Tato prezentace je součástí wiki-prezentace Metoda GUHA, LISp-Miner a typové úlohy
Je dostupná z této adresy
Verse 17. 8. 2019

Typ úlohy: Vzdálenost histogramů od vzoru malá

Data: Hotel

Problém: Je dán histogram atributu jako vzor a zajímá nás, za jakých okolností jsou vzdálenosti histogramu téhož atributu od vzoru malé (histogramy si jsou podobné)

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

© Jan Rauch

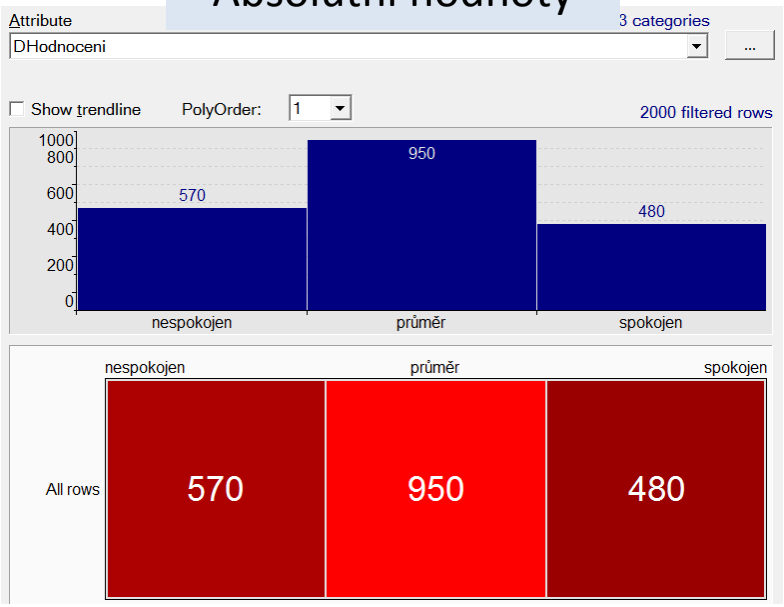
Vzdálenost histogramů

- Vzdálenost histogramů - definice
- Podobnost histogramů
- Poznámky

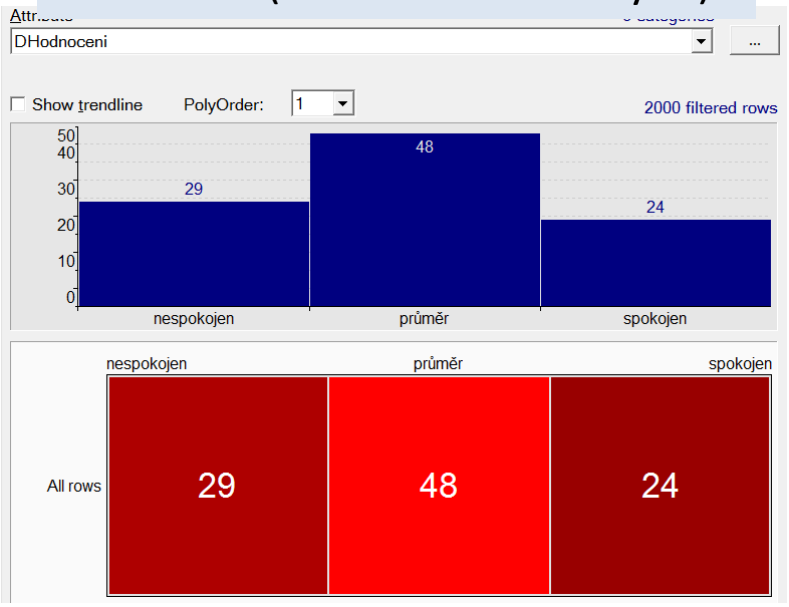
Vzdálenost histogramů – motivace

V celé matici dat je rozložení kategorií atributu DHodnoceni následující:

Absolutní hodnoty



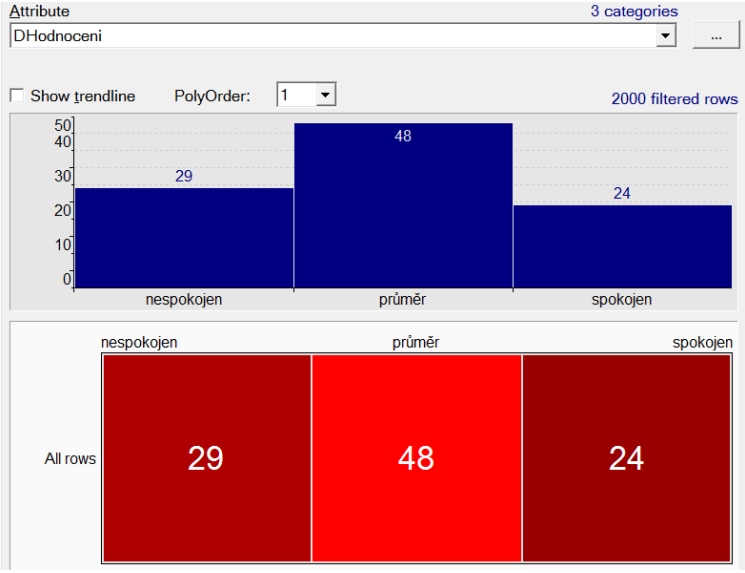
Procenta (zaokrouhlovací chyba)



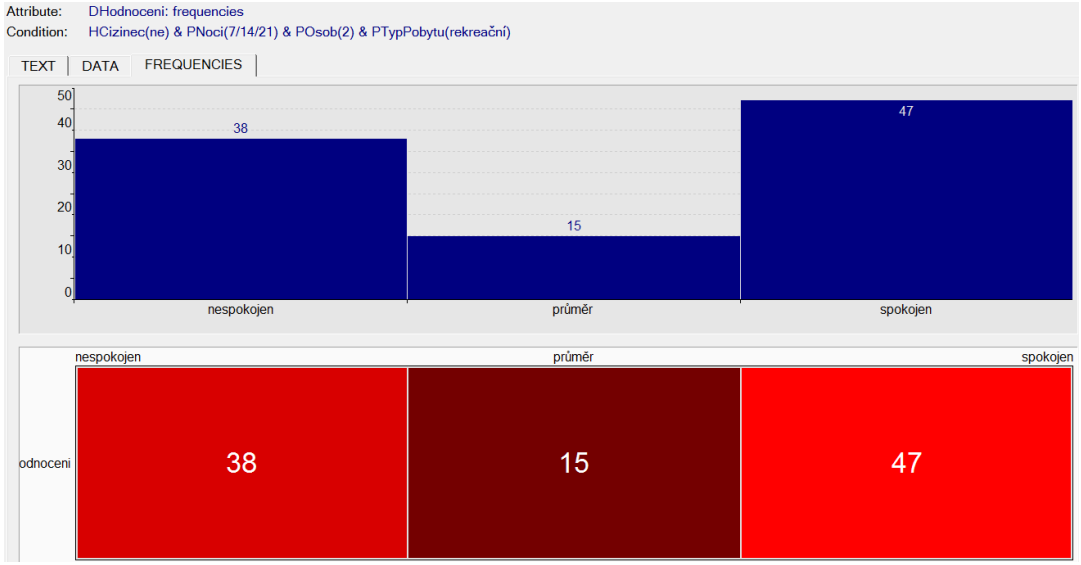
Pro kterou kombinaci údajů o hostovi a jeho bydlišti, o počasí, o počtu nocí a o pobytu se rozložení kategorií atributu DHodnocení nejvíce liší od jejich rozložení pro celou matici?

