Postprocessing of Hypotheses

Martin Kejkula

Supervisor: Jan Rauch

Knowledge Engineering Group, 2007

Outline

- 4ft-hypotheses
  - current handling
- Clustering association rules
  - review of current papers
- Postprocessing 4ft-hypotheses – new approach
- Experiment results
- Discussion

4ft-Hypotheses

- Output of 4ft-Miner (module of LISp-Miner)
- General association rules:

  Ant ~ Succ / Cond:

  - Ant, Succ, Cond
  - literals, several types of coefficients
  - several types of quantifiers
  - ...

4ft-Hypotheses - Description

- Literal/attribute (in all cedents)
- Quantitative characteristics
  - Confidence, support, average difference,...

4ft-Hypotheses – Current Handling

- Sorting
- Filter (and group)
- Importance of literal
- Hypotheses to text report

Clustering Association Rules

- H. Toivonen, M. Klemettinen, P. Ronkainen, K. Hatonen, H. Mannila: Pruning and Grouping Association Rules

- G. Dong, J. Li: Interestingness of Discovered Association Rules in terms of Neighborhood-Based Unexpectedness

- B. Lent, A. Swami, J. Widom: Clustering Association Rules

Toivonen, Mannila: Pruning...

- A1 ~> $_{0.9;\ 0.2}$ S1
- A1 and A2 ~> $_{0.9;\ 0.1}$ S1 (pruned as redundant)

- Clustering of rules:
  - distance based on covered data rows

Prime Hypotheses

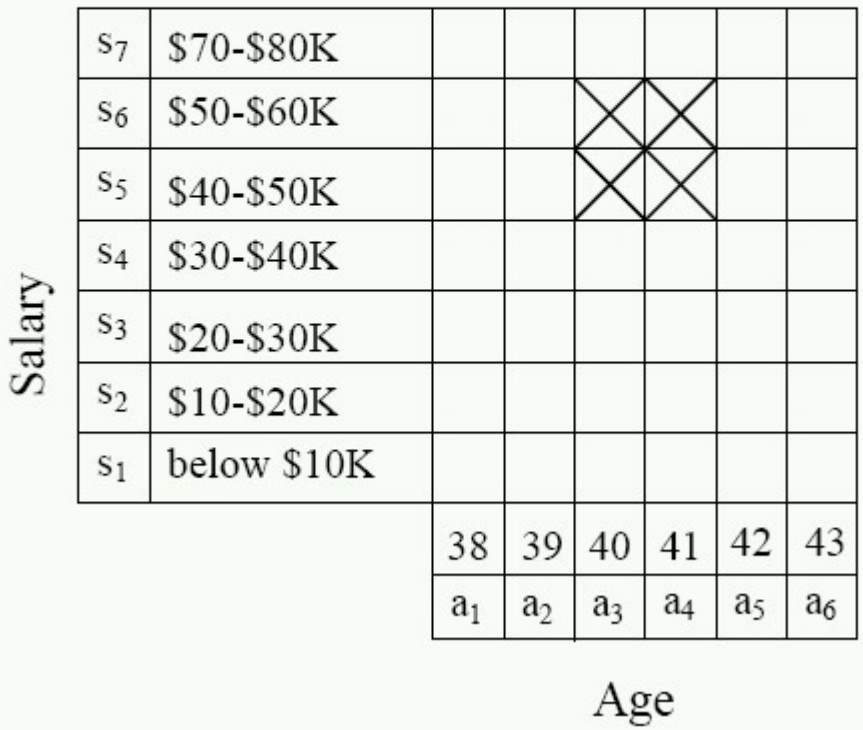
- Output of 4ft-Miner
  - examples:
  - 1. $A1(m) \sim>_{0,9;20} S1(n)$ and $S2(o,p,\mathbf{q,r})$
  - 2. $A1(m) \sim>_{0,9;20} S1(n)$ and $S2(o,p)$

Dong, Li: Neighborhood...

