Fuzzy hypotheses testing and GUHA implicational quantifiers

Martin Holeňa

University of Paderborn, Department of Computer Science – Cadlab
Bahnhofstraße 32, 33102 Paderborn, Germany
(on leave from ICS Prague, Czech Republic) †

Abstract

This paper presents a fuzzy generalization of the concept of implicational quantifiers, one of the key concepts of a sophisticated approach to exploratory data analysis, the general unary hypotheses automaton (GUHA). The proposed generalization is based on a method of fuzzy hypotheses testing, which was elaborated for the statistical tests used in GUHA, and integrated into the context of an observational calculus. After recalling the basic notions pertaining to implicational quantifiers, and explaining the principles of the method used, the concept of a fuzzy-implicational quantifier is introduced, and several results concerning its properties are presented. The proposed fuzzy approach is illustrated on a particular implicational quantifier, commonly encountered in GUHA.

1 Introduction

The intended paper concerns the use of fuzzy logic for *exploratory data analysis,* i.e. for the kind of data analysis the aim of which is to discover possible decision support and interesting relationships captured in the processed data. A particular approach to exploratory data analysis is the one known under the name *General Unary Hypotheses Automaton* (GUHA).

The method of automatic generating of hypotheses, underlying GUHA, attracted much attention in the late seventies ([2], [3], [5]). First simple implementations of the method emerged very soon after ([6], [8]), followed later by more sophisticated systems ([2], [4], [9]).

Basically, the hypotheses generated by the GUHA approach are sentences of an *observational calculus,* as the model of which a dichotomous *data matrix* is used, viewed as a realization of a random sample of binary vectors. Each component of the vectors, i.e. each column of the matrix, corresponds to some *atomic property* of real objects. Actually, only two particular kinds of sentences are considered in the method ([1], [3]):

(i) A *GUHA implication* captures the fact that some combination of present or absent atomic properties implies another such combination. It is defined by means of an implicational quantifier and concerns the conditional probability of the occurrence of the latter combination, conditioned on the occurrence of the former combination.

(ii) A *GUHA association* captures a general dependence between two combinations of present or absent atomic properties. It is defined by means of an associational quantifier and concerns the stochastic independence of the random variables corresponding to the considered combinations.

Most of the procedures for generating GUHA implications and associations, employed in the existing implementations of the approach, are based on statistical hypotheses testing. However, there is an *inner contradiction* inherent to using common statistical tests to this end. Those tests always require a *precise formulation* of the tested conditions on the random variables characterizing the underlying real phenomena. On the other hand, due to its explorative purpose, GUHA is predominantly used in situations in which the user has only rather *vague knowledge* of the underlying phenomena, thus

[†] This paper was presented on Fuzzy workshop which took place in Kočovce (Slovakia)

being unable to make a competent choice of the conditions to be tested. Therefore, it would be advantageous to reflect the vagueness of the user's a priori knowledge in the way in which sentences in GUHA are generated. To generalize the GUHA approach in that way is the objective of the research reported in the paper. The proposed generalization is based on a method of fuzzy hypotheses testing, which was elaborated for the statistical tests used in GUHA, and integrated into the context of an observational calculus. Due to space limitations and the different nature of GUHA implications and GUHA associations, only the former are dealt with.

2 GUHA implicational quantifiers recalled

Definition 1 Let $n \in \mathcal{N}$ and $t = (t_1,\ldots, t_n) \in \mathcal{N}^n$. For each $m \in \mathcal{N}$, denote $\mathcal{M}(m)$, more precisely $\mathcal{M}_{\{0,1\}}(m)$, the set of all $\{0,1\}$-structures of type 1_m with finite natural domain, i.e.