L1-vzdálenost histogramů DHodnoceni v %Cond pro celou matici a pro podmínku χ, příklad

HotelPlusExterni



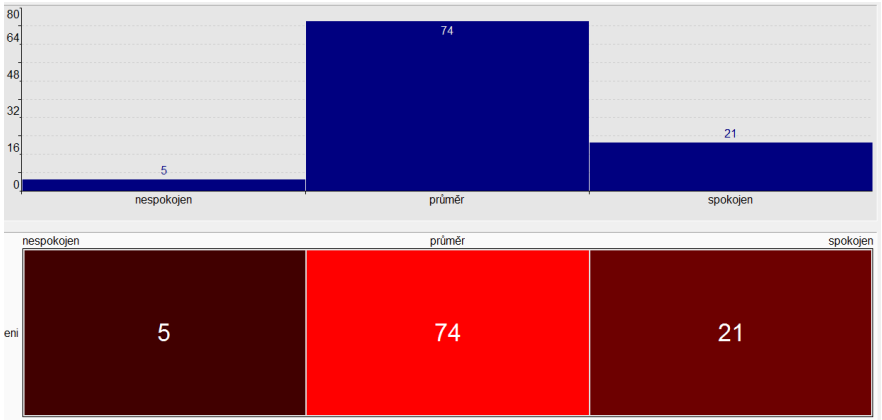
HotelPlusExterni / χ



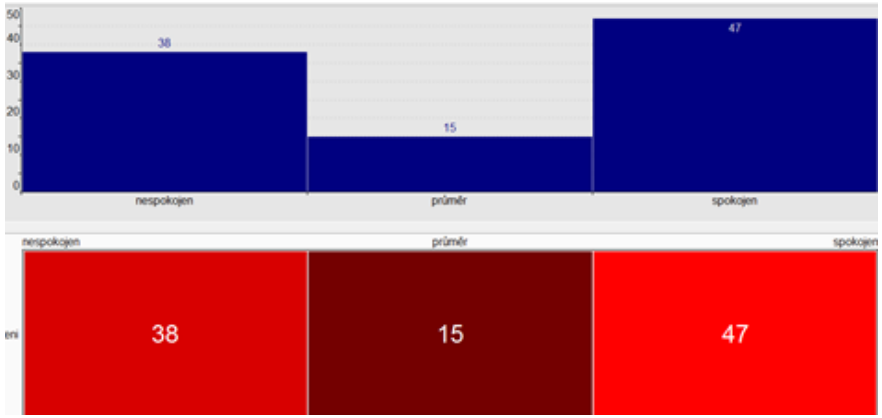
L1(HotelPlusExterni, DHodnoceni, % Cond, all, χ) = |29 - 38| + |48 - 15| + |24 - 47| = 9 + 33 + 23 = 65

L1-vzdálenost histogramů atributu Dhodnoceni v %Cond pro podmínky χ_1 a χ_2, příklad

Dhodnoceni/χ_1 v %Cond



Dhodnoceni/χ_2 v %Cond



L1(HotelPlusExterni, Dhodnoceni, % Cond, χ_1, χ_2) = |5-38| + |74-15| + |21-47| = 33 +59 + 26 = 108

L1-vzdálenost histogramů atributu A v %Cond pro podmínky χ_1 a χ_2

Předpoklady:

- Matice dat M má m řádků, podmínku χ_1 splňuje n_1 řádků, podmínku χ_2 splňuje n_2 řádků.
- Atribut A má kategorie $a_1,\ldots, a_K$

Kategorie atributu A	a_1	…	a_K	Σ
absolutní četnosti kategorií v matici dat M	m_1	…	m_K	m
absolutní četnosti kategorií v matici dat M/χ_1	g_1	…	g_K	n_1
absolutní četnosti kategorií v matici dat M/χ_2	h_1	…	h_K	n_2
velikosti sloupců v histogramu A/all v %Cond (pro celou M)	$100*m_1/m$	…	$100*m_K/m$	100
velikosti sloupců v histogramu A/χ_1 v %Cond	$100*g_1/n_1$	…	$100*g_K/n_1$	100
velikosti sloupců v histogramu A/χ_2 v %Cond	$100*h_1/n_2$	…	$100*f_K/n_2$	100

$$L1(M, A, \%\ Cond, \chi_1, \chi_2) = |100*g_1/n_1 - 100*h_1/n_2| + \ldots + |\ 100*g_K/n_1 - 100*h_K/n_2\ |$$

$$L1(M, A, \%\ Cond, all, \chi_1) = |100*m_1/m - 100*g_1/n_1| + \ldots + |\ 100*m_K/m - 100*g_K/n_1\ |$$

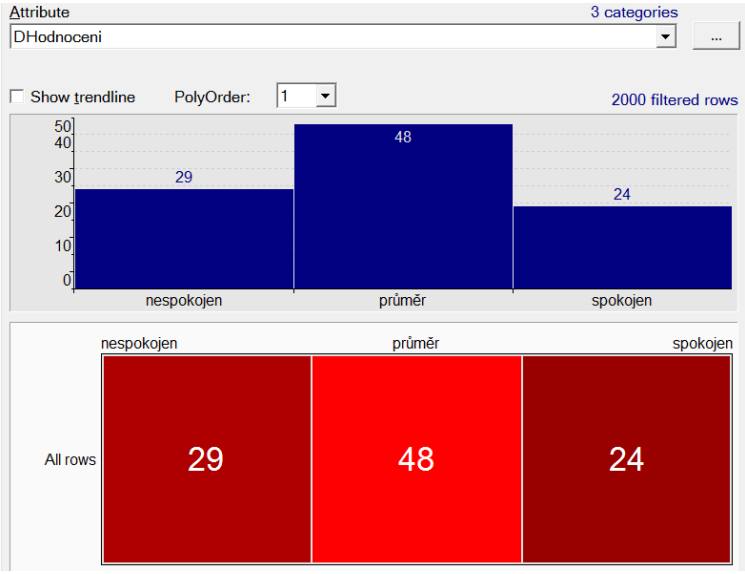
Vzdálenost histogramů

- Vzdálenost histogramů - definice
- Podobnost histogramů
- Poznámky

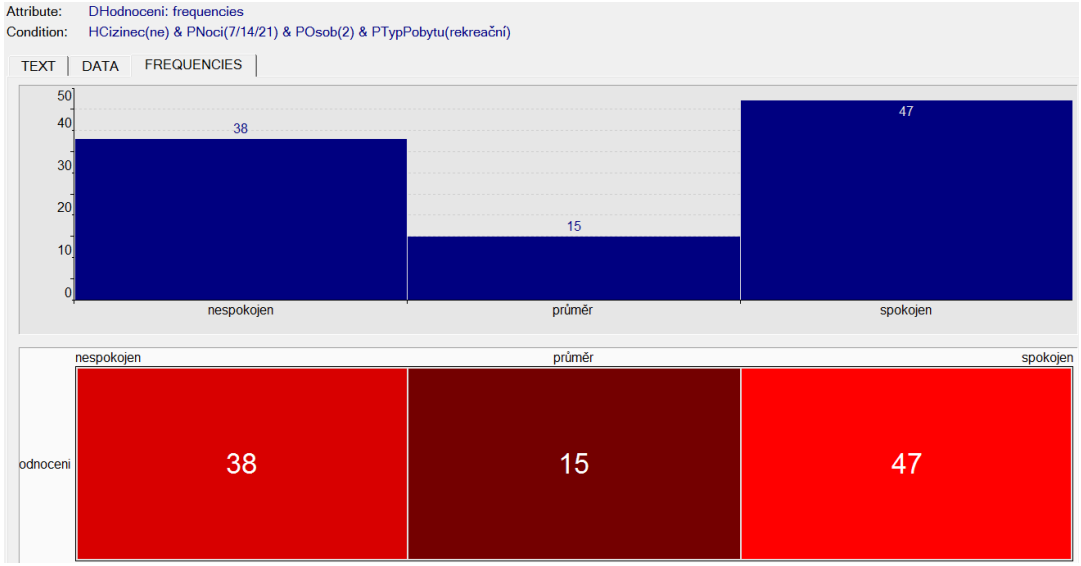
Podobnost histogramů – motivace

Od histogramu atributu DHodnocení se nejvíce liší histogram pro
χ = HCizinec(ne) $\wedge$ PNoci(7/14/21) $\wedge$ POsob(2) $\wedge$ PTypPobytu(rekreační),
viz příklad uvedený zde

