Machine Learning and Association rules

Petr Berka, Jan Rauch
University of Economics, Prague
{berka|rauch}@vse.cz

Tutorial Outline

- Statistics, machine learning and data mining – basic concepts, similarities and differences (P. Berka)
- Machine Learning Methods and Algorithms – general overview and selected methods (P. Berka)
- *Break*
- GUHA Method and LISp-Miner System (J.Rauch)

Part 1

Statistics, machine learning and data mining

Statistics

- A formal science that deals with collection, analysis, interpretation, explanation and presentation of (usually numerical) data.
- The science of making effective use of numerical data relating to groups of individuals or experiments

(wikipedia)

Machine Learning

- „The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.“

  (Mitchell, 1997)

- „Things learn when they change their behavior in a way that makes them perform better in a future.“

  (Witten, Frank, 1999)

Knowledge Discovery in Databases

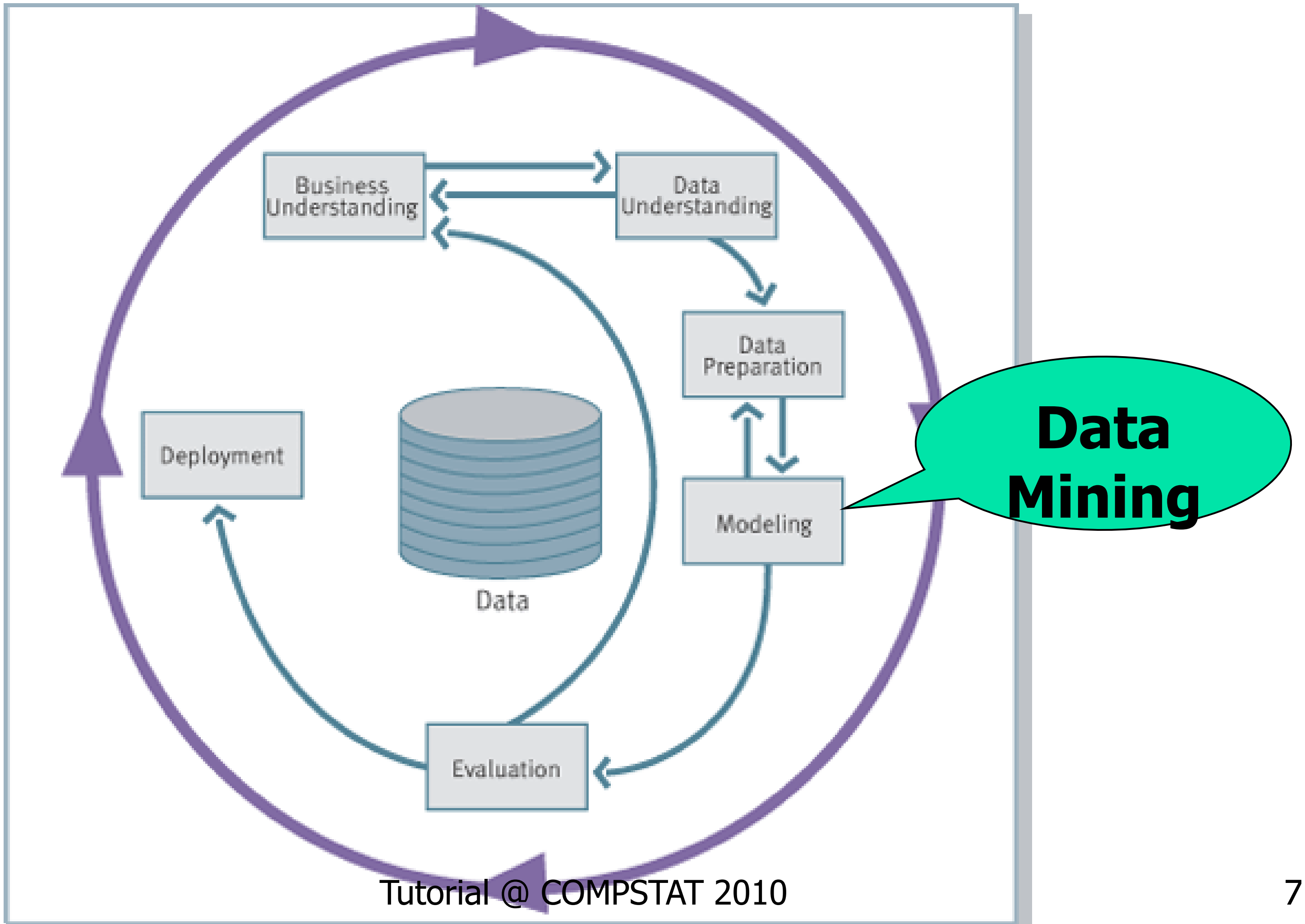
- „Non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns from data."

  (Fayyad et al., 1996)

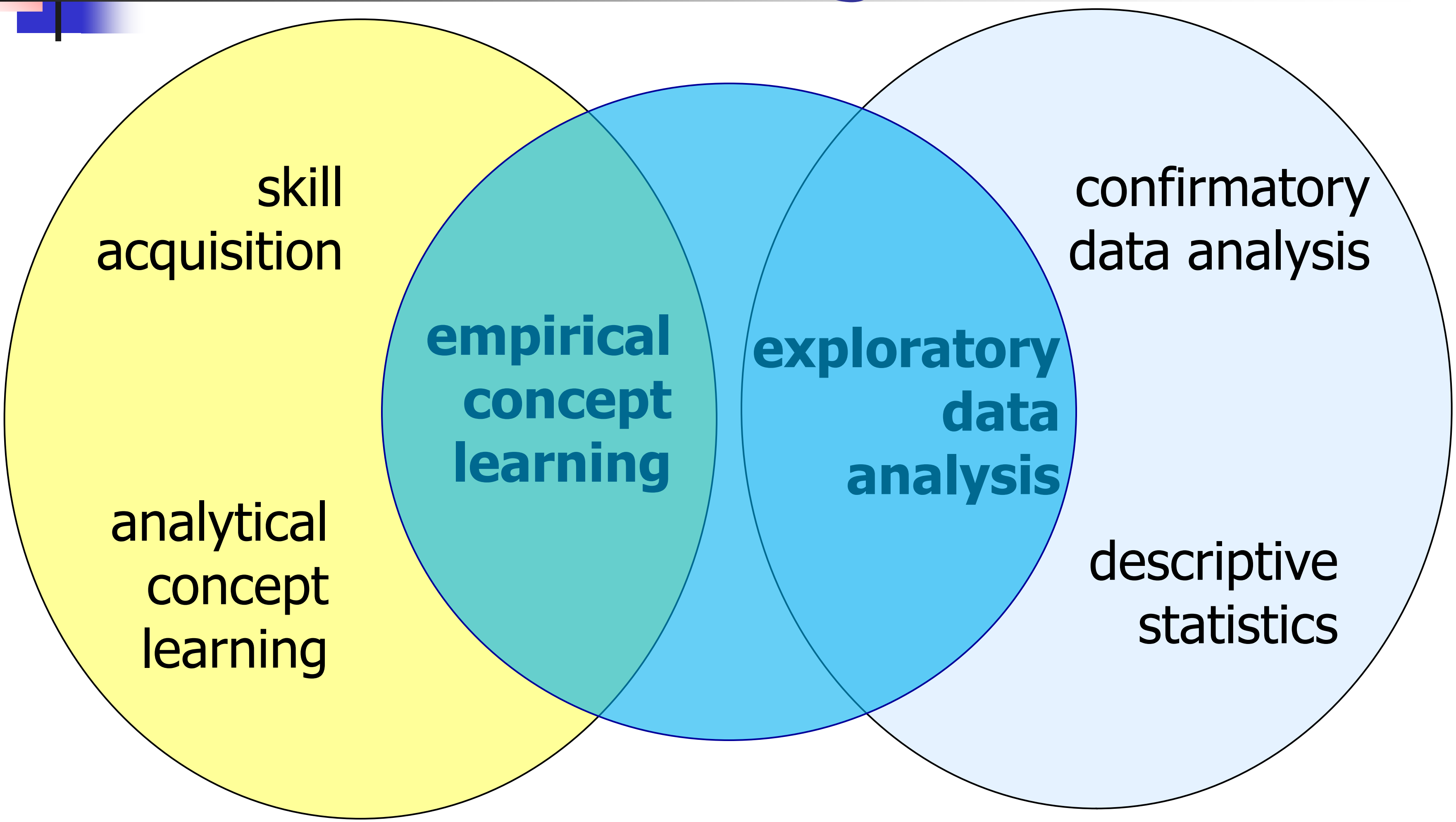
- „Analysis of observational data sets to find unsuspected relationships and summarize data in novel ways that are both understandable and useful to the data owner."

  (Hand, Manilla, Smyth, 2001)

The CRISP-DM Methodology



Machine Learning | Data Mining | Statistics



Statistics vs. Machine Learing

- Hypothesis driven
- Model oriented
  - formulate hypothesis
  - collect data (in a controlled way)
  - analyze data
  - interpret results

- Data driven
- Algorithm oriented
  - formulate a task
  - preprocess available data
  - apply (different) algorithms
  - interpret results

Terminological differences

Machine Learning	Statistics
attribute	variable
target attribute, class	dependent variable, response
input attribute	independent variable, predictor
learning	fitting, parameter estimation
weights (in neural nets)	parameters (in regression)
error	residuum

Similarities

- algorithms
  - decision trees: C4.5 ~ CART
  - neural networks ~ regression
  - nearest neighbor classification
- methods
  - cross-validation test
  - χ^2 test

Part 2

Machine Learning Methods and Algorithms

Learning methods

- rote learning (memoryzing)
- learning from instruction, learning by being told
- learning by analogy, instance-based learning, lazy learning
- explanation-based learning
- learning from examples
- learning from observation and discovery

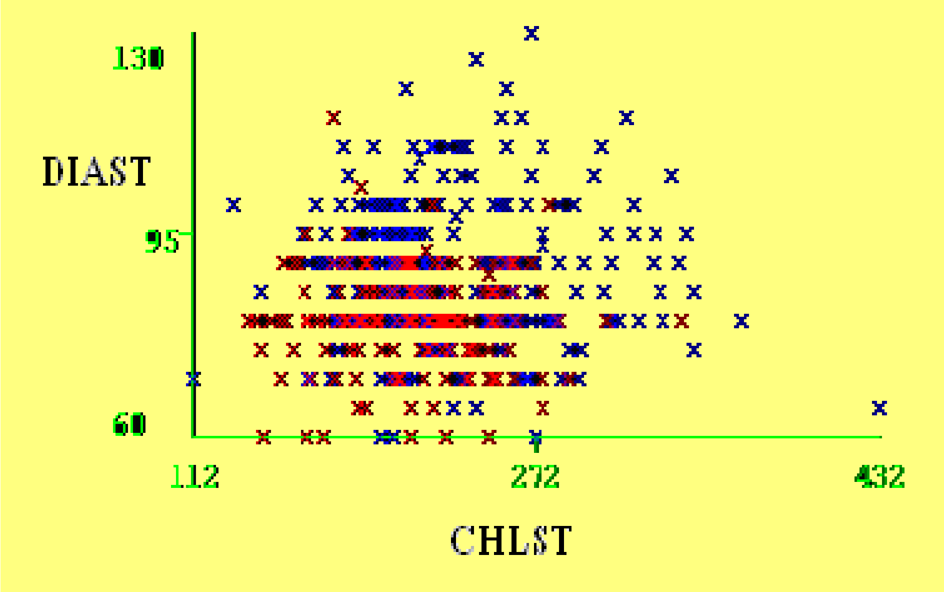
Feedback during learning

- pre-classified examples (**supervised learning**)
- rewards or punishments (**reinforcement learning**)
- indirect hints derived from the behaviour of teacher (**apprenticeship learning**)
- nothing (**unsupervised learning**)

Illustrative Example

Data about pacients with different atherosclerosis risk

Pac-id	DIAST	CHLST	risk
P1	100	300	Ano
P2	85	247	Ne
P3	87	291	Ano
P4	105	259	Ano
P5	81	231	Ne
P6	105	288	Ano
. . .			

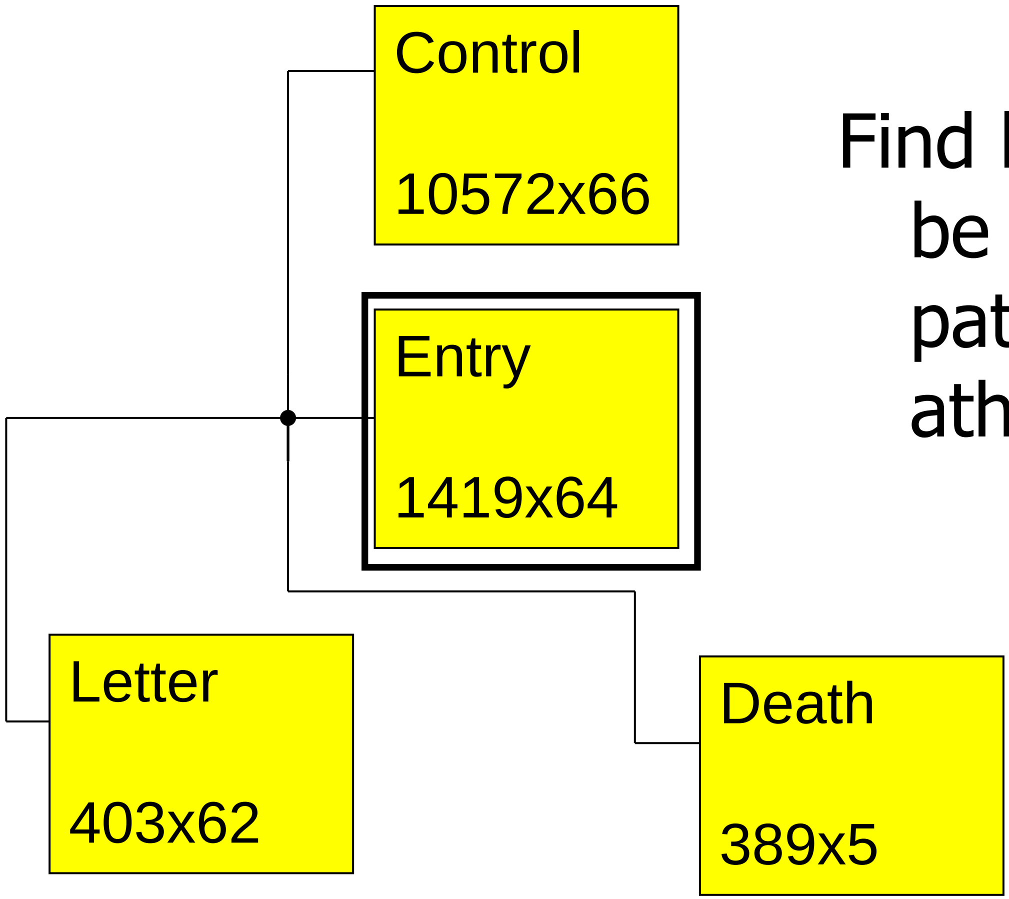


Atherosclerosis risk factors study

Longitudinal (1975-2000) study of atherosclerosis risk factors in the population of middle-aged men divided into three groups (normal, risk, pathological).

- to identify atherosclerosis risk factors prevalence in a population of middle-aged men,
- to follow the development of these risk factors and their impact on the examined men health, especially with respect to atherosclerotic CVD,
- to study the impact of complex risk factors intervention on development of risk factors and CVD mortality,
- to compare (after 10-12 years) risk factors profile and health of the selected men in different groups.

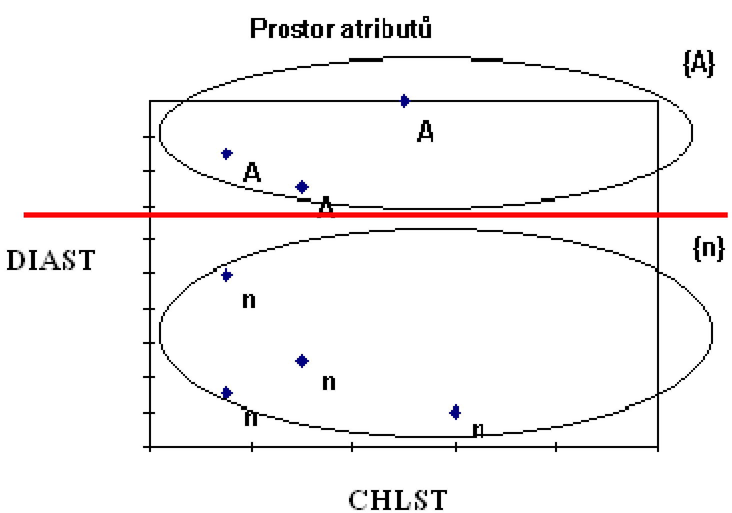
Data STULONG

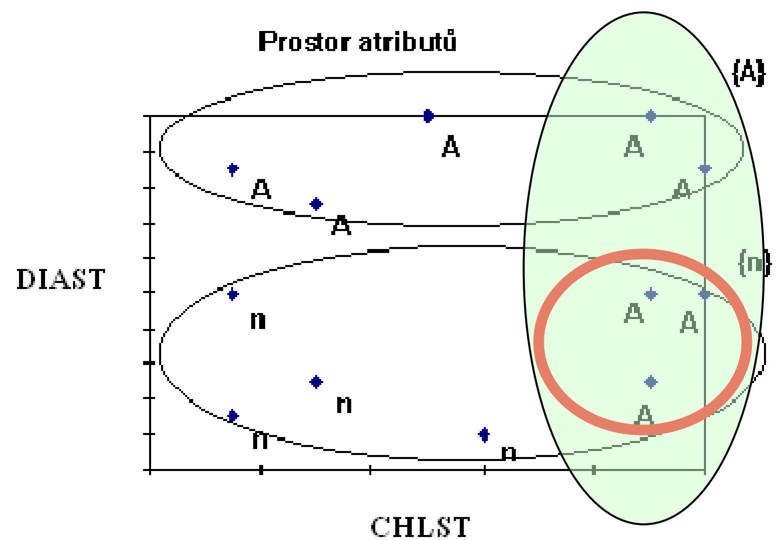


Find knowledge that can be used to classify new patients according to atherosclerosis risk

Empirical concept learning

- examples belonging to the same class have similar characteristics (**similarity-based learning**)
- we infer general knowledge from a finite set of examples (**inductive learning**)





Empirical concept learning from data (1/3)

- Analyzed data

$$\mathbf{D}_{TR} = \begin{bmatrix} x_{11} & x_{12} & \cdots\cdots & x_{1m} & y_1 \\ x_{21} & x_{22} & \cdots\cdots & x_{2m} & y_2 \\ \vdots & \vdots & & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots\cdots & x_{nm} & y_n \end{bmatrix}$$

- Classification task: we search for **knowledge** (represented by a decision function f) $f: \boldsymbol{x} \rightarrow y$, that for input values $\boldsymbol{x}$ of an example infers the value of target attribute $\hat{y} = f(\boldsymbol{x})$.