$$\mathcal{M}(m) = \{(S, f_1, \ldots, f_{m_q}) : S \subset \mathcal{N} \;\&\; |S| < \infty \;\&\; (\forall i \in \{1,\ldots,m\})\, f_i : S \to \{0,1\}\}, \qquad (1)$$

where the symbol $|S|$ stands for the cardinality of a set S.
An *observational predicate calculus* (OPC) of type t is given by

$$(\{P_1,\ldots,P_n,=\}, X, x, J, Q), \text{ where}$$

$\{P_1,\ldots,P_n,=\}$, is a set of *predicates,* $=$ being the identity predicate,

X is an at most countable set of *variables,*

$x \in X$ is a *designated variable,*

$J \overset{\text{def}}{=} \{0, 1, \neg, \&, \vee, \Rightarrow, \Leftrightarrow\}$ is the set of *junctors,*

Q is an at most countable set of *generalized quantifiers,*

and it holds:

(i) $(\forall i \in \{1,\ldots,n\})$ the arity of P_i of is t_i,

(ii) $(\forall q \in Q)(\exists m_q \in \mathcal{N})$ the arity of q is m_q,

(iii) $(\forall q \in Q)$ a unique function $\mathrm{Af}_q : \mathcal{M}(m_q) \to \{0,1\}$ is attached to q, called the *associated function* of q, such that

(3)a) if $M_1, M_2 \in \mathcal{M}(m_q)$ & M_1 and M_2 are isomorphic structures, then $\mathrm{Af}_q(M_1) = \mathrm{Af}_q(M_2)$

(3)b) the function Af defined

$$\mathrm{Af}(q, M) \overset{\text{def}}{=} \mathrm{Af}_q(M) \quad | \quad q \in Q, M \in \mathcal{M}(m_q) \qquad (2)$$

is recursive in both variables.

If in particular $t_1 = \ldots = t_n = 1$, then the OPC is called *monadic.*

Definition 2 Let $\rightarrow$ be a binary generalized quantifier of a monadic OPC, such that its associated function $\mathrm{Af}_\rightarrow$ fulfils

$$(\forall M_1, M_2 \in \mathcal{M}(2))\; a_{M_2} \geq a_{M_1} \;\&\; b_{M_2} \leq b_{M_1} \;\&\; \mathrm{Af}_\rightarrow(M_1) = 1 \Rightarrow \mathrm{Af}_\rightarrow(M_2) = 1, \qquad (3)$$

where for each $M = (S, f_1, f_2) \in \mathcal{M}(2)$,

$$a_M = |\{s : s \in S \;\&\; f_1(s) = f_2(s) = 1\}|, b_M = |\{s : s \in S \;\&\; f_1(s) = 1 \;\&\; f_2(s) = 0\}|. \qquad (4)$$

Then the quantifier $\rightarrow$ is called *implicational.*

Example 1 Let $\alpha \in (0, \frac{1}{2})$, $\theta \in (0,1)$ be given constants. For each $M = (S, f_1, f_2) \in \mathcal{M}(2)$, denote $r_M = a_M + b_M = |\{s : s \in S \;\&\; f_1(s) = 1\}|$. Finally, let $\rightarrow^!$, more precisely $\rightarrow^!_\theta$ be a generalized quantifier defined by means of the associated function

$$\mathrm{Af}_{\to^!_\theta}(M) \overset{\text{def}}{=} \begin{cases} 1 & | \quad M \in \mathcal{M}(2) \;\&\; \sum\limits_{i=a_M}^{r_M} \binom{r_M}{i}\theta^i(1-\theta)^{r_M-i} \leq \alpha \\ 0 & | \quad M \in \mathcal{M}(2) \;\&\; \sum\limits_{i=a_M}^{r_M} \binom{r_M}{i}\theta^i(1-\theta)^{r_M-i} > \alpha. \end{cases} \tag{5}$$

Following the terminology used in [3], $\to^!_\theta$ will be called the quantifier of a ***likely implication*** with a threshold θ.

