Mathware & Soft Computing 6 (1999) 91-108

Fuzzy Max-Min Classifiers Decide locally on the Basis of Two Attributes

Birka von Schmidt[1] & Frank Klawonn[2]
[1] Institute for Flight Guidance. German Aerospace Center
Lilienthalplatz 7. D-38108 Braunschweig, Germany
[2]Dept. of Electrical Engineering and Computer Science
Ostfriesland University of Applied Sciences.
Constantiaplatz 4. D-26723 Emden, Germany

Abstract

Fuzzy classification systems differ from fuzzy controllers in the form of their outputs. For classification problems a decision between a finite number of discrete classes has to be made, whereas in fuzzy control the output domain is usually continuous, i.e. a real interval. In this paper we consider fuzzy classification systems using the max-min inference scheme and classifying an unknown datum on the basis of maximum matching, i.e. assigning it to the class appearing in the consequent of the rule whose premise fits best. We basically show that this inference scheme locally takes only two attributes (variables) into account for the classification decision.

1 Introduction

From a theoretical point of view fuzzy controllers are a method to describe a real function $\mathbb{R}^m \to \mathbb{R}$ (or, in the case of multi-input, multi-output systems, $\mathbb{R}^m \to \mathbb{R}^k$) assigning a real (control) value to a given tuple, point or vector of measured input values. There are a variety of different models of fuzzy controllers like the Mamdani-type controller [12] that uses fuzzy sets in the consequent part of the rules or the Takagi-Sugeno model [17] that allows a (linear) function of the inputs in the consequent part of the rule. For an overview see for instance [9].
In almost all fuzzy control systems, the final crisp output is computed incorporating the outputs of all rules whose premises are satisfied to a degree greater than zero. There are many different ways to aggregate the outputs of the single rules and – in the case of a Mamdani controller – to defuzzify the resulting fuzzy set. Nevertheless, the underlying principal is always that the output is some kind of weighted mean of the outputs of the firing rules.
Fuzzy controllers are well examined as function approximators. Piecewise monotone functions of one variable can be exactly reproduced by a fuzzy controller [1, 11]

and for the multi-dimensional case fuzzy controllers are known to be universal approximators [2, 8, 18] for continuous functions. However, although these positive general results do not apply when the number of rules is restricted. In this case, the set of functions that can be represented by a fuzzy controller is nowhere dense [14].
These results do not apply to fuzzy classification systems. The situation is different, since we have to deal with a function $\mathbb{R}^m \to \mathcal{C}$ where $\mathcal{C}$ is a finite set of discrete classes. We do not assume any kind of structure on $\mathcal{C}$. This means that interpolation between classes does not make any sense. The classes could for instance be different diseases in medical applications, *broken/not broken* in quality control of tiles. Fuzzy classification systems of this type are successfully applied (see for instance [4, 5, 7, 13, 15, 19]), but a systematic experimental or theoretical analysis of these systems was initiated just recently.
Nürnberger et al. [16] investigate the class boundaries of two- and three-dimensional data that can be generated by fuzzy classification systems using different t-norms. Cordón et al. [3] analyse fuzzy classification systems on an experimental basis that do not rely on a classification based on the rule that best fits the input.
A theoretical analysis of fuzzy classification systems is presented in [6]. It was demonstrated that approximate solutions of arbitrary classification problems can already be obtained with crisp sets instead of fuzzy sets. In the case of two-dimensional data, classification problems can exactly be solved, when the classes can be separated by piecewise monotone functions. When the Łukasiewicz t-norm is allowed instead of the minimum or the maximum is replaced by the bounded sum, arbitrary linearly separable classification problems can be solved by fuzzy classification systems, i.e. problems where the classes are separated by a (hyper-)plane. However, fuzzy max-min classification systems cannot solve arbitrary linearly separable classification problems for data with more than two attributes. If the separating hyper-plane depends on more than two variables, fuzzy max-min classification systems can only provide an approximate solution of the corresponding classification problem.
In this paper we generalise this result and show that in principal fuzzy max-min classification systems determine the class locally on the basis of only two attributes.
The paper is organised as follows. The following section briefly reviews the structure of fuzzy max-min classification systems. Then we introduce the basic definitions, that we need, and present our main theorem in section 3. Section 4 contains the construction that proves the main theorem. Some technical requirements needed in the prove can be found in the appendix.

