

# The GUHA method and its meaning for data mining

Petr Hájek, Martin Holeňa, Jan Rauch

**Introduction.** GUHA: a method of exploratory data analysis developed in Prague since mid-sixties of the past century.

GUHA stands for General Unary Hypotheses Automaton. (× the surname Guha)

GUHA has several features of contemporary systems of data mining and knowledge discovery in databases (KDD) and may be well considered as an early example of such systems - developed in ignorance of the GUHA method and its theory.

Already the first English paper on GUHA (Hajek-Havel-Chytil 1966) describes generation of almost true implications with a good antecedent which is very closely related to what is presently called mining association rules.

Aims here: (1) to list basic principles, design choices and properties of GUHA as a data-mining method and system (and of the main GUHA-procedure ASSOC),  
(2) to refer on the LISP-miner activities on the University of Economics Prague (Rauch),  
(3) to refer on some generalizations of GUHA using fuzzy approach (Holeňa).

## **GUHA-its principles and sources**

**(1) The main principle** is formulated in H-H-Ch 1966 as using logic “to describe all the possible assertions which might be hypotheses” , . . . “to generate in some optimal manner all the formal assertions of the special type which should be examined, to verify each such assertion and to output the interesting assertions”. “To summarize, the function of GUHA is to offer hypotheses . . . , not to verify previously formulated hypotheses.

**(2) Data.** Basically: a matrix (relation table) of zeros and ones, each row corresponding to an object and each column corresponding to an attribute (property). (Later versions admit also nominal and real attributes, dichotomized inside GUHA.)

**(3) Composed attributes.** From the very beginning, work with negations: for each (atomic) attribute  $V$ , one has  $\neg V$  (not  $V$ ).

Also: elementary conjunctions of the form  $L_1 \& \dots \& L_k$  where  $L_i$  are literals (atomic attributes or their negations), no attribute occurring both positively and negatively) and elementary disjunctions  $L_1 \vee \dots \vee L_k$ . For example:

SEX: MALE AND AGE: ( $< 30$ ) AND NOT PROFESSION: SCIENTIST

SEX: MALE OR AGE: ( $< 30$ ) OR NOT PROFESSION: SCIENTIST

Dichotomization of this kind in GUHA since H-1973.

**(4) Hypotheses as associations.** A GUHA procedure generates and evaluates hypotheses (rules in DM-slang) on some association of two (atomic or composed) attributes  $A$  (antecedent) and  $S$  (succedent). HHCH 1966: hypotheses of the form

$$A \Rightarrow_{p,s} S$$

( $A$  an elementary conjunction,  $S$  an elementary disjunction) where  $0 < p \leq 1$  and  $s$  is an integer;

$A \Rightarrow_{p,s} S$  is true in the data if at least  $s$  objects satisfy  $A$  and the relative frequency of objects satisfying  $S$  among those satisfying  $A$  is  $\geq p$ . (DM language:  $s$  *minsup* – minimum support and  $p$  the minimal confidence.)

Two frequencies:  $a = fr(A \& S)$ , and  $b = fr(A \& \neg S)$

In general one has to know also  $c = fr(\neg A \& \neg S)$  and  $d = fr(\neg A \& S)$ ;  
*four-fold table* (4ft) of  $A, S$ ; the truth-values of a general association (denoted  $A \sim S$ ) is given by the four-fold table.

	$\psi$	$\neg\psi$	
$\varphi$	$a$	$b$	$r$
$\neg\varphi$	$c$	$d$	$s$
	$k$	$l$	$m$

**(5) Associations and quantifiers.** Each notion of association is given by a function  $q$  assigning to each 4ft  $(a, b, c, d)$  the value 1 (yes, associated) or 0 (no, not associated).

Statisticians:  $q$  is a statistic;

logical standpoint:  $q$  defines a generalized quantifier (has become a part of GUHA-slang) The quantifier  $\Rightarrow_{p,s}$  is called founded implication (FIMPL) and is an example of implicational quantifiers satisfying

IF  $a' \geq a$  AND  $b' \leq b$  AND  $q(a, b, c, d) = 1$  THEN  $q(a', b', c', d') = 1$ .

If  $\Rightarrow^*$  is an implicational quantifier then read  $A \Rightarrow^* S$  as "Many  $A$ 's are  $S$ 's".



More generally, a quantifier  $\sim^*$  is associational if

$a' \geq a, b' \leq b, c' \leq c, d' \geq d$  and  $q(a, b, c, d) = 1$  implies  $q(a', b'c'd') = 1$ .

The simplest example quantifier  $\sim_0$  (called SIMPLE):  $q(a, b, c, d) = 1$  iff  $ad > bc$  (and  $a > BASE$  - a support threshold), equivalently:  $a/(a + b) > c/(c + d)$  saying "S is relatively more frequented in A than in  $\neg A$ ". Note the symmetry:  $A \sim S$  is equivalent to  $S \sim A$ . More sophisticated implicational and symmetric associational quantifiers use test statistics of statistical hypotheses - introduced in GUHA by Havránek in 1971 (LIMPL, UIMPL in GUHA slang) and (by me) in 1968 (FISHER). General theory of implicational and associational quantifiers was initiated in my 1973 paper.

