

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

Je dostupná z [této adresy](#)

Verze 30. 7. 2019

Zadávání statistických KL-kvantifikátorů pro proceduru KL-Miner

Jan Rauch

Katedra informačního a znalostního inženýrství

Vysoká škola ekonomická v Praze

Start tlačítkem Add statistical quantifier

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

- Definice se provádí zadáním parametrů 1 a 2.
- Pro všechny parametry se nabízejí defaultní hodnoty.
- Podmínka kvantifikátoru je definována parametrem 2.

KL Statistical quantifier settings

Interest measure type: Cramer's V coefficient

Relation: Greater than or equal

Threshold value: 0.9

Parameters: Absolute value of TauB for Kendall's coefficient (ie. interval <0;1> only)

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

Category Range: Rows 0 To 100, Columns 0 To 100

Formula:

Note:

Buttons: OK, Cancel

Výchozí KL-tabulka

KL-vztah $R \approx C/\chi$ se statistickým kvantifikátorem \approx se verifikuje na základě KL-tabulky $KL(R,C,M/\chi)$.

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

M/χ	c_1	\dots	c_L	Σ_l
r_1	$n_{1,1}$	\dots	$n_{1,L}$	$n_{1,*}$
\vdots	\vdots		\vdots	\vdots
r_K	$n_{K,1}$	\dots	$n_{K,L}$	$n_{K,*}$
Σ_k	$n_{*,1}$	\dots	$n_{*,L}$	n

Pokud $\chi = True$, pak se KL-vztah týká celé matice dat.

1 - Interest measure type

Vybere se jedna z měr zajímavosti nabízených v menu *Interest measure type* a přiřadí se jí hodnota dle tabulky

Interest measure type
Cramer's V coefficient
Cramer's V coefficient
Kendall's TauB coefficient
Chi-square test
Conditional entropy H(C R)
Mutual information MI(R,C) normalized
Inf. dependence ID(R,C)
Asymmetric information coefficient AIC(R,C)

Interest measure type	Hodnota IM
Cramer's V coefficient	$\sqrt{\frac{\frac{\chi^2}{n}}{\min\{K-1, L-1\}}}$
Kendall's TauB coefficient	$\frac{\sqrt{(n^2 - \sum_{k=1}^K n_{k,*}^2)(n^2 - \sum_{l=1}^L n_{*,l}^2)}}{2(P-Q)}$, kde $P = \sum_{k=1}^K \sum_{l=1}^L n_{k,l} \sum_{i=k+1}^K \sum_{j=l+1}^L n_{i,j}$, $Q = \sum_{k=1}^K \sum_{l=1}^L n_{k,l} \sum_{i=k+1}^K \sum_{j=1}^{l-1} n_{i,j}$
Chi-square test	$\chi^2 = n \left(\sum_{k=1}^K \sum_{l=1}^L \frac{n_{k,l}^2}{n_{k,*} n_{*,l}} - 1 \right)$
Conditional entropy H(C R)	$PodmEntr = \sum_{k=1}^K \sum_{l=1}^L \frac{n_{k,l}}{n} \log_2 \frac{n_{k,l}}{n_{k,*}}$
Mutual information MI(R,C) normalized	$\frac{H(C) - PodmEntr}{\min\{H(C), H(R)\}}$, kde $H(C) = \sum_{l=1}^L \frac{n_{*,l}}{n} \log_2 \frac{n_{*,l}}{n}$, $H(R) = \sum_{k=1}^K \frac{n_{k,*}}{n} \log_2 \frac{n_{k,*}}{n}$
Inf. dependence ID(R,C)	$1 - \frac{PodmEntr}{H(C)}$
Asymmetric information coefficient AIC(R,C)	$1 - \frac{\sum_{k=1}^K n_{k,*} \log_2(n_{k,*}) - \sum_{k=1}^K \sum_{l=1}^L n_{k,l} \log_2(n_{k,l})}{n \log_2(n) - \sum_{l=1}^L n_{*,l} \log_2(n_{*,l})}$

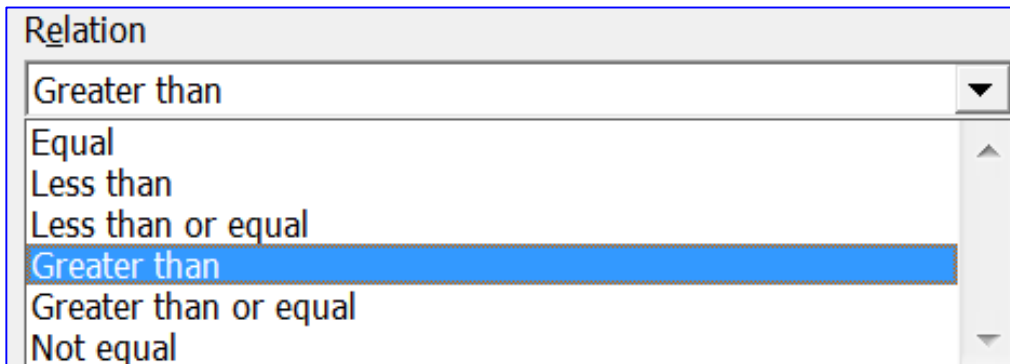
V případě Kendall's TauB coefficient je možno zaškrtnout volbu.
Potom se místo *IM* bere v úvahu $|IM|$.

Parameters

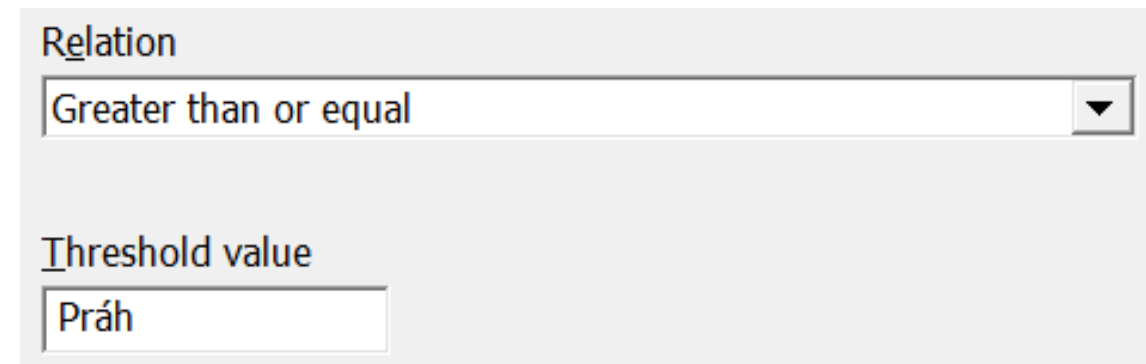
- Absolute value of TauB for Kendall's coefficient (ie. interval <0;1> only)

2 - Relation x Threshold value

Na základě volby v nabídce *Relation* se vybere relace, která se použije pro porovnání hodnoty *IM* vypočtené dle parametru 2 - *Inte*rest measure type s hodnotou *Práh*.



A screenshot of a software interface showing a dropdown menu labeled "Relation". The menu is open, displaying a list of comparison operators: "Greater than", "Equal", "Less than", "Less than or equal", "Greater than", "Greater than or equal", and "Not equal". The "Greater than" option is highlighted with a blue background.



A screenshot of a software interface showing a dropdown menu labeled "Relation" with "Greater than or equal" selected. Below it, a text input field labeled "Threshold value" contains the text "Práh".

Platnost vybrané relace je považována za podmínku definující KL-kvantifikátor zadaný parametry 1 a 2.