Tato prezentace je součástí wiki-prezentace Metoda GUHA a systém LISp-Miner
Je dostupná z této adresy
Verse 31. 7. 2019

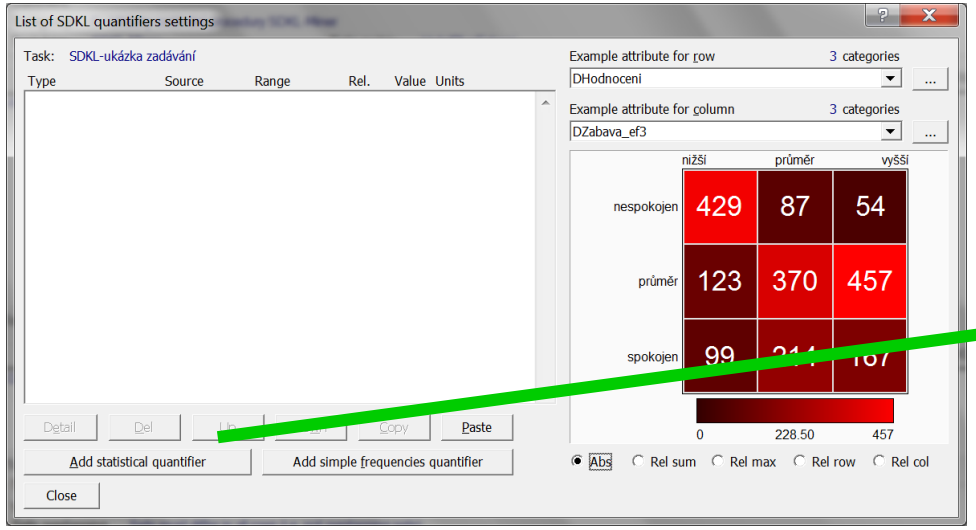
Zadávání statistických SDKL-kvantifikátorů pro proceduru SDKL-Miner

Jan Rauch

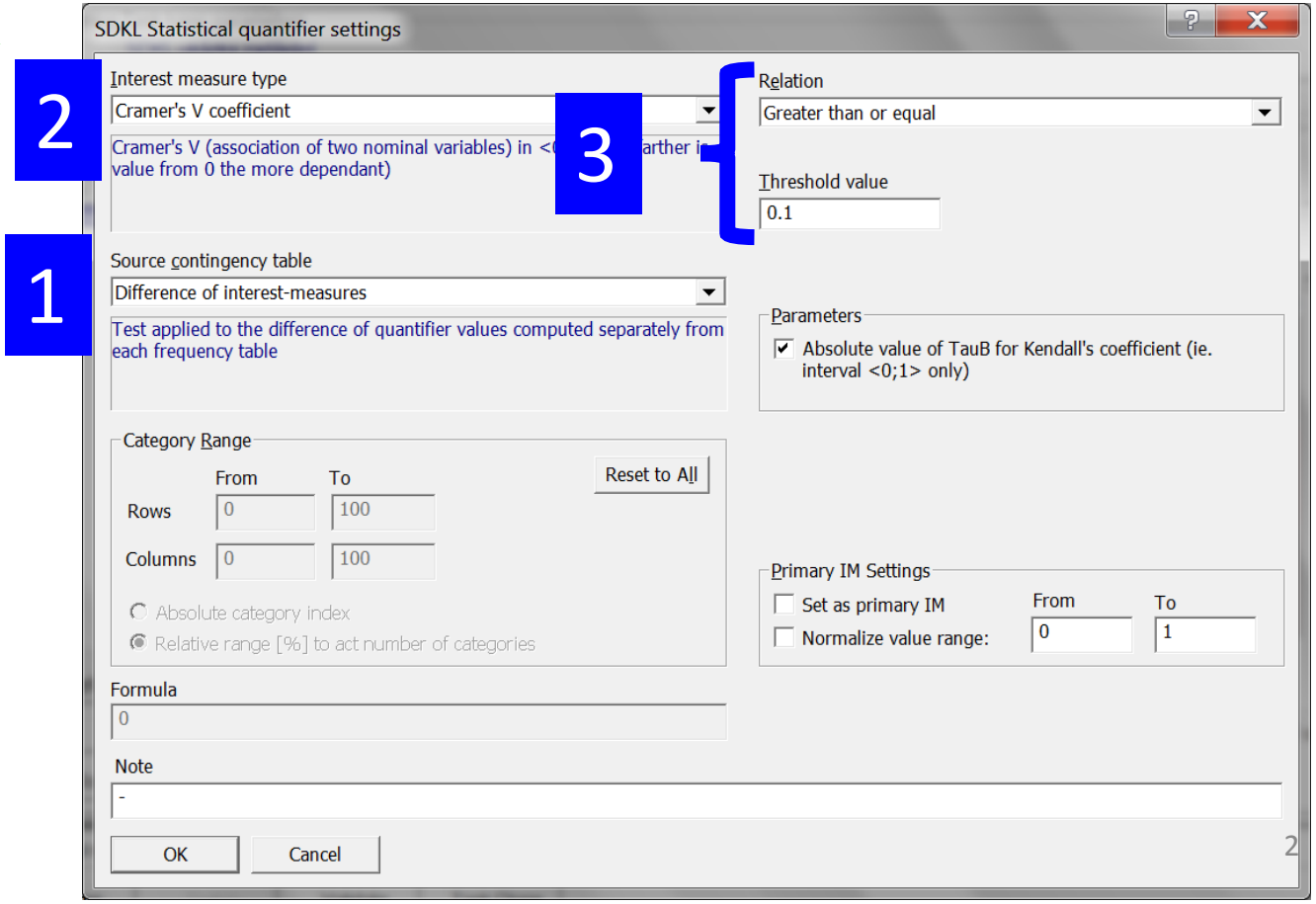
Katedra informačního a znalostního inženýrství