## 6. More on quantifiers.

	$\psi$	$\neg\psi$	
$\varphi$	$a$	$b$	$r$
$\neg\varphi$	$c$	$d$	$s$
	$k$	$l$	$m$

Recall  $a, r$  from the data; assuming  $P(\psi|\varphi) \leq p$  estimate the probability that from  $r$  randomly chosen rows satisfying  $\varphi$ , at least  $a$  satisfy  $\psi$ ; upper bound is

$$LIMPL_p(a, r) = \sum_{i=a}^r \binom{r}{i} p^i (1-p)^{r-i}$$

(Note:  $\binom{r}{i}$  is the number of  $i$ -element subsets of an  $r$ -element set.)

Choose small  $\alpha$  (significance threshold,  $\alpha = 0.05$ ). If  $LIMPL_p(a, r) \leq \alpha$ , reject the hypothesis  $P(\psi|\varphi) \leq p$  (since it implies that what we have observed is very unlikely) - accept  $P(\psi|\varphi) > p$

## Example

240	3
511	73

$$LIMPL_{0.9}(240, 243) = 2.76 \cdot 10^{-8}$$

$$Fr(\psi|\varphi) = 0.987654$$

240	3
240	3

$$LIMPL, Fr(\psi|\varphi) \text{ the same;}$$

$$ad = bc$$

*Fact:* The following rules are sound for the simple quantifier  $\sim_0$   
 (and for  $\Leftrightarrow_p, \Leftrightarrow_p^!$ )

$$\text{(Symmetry)} \quad \frac{\varphi \sim_0 \psi}{\psi \sim_0 \varphi}$$

$$\text{(Negation)} \quad \frac{\varphi \sim_0 \psi}{\neg\psi \sim_0 \neg\varphi}$$

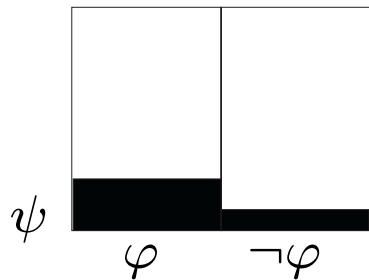
(observe the tables:)

	$\psi$	$\neg\psi$
$\varphi$	$a$	$b$
$\neg\varphi$	$c$	$d$

	$\varphi$	$\neg\varphi$
$\psi$	$a$	$c$
$\neg\psi$	$b$	$d$

	$\neg\psi$	$\psi$
$\varphi$	$d$	$c$
$\neg\varphi$	$b$	$a$

Warning:  $\varphi \sim_0 \psi$  is no implication; only compares relative frequencies



Warning: Implicational quantifiers mostly do not satisfy symmetry, negation:

9	1
8	2

$\varphi \sim_0 \psi$  true,  $\varphi \Rightarrow_{0.9} \psi$  true;  
 $\psi \Rightarrow_{0.9} \varphi$  false,  $\neg\varphi \Rightarrow_{0.9} \neg\psi$  false

*Warning:* A quantifier given by  $\Rightarrow^*$ , i.e.

$(\varphi \Rightarrow^* \psi) \wedge (\neg\varphi \Rightarrow^* \neg\psi)$  satisfies negation but not necessarily symmetry.

9	1
10	90

$$\frac{a}{a+b} = \frac{d}{c+d} = 0.9; \frac{a}{a+c} = \frac{9}{19} \doteq \frac{1}{2}$$

Admittedly, the simple quantifier  $\sim_0$  is not fine enough; there is a very fine statistical variant, testing

$P(\psi|\varphi) > P(\psi)$  - Fisher quantifier

$$\sim_{\alpha}^{FISH}(a, b, c, d) = 1 \text{ iff } ad > bc \text{ and } \sum_{i=0}^{\min(b,c)} \frac{(a+b)!(a+c)!(b+d)!(c+d)!}{m!(a+i)!(c-i)!(c-i)!(c+i)!} \leq \alpha$$

where  $\alpha$  is small (0.05 or so)

Fact: FISHER satisfies symmetry and negation.

Read

$$\varphi \sim_{\alpha}^{FISH} \psi$$

“ $\varphi, \psi$  are mutually associated” (with significance  $\alpha$ )

*Example:*

60	231	291
16	375	391
76	606	682

$$\text{FISHER}(60, 231, 16, 375) = 9.089 * 10^{-12}$$

$$Fr(\psi|\varphi) = 60/291 = 0.2062$$

$$Fr(\psi) = 76/606 = 0.1114$$

$$Fr(\psi|\neg\varphi) = 16/391 = 0.0409$$



Moral:

Rich variety of reasonable kinds of association rules (quantifiers):

Many  $\varphi$ 's's are  $\psi$ 's

Many  $\varphi$ 's are  $\psi$ 's and many  $\psi$ 's are  $f$ 's,

many  $\varphi$ 's are  $\psi$  and many  $\neg\varphi$ 's are  $\neg\psi$ 's

$\varphi$  makes  $\psi$  more frequented (than  $\neg\varphi$  does)

Reasonable logical properties. Various tricks for quick computation

General theory: logic and statistics.