3 Fuzzy hypotheses for the likely implication

Observe that the associated function of the quantifier $\to^!_\theta$ can be expressed as

$$\mathrm{Af}_{\to^!_\theta}(M) \overset{\text{def}}{=} \begin{cases} 1 & | \quad M \in \mathcal{M}(2) \;\&\; T_{(!)} \leq \alpha \\ 0 & | \quad M \in \mathcal{M}(2) \;\&\; T_{(!)} > \alpha. \end{cases} \tag{6}$$

where $T_{(!)} = \sum\limits_{i=a_M}^{r_M} \binom{r_M}{i}\theta^i(1-\theta)^{r_M-i}$ is a realization of a test statistic used for testing the hypothesis $p \leq \theta$, provided the system $(f_1(s), f_2(s))_{s\in S}$ for $M = (S, f_1, f_2) \in \mathcal{M}(2)$ is viewed as a realization of a twodimensional random sample, and

$$p = \Pr(f_2 = 1 | f_1 = 1). \tag{7}$$

To replace that strictly formulated hypothesis by a more vague and relaxed one, the method of fuzzy hypotheses testing proposed in [10] will be employed. According to that method, a null hypothesis to be tested is viewed as a normaeq lized fuzzy set $\tilde{\Pi}$ on the set of admissible parameters, thus in our case

$$\tilde{\Pi} = \{(p, \mu(p)) : p \in (0,1)\} \text{ with } \mu : \Pi \to \langle 0, 1\rangle. \tag{8}$$

The form of the membership function μ of $\tilde{\Pi}$ reflects the nature of the vagueness captured by the null hypothesis. The following situations are possible:

(i) A particular application may suggest a particular value θ. Then only the requirement that $p \leq \theta$ should hold exactly can be relaxed. Instead, we can consider a null hypothesis paraphrased

$$\textit{p is not much higher than } \theta. \tag{9}$$

In that case, the membership function μ of $\tilde{\Pi}$ could have, e.g., one of the forms shown in fig. 1.

(ii) Very often, the value θ in the formulation $p \leq \theta$ of the tested hypothesis is actually immaterial, and its purpose is solely to prevent p from being too high. Then we can relax also the requirement that p is to be compared to a particular value, and consider a null hypothesis paraphrased

$$\textit{p is low.} \tag{10}$$

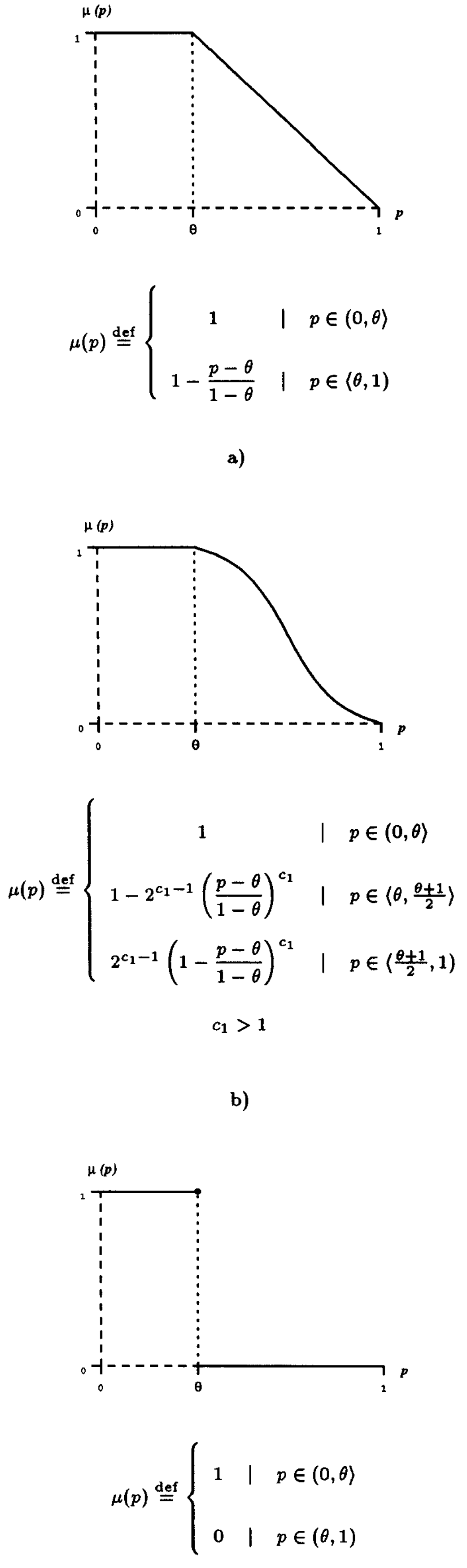
That case is illustrated by the membership functions in fig. 2.

