

Tato prezentace je součástí wiki-prezentace [Metoda GUHA, LISp-Miner a typové úlohy](#)

Je dostupná z [této adresy](#)

Verze 15. 2. 2020

Typ úlohy: Násobné snížení konfidence dodatečnou podmínkou

Data: [Hotel](#)

Problém: *Snížení relativní četnosti extrémních hodnocení dodatečnou podmínkou*

Jan Rauch

Katedra informačního a znalostního inženýrství

Vysoká škola ekonomická v Praze

Snížení relativní četnosti extrémních hodnocení dodatečnou podmínkou

- Motivace
- Princip
- SD4ft-Miner - příklad zadání parametrů
- SD4ft-kvantifikátor
- Second set = dodatečná podmínka
- Přehled výsledků
- Nejsilnější vztah - detail
- Nejsilnější vztah - detail, poznámky

Motivace

DHodnoceni

	nespokojen	průměr	spokojen
ČR	28	44	28
Německo	29	53	18
Polsko	45	39	15
Rakousko	25	53	22
Slovensko	37	41	23

DPersonal

	nižší	průměr	vyšší
ČR	32	31	38
Německo	35	37	29
Polsko	51	31	18
Rakousko	30	37	33
Slovensko	40	32	28

DStrava

	nižší	průměr	vyšší
ČR	31	31	39
Německo	33	37	29
Polsko	44	32	24
Rakousko	32	35	34
Slovensko	43	25	32

DUbytování

	nižší	průměr	vyšší
ČR	31	30	39
Německo	35	36	29
Polsko	49	24	27
Rakousko	30	36	33
Slovensko	38	28	34

DZabava

	nižší	průměr	vyšší
ČR	32	36	32
Německo	35	30	35
Polsko	45	32	23
Rakousko	28	33	39
Slovensko	40	30	30

Relativní četnost nespokojených hostů z ČR je 28%.
Otázka je, jaké dodatečné podmínky sníží tuto relativní četnost nespokojených hostů alespoň o jednu čtvrtinu.

Analogické otázky:

Jaké dodatečné podmínky sníží relativní četnost extrémních hodnocení u jednotlivých států alespoň o jednu čtvrtinu.

Extrémní hodnocení jsou v krajních sloupcích tabulek, jsou různá od průměrných hodnocení

Princip (1)

DHodnoceni

	nespokojen	průměr	spokojen
ČR	28	44	28
Německo	29	53	18
Polsko	45	39	15
Rakousko	25	53	22
Slovensko	37	41	23

Relativní četnost nespokojených hostů z ČR je 28%. Otázka je, jaké dodatečné podmínky sníží tuto relativní četnost nespokojených hostů alespoň o jednu čtvrtinu.

Hotel	DHodnocení(nespokojen)	¬DHodnocení(nespokojen)
HStat(ČR)	a_1	b_1
¬HStat(ČR)	c_1	d_1

Hotel	DHodnocení(nespokojen)	¬DHodnocení(nespokojen)
HStat(ČR) ∧ Podmínka	a_2	b_2
¬(HStat(ČR) ∧ Podmínka)	c_2	d_2

Chceme: $\frac{a_2}{a_2+b_2} \leq \left(1 - \frac{1}{4}\right) \frac{a_1}{a_1+b_1}$, čili $\frac{\frac{a_1}{a_1+b_1}}{\frac{a_2}{a_2+b_2}} \geq 1.33$

Princip (2)

Místo pravidla $HStat(\check{C}R) \approx DHodnocení(nespokojen)$ se čtyřpolní tabulkou

Hotel	DHodnocení(nespokojen)	\neg DHodnocení(nespokojen)
HStat($\check{C}R$)	a_1	b_1
\neg HStat($\check{C}R$)	c_1	d_1

pracujeme s [podmíněným pravidlem](#) $True \approx DHodnocení(nespokojen) / HStat(\check{C}R)$ se čtyřpolní tabulkou

Hotel / HStat($\check{C}R$)	DHodnocení(nespokojen)	\neg DHodnocení(nespokojen)
<i>True</i>	a_1	b_1
$\neg True$	0	0

Zde *True* je [identicky pravdivý booleovský atribut](#).

