Tato prezentace je součástí wiki-prezentace Metoda GUHA, LISp-Miner a typové úlohy
Je dostupná z této adresy
Verse 20. 8. 2019

Typ úlohy: Využití koeficientů

Data: Adult

Problém: *Hledání segmentů osob s extrémními hodnotami zisku a segmentů osob s extrémními hodnotami zisku a zároveň s vysokým příjmem.*

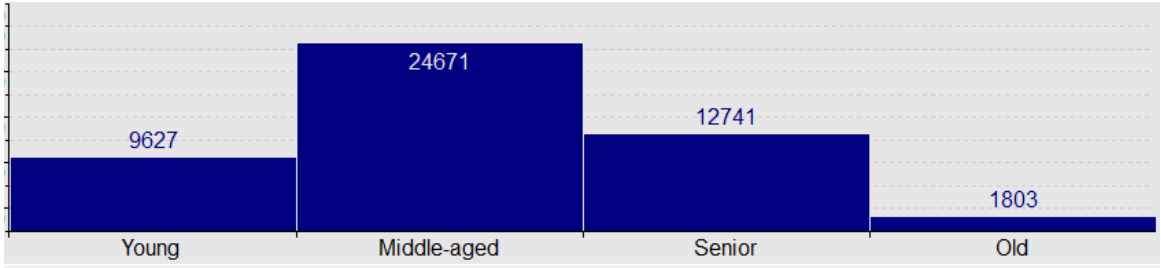
Jan Rauch
Katedra informačního a znalostního inženýrství
Vysoká škola ekonomická v Praze

Příklady využití koeficientů v datech Adult

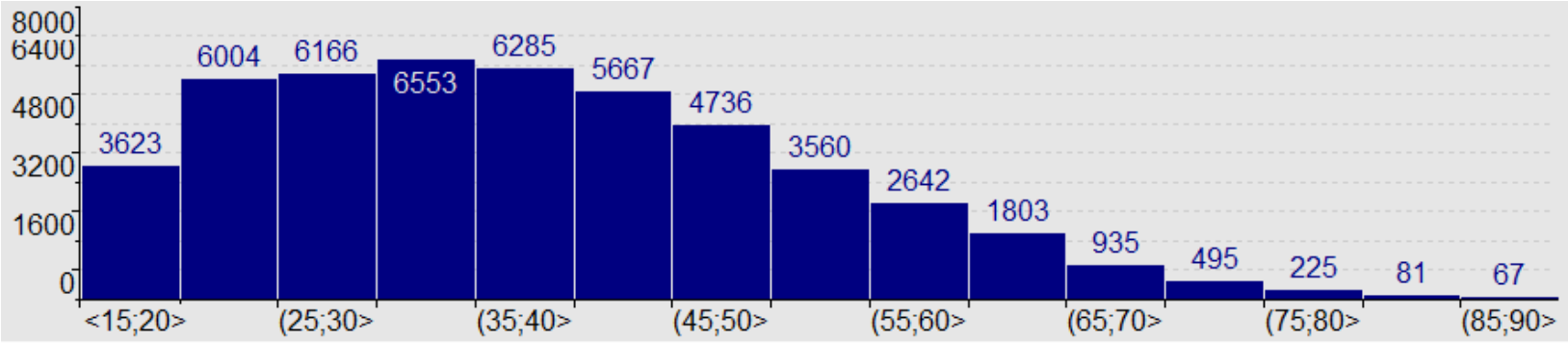
- Nové kategorizace atributů a využití kategorií
- Lift a AA-míra
- Segmenty osob s extrémními hodnotami zisku
- Segmenty osob s extrémními hodnotami zisku a zároveň s vysokým příjmem

Nová kategorizace atributu Age

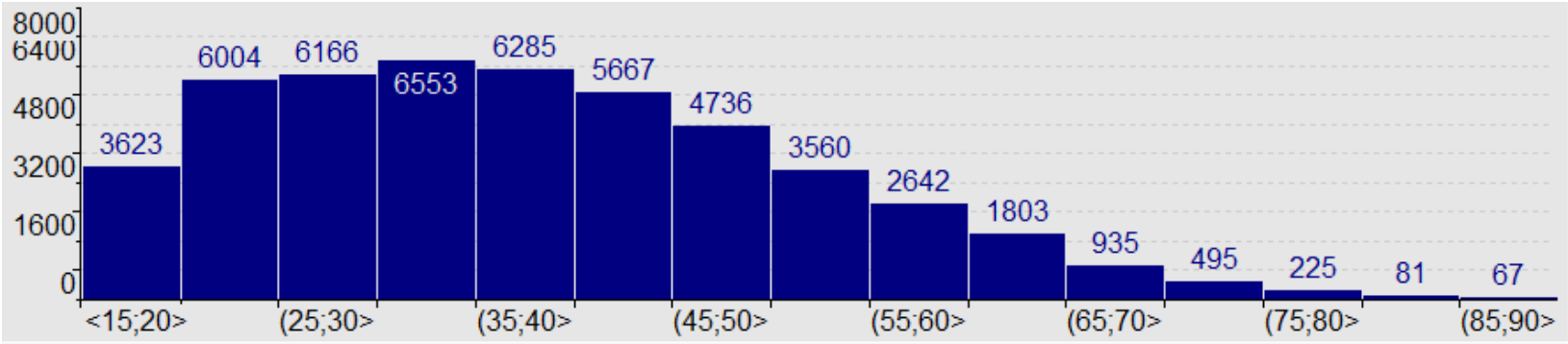
Age použito:



Nová kategorizace Age. 15 kategorií, po pěti letech počínaje 15:

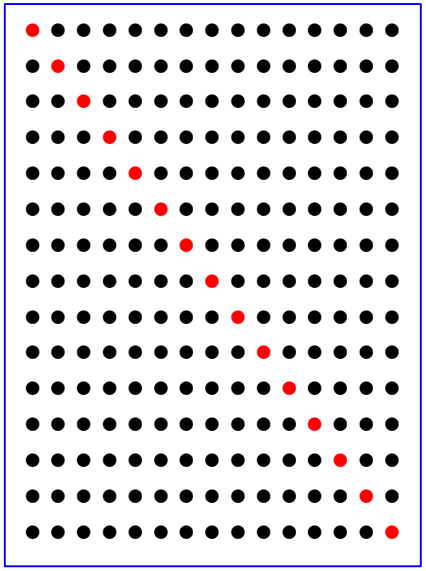


Nová kategorizace atributu Age - použití sequencí

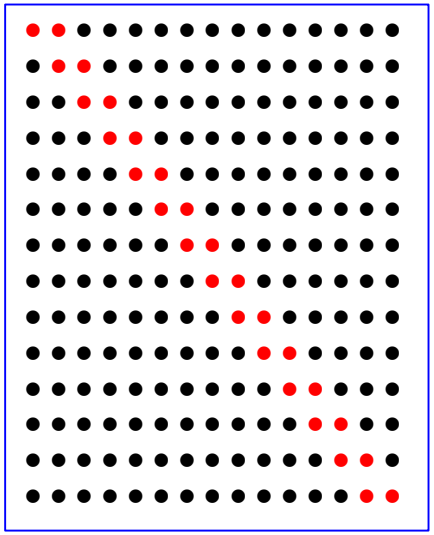


Sequence délky 1 až 4: 15 + 14 + 13 + 12 = 54 literálů

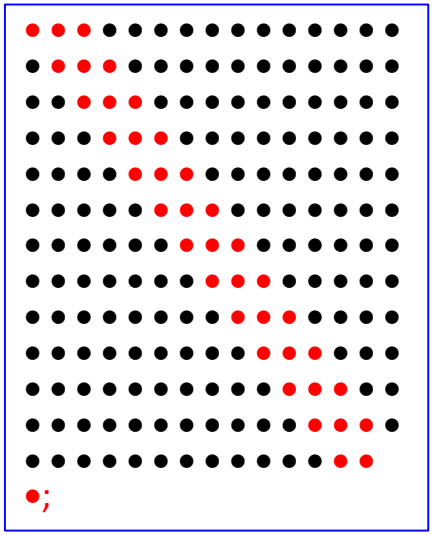
délka 1 - 15 literálů



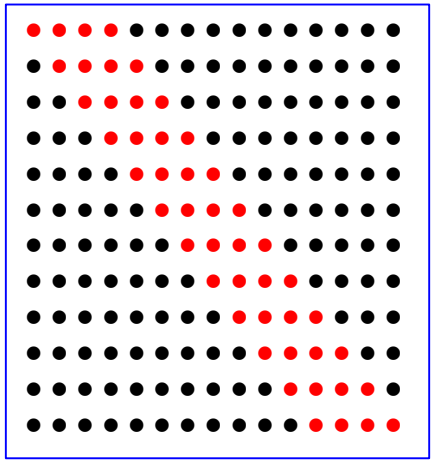
délka 2 - 14 literálů



délka 3 - 13 literálů

