Relevance Test for Fuzzy Rules

A. Krone
Department of Control Engineering
(Prof. Dr. H. Kiendl)
Faculty of Electrical Engineering
University of Dortmund

H. Taeger
Department of Statistics
(Prof. Dr. Hering, Dr. Poehlmann)
Univerity of Dortmund

1 Introduction

Applying data-based fuzzy modelling methods to industrial applications, the following points have to be considered:

- There are often many possible input quantities resulting in enormous search spaces and rule bases.
- The learning data are often contradictory.
- On the part of the industrial operators, interpretable results are desired that allow insight.

A relevance test, that decides whether an IF/THEN statement represents a relevant aspect of the dependencies between the input and output variables, can help in these points [17,20]:

- It allows to break down the problem of finding an appropriate rule base to the problem of finding single relevant rules. The rule base is then generated by the incremental collection of the relevant rules [16,21]. Additionally, the relevance test allows to decide if several input situations can be combined in one premise so that the final rule base is preferably small [15,19].
- It can handle contradictions in the data by using statistical calculations [18,22].
- It supports the interpretability of the individual rules. If only the input/output behaviour was optimized, the inference of the rules would model the output values but the individual rules could be senseless.

Such a relevance test has been introduced by Kiendl and Krabs [12–14] for crisp rule-based modelling. Its idea is presented in Section 2. For fuzzy modelling the algorithm can be extended to the use of fuzzy values between 0 and 1 (Section 3) [13]. By this, the statistical verification gets lost. Therefore, a fuzzy relevance test is developped in Section 4. In Section 5, comparative results are shown.

Beside the here proposed relevance test, other methods of separating between desired and not desired rules may be appropriate depending on the application, for example separating by the hit rate of a rule [9]. The here proposed relevance test aims at handling situations where we are basically interested in detecting causal relations.

2 Crisp Relevance Test

A statement of the following form should be examined:

IF S THEN C

S represents an input situation, C an output event. The input situation, resp. the output event is true or not true. The corresponding characteristic functions $\mathbf{I}_S$ and $\mathbf{I}_C$ take the value 1 if the input situation, resp. the output event is true and the value 0 if the input situation, resp. the output event is not true:

$$\mathbf{I}_S(X(k)) = \begin{cases} 1 & : \quad S \text{ is true for the data sample } d_k \\ 0 & : \quad S \text{ is not true for the data sample } d_k \end{cases}$$

$$\mathbf{I}_C(Y(k)) = \begin{cases} 1 & : \quad C \text{ is true for the data sample } d_k \\ 0 & : \quad C \text{ is not true for the data sample } d_k \end{cases}$$

$X = (X_1, X_2, ...)$ is the vector with the input variables. Y is the output variable. The premise and the conclusion refer to the same data sample $d_k = (x_1(k), x_2(k), ..., y(k)) = (x(k), y(k))$. It includes the realizations of $X(k)$, $Y(k)$ belonging together, for example the observations at the date k.

Example:
X_1 is the heating temperature, X_2 is the outdoor temperature, Y is the room temperature. There are five data samples ($n = 5$) with $d_1 = (50^o\text{C}, -10^o\text{C}, 14^o\text{C})$, $d_2 = (35^o\text{C}, -5^o\text{C}, 12^o\text{C})$, $d_3 = (25^o\text{C}, -2^o\text{C}, 8^o\text{C})$, $d_4 = (20^o\text{C}, -5^o\text{C}, 5^o\text{C})$, $d_5 = (25^o\text{C}, -10^o\text{C}, 3^o\text{C})$. A possible IF/THEN statement is: IF ((heating temperature is lower than 30^oC) $\wedge$ (outdoor temperature is lower than 0^oC)) THEN (room temperature is under 10^oC). This corresponds to: IF ($X_1 < 30^o\text{C} \wedge X_2 < 0^o\text{C}$) THEN ($Y < 10^o\text{C}$). The characteristic functions are defined by

$$\mathbf{I}_S(X(k)) = \begin{cases} 1 & : \quad X_1(k) < 30^o\text{C} \wedge X_2(k) < 0^o\text{C} \\ 0 & : \quad \text{else} \end{cases} \qquad \mathbf{I}_C(Y(k)) = \begin{cases} 1 & : \quad Y(k) < 10^o\text{C} \\ 0 & : \quad \text{else} \end{cases}$$

For the five data samples the following values result: $\mathbf{I}_S(x(1)) = 0$, $\mathbf{I}_S(x(2)) = 0$, $\mathbf{I}_S(x(3)) = 1$, $\mathbf{I}_S(x(4)) = 1$, $\mathbf{I}_S(x(5)) = 1$, $\mathbf{I}_C(y(1)) = 0$, $\mathbf{I}_C(y(2)) = 0$, $\mathbf{I}_C(y(3)) = 1$, $\mathbf{I}_C(y(4)) = 1$, $\mathbf{I}_C(y(5)) = 1$.

The probability that the output event is true ($\mathbf{I}_C(Y(k)) = 1$) is $p = P(C)$. The conditional probability that the output event C is true under the condition that the input situation S is true ($\mathbf{I}_C(Y(k)) = 1 | \mathbf{I}_S(X(k)) = 1$) is $p_\lambda = P(C|S)$. The more these two probabilities differ the more the IF/THEN statement can be seen as relevant.

As these probabilities are not known, they are estimated on the basis of the data samples d_k by the relative frequencies:

$$\hat{p} = \frac{m}{n} \quad \text{and} \quad \hat{p}_\lambda = \frac{m_\lambda}{n_\lambda}$$

with

$n :=$ number of data samples d_k,
$m := \sum_{k=1}^{n} \mathbf{I}_C(Y(k))$,
$n_\lambda := \sum_{k=1}^{n} \mathbf{I}_S(X(k))$,
$m_\lambda := \sum_{k=1}^{n} (\mathbf{I}_S(X(k)) \wedge \mathbf{I}_C(Y(k)))$.

For $X(1), ..., X(n)$ independent identically distributed (i.i.d.) and $Y(1), ..., Y(n)$ i.i.d. it can be proven that $\hat{p}$ and $\hat{p}_\lambda$ are consistent and uniformly minimal-variance unbiased estimators [13, 28].

