Data mining with GUHA – Part 6
LISpMiner’s 4ftTask.exe module I

Esko Turunen

1 Tampere University of Technology
Examples of quantifiers in 4ftTask.exe module I

In this chapter we take a closed look at possibilities to do data mining with the 4ftTask.exe module. There are 15 quantifiers implemented. When suitable, we also explain same of their theoretical properties.

We start by showing another example of the use of the basic implication quantifier (also called founded implication).

In the previous chapter we found attributes that are related only to the attribute Contraceptive(No use). To find attributes related to the attribute Contraceptive(short term) – assuming that there are some! – can be done in many ways, for example as follows.
First open DataSource module -> Database -> Attributes List
We have now a new attribute. Next close this page and go to 4ftTask.exe module.
You may have to modify these values several times before you get any results!
Hypotheses generation survey

Task name: What implies short-term consumption?

Date of output: 20-05-2010

Survey of antecedent attributes:
- Group
  - Antecedent: Age
    - Group: Adult
      - Range: 1-7
  - Antecedent: Children
    - Group: Adult
      - Range: 1-5
  - Antecedent: Education
    - Group: Adult
      - Range: 1-2
  - Antecedent: Husband
    - Group: Adult
      - Range: 1-2
  - Antecedent: Husband or wife
    - Group: Adult
      - Range: 1-2
  - Antecedent: Working (yes)
    - Group: Yes
      - Range: 1-1
  - Antecedent: Married
    - Group: Yes
      - Range: 1-1
  - Antecedent: Religion
    - Group: Yes
      - Range: 1-1

Survey of consequent attributes:
- Group
  - Consequent: Consumption (short-term)
    - Group: 1-1

Survey of 4-tuples:
- 4-tuples and their parameters
  - EACH
    - Parameter: p = 0.000
  - Founded Implication
    - Parameter: p = 0.000

Task parameters:

Close | Copy to clipboard
The image contains a screenshot of a software interface, likely a data analysis or mining tool. The interface is in the language of the user's operating system, possibly Indonesian, based on the text in the software's interface. The screenshot shows a table with rows and columns, indicating some form of data analysis or hypothesis testing. The text within the table appears to be numerical and possibly related to probabilities or statistical measures. There is a URL at the bottom of the image: http://www.vrtuosi.com. The name Esko Turunen is also visible, which might be associated with the software or the user. The screenshot seems to be part of a larger software application used for data mining or hypothesis testing, possibly related to medical or scientific research given the context clues.
There is a subgroup of 58 women (aged 24...30 & ....& high livingstd) and 48 of them use short term contraception method.
Data mining with GUHA – Part 7
4ftTask.exe module II

Esko Turunen

1 Tampere University of Technology
Examples of quantifiers in 4ftTask.exe module II

♣ By above average quantifiers it is possible to do find all (sub)sets containing at least a given number of cases (denoted by base) such that some combination of attributes (predicates) would be at least $p$ times more frequent than in the whole data. In other words: among objects satisfying $\varphi$ there are at least $100 \cdot p\%$ more objects satisfying $\psi$ than there are objects satisfying $\psi$ in the whole data matrix. The exact truth definition of these quantifiers is the following