Princip (3)

Místo pravidla $HStat(\check{C}R) \wedge Podmínka \approx DHodnocení(nespokojen)$ se čtyřpolní tabulkou

Hotel	DHodnocení(nespokojen)	$\neg DHodnocení(nespokojen)$
$HStat(\check{C}R) \wedge Podmínka$	a_2	b_2
$\neg(HStat(\check{C}R) \wedge Podmínka)$	c_2	d_2

pracujeme s podmíněným pravidlem $True \approx DHodnocení(nespokojen) / HStat(\check{C}R) \wedge Podmínka$ se čtyřpolní tabulkou

Hotel / $HStat(\check{C}R) \wedge Podmínka$	DHodnocení(nespokojen)	$\neg DHodnocení(nespokojen)$
$True$	a_2	b_2
$\neg True$	0	0

Zde $True$ je identický pravdivý booleovský atribut.

SD4ft-Miner - příklad zadání parametrů

The screenshot displays the SD4ft-Miner configuration interface with several key sections and annotations:

- ANTECEDENT:** Contains "Prázdný antecedent" (Empty antecedent) with a constraint of "Con, 0 - 0". A blue box highlights "Prázdný antecedent".
- QUANTIFIERS:** A table with columns "Type", "Rel.", "Value", and "Units".

Type	Rel.	Value	Units
a (BASE) FirstSet	>=	50.00	Abs
a (BASE) SecondSet	>=	50.00	Abs
PIM RatioVal	>=	1.33	Abs

A blue box points to this section with the text "SD4ft-kvantifikátor, viz další slide".
- SUCCEDENT:** Contains "Dotazník" (Questionnaire) with a constraint of "Con, 1 - 1". It lists several attributes: "» DHodnoceni (cuts), 1 - 1", "» DPersonal_ef3 (cuts), 1 - 1", "» DStrava_ef3 (cuts), 1 - 1", "» DUbytovani_ef3 (cuts), 1 - 1", and "» DZabava_ef3 (cuts), 1 - 1". A blue box highlights "Extrémní hodnocení u atributů dotazníku" (Extreme evaluation of questionnaire attributes).
- (1) FIRST SET:** Contains "Default Partial Cedent" (Con, 1 - 1) and "» HStat (subset), 1 - 1" (B, pos). A blue box highlights "HStat(?)".
- (2) SECOND SET:** Contains "Host" (Con, 0 - 3), "» HPohlavi (subset), 1 - 1" (B, pos), "» HVek_ef3 (subset), 1 - 1" (B, pos), "» HVek_exp (subset), 1 - 1" (B, pos), and "Začátek pobytu" (Con, 0 - 3). A blue box points to this section with the text "Dodatečná podmínka, viz další slide +1" (Additional condition, see next slide +1).
- CONDITION:** Contains "Default Partial Cedent" (Con, 0 - 5).
- Task parameters:** Includes "Verification mode: The second set is treated as a subset specification to the first set (i.e. Set1 versus Set1 & Set2)", "Sets overlapping: Sets must differ in at least one row (i.e. partially overlapping sets are allowed)", and "Maximal number of hypotheses: 1000". A blue box points to this section with the text "Porovnává se matice Hotel / HStat(ČR) s maticí Hotel / HStat(ČR) ^ Podmínka" (Compares Hotel / HStat(ČR) matrix with Hotel / HStat(ČR) matrix ^ Condition).

