

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

Je dostupná z [této adresy](#)

Verze 30. 7. 2019

Definice pattern difference KL-kvantifikátorů pro proceduru KL-Miner

Jan Rauch

Katedra informačního a znalostního inženýrství

Vysoká škola ekonomická v Praze

Start tlačítkem Add pattern difference quantifier

- Definice kvantifikátoru se provádí zadáním parametrů 1 - 6.
- Pro všechny parametry se nabízejí defaultní hodnoty.
- Podmínka kvantifikátoru je definována parametry 5 a 6.
- Podmínka kvantifikátoru se aplikuje na hodnotu IM spočtenou dle parametru 4.

KL List of quantifiers settings

Task: Zadávání KL-Miner

Type	Source	Range	Rel.	Value	Units

Example attribute for row: DHodnoceni (3 categories)

Example attribute for column: DStrava_edc5 (5 categories)

	<0;19>	<20;39>	<40;59>	<60;79>	<80;100>
nespokojen	154	400	16	0	0
průměr	0	316	372	262	0
spokojen	0	0	0	124	356

Buttons: Detail, Del, Up, Down, Copy, Paste, Add statistical quantifier, Add simple frequencies quantifier, Close, Add pattern-difference quantifier

KL Pattern Difference quantifier settings

Interest measure type: PattDiffMax (6)

Relation: Greater than or equal

Threshold value: 5

Threshold-value units: Absolute number (5)

Source frequencies: Relative [%] to act condition (5)

Pattern used for comparison: Selected stored pattern (of the same size) (1)

Category Range: Rows (From: 0, To: 100), Columns (From: 0, To: 100)

Primary IM Settings: Set as primary IM, Normalize value range: From 0, To 1

Note: -

Buttons: OK, Cancel

Výchozí KL-tabulky

Pattern difference KL-kvantifikátor se pro KL-vztah $R \approx C / \chi$ verifikuje na základě KL-tabulek $KL(R,C,M/\chi)$ a $KL(R,C,M)$:

$$KL(R,C,M/\chi) =$$

M/ χ	c_1	...	c_u	...	c_v	...	c_L	Σ_l
r_1	$n_{1,1}$...	$n_{1,u}$...	$n_{1,v}$...	$n_{1,L}$	$n_{1,*}$
\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	\ddots	\vdots	\vdots
r_p	$n_{p,1}$...	$n_{p,u}$...	$n_{p,v}$...	$n_{p,L}$	$n_{p,*}$
\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	\ddots	\vdots	\vdots
r_q	$n_{q,1}$...	$n_{q,u}$...	$n_{q,v}$...	$n_{q,L}$	$n_{q,*}$
\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	\ddots	\vdots	\vdots
r_K	$n_{K,1}$...	$n_{K,u}$...	$n_{K,v}$...	$n_{K,L}$	$n_{K,*}$
Σ_k	$n_{*,1}$...	$n_{*,u}$...	$n_{*,v}$...	$n_{*,L}$	n

$$KL(R,C,M) =$$

M/ χ	c_1	...	c_u	...	c_v	...	c_L	Σ_l
r_1	$m_{1,1}$...	$m_{1,u}$...	$m_{1,v}$...	$m_{1,L}$	$m_{1,*}$
\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	\ddots	\vdots	\vdots
r_p	$m_{p,1}$...	$m_{p,u}$...	$m_{p,v}$...	$m_{p,L}$	$m_{p,*}$
\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	\ddots	\vdots	\vdots
r_q	$m_{q,1}$...	$m_{q,u}$...	$m_{q,v}$...	$m_{q,L}$	$m_{q,*}$
\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	\ddots	\vdots	\vdots
r_K	$m_{K,1}$...	$m_{K,u}$...	$m_{K,v}$...	$m_{K,L}$	$m_{K,*}$
Σ_k	$m_{*,1}$...	$m_{*,u}$...	$m_{*,v}$...	$m_{*,L}$	n_T

Pokud je jako *Pattern used for comparison* vybrán Selected stored pattern, tak se využijí i jeho frekvence:

Pattern	c_1	...	c_u	...	c_v	...	c_L	Σ_l
r_1	$s_{1,1}$...	$s_{1,u}$...	$s_{1,v}$...	$s_{1,L}$	$s_{1,*}$
\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	\ddots	\vdots	\vdots
r_p	$s_{p,1}$...	$s_{p,u}$...	$s_{p,v}$...	$s_{p,L}$	$s_{p,*}$
\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	\ddots	\vdots	\vdots
r_q	$s_{q,1}$...	$s_{q,u}$...	$s_{q,v}$...	$s_{q,L}$	$s_{q,*}$
\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	\ddots	\vdots	\vdots
r_K	$s_{K,1}$...	$s_{K,u}$...	$s_{K,v}$...	$s_{K,L}$	$s_{K,*}$
Σ_k	$s_{*,1}$...	$s_{*,u}$...	$s_{*,v}$...	$s_{*,L}$	s

1 - Pattern used for comparison

Parametr *Pattern used for comparison* určuje s kterou KL-tabulkou bude vygenerovaný KL-vztah $R \approx C / \chi$ porovnáván.