Interplay of logic and statistics is a fascinating feature in GUHA-style DM/KDD. Pioneering paper Havránek 1971, further his 1975, 1977.

Monograph Hájek-Havránek: Mechanizing hypothesis formation (Springer 1978) - basic theoretical work introducing and developing observational and theoretical logical calculi, their relations to many-valued logics and to statistical inference. Deduction rules on the observational level enabling several optimizations in pruning the tree of hypotheses. Foundations for a general GUHA-procedure generating hypotheses on associations are exhibited, including three methods of handling missing values (originally discussed in Hájek-Bendová-Renc in 1971 and Hájek in 1973. Important: global interpretation of results (Havránek).

The GUHA procedure ASSOC.

The "pre-historical" first implementation of a GUHA procedure by the late I. Havel in 1971 on the Soviet MINSK12. Several implementations during the years. They realize generation of (GUHA)-hypotheses of the form  $A \sim S$  where  $A$  and  $S$  are elementary conjunctions and  $\sim$  is one of six (or more) associational quantifiers, each having several parameters (and possibly also more complex hypotheses, more will be said later).

Three parts:

Pre-processing: input (and possible transformation) of the data and several parameters defining antecedent attributes, succedent attributes, the quantifier used and its parameters, maximal length of antecedent/succedent etc.

Core: Generation and evaluation of hypotheses, with maximal possible pruning of the tree of hypotheses, using various deduction rules. Solution file created.

Post-processing, interpretation. On-line browsing the set of found hypotheses, sorting according to several criteria, printing output reports).

See [www.cs.cas.cz](http://www.cs.cas.cz) (Software, GUHA)

## Generalizing GUHA. (Rauch)

New features of GUHA developed at Faculty of Informatics and Statistics of University of Economics in Prague (UEP).

Motivated by attempts to apply GUHA to data stored in databases. implementation of GUHA method – not using the very known apriori algorithm (Agrawal) but based on representation of analyzed data by suitable strings of bits

These procedures mine for variety of patterns based on verification of several types of contingency tables.

Implemented in the academic software system *LISp-Miner* (<http://lispminer.vse.cz>) developed since 1996 at UEP.

New recent implementations of these six procedures use a visual interface inspired by the Clementine system  
*Ferda* software system.

## LISp-Miner

The idea of the GUHA method : given data, let the computer generate all (or as much as possible) interesting hypotheses of a given logical form that are supported by the data

GUHA-procedure is a computer program, the input of which consists of the analyzed data and of a simple definition of relevant patterns;

it generates each particular relevant pattern (with some optimizations) and tests if it is true in the analyzed data. The output of the procedure consists of all prime patterns.

Prime pattern: true in the analyzed data and does not immediately follow from the other more simple output patterns

All six GUHA procedures implemented in the LISp-Miner system deal with data matrices.

finitely valued attributes.

object i.e. row of $\mathcal{M}$	columns of $\mathcal{M}$ i.e. attributes				examples of literals		cards of categories of attribute $A_1$		
	$A_1$	$A_2$	$\dots$	$A_{100}$	$A_1(1, 2)$	$\neg A_{100}(6)$	$A_1[1]$	$A_1[2]$	$A_1[3]$
$o_1$	1	7	$\dots$	4	$T$	$T$	1	0	0
$o_2$	3	4	$\dots$	6	$F$	$F$	0	0	1
$o_3$	2	9	$\dots$	9	$T$	$T$	0	1	0
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$o_n$	3	1	$\dots$	6	$F$	$F$	0	0	1

*categorical attributes.* Attributes  $A_1$  and  $A_{100}$  are examples of categorical attributes. Attribute  $A_1$  has categories  $\{1,2,3\}$  etc. A Boolean attribute is built using logical connectives from basic Boolean attributes –  $A(\alpha)$  where  $\alpha \subset \{a_1, \dots, a_k\}$  and  $\{a_1, \dots, a_k\}$  is the set of all categories of the attribute  $A$ .  $\alpha$  is the *coefficient* of the literal  $A(\alpha)$ .

Three types of contingency tables:

*4ft-table*  $4ft(\varphi, \psi, \mathcal{M})$  of the Boolean attributes  $\varphi$  and  $\psi$  in  $\mathcal{M}$

*KL-table*  $KL(R, C, \mathcal{M})$  of the categorical attributes  $R$  and  $C$  in  $\mathcal{M}$

*CF-table*  $CF(R, \mathcal{M})$  of the categorical attribute  $R$  in  $\mathcal{M}$ .



$\mathcal{M}$	$\psi$	$\neg\psi$	$\mathcal{M}$	$c_1$	$\dots$	$c_L$
$\varphi$	$a$	$b$	$r_1$	$n_{1,1}$	$\dots$	$n_{1,L}$
$\neg\varphi$	$c$	$d$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
			$r_K$	$n_{K,1}$	$\dots$	$n_{K,L}$
			$\mathcal{M}$	$r_1$	$\dots$	$r_K$
				$n_1$	$\dots$	$n_K$

$$4ft(\varphi, \psi, \mathcal{M}) \quad KL(R, C, \mathcal{M})$$

$$CF(R, \mathcal{M})$$

The computation of all contingency tables is based on representation of analysed data by *cards of categories*. The card of category is a string of bits. Cards of categories are used both to build corresponding cards of Boolean attributes and to compute particular frequencies from contingency tables. Card  $\mathcal{C}(\varphi)$  of Boolean attribute  $\varphi$  is a string of bits. Each row of the data matrix corresponds to one bit of  $\mathcal{C}(\varphi)$ . There is "1" in the  $i$ -th bit if and only if  $\varphi$  is true in row  $o_i$ . It is important that the bit-wise Boolean operations  $\wedge$ ,  $\vee$  and  $\neg$  are carried out by very fast computer instructions.