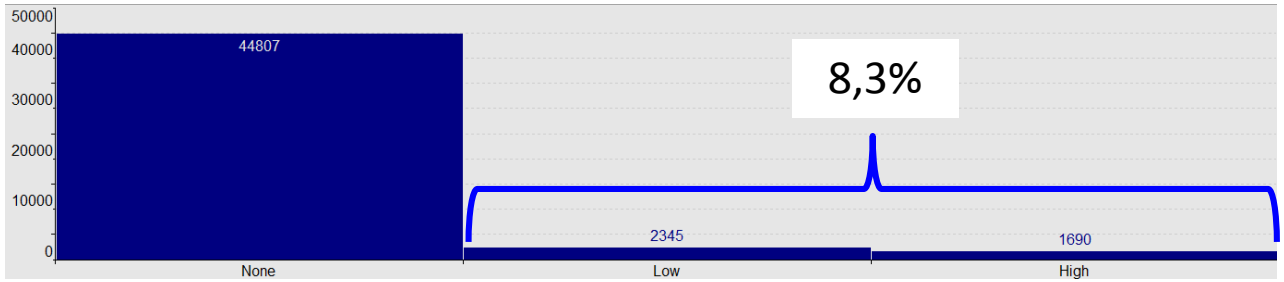


délka 4 - 12 literálů

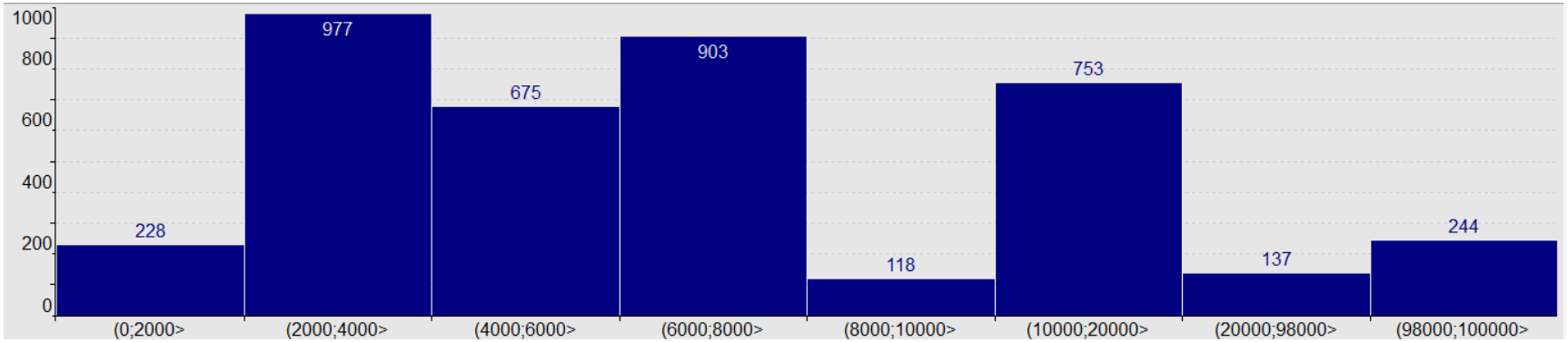


Nová kategorizace atributu Capital_gain

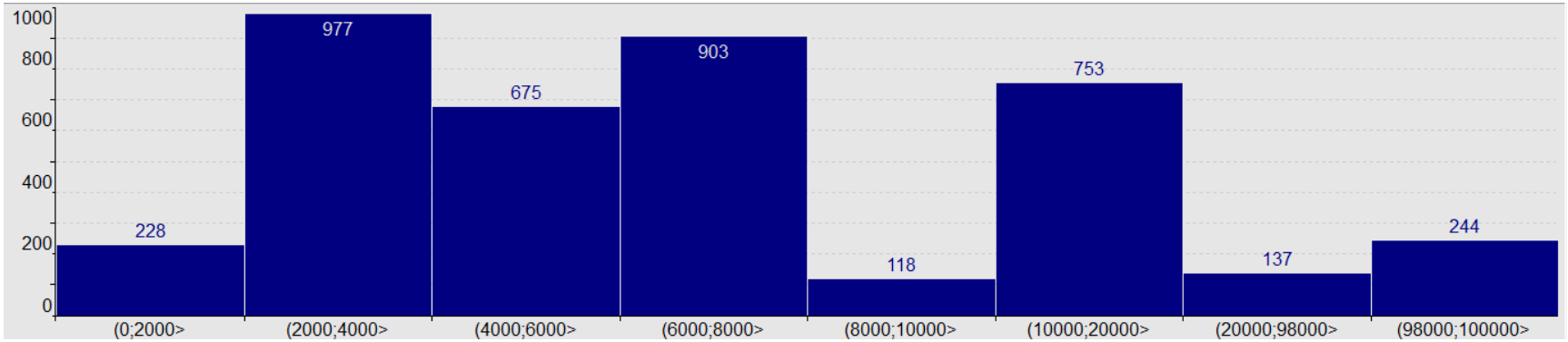
Capital_gain:



Nová kategorizace 8,3% nenulových hodnot. 8 kategorií různé délky:



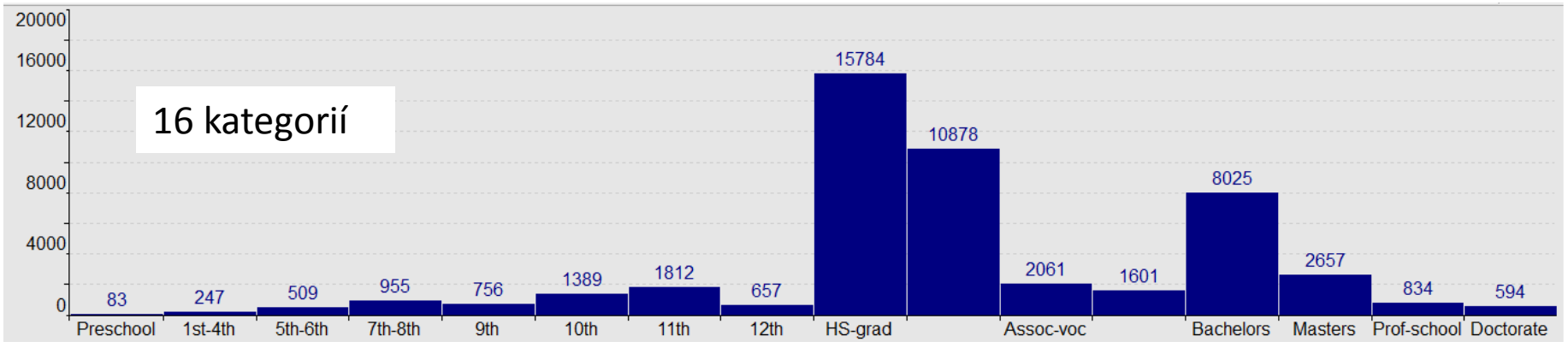
Nová kategorizace atributu Capital_gain - použití pravých řezů