All



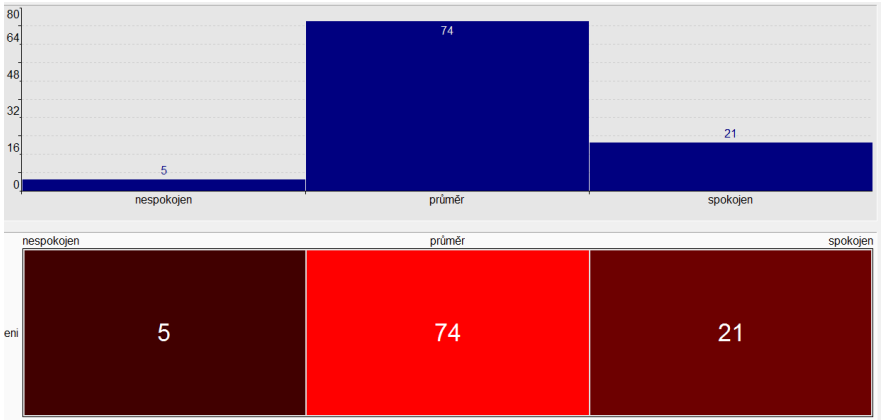
χ = Hcizinec(ne) $\wedge$ Pnoci(7/14/21) $\wedge$ POsob(2) $\wedge$ PTypPobytu(rekreační)



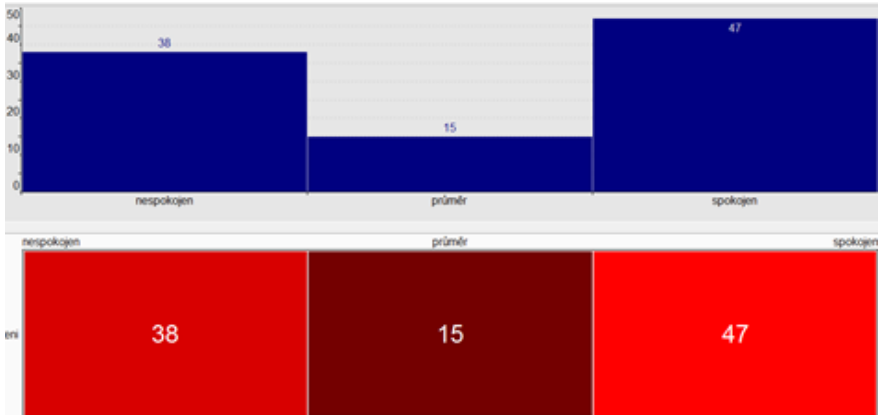
Pro které kombinace údajů o hostovi a jeho bydlišti, o počasí, o počtu nocí a o pobytu se rozložení kategorií atributu DHodnocení podobá rozložení pro χ?

L1-vzdálenost histogramů atributu Dhodnoceni v %Cond pro podmínky χ_1 a χ_2

Dhodnoceni/χ_1 v %Cond



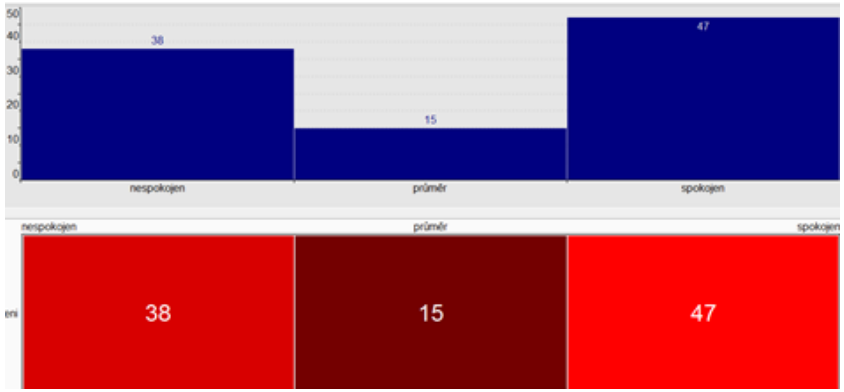
Dhodnoceni/χ_2 v %Cond



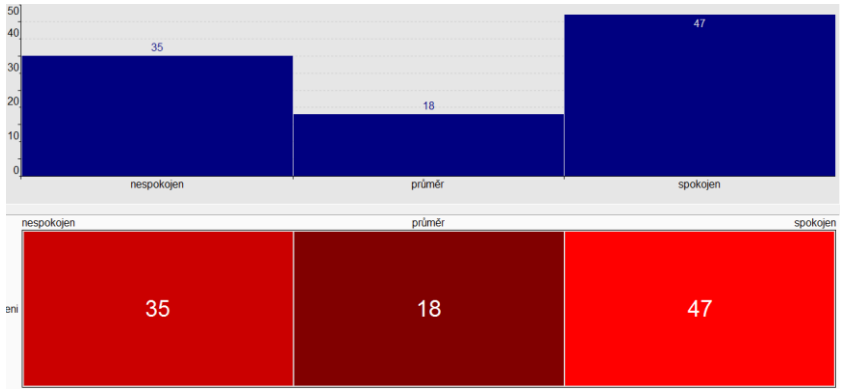
L1(HotelPlusExterni, Dhodnoceni, % Cond, χ_1 , χ_2) = |5-38| + |74-15| + |21-47| = 33 +59 + 26 = 108

Idea – histogramy si jsou podobné, pokud jejich L1-vzdálenost je malá

Dhodnoceni/χ_1 v %Cond



Dhodnoceni/χ_2 v %Cond

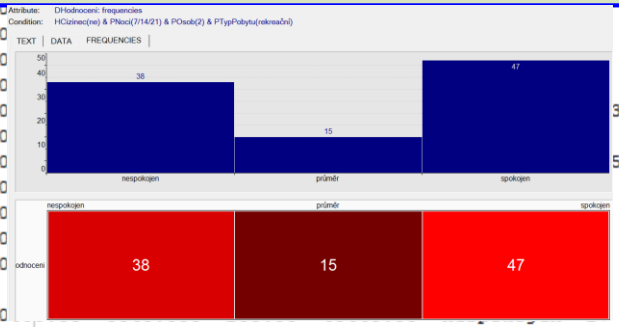


L1(HotelPlusExterni, Dhodnoceni, % Cond, χ_1 , χ_2) = |38-35| + |15-18| + |47-47| = 3 + 3 + 0 = 6

Podobnost histogramů - aplikace procedury CF-Miner, příklad PH1

Hotel.txt

Pro které kombinace údajů o hostovi a jeho bydlišti, o počasí, o počtu nocí a o pobytu je histogram atributu DHodnoceni v % podobný histogramu PŘÍKLAD pro kombinaci χ = Hcizinec(ne) $\wedge$ Pnoci(7/14/21) $\wedge$ POsob(2) $\wedge$ PTypPobytu(rekreační)



? CF: L1 $_{\leq 10\ \%C}$ DHodnoceni / Host, Bydliště, Meteo, PNoci, Pobyt

Meteo.txt

MDatum	MTeplota	MObloha
4.1.2012	-6.3	slunečno
5.1.2012	-6	
6.1.2012	6.	
7.1.2012	1.	
8.1.2012	-1.3	srážky
	7	zataženo
		srážky
		zataženo
12.1.2012	-3.1	srážky
13.1.2012	-8.1	zataženo
14.1.2012	-10.7	srážky
15.1.2012	-5.5	zataženo
16.1.2012	2.3	zataženo
17.1.2012	-1.9	zataženo
18.1.2012	-8.6	zataženo