In general, the membership function μ of as fuzzy null hypothesis $\tilde{\Pi}$ corresponding to the quantifier $\to^!$ will be required to fulfil only the following weak conditions:

$$\begin{aligned} &\text{(i)} && \mu \text{ is nonincreasing on } (0,1), \\ &\text{(ii)} && \lim_{p\to 0+} \mu(p) = 1, \\ &\text{(iii)} && \lim_{p\to 1-} \mu(p) = 0. \end{aligned} \tag{11}$$

Observe that these conditions cover both the examples in fig. 1, corresponding to the fuzzy null hypothesis (9), and the examples in fig. 2, corresponding to (10).

Following [10], a test statistic for a fuzzy null hypothesis will be viewed as a random variable assuming values in the space of real functions on the set of parameters, and a critical region for a fuzzy



$$\mu(p) \overset{\text{def}}{=} \begin{cases} 1 & \mid \quad p \in (0,\theta\rangle \\ 1 - \dfrac{p-\theta}{1-\theta} & \mid \quad p \in \langle\theta,1) \end{cases}$$

a)

$$\mu(p) \overset{\text{def}}{=} \begin{cases} 1 & \mid \quad p \in (0,\theta\rangle \\ 1 - 2^{c_1-1}\left(\dfrac{p-\theta}{1-\theta}\right)^{c_1} & \mid \quad p \in \langle\theta,\frac{\theta+1}{2}\rangle \\ 2^{c_1-1}\left(1 - \dfrac{p-\theta}{1-\theta}\right)^{c_1} & \mid \quad p \in \langle\frac{\theta+1}{2},1) \end{cases}$$

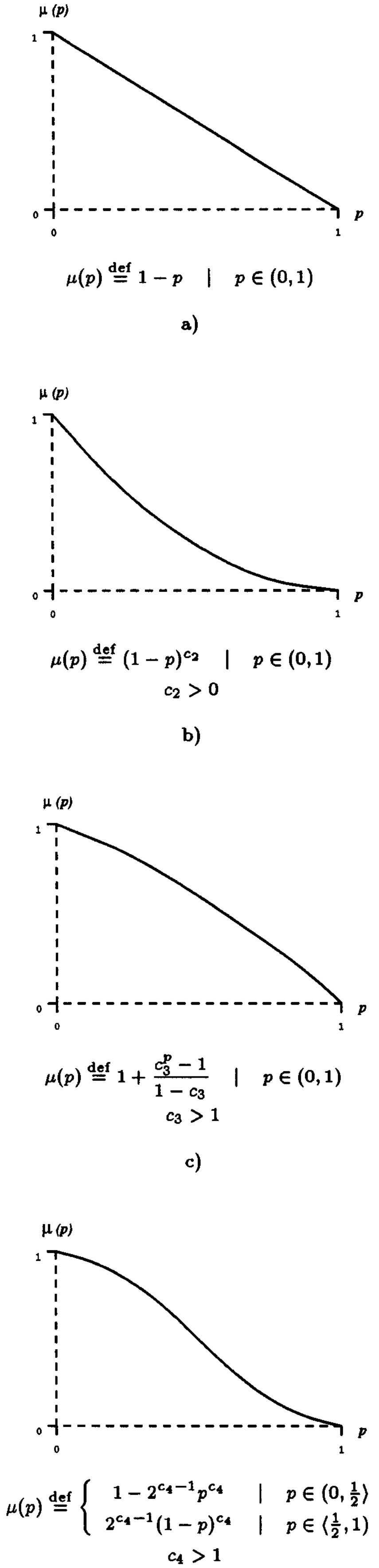
$c_1 > 1$

b)