2 Fuzzy Max-Min Classification Systems

We consider the following classification problem. We have a finite number of classes $\mathcal{C}_1, \ldots, \mathcal{C}_c$. Each class represents a subset of the space $\mathbb{R}^m$ or the unit cube $[0,1]^m$. Therefore, we identify each class with its corresponding subset. We assume that the classes are pairwise disjoint, but we do not require that they cover the whole space, i.e. there might be data that are unclassified.

In practical applications, the situation is usually as follows: A finite set of data including the classes to which the data belong is given. The problem is to find a classifier that – in the best case – assigns all the given (training) data to the corresponding classes and extends the classification also to unknown data in a reasonable way. We do not discuss here, how such a classifier can be learned from data. We are interested in how flexible a fuzzy classifier can be. Therefore, we assume that the corresponding classes are already known for all possible data and examine, whether this classification problem can be solved by a fuzzy classifier.
We restrict our investigations to classification problems with only two classes $\mathcal{C}^+$ and $\mathcal{C}^-$. However, our results can easily be extended to classification problems with more than two classes. In order to see whether the class $\mathcal{C}_1$ can be separated correctly from the other classes $\mathcal{C}_2, \ldots, \mathcal{C}_c$, we simply have to combine the classes $\mathcal{C}_2, \ldots, \mathcal{C}_c$ to one new class and we have a classification with only two classes.
A fuzzy max-min classification system can be formalised as follows. We have a finite set $\mathcal{R}$ of rules of the form

$$R\text{: If } x_1 \text{ is } \mu_R^{(1)} \text{ and } \ldots \text{ and } x_m \text{ is } \mu_R^{(m)} \text{ then class is } \mathcal{C}_R,$$

where $\mathcal{C}_R$ is either $\mathcal{C}^+$ or $\mathcal{C}^-$. The $\mu_R^{(i)}$ are assumed to be fuzzy sets on the X_i, i.e. $\mu_R^{(i)} : X_i \to [0,1]$, where X_i is an interval. In order to keep the notation simple, we incorporate the fuzzy sets $\mu_R^{(i)}$ directly in the rules. In real systems one would replace them by suitable linguistic values like *positive big*, *approximately zero*, etc. and associate the linguistic value with the corresponding fuzzy set.
Each rule is evaluated by interpreting the conjunction in terms of the minimum, i.e.

$$\mu_R(p_1, \ldots, p_m) \;=\; \min_{i \in \{1,\ldots,m\}} \left\{ \mu_R^{(i)}(p_i) \right\} \tag{1}$$

is the degree to which rule R fires.

$$\mu_C^{(\mathcal{R})}(p_1, \ldots, p_m) \;=\; \max \left\{ \mu_R(p_1, \ldots, p_m) \mid \mathcal{C}_R = \mathcal{C} \right\} \tag{2}$$

is the degree to which the point $(p_1, \ldots, p_m)$ is assigned to class $\mathcal{C}$.
Finally, we have to assign the point $(p_1, \ldots, p_m)$ to a unique class (defuzzification) by

$$\mathcal{R}(p_1, \ldots, p_m) \;=\; \begin{cases} \mathcal{C}^+ & \text{if } \mu_{\mathcal{C}^+}^{(\mathcal{R})}(p_1, \ldots, p_m) > \mu_{\mathcal{C}^-}^{(\mathcal{R})}(p_1, \ldots, p_m) \\ \mathcal{C}^- & \text{if } \mu_{\mathcal{C}^-}^{(\mathcal{R})}(p_1, \ldots, p_m) > \mu_{\mathcal{C}^+}^{(\mathcal{R})}(p_1, \ldots, p_m) \\ \textit{unknown} & \text{otherwise.} \end{cases}$$

This means that we assign the point $(p_1, \ldots, p_m)$ to the class of the rule with the maximum firing degree. Note that we denote by $\mathcal{R}$ the set of rules as well as the associated classification mapping based on these classification rules.
When we assume that the fuzzy sets appearing in the rules are continuous, then $\mathcal{C}^+$ and $\mathcal{C}^-$ are open sets. This means that when a point $(p_1, \ldots, p_m)$ is assigned to the class $\mathcal{C}^+$, then there is a neighbourhood of this point in which all points are