Pattern used for comparison

Contingency table of the same attribute pair on the whole matrix

Selected stored pattern (of the same size)

KL Příklad

Normalize pattern frequencies to SUM= 100

Platí:

- Volba *Contingency table of the same attribute pair on the whole matrix* znamená, že bude porovnáván vztah atributů R a C na KL-tabulkách $KL(R,C,M/\chi)$ a $KL(R,C,M)$. Dále značíme $\{w_{i,j}\}_{i=1,\dots,K}^{j=1,\dots,L} = \{m_{i,j}\}_{i=1,\dots,K}^{j=1,\dots,L}$. Platí i $w_{*,i} = m_{*,i}$,
 $w_{j,*} = m_{j,*}$ a $w = n_T$.
- *Selected stored pattern (of the same size)* znamená, že bude porovnáván vztah atributů R a C na KL-tabulkách $KL(R,C,M/\chi)$ a $KL(R,C,Pattern)$. Dále značíme $\{w_{i,j}\}_{i=1,\dots,K}^{j=1,\dots,L} = \{s_{i,j}\}_{i=1,\dots,K}^{j=1,\dots,L}$. Platí i $w_{*,i} = s_{*,i}$, $w_{j,*} = s_{j,*}$ a $w = s$.
- Volba *Normalize pattern frequencies to SUM = 100* znamená, že absolutní frekvence porovnávaných kontingenčních tabulek budou přepočítány na matici dat o 100 řádcích.

2 - Category Range

- Category Range určuje výseky tabulek, na které budou aplikovány ostatní parametry.
- Provádí se způsobem popsáním [zde](#).
- Výseky tabulek je dán čtveřicí $\langle p, q, u, v \rangle$ celých čísel splňujících $1 \leq p \leq q \leq K$ a $1 \leq u \leq v \leq L$
- Výsek daný čtveřicí $\langle p, q, u, v \rangle$ zahrnuje frekvence pro kategorie r_p, \dots, r_q a c_u, \dots, c_v .

Category Range

From
To

Reset to All

Columns

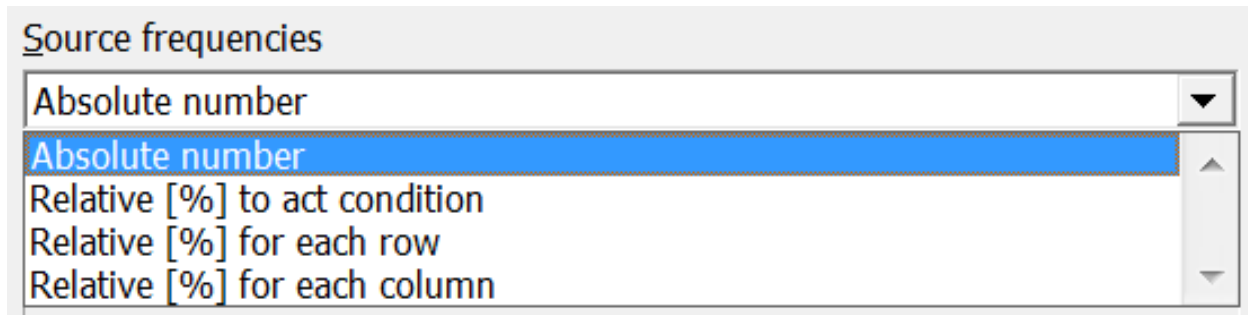
Absolute category index
 Relative range [%] to act number of categories

M/ χ	c_1	...	c_u	...	c_v	...	c_L	Σ_l
r_1	$x_{1,1}$...	$x_{1,u}$...	$x_{1,v}$...	$x_{1,L}$	$m_{1,*}$
\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	\ddots	\vdots	\vdots
r_p	$x_{p,1}$...	$x_{p,u}$...	$x_{p,v}$...	$x_{p,L}$	$x_{p,*}$
\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	\ddots	\vdots	\vdots
r_q	$x_{q,1}$...	$x_{q,u}$...	$x_{q,v}$...	$x_{q,L}$	$x_{q,*}$
\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	\ddots	\vdots	\vdots
r_K	$x_{K,1}$...	$x_{K,u}$...	$x_{K,v}$...	$x_{K,L}$	$x_{K,*}$
Σ_k	$x_{*,1}$...	$x_{*,u}$...	$x_{*,v}$...	$x_{*,L}$	x