$\mathbf{I}_S(X)$ resp. $\mathbf{I}_C(Y)$ are Bernoulli distributed with the parameter p_λ resp. p. On this basis, confidence intervals can be calculated for p and p_λ with the Pearson–Clopper–values [7]. They cover the probabilities p and p_λ each with a given probability $1 - \alpha$ (confidence coefficient).

As only one side of the confidence intervals is interesting in each relevance test, either the one-sided confidence intervals $I^o := [0; p^o]$ and $I^u_\lambda := [p^u_\lambda; 1]$ or $I^o_\lambda := [0; p^o_\lambda]$ and $I^u := [p^u; 1]$ are calculated.

In the case

$$\hat{p} < \hat{p}_\lambda \ \wedge \ p^o < p^u_\lambda$$

the statement 'IF S THEN C' is a **positive relevant rule**. In the case

$$\hat{p} > \hat{p}_\lambda \ \wedge \ p^u > p^o_\lambda$$

the negative statement 'IF S THEN $\neg C$' is a **negative relevant rule** [10]. In all other cases is

$$[p^u; p^o] \cap [p^u_\lambda; p^o_\lambda] \neq \emptyset$$

and thus no relevant rule can be extracted.

After the relevance test the relevant rules can be assigned a rating index [21, 23].

Example:
There are 100 data samples ($n = 100$). The output event C is true in 50 of the 100 data samples ($m = 50$), so that $\hat{p} = 0.5$. The input situation S is true in 20 of the 100 data samples ($n_\lambda = 20$). In 18 of the 20 data samples the output event C is true ($m_\lambda = 18$), so that $\hat{p}_\lambda = 0.9$. As $\hat{p} < \hat{p}_\lambda$, the interval borders p^o and p^u_λ have to be calculated. With a confidence coefficient of 0.95 one gets $p^o = 0.586$ and $p^u_\lambda = 0.717$. The result is visualized in Figure 1. The statement 'IF S THEN C' is a positive relevant rule as the confidence intervals do not intersect. The results for all possible values of m_λ $(0, 1, 2, ..., 20)$ are visualized in Figure 2.

Figure 1: Confidence interval borders for $n = 100$, $m = 50$, $n_\lambda = 20$ and $m_\lambda = 18$

Figure 2: Confidence interval borders for $n = 100$, $m = 50$, $n_\lambda = 20$ and $m_\lambda = 0, 1, 2, ..., 20$

3 Algorithmic Extension of the Crisp Relevance Test

If the input situation and the output event are described by fuzzy sets [11, 29], they can not only be true or not true, but true to a certain degree, normally between 0 and 1. Thus, the characteristic functions $\mathbf{I}_S(X(k))$ and $\mathbf{I}_C(Y(k))$ are substituted by the membership functions:

$$\mu_S(X(k)) \in [0;1] \quad \text{and} \quad \mu_C(Y(k)) \in [0;1].$$

Example:
A crisp definition of the output event C ='room temperature is low' can be

$$\mathbf{I}_C(Y(k)) = \begin{cases} 1 & : \quad Y(k) \in [16^o\mathrm{C}; 19^o\mathrm{C}] \\ 0 & : \quad Y(k) \notin [16^o\mathrm{C}; 19^o\mathrm{C}] \end{cases}$$

A fuzzy definition of the same output event can be

$$\mu_C(Y(k)) = \begin{cases} 0 & : \quad Y(k) < 15^o\mathrm{C} \\ 0.5Y(k) - 7.5 & : \quad Y(k) \in [15^o\mathrm{C}; 17^o\mathrm{C}[\\ 1 & : \quad Y(k) \in [17^o\mathrm{C}; 18^o\mathrm{C}[\\ -0.5Y(k) + 10 & : \quad Y(k) \in [18^o\mathrm{C}; 20^o\mathrm{C}[\\ 0 & : \quad Y(k) \geq 20^o\mathrm{C} \end{cases}$$

Figure 3 illustrates the crisp and fuzzy definition. For example, for $y(1) = 13^o\mathrm{C}$, $y(2) = 15.5^o\mathrm{C}$, $y(3) = 16^o\mathrm{C}$, $y(4) = 17.5^o\mathrm{C}$, $y(5) = 21^o\mathrm{C}$ the degrees of membership are $\mu_C(y(1)) = 0$, $\mu_C(y(2)) = 0.25$, $\mu_C(y(3)) = 0.5$, $\mu_C(y(4)) = 1$, $\mu_C(y(5)) = 0$.

Figure 3: Crisp (a) and fuzzy (b) definition of the event 'room temperature is low'

In the case of fuzzy input situations and fuzzy output events, the formulas of the crisp relevanve test can be extended algorithmicly from integer to real values [13]. This extension is statistically not justified as the $\mu_S(X)$ resp. $\mu_C(Y)$ are no longer Bernoulli distributed. Nevertheless, one gets a kind of interpolating solution that calculates the correct statistical values in the special case of crisp fuzzy sets (characteristic functions).

Then, the estimators are given by

$$\hat{p} = \frac{m}{n} \quad \text{and} \quad \hat{p}_\lambda = \frac{m_\lambda}{n_\lambda}$$

with

$n :=$ number of data samples d_k,
$m := \sum_{k=1}^{n} \mu_C(Y(k))$,
$n_\lambda := \sum_{k=1}^{n} \mu_S(X(k))$,
$m_\lambda := \sum_{k=1}^{n} (\mu_S(X(k)) \wedge \mu_C(Y(k)))$.

The '$\wedge$' operator could be realized by one of the numerous fuzzy AND operators [24]. Reasonably, it should be that one that is also used to calculate $\mu_S(x(k))$ from the individual degrees of activation of the different input values $x_i(k)$. For concrete decision see Section 4.1.

The real values m, n_λ and m_λ are inserted in the formulas for the calculation of the confidence intervals though the formulas are only defined for integer values of m, n_λ and m_λ. In Figure 4 the interpolation between the crisp values is examplarily shown for the interval border p_λ^u.