- Syntax-based distance
- Neighborhoods
- Interesting rules with unexpected confidence

Lent: Clustering Association...

$$(\text{Age} = a_3) \wedge (\text{Salary} = s_5) \Longrightarrow (\text{Group\_label} = A)$$
$$(\text{Age} = a_4) \wedge (\text{Salary} = s_6) \Longrightarrow (\text{Group\_label} = A)$$
$$(\text{Age} = a_4) \wedge (\text{Salary} = s_5) \Longrightarrow (\text{Group\_label} = A)$$
$$(\text{Age} = a_3) \wedge (\text{Salary} = s_6) \Longrightarrow (\text{Group\_label} = A)$$



New approach

to postprocessing of hypotheses

Postprocessing 4ft-Hypotheses

- No information about the set of discovered rules:
  - correlations, above average, confidence (implications),...
  - literals in cedents
  - which/how many rows of data are affected
  - which/how many rows of data are not affected

Postprocessing 4ft-Hypotheses

- Methodology
  - literal/attribute „document“ clustering
  - quantitative characteristics clustering (i.e. two independent clustering)
  - raw (analysed) data „covered“ by clusters
  - results evaluation (new 4ft mining task - metalearning,...)

Literal/Attribute Document Clustering

- Output of 4ft-Miner translation into set of documents
- Attributes (or attributes&categories) are the key-words
- Document clustering
- Output: clusters of similar hypotheses

Hypotheses as Documents (1/2)

- 1 DIAST1 SYST1 AKTPOZAM ALKOHOL CUKR PIVOMN
- 2 DIAST1 SYST1 AKTPOZAM ALKOHOL CUKR IM PIVOMN
- 3 DIAST1 SYST1 AKTPOZAM ALKOHOL CUKR DIABET PIVOMN
- 4 DIAST1 SYST1 AKTPOZAM ALKOHOL CUKR PIVOMN
- 5 …

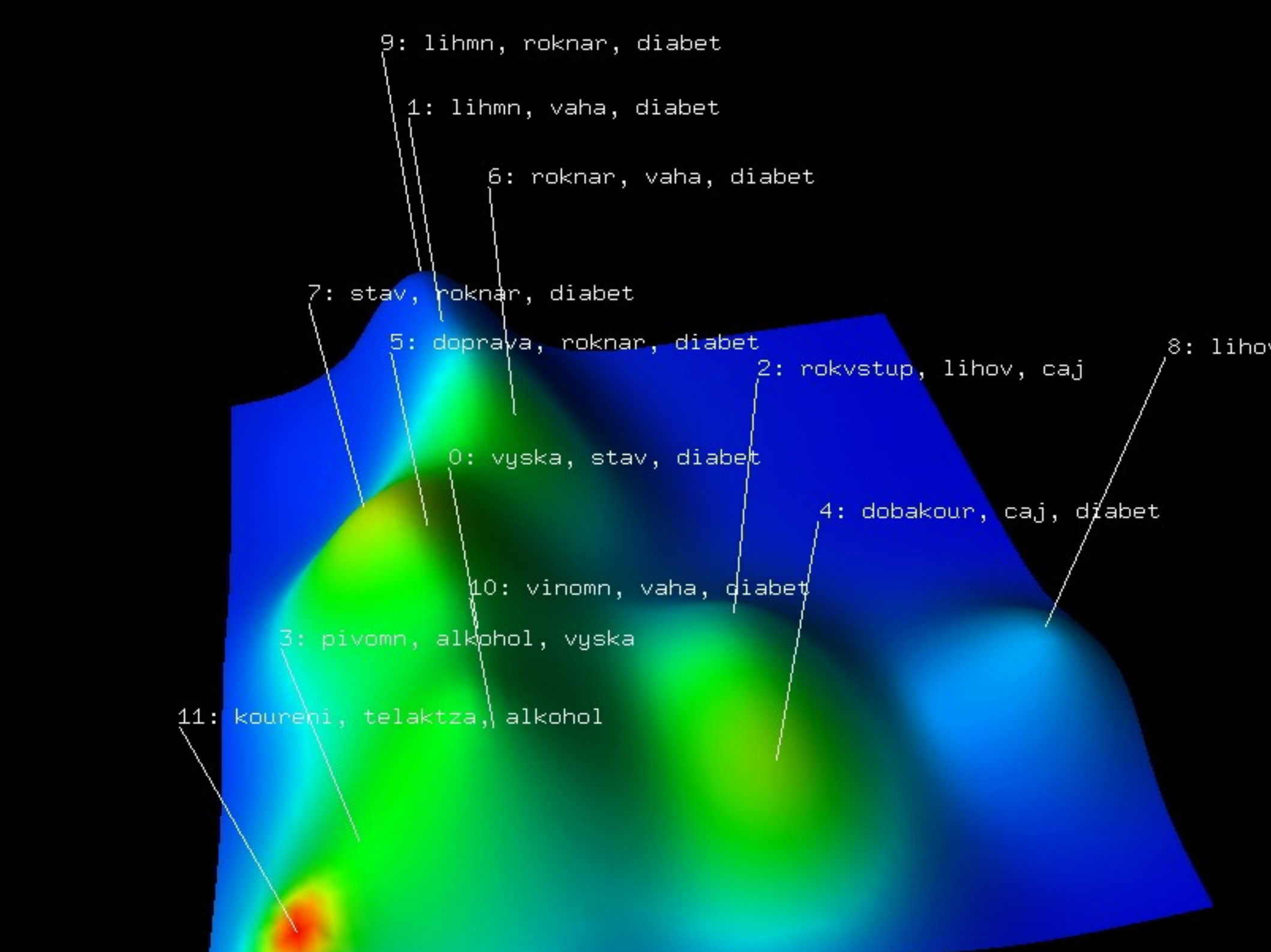
Hypotheses as Documents (2/2)