Empirical concept learning from data (2/3)

- During classification of an example we can make an error $Q_f(\mathbf{o}_i, \hat{y}_i)$:

$$Q_f(\mathbf{o_i}, \hat{y}_i) = (y_i - \hat{y}_i)^2 \qquad Q_f(\mathbf{o_i}, \hat{y}_i) = \begin{cases} 1 & \text{for} \quad y_i \neq \hat{y}_i \\ 0 & \text{for} \quad y_i = \hat{y}_i \end{cases}$$

- For the whole training data $\mathbf{D}_{TR}$ we can compute the total error $Err(f,\mathbf{D}_{TR})$, e.g. as

$$Err(f, D_{TR}) = \frac{1}{n}\sum_{i=1}^{n} Q_f(\mathbf{o_i}, \hat{y}_i)$$

Empirical concept learning from data (3/3)

- The goal of learning is to find such a knowledge f*, that will minimize this error

$$\mathrm{Err}(f^*, D_{TR}) = \min_f \mathrm{Err}(f, D_{TR})$$

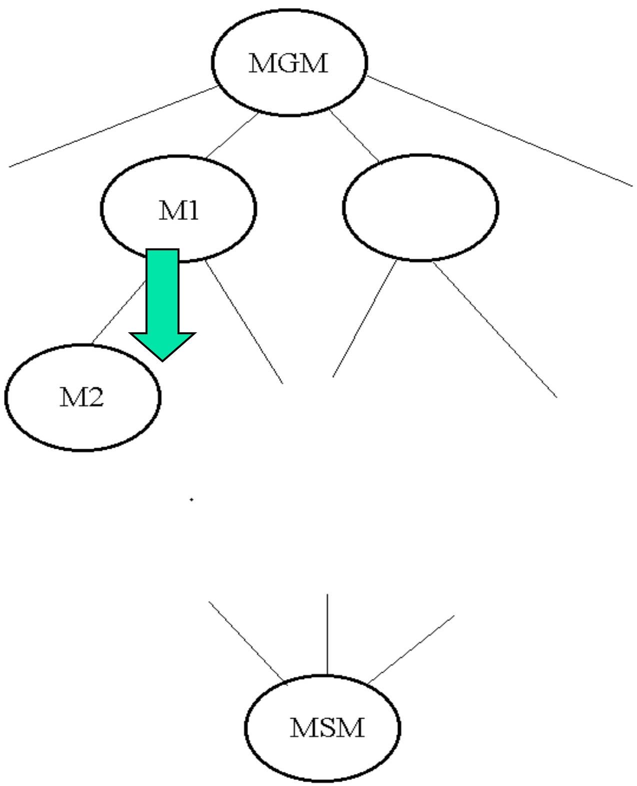
Empirical concept learning as …

- … **search**
  - we are learning both the structure and parameters of a model

- … **approximation**
  - we are learning the parameters of a model

Search (1/2)

- Ordering of models



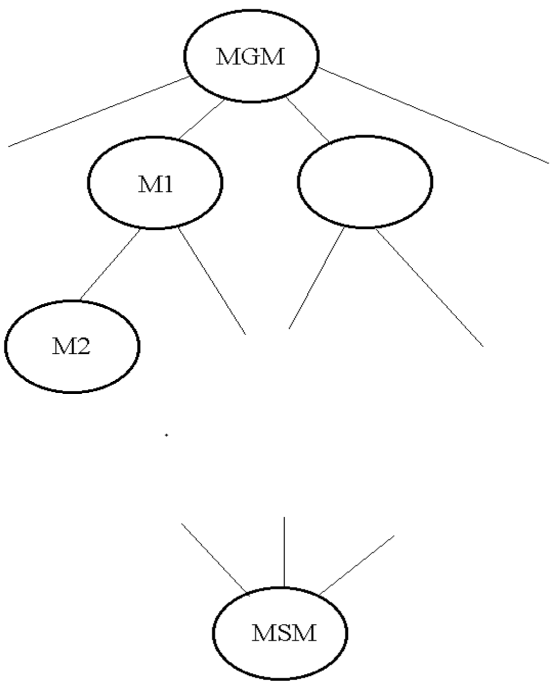
MGM –most general model
(one cluster for all examples)

M1 more general than M2
M2 more specific than M1

MSM – most specific model(s)
(single cluster for each example)

Search (2/2)

- Search methods



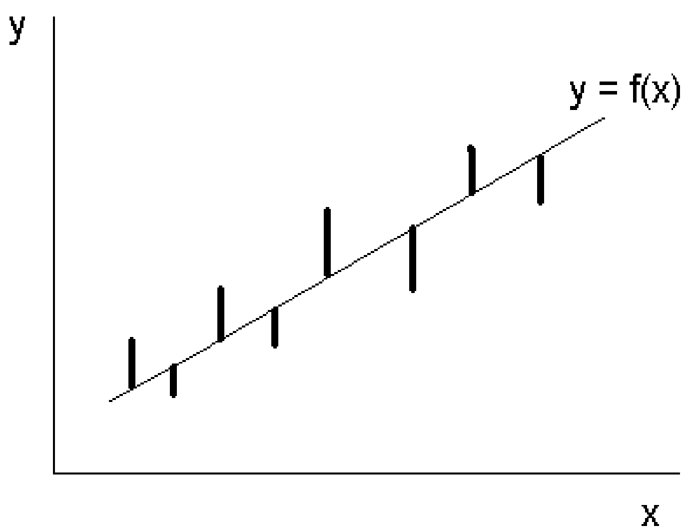
Direction
- top-down
- bottom-up

Strategy
- blind
- heuristic
- random

Breadth
- single
- parallel

Approximation (1/2)

Estimation of the parameters of a model (decision function) $y=f(x)$ using a set of the values $[x_i, y_i]$



Least squares method:

Looking for parameters that minimize the overall error

$$\sum_i (y_i - f(x_i))^2$$

transformed to solving the equation

$$\frac{d}{dq}\sum_i \left(y_i - f(x_i)\right)^2 = 0$$

Approximation (2/2)

- Analytical solution (known type of the function)

  solving a set of equations for the parameters

  - regression
- Numerical solution (unknown type of the function)
  - gradient methods

$$\nabla \mathrm{Err}(\mathbf{q}) = \left[\frac{\partial \mathrm{Err}}{\partial q_0}, \frac{\partial \mathrm{Err}}{\partial q_1}, \ldots, \frac{\partial \mathrm{Err}}{\partial q_Q} \right]$$

Modification of parameters $\mathbf{q} = [q_0, q_1, \ldots, q_Q]$ as $q_j \leftarrow q_j + \Delta q_j$
where

$$\Delta q_j = -\eta \frac{\partial \mathrm{Err}}{\partial q_j}$$

Selected algorithms

- decision trees
- decision rules
- association rules
- neural networks
- genetic algorithms
- bayesian methods
- nearest-neighbor methods

Decision tree algorithms

TDIDT algorithm

1. select the best splitting attribute as a root of the current (sub)tree,
2. divide data in this node into subsets according to the values of the selected attribute and add new node for each this subset,
3. if there is an added node, for which the data do not belong to the same class, goto step 1.

- only categorial attributes
- only data without noise

Splitting criteria

- How to select a splitting attribute?

	Y_1	Y_2	…	Y_S	$\sum$
X_1	a_{11}	a_{12}		a_{1s}	r_1
X_2	a_{21}	a_{22}		a_{2s}	r_2
:	:	:			
X_R	a_{r1}	a_{r2}		a_{r2}	r_r
$\sum$	s_1	s_2		s_s	n

Contingency table
Y class attribute
X input attribute

Entropy (min) – ID3, C4.5

$$H(X) = \sum_{i=1}^{R} \frac{r_i}{n} \left(- \sum_{j=1}^{S} \frac{a_{ij}}{r_i} \log_2 \frac{a_{ij}}{r_i} \right)$$

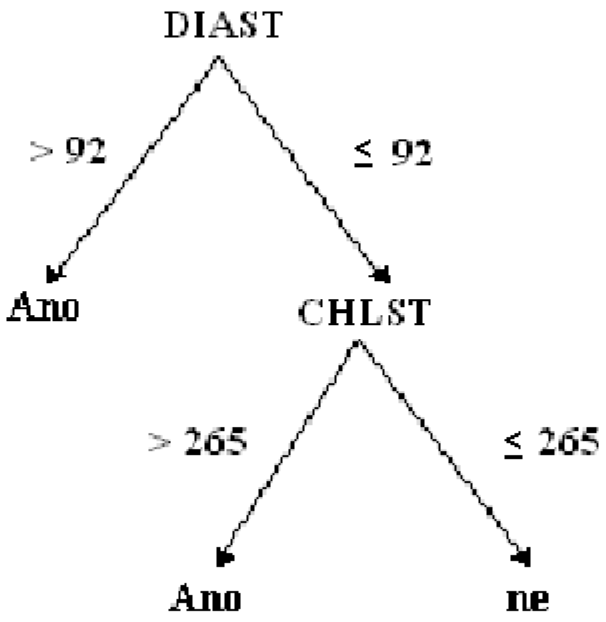
Gini index (min) - CART

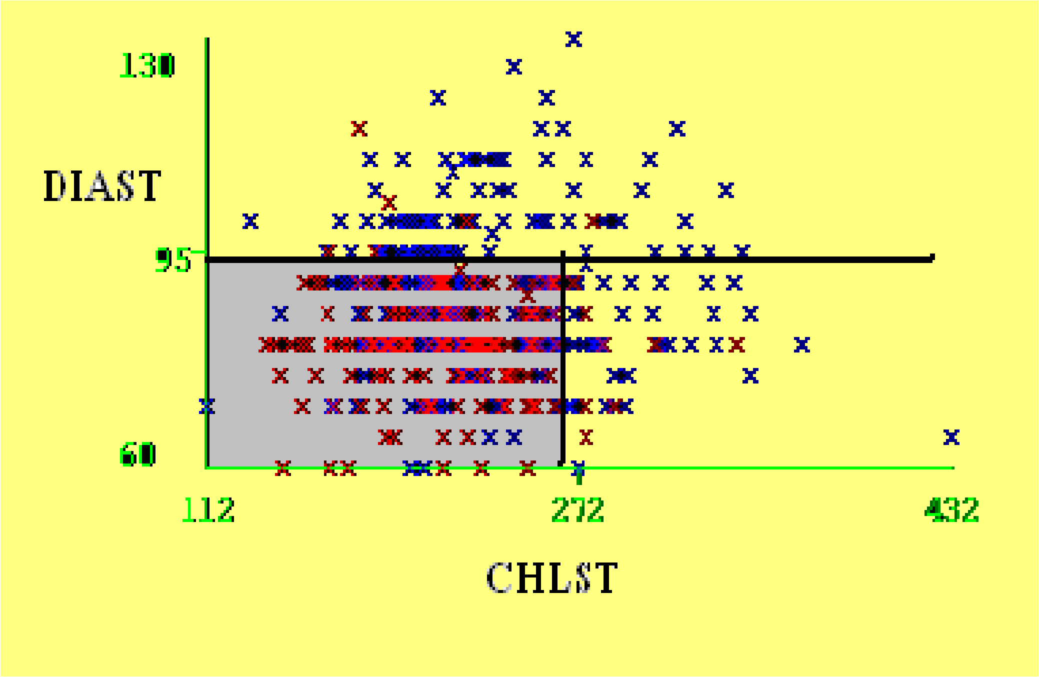
$$Gini(X) = \sum_{i=1}^{R} \frac{r_i}{n} \left(1 - \sum_{j=1}^{S} \left(\frac{a_{ij}}{r_i} \right)^2 \right)$$

χ^2 (max) - CHAID

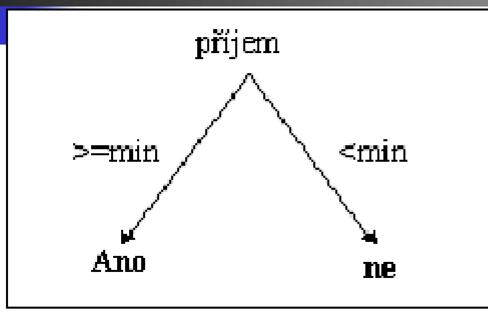
$$\chi^2(X) = \sum_{i=1}^{R} \sum_{j=1}^{S} \frac{\left(a_{ij} - \frac{r_i \cdot s_j}{n} \right)^2}{\frac{r_i \cdot s_j}{n}}$$

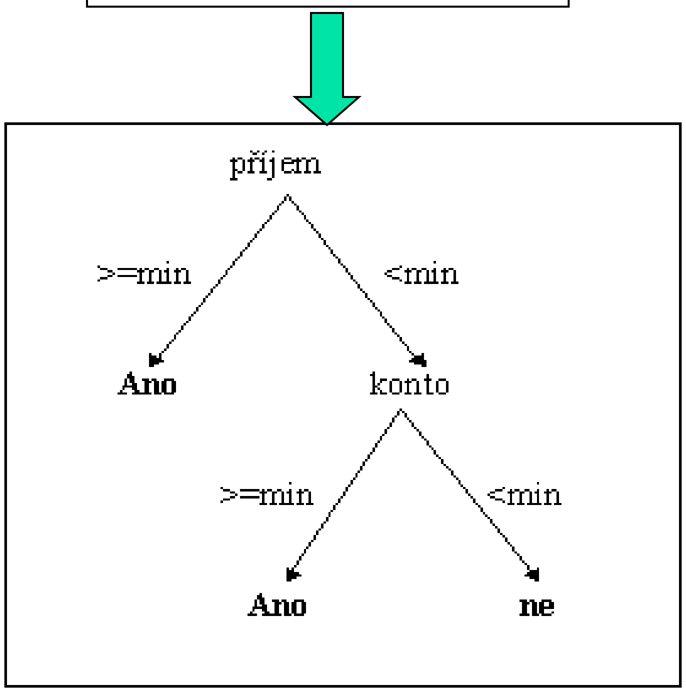
Decision trees in the attribute space





Decision trees (search)





- top-down (TDIDT)
  - single, heuristic
    - ID3, C4.5 (Quinlan), CART (Breiman a kol.)
  - parallel heuristic
    - Option trees (Buntine), Random forrest (Breiman)
- random
  - parallel
    - using genetic programming
- bottom-up additional technique during tree pruning

Decision rules – set covering algorithms

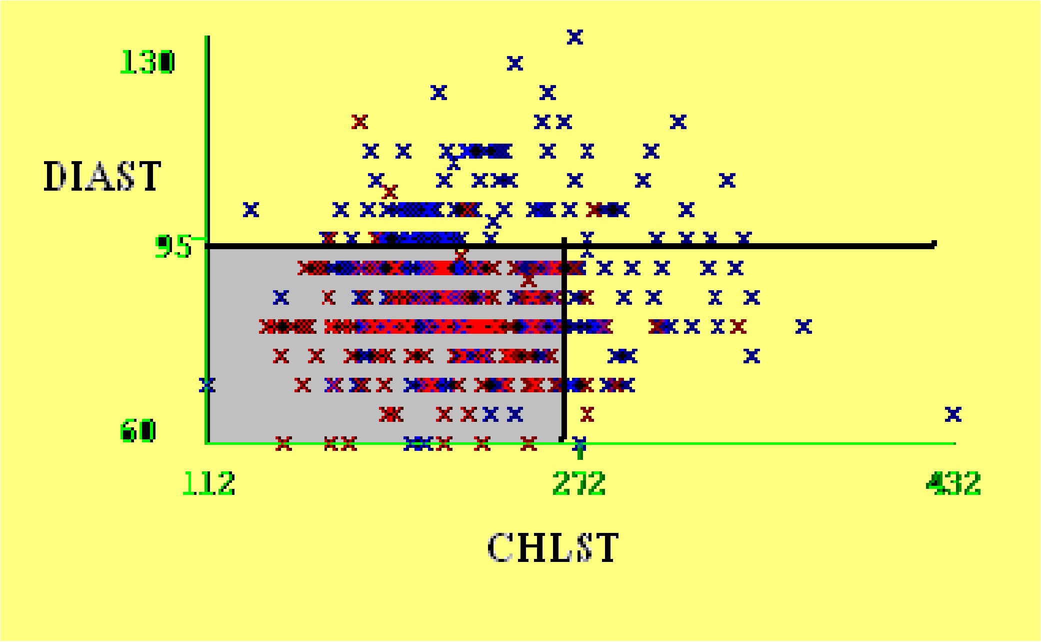
set covering algorithm

1. create a rule that covers some examples of one class and does not cover any examples of other classes
2. remove covered examples from training data
3. if there are some examples not covered by any rule, go to step 1

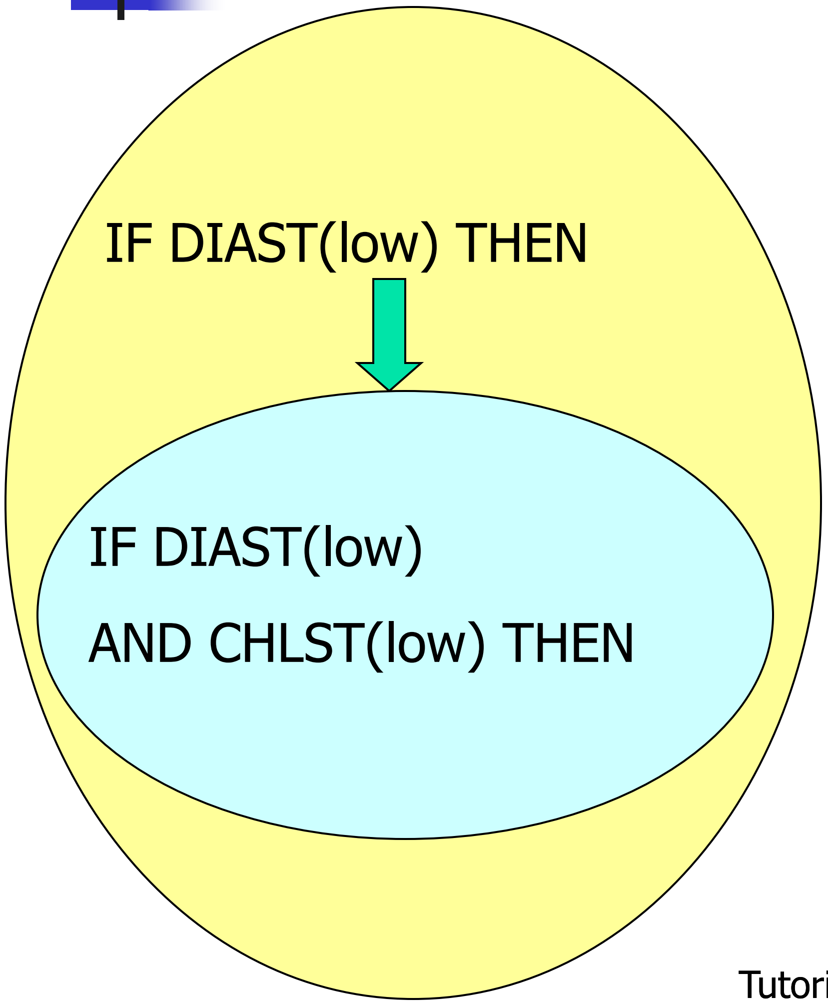
each training example covered by single rule = straightforward use during classification

Decision rules in the attribute space

IF DIASThigh) THEN risk(yes)
IF CHLST(high) THEN risk(yes)
IF DIAST(low) $\wedge$ CHLST(low) THEN risk(no)



Decision rules (search)



- top-down
  - parallel heuristic
    - CN2 (Clark, Niblett), CN4 (Bruha)
- bottom-up
  - single heuristic
    - Find-S (Mitchell)
  - parallel heuristic
    - AQ (Michalski)
- random
  - parallel
    - GA-CN4 (Králík, Bruha)

Decision rules – compositional algorithms (search)

KEX algorithm

1 add empty rule to the rule set *KB*
2 repeat
2.1 find by rule specialization a rule *Ant* $\Rightarrow$ *C* that fulfils the user given criteria on length and validity,
2.2 if this rule significantly improves the set of rules *KB* build so far then add the rule to *KB*

each training example can be covered by more rules = these rules contribute to the final decision during classification

KEX algorithm – more details

KEX algorithm

Initialization

1. forall category (attribute-value pair) A add $A \Rightarrow C$ to $OPEN$
2. add empty rule to the rule set KB

Main loop

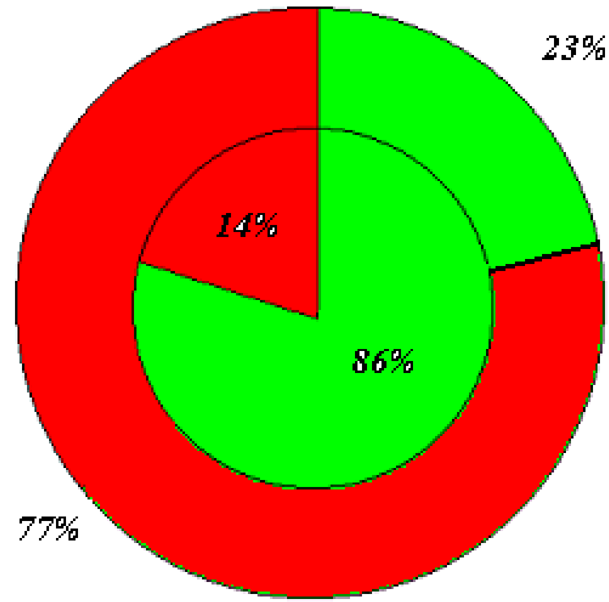
while $OPEN$ is not empty

1. **select** the first implication $Ant \Rightarrow C$ from $OPEN$
2. **test** if this implication significantly improves the set of rules KB built so far (using the χ^2 test, we test the difference between the rule validity and the result of classification of an example covered by Ant) then add it as a new rule to KB
3. for all possible categories A
   (a) **expand** the implication $Ant \Rightarrow C$ by adding A to Ant
   (b) **add** $Ant \wedge A \Rightarrow C$ to $OPEN$ so that $OPEN$ remains ordered according to decreasing frequency of the condition of rules
4. **remove** $Ant \Rightarrow C$ from $OPEN$

Association rules

IF smoking(no) $\wedge$ diast(low) THEN chlst(low)

	SUC	$\neg$SUC	Σ
ANT	257	43	300
$\neg$ANT	66	1036	1102
Σ	323	1079	1402



- support $a/(a+b+c+d)$ = 0.18
- confidence $a/(a+b)$ = 0.86

Association rule (generating as top-down search)

breadth-first

Apriori (Agrawal), LISp-Miner (Rauch)

combination
. . .
4a
4n
5a
5n
1n 2n
1n 2s
1n 2v
1n 3m
1n 3z
. . .