Figure 4: Interpolating values (represented by small circles '$\cdot$') and crisp values (represented by big circles '$\bullet$'), examplarily shown for the interval border p_λ^u: $\bullet \cdots \bullet$ $n_\lambda = 20$ ($m_\lambda = 0, 0.25, 0.5, 0.75, 1, 1.25, ..., 20$), $\cdots$ $n_\lambda = 32.5$ ($m_\lambda = 0, \frac{32.5}{80}, \frac{2 \cdot 32.5}{80}, \frac{3 \cdot 32.5}{80}, ..., 32.5$), $\bullet \cdot \bullet$ $n_\lambda = 40$ ($m_\lambda = 0, 0.5, 1, 1.5, 2, ..., 40$), $\bullet \bullet \bullet$ $n_\lambda = 80$ ($m_\lambda = 0, 1, 2, 3, ..., 80$)

4 Fuzzy Relevance Test

Alternatively to the algorithmic extension of the crisp relevance test, in this Section, a relevance test for fuzzy rules is developped from the statistical point of view. It examines rules of the form

$$\text{IF } S \text{ THEN } C$$

where S represents a fuzzy input situation described by the membership function $\mu_S(X)$ and C a fuzzy output event described by the membership function $\mu_C(Y)$.

Following the idea of the crisp relevance test, first, adaquate probabilities and estimators [23] have to be defined (Section 4.1). Afterwards a method for calculating the confidence intervals has to be developed. As, in contrast to the crisp case, the distribution of $\mu_S(X)$ resp. $\mu_C(Y)$ are not known, an exact parametric calculation of the confidence intervals is not possible. Two different approaches can still be made:

- a non-parametric calculation,
- an asymptotic calculation.

In Section 4.2 the first approach is pursued by using a Bootstrap method for the calculation of the confidence intervals [27]. In Section 4.3 the second approach is examined [27].

4.1 Probabilities and Estimators for Fuzzy Events

Zadeh [30] defines the probability of a fuzzy event A by

$$P(A) = \int_A f(z)dz = \int_{\mathbf{R}} \mu_A(z) f(z)dz = E[\mu_A(Z)]$$

with

Z: random variable,
$f(z)$: density of Z,
$\mu_A(Z)$: membership function for the fuzzy event A,
$E[\cdot]$: the expected value.

Other authors have seized that suggestion [2,26]. In [26] it is proven that the Kolmogoroff axioms of a probability are fulfilled for finite event spaces.

On this basis, the probability of the fuzzy output event C is:

$$P(C) = E[\mu_C(Y)].$$

The conditional probability of the fuzzy output event C under the fuzzy situation S is:

$$P(C|S) = \frac{P(C \cap S)}{P(S)} = \frac{E[\mu_{C\cap S}(Y,X)]}{E[\mu_S(X)]} = \frac{E[\mu_C(Y) \wedge \mu_S(X)]}{E[\mu_S(X)]} \quad \text{with} \quad P(S) \neq 0.$$

Only the algebraic product as '$\wedge$' operator

$$\mu_C(Y) \wedge \mu_S(X) = \mu_C(Y)\mu_S(X)$$

makes sense in the field of probabilities as it is the only operator that can fulfill the two following statistical equations [1]:

1. $P(C \cap S) + P(\bar{C} \cap S) = P(S)$
2. $P(C \cap S) = P(C)P(S)$ if C and S are independent fuzzy events

The compliment is defined by $\mu_{\bar{C}}(Y) = 1 - \mu_C(Y)$.

An estimator for the probability $P(C)$ is [1,2]:

$$\hat{p} = \frac{1}{n}\sum_{k=1}^{n} \mu_C(Y(k)) = \frac{m}{n}.$$

An estimator for the probability $P(C|S)$ is:

$$\hat{p}_\lambda = \frac{\sum_{k=1}^{n}(\mu_C(Y(k))\mu_S(X(k)))}{\sum_{k=1}^{n}\mu_S(X(k))} = \frac{m_\lambda}{n_\lambda}.$$

For $\mu_C(Y(1)), ..., \mu_C(Y(n))$ i.i.d. and $\mu_S(X(1)), ..., \mu_S(X(n))$ i.i.d. it can be proven that $\hat{p}$, m_λ and n_λ are consistent and unbiased estimators [27]. They can be interpreted as average degrees of membership.

Comparing these estimators with those of the algorithmic generalization of the crisp relevance test for fuzzy values, it can be seen that the formulas of the estimators are identical if the algebraic product is chosen as '$\wedge$' operator.

4.2 Bootstrap Method

The Bootstrap methods are resampling methods suggested by Efron [3–5]. Among other applications, they can serve to calculate confidence intervals. The name Bootstrap comes from the English version of the 'Baron von Muenchhausen', about whom is told that he has pulled himself out of a swamp at his bootstraps. For the relevance test the non-parametric Bootstrap method $\mathrm{BC_a}$ (bias-corrected and accelerated) [4,8] is used [27].

From the n data samples $d_k = (x_1(k), x_2(k), ..., y(k))$ represented by $(d_1, d_2, ..., d_k, ..., d_n)$ w random samples of the size n (called Bootstrap samples) are drawn with replacement:

$$\begin{gathered}(d_1^{*(1)}, d_2^{*(1)}, ..., d_n^{*(1)}) \\ (d_1^{*(2)}, d_2^{*(2)}, ..., d_n^{*(2)}) \\ \vdots \\ (d_1^{*(w)}, d_2^{*(w)}, ..., d_n^{*(w)})\end{gathered}$$

For each Bootstrap sample the estimators for $P(C)$ and $P(C|S)$ are calculated:

$$\begin{gathered}\hat{p}^{*(1)}, \hat{p}_\lambda^{*(1)} \\ \hat{p}^{*(2)}, \hat{p}_\lambda^{*(2)} \\ \vdots \\ \hat{p}^{*(w)}, \hat{p}_\lambda^{*(w)}\end{gathered}$$