- 1 AKTPOZAMx1 AKTPOZAM ALKOHOLx2 ALKOHOL CUKRx0 CUKR CUKRx1 CUKRx2 CUKRx3 CUKRx4 PIVOMNx2 PIVOMN DIAST1x95-100 DIAST1 DIAST1x100-105 DIAST1x105-110 SYST1x130-135 SYST1 SYST1x135-140 SYST1x140-145 SYST1x145-150
- 2 AKTPOZAMx1 AKTPOZAM ALKOHOLx2 ALKOHOL CUKRx0 CUKR CUKRx3 CUKRx2 CUKRx1 CUKRx4 IMx2 IM PIVOMNx2 PIVOMN DIAST1x95-100 DIAST1 DIAST1x100-105 DIAST1x105-110 SYST1x130-135 SYST1 SYST1x135-140 SYST1x140-145 SYST1x145-150
- 3 …

Literal/Attribute Document Clustering

- How many clusters we want to get?
- How many attributes we used in 4ft-Task?

  x attributes, then: max x/4 clusters?


9: lihmn, roknar, diabet
1: lihmn, vaha, diabet
6: roknar, vaha, diabet
7: stav, roknar, diabet
5: doprava, roknar, diabet
8: lihov
2: rokvstup, lihov, caj
0: vyska, stav, diabet
4: dobakour, caj, diabet
10: vinomn, vaha, diabet
3: pivomn, alkohol, vyska
11: koureni, telaktza, alkohol

Quantitative Characteristics Clustering (1/2)

- Each hypothesis is described by tuple: <Confidence, Support, ...>
- Do not use Chi-square, Fisher... for clustering
- Normalise before clustering
- Output: clusters of hypotheses with similar quantitative characteristics

Quantitative Characteristics Clustering (2/2)

- How many clusters we want to get?
- What is the interpretation of quantitative characteristics we used for clustering?
  - implicetion, correlation, what else (WRAcc...)?
- max 12 clusters?

Data Covered By Clusters

- Two types of clusters of hypotheses
  - Literal/attribute-clusters
  - Quantitative-characteristics-clusters
- Which rows of data are/are not covered by which clusters?
- Attribute A1: empirical distribution of values: which values are/are not covered by which clusters?

