphi & c & d \end{array}$$

Fundovaná implikace

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

$$\Rightarrow_{p, \text{Base}}$$

$$\frac{a}{a+b} \geq p \wedge a \geq \text{Base}$$

Platnost $\varphi \Rightarrow_{p, \text{Base}} \psi$ znamená:

nejméně $100p$ % z řádků \mathcal{M} splňujících φ splňuje i ψ

a zároveň

nejméně Base řádků splňuje jak φ tak i ψ .

Fundovaná implikace – příklad

HotelPlusExterni	DHodnocení(průměr)	\neg DHodnocení(průměr)
HMesto(Jihlava) \wedge MObloha(slunečno)	34	8
\neg HMesto(Jihlava) \wedge MObloha(slunečno)	916	1042

$$\frac{34}{34 + 8} = 0.81$$

$HMesto(Jihlava) \wedge MObloha(slunečno) \Rightarrow_{0.81, 34} DHodnocení(průměr)$

(Nejméně) 81 % pobytů hostů z Jihlavy začínajících za slunečného počasí je hodnoceno jako průměrné

a zároveň

je takových pobytů (nejméně) 34.

Poznámka: Podle kontextu lze výraz „nejméně“ vynechat, platí i dále.

Fundovaný AA-kvantifikátor

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

$$\sim_{p, \text{Base}}^+$$

$$\frac{a}{a+b} \geq (1+p) \frac{a+c}{a+b+c+d} \wedge a \geq \text{Base}$$

relativní četnost řádků splňujících ψ mezi řádky splňujícími φ

relativní četnost řádků splňujících ψ v celé matici \mathcal{M}

Platnost $\varphi \sim_{p, \text{Base}}^+ \psi$ znamená:

relativní četnost řádků splňujících ψ mezi řádky splňujícími φ je o 100p% vyšší, než relativní četnost řádků splňujících ψ v celé matici \mathcal{M}

a zároveň

nejméně Base řádků splňuje jak φ tak i ψ .

Fundovaný AA-kvantifikátor – příklad

HotelPlusExterni	DUbytování(vyšší)	¬ DUbytování(vyšší)
HStat(ČR) ∧ MObloha(zataženo)	150	156
¬HStat(ČR) ∧ MObloha(zataženo)	549	1145

$$\frac{150}{150+156} = 0.49$$

$$\frac{150+549}{150+156+549+1145} = 0.35$$

$$0.49 = (1 + 0.4) * 0.35$$

HStat(ČR) ∧ MObloha(slunečno) $\Rightarrow_{0.81, 34}$ DUbytování(vyšší)

Relativní četnost pobytů, u nichž je úroveň ubytování hodnocena jako „vyšší“, mezi pobyty hostů z ČR začínajících se zataženou oblohou, je (nejméně) o 40 % vyšší, než relativní četnost pobytů, u nichž je úroveň ubytování hodnocena jako „vyšší“ mezi všemi pobyty a zároveň

je pobytů hostů z ČR začínajících se zataženou oblohou, u nichž je úroveň ubytování hodnocena jako „vyšší“ nejméně 150.

Další příklady 4ft-kvantifikátorů

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

Dvojitá fundovaná implikace: $\Leftrightarrow_{p, \text{Base}}$ $\frac{a}{a+b+c} \geq p \wedge a \geq \text{Base}$

Fundovaná ekvivalence: $\equiv_{p, \text{Base}}$ $\frac{a+d}{a+b+c+d} \geq p \wedge a \geq \text{Base}$

Podpora – support $\rightarrow_{p, s}$ $\frac{a}{a+b} \geq p \wedge \frac{a}{a+b+c+d} \geq s$

Další informace: viz [4ft_Analyticke_otazky.pdf](#) a další přednášky

Kvízy – následuje kvízová otázka

- V průběhu většiny přednášek bude několik kvízových otázek.
- Odpovědi se vkládají přes ISIS pomocí vašeho notebooku nebo chytrého telefonu.
- Kvízové otázky souvisí s přednášenou látkou a jejich správným zodpovězením prokazujete, že rozumíte přednášené látce.
- K výsledkům kvízů bude přihlédnuto v závěrečné klasifikaci.
- Obdobné otázky budou součástí „ostrých“ testů .
- Pro odpověď na kvíz nepoužívejte žádný software.