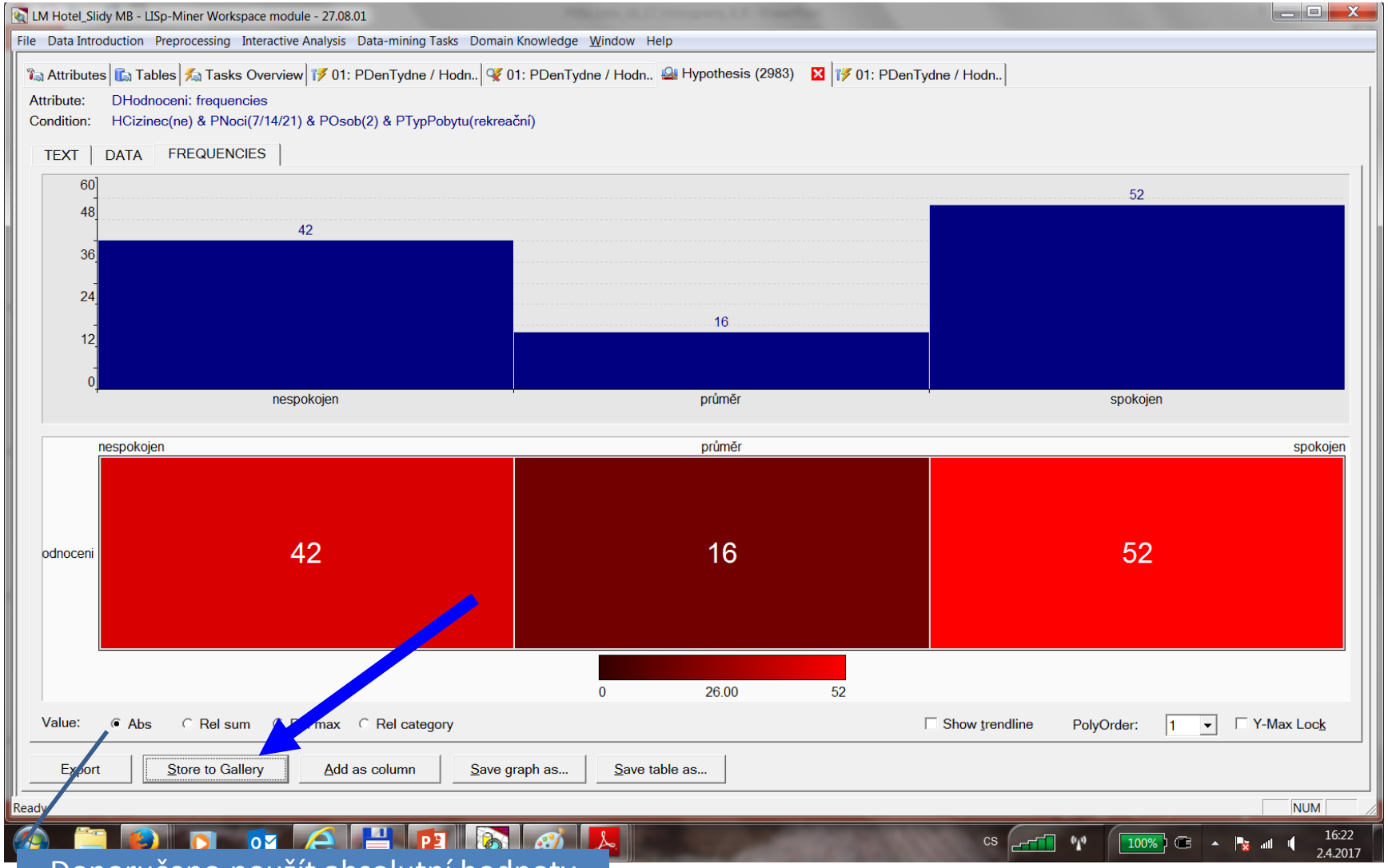
Hledání okolností, za kterých je histogram daného atributu podobný danému histogramu

Uložení daného histogramu do galerie

⇩

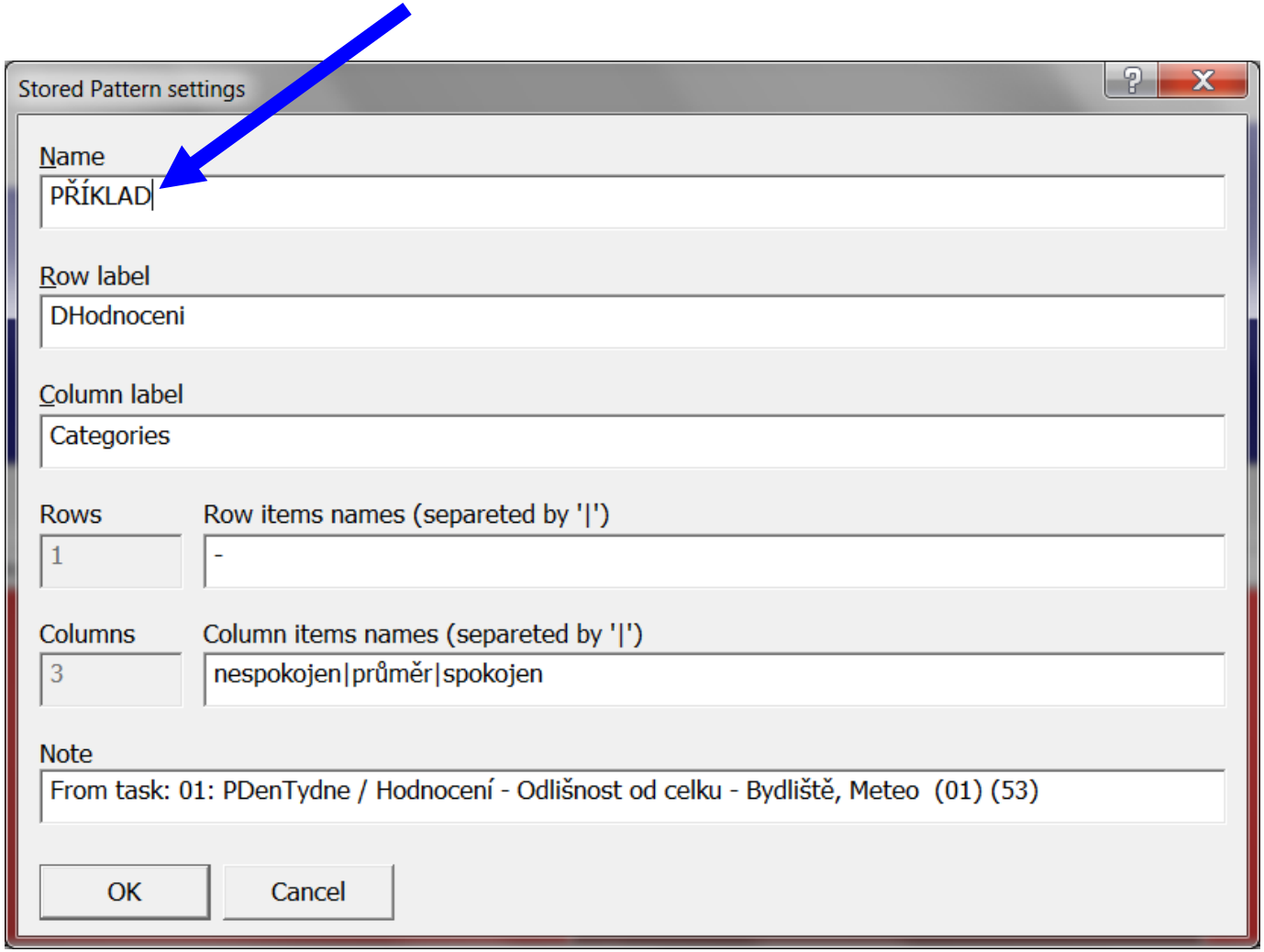
Aplikace procedury CF-Miner s rozdílovým kvantifikátorem (pattern-difference quantifier) a s výběrem daného histogramu uloženého v galerii

Uložení χ = HCizinec(ne) $\wedge$ PNoci(7/14/21) $\wedge$ POsob(2) $\wedge$ PTypPobytu(rekreační) do galerie pod jménem PŘÍKLAD (1)



Doporučeno použít absolutní hodnoty kvůli přesnějším výpočtům

Uložení χ = HCizinec(ne) $\wedge$ Pnoci(7/14/21) $\wedge$ POsob(2) $\wedge$ PTypPobytu(rekreační) do galerie pod jménem PŘÍKLAD (2)



? CF: L1 $_{\leq 10\ \%\ C}$ DHodnocení / Host, Bydliště, Meteo, Pobyt, zadání pro CF-Miner

LM Hotel_Slidy MB - LISp-Miner Workspace module - 27.08.01

File Data Introduction Preprocessing Interactive Analysis Data-mining Tasks Domain Knowledge Window Help

Attributes | Tables | Tasks Overview | 01: PDenTydne / Hodn.. | 01: PDenTydne / Hodn..