Experiment Results

- Mining 4ft-hypotheses in STULONG Entry data set
- 756 hypotheses clustering
  - as 756 documents
  - as 756 tuples of quantitative characteristics
- Clusters evaluation
  - 48 cluster intersections

Postprocessing 4ft-Hypotheses (1/2)

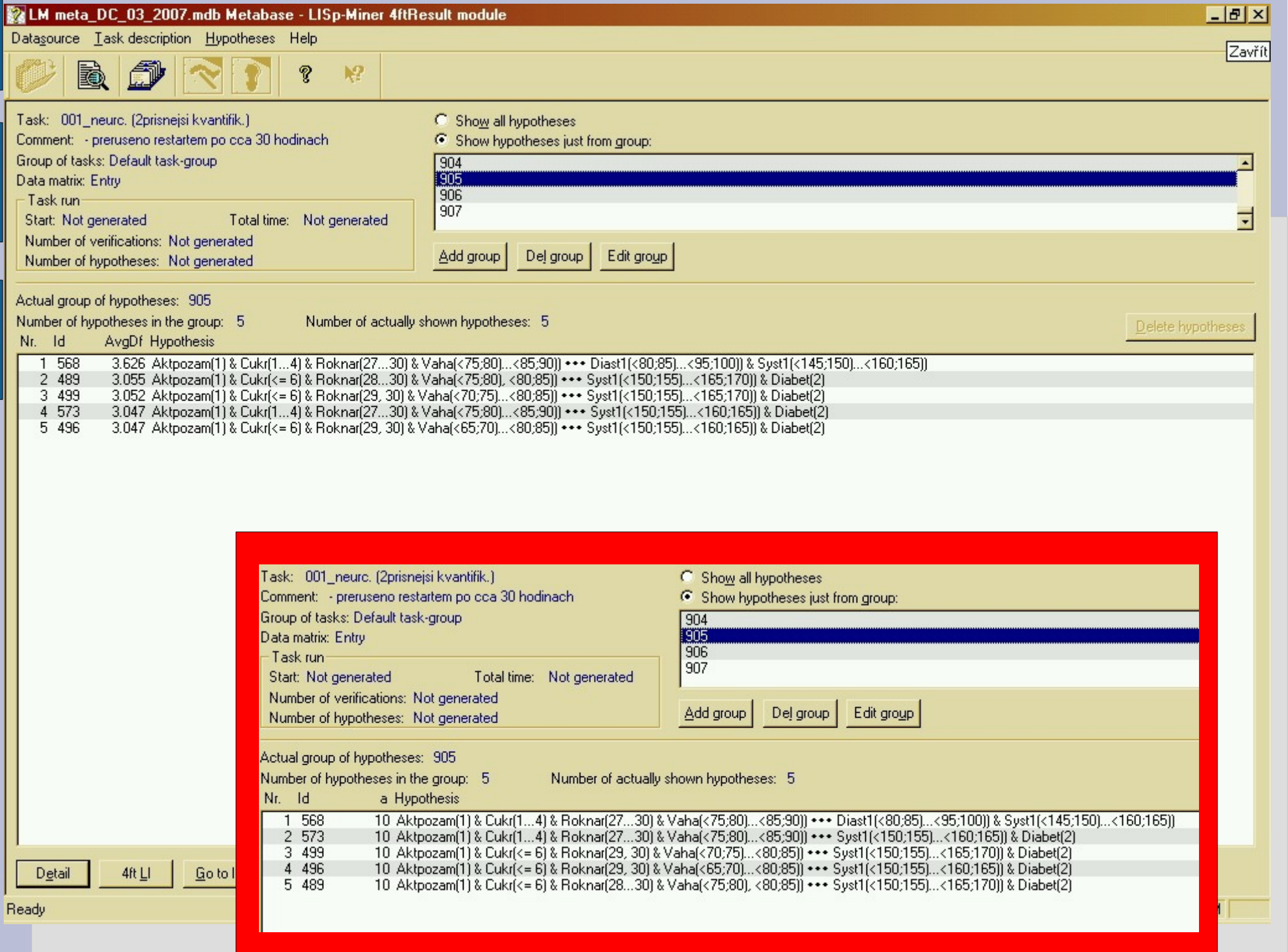
- Normalization of quantitative characteristics
- Clustering of 4ft-rules

  (quantitative characteristics)
- Transformation 4ft-rules into documents
- Clustering of documents