Otázka 4

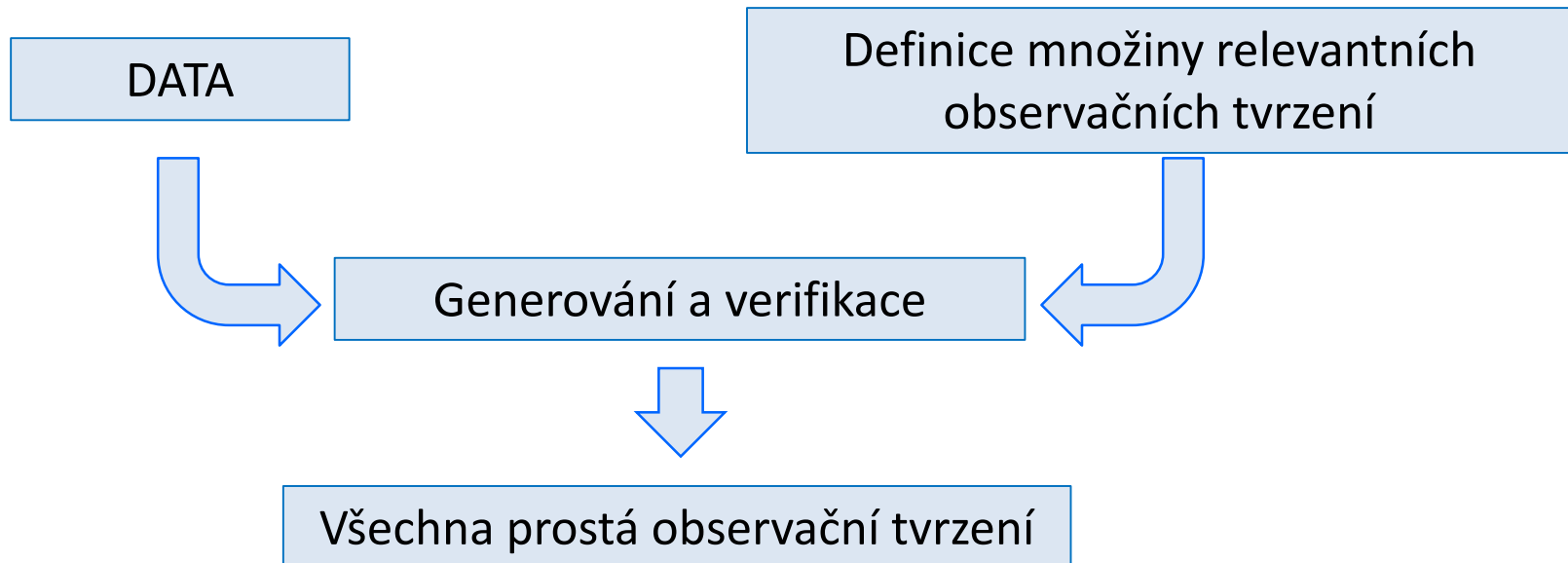
Předpokládejte, že je dána čtyřpolní tabulka:

KLIENT	Úvěr(splácí)	\neg Úvěr(splácí)
Vzdělání(základní) \wedge Stav(svobodný)	p	q
\neg (Vzdělání(základní) \wedge Stav(svobodný))	r	s

Který vzorec vyjadřuje následující tvrzení:

Pro nejméně 75% klientů platí, že $\text{Vzdělání(základní)} \wedge \text{Stav(svobodný)}$ a Úvěr(splácí) mají stejnou hodnotu (true nebo false) a zároveň nejméně 5% klientů splňuje $\text{Vzdělání(základní)} \wedge \text{Stav(svobodný)}$ i Úvěr(splácí) .