also assigned to $\mathcal{C}^+$. The same holds for the class $\mathcal{C}^-$. We are interested in the class boundary, i.e. the set of points that are classified as *unknown*. A typical situation for a point classified as *unknown* is the following: There is exactly one rule firing with the maximum degree for class $\mathcal{C}^+$ and also exactly one rule firing with the maximum degree for class $\mathcal{C}^-$. For each of these rules the firing degree is determined by just one attribute for which the membership degree to the corresponding fuzzy set in the rule yields the firing degree (the minimum in equation (1)). Let us for the moment assume that x_1 is the corresponding attribute for class $\mathcal{C}^+$ and x_2 for class $\mathcal{C}^-$. This means that the firing degree for class $\mathcal{C}^+$ and $\mathcal{C}^-$ does not change, when slightly change any of the values of the attributes $x_3, \ldots, x_m$. In this sense, the classification depends in this situation locally only on the two attributes x_1 and x_2.
However, the above considerations are only correct in this special case where the maximum firing degree for each class is determined by just one rule and the minimum in the corresponding rules is determined by just one variable each.
When there are more than just two rules firing with maximum degree, the situation is different. Let us consider the six rules

R_1: If x_1 is positive small and x_2 is anything and x_3 is anything
then class is $\mathcal{C}^+$

R_2: If x_1 is anything x_2 is positive small and x_3 is anything
then class is $\mathcal{C}^+$

R_3: If x_1 is anything and x_2 is anything and x_3 is positive small
then class is $\mathcal{C}^+$

R_4: If x_1 is negative small and x_2 is anything and x_3 is anything
then class is $\mathcal{C}^-$,

R_5: If x_1 is anything and x_2 is negative small and x_3 is anything
then class is $\mathcal{C}^-$,

R_6: If x_1 is anything and x_2 is anything and x_3 is negative small
then class is $\mathcal{C}^-$,

where the fuzzy sets *positive small* and *negative small* are chosen as illustrated in figure 1 and the fuzzy set corresponding to *anything* is the constant function 1.
When we consider the point $(0, 0, 0)$, we have

$$\mu_{C^+}^{(\mathcal{R})}(0, 0, 0) \;=\; 0.5$$

and

$$\mu_{C^-}^{(\mathcal{R})}(0, 0, 0) \;=\; 0.5,$$

i.e. $(0, 0, 0)$ is classified as unknown. But when we increase any of the three variables x_1, x_2, or x_3, the resulting point is classified to $\mathcal{C}^+$ and when we decrease any of these three variables the resulting point is classified to $\mathcal{C}^-$. This means that the

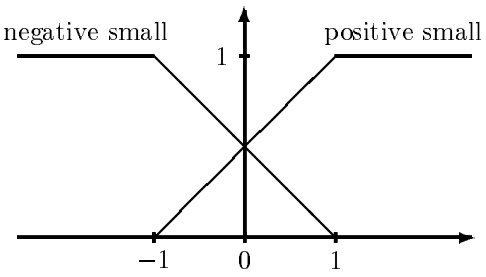


Figure 1: Two fuzzy sets

classification near $(0,0,0)$ depends on all three attributes. If we choose $\varepsilon > 0$, we have $(\varepsilon,0,0),(0,\varepsilon,0),(0,0,\varepsilon) \in \mathcal{C}^+$ and $(-\varepsilon,0,0),(0,-\varepsilon,0),(0,0,-\varepsilon) \in \mathcal{C}^-$. Therefore, we cannot say that fuzzy max-min classification systems generally decide locally on the basis of two variables. However, the above described example can be seen as an exception and we can show the following: When there is a point on the class boundary where the classification depends (locally) on more than two variables, then in any neighbourhood of this point there is another point on the class boundary where the classification depends locally only on at most two variables.
So far we have not made any assumptions on the fuzzy sets. We require that they are continuous and that they have a local one-sided Taylor expansion everywhere. This means the following: If μ is a fuzzy set, then for any $x_0 \in \mathbb{R}$ there is $\varepsilon > 0$ and there exist power series

$$\sum_{k=0}^{\infty} a_k^{(l)} (x - x_0)^k \quad \text{and} \quad \sum_{k=0}^{\infty} a_k^{(r)} (x - x_0)^k,$$

so that

$$\mu(x_0 - h) = \sum_{k=0}^{\infty} a_k^{(l)} h^k$$

and

$$\mu(x_0 + h) = \sum_{k=0}^{\infty} a_k^{(r)} h^k$$

hold for all $0 < h < \varepsilon$.
Note that typical membership functions used in application like piecewise linear functions (for instance triangular or trapezoidal fuzzy sets) or Gaussian membership functions fullfill this property. In the following section we will see, why we need this technical condition.