SD4ft-kvantifikátor

Hotel / HStat(ČR)	DHodnocení(nespokojen)	¬DHodnocení(nespokojen)
<i>True</i>	a_1	b_1
<i>¬True</i>	0	0

Hotel / HStat(ČR) \wedge Podmínka	DHodnocení(nespokojen)	¬DHodnocení(nespokojen)
<i>True</i>	a_2	b_2
<i>¬True</i>	0	0

QUANTIFIERS			
Type	Rel.	Value	Units
a (BASE) FirstSet	>=	50.00	Abs
a (BASE) SecondSet	>=	50.00	Abs
PIM RatioVal	>=	1.33	Abs

$$a_1 \geq 50 \wedge a_2 \geq 50$$

SD4ft Statistical quantifier settings

Interest measure type: p-Implication
a/(a+b) >= p ... at least 100*p [%] of objects satisfying A satisfy also S

Relation: Greater than or equal

Threshold value: 1.33

Operation mode: Ratio of interest-measures
Test applied to the ratio of interest-measures computed separately from each frequency table

Parameters:



$$\text{Chceme: } \frac{a_2}{a_2+b_2} \leq \left(1 - \frac{1}{4}\right) \frac{a_1}{a_1+b_1}, \quad \text{čili } \frac{\frac{a_1}{a_1+b_1}}{\frac{a_2}{a_2+b_2}} \geq 1.33$$

Second set = dodatečná podmínka

(2) SECOND SET

Host Con, 0 - 3

- » HPohlavi (subset), 1 - 1 B, pos
- » HVek_ef3 (subset), 1 - 1 B, pos
- » HVek_exp (subset), 1 - 1 B, pos

Začátek pobytu Con. 0 - 3

Total length: 1 - 4

4ft Second set Partial cedent Settings

Basic parameters

Name: Host

Min. length: 0 Max. length: 3 Literals boolean operation type: Conjunction

Options

Allow only a consecutive sequence of literals in cedent (only neighbouring literals): No

Linked coefficients (all literals must have the same coefficient as in the first one): No

Literals Settings

Underlying attribute	Categories	X-cat	Coefficient type	Length	+/-	B/R	Class of equiv.
HPohlavi	2	No	Subsets	1 - 1	pos	Basic	-
HVek_ef3	3	No	Subsets	1 - 1	pos	Basic	Vek
HVek_exp	4	No	Subsets	1 - 1	pos	Basic	Vek

4ft Second set Partial cedent Settings

Basic parameters

Name: Začátek pobytu

Min. length: 0 Max. length: 3 Literals boolean operation type: Conjunction

Options

Allow only a consecutive sequence of literals in cedent (only neighbouring literals): No

Linked coefficients (all literals must have the same coefficient as in the first one): No

Literals Settings

Underlying attribute	Categories	X-cat	Coefficient type	Length	+/-	B/R	Class of equiv.
PDenTydne	7	No	Subsets	1 - 1	pos	Basic	-
PMesic	12	No	Subsets	1 - 1	pos	Basic	-
PRok	2	No	Subsets	1 - 1	pos	Basic	-

4ft Second set Partial cedent Settings

Basic parameters

Name: Cena pobytu

Min. length: 0 Max. length: 3 Literals boolean operation type: Conjunction

Options

Allow only a consecutive sequence of literals in cedent (only neighbouring literals): No

Linked coefficients (all literals must have the same coefficient as in the first one): No

Literals Settings

Underlying attribute	Categories	X-cat	Coefficient type	Length	+/-	B/R	Class of equiv.
PCenaCelkem	3	No	Subsets	1 - 1	pos	Basic	-
PCenaStrava	3	No	Subsets	1 - 1	pos	Basic	-
PCenaUbytovani	3	No	Subsets	1 - 1	pos	Basic	-