$$\mu(p) \overset{\text{def}}{=} \begin{cases} 1 & \mid \quad p \in (0,\theta\rangle \\ 0 & \mid \quad p \in (\theta,1) \end{cases}$$

c)

Figure 1



$$\mu(p) \overset{\text{def}}{=} 1 - p \quad \mid \quad p \in (0,1)$$

a)

$$\mu(p) \overset{\text{def}}{=} (1-p)^{c_2} \quad \mid \quad p \in (0,1)$$

$c_2 > 0$

b)

$$\mu(p) \overset{\text{def}}{=} 1 + \frac{c_3^p - 1}{1 - c_3} \quad \mid \quad p \in (0,1)$$

$c_3 > 1$

c)

$$\mu(p) \overset{\text{def}}{=} \begin{cases} 1 - 2^{c_4-1}p^{c_4} & \mid \quad p \in (0,\frac{1}{2}\rangle \\ 2^{c_4-1}(1-p)^{c_4} & \mid \quad p \in \langle\frac{1}{2},1) \end{cases}$$

$c_4 > 1$

d)

Figure 2

null hypothesis will be viewed as a mapping assigning to each significance level and each parameter a Borel set of reals. In the particular case of the quantifier $\rightarrow^{!}$, we get the test statistic defined

$$T_{(!)}(p) \overset{\text{def}}{=} \sum_{i=a}^{r_M} \binom{r_M}{i} p^i (1-p)^{r_M - i} \quad | \quad p \in (0,1), \tag{12}$$

and the critical region

$$K(\alpha, p) \overset{\text{def}}{=} (0, \alpha\rangle \quad | \quad \alpha \in (0, \tfrac{1}{2}\rangle, p \in (0,1). \tag{13}$$

4 Fuzzy-implicational quantifiers

Definition 3 Let $\rightarrow$ be a binary generalized quantifier of a monadic OPC, $n \in \mathcal{N}$, $\Pi \subset \Re^n$ be a set of parameters, and $\tilde{\Pi}$ be a particular normalized fuzzy set on Π.
Then $\rightarrow$ is called *fuzzy-implicational* w.r.t. $\tilde{\Pi}$ iff in addition to the associated function $\mathrm{Af}_{\rightarrow}$, a parametrized associated function $\mathrm{Paf}_{\rightarrow} : \mathcal{M}(2) \otimes \Pi \rightarrow \{0,1\}$ is attached to it, with the following properties:

(i) $(\exists \pi \in \Pi)\ \mathrm{Paf}_{\rightarrow}(\cdot, \pi) = \mathrm{Af}_{\rightarrow}$,

(ii) the functions $\mathrm{Faf}^{a}_{\rightarrow} : \mathcal{M}(2) \rightarrow \langle 0,1\rangle$ and $\mathrm{Faf}^{r}_{\rightarrow} : \mathcal{M}(2) \rightarrow \langle 0,1\rangle$, defined

$$\mathrm{Faf}^{a}_{\rightarrow}(M) \overset{\text{def}}{=} \begin{cases} \sup\{\mu(\pi) : \pi \in \Pi \ \&\ \mathrm{Paf}_{\rightarrow}(M,\pi) = 1\} & | \begin{array}{c} M \in \mathcal{M}(2)\ \& \\ (\exists \pi \in \Pi)\ \mathrm{Paf}_{\rightarrow}(M,\pi) = 1 \end{array} \\ 0 & | \begin{array}{c} M \in \mathcal{M}(2)\ \& \\ (\forall \pi \in \Pi)\ \mathrm{Paf}_{\rightarrow}(M,\pi) = 0, \end{array} \end{cases} \tag{14}$$