Pravé řezy délky 1 až 8: 8 literálů

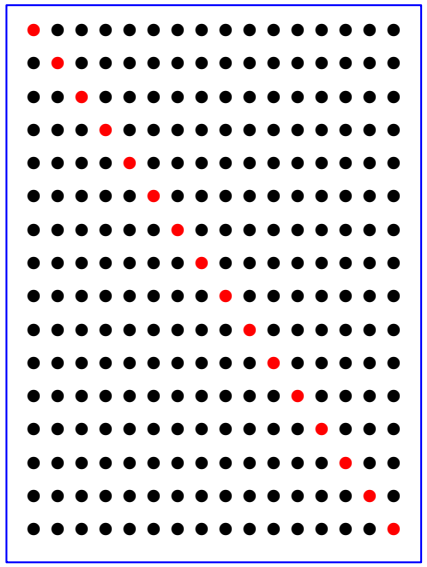
● ● ● ● ● ● ● ●	(98 000; 100 000⟩
● ● ● ● ● ● ● ●	(20 000; 100 000⟩
● ● ● ● ● ● ● ●	(10 000; 100 000⟩
● ● ● ● ● ● ● ●	(8 000; 100 000⟩
● ● ● ● ● ● ● ●	(6 000; 100 000⟩
● ● ● ● ● ● ● ●	(4 000; 100 000⟩
● ● ● ● ● ● ● ●	(2 000; 100 000⟩
● ● ● ● ● ● ● ●	> 0

Kategorizace atributu Education - použití sequencí

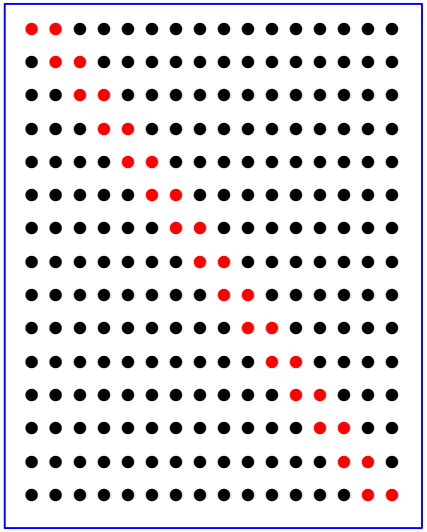


Sequence délky 1 až 4: 16 + 15 + 14 + 13 = 58 literálů

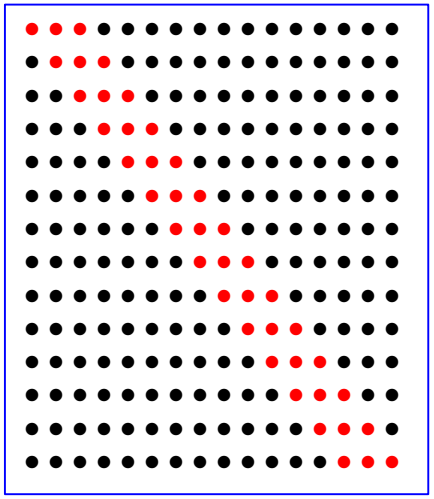
délka 1 - 16 literálů



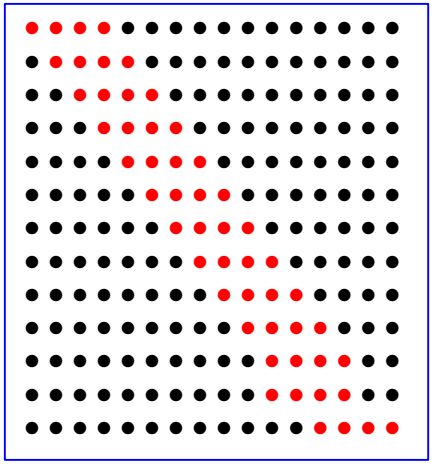
délka 2 - 15 literálů



délka 3 - 14 literálů



délka 4 - 13 literálů



Příklady využití koeficientů v datech Adult

- Nové kategorizace atributů a využití kategorií
- Lift a AA-míra
- Segmenty osob s extrémními hodnotami zisku
- Segmenty osob s extrémními hodnotami zisku a zároveň s vysokým příjmem

Lift asociačního pravidla

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

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

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

- 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

Lift - odvozené 4ft kvantifikátory (1)

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

$\sim^{+}_{p,s}$ - podmínka $\frac{a(a+b+c+d)}{(a+b)(a+c)} \geq p \wedge \frac{a}{a+b+c+d} \geq s$ $0 \leq p \leq 1, 0 \leq s \leq 1$

$\varphi \sim^{+}_{p,s} \psi$: zároveň platí:

- relativní četnost ψ pokud platí φ je nejméně *p*-krát vyšší, než relativní četnost ψ v celé matici dat
- relativní četnost řádků splňujících $\varphi \wedge \psi$ je nejméně *s*

Lift - odvozené 4ft kvantifikátory (2)

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

$\sim^{+}_{p,Base}$ - podmínka $\frac{a(a+b+c+d)}{(a+b)(a+c)} \geq p \wedge a \geq Base$ $0 \leq p \leq 1$, $Base \geq 1$, celé

$\varphi \sim^{+}_{p,\mathrm{Base}} \psi$: zároveň platí:

- relativní četnost ψ pokud platí φ je nejméně *p*-krát vyšší, než relativní četnost ψ v celé matici dat
- počet řádků splňujících $\varphi \wedge \psi$ je nejméně *Base*

AA míra asociačního pravidla

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

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

- AA-míra = lift – 1, neboli

AA-míra = $\dfrac{\textit{relativní četnost } \psi \textit{ pokud platí } \varphi}{\textit{relativní četnost } \psi \textit{ v celé matici dat}}$ – 1, 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

AA míra - odvozené 4ft kvantifikátory (1)

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

$\rightarrow^{+}_{p,s}$ - podmínka $\frac{a}{a+b} \geq (1+p)\frac{a+c}{a+b+c+d} \wedge \frac{a}{a+b+c+d} \geq s$ $0 \leq p \leq 1$, $0 \leq s \leq 1$

$\varphi \rightarrow^{+}_{p,s} \psi$: zároveň platí:

- relativní četnost ψ pokud platí φ je nejméně o $100p$ % vyšší, než relativní četnost ψ v celé matici dat
- relativní četnost řádků splňujících $\varphi \wedge \psi$ je nejméně s

AA míra - odvozené 4ft kvantifikátory (2)

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

$\Rightarrow^{+}_{p,Base}$ - podmínka $\frac{a}{a+b} \geq (1+p)\frac{a+c}{a+b+c+d} \wedge a \geq Base$ $0 \leq p \leq 1$, $Base \geq 1$, celé

$\varphi \Rightarrow^{+}_{p,s} \psi$: zároveň platí:

- relativní četnost ψ pokud platí φ je nejméně o $100p$ % vyšší, než relativní četnost ψ v celé matici dat
- počet řádků splňujících $\varphi \wedge \psi$ je nejméně *Base*

AA míra - poznámka

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

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

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

lift: $$\boxed{\frac{a(a+b+c+d)}{(a+b)(a+c)}} - 1 \geq p$$

ekvivalentní

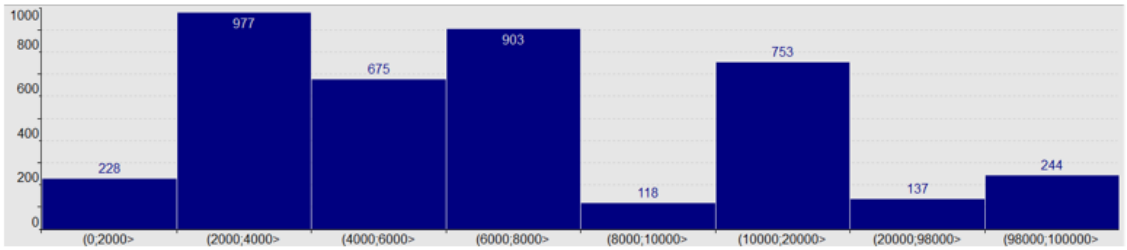
Příklady využití koeficientů v datech Adult

- Nové kategorizace atributů a využití kategorií
- Lift a AA-míra
- Segmenty osob s extrémními hodnotami zisku
- Segmenty osob s extrémními hodnotami zisku a zároveň s vysokým příjmem

Segmenty osob s extrémními hodnotami zisku - definice úlohy

Pro jaké segmenty osob definované kombinacemi hodnot atributů platí, že

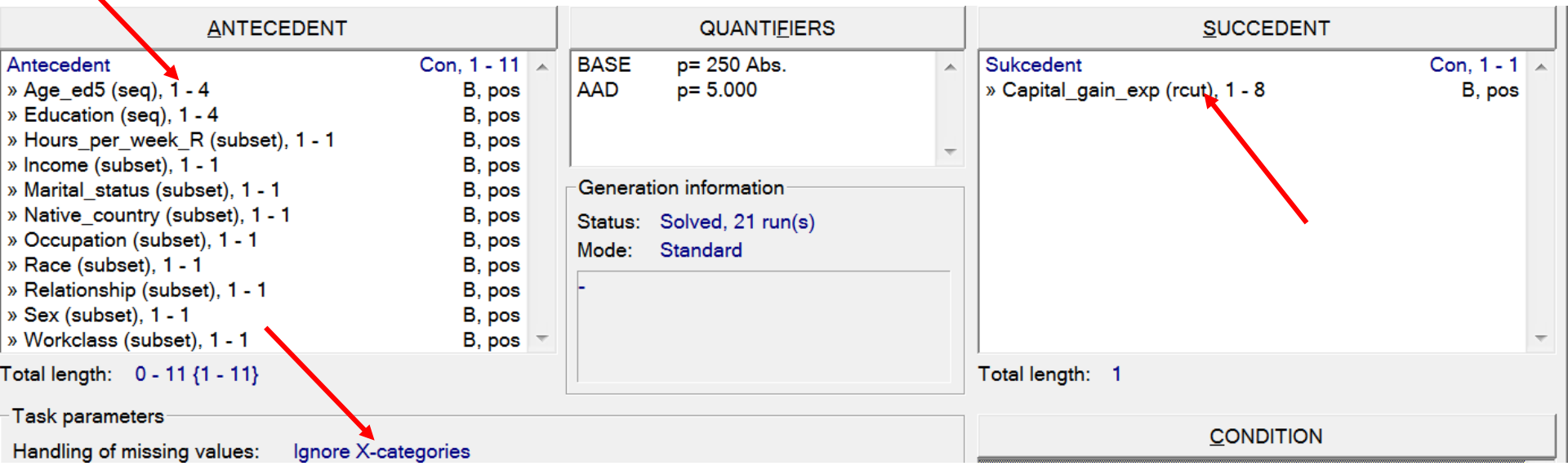
- Segment se týká alespoň 250 osob - řádků matice dat Adult
- Nejvyšší zisky jsou dány pravými řezy atributu Capital_gain
- Četnost osob s daným ziskem v segmentu je alespoň o 500 procent (tedy 6x) větší, než v celých datech?