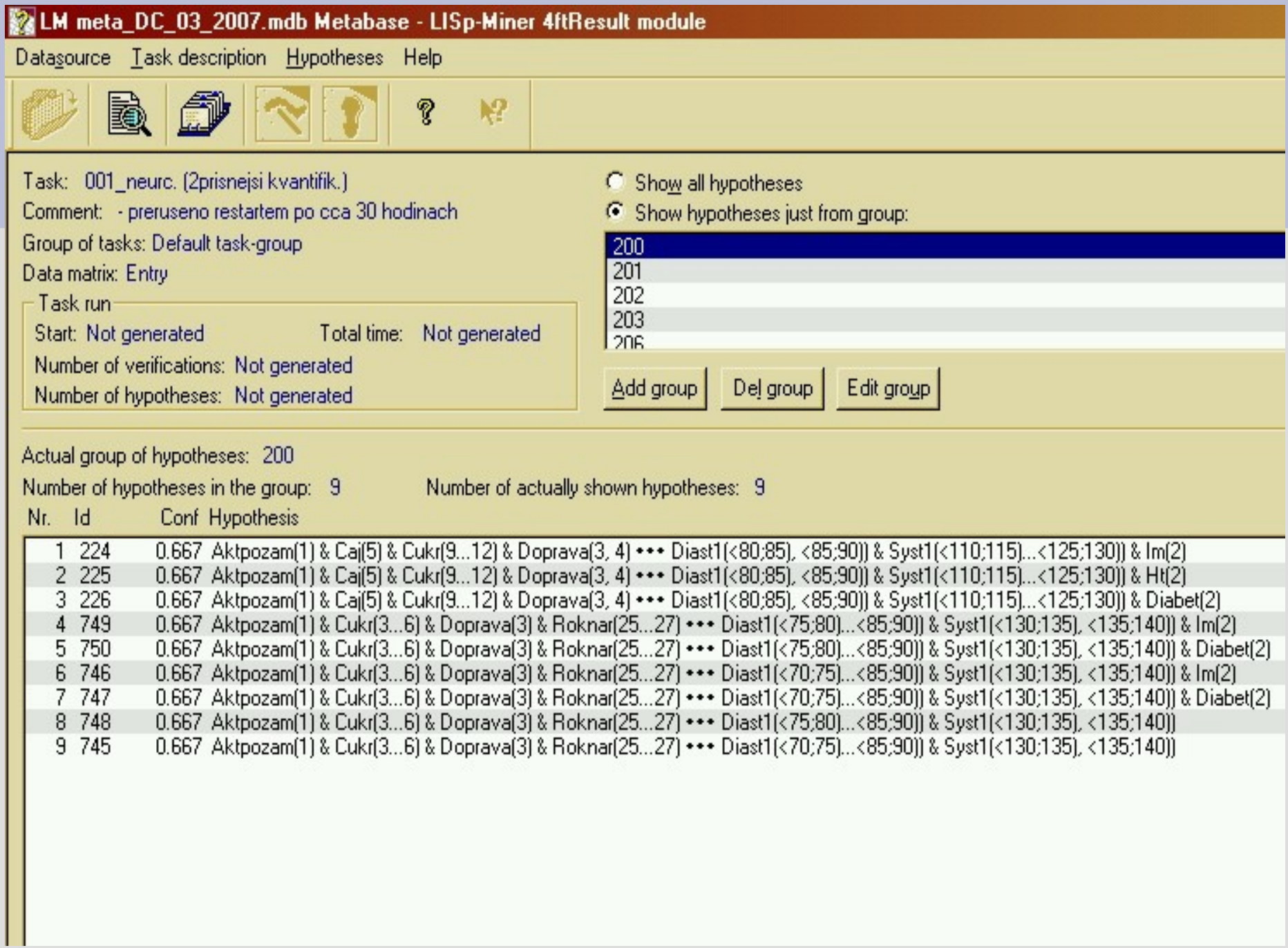
Postprocessing 4ft-Hypotheses (2/2)