$$a \geq \text{base} \quad \text{and} \quad \frac{a}{r} \geq \frac{(1+p)k}{m}$$

For example, we would like to find all groups of at least 15-20 women in the Indonesia data among whom using some of the three contraception methods would be at least 2-3 times more frequent than in the whole data. This can be done by LispMiner in the following way:
These parts you fulfill in the usual way
It turns out that there are no results – try to reduce ‘base’ to, say, 15
<table>
<thead>
<tr>
<th>No</th>
<th>Id</th>
<th>AveGain Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3.147 Age&lt;37, 40 &amp; Childcount&gt;3, 5 &amp; Education&lt;16 &amp; Husband(H) &amp; Husband(1) &amp; I(s)working(No) &amp; MedianAge(Good) &amp; Contraceptive(Leon)</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>3.147 Age&lt;37, 40 &amp; Childcount&gt;3, 5 &amp; Education&lt;16 &amp; Husband(H) &amp; Husband(1) &amp; I(s)working(No) &amp; MedianAge(Good)</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>3.147 Age&lt;37, 40 &amp; Childcount&gt;3, 5 &amp; Education&lt;16 &amp; Husband(H) &amp; Husband(1) &amp; I(s)working(No) &amp; MedianAge(Good)</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3.147 Age&lt;37, 40 &amp; Childcount&gt;3, 5 &amp; Education&lt;16 &amp; Husband(H) &amp; Husband(1) &amp; I(s)working(No) &amp; MedianAge(Good)</td>
</tr>
</tbody>
</table>

Task: Are there subset above average?
Comment: 
Group of tasks: Default group of tasks
Data source: Indonesia
Task run:
Start: 26-4-2010 15:52:03
Total time: 0h 0m 49s
Number of validations: 945428
Number of hypotheses: 4

Actual group of hypotheses: All hypothesis
Number of hypotheses in the group: 4
Number of actually shown hypotheses: 4

Options:
- Show all hypotheses
- Show hypotheses just from group

Add group | Del group | Edit group | Delete Hypotheses
There is a subgroup of 15+1=16 women, aged 37-40, having 3-5 children, highly educated, their husband's occupation group is 1, they are not working outside home: 15 of these women use long-term contraception, which is more than 3 times more frequent than in the whole population.
Below average quantifiers

Below average quantifiers act much in the same way than above average quantifiers do: with them we can find all (sub)sets containing at least a given number of cases (denoted by base) such that some combination of attributes (predicates) would be at least $p$ times less frequent than in the whole data. That is to say: among objects satisfying $\varphi$ there are at least $100 \cdot p\%$ less objects satisfying $\psi$ than there are objects satisfying $\psi$ in the whole data matrix. The exact truth definition is

$$a \geq \text{base} \quad \text{and} \quad \frac{a}{r} \leq \frac{(1-p)k}{m}$$

For example, we would like to find all groups of women in the Indonesia data among whom using one of the contraception methods would be considerably less frequent than in the whole data. This can be done by LispMiner in the following way:
Start a new 4ftTask
First name it and then click ‘OK’
1. Select these predicates e.g. as shown here

2. Adjust below average quantifier’s BASE and \( p \) values

3. Generate
In 3 min 12 sec 4441878 verifications were done and 11 hypotheses were found. Let us see one of them in detail.

<table>
<thead>
<tr>
<th>Nr.</th>
<th>Id</th>
<th>Avg</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>-0.82</td>
<td>Age[36...39] &amp; Childcount[2...5] &amp; Education[4] &amp; Contraceptive(No use)</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>-0.80</td>
<td>Age[36...39] &amp; Childcount[2...5] &amp; Education[4] &amp; Contraceptive(No use)</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>-0.82</td>
<td>Age[36...39] &amp; Childcount[2...5] &amp; Education[4] &amp; Contraceptive(No use)</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>-0.82</td>
<td>Age[36...39] &amp; Childcount[2...5] &amp; Education[4] &amp; Contraceptive(No use)</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>-0.82</td>
<td>Age[36...39] &amp; Childcount[2...5] &amp; Education[4] &amp; Contraceptive(No use)</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>-0.06</td>
<td>Age[36...39] &amp; Childcount[2...5] &amp; Education[4] &amp; Contraceptive(No use)</td>
</tr>
<tr>
<td>7</td>
<td>11</td>
<td>-0.81</td>
<td>Age[36...39] &amp; Childcount[2...5] &amp; Education[4] &amp; Contraceptive(No use)</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>-0.81</td>
<td>Age[36...39] &amp; Childcount[2...5] &amp; Education[4] &amp; Contraceptive(No use)</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>-0.80</td>
<td>Age[36...39] &amp; Childcount[2...5] &amp; Education[4] &amp; Contraceptive(No use)</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>-0.80</td>
<td>Age[36...39] &amp; Childcount[2...5] &amp; Education[4] &amp; Contraceptive(No use)</td>
</tr>
<tr>
<td>11</td>
<td>8</td>
<td>-0.81</td>
<td>Age[36...39] &amp; Childcount[2...5] &amp; Education[4] &amp; Contraceptive(No use)</td>
</tr>
</tbody>
</table>
To see percentages, choose 'Rel row'.
Only 8% of 35-39 year old women with 3-4 children, whose husband is highly educated, do not use any contraceptive.