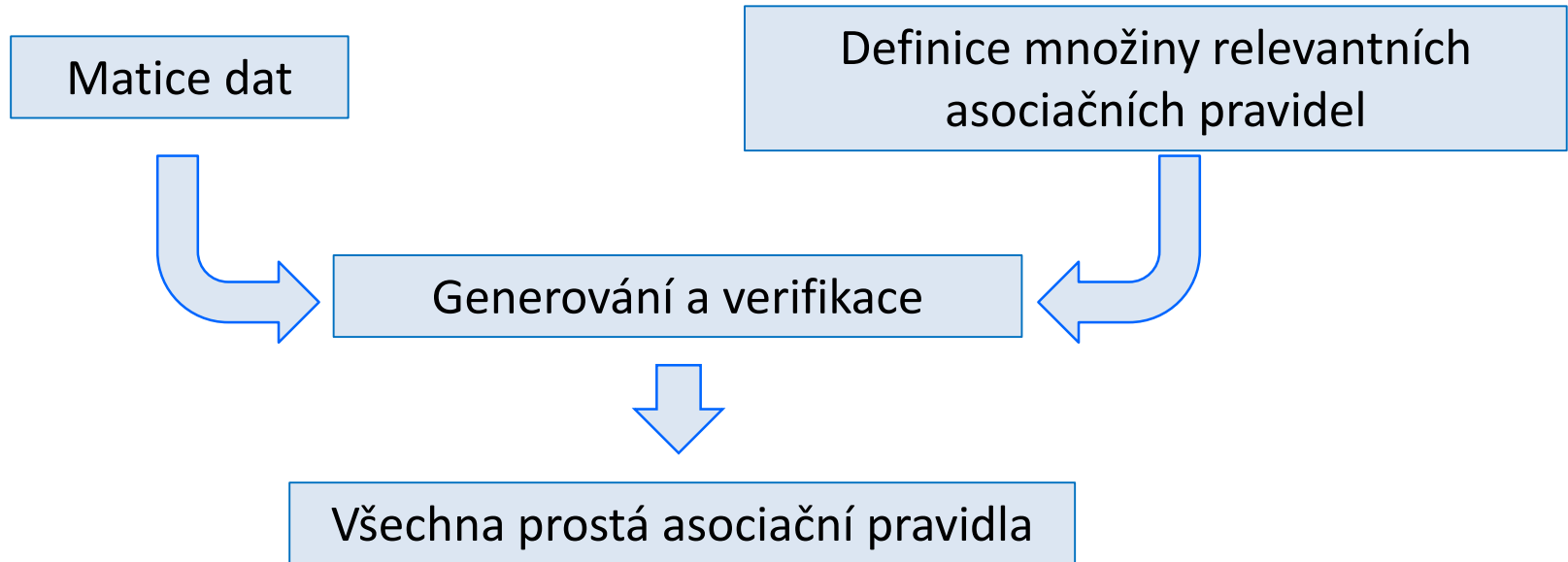
GUHA procedura



Prosté tvrzení: relevantní + pravdivé + nejkratší možné

Obvykle generováno 10^5 a více relevantních observačních tvrzení (= relevantních otázek)

GUHA procedura ASSOC

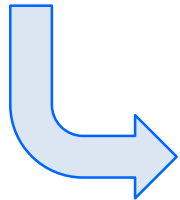


Prosté asociační pravidlo: relevantní + pravdivé + nejkratší možné

4ft-Miner – rozšířená procedura ASSOC

The screenshot shows the 4ft-Miner interface with three main panels: ANTECEDENT, QUANTIFIERS, and SUCCEDENT. The ANTECEDENT panel lists rules like HostBydliště, HCizinec_b, HMesto, and HStat. The QUANTIFIERS panel shows parameters like BASE p=80 Abs. and PIM p=0.700. The SUCCEDENT panel lists rules like Pobyt, PNoci_enum_m, PNoci_exp, PDenTydne, Meteo, and MObloha. A CONDITION panel at the bottom right shows Pobyt and PTypPobytu. The interface also displays generation information (Status: Solved, 2 run(s); Mode: Standard) and task parameters (Handling of missing values: Ignore X-categories; Prime rule test for implications enabled: No; Include succedent extensions of 100% implications: Yes).

Matice dat



Generování a verifikace



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

Podrobnosti: Kapitoly 4.3 a 7 v [1], <http://lispminer.vse.cz/wiki/doku.php?id=mft:start>

Poznámka k seminárním pracím



4ft-Miner



CF-Miner



KL-Miner

- Vzájemné doplňování procedur
- Využití jednoho zadání cedentů a dílčích cedentů ve více procedurách

Dotazník

- » DHodnoceni(3 categories)
- » DPersonal_ef3(3 categories)
- » DStrava_ef3(3 categories)
- » DUbytovani_ef3(3 categories)
- » DZabava_ef3(3 categories)

Pobyt

- » PNoci_enum_m (seq), 1 - 2
- » PDenTydne (subset), 1 - 1
- » POsob (seq), 1 - 3
- » POsobonoci_ef5 (seq), 1 - 2

Host

- » HPohlavi (subset), 1 - 1
- » HVek_exp (seq), 1 - 2

Bydliště

- » HCizinec_b (subset), 1 - 1
- » HMesto (subset), 1 - 1
- » HStat (subset), 1 - 1

Meteo

- » MObloha (subset), 1 - 1
- » MTeplota_exp (seq), 1 - 2

Con, 1 - 4

- B, pos
- B, pos
- B, pos
- B, pos

Con, 0 - 2

- B, pos
- B, pos

Con, 0 - 1

- B, pos
- B, pos
- B, pos

Con, 0 - 2

- B, pos
- B, pos

Asociační pravidla I

- Komentář ke kvízovým otázkám
- Využívání metod data mining
- Asociační pravidla – příklad
- Podmíněná asociační pravidla – příklad
- Asociační pravidla – jiný příklad
- Asociační pravidla – přehledný popis
- Poznámka k seminárním pracím
- **Rekapitulace**
- Doporučení pro zadání 4ft-kvantifikátoru - viz *4ft_Analyticke_otazky.pdf*