Data-mining Task basic parameters

Name: 01: PDenTydne / Hodnocení - Podobnost s PŘÍKLAD — ID: 55

Comment: -

Taskgroup: 01: CF-Miner

Task type: CF-Miner — Data matrix: HotelPlusExterni — Edit

ATTRIBUTES FOR HISTOGRAM

Attributes

» DHodnocení(3 categories)

DHodnocení

QUANTIFIERS

Type	Source	Range	Rel.	Value	Units
SUM	Abs	all	>=	50	Abs
PDSUM	%Cond	all	<=	10	Abs

SUM(Abs) ≥ 50

L1 $_{\leq 10\ \%C}$

CONDITION

Host/Bydliště	Con, 0 - 3
» HCizinec_b (subset), 1 - 1	B, pos
» HMesto (subset), 1 - 1	B, pos
» HStat (subset), 1 - 1	B, pos
» HPohlavi (subset), 1 - 1	B, pos
» HVek_exp (subset), 1 - 1	B, pos
Meteo	Con, 0 - 2
» MObloha (subset), 1 - 1	B, pos
» MTeplota_exp (subset), 1 - 1	B, pos
Počet nocí	Con, 0 - 4
» PNoci_enum_m (subset), 1 - 1	B, pos
» PNoci_exp (subset), 1 - 1	B, pos
» POsob (subset), 1 - 1	B, pos
» POsobonoci_ef5 (subset), 1 - 1	B, pos
Pobyt	Con, 0 - 1
» PPresSobotniNoc (subset), 1 - 1	B, pos
» PTurnus (subset), 1 - 1	B, pos
» PTypPobytu (subset), 1 - 1	B, pos
» PDenTydne (subset), 1 - 1	B, pos
» PMesic (subset), 1 - 1	B, pos
» PRok (subset), 1 - 1	B, pos

Total length: 1 - 10

Generation information

Status: Solved, 31 run(s)

Mode: Standard

Aggregate function

Type: Count(*)

Attribute: -

Task parameters

Include 'worse' extensions of condition: Yes — Maximal number of hypotheses: 1000

Include extensions of coefficients with no change in the histogram: Yes

Include extensions of cedents with no change in the histogram: Yes — Extensions minimal length check: Yes

Params | Validate | Task Clone | Run | Bkgrnd Run | Grid Run | Show Results

Ready

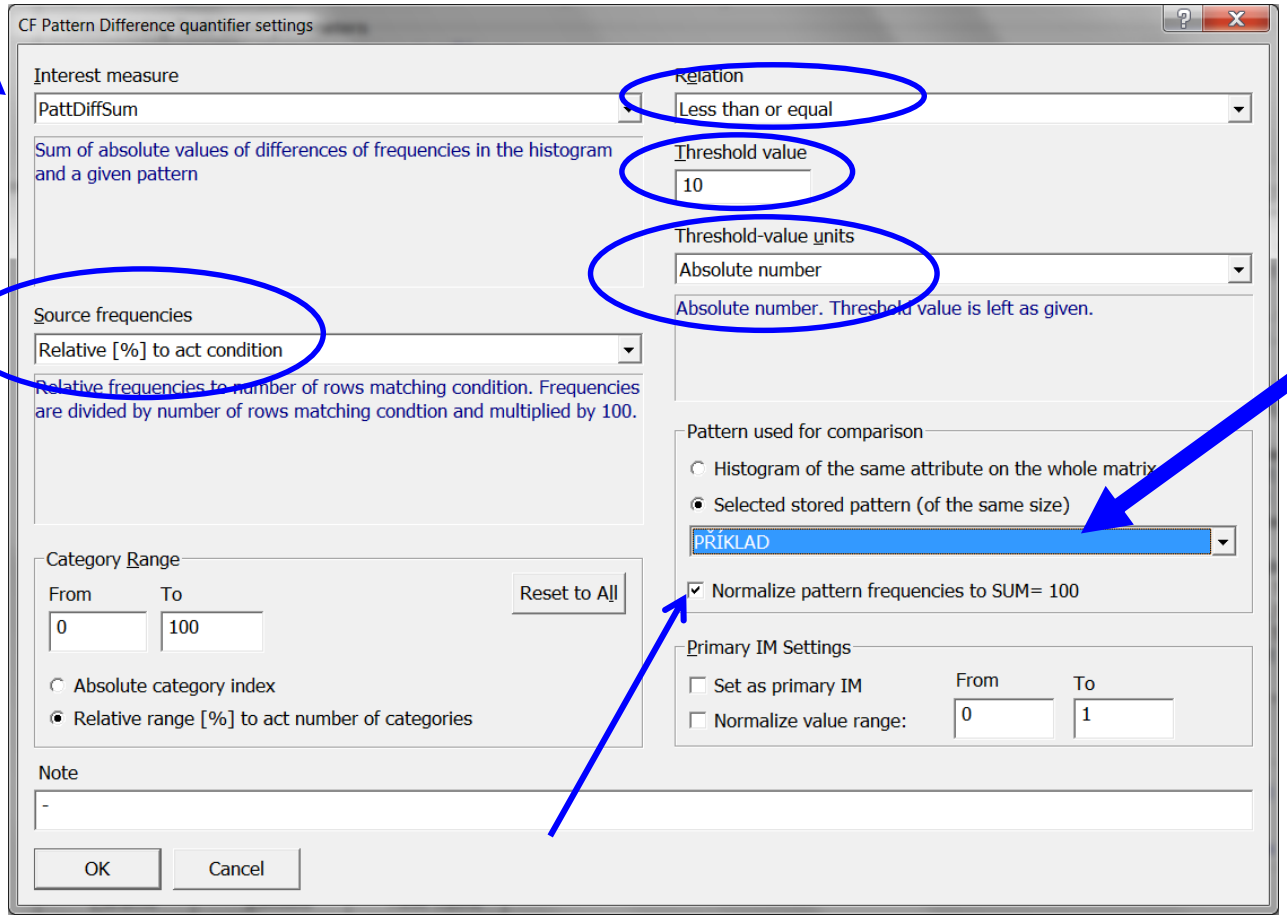
Host/Bydliště(*) ∧ Meteo(*)
Pobyt (*) ∧ Pobyt(*)

Jedno z mnoha možných zadání pro řešení dané analytické otázky.

CF-kvantifikátor L1 ≤ 10 %C

QUANTIFIERS

Type	Source	Range	Rel.	Value	Units
SUM	Abs	all	>=	50	Abs
PDSUM	%Cond	all	<=	10	Abs



? CF: L1 $_{\leq 10\,\%\,C}$ DHodnocení / Host, Bydliště, Meteo, Pobyt, výstup procedury CF-Miner

Task run

Start: 2.4.2017 17:03:37 Total time: 0h 0m 1s

Number of verifications: 21197

Number of hypotheses: 96 Mode: Standard

Add group | Del group | Edit group

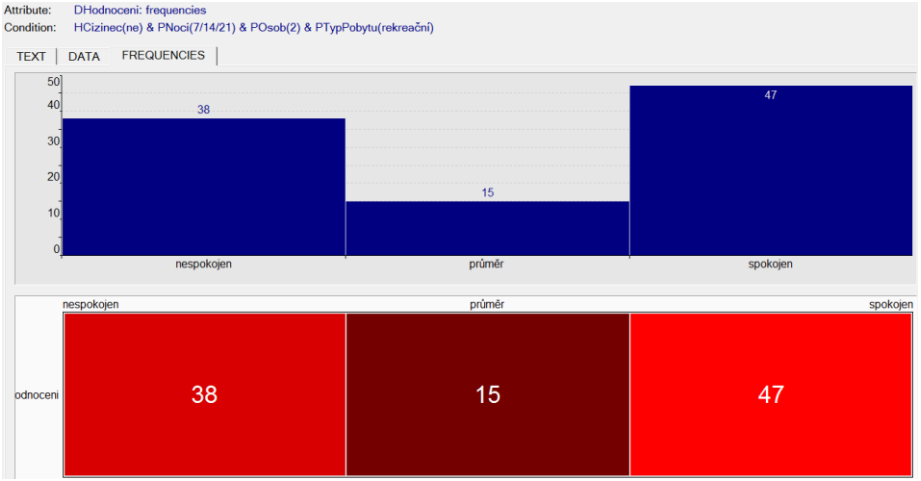
Actual group of hypotheses: All hypotheses