Among the rest of women, the proportion of those who do not use any contraceptive is 45%.
Founded equivalence quantifiers

Founded equivalence quantifiers are created to search for a kind of partial equivalence in data: with these quantifiers we can find all (sub)sets such that, if some combination of attributes $\phi$ is present then also another combination of attributes $\psi$ is present, and if $\phi$ is not present then nor $\psi$ is present. The exact truth definition is

$$a \geq \text{base} \quad \text{and} \quad \frac{a+d}{a+b+c+d} \geq p, \quad p \in (0, 1].$$

As an example we can carry out the following somewhat undetermined search:
We start by opening and naming a new task in 4fTask module.
1. Notice that Antecedent and Succedent parts can contain even same predicates, as here is the case – LispMiner will not write out 'trivial' results'.

2. Select Founded equivalence and adjust parameters BASE and $p$.

3. Generate
One of the 1508 hypotheses is that highly educated women’s group is almost (up to a degree 0.763) equivalent with the group of women who have high living standard and whose husband is also highly educated.
Double founded implication quantifiers act much in the same way as founded equivalence quantifiers do, the only difference is that the $d$-value is not taken into account. The exact truth definition is

$$a \geq \text{base} \quad \text{and} \quad \frac{a}{a+b+c} \geq p, \ p \in (0, 1].$$

Thus, these quantifiers partially mimic a logical form

$$\varphi \text{ implies } \psi \text{ and } \psi \text{ implies } \varphi.$$ 

As an example let us modify the previous example by changing the quantifier:
1. Change to Double founded implication and adjust $p = 0.7$

2. Generate
The results are quite different from those generated by Founded equivalence quantifier: there are only 16 hypotheses. One of them says: If religion is Islam then child count is 1…6, and vice versa. This is true to a degree 0.715.
Simple deviation quantifiers are introduced and studied, in one form or in another, in many data mining frameworks, not only as a part of GUHA. The exact truth definition of simple deviation quantifiers is

\[ ad > e^{\delta} bc, \text{ where the parameter } \delta \geq 0 \]

As an example let us see the following search by LISpMiner:

\[^a\text{There is a tiny typo in LISpMiner’s present version, the exponent there is } \sigma!\]
In this new 4ftTask, choose Simple Deviation quantifier with parameter value $\delta = 1.5$ and click ‘Generate’.
There were 26583 verifications, 14 of them are positive, i.e. GUHA hypotheses. One of them states that there is a strong association between not having any children and not using any contraceptive (a closer look at data shows that most of women in this group are very young).
Data mining with GUHA – Part 8
Statistic quantifiers in 4ftTask.exe module

Esko Turunen

1 Tampere University of Technology
Founded implication, founded equivalence and double founded implication are inspired by logic, though they are not logic implication or equivalence in a deep sense. Above and below average quantifiers have another and rather obvious origin. GUHA matrices/searches can be seen from a statistical point of view, too. We will now study statistically motivated quantifiers. For example, we may ask

Is the coincidence of $\varphi(x)$ and $\psi(x)$ just random or is there some statistically justified dependence between them?

