Tato prezentace je součástí wiki-prezentace Metoda GUHA a systém LISp-Miner
Je dostupná z této adresy
Verse 28. 7. 2019

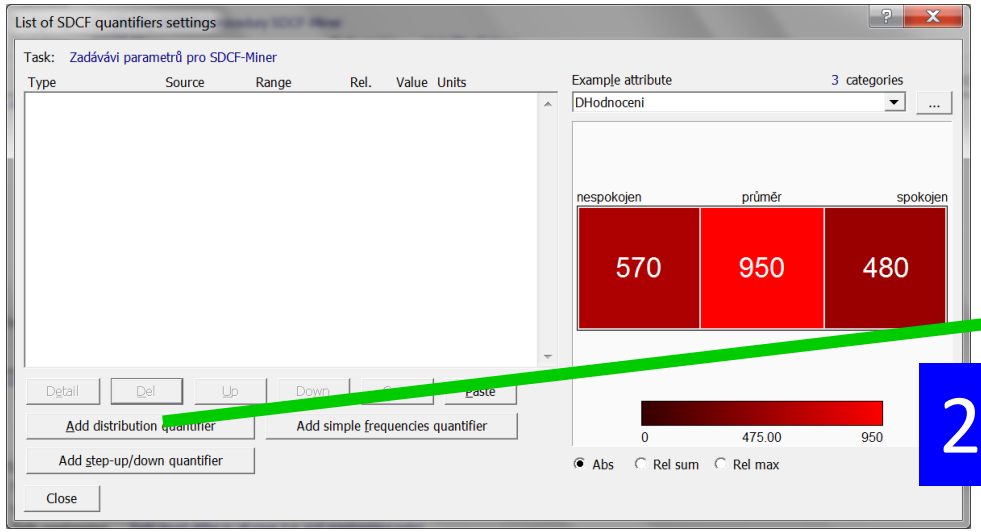
Zadávání distribučních SDCF-kvantifikátorů pro proceduru SDCF-Miner

Jan Rauch

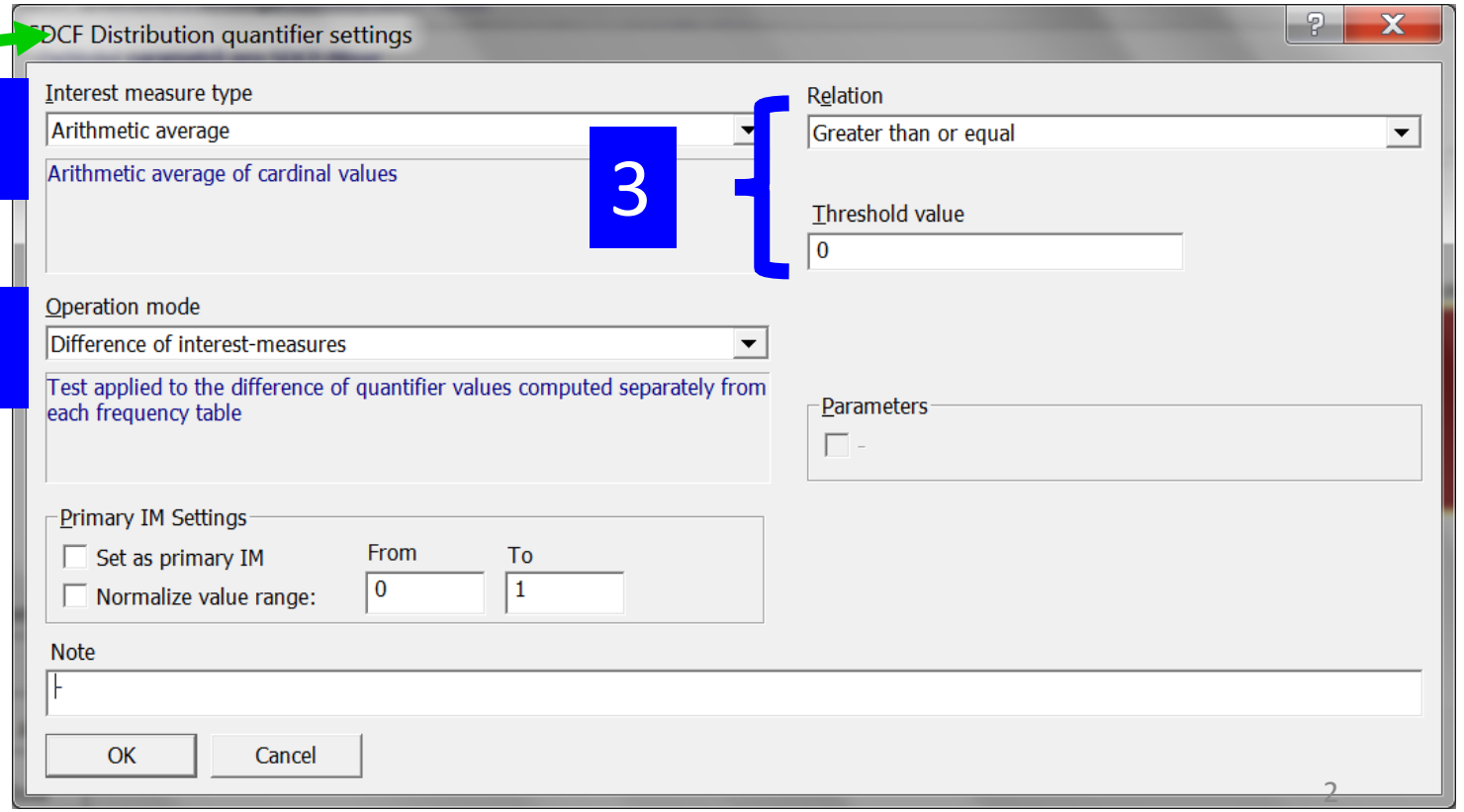
Katedra informačního a znalostního inženýrství