The Bootstrap replications $\hat{p}^{*(1)}, ..., \hat{p}^{*(w)}$ and $\hat{p}_\lambda^{*(1)}, ..., \hat{p}_\lambda^{*(w)}$ are sorted in ascending order. Then, the borders of the one-sided confidence intervals are the following:

$$\begin{aligned}p^u &= \hat{p}^{*(g_u)} \;\; (= \text{the } g_u\text{th smallest value of } \hat{p}^{*(1)}, ..., \hat{p}^{*(w)}) \\ p^o &= \hat{p}^{*(g_o)} \;\; (= \text{the } g_o\text{th smallest value of } \hat{p}^{*(1)}, ..., \hat{p}^{*(w)}) \\ p_\lambda^u &= \hat{p}_\lambda^{*(g_{\lambda u})} \;\; (= \text{the } g_{\lambda u}\text{th smallest value of } \hat{p}_\lambda^{*(1)}, ..., \hat{p}_\lambda^{*(w)}) \\ p_\lambda^o &= \hat{p}_\lambda^{*(g_{\lambda o})} \;\; (= \text{the } g_{\lambda o}\text{th smallest value of } \hat{p}_\lambda^{*(1)}, ..., \hat{p}_\lambda^{*(w)})\end{aligned}$$

with

$g_u := \text{trunc}(\beta_u(w+1))$,
$g_o := \text{trunc}(\beta_o(w+1))$,
$g_{\lambda u} := \text{trunc}(\beta_{\lambda u}(w+1))$,
$g_{\lambda o} := \text{trunc}(\beta_{\lambda o}(w+1))$,
$\text{trunc}(v) :=$ whole-numbered part of v.

The β values are calculated by the distribution function Φ of the standard normal distribution:

$$\beta_u = \Phi(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(\alpha)})})$$

$$\beta_o = \Phi(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(1-\alpha)})})$$

$$\beta_{\lambda u} = \Phi(\hat{z}_{\lambda 0} + \frac{\hat{z}_{\lambda 0} + z^{(\alpha)}}{1 - \hat{a}_\lambda(\hat{z}_{\lambda 0} + z^{(\alpha)})})$$

$$\beta_{\lambda o} = \Phi(\hat{z}_{\lambda 0} + \frac{\hat{z}_{\lambda 0} + z^{(1-\alpha)}}{1 - \hat{a}_\lambda(\hat{z}_{\lambda 0} + z^{(1-\alpha)})})$$

with

$z^{(\alpha)}$, $z^{(1-\alpha)}$:= α-, $(1-\alpha)$-quantile of the standard normal distribution,
$1-\alpha$:= confidence coefficient,
$\hat{z}_0$, $\hat{z}_{\lambda 0}$:= bias parameters,
$\hat{a}$, $\hat{a}_\lambda$:= acceleration parameters.

The bias parameters are calculated by the following quantiles of the standard normal distribution

$$\hat{z}_0 = z^{\left(\frac{r}{w}\right)}$$
$$\hat{z}_{\lambda 0} = z^{\left(\frac{r_\lambda}{w}\right)}$$

with

r := number of Bootstrap replications $\hat{p}^{*(\cdot)}$ that are lower than $\hat{p}$,
r_λ := number of Bootstrap replications $\hat{p}_\lambda^{*(\cdot)}$ that are lower than $\hat{p}_\lambda$.

The acceleration parameters are calculated by

$$\hat{a} = \frac{\sum_{l=1}^{n} \left(\hat{p}^{(-)} - \hat{p}^{(l)}\right)^3}{6\left[\sum_{l=1}^{n} \left(\hat{p}^{(-)} - \hat{p}^{(l)}\right)^2\right]^{3/2}}$$
$$\hat{a}_\lambda = \frac{\sum_{l=1}^{n} \left(\hat{p}_\lambda^{(-)} - \hat{p}_\lambda^{(l)}\right)^3}{6\left[\sum_{l=1}^{n} \left(\hat{p}_\lambda^{(-)} - \hat{p}_\lambda^{(l)}\right)^2\right]^{3/2}}$$

with

$\hat{p}^{(l)}$, $\hat{p}_\lambda^{(l)}$:= estimators on the basis of the lth Jacknife sample $(d_1, ..., d_{l-1}, d_{l+1}, ..., d_n)$,
$\hat{p}^{(-)} := \frac{1}{n}\sum_{l=1}^{n} \hat{p}^{(l)}$,
$\hat{p}_\lambda^{(-)} := \frac{1}{n}\sum_{l=1}^{n} \hat{p}_\lambda^{(l)}$.

The $\mathrm{BC_a}$ confidence intervals are second-order accurate [25].

In this context, an essential drawback of the Bootstrap method is the necessary calculating time, as a minimum number of Bootstrap samples for the calculation of confidence intervals is $w = 1000$ [5]. Consequently, for high values of n and a high number of IF/THEN statements the method is not practicable.

Diagrams like in Figure 2 are not possible for the Bootstrap method as the results are depending on the concrete data samples. Results are examplarily calculated in Section 5.

4.3 Asymptotic Calculation of the Confidence Intervals

The conventional distribution functions are not adequate for $\mu_S(X)$ and $\mu_C(Y)$. The beta distribution comes nearest, as it has values between 0 and 1. But it ignores that $\mu_S(X)$ and $\mu_C(Y)$ are partly descretely (0, 1) and partly continuously (]0; 1[) distributed. Nevertheless, one could calculate confidence intervals for $E[\mu_C(Y)]$ that is distributed according to the sum of beta distributed variables. However, the resulting distribution of the quotient $\frac{E[\mu_C(Y)\mu_S(X)]}{E[\mu_S(X)]}$ can not be derived easily, so that confidence intervals for the conditional probability can not be calculated.

Another possibility is to assume forthwith a distribution for $E[\mu_C(Y)$ instead for $\mu_C(Y)$. According to the central limit theorem [7], the distribution of the sum of any distributed variables converges to a normal distribution for n converging to infinity (under easily fulfillable assumptions), so it can be shown that the following is valid:

$$\frac{\frac{1}{n}\sum_{k=1}^{n}\left(\mu_C(Y(k)) - E[\mu_C(Y(k))]\right)}{\sqrt{VAR[\mu_C(Y(k))]}}\sqrt{n} \quad \overset{n\to\infty}{\sim} \quad N(0,1).$$

with

$n :=$ number of data samples d_k,
$VAR[\cdot] :=$ variance,
$N(0,1) :=$ standard normal distribution.

An approximation to a normal distribution can already be obtained for smaller values of n.