Formální zápis

Asociační pravidlo:

$$\varphi \approx \psi$$

Podmíněné associační pravidlo:

$$\varphi \approx \psi / \chi$$

- φ , ψ a χ jsou booleovské atributy
 - φ se nazývá antecedent
 - ψ se nazývá sukcedent (konsekvent)
 - χ se nazývá podmínka
- \approx je 4ft-kvantifikátor, vyjadřuje vztah φ a ψ
- φ , ψ a χ se souhrnně nazývají cedenty
- obecnější definice než ta zavedená v souvislosti s analýzou nákupního košíku

Čtyřpolní tabulka

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

\mathcal{M} - matice dat

φ, ψ - booleovské atributy

- Čtyřpolní tabulka φ a ψ v matici \mathcal{M} se značí $4ft(\varphi, \psi, \mathcal{M})$
- Platí $4ft(\varphi, \psi, \mathcal{M}) = \langle a, b, c, d \rangle$ kde
 - a - počet řádků splňujících φ i ψ
 - b - počet řádků splňujících φ a nespňujících ψ
 - c - počet řádků nespňujících φ a splňujících ψ
 - d - počet řádků nespňujících φ ani ψ

Míry zajímavosti asociačního pravidla

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

\mathcal{M} - matice dat

φ, ψ - booleovské atributy

- Spolehlivost $\frac{a}{a+b}$
- Podpora $\frac{a}{a+b+c+d}$
- Jaccardova míra $\frac{a}{a+b+c}$
- Přesnost $\frac{a+d}{a+b+c+d}$
- Lift $\frac{a(a+b+c+d)}{(a+b)(a+c)}$
- AA-míra $\frac{a(a+b+c+d)}{(a+b)(a+c)} - 1$
- BA-míra $1 - \frac{a(a+b+c+d)}{(a+b)(a+c)}$
- Další viz [1]

4ft-kvantifikátory

- Symbol \approx , součást asociačního pravidla – výrazu $\varphi \approx \psi$, definuje vztah φ a ψ
- Každému 4ft-kvantifikátoru je přiřazena podmínka týkající se čtyřpolních tabulek $\langle a, b, c, d \rangle$
- Různé typy podmínek:
 - *míra zajímavosti* $\varphi \approx \psi \geq \text{parametr}$
 - testy hypotéz
 - jednoduché podmínky na frekvence a, b, c, d
- Podmínku přiřazenou \approx chápeme jako $\{0,1\}$ -hodnotovou funkci $F_{\approx}(a, b, c, d)$
- Většinou píšeme pouze $\approx(a, b, c, d)$ místo $F_{\approx}(a, b, c, d)$
- Funkce $F_{\approx}(a, b, c, d)$ se nazývá asociovaná funkce 4ft-kvantifikátoru, používá se pro definici pravdivosti asociačního pravidla

Asociační pravidlo je pravdivé v matici dat

- Asociační pravidlo $\varphi \approx \psi$ je pravdivé v matici dat \mathcal{M} pokud platí $F_{\approx}(a,b,c,d) = 1$ kde $\langle a,b,c,d \rangle = 4ft(\varphi,\psi,\mathcal{M})$

$$4ft(\varphi,\psi,\mathcal{M}) = \begin{array}{c|c|c} \mathcal{M} & \psi & \neg\psi \\ \hline \varphi & a & b \\ \hline \neg\varphi & c & d \end{array}$$

- Asociační pravidlo $\varphi \approx \psi$ je pravdivé v matici dat \mathcal{M} pokud je v této matici dat pro $4ft(\varphi,\psi,\mathcal{M})$ splněna podmínka přiřazená 4ft-kvantifikátoru \approx .

Podmíněné asociační pravidlo je pravdivé v matici dat

- Asociační pravidlo $\varphi \approx \psi / \chi$ je pravdivé v matici dat \mathcal{M} , pokud je asociační pravidlo $\varphi \approx \psi$ pravdivé v matici dat \mathcal{M} / χ
- Matice \mathcal{M} / χ vznikne z matice dat \mathcal{M} vynecháním všech řádků které nesplňují χ
- $\varphi \approx \psi / \chi$ je pravdivé v matici dat \mathcal{M} pokud platí $F_{\approx}(a,b,c,d) = 1$ kde $\langle a,b,c,d \rangle = 4ft(\varphi,\psi,\mathcal{M} / \chi)$

$$4ft(\varphi,\psi,\mathcal{M} / \chi) =$$

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

Asociační pravidla I

- Komentář ke kvízovým otázkám
- Využívání metod data mining
- Asociační pravidla – příklad
- Podmíněná asociační pravidla – příklad
- Asociační pravidla – jiný příklad
- Asociační pravidla – přehledný popis
- Poznámka k seminárním pracím
- Rekapitulace
- Doporučení pro zadání 4ft-kvantifikátoru - viz [4ft_Analyticke_otazky.pdf](#)