depth-first

combination
1n
1n 2n
1n 2n 3m
1n 2n 3m 4a
1n 2n 3m 4a 5a
1n 2n 3m 4a 5n
1n 2n 3m 4n
1n 2n 3m 4n 5a
1n 2n 3m 4n 5n
1n 2n 3m 5a
1n 2n 3m 5n

heuristic

KAD (Ivánek, Stejskal)

combination
5a
1n
3m
3z
4a
4n
1v
1n 4a
4n 5a
1v 5a
2v

Association rules algorithm

apriori algorithm

1. set k=1 and add all items that reach *minsup* into L
2. repeat
    1. increase k
    2. consider an itemset C of length k
    3. if all subsets of length k-1 of the itemset C are in L then
        if C reaches *minsup* then add C into L

apriori – more details

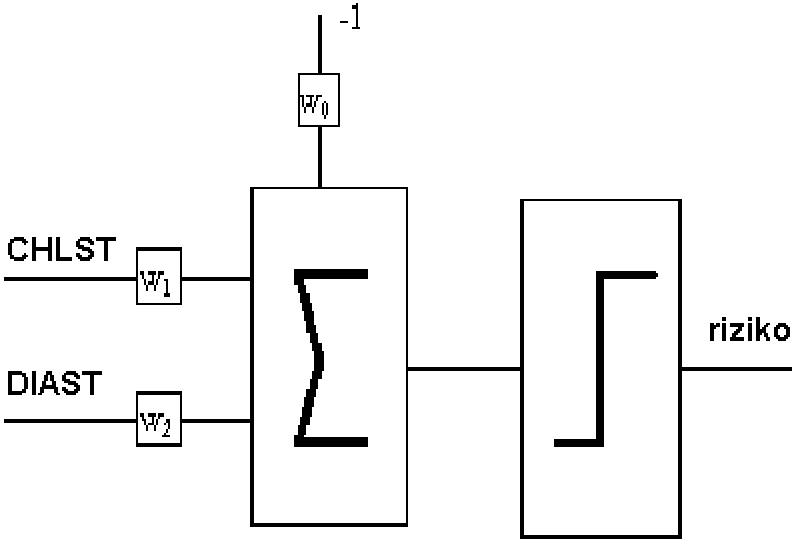
algorithm apriori

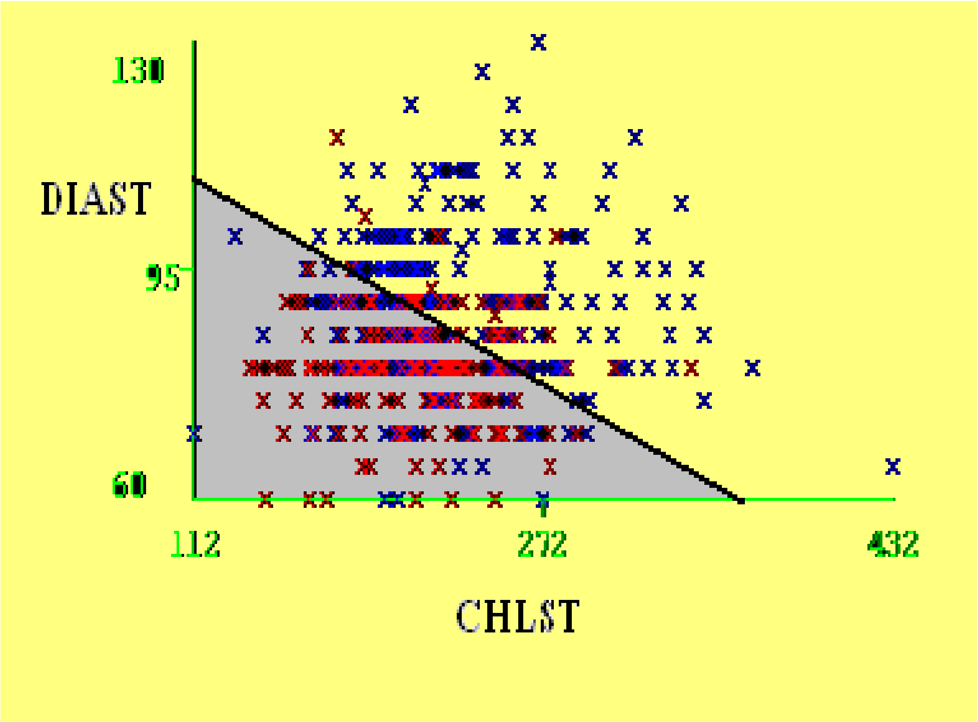
1. assign all items that reached the support of $minsup$ into L_1
2. let $k = 2$
3. while $L_{k-1} \neq \oslash$
   3.1 using the function **apriori-gen** create a set of candidates C_k from L_{k-1}
   3.2 assign all itemsets from C_k that reached the support of $minsup$ into L_k
   3.3 increase k by 1

Function apriori-gen(L_{k-1})

1. for all itemsets $Comb_p$, $Comb_q$ from L_{k-1}
   if $Comb_p$ and $Comb_q$ share $k-2$ items, then add $Comb_p \wedge Comb_q$ to C_k
2. for all itemsets $Comb$ from C_k
   if any subset with a length $k-1$ of $Comb$ is not included in L_{k-1} then remove $Comb$ from C_k

Neural networks – single neuron

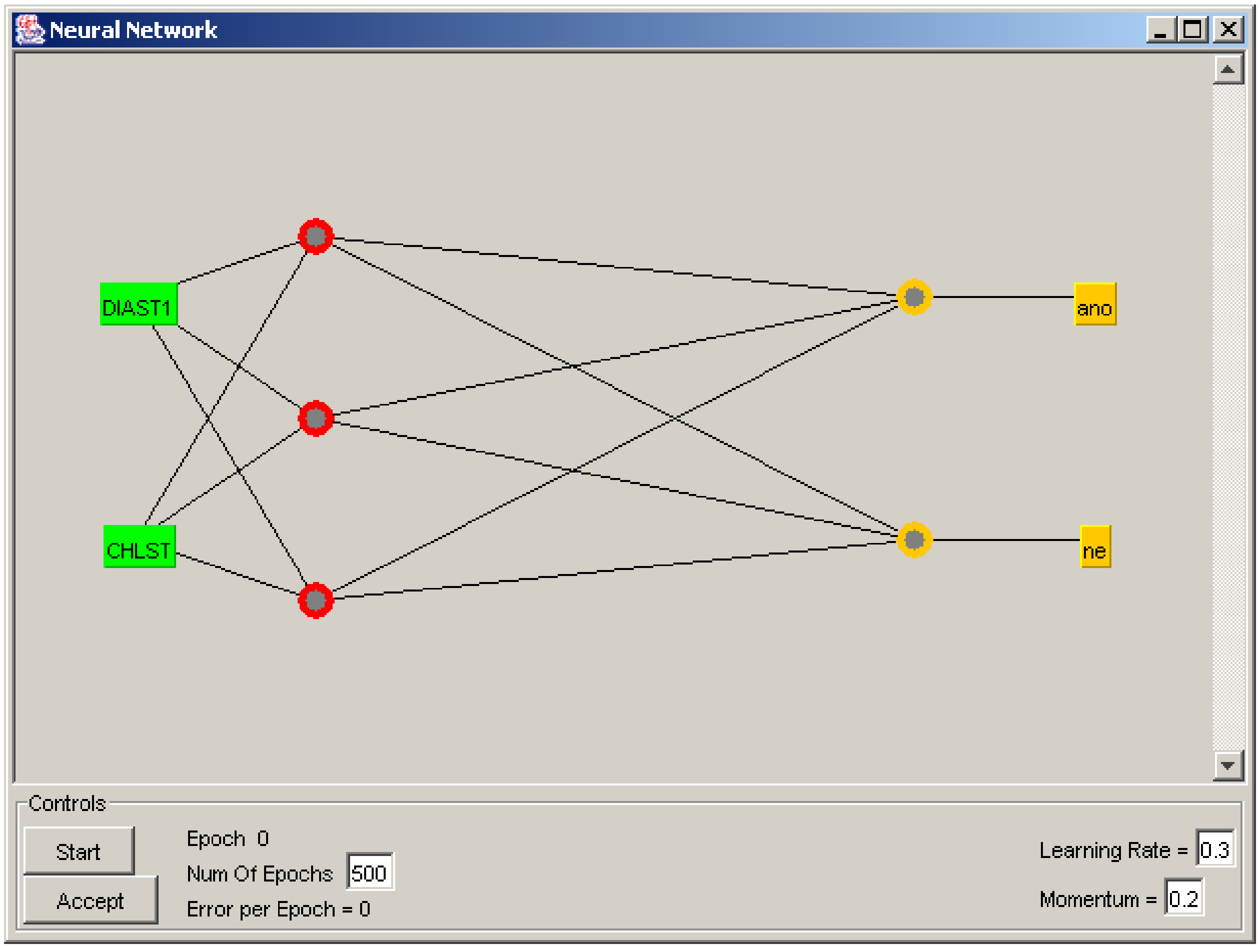




$$y' = 1 \quad \text{for} \quad \sum_{i=1}^{m} w_i x_i \geq w_0$$

$$y' = 0 \quad \text{for} \quad \sum_{i=1}^{m} w_i x_i < w_0$$

Neural networks -multilayer perceptron



Backpropagation algorithm = approximation

Error backpropagation algorithm

1. inicialize the weights in the network with small random numbers (e.g. from the interval $[-0.05, 0.05]$)
2. while the stopping condition is not satisfied for every training example $[\mathbf{x}, y]$ do
   2.1 compute the output out_u for every neuron u
   2.2 for every neuron o from the output layer compute the error

$$error_o = out_o(1 - out_o)(y_o - out_o)$$

   2.3 for every neuron h from the hidden layer compute the error

$$error_h = out_h(1 - out_h) \sum_{o \in output} (w_{ho} error_o)$$

   2.4 for every connection from neuron j to neuron k modify the weight of the connection

$$w_{jk} = w_{jk} + \Delta w_{jk},$$

   where

$$\Delta w_{jk} = \eta \, error_k \, x_{jk}$$

Genetic algorithms = parallel random search

Genetic algorithm(fit,N,K,M)

Initialization

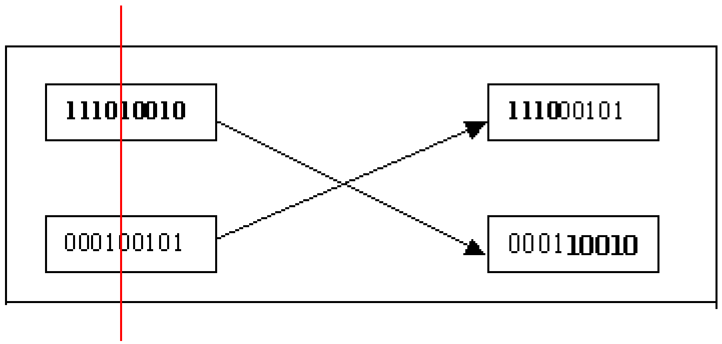
1. assign t:= 0
2. randomly create the initial population $Q(t)$ which contains N individuals
3. compute $fit(h)$ for every individual $h \in Q(t)$

Main loop

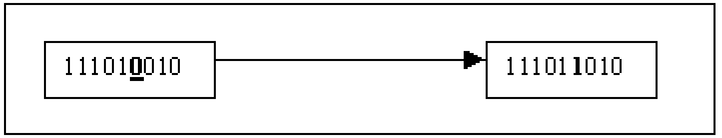
1. while the stopping condition is not satisfied do
   1.1 **selection:** select individuals h from the population $Q(t)$ that will be directly inserted into the population $P(t+1)$
   1.2 **crossover:** choose pairs of individuals (with probability K) from the population $Q(t)$, perform crossover on each pair and insert the offsprings into the population $Q(t+1)$
   1.3 **mutation:** choose individuals h (with probability M) from the population $Q(t+1)$ and mutate them
   1.4 assign t := t + 1
   1.5 compute $fit(h)$ for every individual $h \in Q(t)$
2. return the individual h with the highest value of $fit(h)$

Genetic algorithms

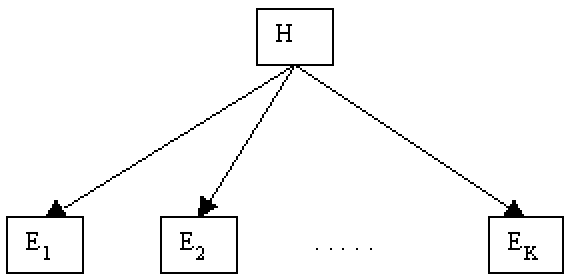
- Genetic operations
  - Selection
  - Cross-over



  - Mutation

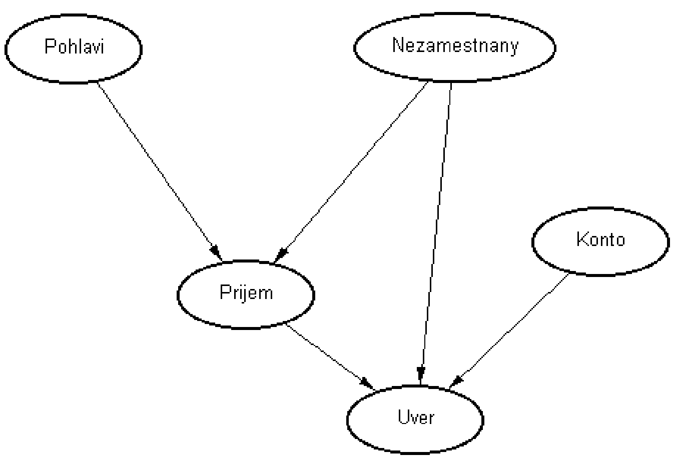


Bayesian methods



- Naive bayesian classifier (approximation)

$$P(H \mid E_1,\ldots,E_K) = \frac{\prod_{k=1}^{K} P(E_k \mid H)\;\; P(H)}{\cancel{P(\;)}}$$



- Bayesian network (search, approximace)

$$P(u_1,\ldots,u_n) = \prod_{ii-1}^{n} P(u_i \mid rodiče(u_i))$$

Naive bayesian classifier

- Computing the probabilities

  P(risk=yes) = 0.71 P(risk=no) = 0.19

  P(smoking=yes)|risk=yes) = 0.81

  P(smoking=no)|risk=no) = 0.19

  . . .

- Classification

  Class H_i with highest value of $\prod_k P(E_k|H_i)\, P(H_i)$

Nearest-neighbor methods

Algorithm k-NN

Learning

Add examples [$\boldsymbol{x}_i$, y_i] into case base

Classification

1. For a new example $\boldsymbol{x}$

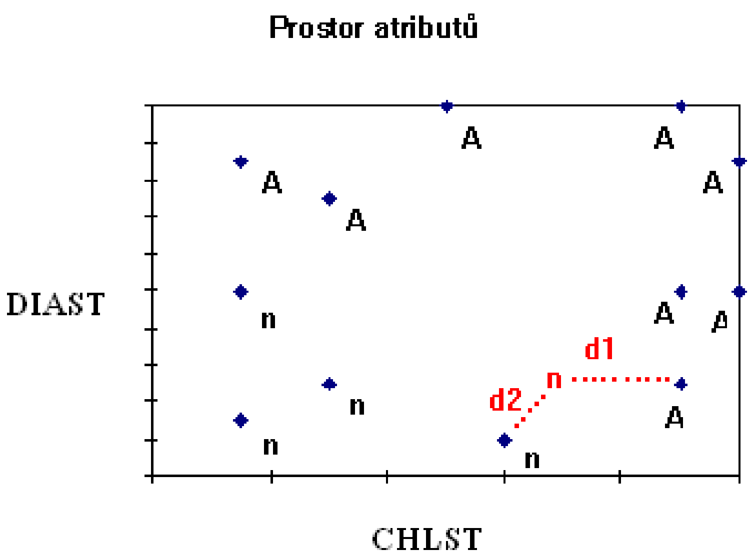
   1.1. Find $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_K$ K nearest neighbors

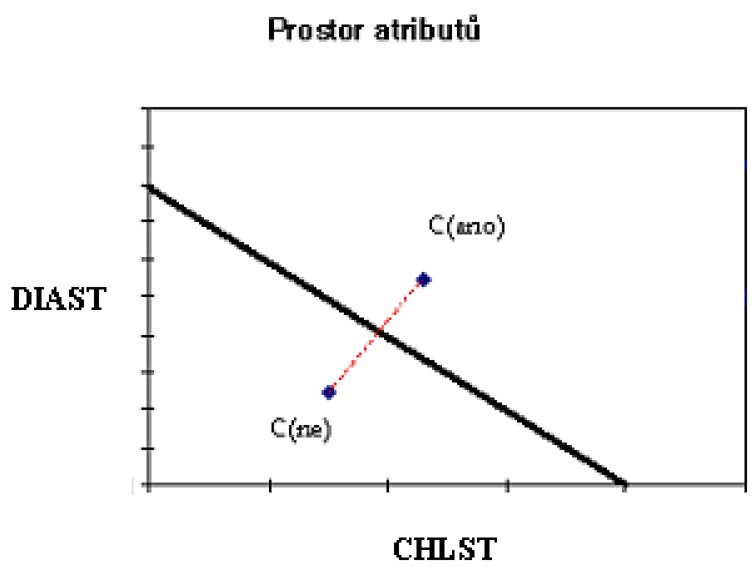
   1.2. assign

   $$y = \hat{y}' \Leftrightarrow y' \text{ is the majority class of } \boldsymbol{x}_1, \ldots \boldsymbol{x}_K,$$

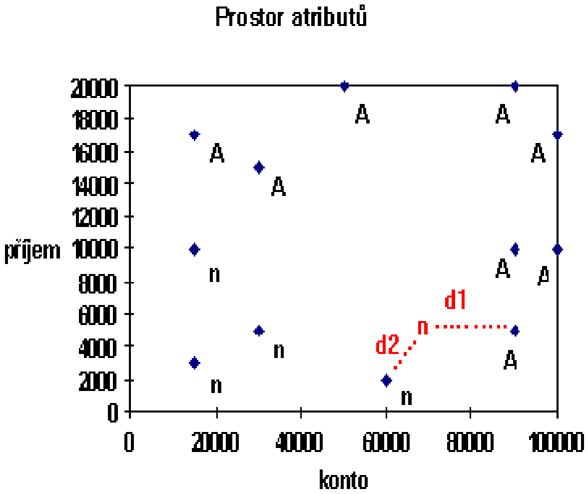
Nearest-neighbors in the attribute space

- Using examples
- Using centroids





Nearest-neighbor methods



- Selecting instances to be added
  - no search
    - IB1 (Aha)
  - simple heuristic top-down search
    - IB2, IB3 (Aha)
- clustering (identifying centroids)
  - simple heuristic top-down search
    - top-down (divisive)
    - bottom-up (aglomerative)
  - approximation
    - K-NN (given number of clusters)

Further readings

- T. Mitchell: Machine Learning. McGraw-Hill, 1997
- J. Han, M. Kerber: Data Mining, Concepts and Techniques. Morgan Kaufmann, 2001
- I. Witten, E. Frank: Data Mining, Practical Machine Learning tools and Techniques with Java. 2 edition. Morgan Kaufmann, 2005

- http://www.aaai.org/AITopics
- http://www.kdnuggets.com

Break

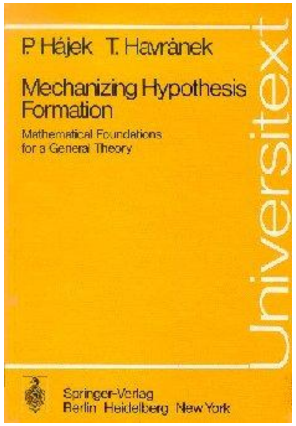
Part 3

GUHA Method and LISp-Miner System

GUHA Method and LISp-Miner System

Why here?

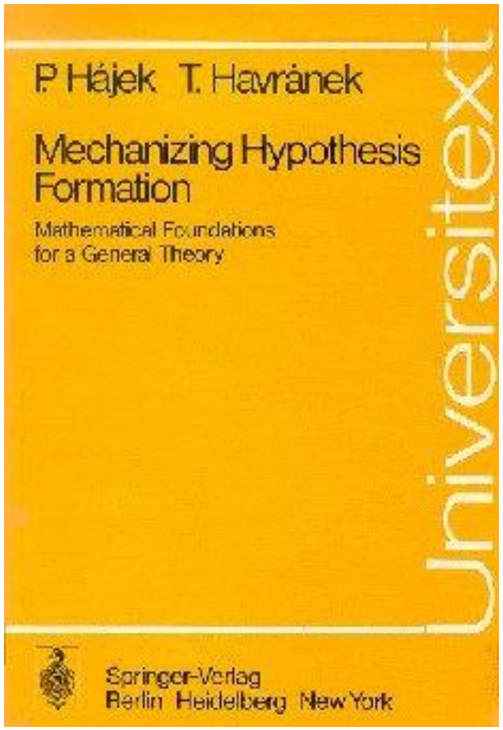
- Association rules coined by Agrawal in 1990’s
- More general rules studied since 1960’s
- GUHA method of mechanizing hypothesis formation
- Theory based on combination of
  - Mathematical logic
  - Mathematical statistics
- Several implementations
  - LISp-Miner system
- Relevant tools and theory



Outline

- GUHA – main features
- Association rule – couple of Boolean attributes
- GUHA procedure ASSOC
- LISp-Miner system
- Related research

GUHA – main features



1978

Starting questions:

Can computers formulate and verify scientific hypotheses?

Can computers in a rational way analyse empirical data and produce reasonable reflection of the observed empirical world? Can it be done using mathematical logic and statistics?

Examples of hypothesis formation

(1) This crow is black.
That crow is black.

All observed crows are black.

All crows are black.

(2) This crow is black.
That crow is black.

Many crows have been observed; relative frequency of black ones is high.

Crows have a considerable change of being black.

(3)

rat no.	weight g	weight of the kidney mg
1	362	1432
2	372	1601
3	376	1436
4	407	1633
5	411	2262

The observed weights of the kidneys have the same order as the weights of the rats with one exception.

The weight of rat's kidney is positively dependent on the weight of the rat.

Evidence

Observational statement

Theoretical statement

(1): Theoretical statement $\Rightarrow$ observational statement

(2), (3) : Theoretical statement ??? observational statement

From an observational statement to a theoretical statement

Evidence

Observational statement

Theoretical statement

(1): Theoretical statement $\Rightarrow$ observational statement

(2), (3) : Theoretical statement ??? observational statement

- Justified by some *rules of rational inductive inference*
- Some philosophers reject any possibility of formulating such rules
- Nobody believes that there can be universal rules
- There are non-trivial rules of inductive inference applicable under some well described circumstances
- Some of them are useful in mechanized inductive inference

Scheme of inductive inference: $\dfrac{\text{theoretical assumptions, observational statement}}{\text{theoretical statement}}$

Logic of discovery

Five questions:

Scheme of inductive inference: $\dfrac{\text{theoretical assumptions, observational statement}}{\text{theoretical statement}}$

L0: In what languages does one formulate observational and theoretical statements? (What is the syntax and semantics of these languages? What is their relation to the classical first order predicate calculus?)

L1: What are rational inductive inference rules bridging the gap between observational and theoretical sentences? (What does it mean that a theoretical statement is justified?)

