Tato prezentace je součástí wiki-prezentace Metoda GUHA, LISp-Miner a typové úlohy
Je dostupná z této adresy
Verse 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: $\frac{a_2}{a_2+b_2} \geq (1+\frac{1}{2})\frac{a_1}{a_1+b_1}$, čili $\frac{\frac{a_1}{a_1+b_1}}{\frac{a_2}{a_2+b_2}} \leq 0.66$

Princip (2)

Místo pravidla HStat(ČR) ≈ DHodnocení(spokojen) se čtyřpolní tabulkou

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

pracujeme s podmíněným pravidlem *True* ≈ DHodnocení(spokojen) / HStat(ČR) se čtyřpolní tabulkou

Hotel / HStat(ČR)	DHodnocení(spokojen)	¬DHodnocení(spokojen)
True	a_1	b_1
¬*True*	*0*	*0*

.

Zde *True* je identicky pravdivý booleovský atribut.

Princip (3)

Místo pravidla HStat(ČR) $\wedge$ Podmínka $\approx$ DHodnocení(spokojen) se čtyřpolní tabulkou

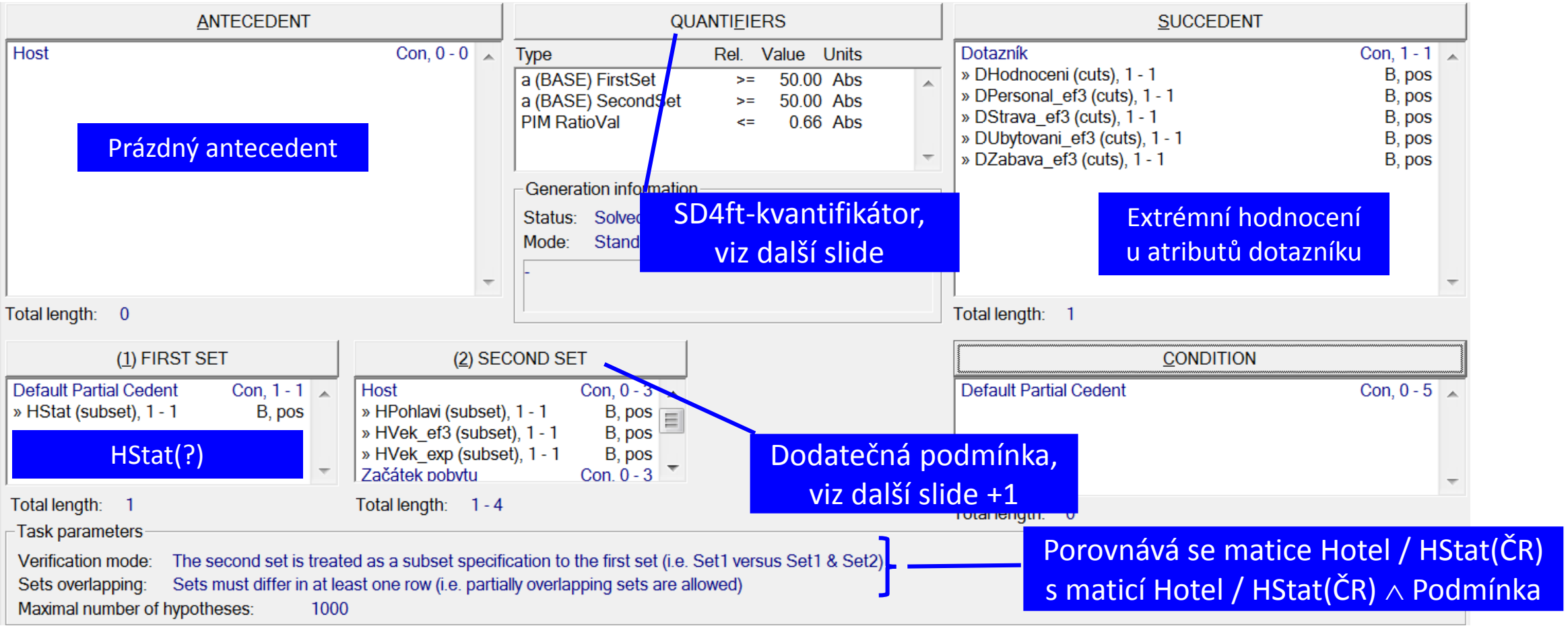
Hotel	DHodnocení(spokojen)	$\neg$DHodnocení(spokojen)
HStat(ČR) $\wedge$ Podmínka	a_2	b_2
$\neg$(HStat(ČR) $\wedge$ Podmínka)	c_2	d_2

pracujeme s podmíněným pravidlem *True* $\approx$ DHodnocení(spokojen) / HStat(ČR) $\wedge$ Podmínka se čtyřpolní tabulkou