Přehled výsledků

Task run

Start: 16.2.2020 11:22:23

Total time: 0h 0m 5s

Number of verifications: 198850

Number of hypotheses: 12

Mode: Standard

Add group

Del group

Edit group

Actual group of hypotheses: All hypotheses

Hypotheses in group: 12

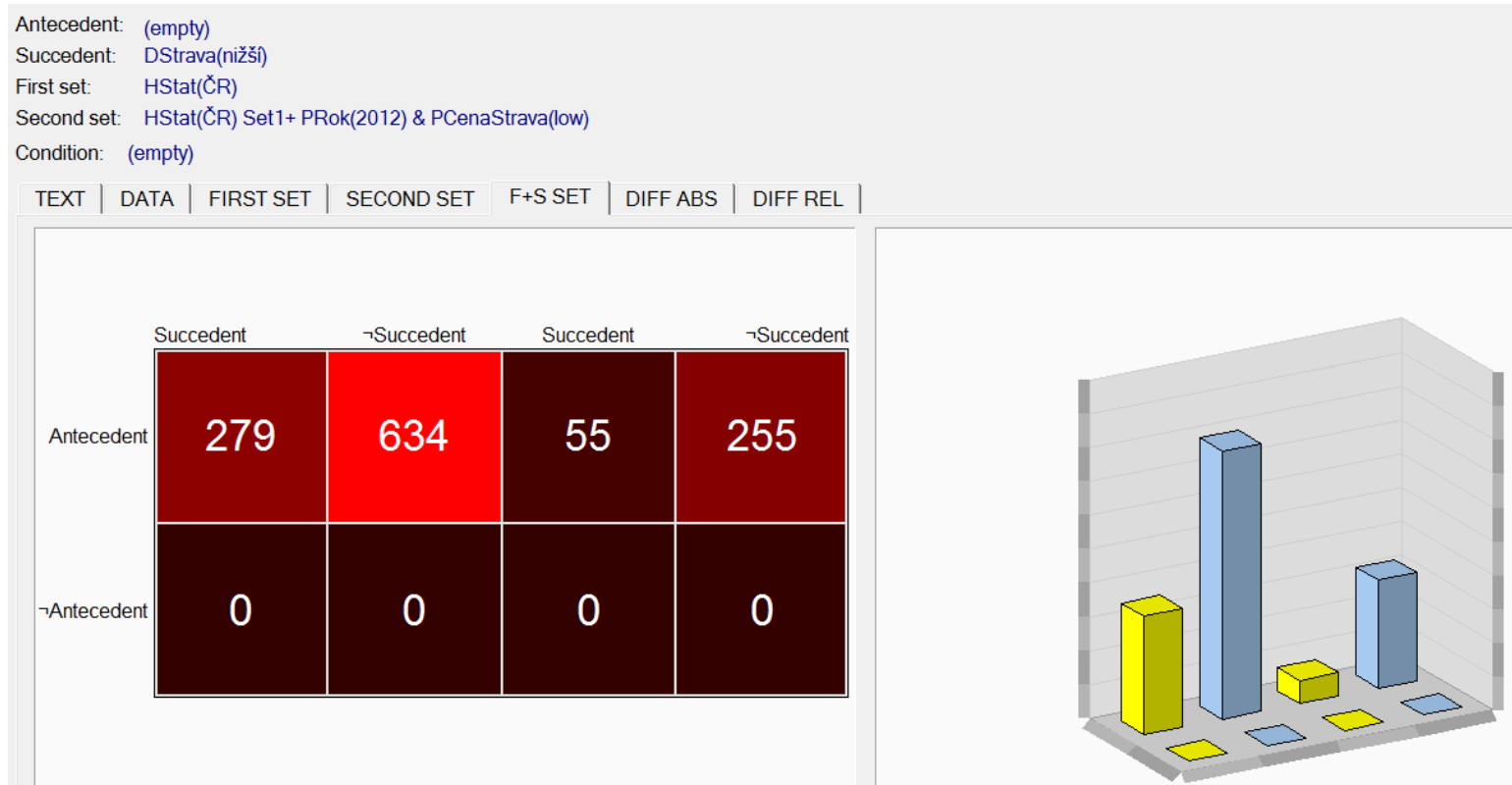
Shown hypotheses: 12

Highlighted: 0

Nr. Id R-Conf 1:Conf 2:Conf Hypothesis

1	6	1.722	0.306	0.177	(empty) >+< DStrava(nižší) : HStat(ČR) × Set1 & PRok(2012) & PCenaStrava(low)
2	4	1.655	0.306	0.185	(empty) >+< DStrava(nižší) : HStat(ČR) × Set1 & PRok(2012)
3	3	1.551	0.314	0.203	(empty) >+< DUbytovani(nižší) : HStat(ČR) × Set1 & PRok(2013) & PCenaStrava(low)
4	5	1.531	0.324	0.212	(empty) >+< DZabava(vyšší) : HStat(ČR) × Set1 & PRok(2012)
5	11	1.504	0.303	0.201	(empty) >+< DPersonal(nižší) : HStat(Rakousko) × Set1 & PRok(2013)
6	1	1.474	0.314	0.213	(empty) >+< DUbytovani(nižší) : HStat(ČR) × Set1 & PRok(2013)
7	12	1.461	0.388	0.266	(empty) >+< DZabava(vyšší) : HStat(Rakousko) × Set1 & PRok(2012)
8	7	1.436	0.324	0.226	(empty) >+< DZabava(vyšší) : HStat(ČR) × Set1 & PRok(2012) & PCenaStrava(low)
9	10	1.402	0.315	0.225	(empty) >+< DPersonal(nižší) : HStat(ČR) × Set1 & HPohlavi(žena) & PRok(2013)
10	9	1.384	0.314	0.227	(empty) >+< DUbytovani(nižší) : HStat(ČR) × Set1 & HPohlavi(muž) & PRok(2013)
11	2	1.376	0.315	0.229	(empty) >+< DPersonal(nižší) : HStat(ČR) × Set1 & PRok(2013) & PCenaStrava(low)
12	8	1.332	0.306	0.229	(empty) >+< DStrava(nižší) : HStat(ČR) × Set1 & PDenTydne(Pá) & PRok(2012)

Nejsilnější vztah - detail