K porovnání	c_1	...	c_u	...	c_v	...	c_L	Σ_l
r_1	$y_{1,1}$...	$y_{1,u}$...	$y_{1,v}$...	$y_{1,L}$	$m_{1,*}$
\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	\ddots	\vdots	\vdots
r_p	$y_{p,1}$...	$y_{p,u}$...	$y_{p,v}$...	$y_{p,L}$	$y_{p,*}$
\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	\ddots	\vdots	\vdots
r_q	$y_{q,1}$...	$y_{q,u}$...	$y_{q,v}$...	$y_{q,L}$	$y_{q,*}$
\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	\ddots	\vdots	\vdots
r_K	$y_{K,1}$...	$y_{K,u}$...	$y_{K,v}$...	$y_{K,L}$	$y_{K,*}$
Σ_k	$y_{*,1}$...	$y_{*,u}$...	$y_{*,v}$...	$y_{*,L}$	y

3 - Source frequencies

Vybírá se jedna z voleb nabízených v menu *Source frequencies*.

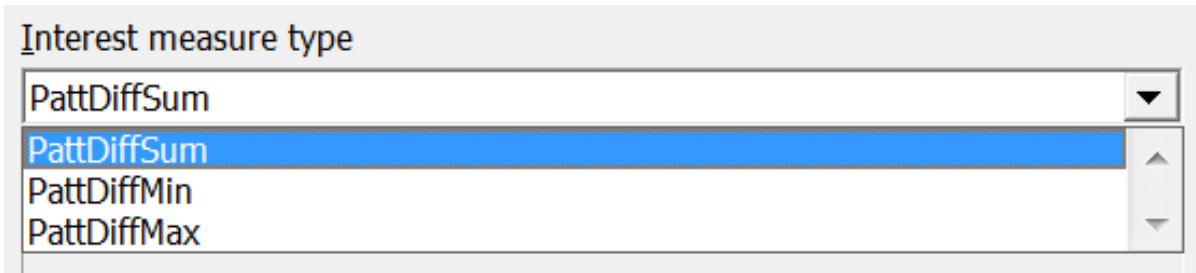


Na základě této volby se určí výseky tabulek $\{x_{i,j}\}_{i=p,\dots,q}^{j=u,\dots,v}$ a $\{y_{i,j}\}_{i=p,\dots,q}^{j=u,\dots,v}$ použité při verifikaci. Dále píšeme pouze $x_{i,j}$ místo $\{x_{i,j}\}_{i=p,\dots,q}^{j=u,\dots,v}$, atd. Platí:

- pro *Absolute number*
 - pokud není zaškrtnuta volba *Normalize pattern frequencies to SUM = 100*: $x_{i,j} = n_{i,j}$ a $y_{i,j} = w_{i,j}$
 - pokud je zaškrtnuta volba *Normalize pattern frequencies to SUM = 100*: $x_{i,j} = \frac{100}{n} n_{i,j}$ a $y_{i,j} = \frac{100}{w} w_{i,j}$
- pro *Relative [%] to act condition*: $x_{i,j} = \frac{n_{i,j}}{n}$ a $y_{i,j} = \frac{w_{i,j}}{w}$
- pro *Relative [%] for each row*: $x_{i,j} = \frac{n_{i,j}}{n_{i,*}}$ a $y_{i,j} = \frac{w_{i,j}}{w_{i,*}}$
- pro *Relative [%] for each column* $x_{i,j} = \frac{n_{i,j}}{n_{*,j}}$ a $y_{i,j} = \frac{w_{i,j}}{w_{*,j}}$

4 - Interest measure type

Vybírá se jedna z voleb nabízených v menu *Interest measure type*.

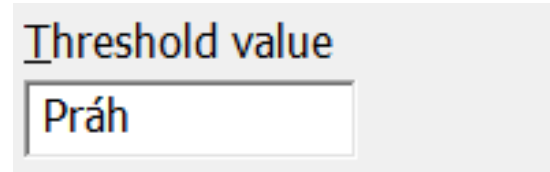
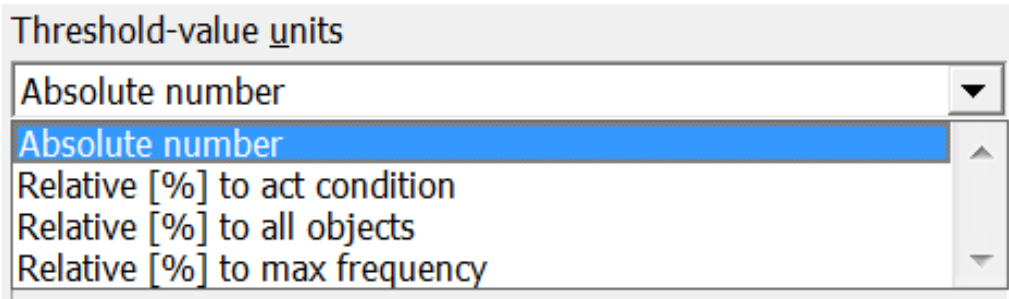