$$\mathrm{Faf}^{r}_{\rightarrow}(M) \overset{\text{def}}{=} \begin{cases} 1 - \sup\{\mu(\pi) : \pi \in \Pi \ \&\ \mathrm{Paf}_{\rightarrow}(M,\pi) = 0\} & | \begin{array}{c} M \in \mathcal{M}(2)\ \& \\ (\exists \pi \in \Pi)\ \mathrm{Paf}_{\rightarrow}(M,\pi) = 0 \end{array} \\ 1 & | \begin{array}{c} M \in \mathcal{M}(2)\ \& \\ (\forall \pi \in \Pi)\ \mathrm{Paf}_{\rightarrow}(M,\pi) = 1, \end{array} \end{cases} \tag{15}$$

fulfil the condition

$$\begin{gathered} (\forall M_1, M_2 \in \mathcal{M}(2))\ a_{M_2} \geq a_{M_1} \ \&\ b_{M_2} \leq b_{M_1} \Rightarrow \\ \Rightarrow \mathrm{Faf}^{a}_{\rightarrow}(M_2) \geq \mathrm{Faf}^{a}_{\rightarrow}(M_1) \ \&\ \mathrm{Faf}^{r}_{\rightarrow}(M_2) \geq \mathrm{Faf}^{r}_{\rightarrow}(M_1). \end{gathered} \tag{16}$$

The function $\mathrm{Faf}^{a}_{\rightarrow}$ will be called an *accepting fuzzy associated function* of $\rightarrow$ w.r.t. $\tilde{\Pi}$, and $\mathrm{Faf}^{r}_{\rightarrow}$ will be called a *rejecting fuzzy associated function* of $\rightarrow$ w.r.t. $\tilde{\Pi}$.

The fuzzy-implicationality of quantifiers is a generalization of their implicationality, in the sense that for $\tilde{\Pi}$ corresponding to a one-element crisp set, definition 3 coincides with definition 2. Moreover, both the accepting fuzzy associated function and the rejecting fuzzy associated function of a binary generalized quantifier coincide with its usual ($\{0,1\}$-valued) associated function in that case.

Theorem 1 Let $M \in \mathcal{M}(2)$, a_M and r_M be the numbers assigned to M according to (4), and $T_{(!)}$ be a realization of the test statistic introduced in (12), obtained through replacing the random variable a with its realization a_M. Let further α be the constant from (5), and K be the critical region introduced in (13). Finally, let $\tilde{\Pi}$ be a normalized fuzzy set on $(0,1)$, such that for its membership function μ the conditions (11) are valid.
Then the generalized quantifier $\rightarrow^{!}$ is fuzzy-implicational w.r.t. $\tilde{\Pi}$, and its fuzzy associated functions w.r.t. $\tilde{\Pi}$ fulfil

$$\mathrm{Faf}^{a}_{\rightarrow^{!}}(M) = 1, \tag{17}$$

$$a_M > 0 \Rightarrow (\exists p_M > \in (0,1))\ T^{-1}_{(!)}(\alpha) = \{p_M\} \ \&\ \mathrm{Faf}^{r}_{\rightarrow^{!}}(M) = 1 - \lim_{p \rightarrow p_M^{+}} \mu(p). \tag{18}$$

If in addition μ is right-continuous on $(0,1)$, (18) simplifies to

$$a_M > 0 \Rightarrow (\exists p_M > \in (0,1))\ T^{-1}_{(!)}(\alpha) = \{p_M\} \ \&\ \mathrm{Faf}^{r}_{\rightarrow^{!}}(M) = 1 - \mu(p_M). \tag{19}$$

A proof of this theorem, as well as a proof of the theorem 2 below, can be found in [7].