Example: In Figure 5 approximations for $n = 20$ and $n = 40$ are exemplarily shown. First, 250 samples with $n = 20$ are taken out of 35000 measured data values. For each sample $\sum_{k=1}^{n}\mu_C(y(k))$ is calculated. Here, the output event C is described by a triangular membership function. Afterwards, the same is done with $n = 40$. For both cases histograms are calculated and normal distributions are adjusted. The approximation for $n = 40$ is better than for $n = 20$ as expected.

As a conclusion, p^u and p^o can be calculated asymptotically for $E[\mu_C(Y)]$ for $Y(1), ..., Y(n)$ i.i.d. by

$$\begin{aligned} p^u &= \max\left\{0, \frac{m}{n} - \frac{t_{n-1;1-\alpha}}{\sqrt{n}}S_C\right\} \\ p^o &= \min\left\{\frac{m}{n} + \frac{t_{n-1;1-\alpha}}{\sqrt{n}}S_C, 1\right\} \end{aligned}$$

with

$n :=$ number of data samples d_k,
$t_{n-1;1-\alpha} := (1-\alpha)$ quantile of the t distribution with $(n-1)$ degrees of freedom,
$1-\alpha :=$ confidence coefficient,
$S_C := \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}\left(\mu_C(Y(k) - \frac{m}{n}\right)^2}$ (estimator for standard deviation),
$m := \sum_{k=1}^{n}\mu_C(Y(k))$.

(a) 20 data samples (b) 40 data samples

Figure 5: Adaption of normal distributions to histograms

For the conditional probability the calculation is more difficult because of the quotient. According to the central limit theorem, m_λ and n_λ are asymptotically normal distributed. Then, the Fieller method [6] can be applied for $X(1), ..., X(n)$ i.i.d. and $Y(1), ..., Y(n)$ i.i.d. [27].

The following asymptotic confidence interval borders result from this approach:

$$p_\lambda^u = \max\left\{ 0, \frac{\frac{m_\lambda n_\lambda}{n} - t^2_{n-1;1-\alpha} S_{CS,S}}{\frac{n_\lambda^2}{n} - t^2_{n-1;1-\alpha} S_S^2} - \right.$$

$$\left. - \frac{\sqrt{\left(\frac{m_\lambda n_\lambda}{n^2} - \frac{t^2_{n-1;1-\alpha}}{n} S_{CS,S}\right)^2 - \left(\left(\frac{m_\lambda}{n}\right)^2 - \frac{t^2_{n-1;1-\alpha}}{n} S^2_{CS}\right)\left(\left(\frac{n_\lambda}{n}\right)^2 - \frac{t^2_{n-1;1-\alpha}}{n} S^2_S\right)}}{\left(\frac{n_\lambda}{n}\right)^2 - \frac{t^2_{n-1;1-\alpha}}{n} S^2_S} \right.$$

$$p_\lambda^o := \min\left\{ \frac{\frac{m_\lambda n_\lambda}{n} - t^2_{n-1;1-\alpha} S_{CS,S}}{\frac{n_\lambda^2}{n} - t^2_{n-1;1-\alpha} S_S^2} + \right.$$

$$\left. + \frac{\sqrt{\left(\frac{m_\lambda n_\lambda}{n^2} - \frac{t^2_{n-1;1-\alpha}}{n} S_{CS,S}\right)^2 - \left(\left(\frac{m_\lambda}{n}\right)^2 - \frac{t^2_{n-1;1-\alpha}}{n} S^2_{CS}\right)\left(\left(\frac{n_\lambda}{n}\right)^2 - \frac{t^2_{n-1;1-\alpha}}{n} S^2_S\right)}}{\left(\frac{n_\lambda}{n}\right)^2 - \frac{t^2_{n-1;1-\alpha}}{n} S^2_S}, 1 \right.$$

with

$n :=$ number of data samples d_k,
$m_\lambda := \sum_{k=1}^{n} \mu_C(Y(k))\mu_S(X(k))$,
$n_\lambda := \sum_{k=1}^{n} \mu_S(X(k))$,
$S_{CS}^2 := \frac{1}{n-1} \sum_{k=1}^{n} (\mu_C(Y(k))\mu_S(X(k)) - \frac{m_\lambda}{n})^2$,
$S_S^2 := \frac{1}{n-1} \sum_{k=1}^{n} (\mu_S(X(k)) - \frac{n_\lambda}{n})^2$,
$S_{CS,S} := \frac{1}{n-1} \sum_{k=1}^{n} (\mu_C(Y(k))\mu_S(X(k)) - \frac{m_\lambda}{n})(\mu_S(Y(k)) - \frac{n_\lambda}{n})$,
$t_{n-1;1-\alpha} := (1-\alpha)$ quantile of the t-distribution with $(n-1)$ degrees of freedom,
$1-\alpha$:=confidence coefficient.

It can be shown that the result for the unconditional probability is a special case of the result for the conditional probability with $S = \Omega$ and $\mu_\Omega = 1$ [27].

A main difference to the crisp relevance test is that the quantities m, n, m_λ, n_λ are not sufficient to calculate the confidence intervals. Additionally, the estimated variances S_C^2, S_S^2, S_{CS}^2, $S_{CS,S}$ are necessary. Thus, for one combination of m, n, m_λ, n_λ an infinity number of values for the confidence interval borders is possible.

For the unconditional probability, the smallest confidence intervals are achieved for $S_C^2 = 0$. Then, the confidence interval borders are $p^u = p^o = \hat{p}$. The largest confidence intervals are achieved if $\mu_C(Y(k)) \in \{0;1\}$ and $\mu_S(X(k)) \in \{0;1\}$ is valid for all values of k. Then, the variance S_C^2 becomes maximum. So, the range of values for the confidence interval borders of p is given by:

$$\max\left\{0, \frac{m}{n} - \frac{t_{n-1;1-\alpha}}{\sqrt{n(n-1)}}\sqrt{m - \frac{m^2}{n}}\right\} \leq p^u \leq \frac{m}{n}$$

$$\frac{m}{n} \leq p^o \leq \min\left\{\frac{m}{n} + \frac{t_{n-1;1-\alpha}}{\sqrt{n(n-1)}}\sqrt{m - \frac{m^2}{n}}, 1\right\}$$