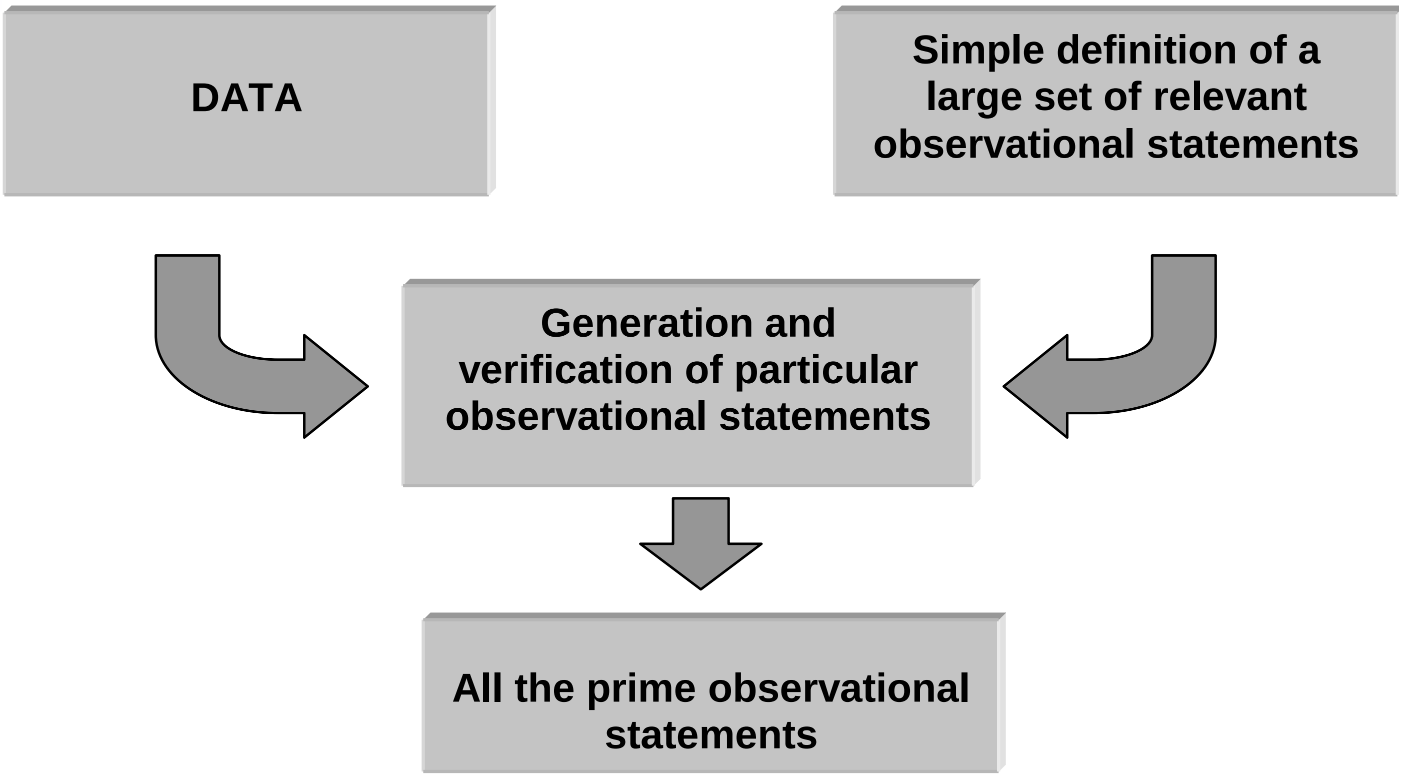
L2: Are there rational methods for deciding whether a theoretical statement is justified (on the basis of given theoretical assumptions and observational statements)?

L3: What are the conditions for a theoretical statement or a set of theoretical statements to be of interest (importance) with respect to the task of scientific cognition?

L4: Are there methods for suggesting such a set of statements, which is as interesting, as possible?

L0 – L2: Logic of induction L3 – L4: Logic of suggestion L0 – L4: Logic of discovery

GUHA Procedure



Observational : Theoretical statement = 1:1

Outline

- GUHA – main features
- Association rule – couple of Boolean attributes
  - Data matrix and Boolean attributes
  - Association rule
  - 4ft-quantifiers
- GUHA procedure ASSOC
- LISp-Miner
- Related research

Data matrix and Boolean attributes

A_1	A_2	…	A_m	$A_1(3)$	$A_2(7,9)$	$A_1(3) \wedge A_2(7,9)$	…
3	9	…	6	1	1	1	…
7	5	…	7	0	0	0	…
…	…	…	…	…	…	…	…
4	7	…	5	0	1	0	…

Data matrix $\mathcal{M}$

Boolean attributes φ, ψ, χ

Association rule

antecedent

succedent

4ft quantifier

$$\varphi \approx \psi$$

$\mathcal{M}$	ψ	$\neg\psi$
φ	a	b
$\neg\,\varphi$	c	d

$$F_{\approx}(a,b,c,d) = \begin{cases} 1 \ldots \varphi \approx \psi \text{ is true in } \mathcal{M} \\ 0 \ldots \varphi \approx \psi \text{ is false in } \mathcal{M} \end{cases}$$

Important simple 4ft-quantifiers (1)

$\mathcal{M}$	ψ	$\neg\psi$
φ	a	b
$\neg\varphi$	c	d

Founded implication: $\varphi \Rightarrow_{p,Base} \psi$ $\quad \dfrac{a}{a+b} \geq p \wedge a \geq Base$

Double founded implication: $\varphi \Leftrightarrow_{p,Base} \psi$ $\quad \dfrac{a}{a+b+c} \geq p \wedge a \geq Base$

Founded equivalence: $\varphi \equiv_{p,Base} \psi$ $\quad \dfrac{a+d}{a+b+c+d} \geq p \wedge a \geq Base$

Important simple 4ft-quantifiers (2)

$\mathcal{M}$	ψ	$\neg\psi$
φ	a	b
$\neg\varphi$	c	d

Above Average: $\varphi \Rightarrow^{+}{}_{p,Base}\ \psi$ $\quad \dfrac{a}{a+b} \geq (1+p)\dfrac{a+c}{a+b+c+d} \wedge a \geq Base$

„Classical": $\varphi \rightarrow_{C,S}\ \psi$ $\quad \dfrac{a}{a+b} \geq C \wedge \dfrac{a}{a+b+c+d} \geq S$

4ft-quantifiers – statistical hypothesis tests (1)

$\mathcal{M}$	ψ	$\neg\psi$
φ	a	b
$\neg\varphi$	c	d

Lower critical implication for $0 < p \leq 1$, $0 < \alpha < 0.5$

$$\varphi \Rightarrow^{!}_{p,\alpha,Base} \psi$$

$$\sum_{i=a}^{a+b} \binom{a+b}{i} p^i (1-p)^{a+b-i} \leq \alpha \wedge a \geq Base$$

The rule $\varphi \Rightarrow^{!}_{p;\alpha} \psi$ corresponds to the statistical test (on the level α) of the null hypothesis H_0: $P(\psi \mid \varphi) \leq p$ against the alternative one H_1: $P(\psi \mid \varphi) > p$. Here $P(\psi \mid \varphi)$ is the conditional probability of the validity of ψ under the condition φ.

4ft-quantifiers – statistical hypothesis tests (2)

$\mathcal{M}$	ψ	$\neg\psi$
φ	a	b
$\neg\varphi$	c	d

Fisher's quantifier for $0 < \alpha < 0.5$

$$\varphi \sim_{\alpha,Base} \psi$$

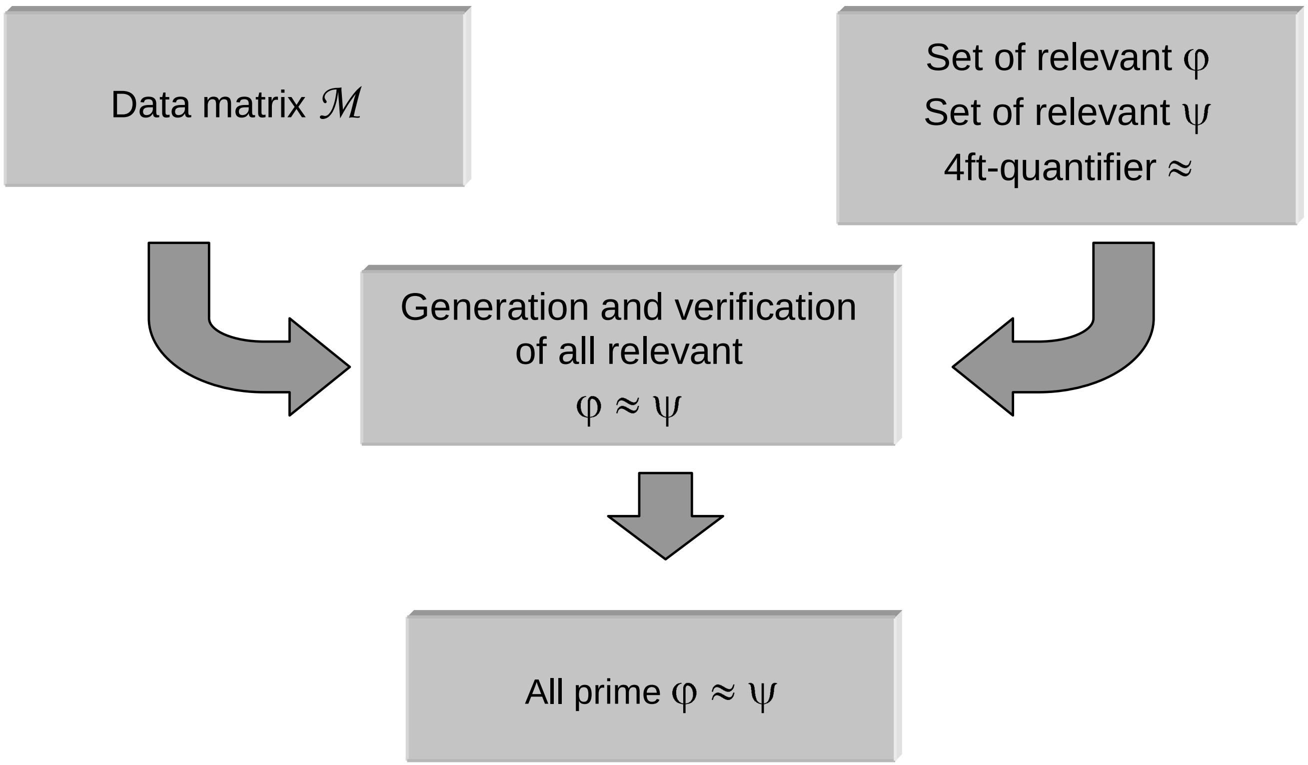
$$\sum_{i=a}^{\min(r,k)} \frac{\binom{k}{i}\binom{n-k}{r-i}}{\binom{r}{n}} \leq \alpha \ \wedge \ ad > bc \ \wedge \ a \geq Base$$

The rule $\varphi \sim_{\alpha,Base} \psi$ corresponds to the statistical test (on the level α of the null hypothesis of independence of φ and ψ against the alternative one of the positive dependence.

Outline

- GUHA – main features
- Association rule – couple of Boolean attributes
- GUHA procedure ASSOC
- LISp-Miner
- Related research

GUHA procedure ASSOC



GUHA – selected implementations (1)

- 1966 - MINSK 22 (I. Havel)
  Boolean data matrix
  simplified version
  association rules
  punch tape

- end of 1960s - IBM 7040 (I. Havel)

- 1976 IBM 370 (I. Havel, J. Rauch)
  Boolean data matrix
  association rules
  statistical quantifiers
  bit strings
  punch cards

GUHA – selected implementations (2)

- Early 1990s – PC-GUHA
  MS DOS
  A. Sochorová, P. Hájek, J. Rauch

- Since 1995 GUHA+-
  Windows
  D. Coufal + all.

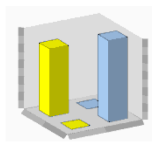
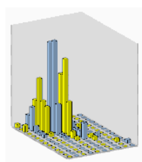
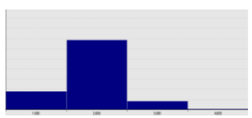
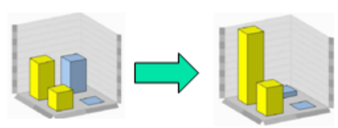
- Since 1996 LISp-Miner
  Windows
  M. Šimůnek + J. Rauch + all.
  7 GUHA procedures
  KEX
  related research

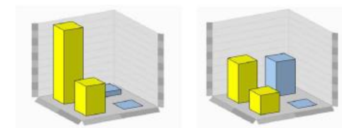
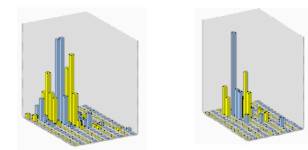
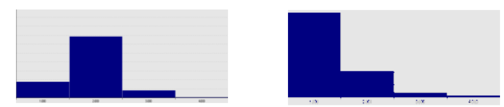
- Since 2006 Ferda, M. Ralbovský + all.

Outline

- GUHA – main features
- Association rule – couple of Boolean attributes
- GUHA procedure ASSOC
- LISp-Miner
  - Overview
  - Application examples
- Related research

http://lispminer.vse.cz LISp-Miner overview

- 4ft-Miner 
- KL-Miner 
- CF-Miner 
- 4ftAction-Miner 

- SD4ft-Miner 
- SDKL-Miner 
- SDCF-Miner 

i.e. 7 GUHA procedures

KEX

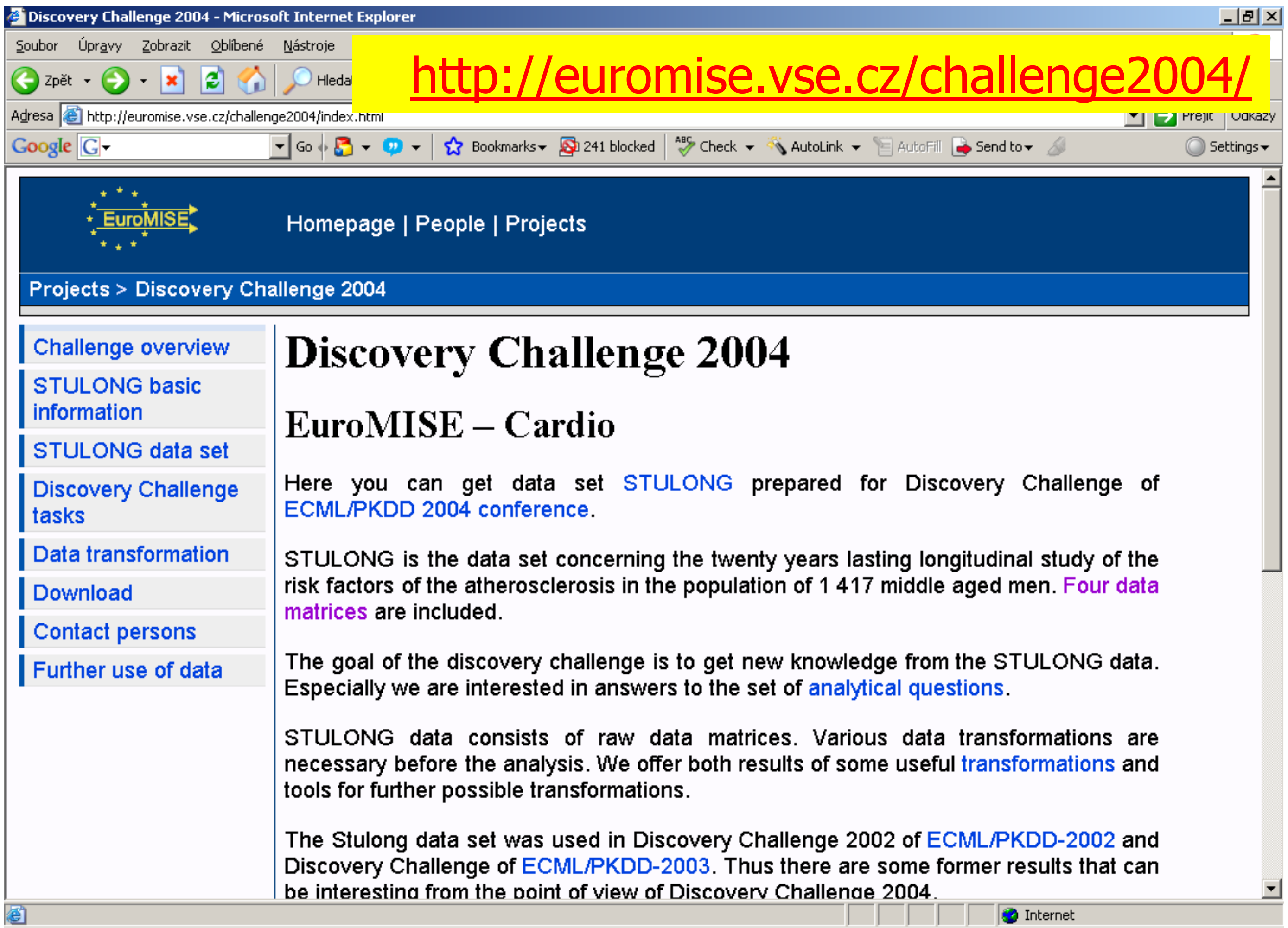
LMDataSource

LISp-Miner, application examples

- Stulong data set
- 4ft-Miner (enhanced ASSOC procedure):
  - $\mathcal{B}$(Physical, Social) $\approx^?$ $\mathcal{B}$(Biochemical)
- SD4ft-Miner:
  - normal $\otimes$ risk: $\mathcal{B}$(Physical, Social) $\approx^?$ $\mathcal{B}$(Biochemical)

Stulong data set (1)

http://euromise.vse.cz/challenge2004/



Stulong data set (2)

http://euromise.vse.cz/challenge2004/data/entry/

- A survey of attributes
- Identification data
- Social characteristics
- Activities
- Smoking
- Alcohol
- Sugar, coffee, tea
- Personal anamnesis
- Questionnaire A_2
- Physical exam.
- Biochemical exam.
- Risk factors

Entry examination – a survey of attributes

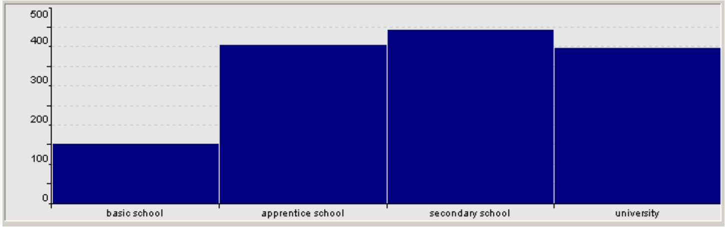
1 417 men have been examined during the entry examination. Values of 244 attributes have been surveyed with each patient. Values of 64 attributes are either codes or results of size measurements of different variables or results of transformations of the rest of the attributes. Values of all these 64 attributes are stored in the data matrix Entry. Attributes can be divided into groups according to the Table 1.