Pravé řezy délky 1 až 8: 8 literálů

●●●●●●●●	(98 000; 100 000⟩
●●●●●●●●	(20 000; 100 000⟩
●●●●●●●●	(10 000; 100 000⟩
●●●●●●●●	(8 000; 100 000⟩
●●●●●●●●	(6 000; 100 000⟩
●●●●●●●●	(4 000; 100 000⟩
●●●●●●●●	(2 000; 100 000⟩
●●●●●●●●	> 0

Segmenty klientů s extrémními hodnotami zisku - zadání úlohy pro 4ft-Miner



- Relevantní segmenty jsou definované antecedenty - konjunkcemi délky 1 až 11
- Pro atributy Age_ed5 a Education se používají koeficienty - sequence délky 1 až 4
- Extrémní hodnoty zisku jsou dány pravými řezy
- Segment je zajímavý pokud současně platí
  - patří do něho alespoň 250 osob, viz BASE p= 250 Abs.
  - četnost osob se ziskem daným pravým řezem je alespoň o 500 procent tedy 6x větší v segmentu, než v celých datech, viz AAD p= 5.000
- Použito Ignore X-categories, viz téma asociační pravidla a neúplná informace

Segmenty klientů s extrémními hodnotami zisku - ukázka přehledného výstupu

Task run

Start: 7.1.2019 22:58:49 Total time: 0h 4m 10s

Number of verifications: 8272873

Number of hypotheses: 35 Mode: Standard

Add group | Del group | Edit group

Actual group of hypotheses: All hypotheses

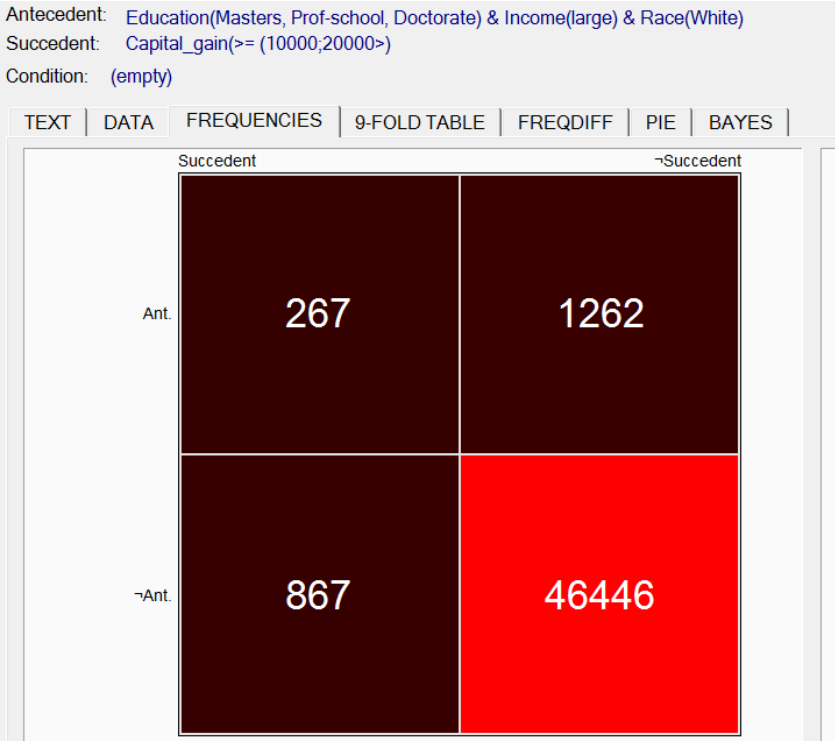
Hypotheses in group: 35 Shown hypotheses: 35 Highlighted: 0

Delete hypothese

Nr.	Id	Lift	Hypothesis
1	33	7.521	**Education**(*Masters*,*Prof-school*,*Doctorate*) & **Income**(*large*) & **Race**(*White*) >÷< **Capital_gain**(> *10000*)
2	30	7.434	**Education**(*Masters*,*Prof-school*,*Doctorate*) & **Income**(*large*) & **Native_country**(*United-States*) >÷< **Capital_gain**(> *10000*)
3	28	7.400	**Education**(*Masters*,*Prof-school*,*Doctorate*) & **Income**(*large*) >÷< **Capital_gain**(> *10000*)
4	27	7.170	**Education**(*Masters*,*Prof-school*) & **Income**(*large*) >÷< **Capital_gain**(> *8000*)
5	31	7.074	**Education**(*Masters*,*Prof-school*,*Doctorate*) & **Income**(*large*) & **Native_country**(*United-States*) >÷< **Capital_gain**(> *8000*)
6	32	7.053	**Education**(*Masters*,*Prof-school*,*Doctorate*) & **Income**(*large*) & **Native_country**(*United-States*) & **Race**(*White*) >÷< **Capital_gain**(> *8000*)
7	34	7.042	**Education**(*Masters*,*Prof-school*,*Doctorate*) & **Income**(*large*) & **Race**(*White*) >÷< **Capital_gain**(> *8000*)
8	29	7.003	**Education**(*Masters*,*Prof-school*,*Doctorate*) & **Income**(*large*) >÷< **Capital_gain**(> *8000*)
9	35	6.884	**Education**(*Prof-school*,*Doctorate*) >÷< **Capital_gain**(> *8000*)
10	14	6.399	**Education**(*Bachelors*,*Masters*,*Prof-school*) & **Hours**(*Over-time*) & **Income**(*large*) >÷< **Capital_gain**(> *10000*)
11	18	6.379	**Education**(*Bachelors*,*Masters*,*Prof-school*,*Doctorate*) & **Hours**(*Over-time*) & **Income**(*large*) >÷< **Capital_gain**(> *10000*)
12	24	6.358	**Education**(*Bachelors*,*Masters*,*Prof-school*,*Doctorate*) & **Hours**(*Over-time*) & **Income**(*large*) & **Race**(*White*) >÷< **Capital_gain**(> *10000*)
13	20	6.342	**Education**(*Bachelors*,*Masters*,*Prof-school*,*Doctorate*) & **Hours**(*Over-time*) & **Income**(*large*) & **Native_country**(*United-States*) >÷< **Capital_gain**(> *10000*)
14	22	6.310	**Education**(*Bachelors*,*Masters*,*Prof-school*,*Doctorate*) & **Hours**(*Over-time*) & **Income**(*large*) & **Native_country**(*United-States*) & **Race**(*White*) >÷< **Capital_gain**(> *100*
15	6	6.288	**Age**(*(40;60>*) & **Education**(*Bachelors*,*Masters*,*Prof-school*,*Doctorate*) & **Income**(*large*) & **Native_country**(*United-States*) & **Race**(*White*) >÷< **Capital_gain**(> *10000*)

- Žádné z vystupujících pravidel nelze získat aplikací arules pro zadané kategorie atributů.
- Při použití arules je nutné složité předzpracování a násobné aplikace.