There are six GUHA procedures implemented in the LISP-Miner system:

4ft-Miner, KL-Miner, CF-Miner, SD4ft-Miner, SDKL-Miner, and SDCF-Miner.

The procedure **4ft-Miner** is the enhanced procedure ASSOC. The *4ft-Miner* mines for association rules  $\varphi \approx \psi$  and for conditional association rules  $\varphi \approx \psi/\gamma$ . The *association rule*  $\varphi \approx \psi$  means that the Boolean attributes  $\varphi$  and  $\psi$  are associated in the way given by the symbol  $\approx$ . This symbol is called *4ft-quantifier*. Its semantics is given by a function associating with each 4ft-table the truth value 0 or 1.

The association rule  $\varphi \approx \psi$  is true in the data matrix  $\mathcal{M}$  if the condition corresponding to the 4ft-quantifier  $\approx$  is satisfied in the 4ft-table  $4ft(\varphi, \psi, \mathcal{M})$ . The quantifier  $\Rightarrow_{p,t}$  and the quantifiers motivated by the statistical hypothesis tests are examples of 4ft-quantifiers. An additional example of the 4ft-quantifier is the 4ft-quantifier  $\sim_{p,t}^+$  of *above average dependence*. It is defined for  $0 < p$  and  $t > 0$  by the condition  $\frac{a}{a+b} \geq (1+p)\frac{a+c}{a+b+c+d} \wedge a \geq t$ . There are 17 types of the 4ft-quantifiers implemented in the 4ft-Miner procedure.

The procedure 4ft-Miner mines also for *conditional association rules* of the form  $\varphi \approx \psi/\gamma$ . The conditional association rule  $\varphi \approx \psi/\gamma$  means that  $\varphi$  and  $\psi$  are in the relation given by the 4ft-quantifier  $\approx$  when the condition given by the Boolean attribute  $\gamma$  is satisfied. The rule  $\varphi \approx \psi/\gamma$  is true in the data matrix  $\mathcal{M}$  if the condition corresponding to the 4ft-quantifier  $\approx$  is satisfied in the 4ft-table  $4ft(\varphi, \psi, \mathcal{M}/\gamma)$ . The term  $\mathcal{M}/\gamma$  denotes the data matrix consisting of all rows of data matrix  $\mathcal{M}$  satisfying the condition  $\gamma$ .

The procedure **KL-Miner** mines for KL-patterns  $R \sim C/\gamma$ . The *KL-pattern*  $R \sim C/\gamma$  means that the categorial attributes  $R$  and  $C$  are in a relation given by the symbol  $\sim$  when the condition given by the Boolean attribute  $\gamma$  is satisfied. The symbol  $\sim$  is called *KL-quantifier*. A KL-quantifier corresponds to a condition concerning the *KL-table*  $KL(R, C, \mathcal{M}')$  of the categorial attributes  $R$  and  $C$  in the data matrix  $\mathcal{M}'$  in question. The *KL-pattern*  $R \sim C/\gamma$  is true in the data matrix  $\mathcal{M}$  if the condition corresponding to  $\sim$  is satisfied in the KL-table  $KL(R, C, \mathcal{M}/\gamma)$  of the categorial attributes  $R$  and  $C$  in the data matrix  $\mathcal{M}/\gamma$ .

The procedure **CF-Miner** mines for CF-patterns of the form  $\sim R/\gamma$ . A *CF-pattern*  $\sim R/\gamma$  means that frequencies of categories of attribute  $R$  satisfy the condition given by the symbol  $\sim$  when an other condition given by the Boolean attribute  $\gamma$  is satisfied. The symbol  $\sim$  is called *CF-quantifier* here. The CF-quantifier corresponds to a condition concerning the *CF-table*  $CF(R, \mathcal{M}')$  of the categorial attribute  $R$  in the data matrix  $\mathcal{M}'$  in question. The *CF-pattern*  $\sim R/\gamma$  is true in the data matrix  $\mathcal{M}$  if the condition corresponding to  $\sim$  is satisfied in the CF-table  $CF(R, \mathcal{M}/\gamma)$  of the categorial attribute  $R$  in the data matrix  $\mathcal{M}/\gamma$ .