Table 1: Groups of the attributes in the entry examination

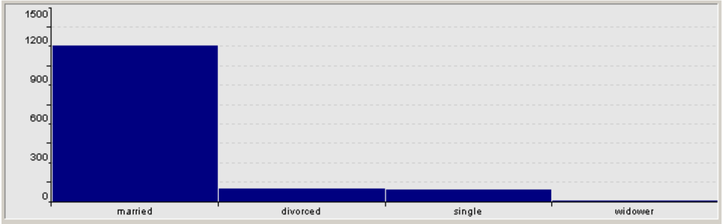
Groups of attributes	Number of attributes
identification data	2
social characteristics	6
physical activity	4
smoking	3
drinking of alcohol	9
sugar, coffee, tea	3
personal anamnesis	18
questionnaire A_2	3
physical examination	8
biochemical examination	3
risk faktors	5

Social characteristcs

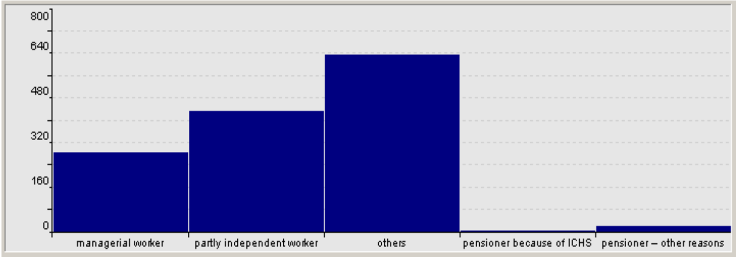
Education



Marital status

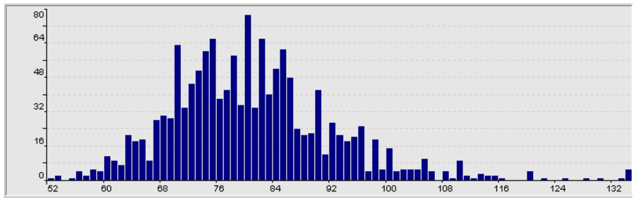


Responsibility in a job

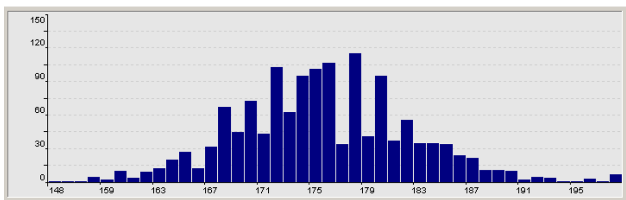


Physical examinations

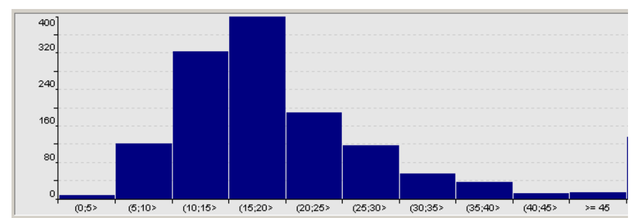
Weight [kg]



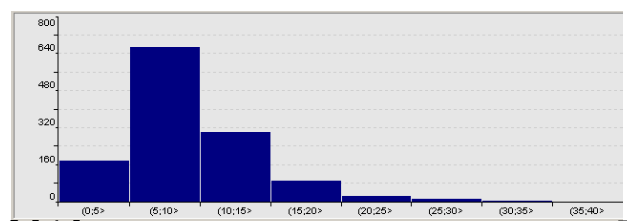
Height [cm]



Skinfold above musculus triceps [mm]



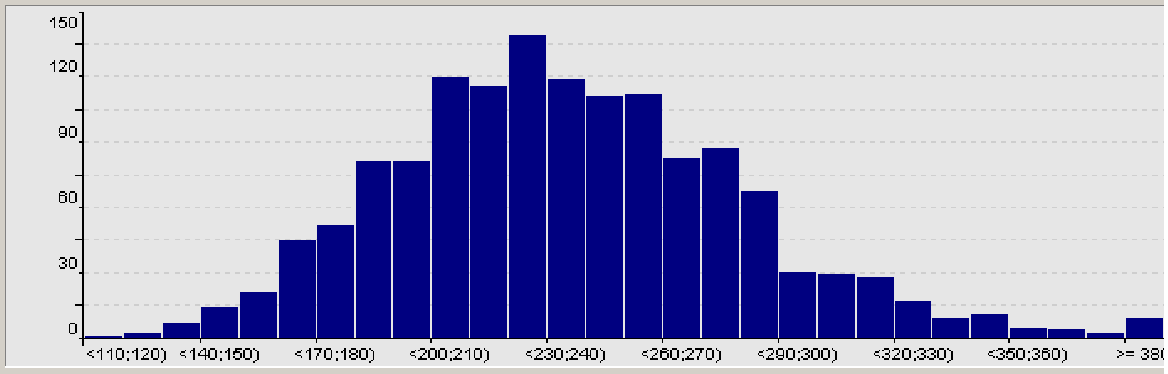
Skinfold above musculus subscapularis [mm]



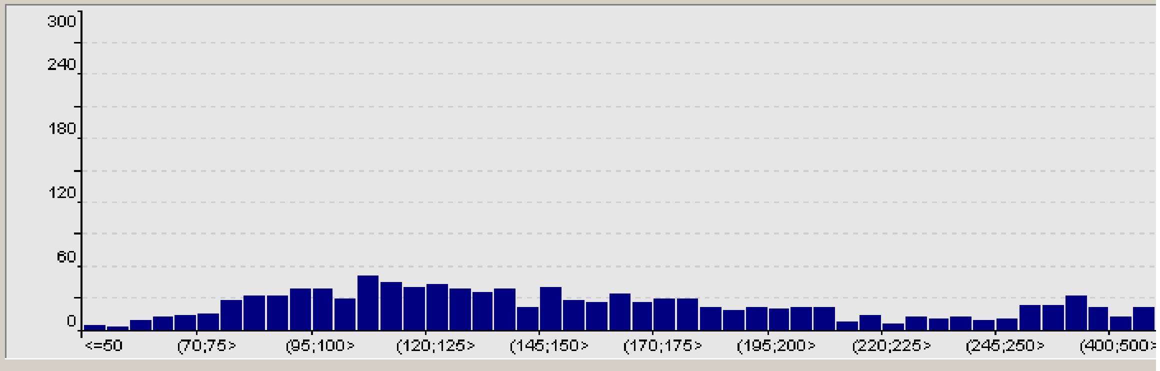
…… additional attributes

Biochemical examinations

Cholesterol [mg%]



Triglycerides in mg%



LISp-Miner, application examples

- Stulong data set
- 4ft-Miner (enhanced ASSOC procedure):
  - $\mathcal{B}$(Physical, Social) $\approx^?$ $\mathcal{B}$(Biochemical)
- SD4ft-Miner:
  - normal $\otimes$ risk: $\mathcal{B}$(Physical, Social) $\approx^?$ $\mathcal{B}$(Biochemical)

$\mathcal{B}$ (Physical, Social) $\approx^?$ $\mathcal{B}$ (Biochemical)

In the ENTRY data matrix,

are there some interesting relations between Boolean attributes describing combination of results of Physical examination and Social characteristics and results of Biochemical examination?

$$\varphi \approx^? \psi$$

$\varphi \in \mathcal{B}$ (Physical, Social)

$\psi \in \mathcal{B}$ (Biochemical)

$\approx^?$ evaluated using 4-fould table

ENTRY	ψ	$\neg\psi$
φ	a	b
$\neg\varphi$	c	d

Applying GUHA procedure 4ft-Miner

$\mathcal{B}$ (Physical, Social) $\approx^{?}$ $\mathcal{B}$ (Biochemical)

$\varphi \approx^{?} \psi$

$\varphi \in \mathcal{B}$ (Physical, Social)

$\psi \in \mathcal{B}$ (Biochemical)

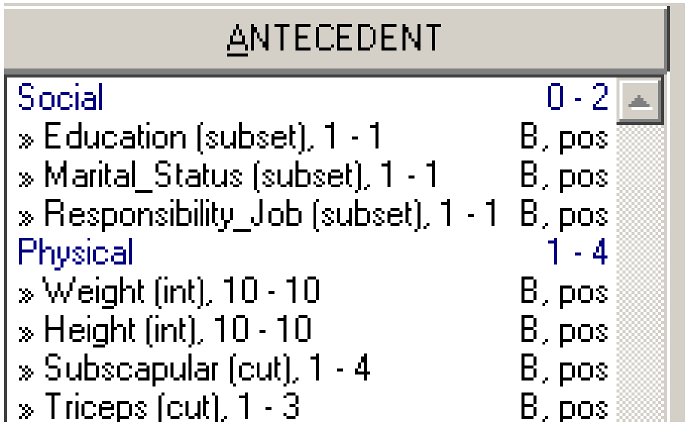
Entry data matrix

$\mathcal{B}$ (Physical, Social)
$\mathcal{B}$ (Biochemical)
$\approx^{?}$

Generation and verification of
$\varphi \approx^{?} \psi$

All prime $\varphi \approx^{?} \psi$

Defining $\mathcal{B}$(Social, Physical) (1)

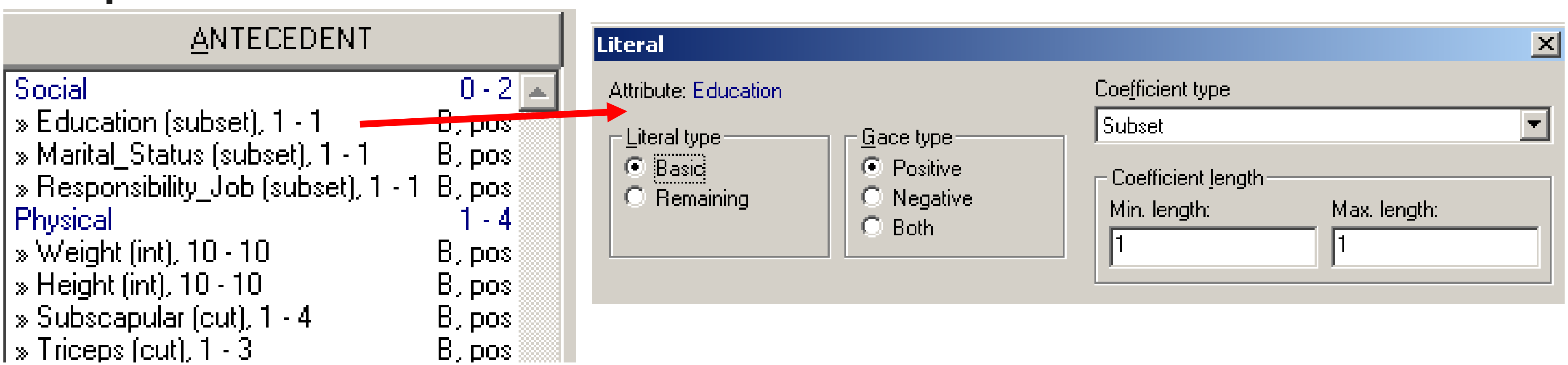


$$\mathcal{B}(\text{Social, Physical}) = \mathcal{B}(\text{Social}) \wedge \mathcal{B}(\text{Physical})$$

$$\mathcal{B}(\text{Social}) = \bigwedge_0^2 [\mathcal{B}(\text{Education}), \mathcal{B}(\text{Marital Status}), \mathcal{B}(\text{Responsibility\_Job})]$$

$$\mathcal{B}(\text{Physical}) = \bigwedge_1^4 [\mathcal{B}(\text{Weight}), \mathcal{B}(\text{Height}), \mathcal{B}(\text{Subscapular}), \mathcal{B}(\text{Triceps})]$$

Defining $\mathcal{B}$ (Social, Physical) (2)



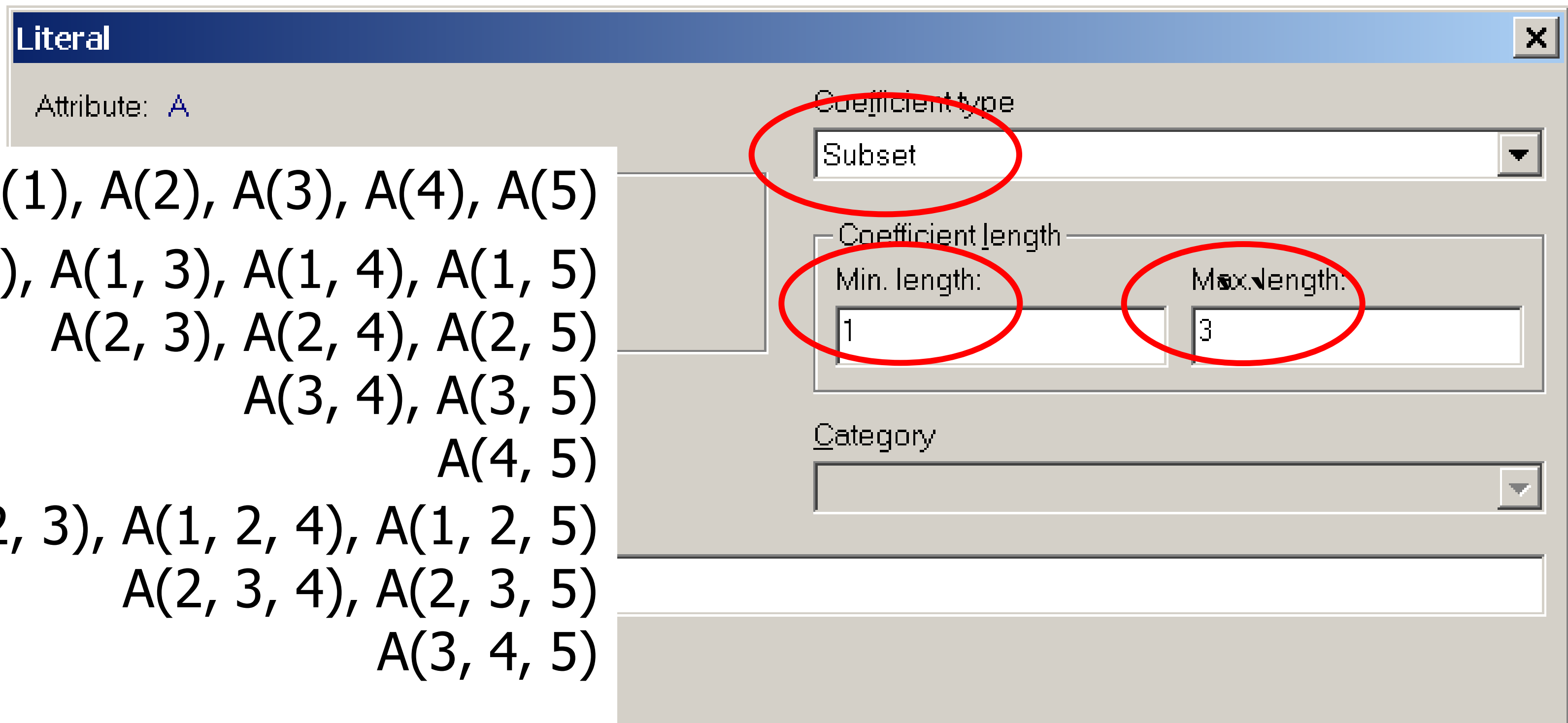
Education: basic school, apprentice school, secondary school, university

$\mathcal{B}$ (Education): Subsets of length 1 - 1

Education (basic school), Education (apprentice school)

Education (secondary school), Education (university)

Note: Attribute A with categories 1, 2, 3, 4, 5 Literals with coefficients Subset (1 – 3):

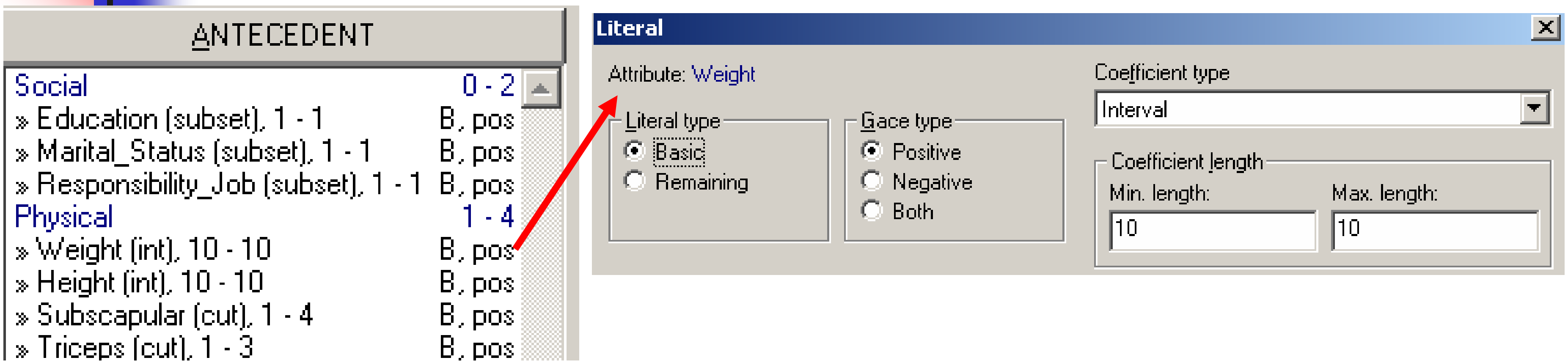


A(1), A(2), A(3), A(4), A(5)

A(1, 2), A(1, 3), A(1, 4), A(1, 5)
A(2, 3), A(2, 4), A(2, 5)
A(3, 4), A(3, 5)
A(4, 5)

A(1, 2, 3), A(1, 2, 4), A(1, 2, 5)
A(2, 3, 4), A(2, 3, 5)
A(3, 4, 5)

Defining $\mathcal{B}$ (Social, Physical) (3)



Set of categories of Weight: 52, 53, 54, 55, ……., 130, 131, 132, 133

$\mathcal{B}$ (Weight): Intervals of length 10 - 10: Weight(52 – 61), Weight(53 – 62), …

52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63,…., 128, 129, 130, 131, 132, 133

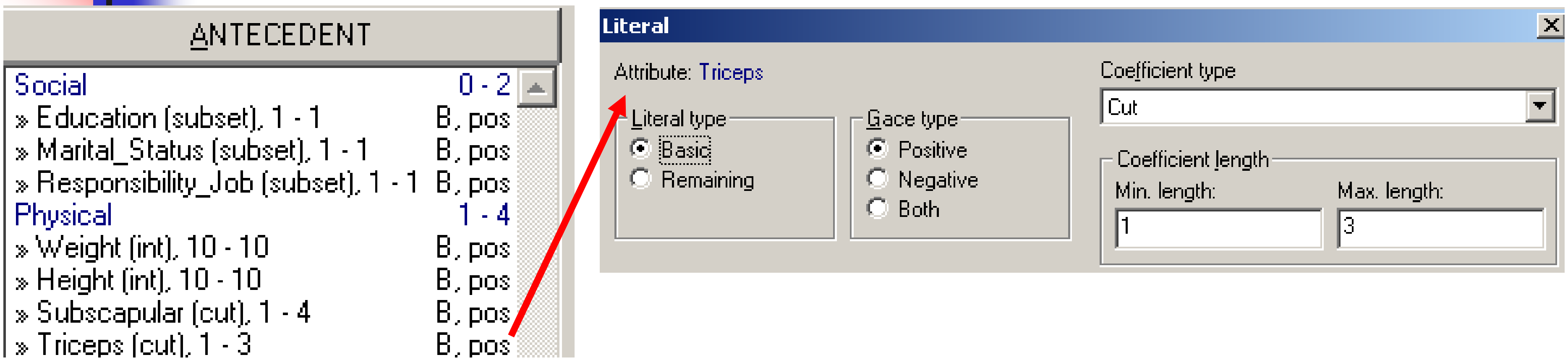
52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63,…., 128, 129, 130, 131, 132, 133

52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63,…., 128, 129, 130, 131, 132, 133

,….,

52, 53, 54, 55, 56, …., 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133

Defining $\mathcal{B}$ (Social, Physical) (4)



Set of categories of Triceps: (0;5〉, (5;10〉, (10;15〉, ..., (25;30〉 (30;35〉 (35;40〉

$\mathcal{B}$ (Triceps): Cuts 1 - 3

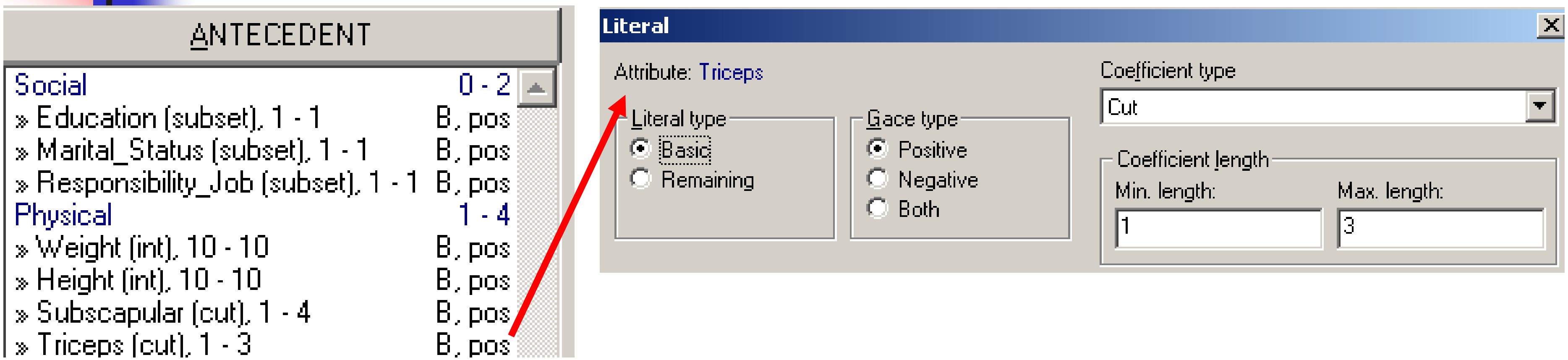
Left cuts 1 – 3 **i.e. Triceps(low)**

(0;5〉, (5;10〉, (10;15〉, (15;20〉, (20;25〉, (25;30〉, (30;35〉, (35;40 〉 i.e. Triceps(1 – 5)

(0;5〉, (5;10〉, (10;15〉, (15;20〉, (20;25〉, (25;30〉, (30;35〉, (35;40 〉 i.e. Triceps(1 – 10)

(0;5〉, (5;10〉, (10;15〉, (15;20〉, (20;25〉, (25;30〉, (30;35〉, (35;40 〉 i.e. Triceps(1 – 15)

Defining $\mathcal{B}$ (Social, Physical) (5)



Set of categories of Triceps: (0;5〉, (5;10〉, (10;15〉, …, (25;30〉 (30;35〉 (35;40〉

$\mathcal{B}$ (Triceps): Cuts 1 - 3

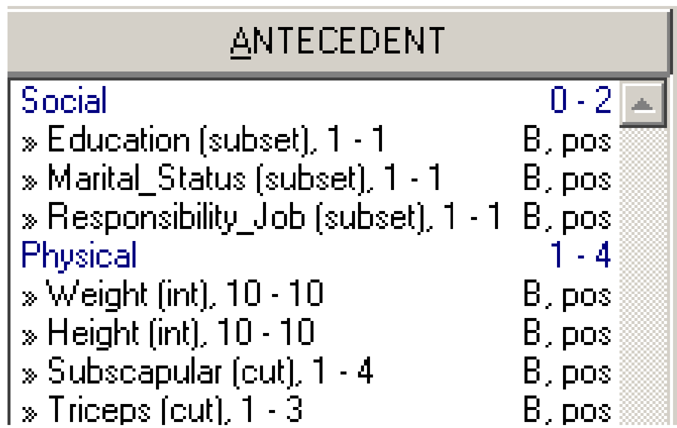
Right cuts 1 – 3 **i.e. Triceps(high)**

(0;5〉, (5;10〉, (10;15〉, (15;20〉, (20;25〉, (25;30〉, (30;35〉, (35;40 〉 i.e. Triceps(35 – 40)

(0;5〉, (5;10〉, (10;15〉, (15;20〉, (20;25〉, (25;30〉, (30;35〉, (35;40 〉 i.e. Triceps(30 – 40)

(0;5〉, (5;10〉, (10;15〉, (15;20〉, (20;25〉, (25;30〉, (30;35〉, (35;40 〉 i.e. Triceps(25 – 45)

Defining $\mathcal{B}$(Social, Physical) (6)



Examples of $\varphi \in \mathcal{B}$(Social, Physical):