- Clusters intersections
- Transformation 4ft-rules into SQL-query:
  - which rows are affected by 4ft-rule
- Integration results & delivery to user
  - Each hypotheses is in one cluster intersection
  - One row of data can be covered by one or more hypotheses from one or more clusters intersections

popis shluku dokumentu	ID shluku dokumentu	ID shluku kvantitativ	prumerna Confidence	prumerna Dconfidence	prumerna Econfidence	prumerny Support	prumerna Completeness	prumerna average difference	počet hypotez	počet ICO – zaznamu v tabulce ENTRY	
pivomn, alkohol, im	5	0	0,833	0,036	0,808	0,007	0,036	3,228	6	10	1,67
	5	1	0,204	0,095	0,932	0,007	0,156	3,361	35	12	0,34
	5	2	0,215	0,107	0,940	0,007	0,178	4,307	36	12	0,33
	5	6	0,364	0,074	0,909	0,007	0,086	3,254	10	29	2,9
	5	7	0,402	0,077	0,912	0,007	0,087	3,745	11	29	2,64
vyska, stav, im	6	0	0,673	0,042	0,840	0,007	0,043	3,111	8	10	1,25
	6	1	0,227	0,092	0,929	0,007	0,136	3,232	48	40	0,83
	6	2	0,241	0,101	0,937	0,007	0,149	4,079	6	13	2,17
	6	7	0,464	0,061	0,892	0,007	0,066	3,331	5	11	2,2
vinomn, vaha, rokvstup	7	3	0,588	0,051	0,869	0,007	0,053	3,427	15	21	1,4
	7	5	0,526	0,053	0,872	0,007	0,055	3,089	18	14	0,78
	7	6	0,378	0,068	0,902	0,007	0,077	3,076	6	10	1,67
	7	7	0,450	0,062	0,892	0,007	0,067	3,222	19	16	0,84
lihov, rokvstup, caj	8	0	0,672	0,049	0,840	0,008	0,050	3,092	11	12	1,09
	8	1	0,189	0,092	0,930	0,007	0,152	3,048	2	0	0
	8	3	0,556	0,054	0,877	0,007	0,057	3,474	4	14	3,5
	8	5	0,526	0,054	0,876	0,007	0,057	3,238	4	15	3,75
	8	6	0,359	0,071	0,907	0,007	0,082	3,133	14	50	3,57
	8	7	0,424	0,068	0,897	0,007	0,075	3,249	37	40	1,08
roknar, vaha, diabet	9	0	0,748	0,043	0,833	0,007	0,044	3,345	8	13	1,63
	9	1	0,225	0,096	0,928	0,008	0,145	3,260	14	12	0,86
	9	2	0,233	0,106	0,935	0,008	0,165	3,905	7	12	1,71
	9	3	0,591	0,051	0,863	0,007	0,053	3,249	10	18	1,8
	9	4	0,526	0,060	0,890	0,007	0,064	3,744	1	10	10
	9	5	0,538	0,052	0,871	0,007	0,055	3,166	5	19	3,8
	9	6	0,328	0,080	0,913	0,008	0,096	3,145	5	11	2,2
	9	7	0,451	0,062	0,893	0,007	0,067	3,286	10	40	4