Vysoká škola ekonomická v Praze

Start tlačítkem Add statistical quantifier



- Definice se provádí zadáním parametrů 1 - 3.
- Pro všechny parametry se nabízejí defaultní hodnoty.
- Podmínka kvantifikátoru je definována parametrem 3.



Výchozí KL-tabulky

Statistický SDKL-kvantifikátor se aplikuje na dvojici KL-tabulek $\mathrm{TKL}_\alpha = \mathrm{KL}(\mathrm{R},\mathrm{C},\mathrm{M}/\chi\wedge\alpha)$ a $\mathrm{TKL}_\beta = \mathrm{KL}(\mathrm{R},\mathrm{C},\mathrm{M}/\chi\wedge\beta)$.

$\mathrm{M}/\chi\wedge\alpha$	c_1	$\dots$	c_L	Σ_l
r_1	$n_{\alpha,1,1}$	$\dots$	$n_{\alpha,1,L}$	$n_{\alpha,1,*}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
r_K	$n_{\alpha,K,1}$	$\dots$	$n_{\alpha,K,L}$	$n_{\alpha,K,*}$
Σ_k	$n_{\alpha,*,1}$	$\dots$	$n_{\alpha,*,L}$	n_α

$\mathrm{TKL}_\alpha = \mathrm{KL}(\mathrm{R},\mathrm{C},\mathrm{M}/\chi\wedge\alpha)$

$\mathrm{M}/\chi\wedge\beta$	c_1	$\dots$	c_L	Σ_l
r_1	$n_{\beta,1,1}$	$\dots$	$n_{\beta,1,L}$	$n_{\beta,1,*}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
r_K	$n_{\beta,K,1}$	$\dots$	$n_{\beta,K,L}$	$n_{\beta,K,*}$
Σ_k	$n_{\beta,*,1}$	$\dots$	$n_{\beta,*,L}$	n_β

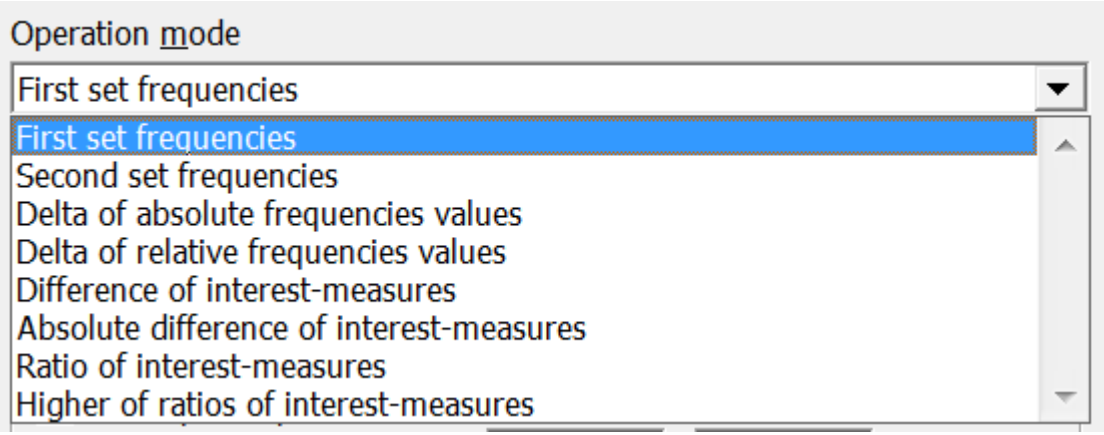
$\mathrm{TKL}_\beta = \mathrm{KL}(\mathrm{R},\mathrm{C},\mathrm{M}/\chi\wedge\beta)$