3 Basic Definitions

We consider a fuzzy max-min classification system as it was described in the previous section.

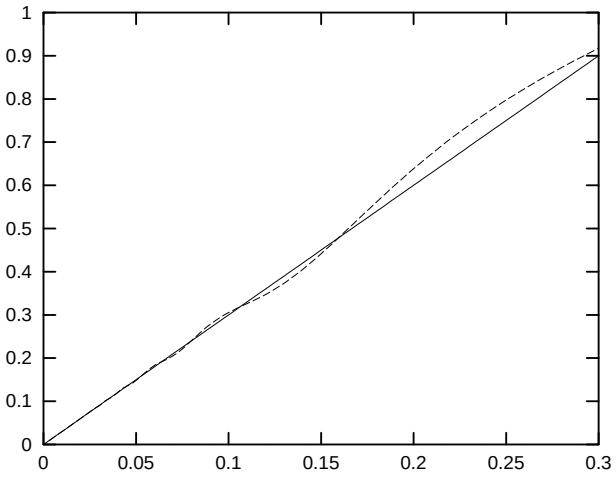


Figure 2: Two fuzzy sets

Definition 1 *The set $\mathcal{D}$ of the points that have the same membership degree to $\mathcal{C}^+$ as to $\mathcal{C}^-$ is called separating set.*

$\delta^{(i)}$ denotes the vector that has δ as the i^{th} component and 0 for the other components.
As we have already mentioned in the previous section, we require that the fuzzy sets have a local one-sided Taylor expansion. In order to illustrate what can happen, if we refrain from this restriction, we consider the following example.

Example 1 *We consider a fuzzy classification system for data with just one attribute x with the following two rules:*

R_1: If x is μ_1 then class is $\mathcal{C}^+$
R_2: If x is μ_2 then class is $\mathcal{C}^-$

where the fuzzy sets μ_1 and μ_2 are defined by

$$\mu_1(x) = \begin{cases} 0 & \text{if } x < 0 \\ 3x & \text{if } 0 \leq x \leq \frac{1}{3} \\ 1 & \text{otherwise} \end{cases}$$

and

$$\mu_2(x) = \begin{cases} 0 & \text{if } x < 0 \\ 3x - x^2 \cdot \sin(\frac{1}{x}) & \text{if } 0 \leq x \leq \frac{1}{3} \\ 1 - \frac{1}{9} \cdot \sin(3) & \text{otherwise.} \end{cases}$$

Figure 2 illustrates these two fuzzy sets.
Note that μ_2 is continuous, even differentiable, and monotonous (the first derivative is positive), but does not have a local one-sided Taylor expansion at $x_0 = 0$.

The point 0 is a separating point for this fuzzy classification system. But for any interval $[0, \varepsilon]$*, no matter how small we choose* $\varepsilon > 0$*, there are points in this interval that are classified to* $\mathcal{C}^+$ *and also points that are classified to* $\mathcal{C}^-$*.*

By requiring that each fuzzy set is continuous and has a local one-sided Taylor expansion everywhere, we can not have such a strange situation as in the above described example. When we are in a separating point and consider one variable that we want to change in one direction by a very small value, then we can say that we always end up in the same class (or always remain in the separating set), as long as the change is small enough. An 'oscillation' between the classes as in the example is not possible. The following shows that our fuzzy classification systems have this property.

Lemma 1 *For each point* p *and for each coordinate* p_i *of* p*, we have*

$$(\exists\, A, B \in \{\mathcal{C}^+, \mathcal{C}^-, \mathcal{D}\})(\exists\, \varepsilon > 0)\ (\forall\, 0 < \delta < \varepsilon)(p + \delta^{(i)} \in A\ \wedge\ p - \delta^{(i)} \in B). \quad (3)$$

Proof: If p is not a separating point, then p belongs to $\mathcal{C}^+$ or $\mathcal{C}^-$. Since these sets are open, a sufficiently small variation of any variable will not lead out of these sets. Therefore, we only need to consider separating points. Let us assume, that we want to increase the variable p_i. It is easy to determine, how a (sufficiently) small increase of p_i will alter the firing degree $\mu^{(\mathcal{R})}_{\mathcal{C}^+}(p)$ for class $\mathcal{C}^+$. When the change of p_i influences the firing degree $\mu^{(\mathcal{R})}_{\mathcal{C}^+}(p)$ at all, we only need to consider the rules firing for class $\mathcal{C}^+$ in which p_i determines the minimum. It is now easy to determine which will take over when we increase p_i: We need to know which fuzzy set for p_i will yield the strongest change. Since we can compute the Taylor expansions of each fuzzy set for p_i, we can easily solve this problem using lemma 5 in the appendix. We can do the same for the rules for class $\mathcal{C}^-$. Finally, we have to decide for which class we have the strongest change. But this can be done again by making use of lemma 5. □

Note that we only need the Taylor expansions for the proof of lemma 1. The essential property that we are interested in is (3).