For the conditional probability, the smallest confidence intervals are achieved for $S_S^2 = 0$, $S_{CS}^2 = 0$, $S_{CS,S} = 0$. Then, the confidence interval borders are $p_\lambda^u = p_\lambda^o = \hat{p}_\lambda$. Analyses show that the largest confidence intervals are achieved if $\mu_C(Y(k)) \in \{0;1\}$ and $\mu_S(X(k)) \in \{0;1\}$ is valid for all values of k. Then, the variances S_S^2, S_{CS}^2, $S_{CS,S}$ become maximum. A proof could not be supplied yet. Assuming the correctness of that relationship, the range of values for the confidence interval borders of p_λ is given by:

$$\max\left\{0, \frac{m_\lambda}{n_\lambda} - \sqrt{\frac{m_\lambda^2}{n_\lambda^2} - \frac{m_\lambda^2\left[\left(1 - \frac{t_{n-1;1-\alpha}^2}{m_\lambda}\right) n - 1 + t_{n-1;1-\alpha}^2\right]}{n_\lambda^2\left[\left(1 - \frac{t_{n-1;1-\alpha}^2}{n_\lambda}\right) n - 1 + t_{n-1;1-\alpha}^2\right]}}\right\} \leq p_\lambda^u \leq \frac{m_\lambda}{n_\lambda}$$

$$\frac{m_\lambda}{n_\lambda} \le p^o_\lambda \le \min\left\{ \frac{m_\lambda}{n_\lambda} + \sqrt{\frac{m_\lambda^2}{n_\lambda^2} - \frac{m_\lambda^2\left[\left(1-\frac{t^2_{n-1;1-\alpha}}{m_\lambda}\right)n-1+t^2_{n-1;1-\alpha}\right]}{n_\lambda^2\left[\left(1-\frac{t^2_{n-1;1-\alpha}}{n_\lambda}\right)n-1+t^2_{n-1;1-\alpha}\right]}}, 1\right\}$$

The range of possible values for the conditional probability is almost identical to the range of possible values for the unconditional probability if $m_\lambda = m$ and $n_\lambda = n$. The difference is getting smaller with n_λ increasing and n decreasing.

Example:
In Figure 6 the possible results for p^u and p^o are shown exemplarily for $n = 60$ and $0 < m < 60$. The possible values for p^u lie in the lower marked area and the possible values for p^o lie in the upper marked area. (The same diagram is achieved for p^u_λ and p^o_λ for $n_\lambda = 60$ and $0 < m_\lambda < 60$.) The circles represent the results for $\mu_C(Y(k)) \in \{0;1\} \wedge \mu_S(X(k)) \in \{0;1\}$.

Figure 6: Possible values for the confidence interval borders for $n = 60$ and $0 < m < 60$

For a more detailed view a cutting of the upper area is presented in Figure 7. The range of values of m is $38 < m < 41$. The big circles are adopted from Figure 6. The dotted line is the upper limit of p^o. The stars represent the results for the following values of

Figure 7: Exemplary values of p^o for $n = 60$ and $38 < m < 41$

$\mu_C(y(k))$:

$$\begin{aligned}
\mu_C(y(1)) &= 0 \\
\mu_C(y(2)) &= 0 \\
&\vdots \\
\mu_C(y(20)) &= 0 \\
\mu_C(y(21)) &= 1 \\
\mu_C(y(22)) &= 1 \\
&\vdots \\
\mu_C(y(59)) &= 1 \\
\mu_C(y(60)) &= 1 - q/5
\end{aligned}$$

with $q = 1, 2, 3, 4$. The data samples are constructed in that way that the value of $\hat{p}$ is decreased iteratively from 40/60 to 39/60 by changing only one data sample.

The small circles represent the results for the following values of $\mu_C(y(k))$:

$$\begin{aligned}
\mu_C(y(1)) &= 0 + r/12 \\
\mu_C(y(2)) &= 0 + r/12 \\
&\vdots \\
\mu_C(y(20)) &= 0 + r/12 \\
\mu_C(y(21)) &= 1 - r/24 \\
\mu_C(y(22)) &= 1 - r/24 \\
&\vdots \\
\mu_C(y(60)) &= 1 - r/24
\end{aligned}$$

with $r = 1, 2, ..., 8$. The data samples are constructed in that way that $\hat{p}$ remains constant $40/60$ while all data samples of zero (one) are simultaneously increased (decreased) until all data samples have the same value of $2/3$. For the conditional probability an equivalent diagram to Figure 7 can be constructed.

Usually, the range of values leading to degrees of membership greater than zero is only a part of the whole range of values covered by data samples. Then, the confidence interval borders lie mainly near the maximum values. This is illustrated in Figure 8. For each diagram, 60 data samples are drawn randomly (identically distributed) 10000 times. Then, for each of the 10000 sets of data samples, the upper border p^o of the confidence interval is calculated for the fuzzy event 'room temperature is low' (Figure 3). The whole upper area of possible values of p^o (Figure 6) is devided in five sections. They are defined by building five equidistant intervals for each given value of $\hat{p}$. The first section includes the lowest possible values ($\hat{p}$), the fifth section the greatest possible values (upper border of the upper area). The 10000 results of p^o are assigned to the different sections.

(a) random data samples between 10°C and 30°C

(b) random data samples between 14°C and 22°C

(c) random data samples between 18°C and 20°C

(d) random data samples between 18.5°C and 19.5°C

Figure 8: Number of values of p^o that lie in the equidistant interval $1, 2, ..., 5$ between the lower and upper limit of p^o for 10000 sets of random (identically distributed) data samples of the size $n = 60$

In Figure 8(a) the random temperature values lie between 10°C and 30°C, so that the designed fuzzy set covers only a part of the range of values. Consequently, the most results for p^o lie in the fifth section.

In the case that the random values lie between 14°C and 22°C, almost the whole range of values is covered by the designed fuzzy set. Then, the values of p^o lie about halfly in the fourth and halfly in the fifth section (Figure 8(b)).

In the case that the random values lie between 18°C and 20°C, all values lie on the descending edge of the fuzzy set, so that the degrees of membership of 0 and 1 have each a probability of zero. In this exceptional case the values of p^o lie mainly in the third section (Figure 8(c)).