Hotel / HStat(ČR) $\wedge$ 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



SD4ft-kvantifikátor

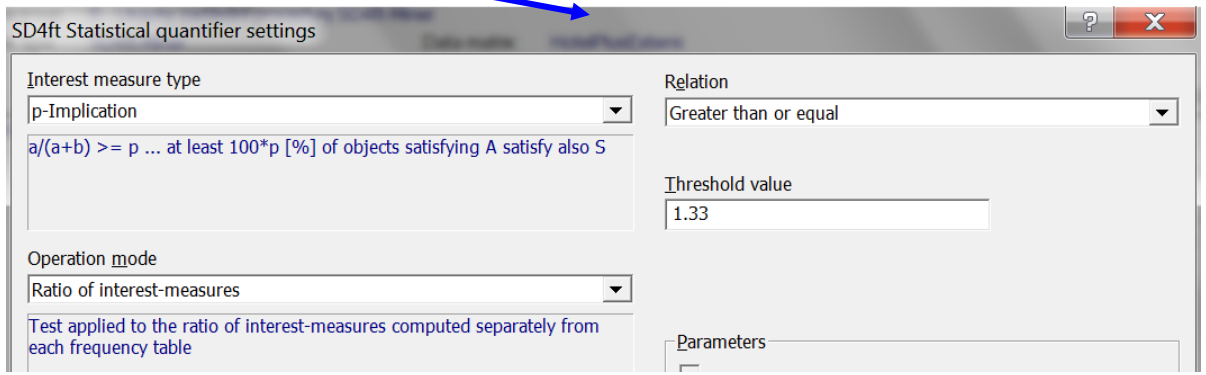
Hotel / HStat(ČR)	DHodnocení(spokojen)	¬DHodnocení(spokojen)
True	a_1	b_1
¬True	*0*	*0*

Hotel / HStat(ČR) ∧ Podmínka	DHodnocení(spokojen)	¬DHodnocení(spokojen)
True	a_2	b_2
¬True	*0*	*0*

QUANTIFIERS

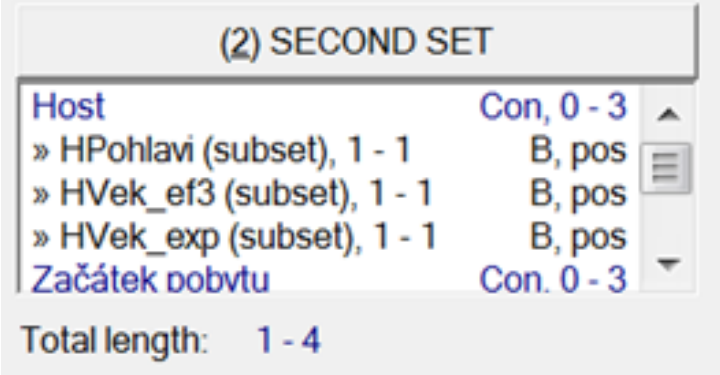
Type	Rel.	Value	Units
a (BASE) FirstSet	>=	50.00	Abs
a (BASE) SecondSet	>=	50.00	Abs
PIM RatioVal	<=	0.66	Abs