Definition 2 *Let* p *be a point of the separating set* $\mathcal{D}$*.* p *is called a proper separating point, when*

$$(\forall\, \varepsilon > 0)(\ \exists\, p', p'' \in \mathcal{N}_\varepsilon(p)) : (p' \in \mathcal{C}^+ \ \textit{and}\ p'' \in \mathcal{C}^-)$$

holds, where $\mathcal{N}_\varepsilon(p)$ *denotes the* ε*-neighbourhood of* p*.*

This means that for $\mathcal{C}^-$ as well as for $\mathcal{C}^+$ there exists a direction in which the set can be reached in an arbitrarily small distance from p.

Definition 3 *Let* p *be a proper separating point and* x_i *a single variable with the value* p_i *for* p*.*

1. *x_i is called relevant at p iff*

$$(\exists\, \varepsilon > 0)(\forall\, 0 < \delta < \varepsilon): \quad \{p + \delta^{(i)}, p - \delta^{(i)}\} \cap \mathcal{C}^+ \neq \emptyset \text{ and } \{p + \delta^{(i)}, p - \delta^{(i)}\} \cap \mathcal{C}^- \neq \emptyset.$$

   This means that increasing p_i leads into one set and decreasing p_i leads into the other one.

2. *x_i is called semi-relevant (for $\mathcal{C}^+$) at p iff*

$$(\exists\, \varepsilon > 0)(\forall\, 0 < \delta < \varepsilon): \quad \{p + \delta^{(i)}, p - \delta^{(i)}\} \subset \mathcal{C}^+ \cup \mathcal{D} \text{ and } \{p + \delta^{(i)}, p - \delta^{(i)}\} \cap \mathcal{C}^+ \neq \emptyset$$

   This means that we can reach only $\mathcal{C}^+$ and not $\mathcal{C}^-$ by varying p_i by an arbitrarily small distance. In the same way we define the notion "semi-relevant for $\mathcal{C}^-$".

3. *x_i is called irrelevant at p iff*

$$(\exists\, \varepsilon > 0)(\forall\, 0 < \delta < \varepsilon): \{p + \delta^{(i)}; p - \delta^{(i)}\} \subseteq \mathcal{D}.$$

   This means that it is impossible to reach either $\mathcal{C}^+$ or $\mathcal{C}^-$ by varying p_i by an arbitrarily small amount.

Lemma 1 guarantees that a proper separating point is either relevant, semi-relevant or irrelevant.

Theorem 2 (main theorem) *Let p be a proper separating point and $\varepsilon > 0$. Then there exists a proper separating point p' in the neighbourhood $\mathcal{N}_\varepsilon(p)$ of p that has at most two variables that are relevant or semi-relevant.*

The proof of this theorem will be given in the next section, where we actually show constructively, how to obtain the point $p' \in \mathcal{N}_\varepsilon(p)$ that has only two relevant or semi-relevant variables.

Remark 1 *There is no reason in considering a point p that does not belong to $\mathcal{D}$. Because of $\mathcal{C}^+$ and $\mathcal{C}^-$ being open sets, there is always a neighbourhood of p that is completely contained in $\mathcal{C}^+$, respectively $\mathcal{C}^-$. In this sense, points that are in one of the classes, do not have relevant variables.*
In case of p being an inner point of $\mathcal{D}$ we can use the same argumentation.

4 Finding a Point with only Two Relevant Variables

Without loss of generality we assume that $x_1, \cdots, x_n$ are the relevant and semi-relevant variables and $x_{n+1}, \cdots, x_m$ are the irrelevant ones.