For example, it is custom to use the $\chi^2$ test to compare observed and expected values; a genetic experiment might hypothesize that the next generation of plants will exhibit a certain set of colors. By comparing the observed results with the expected ones, we can decide whether our original hypothesis is valid.
Fisher quantifiers\(^a\) \(\approx_{\alpha}, 0 < \alpha \leq 0.5\). Fisher quantifier corresponds to the test of hypothesis

\[
\text{Probability}(\varphi(x) \mid \psi(x)) > \text{Probability}(\varphi(x) \mid \neg \psi(x))
\]

with significance \(\alpha\).

For example, our data may concern health and smoking. Let \(\nu(\varphi(x)) = \text{TRUE}\) mean \(x\) is smoker and \(\nu(\psi(x)) = \text{TRUE}\) mean \(x\) has cancer. If LISpMiner procedures \(\varphi(x) \approx_{0.01} \psi(x)\), we accept the hypothesis There is a dependence between smoking and cancer, and the probability that we are mistaken is maximally 0.01.

More precisely, a Fisher quantifiers \(\approx_{\alpha}, 0 < \alpha \leq 0.5\) are defined such that, for any model \(M\), \(\nu(\varphi(x) \approx_{\alpha} \psi(x)) = \text{TRUE}\) iff \(ad > bc\) and

\[
\sum_{i=0}^{\min\{b,c\}} \frac{r!s!k!l!}{m!(a+1)!(b-1)!(c-1)!(d+1)!} \leq \alpha.
\]

\(^a\)See http://faculty.vassar.edu/lowry/fisher.html
To find all such attribute pairs that are not independent, we can use Fisher quantifiers. Open a new 4ft Task, select Antecedent and Succeedent parts and Fisher quantifier’s parameters. Here we have $\alpha = 0.005$ and BASE = 20%. Then click ‘Generate’

Notice that we added a Condition ‘Low education’
This is a new attribute defined by ‘education = 1’
53 hypotheses were found. For example, among low educated women, there is a high probability of positive dependence between short-term birth control method and having 2...4 children.
Fisher tests are **exact** tests for testing dependencies, however, they are computationally quite time-consuming. Therefore they are often replaced by $\chi^2$-tests, which are computationally easier (but are only asymptotic, not exact tests). $\chi^2$ quantifiers $\approx_\alpha$, $0 \leq \alpha \leq 1$ are defined such that, for any model $M$, $v(\varphi(x) \approx_\alpha \psi(x)) = \text{TRUE}$ iff $ad > bc$ and

\[
\frac{m(ad - bc)^2}{rskl} \leq \alpha.
\]

$\chi^2$ tests\(^a\) reject on the level $\alpha$ the null hypothesis that $\varphi$ and $\psi$ are independent in favor of their positive logarithmic interaction.

If we repeat the previous LISpMiner search by replacing Fisher quantifier by $\chi^2$ quantifier but leave the other parameters unchanged, we will see the same hypotheses and some new ones. Let us take another example.

\(^a\)http://www.statsdirect.com/help: select Chi-square Tests
To find attribute pairs that probably are not independent from each other, set BASE = 300 cases and $\alpha = 0.001$ Then click ‘Generate’.
One of the 38 generated hypotheses says that having at most 2 children and not using any contraceptive are not independent attributes. Thus, the data supports a hypothesis that these attributes depend on each other. – Click ‘Rel row’ to see the percentual shares!
Quantifiers based on binomial distribution

Other statistically justified quantifiers in LISpMiner for testing dependencies are based on binomial distribution\(^a\). In LISpMiner there are four classes of such quantifiers.

Lower critical implication quantifiers \(\varphi(x) \approx_{p,\alpha} \psi(x)\) corresponding to a test (on the level of \(\alpha\)) of a null hypothesis

\(H_0 : P(\varphi(x) | \psi(x)) \leq p\) against the alternative one

\(H_1 : P(\varphi(x) | \psi(x)) > p.\)

If the association rule \(\varphi(x) \approx_{p,\alpha} \psi(x)\) is true in data then the alternative hypothesis is accepted.