Hypotheses in group: 96 Shown hypotheses: 96 Highlighted: 0

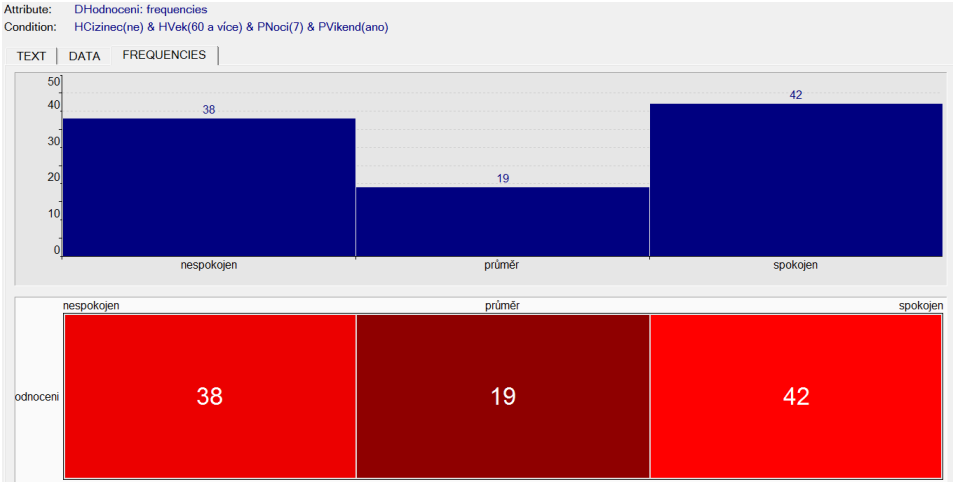
Nr.	Id	PDSUM	Hypothesis
1	34	9.930	**DHodnoceni / HCizinec**(*ne*) & **HVek**(*60 a více*) & **PNoci**(*7*) & **PVikend**(*ano*)
2	35	9.930	**DHodnoceni / HStat**(*ČR*) & **HVek**(*60 a více*) & **PNoci**(*7*) & **PVikend**(*ano*)
3	58	9.800	**DHodnoceni / HCizinec**(*ne*) & **PNoci**(*7/14/21*) & **POsobonoci**(*vyšší*) & **PTurnus**(*ano*)
4	59	9.800	**DHodnoceni / HCizinec**(*ne*) & **PNoci**(*7/14/21*) & **POsobonoci**(*vyšší*) & **PDenTydne**(*So*)
5	53	9.800	**DHodnoceni / HCizinec**(*ne*) & **POsobonoci**(*vyšší*) & **PTurnus**(*ano*)
6	15	9.800	**DHodnoceni / HMesto**(*Praha*) & **HPohlavi**(*muž*) & **PTypPobytu**(*rekreační*)
7	89	9.800	**DHodnoceni / HStat**(*ČR*) & **PNoci**(*7/14/21*) & **POsobonoci**(*vyšší*) & **PTurnus**(*ano*)
8	90	9.800	**DHodnoceni / HStat**(*ČR*) & **PNoci**(*7/14/21*) & **POsobonoci**(*vyšší*) & **PDenTydne**(*So*)
9	84	9.800	**DHodnoceni / HStat**(*ČR*) & **POsobonoci**(*vyšší*) & **PTurnus**(*ano*)
10	20	9.697	**DHodnoceni / HCizinec**(*ne*) & **HPohlavi**(*žena*) & **PNoci**(*7/14/21*) & **POsob**(*2*) & **PTypPobytu**(*rekreační*)
11	23	9.697	**DHodnoceni / HStat**(*ČR*) & **HPohlavi**(*žena*) & **PNoci**(*7/14/21*) & **POsob**(*2*) & **PTypPobytu**(*rekreační*)
12	30	9.614	**DHodnoceni / HCizinec**(*ne*) & **HVek**(*od 28 do 60*) & **MTeplota**(*zima*) & **PRok**(*2012*)
13	33	9.614	**DHodnoceni / HStat**(*ČR*) & **HVek**(*od 28 do 60*) & **MTeplota**(*zima*) & **PRok**(*2012*)
14	4	9.545	**DHodnoceni / MTeplota**(*zima*) & **PNoci**(*7/14/21*) & **PMesic**(*leden*)
15	66	9.483	**DHodnoceni / HMesto**(*Plzeň*) & **PTypPobytu**(*rekreační*)

? CF: L1 $_{\leq 10\ \%\ C}$ DHodnocení / Host, Bydliště, Meteo, Pobyt, podrobnosti k prvnímu histogramu

PŘÍKLAD
HCizinec(ne) ∧ PNoci(7/14/21) ∧ POsob(2) ∧ PTypPobytu(rekreační)



První histogram χ
HCizinec(ne) ∧ HVek(60 a více) ∧ PNoci(7) ∧ PVikend(ano)



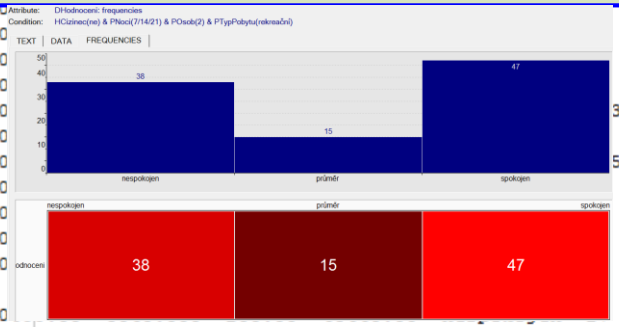
L1(HotelPlusExterni, DHodnoceni, % Cond, PŘÍKLAD, χ) = |38-38| + |15-19| + |47-42| = 0 + 4 + 5 = 9

Poznámka: Pokud použijeme zaokrouhlené hodnoty na celá procenta, dochází k rozdílům od hodnot PDSUM uvedeným v přehledovém výstupu procedury CF-Miner i k součtu procent odlišnému od 100.

Podobnost histogramů - aplikace procedury CF-Miner, příklad PH2

Hotel.txt

Pro které kombinace údajů o hostovi a jeho bydlišti, o počasí, o počtu nocí a o pobytu je histogram atributu DHodnoceni v % podobný histogramu PŘÍKLAD pro kombinaci χ = Hcizinec(ne) $\wedge$ PNoci(7/14/21) $\wedge$ POsob(2) $\wedge$ PTypPobytu(rekreační)



? CF: L1 $_{\leq 10\ \%C}$ DHodnoceni / Host, Bydliště, Meteo, PNoci, Pobyt

Jako příklad PH1, ale přidán CF-kvantifikátor MAXDiff $\leq$ 5.

Meteo.txt

? CF: L1 $_{\leq 10\,\%\,C}$ DHodnocení / Host, Bydliště, Meteo, Pobyt, zadání pro CF-Miner

LM Hotel_Slidy MB - LISp-Miner Workspace module - 27.08.01

File Data Introduction Preprocessing Interactive Analysis Data-mining Tasks Domain Knowledge Window Help

Attributes | Tables | Tasks Overview | 01: PDenTydne / Hodn.. | 01: PDenTydne / Hodn.. | 01: PDenTydne / Hodn.. | 01: PDenTydne / Hodn.. | Hypothesis (3264) | Hypothesis (3245)

Data-mining Task basic parameters

Name: 01: PDenTydne / Hodnocení - Podobnost s PŘÍKLAD Kviz — ID: 57

Comment: -

Taskgroup: 01: CF-Miner

Task type: CF-Miner — Data matrix: HotelPlusExterni — Edit

ATTRIBUTES FOR HISTOGRAM

Attributes
» DHodnoceni(3 categories)

DHodnocení

QUANTIFIERS

Type	Source	Range	Rel.	Value	Units
SUM	Abs	all	>=	50	Abs
PDSUM	%Cond	all	<=	10	Abs
PDMAX	%Cond	all	<=	5	Abs

SUM(Abs) $\geq$ 50

L1 $_{\leq 10\,\%C}$

MaxDiff $\leq$ 5

CONDITION

Host/Bydliště — Con, 0 - 3
» HCizinec_b (subset), 1 - 1 — B, pos
» HMesto (subset), 1 - 1 — B, pos
» HStat (subset), 1 - 1 — B, pos
» HPohlavi (subset), 1 - 1 — B, pos
» HVek_exp (subset), 1 - 1 — B, pos
Meteo — Con, 0 - 2
» MObloha (subset), 1 - 1 — B, pos
» MTeplota_exp (subset), 1 - 1 — B, pos
Počet nocí — Con, 0 - 4
» PNoci_enum_m (subset), 1 - 1 — B, pos
» PNoci_exp (subset), 1 - 1 — B, pos
» POsob (subset), 1 - 1 — B, pos
» POsobonoci_ef5 (subset), 1 - 1 — B, pos
Pobyt — Con, 0 - 1
» PPresSobotniNoc (subset), 1 - 1 — B, pos
» PTurnus (subset), 1 - 1 — B, pos
» PTypPobytu (subset), 1 - 1 — B, pos
» PDenTydne (subset), 1 - 1 — B, pos
» PMesic (subset), 1 - 1 — B, pos
» PRok (subset), 1 - 1 — B, pos
Total length: 1 -

Host/Bydliště(*) $\wedge$ Meteo(*)
Pobyt (*) $\wedge$ Pobyt(*)

Aggregate function

Type: Count(*)

Attribute: -

Task parameters

Include 'worse' extensions of condition: Yes — Maximal number of hypotheses: 1000

Include extensions of coefficients with no change in the histogram: Yes

Include extensions of cedents with no change in the histogram: Yes — Extensions minimal length check: Yes

Params | Validate | Task Clone

Run | Bkgrnd Run | Grid Run | Show Results

Ready — NUM — 19:29 3.4.2017

Jedno z mnoha možných zadání pro řešení dané analytické otázky.

