

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

Je dostupná z [této adresy](#)

Verze 26. 2. 2020

Typ úlohy: Násobné zvýšení konfidence dodatečnou podmínkou

Data: Hotel

Problém: *Zvýšení relativní četnosti extrémních hodnocení dodatečnou podmínkou*

Jan Rauch

Katedra informačního a znalostního inženýrství

Vysoká škola ekonomická v Praze

# Zvýšení 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 spokojených hostů z ČR je 28%.  
Otázka je, jaké dodatečné podmínky zvýší tuto relativní četnost spokojených hostů alespoň o jednu polovinu.

Analogické otázky:

Jaké dodatečné podmínky zvýší relativní četnost extrémních hodnocení u jednotlivých států alespoň o jednu polovinu.

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

Hotel	DHodnocení(spokojen)	¬DHodnocení(spokojen)
HStat(ČR)	$a_1$	$b_1$
¬HStat(ČR)	$c_1$	$d_1$

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

Chceme:  $\frac{a_2}{a_2+b_2} \geq \left(1 + \frac{1}{2}\right) \frac{a_1}{a_1+b_1}$ , čili  $\frac{\frac{a_1}{a_1+b_1}}{\frac{a_2}{a_2+b_2}} \leq 0.66$

## Princip (2)

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

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

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

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

Zde *True* je identicky pravdivý booleovský atribut.

## Princip (3)

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

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

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

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

Zde  $True$  je identicky pravdivý booleovský atribut.

# SD4ft-Miner - příklad zadání parametrů

**ANTECEDENT**

Host Con, 0 - 0

Prázdný antecedent

Total length: 0

**QUANTIFIERS**

Type	Rel.	Value	Units
a (BASE) FirstSet	>=	50.00	Abs
a (BASE) SecondSet	>=	50.00	Abs
PIM RatioVal	<=	0.66	Abs

Generation information

Status: Solved

Mode: Stand

SD4ft-kvantifikátor, viz další slide

**SUCCEEDENT**

Dotazník Con, 1 - 1

- » DHodnoceni (cuts), 1 - 1 B, pos
- » DPersonal\_ef3 (cuts), 1 - 1 B, pos
- » DStrava\_ef3 (cuts), 1 - 1 B, pos
- » DUbytovani\_ef3 (cuts), 1 - 1 B, pos
- » DZabava\_ef3 (cuts), 1 - 1 B, pos

Extrémní hodnocení u atributů dotazníku

Total length: 1

**(1) FIRST SET**

Default Partial Cedent Con, 1 - 1

- » HStat (subset), 1 - 1 B, pos

HStat(?)

Total length: 1

**(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 pobvtu Con, 0 - 3

Dodatečná podmínka, viz další slide +1

Total length: 1 - 4

**CONDITION**

Default Partial Cedent Con, 0 - 5

Total length: 0

Task parameters

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)

Maximal number of hypotheses: 1000

Porovnává se matice Hotel / HStat(ČR) s maticí Hotel / HStat(ČR) ^ Podmínka

# SD4ft-kvantifikátor

Hotel / HStat(ČR)	DHodnocení(spokojen)	¬DHodnocení(spokojen)
<i>True</i>	$a_1$	$b_1$
$\neg True$	$0$	$0$

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

QUANTIFIERS			
Type	Rel.	Value	Units
a (BASE) FirstSet	>=	50.00	Abs
a (BASE) SecondSet	>=	50.00	Abs
PIM RatioVal	<=	0.66	Abs

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

SD4ft Statistical quantifier settings

Interest measure type: p-implication  
 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

Chceme:  $\frac{a_2}{a_2+b_2} \geq \left(1 + \frac{1}{2}\right) \frac{a_1}{a_1+b_1}$ , čili  $\frac{\frac{a_1}{a_1+b_1}}{\frac{a_2}{a_2+b_2}} \leq 0.66$