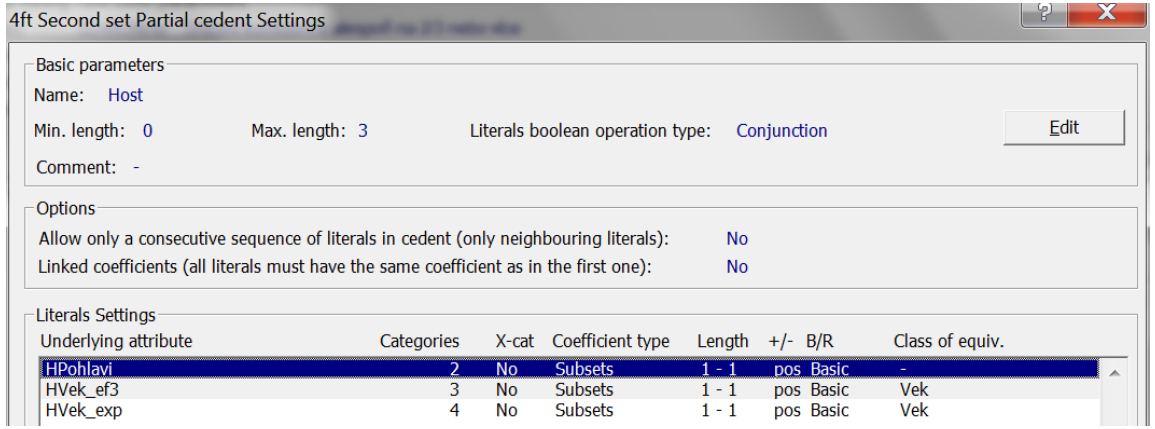
$a_1 \geq 50 \wedge a_2 \geq 50$

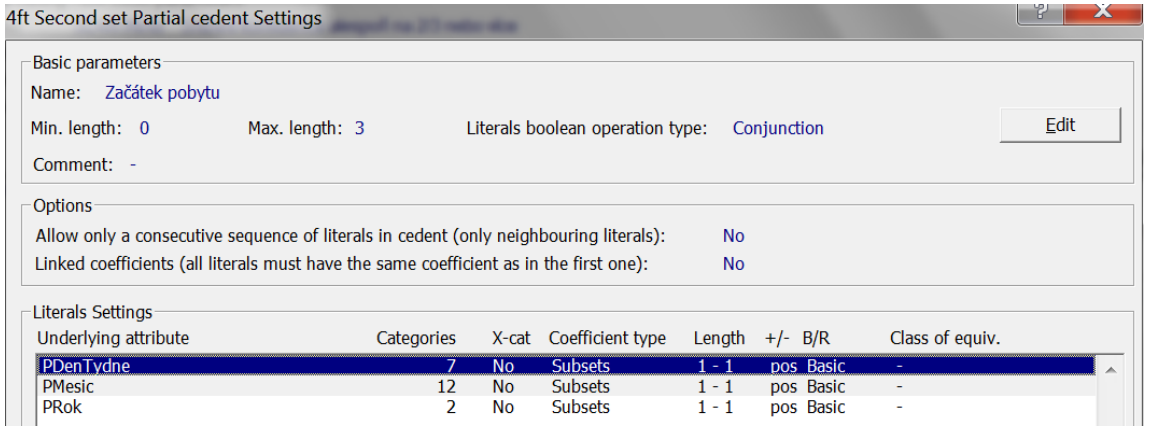


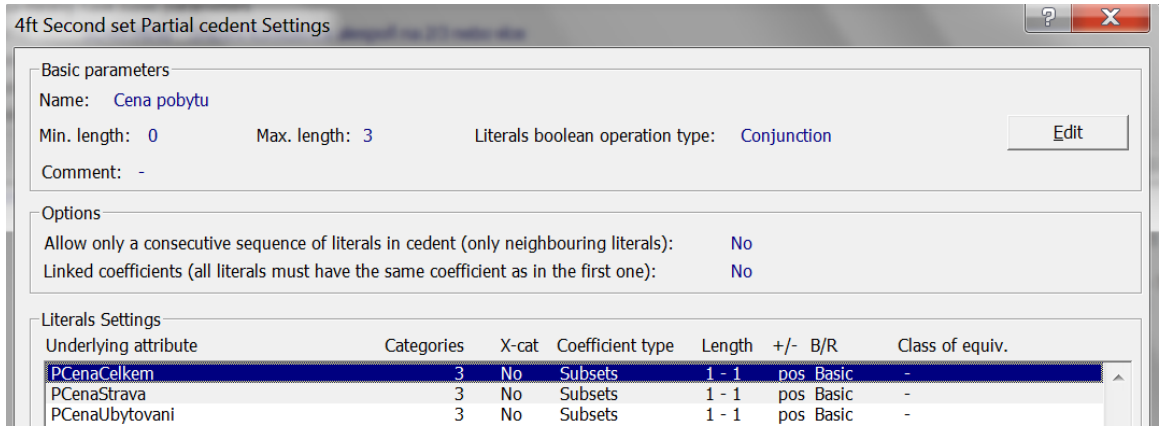
Chceme: $\frac{a_2}{a_2+b_2} \geq (1+\frac{1}{2})\frac{a_1}{a_1+b_1}$, čili $\frac{\frac{a_1}{a_1+b_1}}{\frac{a_2}{a_2+b_2}} \leq 0.66$