CF-kvantifikátor MAXDiff ≤ 5 %C

QUANTIFIERS

Type	Source	Range	Rel.	Value	Units
SUM	Abs	all	>=	50	Abs
PDSUM	%Cond	all	<=	10	Abs
PDMAX	%Cond	all	<=	5	Abs

Rozdíl výška histogramů v % pro jednotlivé kategorie je maximálně 5.

CF Pattern Difference quantifier settings

Interest measure

PattDiffMax

The maximal of absolute values of differences of frequencies in the histogram and a given pattern

Source frequencies

Relative [%] to act condition

Relative frequencies to number of rows matching condition. Frequencies are divided by number of rows matching condtion and multiplied by 100.

Category Range

From 0 To 100

Reset to All

Absolute category index

Relative range [%] to act number of categories

Relation

Less than or equal

Threshold value

5

Threshold-value units

Absolute number

Absolute number. Threshold value is left as given.

Pattern used for comparison

Histogram of the same attribute on the whole matrix

Selected stored pattern (of the same size)

PŘÍKLAD

Normalize pattern frequencies to SUM= 100

Primary IM Settings

Set as primary IM

Normalize value range: From 0 To 1

Note

-

OK Cancel

? CF: L1 $_{\leq 10\,\%\,C}$ DHodnocení / Host, Bydliště, Meteo, Pobyt, výstup procedury CF-Miner

Task run

Start: 3.4.2017 19:32:10 Total time: 0h 0m 1s

Number of verifications: 21197

Number of hypotheses: 96 Mode: Standard

Počet vystupujících histogramů je 96, tedy stejný jako v příkladu PH1. Jsou však v pořadí podle PDMAX

Actual group of hypotheses: All hypotheses

Hypotheses in group: 96 Shown hypotheses: 96 Highlighted: 0

Nr.	Id	PDMAX	Hypothesis
1	34	4.965	**DHodnoceni** / **HCizinec**(*ne*) & **HVek**(*60 a více*) & **PNoci**(*7*) & **PVikend**(*ano*)
2	35	4.965	**DHodnoceni** / **HStat**(*ČR*) & **HVek**(*60 a více*) & **PNoci**(*7*) & **PVikend**(*ano*)
3	58	4.900	**DHodnoceni** / **HCizinec**(*ne*) & **PNoci**(*7/14/21*) & **POsobonoci**(*vyšší*) & **PTurnus**(*ano*)
4	59	4.900	**DHodnoceni** / **HCizinec**(*ne*) & **PNoci**(*7/14/21*) & **POsobonoci**(*vyšší*) & **PDenTydne**(*So*)
5	53	4.900	**DHodnoceni** / **HCizinec**(*ne*) & **POsobonoci**(*vyšší*) & **PTurnus**(*ano*)
6	15	4.900	**DHodnoceni** / **HMesto**(*Praha*) & **HPohlavi**(*muž*) & **PTypPobytu**(*rekreační*)
7	89	4.900	**DHodnoceni** / **HStat**(*ČR*) & **PNoci**(*7/14/21*) & **POsobonoci**(*vyšší*) & **PTurnus**(*ano*)
8	90	4.900	**DHodnoceni** / **HStat**(*ČR*) & **PNoci**(*7/14/21*) & **POsobonoci**(*vyšší*) & **PDenTydne**(*So*)
9	84	4.900	**DHodnoceni** / **HStat**(*ČR*) & **POsobonoci**(*vyšší*) & **PTurnus**(*ano*)
10	20	4.848	**DHodnoceni** / **HCizinec**(*ne*) & **HPohlavi**(*žena*) & **PNoci**(*7/14/21*) & **POsob**(*2*) & **PTypPobytu**(*rekreační*)
11	23	4.848	**DHodnoceni** / **HStat**(*ČR*) & **HPohlavi**(*žena*) & **PNoci**(*7/14/21*) & **POsob**(*2*) & **PTypPobytu**(*rekreační*)
12	30	4.807	**DHodnoceni** / **HCizinec**(*ne*) & **HVek**(*od 28 do 60*) & **MTeplota**(*zima*) & **PRok**(*2012*)
13	33	4.807	**DHodnoceni** / **HStat**(*ČR*) & **HVek**(*od 28 do 60*) & **MTeplota**(*zima*) & **PRok**(*2012*)
14	4	4.773	**DHodnoceni** / **MTeplota**(*zima*) & **PNoci**(*7/14/21*) & **PMesic**(*leden*)
15	66	4.742	**DHodnoceni** / **HMesto**(*Plzeň*) & **PTypPobytu**(*rekreační*)

? CF: L1 $_{\leq 10\ \%\ C}$ DHodnocení / Host, Bydliště, Meteo, Pobyt, zadání pro CF-Miner – Příklad PH2

LM Hotel_Slidy MB - LISp-Miner Workspace module - 27.08.01

File Data Introduction Preprocessing Interactive Analysis Data-mining Tasks Domain Knowledge Window Help

Attributes | Tables | Tasks Overview | 01: PDenTydne / Hodn.. | 01: PDenTydne / Hodn.. | 01: PDenTydne / Hodn.. | 01: PDenTydne / Hodn

Data-mining Task basic parameters

Name: 01: PDenTydne / Hodnocení - Podobnost s PŘÍKLAD Kviz

Comment: -

Taskgroup: 01: CF-Miner

Task type: CF-Miner Data matrix: HotelPlusExterni

ATTRIBUTES FOR HISTOGRAM

Attributes

» DHodnoceni(3 categories)

DHodnocení

QUANTIFIERS

Type	Source	Range	Rel.	Value	Units
SUM	Abs	all	>=	50	Abs
PDSUM	%Cond	all	<=	10	Abs
PDMAX	%Cond	all	<=	4	Abs

SUM(Abs) $\geq$ 50

L1 $_{\leq 10\ \%C}$

MaxDiff $\leq$ 4

CONDITION

Host/Bydliště Con, 0 - 3
» HCizinec_b (subset), 1 - 1 B, pos
» HMesto (subset), 1 - 1 B, pos
» HStat (subset), 1 - 1 B, pos
» HPohlavi (subset), 1 - 1 B, pos
» HVek_exp (subset), 1 - 1 B, pos
Meteo Con, 0 - 2
» MObloha (subset), 1 - 1 B, pos
» MTeplota_exp (subset), 1 - 1 B, pos
Počet nocí Con, 0 - 4
» PNoci_enum_m (subset), 1 - 1 B, pos
» PNoci_exp (subset), 1 - 1 B, pos
» POsob (subset), 1 - 1 B, pos
» POsobonoci_ef5 (subset), 1 - 1 B, pos
Pobyt Con, 0 - 1
» PPresSobotniNoc (subset), 1 - 1 B, pos
» PTurnus (subset), 1 - 1 B, pos
» PTypPobytu (subset), 1 - 1 B, pos
» PDenTydne (subset), 1 - 1 B, pos
» PMesic (subset), 1 - 1 B, pos
» PRok (subset), 1 - 1 B, pos
» PSezona_b (subset), 1 - 1 B, pos
Total length: 1 - 1

Aggregate function

Type: Count(*)

Attribute: -

Task parameters

Include 'worse' extensions of condition: Yes

Include extensions of coefficients with no change in the histogram: Yes

Include extensions of cedents with no change in the histogram: Yes

Maximal number of hypotheses: 1000

Extensions minimal length check: Yes

Params | Validate | Task Clone

Run | Bkgrnd Run | Grid Run | Show Results

Ready NUM 19:38 3.4.2017

Příklad PH2 – změna z MaxDiff $\leq$ 5 na MaxDiff $\leq$ 4.