Remark 2 *We consider a point $p = (p_1, \cdots, p_m) \in \mathbb{R}^m$. The set of the rules that are firing in p with the maximum degree $\mu^{(\mathcal{R})}_{\mathcal{C}^+}(p) = \mu^{(\mathcal{R})}_{\mathcal{C}^-}(p)$ is denoted by $\mathcal{R}_p$.*

1. *When we vary an attribute of p, the variation has to be sufficiently small, so that for the new point $p' = (p_1, \cdots, p'_i, \cdots, p_m)$ there is no new rule R with $R \in \mathcal{R}_{p'}$ but $R \notin \mathcal{R}_p$.*

2. *When we consider two fuzzy-sets $\mu = \mu^{(i)}_R$ and $\nu = \nu^{(i)}_{R'}$ for one variable $x = x_i$, we want to know which one is increasing faster, when we vary x. Since the Taylor expansions at x_0 in the considered direction of the two functions exist, we can compute the (directed) derivatives at x_0. When the first n derivatives are equal, but for the $(n+1)^{th}$ derivative we have $\mu^{(n+1)}(x_0) > \nu^{(n+1)}(x_0)$, then within $\mathcal{N}_\varepsilon(x_0)$ μ is increasing faster than ν, when increasing x, and the other way round, when decreasing x.*

   For this comparison we only need the coefficients of the Taylor expansion up to that term that is different for μ and ν. We will explain this more detailed in the appendix.

3. *The variation ε must also be sufficiently small, so that μ and ν do not 'overtake' each other. This means the following: When $\mu^{(n+1)}(x_0) > \nu^{(n+1)}(x_0)$, then $\mu(x) > \nu(x)$ for all $x > x_0$, $x \in \mathcal{N}_\varepsilon(x_0)$, and $\mu(x) < \nu(x)$ for all $x < x_0$, $x \in \mathcal{N}_\varepsilon(x_0)$.*

Because of part 1 of remark 2 the irrelevant variables stay irrelevant, so that we do not need to consider them at all. For the proof of our main theorem we only need to consider the case that there are at least three relevant or semi-relevant variables at the point p.

4.1 First case: Every rule contributing to the maximum firing degree has just one dominating variable

Definition 4 *Let R be a rule and let x_i be a variable. x_i is called dominating at point p iff*

$$\mu_R(p) \;=\; \mu^{(R)}_i(p_i).$$

The first case to be considered is the most simple one, where each rule of $\mathcal{R}_p$ has only one dominating variable at point p.
Changing one variable x_i and considering a single rule (without loss of generality firing for $\mathcal{C}^+$), we have the following: When μ_i is increasing with the change of x_i, then the rule leads to a decision for $\mathcal{C}^+$; the case of μ_i remaining constant is trivial, and when μ_x is decreasing, the rule is not relevant any more for the classification of p.

4.1.1 Λ_i^+ and Λ_i^-

$\mathcal{R}^{(i)}_p$ denotes the set of rules firing in p with the dominating variable x_i, i.e. the rules $R \in \mathcal{R}_p$ with $\mu^{(i)}_R(p_i) = \mu_R(p)$. For our purpose, it is possible to combine

the firing degrees given by the rules of $\mathcal{R}_p^{(i)}$ firing for $\mathcal{C}^+$ into one function Λ_i^+. Λ_i^+ describes the membership degree for the class $\mathcal{C}^+$, when we vary the attribute p_i of p and consider only the rules of $\mathcal{R}_p^{(i)}$:

$$\Lambda_i^+(\delta) \;:=\; \mu_{\mathcal{C}^+}^{(\mathcal{R}_p^{(i)})}(p+\delta^{(i)}) \;=\; \max\{\mu_R^{(i)}(p+\delta^{(i)}) \mid R \in \mathcal{R}_p^{(i)},\; C_R = \mathcal{C}^+\}$$

The Λ_i^- are defined analogously. When there are different rules firing with maximum degree at p that have all x_i as the only dominating variable, then we normally get a sharp bend at $\delta = 0$ (see figure 3).

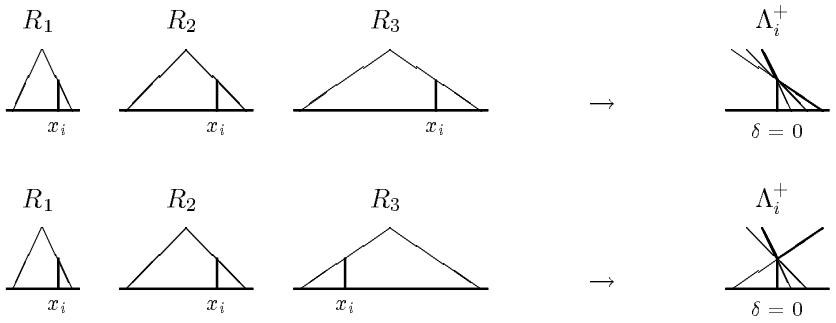


Figure 3: Two examples for the construction of Λ_i^+

When all rules firing at p for $\mathcal{C}^+$ with the dominating variable x_i are combined into Λ_i^+, then we can consider Λ_i^+ as the only function giving the degree to which p belongs to $\mathcal{C}^+$ with respect to the dominating variable x_i.
Λ_i^- is constructed in the same way, and if e.g. for $\mathcal{C}^+$ and x_i there is no such rule, then we have $\Lambda_i^+ \equiv 0$.