The procedure **SD4ft-Miner** mines for SD4ft-patterns of the form  $\alpha \bowtie \beta : \varphi \approx^{SD} \psi/\gamma$ . Here  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\varphi$ , and  $\psi$  are Boolean attributes derived from the columns of analyzed data matrix  $\mathcal{M}$ . The attributes  $\alpha$  and  $\beta$  define two subsets of rows. The attribute  $\gamma$  defines a condition. The attributes  $\varphi$  and  $\psi$  are *antecedent* and *succedent* of the association rule  $\varphi \approx \psi$  in question. The SD4ft-pattern  $\alpha \bowtie \beta : \varphi \approx^{SD} \psi/\gamma$  means that the subsets given by the Boolean attributes  $\alpha$  and  $\beta$  differ as for validity of association rule  $\varphi \approx \psi$  when the condition given by the Boolean attribute  $\gamma$  is satisfied. A measure of difference is defined by the symbol  $\approx^{SD}$  that is called *SD4ft-quantifier*. The SD4ft-quantifier corresponds to a condition concerning two 4ft-tables  $\langle a, b, c, d \rangle$  and  $\langle a', b', c', d' \rangle$ . The SD4ft-pattern  $\alpha \bowtie \beta : \varphi \approx^{SD} \psi/\gamma$  is true in the data matrix  $\mathcal{M}$  if the condition corresponding to  $\approx^{SD}$  is satisfied for the 4ft-tables  $4ft(\varphi, \psi, \mathcal{M}/(\alpha \wedge \gamma))$  and  $4ft(\varphi, \psi, \mathcal{M}/(\beta \wedge \gamma))$ .

The input of the GUHA procedure: the analyzed data and of a simple definition of a large set of relevant patterns. The analyzed data has the form of data matrix for all GUHA procedures implemented in the LISp-Miner system. The set of relevant association rules  $\varphi \approx \psi$  to be generated and verified by the procedure 4f-Miner is given by the definition of the set  $\mathcal{B}_{ant}$  of *relevant antecedents* (i.e. set of Boolean attributes), the definition of the set  $\mathcal{B}_{suc}$  of *relevant succedents* and by the 4ft-quantifier  $\approx$ . All rules  $\varphi \approx \psi$  such that  $\varphi \in \mathcal{B}_{\varphi}$  and  $\psi \in \mathcal{B}_{\psi}$  are verified. The set of relevant conditional association rules  $\varphi \approx \psi/\gamma$  is defined in a similar way by  $\mathcal{B}_{ant}$ ,  $\mathcal{B}_{suc}$ ,  $\mathcal{B}_{cond}$  and  $\approx$  where  $\mathcal{B}_{cond}$  is *the set of relevant conditions*. Similarly for the set of relevant SD4ft-patterns  $\alpha \bowtie \beta : \varphi \approx^{SD} \psi/\gamma$  and the set of KL-patterns  $R \sim C/\gamma$  to be generated and verified by the KL-Miner procedure etc.



The GUHA procedures of the LISp-Miner system were many times used to analyze real data. We give a short example of its applications in the medical data STULONG. The example concerns the task "What are the differences between the groups of normal and of risk patients what concerns relation of Boolean characteristics of patient's examination and of patient's blood pressure. The alcohol consumption can be considered as an additional condition." One way to solve this task is to search for all SD4ft-patterns  $normal \bowtie risk : \varphi \Rightarrow_{30,30,0.4}^{SD} \psi/\gamma$  such that  $\varphi \in \mathcal{B}_{patient}$ ,  $\psi \in \mathcal{B}_{blood}$  and  $\gamma \in \mathcal{B}_{alcohol}$ . The SD4ft quantifier  $\Rightarrow_{30,30,0.4}^{SD}$  is defined by the condition  $a \geq 30 \wedge a' \geq 30 \wedge \left| \frac{a}{a+b} - \frac{a'}{a'+b'} \right| \geq 0.4$ . Informally speaking, the core of this condition is that the confidence of the association rule  $\varphi \Rightarrow^* \psi$  on the set of normal patients differs from the confidence of the same rule on the set of risk patients, we however consider only the patients satisfying  $\gamma$ .

We solved the task in the data matrix STULONG concerning about 1417 patients. More than  $12 * 10^6$  SD4ft patterns were generated and verified in about 6 minutes at PC with 1.58 GHz and 512 MB RAM. There are 20 output SD4ft-patterns. The strongest one (i.e., the pattern with the highest difference of confidences  $\frac{a}{a+b} - \frac{a'}{a'+b'}$ ) is the pattern

$$normal \bowtie risk : Patient \Rightarrow_{31,31,0.45}^{SD} Blood / Vine(not) .$$

It is  $Patient = Weight\langle 71, 80 \rangle \wedge Height\langle 168, 177 \rangle$  and  $Blood = Diastolic\langle 60, 90 \rangle \wedge Systolic\langle 110, 140 \rangle$ .

The corresponding 4ft-tables are:

$normal \wedge Vine(not)$	$Blood$	$\neg Blood$
$Patient$	31	2
$\neg Patient$	43	23

$risk \wedge Vine(not)$	$Blood$	$\neg Blood$
$Patient$	31	32
$\neg Patient$	109	155

Note that the confidence of the rule  $Patient \Rightarrow^* Blood$  is 0.94 for normal patients not drinking vine and 0.49 for risk patients not drinking vine i. e. the difference of confidences is 0.45. This difference of confidences is very high and points to difference between normal and risk patients that is interesting from the medical point of view.