# 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 (18 nejsilnějších z celkem 29)

Task run						
Start: 16.2.2020 20:36:02		Total time: 0h 0m 5s				
Number of verifications: 198850						
Number of hypotheses: 29		Mode: Standard				
<input type="button" value="Add group"/> <input type="button" value="Del group"/> <input type="button" value="Edit group"/>						
Actual group of hypotheses: All hypotheses						
Hypotheses in group: 29		Shown hypotheses: 29		Highlighted: 0		
Nr.	Id	R-Conf	1:Conf	2:Conf	Hypothesis	
1	19	0.583	0.331	0.569	(empty)	>>< DStrava( <i>nižší</i> ) : HStat(Německo) × Set1 & PRok(2013) & PCenaCelkem( <i>vyšší</i> )
2	21	0.583	0.331	0.569	(empty)	>>< DStrava( <i>nižší</i> ) : HStat(Německo) × Set1 & PRok(2013) & PCenaCelkem( <i>vyšší</i> ) & PCenaUbytovani( <i>high</i> )
3	18	0.583	0.331	0.569	(empty)	>>< DStrava( <i>nižší</i> ) : HStat(Německo) × Set1 & PRok(2013) & PCenaUbytovani( <i>high</i> )
4	6	0.596	0.279	0.469	(empty)	>>< DHodnoceni( <i>nespokojen</i> ) : HStat(ČR) × Set1 & PRok(2012) & PCenaCelkem( <i>vyšší</i> )
5	8	0.596	0.279	0.469	(empty)	>>< DHodnoceni( <i>nespokojen</i> ) : HStat(ČR) × Set1 & PRok(2012) & PCenaCelkem( <i>vyšší</i> ) & PCenaUbytovani( <i>high</i> )
6	4	0.596	0.279	0.469	(empty)	>>< DHodnoceni( <i>nespokojen</i> ) : HStat(ČR) × Set1 & PRok(2012) & PCenaUbytovani( <i>high</i> )
7	7	0.603	0.315	0.523	(empty)	>>< DPersonal( <i>nižší</i> ) : HStat(ČR) × Set1 & PRok(2012) & PCenaCelkem( <i>vyšší</i> )
8	9	0.603	0.315	0.523	(empty)	>>< DPersonal( <i>nižší</i> ) : HStat(ČR) × Set1 & PRok(2012) & PCenaCelkem( <i>vyšší</i> ) & PCenaUbytovani( <i>high</i> )
9	5	0.603	0.315	0.523	(empty)	>>< DPersonal( <i>nižší</i> ) : HStat(ČR) × Set1 & PRok(2012) & PCenaUbytovani( <i>high</i> )
10	2	0.617	0.279	0.452	(empty)	>>< DHodnoceni( <i>nespokojen</i> ) : HStat(ČR) × Set1 & PCenaCelkem( <i>vyšší</i> ) & PCenaStrava( <i>high</i> )
11	3	0.617	0.279	0.452	(empty)	>>< DHodnoceni( <i>nespokojen</i> ) : HStat(ČR) × Set1 & PCenaCelkem( <i>vyšší</i> ) & PCenaStrava( <i>high</i> ) & PCenaUbytovani( <i>high</i> )
12	1	0.617	0.279	0.452	(empty)	>>< DHodnoceni( <i>nespokojen</i> ) : HStat(ČR) × Set1 & PCenaStrava( <i>high</i> ) & PCenaUbytovani( <i>high</i> )
13	20	0.623	0.331	0.532	(empty)	>>< DStrava( <i>nižší</i> ) : HStat(Německo) × Set1 & PRok(2013) & PCenaStrava( <i>high</i> )
14	15	0.629	0.314	0.500	(empty)	>>< DUbytovani( <i>nižší</i> ) : HStat(ČR) × Set1 & HVek( <i>průměr</i> ) & PRok(2012)
15	25	0.631	0.318	0.505	(empty)	>>< DStrava( <i>nižší</i> ) : HStat(Rakousko) × Set1 & HPohlavi( <i>žena</i> ) & PRok(2013) & PCenaStrava( <i>low</i> )
16	24	0.634	0.305	0.481	(empty)	>>< DUbytovani( <i>nižší</i> ) : HStat(Rakousko) × Set1 & PRok(2012) & PCenaStrava( <i>low</i> )
17	29	0.639	0.305	0.477	(empty)	>>< DUbytovani( <i>nižší</i> ) : HStat(Rakousko) × Set1 & HVek( <i>od 28 do 60</i> ) & PRok(2012)
18	13	0.642	0.306	0.476	(empty)	>>< DStrava( <i>nižší</i> ) : HStat(ČR) × Set1 & PDenTydne( <i>So</i> ) & PRok(2013) & PCenaCelkem( <i>vyšší</i> )