When each attribute p_i is changed by δ_i, then we denote the vector incorporating all changes by $\bar{\delta} := \sum_{i=1}^n \delta_i^{(i)}$. The functions Λ_M^+ and Λ_M^- of the total membership degree of $p' := p + \bar{\delta}$ to $\mathcal{C}^+$ and $\mathcal{C}^-$ are calculated in the following way:

$$\Lambda_M^+(p') \;=\; \Lambda_M^+(p+\bar{\delta}) \;:=\; \max\{\Lambda_i^+(\delta_i) \mid i \in \{1,\cdots,n\}\},$$

and analogously for Λ_M^-.

4.1.2 Moving towards a point with only two relevant variables

We consider Λ_i^+ and Λ_i^- instead of the individual rules.
For the procedure of finding $p' \in \mathcal{N}_\varepsilon(p) \cap \mathcal{D}$ with p' having just two relevant variables we choose a variable x_i that is relevant or semi-relevant for $\mathcal{C}^+$ and another variable x_j that is relevant or semi-relevant for $\mathcal{C}^-$. We take p_i and vary it by the distance $\delta_i \neq 0$, so that $\Lambda_i^+(\delta_i) > \Lambda_i^-(\delta_i)$ and $\Lambda_i^+(\delta_i) > \Lambda_k^+(0) = \Lambda_k^-(0)$ for $k \neq i$.
We obtain a point $p'' = p + \delta_i^{(i)}$ that belongs to $\mathcal{C}^+$, but we are looking for a point $p' \in \mathcal{D}$. Now we can alter the attribute p_j of p by $\delta_j \neq 0$ towards that direction

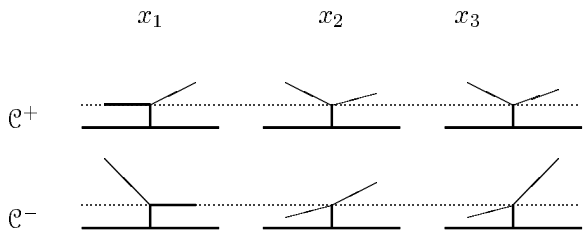


Figure 4: An example with three variables for Λ_i^+ and Λ_i^-

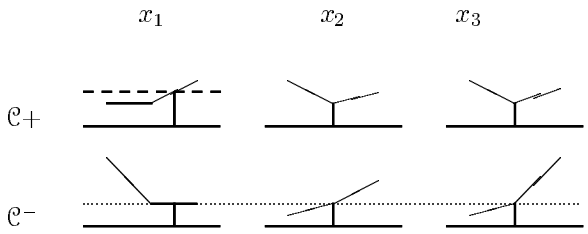


Figure 5: The same example after pushing x_1

where Λ_j^- grows faster than Λ_j^+, so that $\Lambda_j^-(\delta_j) > \Lambda_j^+(\delta_j)$ holds. We choose δ_j in such a way that after this step we have $\Lambda_j^-(\delta_j) = \Lambda_i^+(\delta_i) > \Lambda_k^+(0) = \Lambda_k^-(0)$ for $i \neq k \neq j$. Then $p' = p + \delta_i^{(i)} + \delta_j^{(j)}$ is an element of $\mathcal{D}$.