Vysoká škola ekonomická v Praze

Start tlačítkem Add simple frequencies quantifier



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



Výchozí CF-tabulky

Distribuční SDCF-kvantifikátor se aplikuje na CF-tabulky $CF_\alpha = CF(A, \chi \wedge \alpha, M)$ a $CF_\beta = CF(A, \chi \wedge \beta, M)$

$CF_\alpha =$

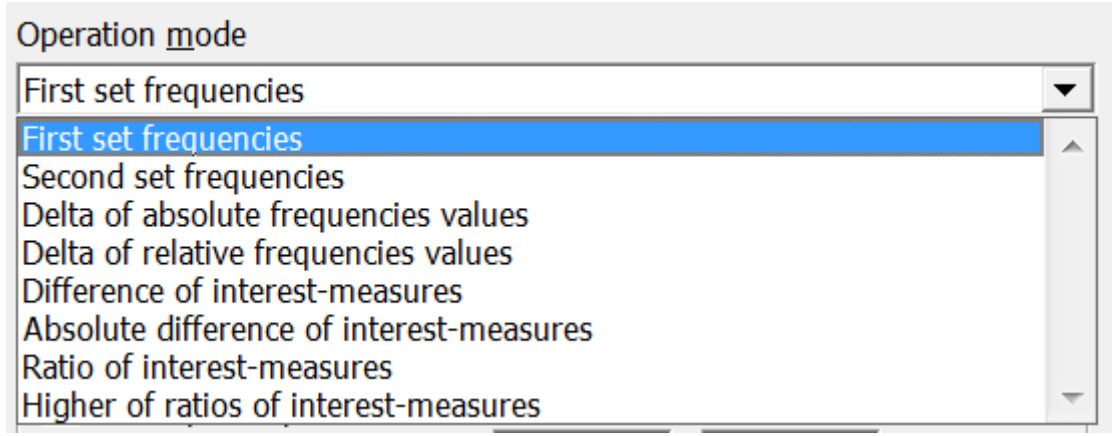
četnosti kategorií pro atribut A	a_1	…	a_K	Σ
absolutní četnosti v matici $M/\chi \wedge \alpha$	$n_{\alpha,1}$	…	$n_{\alpha,K}$	n_α

$CF_\beta =$

četnosti kategorií pro atribut A	a_1	…	a_K	Σ
absolutní četnosti v matici $M/\chi \wedge \beta$	$n_{\beta,1}$	…	$n_{\beta,K}$	n_β