- relativní četnost hodnocení DStrava(nižší) od hostů splňujících HStat(ČR) je $\frac{279}{279+634} = 0.31$
- relativní četnost hodnocení DStrava(nižší) od hostů splňujících $HStat(ČR) \wedge PRok(2012) \wedge PCenaStrava(low)$ je , $\frac{55}{55+255} = 0.18$, tedy o 42% nižší.

Nejsilnější vztah - detail, poznámky

Antecedent: (empty)
Succedent: DStrava(nižší)
First set: HStat(ČR)
Second set: HStat(ČR) Set1+ PRok(2012) & PCenaStrava(low)
Condition: (empty)

TEXT | DATA | FIRST SET | SECOND SET | F+S SET | DIFF ABS | DIFF REL

	Succedent	¬Succedent	Succedent	¬Succedent
Antecedent	279	634	55	255
¬Antecedent	0	0	0	0

Relativní četnost hodnocení DStrava(nižší) od hostů splňujících HStat(ČR) =

- konfidence podmíněného pravidla $True \approx DStrava(nižší)/HStat(ČR)$
- konfidence pravidla $HStat(ČR) \approx DStrava(nižší)$
- charakteristika PIM podmíněného pravidla $True \approx DStrava(nižší)/HStat(ČR)$
- charakteristika PIM pravidla $HStat(ČR) \approx DStrava(nižší)$
- $\frac{279}{279+634} = 0.31$.

Relativní četnost hodnocení DStrava(nižší) od hostů splňujících od hostů splňujících $HStat(ČR) \wedge PRok(2012) \wedge PCenaStrava(low) =$

- konfidence podmíněného pravidla $True \approx DStrava(nižší)/HStat(ČR) \wedge PRok(2012) \wedge PCenaStrava(low)$
- konfidence pravidla $HStat(ČR) \wedge PRok(2012) \wedge PCenaStrava(low) \approx DStrava(nižší)$
- charakteristika PIM podmíněného pravidla $True \approx DStrava(nižší)/HStat(ČR) \wedge PRok(2012) \wedge PCenaStrava(low)$
- charakteristika PIM pravidla $HStat(ČR) \wedge PRok(2012) \wedge PCenaStrava(low) \approx DStrava(nižší)$
- $\frac{55}{55+255} = 0.18$.