## Research projects related to LISp-Miner

Big challenge: automatized chaining of particular procedures of LISp-Miner to solve given problem. Recall GUHA80 (never realized). One of current research activities is building an open system called *EverMiner* of tools to facilitate solving real problems using all procedures implemented in the LISp-Miner system. The *EverMiner* system will consist of typical tasks, scenarios, repositories of partial cedents etc.

The efficient applications of particular GUHA procedures requires however also background knowledge related to area of applications. An approach to it is based on storing relevant background knowledge in the special part of the LISp-Miner system, this part is called *LISp-Miner Knowledge Base*

There are *10 challenging problems in data mining research* listed at the home-page of *IEEE International conference on data mining*, see <http://www.cs.uvm.edu/~icdm/>. It is among other emphasized that results of data data mining must be related to the real world decisions they affect. We attempt to produce analytical reports automatically. Such analytical reports are natural candidates for Semantic Web. Thus the research project called *SEWEBAR* was launched.

## **Fuzzy hypothesis testing for GUHA. (Holeña)**

In the period 1995–2004, an important direction of theoretical research connected to the method GUHA was the generalization of hypotheses tests used in many generalized quantifiers to fuzzy hypotheses testing. Recall: overwhelming majority of generalized quantifiers used in the GUHA method have a statistical motivation. GUHA adopts the statistical point of view of data concerning a set of objects as a random sample from some probability distribution. The truth functions of generalized quantifiers are then defined in such a way that they correspond to the application of some statistical method to such random samples. In particular, two key types of statistical methods: parameter estimation and hypotheses testing.

1. In the case of *generalized quantifiers based on parameter estimation*, the truth function states that some estimator of some parameter of the probability distribution of the random sample  $\|\varphi\|, \|\psi\|, \dots$  fulfils some prescribed condition.

The best known quantifier of that kind is the *founded implication*  $\Rightarrow_\theta$ , for which the conditional probability  $P(\|\psi\| = 1 \mid \|\varphi\| = 1)$  – based on the unbiased estimator  $\frac{a}{a+b}$ ; recall  $q \Rightarrow_\theta(a, b, c, d) = 1$  if  $\frac{a}{a+b} \geq \theta$  &  $a \geq A$  and  $= 0$  else.

2. *generalized quantifiers based on hypotheses testing*: the truth function corresponds to the result of some statistical test of some null hypothesis  $H_0$  concerning the probability distribution of the random sample  $\|\varphi\|, \|\psi\|, \dots$  against some alternative  $H_1$ . Hypotheses concerning the four-fold table of that random sample are tested (for example, the hypothesis that the probability distributions of the columns and the rows of the table are independent). Example: the quantifier of lower critical implication  $\Rightarrow_{\theta, \alpha}^!$ , in which  $\theta \in (\frac{1}{2}, 1)$  is a given constant, is defined for testing the hypothesis  $H_0 : P(\|\psi\| = 1 \mid \|\varphi\| = 1) \in \langle 0, \theta \rangle$  against the alternative  $H_1 : P(\|\psi\| = 1 \mid \|\varphi\| = 1) \in (\theta, 1)$  by means of the binomial test. That test has the test statistics  $\sum_{i=a}^{a+b} \binom{a+b}{i} \theta^i (1 - \theta)^{a+b-i}$  and the critical region  $C_\alpha = \langle 0, \alpha \rangle$ , which leads to the truth function

$$q_{\Rightarrow_{\theta, \alpha}^!} (a, b, c, d) = 1 \text{ if } \sum_{i=a}^{a+b} \binom{a+b}{i} \theta^i (1 - \theta)^{a+b-i} \leq \alpha, \text{ else } = 0.$$