Tabulky zapisujeme i ve formě $\mathrm{TKL}_\alpha = \{n_{\alpha,i,j}\}_{i=1,\dots,K}^{j=1,\dots,L}$ a $\mathrm{TKL}_\beta = \{n_{\beta,i,j}\}_{i=1,\dots,K}^{j=1,\dots,L}$.

Pokud není nebezpečí nedorozumění, píšeme pouze $\mathrm{TKL}_\alpha = \{n_{\alpha,i,j}\}$ a $\mathrm{TKL}_\beta = \{n_{\beta,i,j}\}$.

1 - Operation mode

Vybírá se jeden z operačních módů nabízených v menu *Operation mode.* Ten určuje, jakým způsobem bude aplikována vybraná míra zajímavosti na KL-tabulky TKL_α a TKL_β.



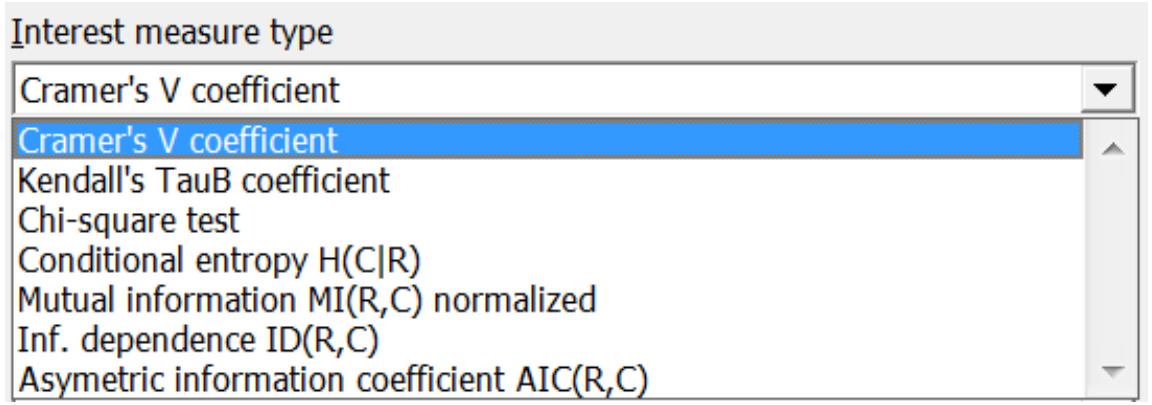
Pro zbývající čtyři nabízené módy se vybraná míra zajímavosti aplikuje zvlášť na každou z tabulek TKL_α a TKL_β. Poté se výsledky dále zpracují.

Pro první čtyři nabízené Operation mode se nejprve vytvoří KL-tabulka $\{z_{i,j}\}_{i=1,\dots,K}^{j=1,\dots,L}$ takto:

- pro *First set frequencies* $\{z_{i,j}\} = \{n_{\alpha,i,j}\}$
- pro *Second set frequencies* $\{z_{i,j}\} = \{n_{\beta,i,j}\}$
- pro *Delta of absolute frequencies values* $\{z_{i,j}\} = \{|n_{\alpha,i,j} - n_{\beta,i,j}|\}$
- pro *Delta of relative frequencies values* $\{z_{i,j}\} = \{|\frac{n_{\alpha,i,j}}{n_\alpha} - \frac{n_{\beta,i,j}}{n_\beta}|\}$.

2 - Interest measure type (A)

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



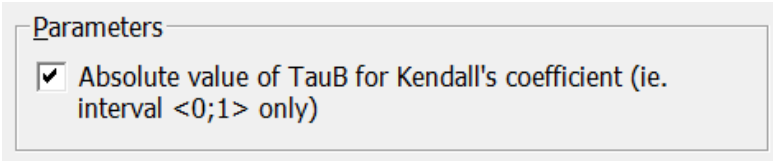
Pro výpočet hodnoty *IM* pro první čtyři operační módy

- *First set frequencies*
- *Second set frequencies*
- *Delta of absolute frequencies values,*
- *Delta of relative frequencies values*

se pro jednotlivé volby *Interest measure type* použije KL-tabulka $\{z_{i,j}\}$ přiřazená výše uvedeným způsobem.