For any model \(M\), \(v(\varphi(x) \approx_{p,\alpha} \psi(x)) = \text{TRUE}\) iff \(a \geq \text{BASE}\) and

\[
\sum_{i=a}^{a+b} \frac{(a+b)!}{i!(a+b-i)!} p^i (1 - p)^{a+b-i} \leq \alpha,
\]

where \(0 < p \leq 1, 0 < \alpha \leq 0.5, \text{BASE} > 0.\)

Let us start a new 4ftTask and see an example:

\(^a\)http://www.statsdirect.com/help/: select Binomial distribution
1. Select the Antecedent and Succedent variables.
2. Open Quantifiers and select Lower Critical Implication, adjust parameters $p = 0.9$, $\alpha = 0.1$ and BASE = 5%.
3. Click ‘Generate’.
One of the 4 founded hypotheses supports the dependence of not having any children and not using any contraceptive.
Upper critical implication quantifiers

A kind of ‘opposite’ to lower critical implication quantifiers are

Upper critical implication quantifiers $\phi(x) \approx_{p,\alpha} \psi(x)$

corresponding to a test (on the level of $\alpha$) of a null hypothesis

$H_0 : P(\phi(x)|\psi(x)) \geq p$ against the alternative one

$H_1 : P(\phi(x)|\psi(x)) < p$.

If the association rule $\phi(x) \approx_{p,\alpha} \psi(x)$ is true in data then the alternative hypothesis is accepted.

For any model $M$, $v(\phi(x) \approx_{p,\alpha} \psi(x)) = \text{TRUE}$ iff $a \geq \text{BASE}$ and

$$\sum_{i=a}^{a+b} \frac{(a+b)!}{i!(a+b-i)!} p^i (1 - p)^{a+b-i} > \alpha,$$

where $0 < p \leq 1$, $0 < \alpha \leq 0.5$, $\text{BASE} > 0$.

Let us start a new 4ftTask and see an example:
1. Select Antecedent and Succeedent variables
2. Select Upper Critical Implication with parameters $p = 0.8$, $\alpha = 0.1$ and $\text{BASE} = 2\%$
3. Click ‘Generate’.
All of the produced 132 results are connected to child count. For example, ‘23 year old women’ and ‘having at most 3 children’ are associated attributes.
Lower critical equivalence quantifiers

Much like to lower critical implication quantifiers are lower critical equivalence quantifiers \( \varphi(x) \approx_{p,\alpha} \psi(x) \). However, they are computationally much more time-consuming.

For any model \( M \), \( \nu(\varphi(x) \approx_{p,\alpha} \psi(x)) = \text{TRUE} \) iff \( a \geq \text{BASE} \) and

\[
\sum_{i=a}^{a+b+c+d} \frac{(a+b+c+d)!}{i!(a+b+c+d-i)!} p^i (1 - p)^{a+b+c+d-i} \leq \alpha,
\]

where \( 0 < p \leq 1, 0 < \alpha \leq 0.5, \text{BASE} > 0 \).

Create a new 4ftTask and see an example:
Select as here and generate!
After 9:32:59 minutes and 13,609,257 verifications 24 hypotheses were found. Here one of them is expressed as percentages (use ‘Rel row’). Attributes ‘husband education = 1’ and ‘child count = 5…7 & education = 1’ are associated.
Upper critical equivalence quantifiers

Much like to upper critical implication quantifiers are upper critical equivalence quantifiers \( \varphi(x) \approx_{p,\alpha} \psi(x) \). However, they are again computationally much more time-consuming.

For any model \( M \), \( v(\varphi(x) \approx_{p,\alpha} \psi(x)) = \text{TRUE} \) iff \( a \geq \text{BASE} \) and

\[
\sum_{i=a}^{a+b+c+d} \frac{(a+b+c+d)!}{i!(a+b+c+d-i)!} p^i (1 - p)^{a+b+c+d-i} \geq \alpha,
\]

where \( 0 < p \leq 1, 0 < \alpha \leq 0.5, \text{BASE} > 0 \).