Host/Bydliště(*) $\wedge$ Meteo(*) Pobyt (*) $\wedge$ Pobyt(*)

Jedno z mnoha možných zadání pro řešení dané analytické otázky.

? CF: L1 $_{\leq 10\ \%\ C}$ DHodnocení / Host, Bydliště, Meteo, Pobyt, výstup procedury CF-Miner

Task run

Start: 3.4.2017 19:50:47 Total time: 0h 0m 1s

Number of verifications: 21197

Number of hypotheses: 73 Mode: Standard

Počet vystupujících histogramů je 73, tedy menší než v příkladu PH2. Jsou v pořadí podle PDMAX

Actual group of hypotheses: All hypotheses

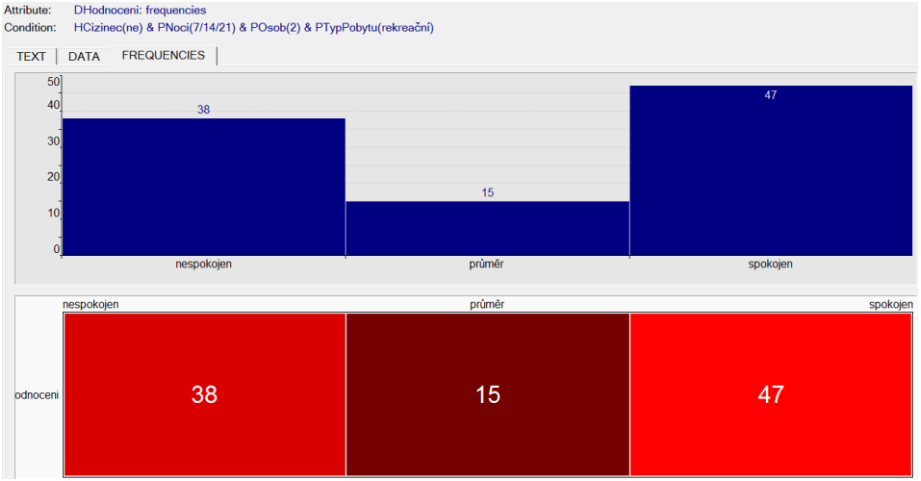
Hypotheses in group: 73 Shown hypotheses: 73 Highlighted: 0

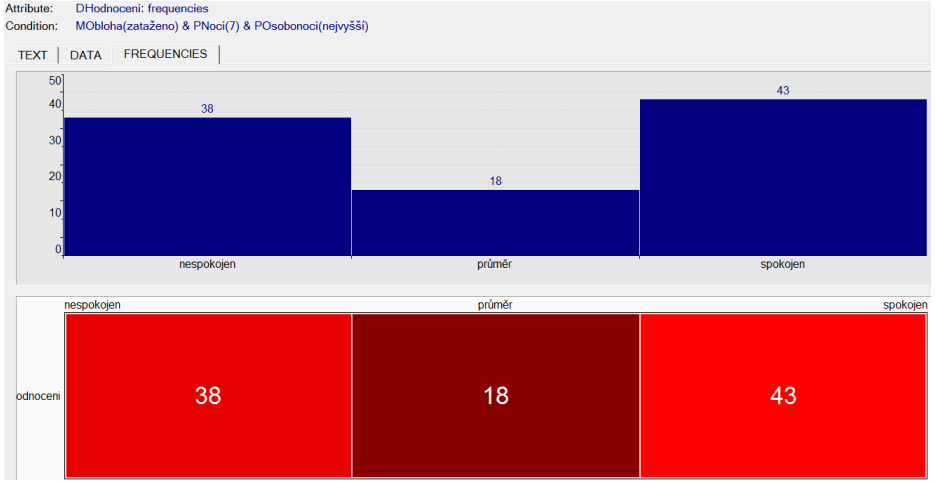
Nr.	Id	PDMAX	Hypothesis
1	4	3.939	**DHodnoceni / MObloha**(*zataženo*) & **PNoci**(*7*) & **POsobonoci**(*nejvyšší*)
2	5	3.939	**DHodnoceni / MObloha**(*zataženo*) & **PNoci**(*7*) & **POsobonoci**(*nejvyšší*) & **PPresSobotniNoc**(*ano*)
3	6	3.939	**DHodnoceni / MObloha**(*zataženo*) & **PNoci**(*7*) & **POsobonoci**(*nejvyšší*) & **PTypPobytu**(*rekreační*)
4	19	3.923	**DHodnoceni / HVek**(*od 28 do 60*) & **MObloha**(*zataženo*) & **PNoci**(*7*) & **PTypPobytu**(*rekreační*)
5	47	3.877	**DHodnoceni / HCizinec**(*ne*) & **MTeplota**(*teplo*) & **POsobonoci**(*nejvyšší*) & **PTurnus**(*ne*)
6	73	3.877	**DHodnoceni / HStat**(*ČR*) & **MTeplota**(*teplo*) & **POsobonoci**(*nejvyšší*) & **PTurnus**(*ne*)
7	7	3.822	**DHodnoceni / MObloha**(*zataženo*) & **PNoci**(*7*) & **PTurnus**(*ano*)
8	8	3.822	**DHodnoceni / MObloha**(*zataženo*) & **PNoci**(*7*) & **PDenTydne**(*So*)
9	29	3.788	**DHodnoceni / HCizinec**(*ne*) & **PNoci**(*7*) & **POsob**(*2*) & **POsobonoci**(*vyšší*) & **PVikend**(*ano*)
10	37	3.788	**DHodnoceni / HCizinec**(*ne*) & **PNoci**(*7*) & **POsob**(*2*) & **PVikend**(*ano*)
11	42	3.788	**DHodnoceni / HCizinec**(*ne*) & **PNoci**(*7*) & **POsobonoci**(*vyšší*) & **PVikend**(*ano*)
12	33	3.788	**DHodnoceni / HCizinec**(*ne*) & **PNoci**(*7/14/21*) & **POsob**(*2*) & **POsobonoci**(*vyšší*) & **PVikend**(*ano*)
13	55	3.788	**DHodnoceni / HStat**(*ČR*) & **PNoci**(*7*) & **POsob**(*2*) & **POsobonoci**(*vyšší*) & **PVikend**(*ano*)
14	63	3.788	**DHodnoceni / HStat**(*ČR*) & **PNoci**(*7*) & **POsob**(*2*) & **PVikend**(*ano*)
15	68	3.788	**DHodnoceni / HStat**(*ČR*) & **PNoci**(*7*) & **POsobonoci**(*vyšší*) & **PVikend**(*ano*)

? CF: L1 $_{\leq 10\ \%\ C}$ DHodnocení / Host, Bydliště, Meteo, Pobyt, podrobnosti k prvnímu histogramu, příklad 10

PŘÍKLAD
HCizinec(ne) ∧ PNoci(7/14/21) ∧ POsob(2) ∧ PTypPobytu(rekreační)

První histogram χ
MObloha(zataženo) ∧ PNoci(7) ∧ POsobonoci(nejvyšší)





L1(HotelPlusExterni, DHodnoceni, % Cond, PŘÍKLAD, χ) = |38-38| + |15-18| + |47-43| = 0 + 3 + 4 = 7

MaxDiff(HotelPlusExterni, DHodnoceni, % Cond, PŘÍKLAD, χ) = Max{|38-38|, |15 -18|, |47- 43|} = 4

Poznámka: Pokud použijeme zaokrouhlené hodnoty na celá procenta, dochází k rozdílům od hodnot PDMAX uvedeným v přehledovém výstupu procedury CF-Miner.

Vzdálenost histogramů - poznámky

- Kvantifikátory využívající vzdálenost histogramů nejsou zatím popsány v literatuře.
- Hledání podmínek, za kterých je vzdálenost histogramu daného atributu od histogramu téhož atributu na celé matici dat větší než daný práh lze chápat i jako hledání výjimek od histogramu daného atributu na celé matici dat. Čím větší vzdálenost, tím větší výjimka.
- Takové hledání výjimek lze aplikovat i v případě, kdy pro daný atribut
  - je dán histogram na podmatici M/ω.
  - hledáme podmínky κ, za kterých je vzdálenost histogramu na podmatici $M/(\omega \wedge \kappa)$ od histogramu na podmatici M/ω větší než daný práh.
- Je řada dalších, dosud neprozkoumaných způsobů, jak využít kvantifikátory založené na vzdálenosti histogramů.