In the case that the random values lie between 18.5°C and 19.5°C, all values lie on the middle of the descending edge of the fuzzy set, so that the degrees of membership are between 0.75 and 0.25. In this extremely exceptional case the values of p^o lie all in the second section (Figure 8(d)).

The same effect can be observed for the confidence interval borders of the conditional probability.

As the calculations are asymptotically, problems arise for smaller numbers of data samples and here, especially, for the calculation of p^u_λ if $\hat{p}_\lambda \approx 1$ (positive rule) and for the calculation of p^o_λ if $\hat{p}_\lambda \approx 0$ (negative rule). This results from the fact that for $\hat{p}_\lambda = 0$ there is $p^o_\lambda = 0$ and for $\hat{p}_\lambda = 1$ there is $p^u_\lambda = 1$ – independent of the number of data samples. Consequently, rules that are correct for almost all data samples will be seen as relevant even if the number of data samples is small.

5 Comparison

In this Section, first, the results of the algorithmic extension of the crisp relevance test are compared with the results of the asymptotic fuzzy relevance test. Afterwards, a concrete set of data samples of a chemical reactor is taken to compare all three approaches exemplarily by means of three IF/THEN statements.

In Figure 9 the interpolating values of p^o and p^u of the extension of the crisp relevance test are represented together with the possible values of p^o and p^u of the asymptotic fuzzy relevance test of Figure 6 ($n = 60$ and $0 \leq m \leq 60$). The interpolating values of the crisp relevance test lie near the lower and upper border of possible values of the asymptotic fuzzy relevance test. A further comparison is interesting with respect to two viewpoints:

- How good is the result of the asymptotic fuzzy relevance test in the special case of crisp sets ($\mu_C(Y(k)) \in \{0; 1\} \wedge \mu_S(X(k)) \in \{0; 1\}$)?
  The crisp relevance test supplies the exact results. The results of the asymptotic fuzzy relevance test are given by the circles of Figure 9. The difference between the results is the error of the asymptotic fuzzy relevance test in the case of crisp sets. The error is decreasing monotonously to zero for an increasing n.

- How good is the interpolation by the extension of the crisp relevance test in the case of fuzzy sets $(\mu_C(Y(k)) \in [0;1] \wedge \mu_S(X(k)) \in [0;1])$?
  Considering Figure 8, the error is small if only a part of the data samples realizes in the range of the fuzzy set of the input situation and in the range of the fuzzy set of the output event. Then, the variances are high and the confidence interval borders are near the lower and upper border of the possible values. If the data samples realize mainly at the increasing or decreasing edge of the fuzzy set of the input situation and the fuzzy set of the output event, the variances are low and the error is greater. Principally, triangular fuzzy sets will lead to smaller errors than trapezium fuzzy sets and fuzzy sets with low density to greater errors than fuzzy sets with high density.

Figure 9: Comparison of the confidence interval borders p^u and p^o of the asymptotic fuzzy relevance test and the crisp relevance test for fuzzy values for $n = 60$ and $0 < m < 60$

Using the extension of the crisp relevance test, the confidence intervals are mostly greater than necessary. This can be interpreted as a conservative relevance test that aspires a minimum confidence coefficient of $1 - \alpha$. Consequently, it can happen that statements are not accepted as rules that would be accepted if an exact confidence coefficient of $1 - \alpha$ is demanded.

In accordance with the more complex formula, the computing time of the asymptotic fuzzy relevance test is a little bit longer than the computing time of the crisp relevance test. Whereas, the computing time of the Bootstrap fuzzy relevance test is not practical for testing a higher number of statements. Nevertheless, the Bootsprap fuzzy relevance test can be used to judge the results of the other two relevance tests, as it supplies very good results [5, 25].

In Figure 10 the results of the three relevance tests of three different statements are shown. The confidence intervals are calculated with a confidence coefficient of 0.95. Measured by the results of the Bootstrap fuzzy relevance test, the results of the asymptotic fuzzy relevance test are very good. The confidence intervals of the extension of the crisp relevance test are larger. The first two statements are seen as relevant by all three tests as the confidence intervals do not overlap. The first statement represents a negative relevant rule, the second statement a positive relevant rule. The third statement is seen as a positive relevant rule by the fuzzy relevance tests, but not by the algorithmic extension of the crisp relevance test. Here, the larger confidence intervals cause an overlapping.

(a) IF (oil temperature is in set 1) THEN (reactor temperature is in set 4) ($n = 206$, $m = 93.86$, $n_\lambda = 15.51$, $m_\lambda = 0.13$)

(b) IF (oil temperature is in set 4) THEN (reactor temperature is in set 4) ($n = 206$, $m = 93.86$, $n_\lambda = 25.69$, $m_\lambda = 17.91$)

(c) IF (oil temperature is in set 2) THEN (reactor temperature is in set 5) ($n = 206$, $m = 92.85$, $n_\lambda = 98.95$, $m_\lambda = 57.76$)

Figure 10: Comparison of confidence interval borders for process data of 206 data samples for three selected IF/THEN statements: ']','[' results of the extension of the crisp relevance test, '↑' results of the Bootstrap fuzzy relevance test, '↓' results of the asymptotic fuzzy relevance test

6 Conclusions

In the field of data-based fuzzy modelling, the incremental collecting of single relevant rules allows to handle complex problems. To decide if an IF/THEN statement is a relevant rule a relevance test is necessary. A statistical approach is given by the demand that the confidence intervals of the probabilities p and p_λ do not overlap – with p the probability that the output event of the conclusion is true and p_λ the probability that the output event is true under the condition that the input situation of the premise is true.

For crisp rule-based modelling the confidence intervals can be calculated by conventional statistical formulas. For fuzzy modelling problems arise. Three different methods are proposed in this paper: an algorithmic extension of the crisp relevance test, a Bootstrap fuzzy relevance test and an asymptotic fuzzy relevance test.

The results of the Bootstrap fuzzy relevance test are very good, but the high computing time makes its application only practicable for a small number of data samples. The asymptotic fuzzy relevance test supplies good results for a higher number of data samples. The algorithmic extension of the crisp relevance test tends to calculate too large confidence intervals, but has the smallest computing time.