Now create the last 4ftTask:
We use Upper Critical Equivalence to find attributes that are somewhat equivalent to certain amount of children, thus we put the following adjustments.
Attributes ’women in age 20...24’ and ‘child count = 1’ are associated. Notice that, as the used quantifier is a kind of equivalence, the value $d = 1057$ plays an essential role. Had we used a founded implication quantifier, only a very low $p$-value $0.42 (= a/(a+b))$ would have produced this result.
Data mining with GUHA – Part 9
Set differs from Set

Esko Turunen¹

¹Tampere University of Technology
Set differs from Set – SD4ftTask.exe module

The SD4ft–Miner module (SD4ftTask.exe file) is similar to 4ft–Miner, because it searches for association rules and offers many quantifiers such as founded implication, double founded implication etc. The difference is that SD4ft–Miner counts values of the quantifiers for two different sets and then compares these two results – in fact, this generalizes the idea behind above average quantifiers.

This part we

- download SD4ftTask.exe file to LispMiner2010 folder
- give the truth definitions deleted to various SDS-quantifiers
- show some practical examples.
To define the truth of a formula $\phi \approx \phi$, SD4ft–Miner counts values of the quantifiers for two different sets and then compares these two results. For a quantifier there are four alternatives to choose from: basic implication, basic double implication, basic equivalence and above average quantifier, moreover, the comparison can be done in 6 different ways. For example, for basic implication one such alternative is

$$\left| \frac{a_1}{a_1+b_1} - \frac{a_2}{a_2+b_2} \right| \geq p, \ a_1 \geq \text{base}_1, \ a_2 \geq \text{base}_2$$

**Exercises**

18. How would you define ’Model $N$ is implicationally better than model $M$’ in SDS framework?

19. Based on Exercise 18, discuss how to define implicational SDS –quantifiers.

20. What are the (obvious) truth conditions for the other quantifiers available in SD4ft–Miner?
download LM.SD4ft.zip
unzip it to your LispMiner2010 folder
Your C:\LispMiner2010 should look like this. Open SD4ftTask.exe and select there (as usual!) Indonesia1.mdb MB
SD4ft-tasks are created much in the same way than 4ft-tasks are. Here we have created three examples. You get them by first clicking ‘Add’ and then filling empty fields as follows:
The analytical question

Difference in founded implication greater than 0.3

Both subsets have at least 10 elements

Age and child count are in 1 – 4 intervals, others in 1-1 subsets

Subsets whose possible difference is studied
One of the 22 hypotheses says: Among 40–42 year old women with 4–6 children, among Islamite only 15 out of 33 use long term contraception, while among non-Islamite 10 out of 13 do use long term contraception.
If you construct several more or less similar tasks, a good idea is to take a clone of an existing task. Here B. Example was done by a clone of A. Example.

In B. Example we have also introduced two new predicates, namely ‘Low education’ (Education = 1) and ‘High education’ (Education = 4), see from Part 6 how this can be done.
This task is much similar to the previous one, there we studied the influence of religion to contraception method.

Here ‘religion’ is replaced by ‘education’.

The required minimal subsets are now a bit bigger (at least 20 cases), while the quantifier is the same.
58 hypotheses were found. For example, among families with maximally one child there is a big difference between women with low education and high education. In the first set nearly nobody uses any contraception, while in the second set almost every second woman uses some contraceptives.
The third example is created by the previous one by first taking a **clone** of B. Example.

Here the analytic question sounds: Does husband’s education have an effect on child count? Therefore we need two new predicates ‘Husbands education low’ (= 1) and ‘Husband’s education high’ (= 4), see Part 6 how these can be created.

The antecedent part is now ‘child count’

We use quantifier ‘Absolute difference in Above average values’
The Succedent part is 'Child count', here from 1 to 3 predicates, each an interval of length 1 to 3, so the widest possible output could be a 9 year age interval.