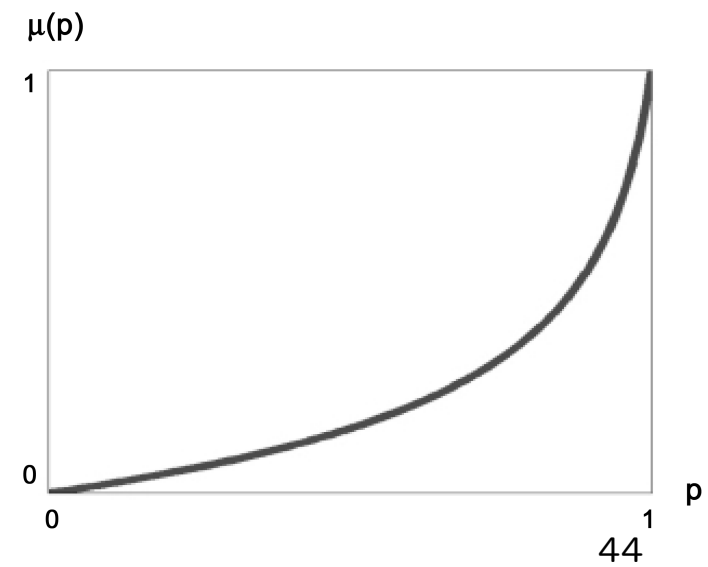
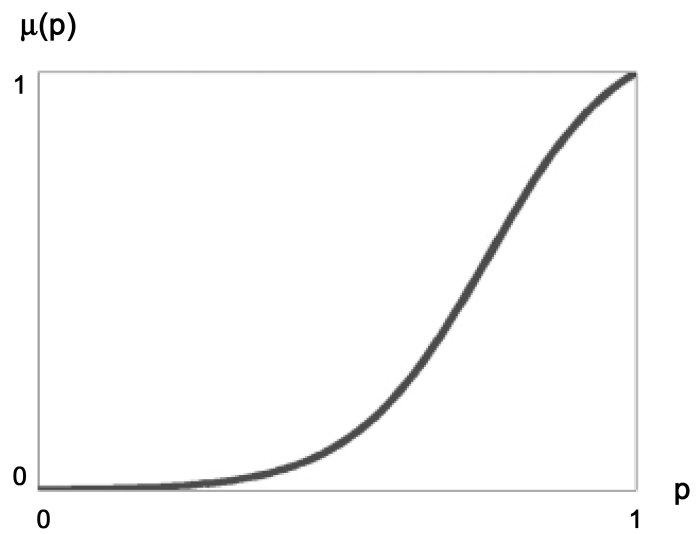
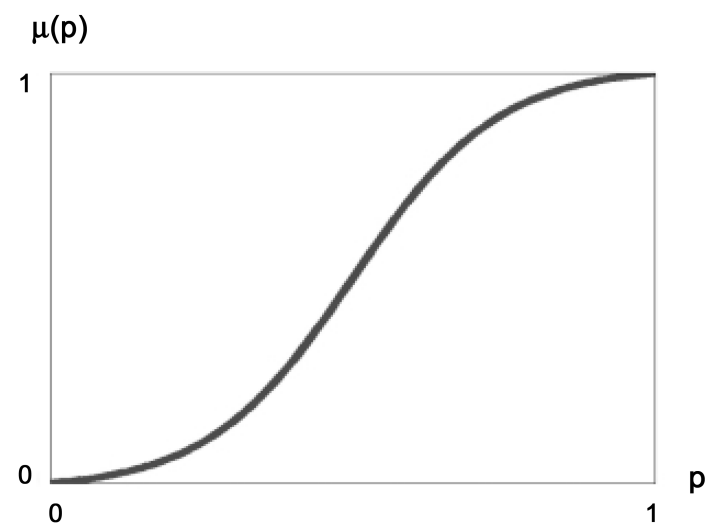
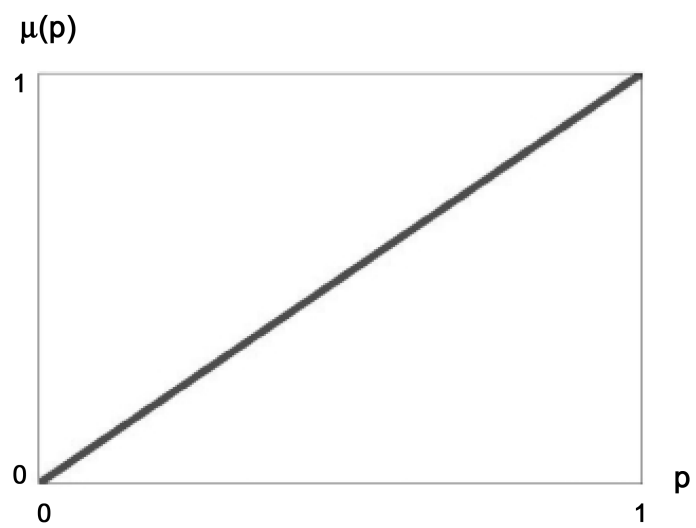


For which  $\varphi$  and  $\psi$  the sentence  $\varphi \Rightarrow_{\theta, \alpha}^! \psi$  is valid, depends on the choice of the constant  $\theta > \frac{1}{2}$  (e.g.,  $\theta = 0.8$ ,  $\theta = 0.9$ ). However, data mining is typically performed in situations when there is only very little knowledge available about the probability distribution governing the data, thus no clue for the choice of the constant  $\theta$ . Therefore, using the lower critical implication (as well as several other GUHA quantifiers) entails a large amount of subjectivity. How to decrease this subjectivity? Possibility to replace hypotheses described with traditional, crisp sets by hypotheses described with fuzzy sets (therefore called *fuzzy hypotheses*).

In the particular case of the lower critical implication, the interval  $(\theta, 1)$ ,  $\theta > \frac{1}{2}$  for the conditional probability  $P(\|\psi\| = 1 \mid \|\varphi\| = 1)$  in  $H_1$  is replaced by a fuzzy set with the intended meaning "high probability". The complementary interval  $\langle 0, \theta \rangle$  for  $P(\|\psi\| = 1 \mid \|\varphi\| = 1)$  in  $H_0$  is then replaced with a fuzzy set with the intended meaning "probability that is not high". Consequently, the sentence  $\varphi \Rightarrow_{\theta, \alpha}^! \psi$  can be interpreted "with probability at least  $1 - \alpha$  the conditional probability  $P(\|\psi\| = 1 \mid \|\varphi\| = 1)$  is high". Possibly replace even the significance level  $\alpha$  by a fuzzy set. If that fuzzy set has the intended meaning "high significance level", the sentence  $\varphi \Rightarrow_{\theta, \alpha}^! \psi$  can be finally interpreted "at a high significance level, the conditional probability  $P(\|\psi\| = 1 \mid \|\varphi\| = 1)$  is high".