1 - Operation mode

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



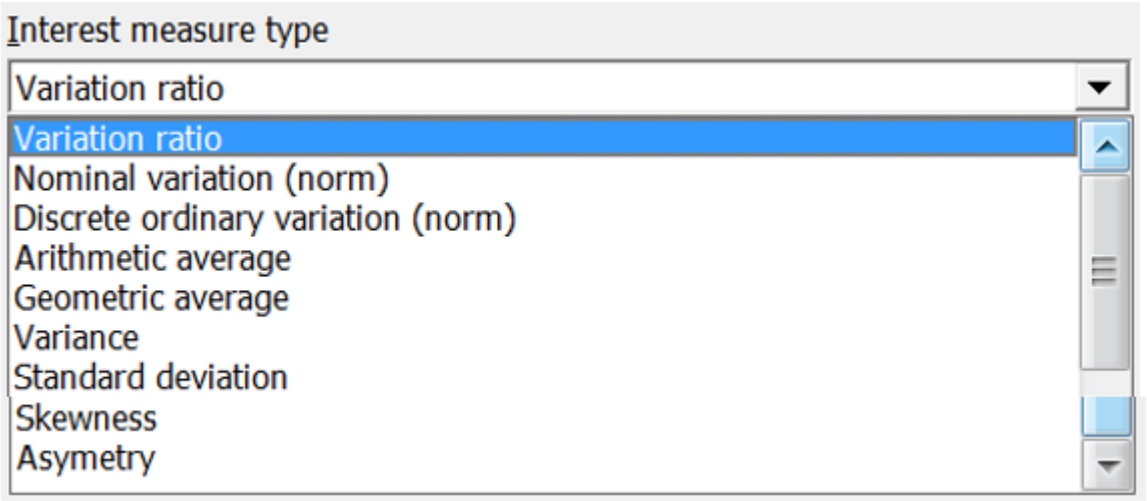
Pro zbývající čtyři nabízené módy se vybraná míra zajímavosti aplikuje zvlášť na každou z tabulek CF_α a CF_β. Poté se výsledky dále zpracují.

Pro první čtyři nabízené Operation mode se nejprve vytvoří CF-tabulka $\langle z_1, \ldots, z_K \rangle$ takto:

- pro *First set frequencies* $\langle z_1, \ldots, z_K \rangle = \langle n_{\alpha,1}, \ldots, n_{\alpha,K} \rangle$
- pro *Second set frequencies* $\langle z_1, \ldots, z_K \rangle = \langle n_{\beta,1}, \ldots, n_{\beta,K} \rangle$
- pro *Delta of absolute frequencies values* $\langle z_1, \ldots, z_K \rangle = \langle |n_{\alpha,1} - n_{\beta,1}|, \ldots, |n_{\alpha,K} - n_{\beta,K}| \rangle$
- pro *Delta of relative frequencies values* $\langle z_1, \ldots, z_K \rangle = \langle \left|\frac{n_{\alpha,1}}{n_\alpha} - \frac{n_{\beta,1}}{n_\beta}\right|, \ldots, \left|\frac{n_{\alpha,K}}{n_\alpha} - \frac{n_{\beta,K}}{n_\beta}\right| \rangle$.

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 CF-tabulka $\langle z_1, \ldots, z_K \rangle$ přiřazená výše uvedeným způsobem.

Interest measure type	Hodnota IM	Poznámka
Variation ratio	$1 - \max\{z_1, \ldots, z_K\}$	
Nominal variation (norm)	$\frac{K}{K-1}\Sigma_{i=1}^{K} z_i(1-z_i)$	
Discrete ordinary variation (norm)	$\frac{2}{K-1}\Sigma_{i=1}^{K} F_i(1-F_i)$	$F_i = \Sigma_{j=1}^{i} z_j, i = 1, \ldots, K$
Arithmetic average	$\Sigma_{i=1}^{K} z_i a_i$	pouze pro kardinální atributy
Geometric average	$\Pi_{i=1}^{K} a_i^{z_i}$	pouze pro kardinální atributy
Variance	$\Sigma_{i=1}^{K} f_i(a_i - AvgA)^2$	$AvgA = \Sigma_{i=1}^{K} z_i a_i$
Standard deviation	$\sqrt{\Sigma_{i=1}^{K} f_i(a_i - AvgA)^2}$	dále značíme jako $StdDev$
Skewness	$\frac{\Sigma_{i=1}^{K} f_i(a_i - AvgA)^3}{StdDev^3}$	
Asymetry	$\frac{N_2 - N_1}{n}$	$N_1 = \Sigma z_i$ pro $a_i < AvgA$ $N_2 = \Sigma z_i$ pro $a_i > AvgA$

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*:

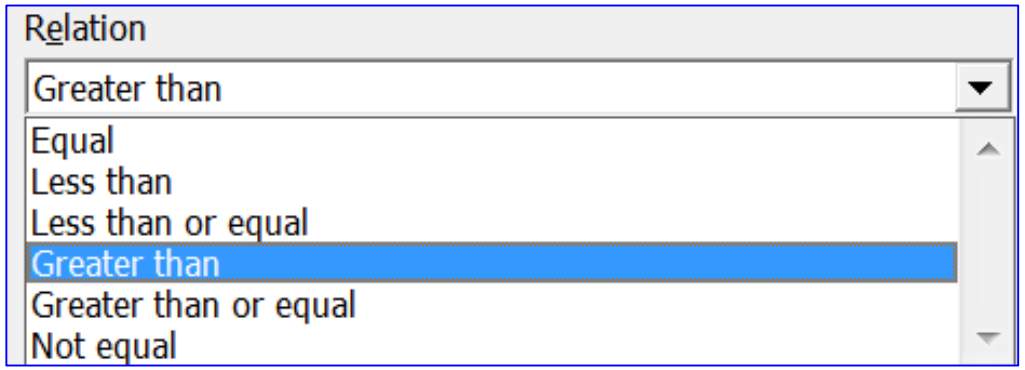
Interest measure type	Hodnota IM_α	Hodnota IM_β
Variation ratio	$1 - \max\{n_{\alpha,1}, \ldots, n_{\alpha,K}\}$	$1 - \max\{n_{\beta,1}, \ldots, n_{\beta,K}\}$
Nominal variation (norm)	$\frac{K}{K-1}\Sigma_{i=1}^{K} n_{\alpha,i}(1 - n_{\alpha,i})$	$\frac{K}{K-1}\Sigma_{i=1}^{K} n_{\beta,i}(1 - n_{\beta,i})$
Discrete ordinary variation (norm)	$\frac{2}{K-1}\Sigma_{i=1}^{K} F_{\alpha,i}(1 - F_{\alpha,i})$, kde $F_{\alpha,i} = \Sigma_{j=1}^{i} n_{\alpha,j}$	$\frac{2}{K-1}\Sigma_{i=1}^{K} F_{\beta,i}(1 - F_{\beta,i})$, kde $F_{\beta,i} = \Sigma_{j=1}^{i} n_{\beta,j}$
Arithmetic average ($= AvgA_{\alpha,\beta}$)	$\Sigma_{i=1}^{K} n_{\alpha,i} a_i$	$\Sigma_{i=1}^{K} n_{\beta,i} a_i$
Geometric average	$\Pi_{i=1}^{K} a_i^{n_{\alpha,i}}$	$\Pi_{i=1}^{K} a_i^{n_{\beta,i}}$
Variance	$\Sigma_{i=1}^{K} f_i (a_i - AvgA_\alpha)^2$	$\Sigma_{i=1}^{K} f_i (a_i - AvgA_\beta)^2$
Standard deviation ($= StdDev_{\alpha,\beta}$)	$\sqrt{\Sigma_{i=1}^{K} f_i (a_i - AvgA_\alpha)^2}$	$\sqrt{\Sigma_{i=1}^{K} f_i (a_i - AvgA_\beta)^2}$
Skewness	$\frac{\Sigma_{i=1}^{K} f_i (a_i - AvgA_\alpha)^3}{StdDev_\alpha^3}$	$\frac{\Sigma_{i=1}^{K} f_i (a_i - AvgA_\beta)^3}{StdDev_\beta^3}$
Asymetry	$\frac{N_{\alpha,2} - N_{\alpha,1}}{n}$ $N_{\alpha,1} = \Sigma n_{\alpha,i};\ a_i < AvgA_\alpha$ $N_{\alpha,2} = \Sigma n_{\alpha,i};\ a_i > AvgA_\alpha$	$\frac{N_{\beta,2} - N_{\beta,1}}{n}$ $N_{\beta,1} = \Sigma n_{\beta,i};\ a_i < AvgA_\beta$ $N_{\beta,2} = \Sigma n_{\beta,i};\ a_i > AvgA_\beta$

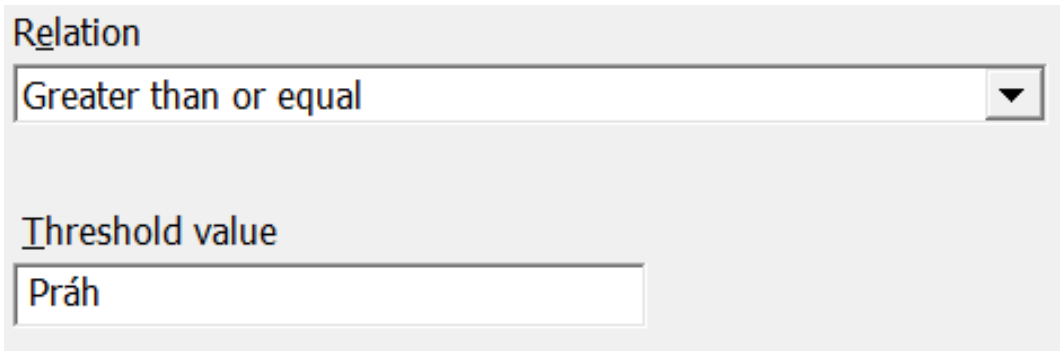
Poté se vypočte hodnota IM v závislosti na vybraném *Operation mode*:

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 a 2.