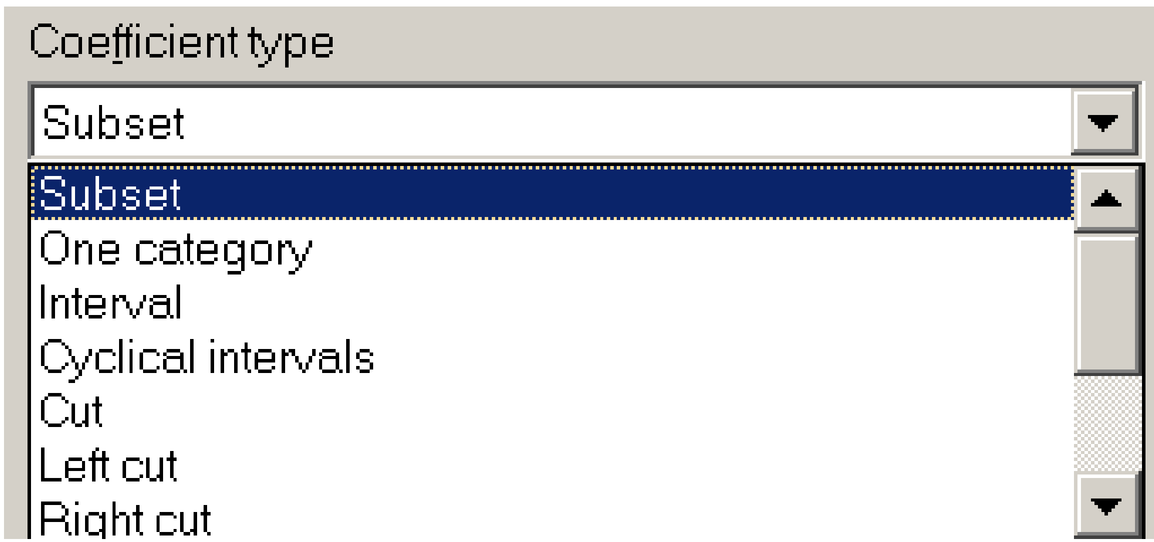
Education (basic school)

Education (university) $\wedge$ Marital_Status(single) $\wedge$ Weight (52 – 61)

Marital_Status(divorced) $\wedge$ Weight (52 – 61) $\wedge$ Triceps (25 – 45)

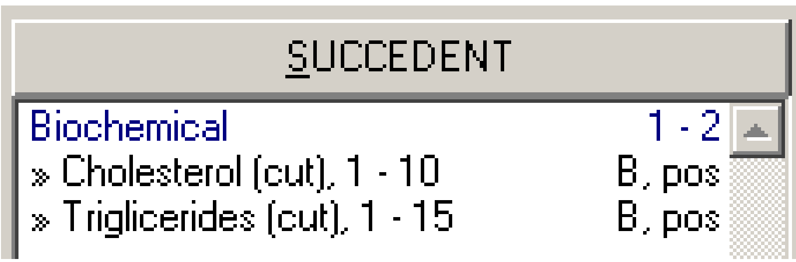
Weight (52 – 61) $\wedge$ Height (52 – 61) $\wedge$ Subscapular(0 – 10) $\wedge$ Triceps (25 – 45)

Note: Types of coefficients



See examples above

Defining $\mathcal{B}$ (Biochemical)



Analogously to $\mathcal{B}$ (Social, Physical)

Examples of $\psi \in \mathcal{B}$(Biochemical):

Cholesterol (110 – 120), Cholesterol (110 – 130), …, Cholesterol (110 – 210)

Cholesterol ($\geq$ 380), Cholesterol ($\geq$ 370), …, Cholesterol ($\geq$ 290)

Cholesterol ($\geq$ 380) $\wedge$ Triglicerides ($\leq$ 50), …

Cholesterol ($\geq$ 380) $\wedge$ Triglicerides ($\leq$ 300), …

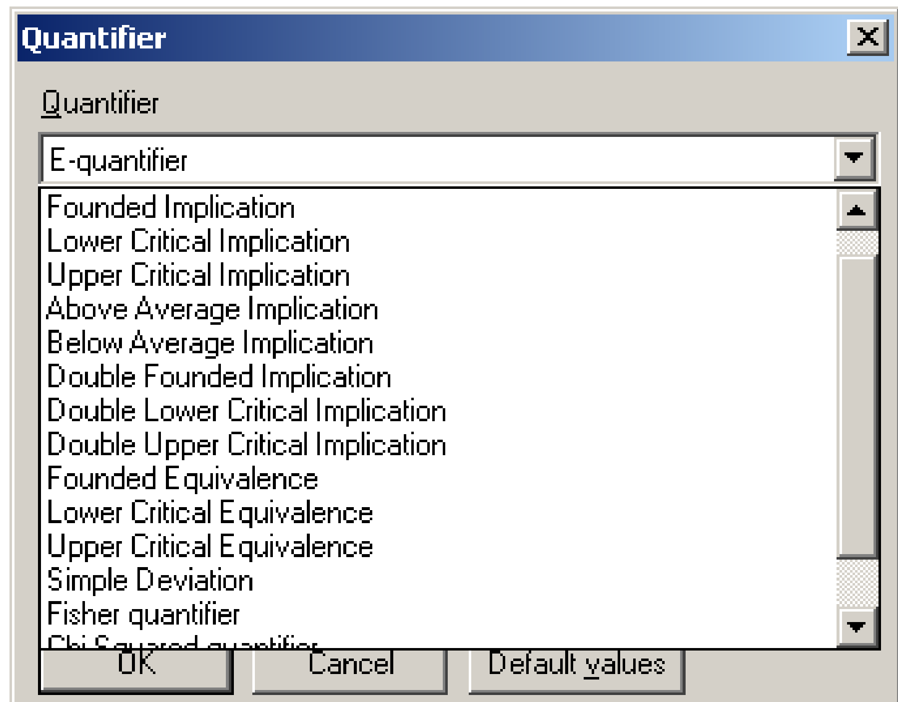
…,

Defining $\approx^?$ in $\varphi \approx^? \psi$

$\approx^?$ corresponds to a condition concerning 4ft(φ, ψ, $\mathcal{M}$)

$\mathcal{M}$	ψ	$\neg\psi$
φ	a	b
$\neg\varphi$	c	d

17 types of 4ft-quantifiers



Two examples of $\approx^?$

$\mathcal{M}$	ψ	$\neg\psi$
φ	a	b
$\neg\varphi$	c	d

Founded implication $\Rightarrow_{p,B}$ $\quad \dfrac{a}{a+b} \geq p \wedge a \geq B$

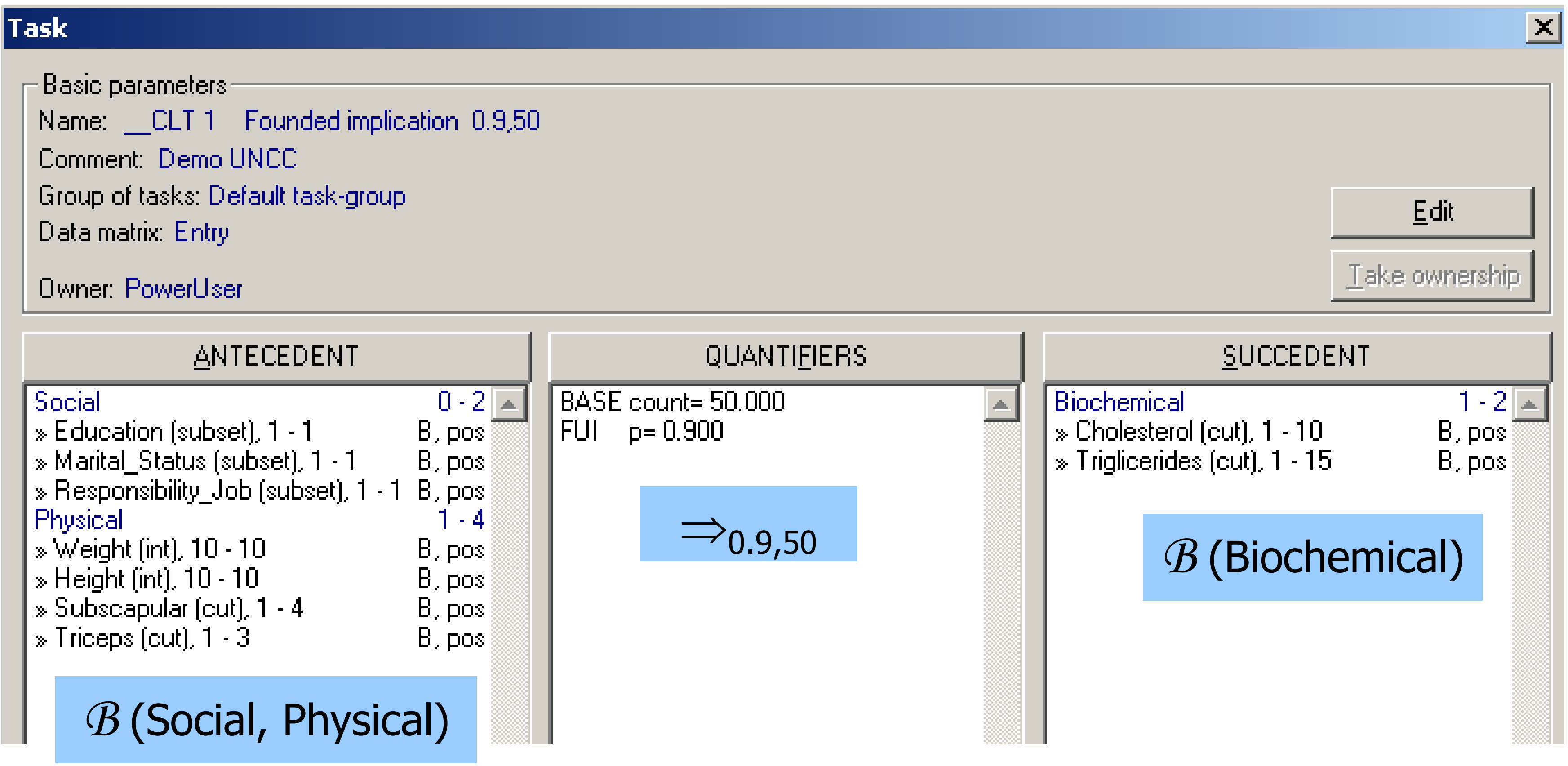
$\varphi \Rightarrow_{p,B} \psi$: at least 100p per cent of objects of $\mathcal{M}$

φsatisfying φ satisfy also ψ φ and there are at least Base objects satisfying both φ and ψ

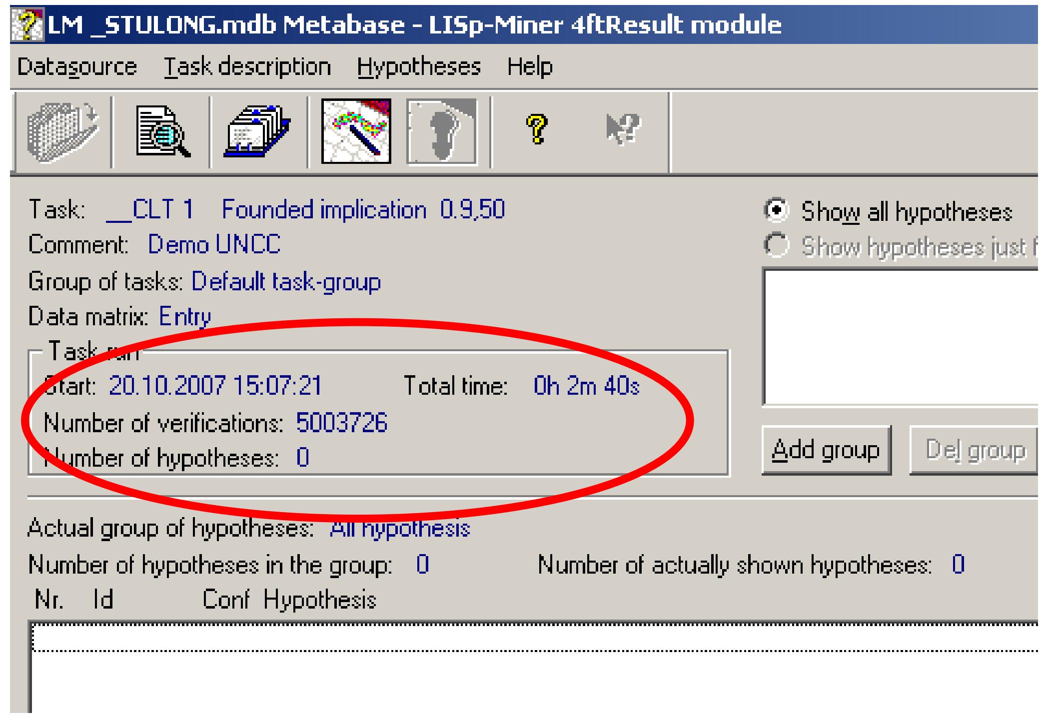
Above average $\Rightarrow^{+}_{p,B}$ $\quad \dfrac{a}{a+b} \geq (1+p) \wedge \dfrac{a+c}{a+b+c+d} \wedge a \geq B$

$\varphi \Rightarrow^{+}_{p,B} \psi$: the relative frequency of objects of M satisfying ψ among the objects satisfying φ is at least 100p per cent higher than the relative frequency of ψ in the whole data matrix $\mathcal{M}$ and there are at least Base objects satisfying both φ and ψ

Solving $\mathcal{B}$(Social, Physical) $\Rightarrow_{0.9,50}$ $\mathcal{B}$(Biochemical) (1)



Solving $\mathcal{B}$(Social, Physical) $\Rightarrow_{0.9,50}$ $\mathcal{B}$(Biochemical) (2)



PC with 1.66 GHz, 2 GB RAM

2 min. 40 sec.

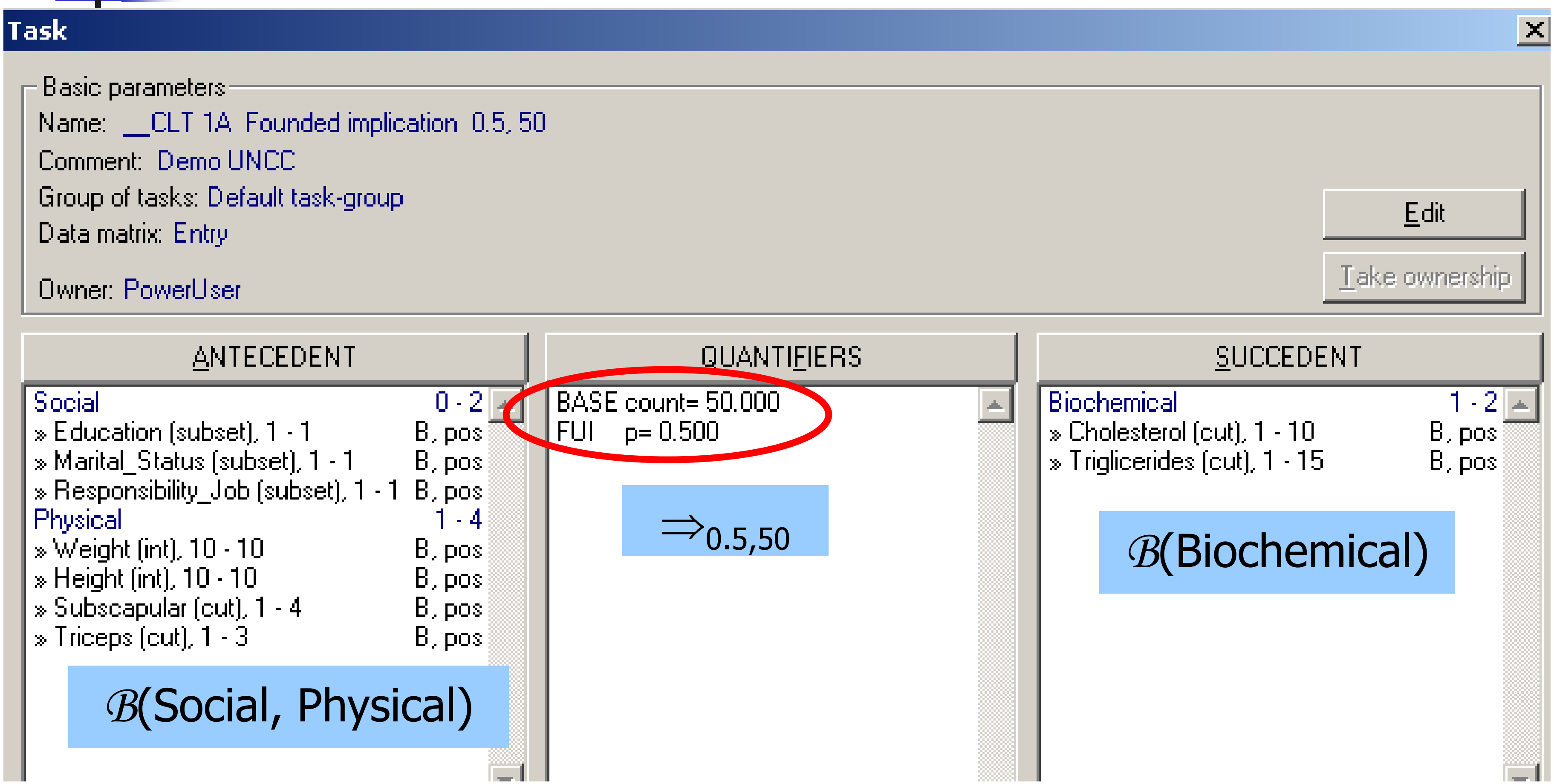
5 . 10^6 rules verified

0 true rules

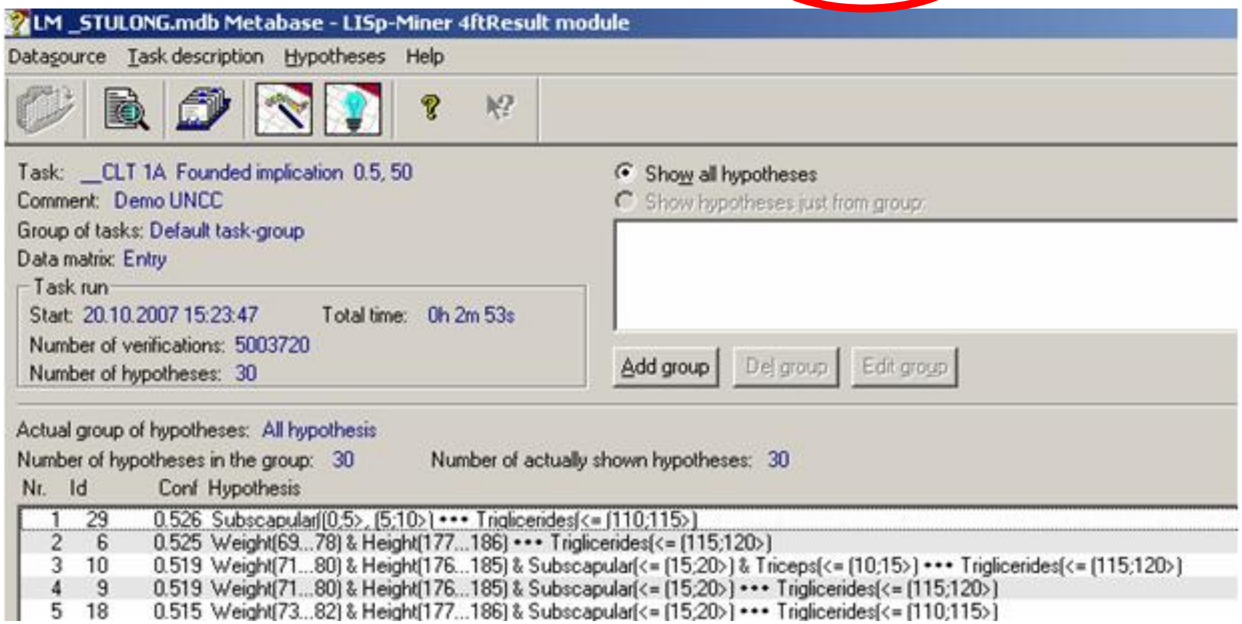
Problem: Confidence 0.9 in $\Rightarrow_{0.9,50}$ too high

Solution: Use confidence 0.5

Solving $\mathcal{B}$(Social, Physical) $\Rightarrow_{0.5,50}$ $\mathcal{B}$(Biochemical) (1)



Solving $\mathcal{B}$(Social, Physical) $\Rightarrow_{0.5,50}$ $\mathcal{B}$(Biochemical) (2)



30 rules with confidence $\geq$ 0.5

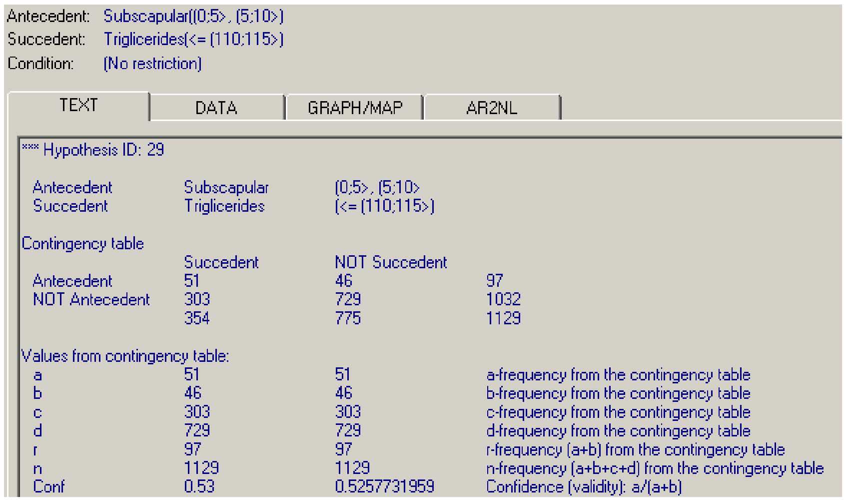
Problem: The strongest rule has confidence only 0.526, see detail