- a) The definition of a fuzzy set must not allow an interpretation that would contradict the intended meaning.
- b) All fuzzy sets with the same intended meaning usually have to fulfil certain requirements. For example, for the fuzzy sets  $\mu_1$  on  $(0,1)$  with the intended meaning "high probability", should satisfy

$\mu$  is nondecreasing,  $\lim_{p \rightarrow 0+} \mu_1(p) = 0$ ,  $\lim_{p \rightarrow 1-} \mu_1(p) = 1$ .



The research reported can be divided in two distinct phases in which fuzzy hypotheses testing was studied

In the early phase, this was the framework of fuzzy set theory. In that framework, the result of testing a fuzzy hypothesis  $H_0$  is viewed as a fuzzy set on the pair of possibilities  $\{ "H_0 \text{ is rejected}" , "H_0 \text{ is not rejected}" \}$ . The membership grades of those possibilities are in the case of binary quantifiers the values of two functions that generalize the truth functions.

In the more recent phase, fuzzy hypotheses testing was studied in the framework of fuzzy logic, in which the result of testing  $H_0$  is viewed as an evaluation of the sentence  $R_{FL\forall}$  that states the rejection of  $H_0$  in an appropriate fuzzy predicate logic. This sentence is evaluated in a model that comprises crisp sets evaluating predicates of traditional observational logic on data, together with fuzzy sets evaluating fuzzy hypotheses  $H_0$  and  $H_1$ , and a fuzzy significance level.

*fuzzy logic in broader sense versus fuzzy logic in narrow sense*  
(mathematical fuzzy logic)

Main results of the more recent phase of the research reported in this section can be summarized as follows:

Formal definition of a *fuzzy predicate logic suitable for hypotheses testing*,  $FL\forall$ , which includes Boolean predicates  $\varphi_1(x), \dots, \varphi_k$  from the observational logic, fuzzy predicates  $H_0, H_1$  for hypotheses and  $S$  for a significance level, and a generalized quantifier  $\nabla()$  for testing  $H_0$  against  $H_1$ .





Formal definition of the *representation of a statistical test of  $H_0$  against  $H_1$  with the quantifier  $\nabla()$*  in a model of  $\text{FL}\forall$  that includes evaluations of the predicates of the observational logic  $\text{FL}\forall$  on data, and fuzzy sets evaluating  $H_0$ ,  $H_1$  and  $S$ , as well as formal definition of the *degree of rejection of  $H_0$*  as the evaluation of a particular sentence,  $R_{\text{FL}\forall}$ , of  $\text{FL}\forall$  in that model. Since the interpretation of fuzzy predicates in any model of  $\text{FL}\forall$  are fuzzy sets, each such model is actually a bridge between the approach to fuzzy hypotheses testing based on fuzzy logic in narrow sense, and the traditional approach relying on the fuzzy set theory.

(and some further results).



## Elbe river groyne fields ecology - DB "biodat10"

fuzzy lower critical implication: truth grades for 4 definitions of "p is high",  
above the threshold 0.95 (according to the definition 1)

				Species	Ecological factors	
1	1	1	0.91892	Cladocera	flow velocity = 50-70 cm/s	
0.98154	0.9997	0.9943	0.65362	Copepoda	grain diameter - Führböter's method = 0.4-0.6 mm	
1	1	1	0.91892	Copepoda	flow velocity = 50-70 cm/s	
0.98154	0.9997	0.9943	0.65362	Nais	grain diameter - Führböter's method = 0.4-0.6 mm	
1	1	1	0.91892	Nais	flow velocity = 50-70 cm/s	
1	1	1	0.91892	Robackia	flow velocity = 50-70 cm/s	
0.97816	0.99858	0.97911	0.73121	Tubificidae	glowable proportion = below 10 %	
1	1	1	0.91892	Cladocera	glowable proportion = below 10 %	flow velocity = 50-70 cm/s
1	1	1	0.91892	Copepoda	glowable proportion = below 10 %	flow velocity = 50-70 cm/s

**Summary:** what can data mining systems learn from the  
**GUHA method:** (besides historical priority)  
**method:** systematic generation and verification of hypotheses  
(associations), applying:  
**logic** - logical form of associations, use of logical connectives  
(conjunction, disjunction, negation,...), use of generalized quan-  
tifiers and deduction rules; many-valued and fuzzy logics  
**statistics** - statistical tests as generalized quantifiers, their log-  
ical and probabilistic properties, global interpretation of results  
of data mining  
**implementation** bit-string based, different from the well known  
Apriori algorithm and allowing good work with complex formulas.

Book Hájek: Mechanizing hypothesis formation, published 1978 by Springer, now legally freely obtainable from my web page [www.cs.cas.cz/hajek](http://www.cs.cas.cz/hajek);  
paper Hájek, Holeňa, Rauch to appear - 114 citations.

**Thank you for your attention!**