LM meta_DC_03_2007.mdb Metabase - LISp-Miner 4ftResult module
Datasource Task description Hypotheses Help
Zavřít
Task: 001_neurc. (2prisnejsi kvantifik.)
Comment: - preruseno restartem po cca 30 hodinach
Group of tasks: Default task-group
Data matrix: Entry
Task run
Start: Not generated Total time: Not generated
Number of verifications: Not generated
Number of hypotheses: Not generated
Show all hypotheses
Show hypotheses just from group:
904
905
906
907
Add group
Del group
Edit group
Actual group of hypotheses: 905
Number of hypotheses in the group: 5 Number of actually shown hypotheses: 5
Delete hypotheses
Nr. Id AvgDf Hypothesis
1 568 3.626 Aktpozam(1) & Cukr(1...4) & Roknar(27...30) & Vaha(<75;80)...<85;90)) ••• Diast1(<80;85)...<95;100)) & Syst1(<145;150)...<160;165))
2 489 3.055 Aktpozam(1) & Cukr(<= 6) & Roknar(28...30) & Vaha(<75;80), <80;85)) ••• Syst1(<150;155)...<165;170)) & Diabet(2)
3 499 3.052 Aktpozam(1) & Cukr(<= 6) & Roknar(29, 30) & Vaha(<70;75)...<80;85)) ••• Syst1(<150;155)...<165;170)) & Diabet(2)
4 573 3.047 Aktpozam(1) & Cukr(1...4) & Roknar(27...30) & Vaha(<75;80)...<85;90)) ••• Syst1(<150;155)...<160;165)) & Diabet(2)
5 496 3.047 Aktpozam(1) & Cukr(<= 6) & Roknar(29, 30) & Vaha(<65;70)...<80;85)) ••• Syst1(<150;155)...<160;165)) & Diabet(2)
Detail
4ft LI
Go to
Ready
Task: 001_neurc. (2prisnejsi kvantifik.)
Comment: - preruseno restartem po cca 30 hodinach
Group of tasks: Default task-group
Data matrix: Entry
Task run
Start: Not generated Total time: Not generated
Number of verifications: Not generated
Number of hypotheses: Not generated
Show all hypotheses
Show hypotheses just from group:
904
905
906
907
Add group
Del group
Edit group
Actual group of hypotheses: 905
Number of hypotheses in the group: 5 Number of actually shown hypotheses: 5
Nr. Id a Hypothesis
1 568 10 Aktpozam(1) & Cukr(1...4) & Roknar(27...30) & Vaha(<75;80)...<85;90)) ••• Diast1(<80;85)...<95;100)) & Syst1(<145;150)...<160;165))
2 573 10 Aktpozam(1) & Cukr(1...4) & Roknar(27...30) & Vaha(<75;80)...<85;90)) ••• Syst1(<150;155)...<160;165)) & Diabet(2)
3 499 10 Aktpozam(1) & Cukr(<= 6) & Roknar(29, 30) & Vaha(<70;75)...<80;85)) ••• Syst1(<150;155)...<165;170)) & Diabet(2)
4 496 10 Aktpozam(1) & Cukr(<= 6) & Roknar(29, 30) & Vaha(<65;70)...<80;85)) ••• Syst1(<150;155)...<160;165)) & Diabet(2)
5 489 10 Aktpozam(1) & Cukr(<= 6) & Roknar(28...30) & Vaha(<75;80), <80;85)) ••• Syst1(<150;155)...<165;170)) & Diabet(2)

popis shluku dokumentu	ID shluku dokumentu	ID shluku kvantitativ	prumerna Confidence	prumerna Dconfidence	prumerna Econfidence	prumerny Support	prumerna Completeness	prumerna average difference	počet hypotez	počet ICO – zaznamu v tabulce ENTRY
	0	3	0,556	0,053	0,874	0,007	0,056	3,348	4	13
	0	4	0,500	0,072	0,899	0,008	0,078	3,958	4	11
lihmn, roknar, diabet	0	5	0,534	0,056	0,872	0,008	0,059	3,129	12	14
	0	6	0,340	0,076	0,913	0,007	0,089	3,257	19	13
	0	7	0,440	0,066	0,895	0,007	0,072	3,274	30	14
	1	0	0,709	0,044	0,842	0,007	0,045	3,355	11	12
	1	3	0,593	0,051	0,867	0,007	0,053	3,402	16	29
lihmn, vaha, diabet	1	4	0,516	0,062	0,892	0,007	0,065	3,755	10	20
	1	5	0,527	0,058	0,878	0,008	0,061	3,301	32	35
	1	6	0,345	0,073	0,909	0,007	0,084	3,086	3	10
	1	7	0,474	0,062	0,889	0,007	0,066	3,299	13	21
	2	0	0,667	0,044	0,848	0,007	0,045	3,292	9	20
	2	1	0,234	0,090	0,928	0,007	0,129	3,238	28	39
doprava, roknar, diabet	2	2	0,250	0,103	0,938	0,007	0,149	4,272	5	10
	2	3	0,588	0,047	0,858	0,007	0,049	3,069	7	12
	2	6	0,344	0,075	0,911	0,007	0,087	3,191	15	33
	2	7	0,423	0,076	0,898	0,008	0,085	3,280	19	51
dobakour, caj, ht	3	0	0,667	0,042	0,840	0,007	0,043	3,084	3	10
	3	7	0,450	0,063	0,894	0,007	0,069	3,302	22	21
stav, roknar, diabet	4	1	0,235	0,095	0,928	0,008	0,139	3,308	38	20
	4	2	0,281	0,109	0,936	0,008	0,152	4,464	28	12
	5	0	0,833	0,036	0,808	0,007	0,036	3,228	6	10
	5	1	0,204	0,095	0,932	0,007	0,156	3,361	35	12
pivomn, alkohol, im	5	2	0,215	0,107	0,940	0,007	0,178	4,307	36	12
	5	6	0,364	0,074	0,909	0,007	0,086	3,254	10	29
	5	7	0,402	0,077	0,912	0,007	0,087	3,745	11	29
	6	0	0,673	0,042	0,840	0,007	0,043	3,111	8	10
	6	1	0,227	0,092	0,929	0,007	0,136	3,232	48	40
vyska, stav, im	6	2	0,241	0,101	0,937	0,007	0,149	4,079	6	13
	6	7	0,464	0,061	0,892	0,007	0,066	3,331	5	11