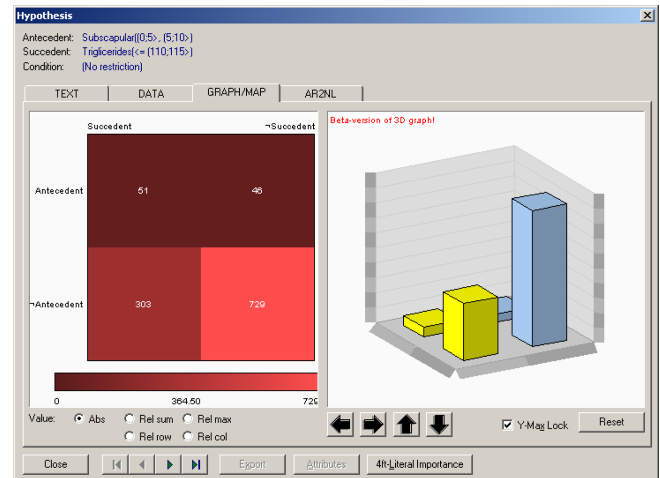
Solution: Search for rules expressing 70% higher relative frequency than average

It means to use $\Rightarrow^{+}_{0.7,50}$ instead of $\Rightarrow_{0.5,50}$

Solving $\mathcal{B}$(Social, Physical) $\Rightarrow_{0.5,50}$ $\mathcal{B}$(Biochemical) (3)

Detail of results - the strongest rule

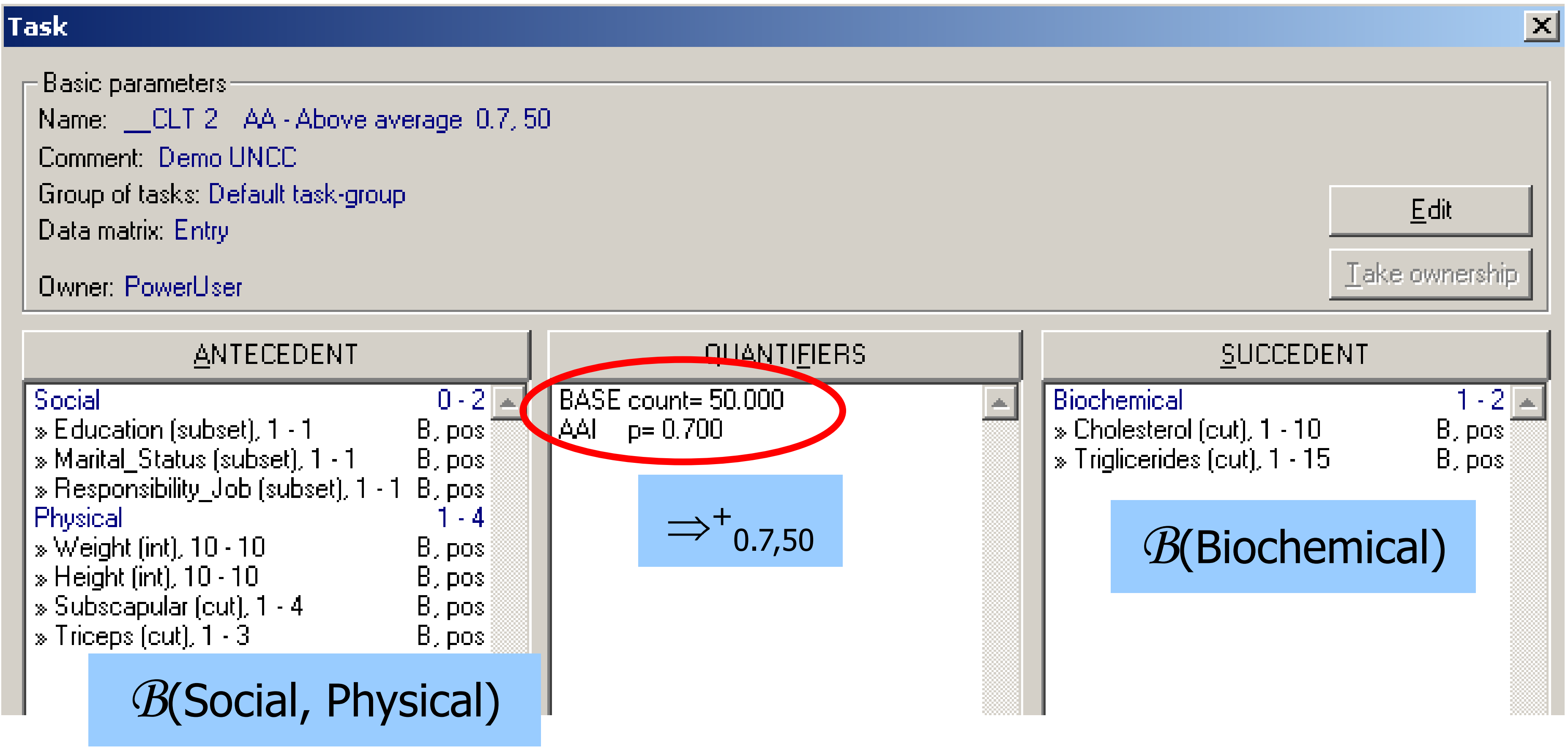




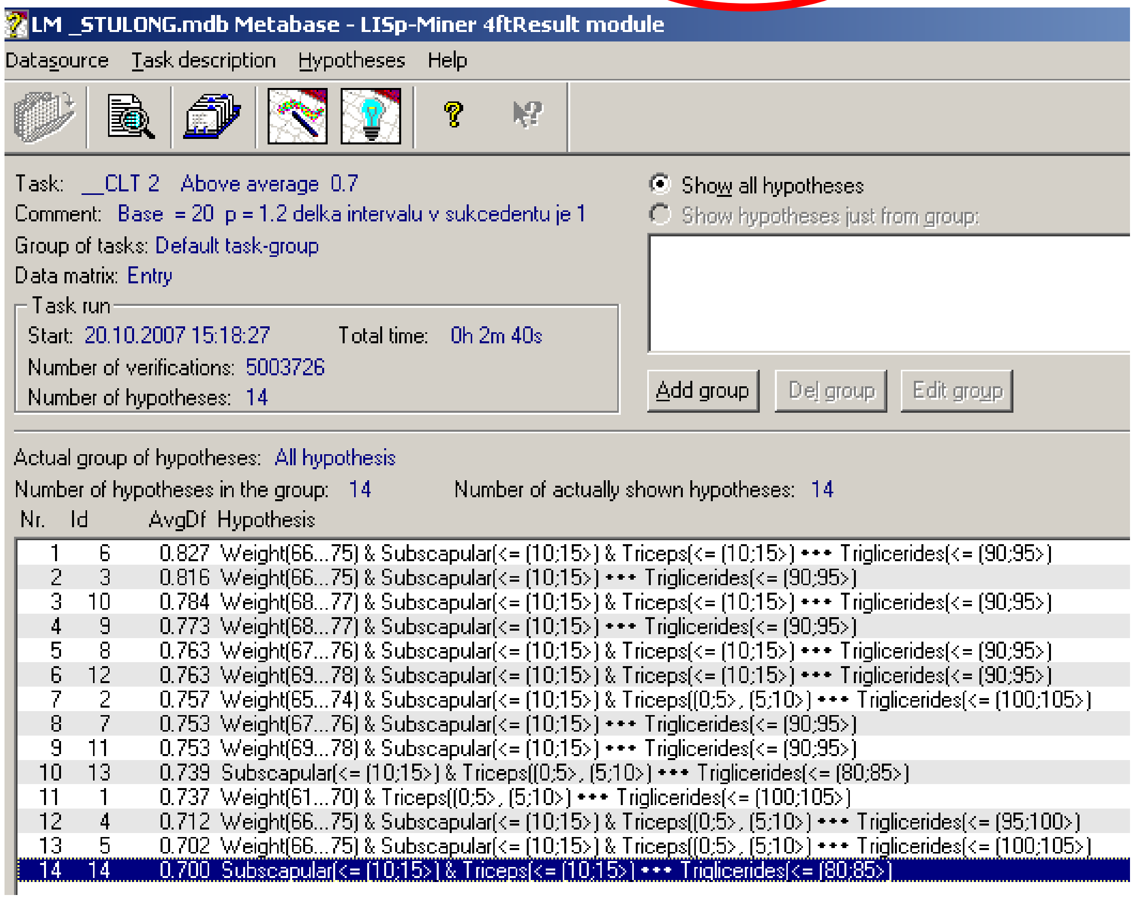
Entry	Triglicerides(≤115)	¬ Triglicerides(≤115)
Subscapular(0;10⟩	51	46
¬ Subscapular(0;10⟩	303	729

Subscapular(0;10⟩ $\Rightarrow_{0.53,\ 51}$ Triglicerides(≤115)

Solving $\mathcal{B}$(Social, Physical) $\Rightarrow^{+}_{0.7,50}$ $\mathcal{B}$(Biochemical) (1)



Solving $\mathcal{B}$(Social, Physical) $\Rightarrow^{+}_{0.7,50}$ $\mathcal{B}$(Biochemical) (2)



14 rules with relative frequency of succedent $\geq$ 0.7 than average, example – see detail

Solving $\mathcal{B}$(Social, Physical) $\Rightarrow^+_{0.7,50}$ $\mathcal{B}$(Biochemical) (3)

Detail of results - the strongest rule

φ: Weight (65;75⟩ ∧ Subscapular(≤15) ∧ Triceps(≤15)

ψ: Triglicerides (≤ 95)

confidence = 51 / 165 = 0.31 (not interesting!)

Entry	ψ	$\neg\psi$	
φ	51	114	165
$\neg\varphi$	140	824	964
	191	938	1129

relative frequency of patients satisfying ψ in the whole data matrix: $$\frac{51+140}{51+114+140+824} = 0.17$$

relative frequency of patients satisfying ψ among the patients satisfying φ: $$\frac{51}{51+114} = 0.31$$

i.e. 82 % higher

$$\frac{51}{51+114} = (1+0.82)\frac{51+140}{51+114+140+824}$$

thus $\varphi \Rightarrow^+_{0.82,51} \psi$

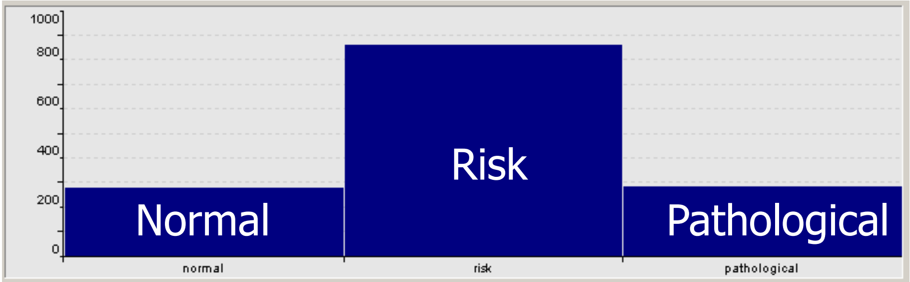
4ft-Miner, summary

- mines for rules $\varphi \approx \psi$ / χ and conditional rules $\varphi \approx \psi$ / χ
- very fine tools to define set of relevant φ, ψ, χ
- elements of semantics ..... Right cuts 1 – 3 i.e. Triceps(high
- measures of association $\approx$ on 4ft(φ, ψ, $\mathcal{M}$) = $\langle a, b, c, d \rangle$
- works very fast
- does not use Apriori, uses bit string approach

LISp-Miner, application examples

- Stulong data set
- 4ft-Miner (enhanced ASSOC procedure):
  - $\mathcal{B}$(Physical, Social) $\approx^?$ $\mathcal{B}$(Biochemical)
- SD4ft-Miner:
  - normal $\otimes$ risk: $\mathcal{B}$(Physical, Social) $\approx^?$ $\mathcal{B}$(Biochemical)

SD4ft-Miner Motivation



Is there any difference between normal and risk patients what concerns

$\mathcal{B}$ (Social, Physical) $\approx^?$ $\mathcal{B}$ (Biochemical)?

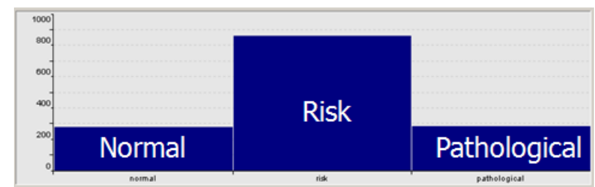
normal $\otimes$ risk: $\mathcal{B}$ (Social, Physical) $\approx^?$ $\mathcal{B}$ (Biochemical)

Normal $\otimes$ Risk: $\mathcal{B}$ (Social, Physical) $\approx^{?}$ $\mathcal{B}$ (Biochemical) (1)

Is there any difference between normal and risk what concerns $\varphi \Rightarrow_{p,\,B} \psi$?

normal	ψ	$\neg\psi$
φ	a_1	b_1
$\neg\varphi$	c_1	d_1

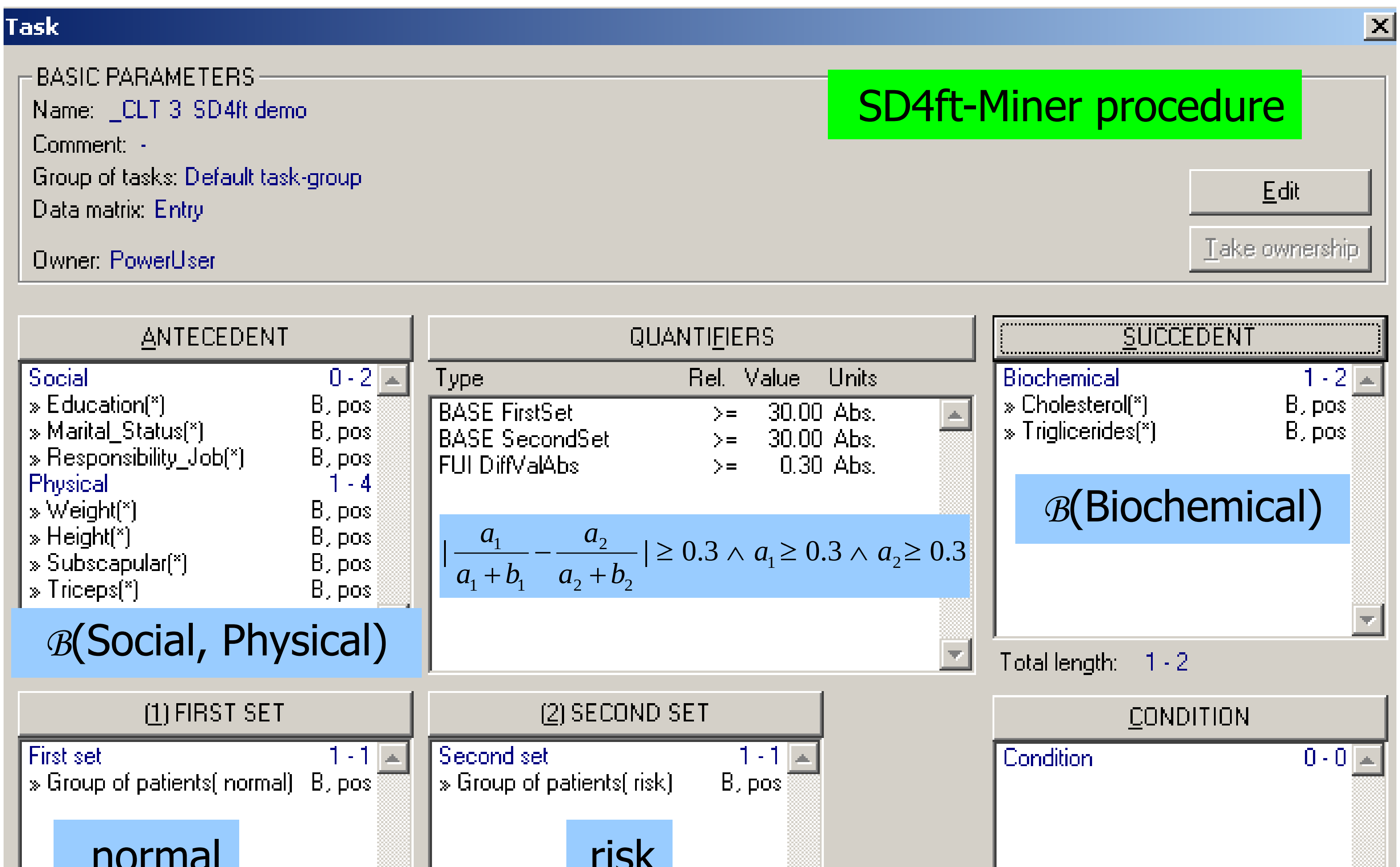
risk	ψ	$\neg\psi$
φ	a_2	b_2
$\neg\varphi$	c_2	d_2



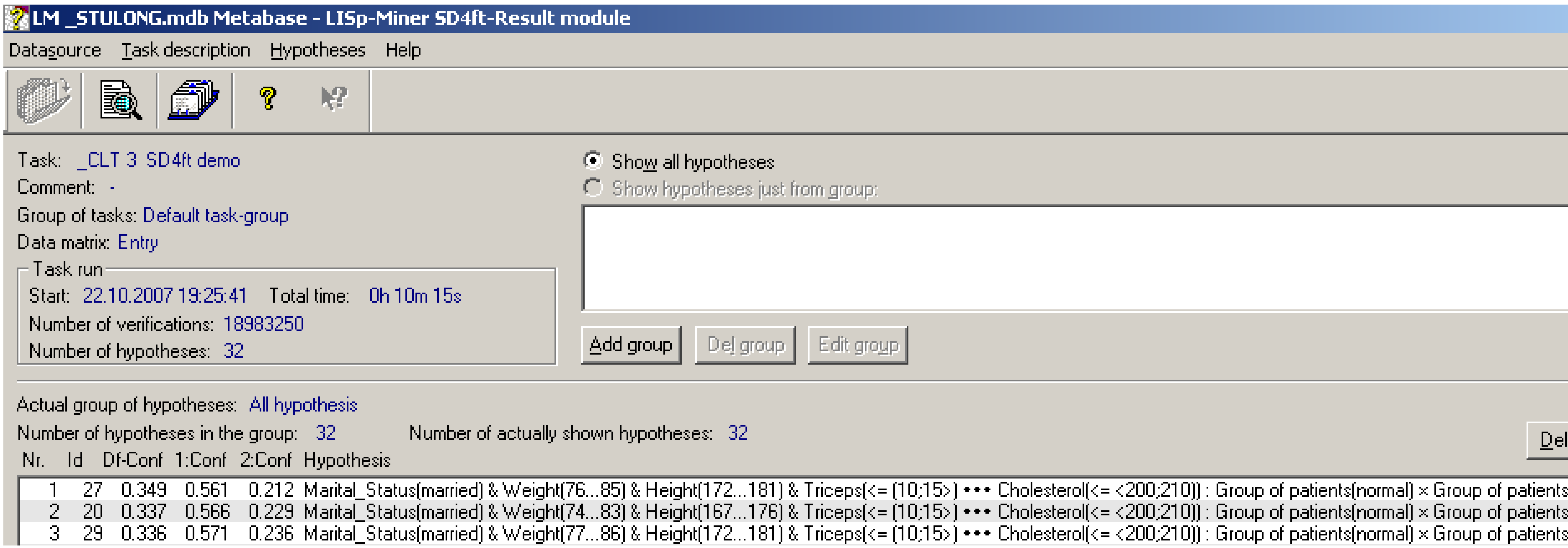
Example of difference: $|\text{confidence}_{\text{normal}} - \text{confidence}_{\text{risk}}| \geq 0.3$

Condition of interestingness: $$\left| \frac{a_1}{a_1 + b_1} - \frac{a_2}{a_2 + b_2} \right| \geq 0.3 \wedge a_1 \geq 30 \wedge a_2 \geq 30$$

Normal ⊗ Risk: $\mathcal{B}$ (Social, Physical) $\approx^?$ $\mathcal{B}$ (Biochemical) (2)



Normal $\otimes$ Risk: $\mathcal{B}$ (Social, Physical) $\approx^{?}$ $\mathcal{B}$ (Biochemical) (3)



19 000 000 patterns verified in 10 minutes

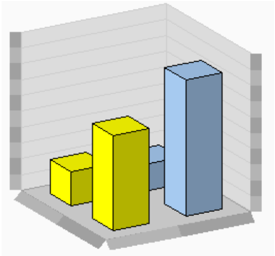
32 patterns found

The strongest one – see detail

normal ⊗ risk: $\mathcal{B}$ (Social, Physical) $\approx^?$ $\mathcal{B}$ (Biochemical) (4)

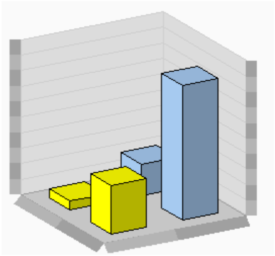
Detail of results - the strongest rule

Entry / normal	ψ	$\neg\psi$
φ	*32*	*25*
$\neg\varphi$	*90*	*129*



$\text{confidence}_{\text{normal}} = 0.56$

Entry / risk	ψ	$\neg\psi$
φ	*32*	*119*
$\neg\varphi$	*188*	*520*



$\text{confidence}_{\text{risk}} = 0.21$

φ: Marital_Status(married) $\wedge$ Weight $(75,85\rangle$ $\wedge$ Height $(172,181\rangle$ $\wedge$ Triceps($\leq$15)

ψ: Cholesterol ($\leq$ 210)