Na základě této volby se určí hodnota IM pro porovnání s prahem dle relace zadané parametrem 6. Platí:

- pro *PattDiffSum* se spočítá hodnota $IM = \sum_{i=p}^q \sum_{j=u}^v |x_{i,j} - y_{i,j}|$
- pro *PattDiffMin* se spočítá hodnota $IM = \min\{|x_i - y_i| \mid i = p, \dots, q; j = u, \dots, v\}$
- pro *PattDiffMax* se spočítá hodnota $IM = \max\{|x_i - y_i| \mid i = p, \dots, q; j = u, \dots, v\}$

5 - Threshold-value units

Vybírá se jedna z voleb nabízených v menu *Threshold-value units*

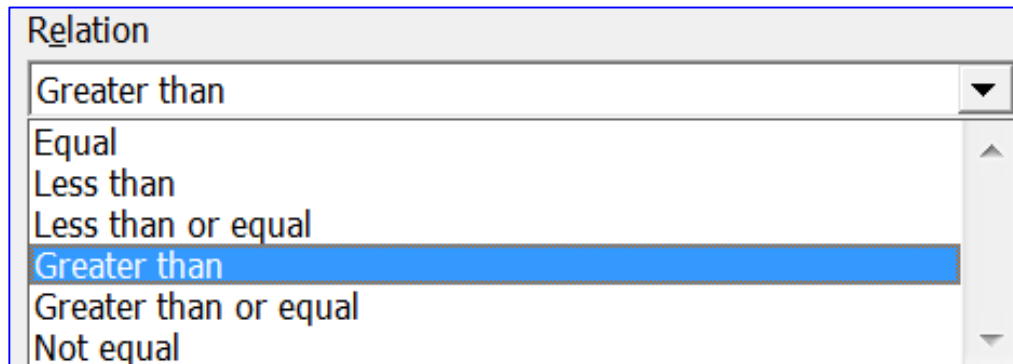


Na základě této volby se určí jednotky, ve kterých je hodnota *Práh* zadaná v poli *Threshold value*. To znamená, že se před porovnáním dle parametru 5 zadaná hodnota *Práh* vynásobí vhodným číslem dle volby v menu. Výslednou hodnotu pro porovnání značíme *PráhAkt*. Platí:

- pro *Absolute number* $PráhAkt = Práh$
- pro *Relative [%] to act condition* $PráhAkt = \frac{100}{n} * Práh$
- pro *Relative [%] to all objects* $PráhAkt = \frac{100}{n_T} * Práh$
- pro *Relative [%] to all objects* $PráhAkt = \frac{100}{Max} * Práh$ kde $Max = \max\{n_{i,j} | i = 1, \dots, K; j = 1, \dots, L\}$

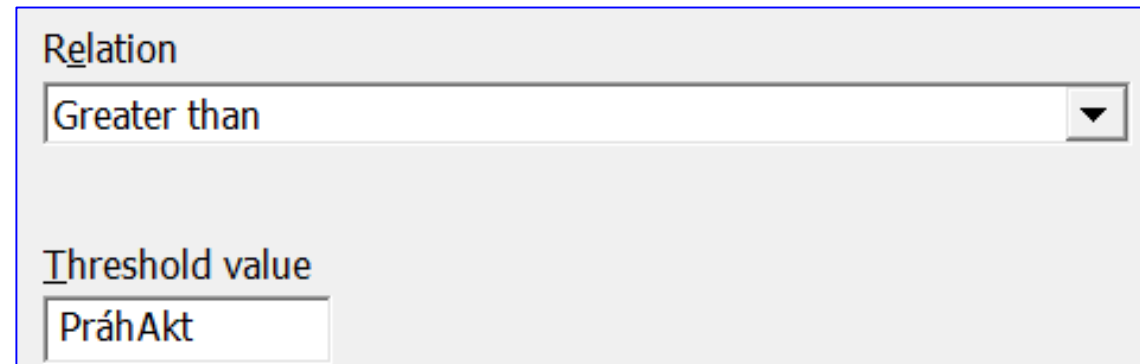
6 - Relation x Threshold value

Na základě volby v nabídce *Relation* se vybere relace, která se použije pro porovnání hodnoty *IM* vypočtené dle parametru 4 - *Inte*nterest measure type s hodnotou *PráhAkt* vypočtenou dle parametru 5 - *Threshold-value units*.



Relation

- Greater than
- Equal
- Less than
- Less than or equal
- Greater than
- Greater than or equal
- Not equal



Relation

Greater than

Threshold value

PráhAkt

Platnost vybrané relace je považována za podmínku definující KL-kvantifikátor zadaný parametry 1 až 6.