Segmenty klientů s extrémními hodnotami zisku - ukázka detailního výstupu nejsilnějšího pravidla



Antecedent - segment osob splňujících Education(Masters, Prof-school,Doctorate) $\wedge$ Income(large) $\wedge$ Race(White)

Sukcedent: Capital_gain: $\geq$ 10 000

Relativní četnost osob v segmentu s Capital_gain: $\geq$ 10 000 je 7.521 krát větší, než relativní četnost takových osob v celých datech

$$\text{Lift} = \frac{\frac{267}{267 + 1\,262}}{\frac{267 + 867}{267 + 1\,262 + 867 + 46\,446}} = \frac{\frac{267}{1529}}{\frac{1134}{48842}} = \frac{267 * 48842}{1529 * 1134} = 7.521$$

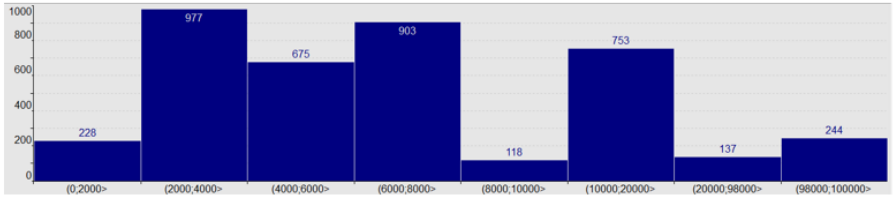
Příklady využití koeficientů v datech Adult

- Nové kategorizace atributů a využití kategorií
- Lift a AA-míra
- Segmenty osob s extrémními hodnotami zisku
- Segmenty osob s extrémními hodnotami zisku a zároveň s vysokým příjmem

Extrémní zisk a zároveň vysoký příjem - definice úlohy

Pro jaké segmenty osob definované kombinacemi hodnot atributů platí, že

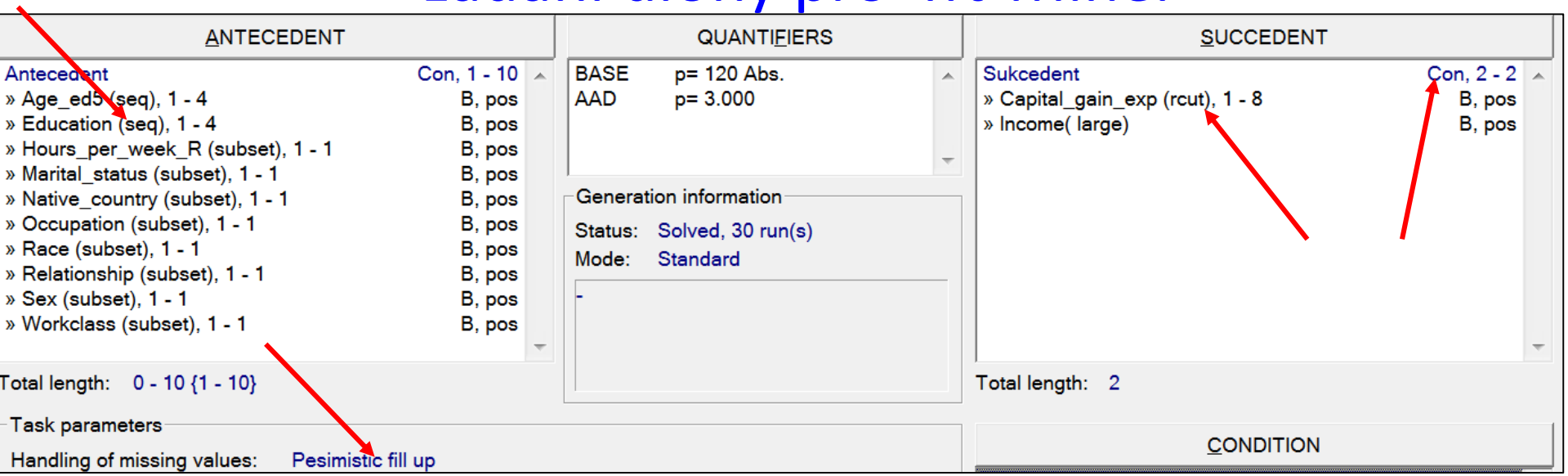
- Segment se týká alespoň 120 osob - řádků matice dat Adult
- Nejvyšší zisky jsou dány pravými řezy atributu Capital_gain
- Četnost osob splňujících Income(large) a vykazujících některý z maximálních zisků je v segmentu alespoň o 300 procent (tedy 4x) větší, než v celých datech



Pravé řezy délky 1 až 8: 8 literálů

••••••••	(98 000; 100 000⟩
••••••••	(20 000; 100 000⟩
••••••••	(10 000; 100 000⟩
••••••••	(8 000; 100 000⟩
••••••••	(6 000; 100 000⟩
••••••••	(4 000; 100 000⟩
••••••••	(2 000; 100 000⟩
••••••••	> 0