## Nejsilnější vztah - detail

Antecedent: (empty)

Succedent: DStrava(nižší)

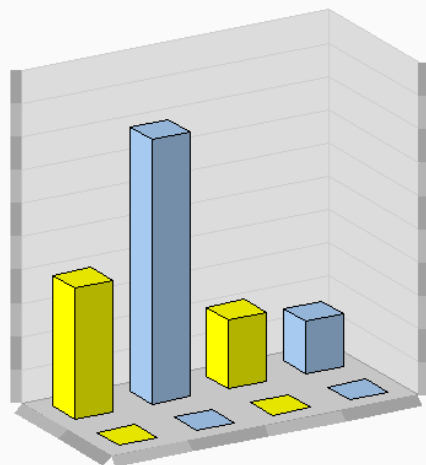
First set: HStat(Německo)

Second set: HStat(Německo) Set1+ PRok(2013) & PCenaCelkem(vyšší)

Condition: (empty)

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

	Succedent	¬Succedent	Succedent	¬Succedent
Antecedent	118	238	62	47
¬Antecedent	0	0	0	0



- relativní četnost hodnocení DStrava(nižší) od hostů splňujících HStat(Německo) je  $\frac{118}{118+238} = 0.33$
- relativní četnost hodnocení DStrava(nižší) od hostů splňujících  $HStat(Německo) \wedge PRok(2013) \wedge PCenaCelkem(vyšší)$  je  $\frac{62}{62+47} = 0.57$ , tedy o 73% vyšší

# Nejsilnější vztah - detail, poznámky

Antecedent: (empty)  
Succedent: DStrava(nižší)  
First set: HStat(Německo)  
Second set: HStat(Německo) Set1+ PRok(2013) & PCenaCelkem(vyšší)  
Condition: (empty)

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

	Succedent	¬Succedent	Succedent	¬Succedent
Antecedent	118	238	62	47
¬Antecedent	0	0	0	0

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

- konfidence podmíněného pravidla  $True \approx DStrava(nižší)/HStat(Německo)$
- konfidence pravidla  $HStat(Německo) \approx DStrava(nižší)$
- charakteristika PIM podmíněného pravidla  $True \approx DStrava(nižší)/HStat(Německo)$
- charakteristika PIM pravidla  $HStat(Německo) \approx DStrava(nižší)$
- $\frac{118}{118+238} = 0.33.$

Relativní četnost hodnocení DStrava(nižší) od hostů splňujících od hostů splňujících  $HStat(Německo) \wedge PRok(2013) \wedge PCenaCelkem(vyšší) =$

- konfidence podmíněného pravidla  $True \approx DStrava(nižší)/ HStat(Německo) \wedge PRok(2013) \wedge PCenaCelkem(vyšší)$
- konfidence pravidla  $HStat(Německo) \wedge PRok(2013) \wedge PCenaCelkem(vyšší) \approx DStrava(nižší)$
- charakteristika PIM podmíněného pravidla  $True \approx DStrava(nižší)/ HStat(Německo) \wedge PRok(2013) \wedge PCenaCelkem(vyšší)$
- charakteristika PIM pravidla  $HStat(Německo) \wedge PRok(2013) \wedge PCenaCelkem(vyšší) \approx DStrava(nižší)$
- $\frac{62}{62+47} = 0.57.$