Observe that the null hypothesis $p \leq \theta$, corresponding to the likely implication, can be viewed as a fuzzy null hypothesis with the membership function from figure 1c). Though this is not a membership function of a one-element set, the following theorem shows that using it implies a coincidence between the usual associated function of $\rightarrow^!$, and its nonconstant fuzzy associated function $\mathrm{Faf}^r_{\rightarrow^!}$. Consequently, testing the hypothesis $p \leq \theta$ is equivalent to testing the fuzzy hypothesis corresponding to the crisp set $(0, \theta\rangle$. Similar results can be obtained also for other commonly encountered implicational quantifiers.

Theorem 2 Let $\theta \in (0, 1)$, and p_M for $M \in \mathcal{M}(2)$ be the parameter value introduced in the theorem 1. Denote

$$\tilde{\Pi}_{(!)} \overset{\text{def}}{=} \{(p, \mu(p)) : p \in (0,1) \,\&\, \mu|(0,\theta\rangle = 1 \,\&\, \mu|(\theta, 1) = 0\}. \tag{20}$$

Then for the quantifier $\rightarrow^! = \rightarrow^!_\theta$, the rejecting fuzzy associated function w.r.t. $\tilde{\Pi}_{(!)}$ fulfils

$$(\forall M \in \mathcal{M}(2))\ \mathrm{Faf}^r_{\rightarrow^!}(M) = \mathrm{Af}_{\rightarrow^!}(M) = \begin{cases} 1 & \text{iff} \quad p_M \geq \theta \\ 0 & \text{iff} \quad p_M < \theta. \end{cases} \tag{21}$$

5 Conclusion

The results presented in this paper show that the usual method of generating GUHA implications is actually a special case of the proposed fuzzy approach, i.e. that approach can be indeed considered a fuzzy generalization of the traditional GUHA. However, it is only the basic principle of this generalization which is outlined here. To become practically applicable, the approach must be extended to cover a number of other concepts and issues.

Two such extensions, to GUHA associations and to the concept of a power function, are actually already the matter of an ongoing research. Intended future extensions should include at least foundedness, restricted sentences, and missing data. From the effectiveness point of view, it is important to extend the approach to the concepts of hopeless antecedents and succedents, and to the notion of improving predicates.

References

[1] P. Hájek. The new version of the GUHA procedure ASSOC (generating hypotheses on associations) – mathematical foundations. In *COMPSTAT 1984 – Proceedings in Computational Statistics*, pages 360–365, 1984.

[2] P. Hájek and T. Havránek. On generating of inductive hypotheses. *International Journal of Man Machine Studies*, 9:415–438, 1977.

[3] P. Hájek and T. Havránek. *Mechanizing Hypothesis Formation*. Springer-Verlag, Berlin, 1978.

[4] P. Hájek, A. Sochorová, and J. Zvárová. GUHA for personal computers. *Statistical Software Newsletters*, 1994.

[5] T. Havránek. An alternative approach to missing information in the GUHA method. *Kybernetika*, 16:145–155, 1980.

[6] T. Havránek. Some comments on GUHA procedures. In *Exploratory Data Analysis*. Springer-Verlag, Berlin, 1980.

[7] M. Holeňa. Fuzzy hypotheses for GUHA implications. Submitted for publication, 1995.

[8] J. Rauch. Some remarks on computer realization of the GUHA method. *International Journal of Man Machine Studies*, 10:75–86, 1978.

[9] A. Sochorová and T. Havránek. A new version of the GUHA method for analysing categorial data. In *Computers in Medicine and Health Care*, pages 169–170, 1990.

[10] N. Watanabe and T. Imaizumi. A fuzzy statistical test of fuzzy hypotheses. *Fuzzy Sets and Systems*, 53:167–178, 1993.