The employment of the three relevance tests will depend on the respective application. For high dimensional search spaces with a multitude of relevant rules, the algorithmic extension is acceptable, especially, if for each input and output variable several trapezium fuzzy sets are reasonable. In the other cases, the higher effort of the fuzzy relevance tests can remunerate.

The calculation of estimators and confidence intervals on fuzzy data is also meaningful for other test and rating strategies, e.g. the results can be directly used for the method 'Confident Normalized Hit Rate' of Jessen and Slawinski [9].

Acknowledgement

We thank Prof. Dr. Hering and Dr. Poehlmann for the constructive statistical discussions. The research is sponsored by the Deutsche Forschungsgemeinschaft (DFG), as part of the Collaborative Research Center 'Computational Intelligence' (531) of the University of Dortmund.

References

[1] H. Bandemer and S. Gottwald. *Einfuehrung in Fuzzy-Methoden: Theorie und Anwendung unscharfer Mengen.* Akademie Verlag GmbH, Berlin, 4th edition, 1993.

[2] D. Dubois and H. Prade. *Fuzzy Sets and Systems: Theory and Applications.* Academic Press, New York, 1980.

[3] B. Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics 7*, pages 1–26, 1979.

[4] B. Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association 82*, pages 171–200, 1987.

[5] B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap.* Chapman & Hall, New York, 1993.

[6] E.C. Fieller. Some problems in interval estimation. *Journal Royal Statistical Society Series B 16*, pages 175–185, 1954.

[7] J. Hartung, B. Elpelt, and K-H. Kloesener. *Statistik.* Oldenbourg Verlag, Muenchen, 10th edition, 1995.

[8] J.S.U Hjorth. *Computer Intensive Statistical Methods.* Chapman & Hall, New York, 1994.

[9] H. Jessen and T. Slawinski. Test- and rating strategies for data-based rule generation. In *Reihe Computational Intelligence.* No CI-39/98, Collaborative Research Center 531, University of Dortmund, 1998.

[10] H. Kiendl. Hyperinferenz und Hyperdefuzzifizierung. *Automatisierungstechnik 41, Heft 11*, pages 33–36, 1993.

[11] H. Kiendl. *Fuzzy-Control methodenorientiert.* Oldenbourg Verlag, Muenchen, 1997.

[12] M. Kiendl, H. Krabs. Ein Verfahren zur Generierrung regelbasierter Modelle fuer dynamische Systeme. *Automatisierungstechnik 37*, pages 423–430, 1989.

[13] M. Krabs. *Das ROSA-Verfahren zur Modelierung dynamischer Systeme durch Regeln mit statistischer Relevanzbewertung.* VDI Verlag, 1994.

[14] M. Krabs and H. Kiendl. Anwendungsfelder der automatischen Regelgenerierung mit dem ROSA-Verfahren. *at - Automatisierungstechnik*, 43(6):269–276, 1995.

[15] A. Krone. Advanced rule reduction concepts for optimising efficiency of knowledge extraction. In *EUFIT '96 September 2-5*, pages 919–923, Aachen, 1996.

[16] A. Krone, T. Baeck, and P. Teuber. Evolutionaeres Suchkonzept zum Aufstellen signifikanter Fuzzy-Regeln. *at - Automatisierungstechnik*, 44(8):405–411, 1996.

[17] A. Krone, C. Frenck, and O. Russak. Design of a fuzzy controller for an alkoxylation process using the ROSA method for automatic rule generation. In *Third European Congress on Intelligent Techniques and Soft Computing , EUFIT '95*, pages 760–764, Aachen, 1995.

[18] A. Krone and H. Kiendl. Automatic generation of positive and negative rules for two-way fuzzy controllers. In *Second European Congress on Intelligent Techniques and Soft Computing , EUFIT'94*, pages 438–447, Aachen, 1994.

[19] A. Krone and H. Kiendl. Strategieelemente zur Beherrschung von Komplexitaet bei der rechnergestuetzten regelbasierten Modellierung. In *5. Workshop Fuzzy Control*, pages 56–70. VDI/VDE GMA-UA 1.4.2, Forschungsbericht der Fakultaet fuer Elektrotechnik, Nr. 0295, Universitaet Dortmund, 1995.

[20] A. Krone and H. Kiendl. Rule-based decision analysis with Fuzzy-ROSA method. In *European Workshop on Fuzzy Decision Analysis for Management, Planning and Optimization, EFDAN'96*, pages 109–114, 1996.

[21] A. Krone and H. Kiendl. Evolutionary concept for generating relevant fuzzy rules from data. *International Journal of Knowledge-based Intelligent Engineering Systems*, 1(4):207–213, 1997.

[22] A. Krone and U. Schwane. Generating fuzzy rules from contradictory data of different control strategies and control performances. In *Fifth IEEE International Conference on Fuzzy Systems, FUZZ- IEEE'96*, pages 492–497, New Orleans, 1996.

[23] A. Krone and H. Taeger. Behandlung unscharfer Ereignisse bei der Generierung relevanter Regeln mit dem Fuzzy-ROSA-Verfahren. Forschungsbericht Nr. 0296, Department of Electrical Engineering, University of Dortmund, 1996.

[24] M. Mizumoto. Pictorial representations of fuzzy connectives, part I: Cases of t- norms, t- conorms and averaging operators. *Fuzzy Sets and Systems 31*, pages 217–242, 1989.

[25] J. Shao and D. Tu. *The Jackknife and Bootstrap.* Springer, New York, 1995.

[26] P. Smets. Probability of a fuzzy event: An axiomatic approach. *Fuzzy Sets and Systems 7*, pages 153–164, 1982.

[27] H. Taeger. Relevanzueberpruefung von unscharfen Regeln mittels statistischer Verfahren. Master's thesis, University of Dortmund, Department of Statistics, 1998.

[28] S.-O. Troschke. Kennzahlen der regelbasierten Modellierung in Expertensystemen - Ein Ansatz zur Bewertung von Unsicherheit bei der automatischen Erzeugung von Produktionsregeln. Master's thesis, University of Dortmund, Department of Mathematics, 1992.

[29] L.A. Zadeh. Fuzzy sets. *Information Control 8*, pages 338–353, 1965.

[30] L.H. Zadeh. Probability measures of fuzzy events. *Journal of Mathematical Analysis and Applications 23*, pages 421–427, 1968.