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

V případě Kendall's TauB coefficient je možno zaškrtnout volbu



.

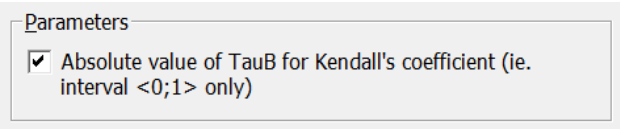
Potom se místo *IM* bere v úvahu | *IM* |.

2 - Interest measure type (B)

Při výpočtu hodnoty *IM* pro zbývající čtyři operační módy se nejprve se vypočtou hodnoty IM_α a IM_β pro vybraný *Interest measure type*:

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

V případě Kendall's TauB coefficient je možno zaškrtnout volbu



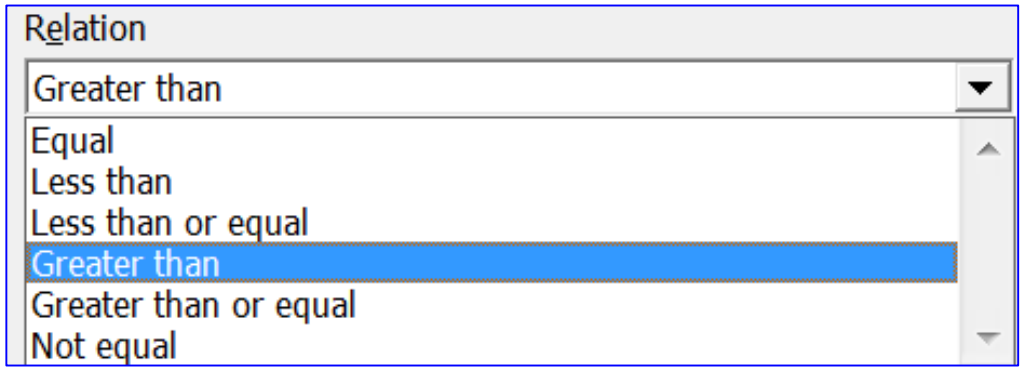
Potom se místo IM_α bere v úvahu $|\,IM_\alpha\,|$ a místo IM_β se bere v úvahu $|\,IM_\beta\,|$.

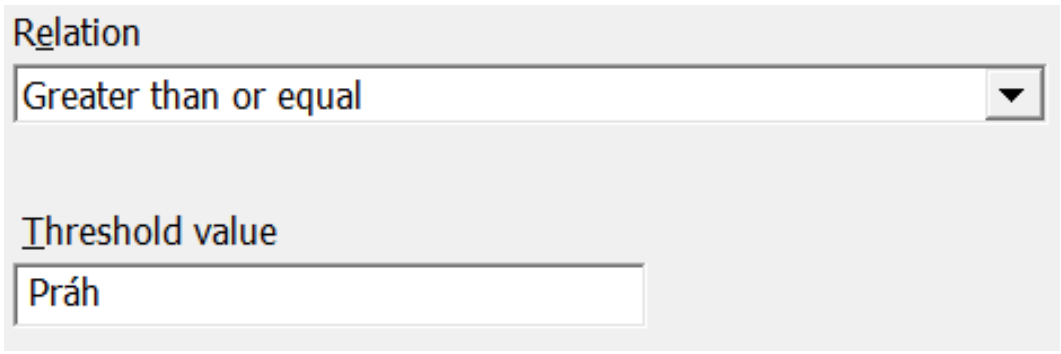
Poté se vypočte hodnota IM v závislosti na vybraném *Operation mode*:

Operation mode	IM
Difference of interest-measures	$IM_\alpha - IM_\beta$
Absolute difference of interest-measures	$\lvert IM_\alpha - IM_\beta\rvert$
Ratio of interest-measures	$\frac{IM_\alpha}{IM_\beta}$
Higher of ratios of interest-measures	$\max\{\frac{IM_\alpha}{IM_\beta}, \frac{IM_\beta}{IM_\alpha}\}$

3 - Relation x Threshold value

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





Platnost vybrané relace je považována za podmínku definující SDCF-kvantifikátor zadaný parametry 1 až 3.