LM meta_DC_03_2007.mdb Metabase - LISp-Miner 4ftResult module
Datasource Task description Hypotheses Help
Task: 001_neurc. (2prisnejsi kvantifik.)
Comment: - preruseno restartem po cca 30 hodinach
Group of tasks: Default task-group
Data matrix: Entry
Task run
Start: Not generated Total time: Not generated
Number of verifications: Not generated
Number of hypotheses: Not generated
Show all hypotheses
Show hypotheses just from group:
200
201
202
203
Add group
Del group
Edit group
Actual group of hypotheses: 200
Number of hypotheses in the group: 9 Number of actually shown hypotheses: 9
Nr. Id Conf Hypothesis
1 224 0.667 Aktpozam(1) & Caj(5) & Cukr(9...12) & Doprava(3, 4) ••• Diast1(<80;85), <85;90)) & Syst1(<110;115)...<125;130)) & Im(2)
2 225 0.667 Aktpozam(1) & Caj(5) & Cukr(9...12) & Doprava(3, 4) ••• Diast1(<80;85), <85;90)) & Syst1(<110;115)...<125;130)) & Ht(2)
3 226 0.667 Aktpozam(1) & Caj(5) & Cukr(9...12) & Doprava(3, 4) ••• Diast1(<80;85), <85;90)) & Syst1(<110;115)...<125;130)) & Diabet(2)
4 749 0.667 Aktpozam(1) & Cukr(3...6) & Doprava(3) & Roknar(25...27) ••• Diast1(<75;80)...<85;90)) & Syst1(<130;135), <135;140)) & Im(2)
5 750 0.667 Aktpozam(1) & Cukr(3...6) & Doprava(3) & Roknar(25...27) ••• Diast1(<75;80)...<85;90)) & Syst1(<130;135), <135;140)) & Diabet(2)
6 746 0.667 Aktpozam(1) & Cukr(3...6) & Doprava(3) & Roknar(25...27) ••• Diast1(<70;75)...<85;90)) & Syst1(<130;135), <135;140)) & Im(2)
7 747 0.667 Aktpozam(1) & Cukr(3...6) & Doprava(3) & Roknar(25...27) ••• Diast1(<70;75)...<85;90)) & Syst1(<130;135), <135;140)) & Diabet(2)
8 748 0.667 Aktpozam(1) & Cukr(3...6) & Doprava(3) & Roknar(25...27) ••• Diast1(<75;80)...<85;90)) & Syst1(<130;135), <135;140))
9 745 0.667 Aktpozam(1) & Cukr(3...6) & Doprava(3) & Roknar(25...27) ••• Diast1(<70;75)...<85;90)) & Syst1(<130;135), <135;140))

Future Work

- Which rows of data are/are not covered by which clusters?
  - Already done, but not presented today
  - How to present to user (user-friendly)?
- Empirical distribution of attribute values: which values are/are not covered by which clusters?

Discussion


Knowledge Engineering Group