Second set = dodatečná podmínka









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

Add group | Del group | Edit group

Actual group of hypotheses: All hypotheses

Hypotheses in group: 29 Shown hypotheses: 29 Highlighted: 0

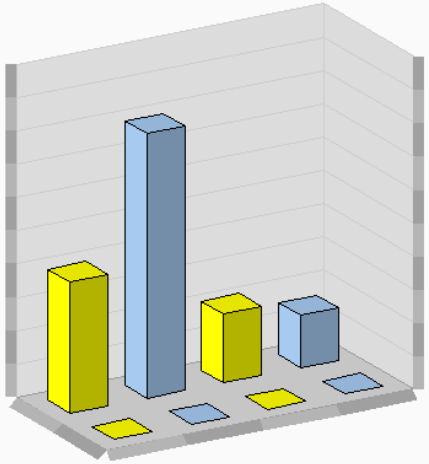
Nr.	Id	R-Conf	1:Conf	2:Conf	Hypothesis
1	19	0.583	0.331	0.569	(empty) >÷< **DStrava**(*nižší*) : **HStat**(*Německo*) × Set1 & **PRok**(*2013*) & **PCenaCelkem**(*vyšší*)
2	21	0.583	0.331	0.569	(empty) >÷< **DStrava**(*nižší*) : **HStat**(*Německo*) × Set1 & **PRok**(*2013*) & **PCenaCelkem**(*vyšší*) & **PCenaUbytovani**(*high*)
3	18	0.583	0.331	0.569	(empty) >÷< **DStrava**(*nižší*) : **HStat**(*Německo*) × Set1 & **PRok**(*2013*) & **PCenaUbytovani**(*high*)
4	6	0.596	0.279	0.469	(empty) >÷< **DHodnoceni**(*nespokojen*) : **HStat**(*ČR*) × Set1 & **PRok**(*2012*) & **PCenaCelkem**(*vyšší*)
5	8	0.596	0.279	0.469	(empty) >÷< **DHodnoceni**(*nespokojen*) : **HStat**(*ČR*) × Set1 & **PRok**(*2012*) & **PCenaCelkem**(*vyšší*) & **PCenaUbytovani**(*high*)
6	4	0.596	0.279	0.469	(empty) >÷< **DHodnoceni**(*nespokojen*) : **HStat**(*ČR*) × Set1 & **PRok**(*2012*) & **PCenaUbytovani**(*high*)
7	7	0.603	0.315	0.523	(empty) >÷< **DPersonal**(*nižší*) : **HStat**(*ČR*) × Set1 & **PRok**(*2012*) & **PCenaCelkem**(*vyšší*)
8	9	0.603	0.315	0.523	(empty) >÷< **DPersonal**(*nižší*) : **HStat**(*ČR*) × Set1 & **PRok**(*2012*) & **PCenaCelkem**(*vyšší*) & **PCenaUbytovani**(*high*)
9	5	0.603	0.315	0.523	(empty) >÷< **DPersonal**(*nižší*) : **HStat**(*ČR*) × Set1 & **PRok**(*2012*) & **PCenaUbytovani**(*high*)
10	2	0.617	0.279	0.452	(empty) >÷< **DHodnoceni**(*nespokojen*) : **HStat**(*ČR*) × Set1 & **PCenaCelkem**(*vyšší*) & **PCenaStrava**(*high*)
11	3	0.617	0.279	0.452	(empty) >÷< **DHodnoceni**(*nespokojen*) : **HStat**(*ČR*) × Set1 & **PCenaCelkem**(*vyšší*) & **PCenaStrava**(*high*) & **PCenaUbytovani**(*high*)
12	1	0.617	0.279	0.452	(empty) >÷< **DHodnoceni**(*nespokojen*) : **HStat**(*ČR*) × Set1 & **PCenaStrava**(*high*) & **PCenaUbytovani**(*high*)
13	20	0.623	0.331	0.532	(empty) >÷< **DStrava**(*nižší*) : **HStat**(*Německo*) × Set1 & **PRok**(*2013*) & **PCenaStrava**(*high*)
14	15	0.629	0.314	0.500	(empty) >÷< **DUbytovani**(*nižší*) : **HStat**(*ČR*) × Set1 & **HVek**(*průměr*) & **PRok**(*2012*)
15	25	0.631	0.318	0.505	(empty) >÷< **DStrava**(*nižší*) : **HStat**(*Rakousko*) × Set1 & **HPohlavi**(*žena*) & **PRok**(*2013*) & **PCenaStrava**(*low*)
16	24	0.634	0.305	0.481	(empty) >÷< **DUbytovani**(*nižší*) : **HStat**(*Rakousko*) × Set1 & **PRok**(*2012*) & **PCenaStrava**(*low*)
17	29	0.639	0.305	0.477	(empty) >÷< **DUbytovani**(*nižší*) : **HStat**(*Rakousko*) × Set1 & **HVek**(*od 28 do 60*) & **PRok**(*2012*)
18	13	0.642	0.306	0.476	(empty) >÷< **DStrava**(*nižší*) : **HStat**(*ČR*) × Set1 & **PDenTydne**(*So*) & **PRok**(*2013*) & **PCenaCelkem**(*vyšší*)

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

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

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