Choose AAI DiffValAbs-quantifier.
There are $10 + 8 + 37 + 179 = 234$ women whose husband’s occupation category is ’2’. In the first set, 18 in number, more than every second has 5 to 7 children, while in the second set, 215 in number, less than every fifth has 5 to 7 children.
Data mining with GUHA – Part 10
Action Miner

Esko Turunen

1 Tampere University of Technology
Thanks to its profound foundations, the GUHA method is flexible for new ideas and approaches. One of such new approach in data mining are action rules invented by Ras and Wieczorkowska in 2000. The idea is to find dynamic features in data; same predicates are considered stable attributes and some others are flexible attributes.

Ac4ftTask.exe module, an action rule miner, has been added to LISp–Miner very recently and is still under development. Action Miner uses existing GUHA quantifiers that are combined in a new way.

Instead of going into theoretical details and definitions, we show examples that hopefully clarify some of the possibilities that are available. – First, however, we have to download Ac4ftTask.exe module.
2. Download LM.Ac4ft.zip, extract all files to C:\LISpMiner2010
3. These are the 4 new items you will need
4. Open Ac4ftTask.exe, select there Indonesia1.mdb MB, add a new task and you should have the following view …
...if everything went well, you should have this view.

The first task is to determine the analytic question we want to search.
Analytic question: If Education and Husband’s Education are stable, how does changes in child count effect on Contraception?

Stable part: Education and Husband’s Education Coefficient is interval 1 - 1

Variable part: Child count Coefficient is interval 1 - 1

Quantifiers are much like in Above average – quantifiers! Set Base before = Base after = 50 Abs. diff. in Founded implic. values is at least 0.5

Stable part: Contraceptive Coefficient is interval 1-1

Leave empty
Ignoring symmetric cases, there are 5 hypotheses, all being subsets of cases where husband is highly educated (HusbEduc = 4).

This one says:
After having first child the use of contraceptives dramatically increases. Indeed, nobody of the 65 women having no children uses any contraceptives, while majority (91 out of 178) of women having one child use contraceptives.
Analytical question to be examined: among young women, if child count and age are stable, which factors increase living standard?

Create predicate ‘Age < 30’, see Part 6!
Age < 30 and child count are in 1 – 5 intervals

HusbEduc, HubbOccp, IsWorking, Education are 1-1 subsets

Choosing quantifier: BASE-values and Founded implication-quantifier values in ‘Before’ and ‘After’ sets should have more or less same values.

Living Std is a 1 – 1 subset
26 non symmetric hypotheses we found. One of them is here and reveals a fact:

Among families with 1 or 2 children, wife 23 to 26 year old and husband highly educated, if husbands occupation changes from status 3 to status 1, then standard of living improves from level 3 to level 4.

Indeed, the distributions of these sets are almost equal: 23 – 11 and 24 – 15.
Some additional remarks

We have seen only an outline of LISpMiner software – there are plenty of details, features and even whole modules we did not touch at all. The best (and only!) way to learn them is self–education. We believe that this teaching material helps to get started.

Besides, GUHA and LISpMiner are subjects of an intensive investigation and experiment, so new subjects and theme may appear. Follow the home page http://lispminer.vse.cz/index.html. – If some of your modules are obsolete, you have to download a new version from this page.
Data mining exercise

Characterize, in a dozen of LISpMiner hypotheses, the use of preservatives among 25 – 35 year old Indonesian women. What is characteristic of them? Among this subset of the whole Indonesia data, what kind of differences among various social, educational religious etc groups can be found? Use at least three different approaches. The report you write should contain

- cover, a page containing your name, country, student number, e-mail address and name of this teaching module
- a brief introduction to GUHA method and to LISpMiner, 2 – 3 pages
- a description of the Indonesia data and the aim of this work, 2 – 3 pages
- your results: annex screen prints from Result module and comment them, 10 – 20 page