$\text{confidence}_{\text{normal}} - \text{confidence}_{\text{risk}} = 0.35$

SD4ft-Miner, Summary

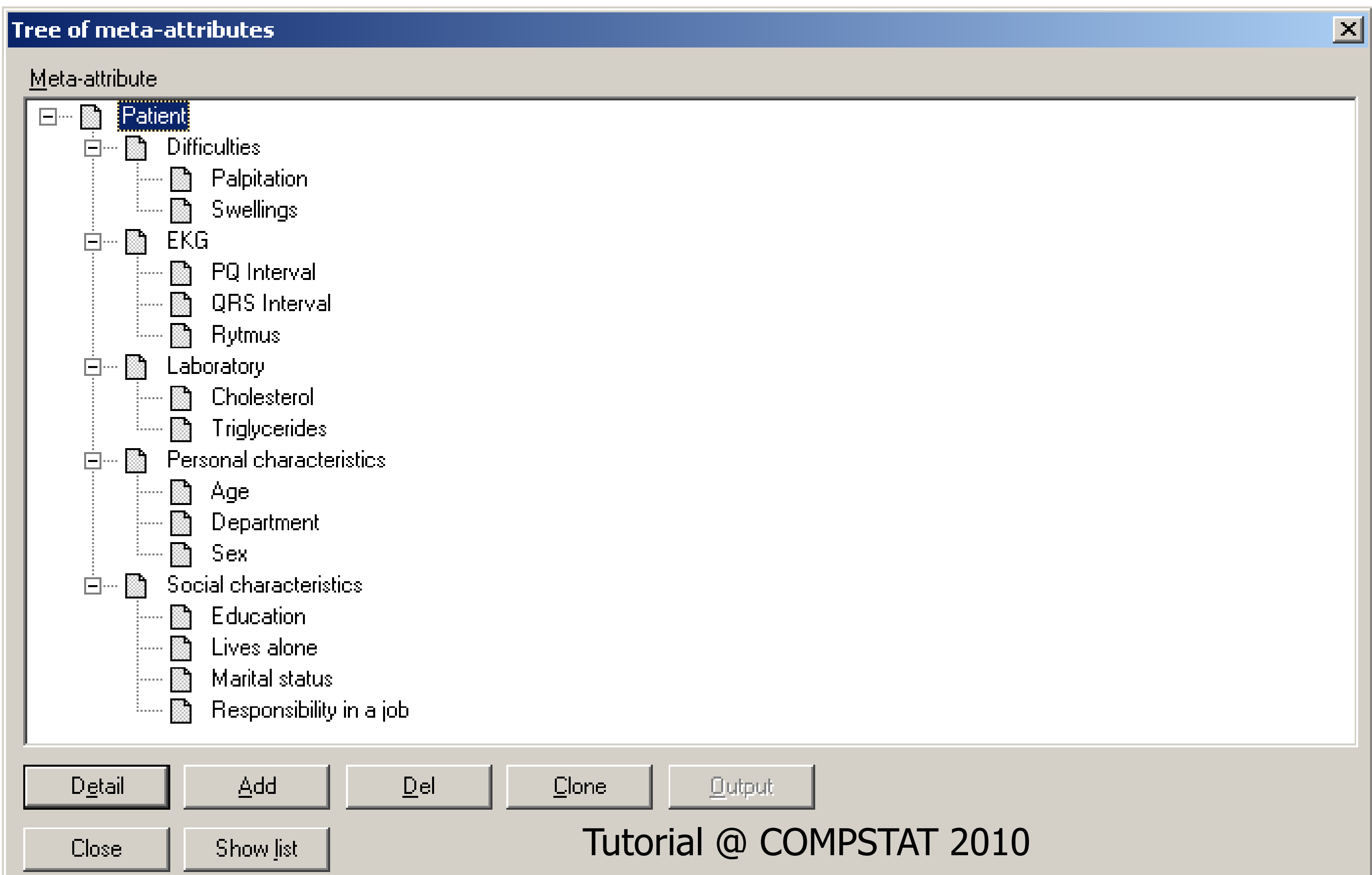
- Mines for patterns $\alpha \otimes \beta: \varphi \approx \psi / \chi$
- Are there any differences between sets α and β what concerns relation of some φ and ψ when condition χ is satisfied?
- Based on same principles as 4ft-Miner
  - definitions of α, β, φ, ψ, χ
  - measures of association on $\langle a, b, c, d \rangle$
- Powerful tool, requires careful applications
- Necessity to use domain knowledge

Outline

- GUHA – main features
- Association rule – couple of Boolean attributes
- GUHA procedure ASSOC
- LISp-Miner
- Related research
  - Domain knowledge
  - SEWEBAR project
  - Observational calculi
  - EverMiner project

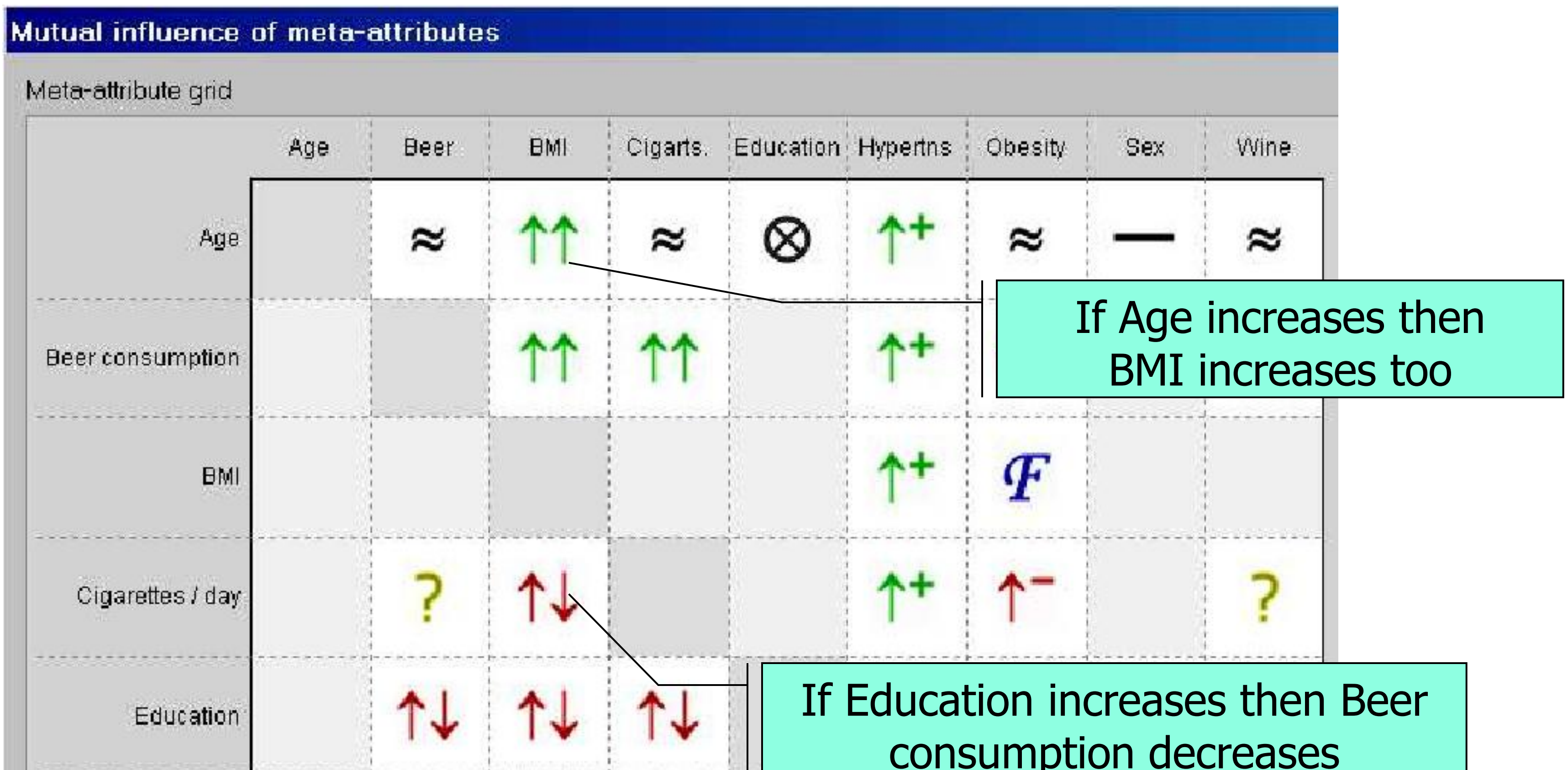
LISp-Miner Knowledge Base (1)

Storing and maintaining groups of attributes:



LISp-Miner Knowledge Base (2)

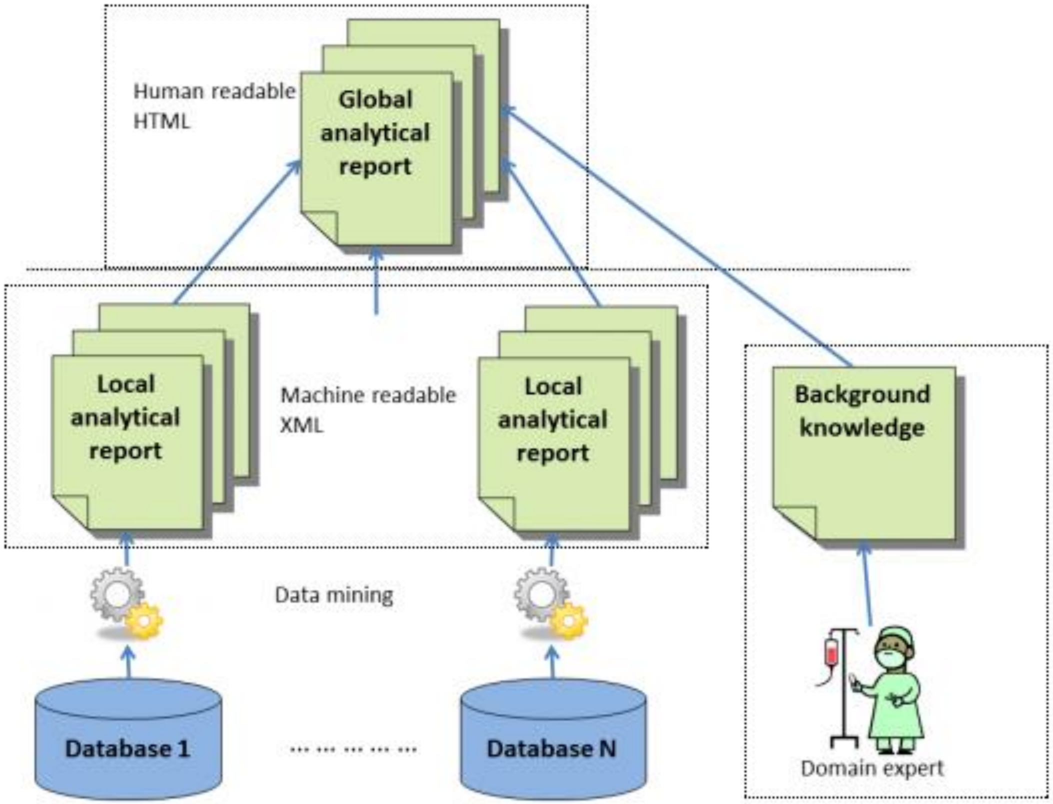
Mutual influence of attributes



SEWEBAR project

http://sewebar.vse.cz/

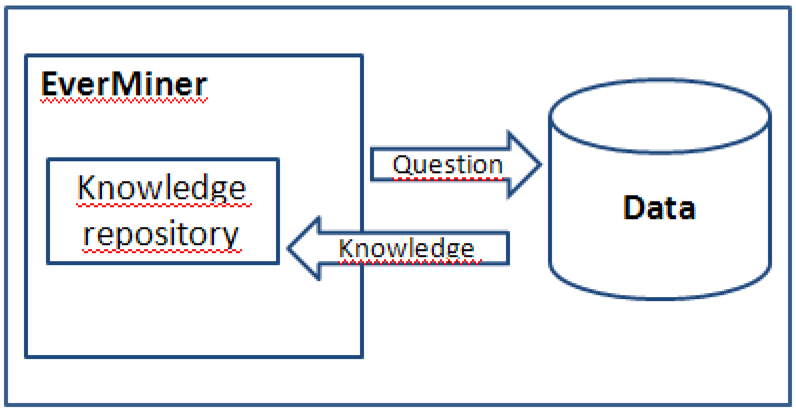
SEWEBAR (SEmantic WEB and Analytical Reports)

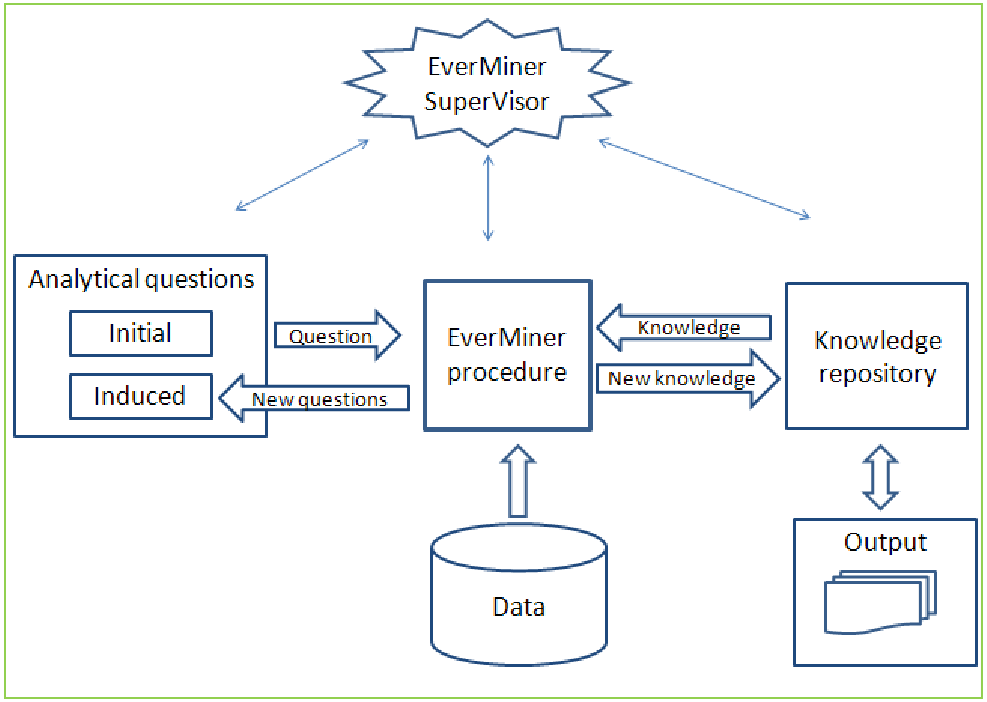


Key Concepts

- LISp-Miner Data Mining System
- Ferda Data Mining System
- Association Rules
- GUHA method

EverMiner project





Observational calculi

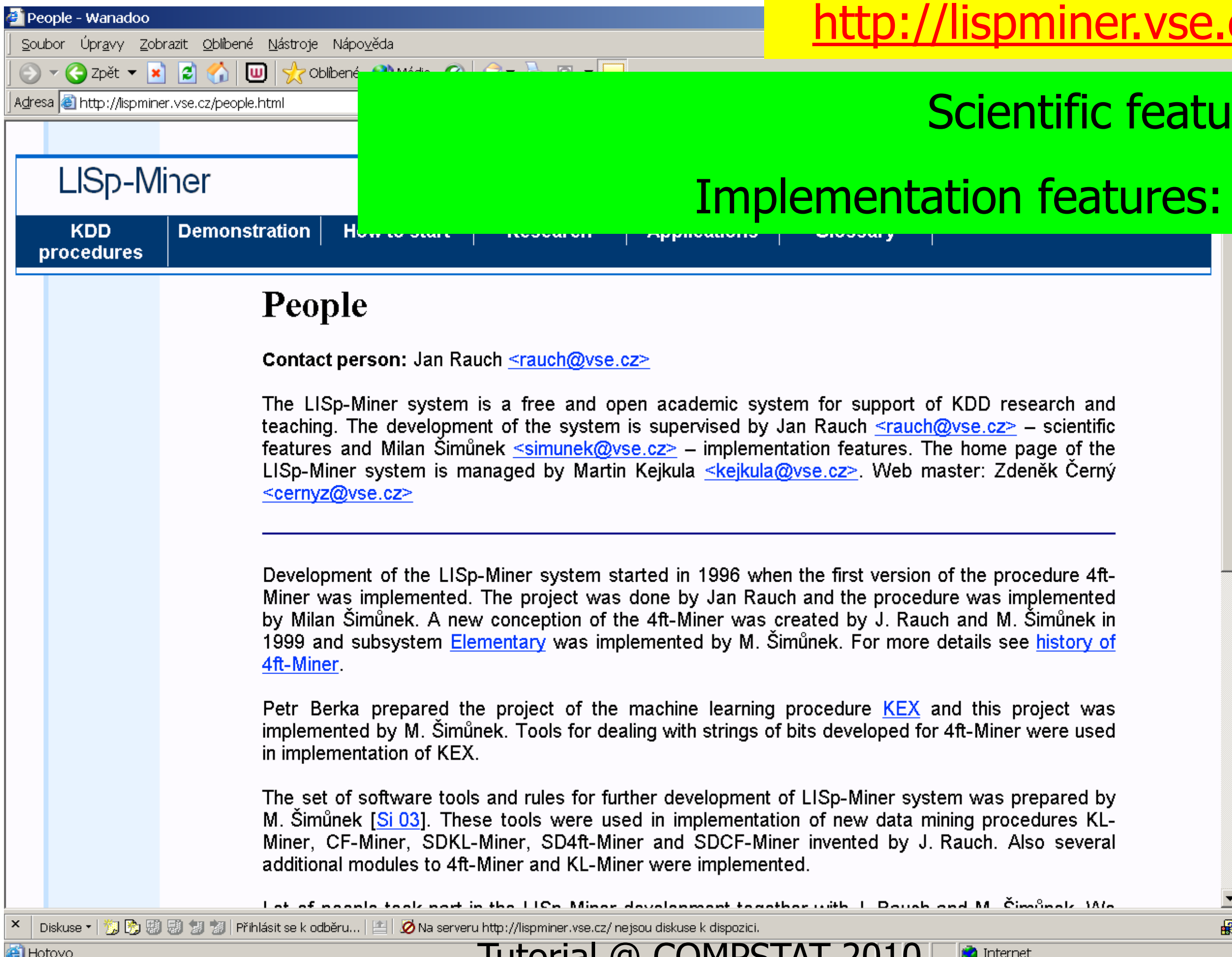
- Logical calculi with formulas – patterns mined from data
- Study of logical properties of such calculi
- Logic of association rules $\varphi \approx \psi$
- Deduction rules between association rules
  - $\dfrac{\varphi \approx \psi}{\varphi' \approx \psi'}$ is correct iff … ; $\dfrac{A(\alpha) \Rightarrow_{0.9,50} B(\beta)}{A(\alpha) \Rightarrow_{0.9,50} B(\beta) \vee C(\gamma)}$ is correct
- Various applications

LISp-Miner - authors

http://lispminer.vse.cz/people.html

Scientific features: Jan Rauch

Implementation features: Milan Šimůnek



People - Wanadoo

Soubor Úpravy Zobrazit Oblíbené Nástroje Nápověda

Adresa http://lispminer.vse.cz/people.html

LISp-Miner

KDD procedures | Demonstration

People

Contact person: Jan Rauch <rauch@vse.cz>

The LISp-Miner system is a free and open academic system for support of KDD research and teaching. The development of the system is supervised by Jan Rauch <rauch@vse.cz> – scientific features and Milan Šimůnek <simunek@vse.cz> – implementation features. The home page of the LISp-Miner system is managed by Martin Kejkula <kejkula@vse.cz>. Web master: Zdeněk Černý <cernyz@vse.cz>

Development of the LISp-Miner system started in 1996 when the first version of the procedure 4ft-Miner was implemented. The project was done by Jan Rauch and the procedure was implemented by Milan Šimůnek. A new conception of the 4ft-Miner was created by J. Rauch and M. Šimůnek in 1999 and subsystem Elementary was implemented by M. Šimůnek. For more details see history of 4ft-Miner.

Petr Berka prepared the project of the machine learning procedure KEX and this project was implemented by M. Šimůnek. Tools for dealing with strings of bits developed for 4ft-Miner were used in implementation of KEX.

The set of software tools and rules for further development of LISp-Miner system was prepared by M. Šimůnek [Si 03]. These tools were used in implementation of new data mining procedures KL-Miner, CF-Miner, SDKL-Miner, SD4ft-Miner and SDCF-Miner invented by J. Rauch. Also several additional modules to 4ft-Miner and KL-Miner were implemented.

Diskuse Přihlásit se k odběru... Na serveru http://lispminer.vse.cz/ nejsou diskuse k dispozici.

Hotovo Internet

Further readings

- Rauch J., Šimůnek M. (2005) An Alternative Approach to Mining Association Rules.In: Lin T Y et al. (eds) Data Mining: Foundations, Methods, and Applications, Springer-Verlag, pp. 219—238

- Šimůnek M. (2003) Academic KDD Project LISp-Miner. In Abraham A. et all (eds) Advances in Soft Computing - Intelligent Systems Design and Applications, Springer, Berlin Heidelberg New York

- Rauch J.: (2005) Logic of Association Rules. Applied Intelligence 22, 9–-28.

- Rauch J., Šimůnek M.(2009) Dealing with Background Knowledge in the SEWEBAR Project. In: Berendt B. et al.: Knowledge Discovery Enhanced with Semantic and Social Information}. Berlin, Springer-Verlag, 2009, pp. 89 – 106

- Kliegr T., Ralbovský M., Sv\'atek V., Šimůnek M., Jirkovský V., Nemrava J., Zemánek, J.(2009) Semantic Analytical Reports: A Framework for Post-processing data Mining Results. In: Foundations of Intelligent Systems. Berlin, Springer Verlag, 2009, pp. 88 — 98.