Extrémní zisk a zároveň vysoký příjem - definice úlohy - zadání úlohy pro 4ft-Miner



- Relevantní segmenty jsou definované antecedenty - konjunkcemi délky 1 až 10
- Pro atributy Age_ed5 a Education se používají koeficienty - sequence délky 1 až 4
- Extrémní hodnoty zisku jsou dány pravými řezy
- Sukcedent je konjunkce pravého řezu - extrémní hodnoty Capital_gain a Income(large),
- Segment je zajímavý pokud současně platí
  - patří do něho alespoň 120 osob, viz BASE p= 120 Abs.
  - četnost osob splňujících sukcedent je v segmentu alespoň o 300 procent, tedy 4x větší než v celých datech, viz AAD p= 3.000
- Použito zabezpečené (pesimistické) doplnění, viz téma asociační pravidla a neúplná informace

Konjunkce atributů v sukcedentu - ukázka přehledného výstupu

Task run

Start: 13.1.2019 22:02:00 Total time: 0h 40m 39s

Number of verifications: 30281480

Number of hypotheses: 164 Mode: Standard

Add group | Del group | Edit group

Actual group of hypotheses: All hypotheses

Hypotheses in group: 164 Shown hypotheses: 164 Highlighted: 0

Nr.	Id	Lift	Hypothesis
1	130	6.641	**Education**(*Prof-school*) >÷< **Capital_gain**(>*10000*) & **Income**(*large*)
2	59	6.541	**Age**(*(35;55>*) & **Education**(*Prof-school*,*Doctorate*) >÷< **Capital_gain**(>*10000*) & **Income**(*large*)
3	134	6.117	**Education**(*Prof-school*,*Doctorate*) & **Marital_status**(*Married-civ-spouse*) >÷< **Capital_gain**(>*10000*) & **Income**(*large*)
4	131	5.990	**Education**(*Prof-school*) >÷< **Capital_gain**(>*8000*) & **Income**(*large*)
5	157	5.978	**Education**(*Prof-school*,*Doctorate*) & **Race**(*White*) & **Sex**(*Male*) >÷< **Capital_gain**(>*10000*) & **Income**(*large*)
6	60	5.946	**Age**(*(35;55>*) & **Education**(*Prof-school*,*Doctorate*) >÷< **Capital_gain**(>*8000*) & **Income**(*large*)
7	138	5.935	**Education**(*Prof-school*,*Doctorate*) & **Marital_status**(*Married-civ-spouse*) & **Relationship**(*Husband*) & **Sex**(*Male*) >÷< **Capital_**
8	136	5.935	**Education**(*Prof-school*,*Doctorate*) & **Marital_status**(*Married-civ-spouse*) & **Relationship**(*Husband*) >÷< **Capital_gain**(>*10000*
9	161	5.935	**Education**(*Prof-school*,*Doctorate*) & **Relationship**(*Husband*) & **Sex**(*Male*) >÷< **Capital_gain**(>*10000*) & **Income**(*large*)
10	159	5.935	**Education**(*Prof-school*,*Doctorate*) & **Relationship**(*Husband*) >÷< **Capital_gain**(>*10000*) & **Income**(*large*)
11	163	5.932	**Education**(*Prof-school*,*Doctorate*) & **Sex**(*Male*) >÷< **Capital_gain**(>*10000*) & **Income**(*large*)
12	140	5.929	**Education**(*Prof-school*,*Doctorate*) & **Marital_status**(*Married-civ-spouse*) & **Sex**(*Male*) >÷< **Capital_gain**(>*10000*) & **Income**(*la*
13	146	5.904	**Education**(*Prof-school*,*Doctorate*) & **Native_country**(*United-States*) & **Race**(*White*) & **Sex**(*Male*) >÷< **Capital_gain**(>*10000*) &
14	148	5.876	**Education**(*Prof-school*,*Doctorate*) & **Native_country**(*United-States*) & **Sex**(*Male*) >÷< **Capital_gain**(>*10000*) & **Income**(*large*)
15	153	5.820	**Education**(*Prof-school*,*Doctorate*) & **Occupation**(*Prof-specialty*) & **Sex**(*Male*) >÷< **Capital_gain**(>*10000*) & **Income**(*large*)

- Žádné z vystupujících pravidel nelze získat aplikací arules pro zadané kategorie atributů.
- Při použití arules je nutné složité předzpracování a násobné aplikace.
- V arules principelně nelze použít zabezpečené (pesimistické) doplnění

Konjunkce atributů v sukcedentu
- ukázka detailního výstupu nejsilnějšího pravidla

Antecedent: Education(Prof-school)
Succedent: Capital_gain(>= (10000;20000>) & Income(large)
Condition: (empty)

TEXT | DATA | FREQUENCIES | 9-FOLD TABLE | FREQDIFF | PIE | BA

	Succedent	¬Succedent
Ant.	121	713
¬Ant.	946	47062

Antecedent - segment osob splňujících Education(Prof-school)

Succedent:
Capital_gain: $\geq$ 10 000 $\wedge$ Income(large)

Relativní četnost osob splňujících Capital_gain: $\geq$ 10 000 $\wedge$ Income(large) mezi osobami splňujícími Education(Prof-school) je 6.64 krát vyšší, než v celých datech.

$$\text{Lift} = \frac{\frac{121}{121+713}}{\frac{121+946}{121+713+946+47\,062}} = \frac{\frac{121}{834}}{\frac{1\,067}{48842}} = \frac{121 * 48842}{1\,067 * 834} = 6.641$$