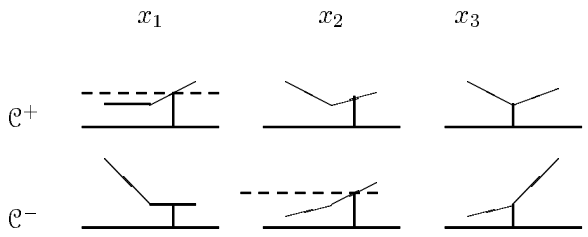


Figure 6: The example after changing x_2

Now the point is reached, where the variables x_k for $i \neq k \neq j$ are not dominating anymore in any rule. Therefore they are irrelevant. Only x_i and x_j are relevant (not semi-relevant) variables in p'. For the total membership degrees Λ_M^+ and Λ_M^- of the point p' to $\mathcal{C}^+$ and $\mathcal{C}^-$, we obtain

$$\Lambda_M^+(p + \bar{\delta}) = \max\{\Lambda_i^+(\delta_i), \Lambda_j^+(\delta_j)\} \text{ and } \Lambda_M^-(p + \bar{\delta}) = \max\{\Lambda_i^-(\delta_i), \Lambda_j^-(\delta_j)\}.$$

Because of having changed the variables sufficiently small, $\mathcal{R}'_p$ does not contain any new rules. Therefore, we have $\Lambda^+_{x_i+\delta_i}(\varepsilon) = \Lambda^+_{x_i}(\varepsilon+\delta_i)$ in $\mathcal{N}_\varepsilon(p)$. This means that within the domain we can still use the same functions of membership degree for p' as for p.
When taking $p'_i = p_i + \delta_i$ and changing it into that direction where the membership degree for $\mathcal{C}^+$ is increasing, this yields a decision for $\mathcal{C}^+$. When moving p'_i towards the other direction the membership degree for $\mathcal{C}^+$ is lowered and we get a decision for $\mathcal{C}^-$. The same applies to $p'_j = p_j + \delta_j$ and $\mathcal{C}^-$, so that x_i and x_j are both relevant variables.

Remark 3 *Because the fuzzy sets have (directed) derivatives, it is possible to vary p_i by an arbitrary small value, so that the value needed to vary p_j is small enough so that p' is in the ε-neighbourhood $\mathcal{N}_\varepsilon(p)$ of p.*

Proof: μ_i and μ_j have a bounded slope on $\mathcal{N}_\varepsilon(p)$, because they are also differentiable on the closure $\overline{\mathcal{N}}_\varepsilon(p)$. Therefore, it is possible to calculate $\max_{p'_i \in \overline{\mathcal{N}}_{\frac{\varepsilon}{n}}(p_i)}\{||\mu_i(p'_i) - \mu_i(p_i)||\}$ and $\max_{p'_j \in \overline{\mathcal{N}}_{\frac{\varepsilon}{n}}(p_j)}\{||\mu_j(p'_j) - \mu_j(p_j)||\}$. We define

$$\Delta\mu := \min\{\max_{p'_i \in \overline{\mathcal{N}}_{\frac{\varepsilon}{n}}(p_i)} ||\mu_i(p'_i) - \mu_i(p_i)||; \max_{p'_j \in \overline{\mathcal{N}}_{\frac{\varepsilon}{n}}(p_j)} ||\mu_j(p'_j) - \mu_j(p_j)||\}.$$

Now we can choose $p'_i \in \mathcal{N}_{\frac{\varepsilon}{n}}(p_i)$, so that $||\mu_i(p'_i) - \mu_i(p_i)|| = \Delta\mu$, and $p'_j \in \mathcal{N}_{\frac{\varepsilon}{n}}(p_j)$ so that $||\mu_j(p'_j) - \mu_j(p_j)|| = \Delta\mu$.
When taking the Euclidian norm, we obtain

$$\begin{aligned} ||p' - p|| &= \sqrt{\textstyle\sum_{k=1}^n (p'_k - p_k)^2} \\ &\leq \sqrt{\textstyle\sum_{k=1}^n \left(\frac{\varepsilon}{n}\right)^2} = \sqrt{n \cdot \left(\frac{\varepsilon}{n}\right)^2} \\ &= \tfrac{1}{\sqrt{n}}\varepsilon < \varepsilon, \end{aligned}$$

which proves that $p' \in \mathcal{N}_\varepsilon(p)$ holds. When considering another norm we just have to take $\frac{\varepsilon}{\alpha}$ with (another α instead of n) instead of $\frac{\varepsilon}{n}$. □

4.2 Second case: There are rules contributing to the maximum firing degree with more than one dominating variable

In this section we consider the case that there are rules with two or more dominating variables. If there are also rules having only one relevant variable, we use Λ^+_i and Λ^-_i as already described for these rules.
First we consider a single rule R with two or more dominating variables. When the dominating variable x_i is varied there are three possibilities: If μ_i is increasing, x_i is not dominating any more, so that there is one dominating variable less in this rule, but the rule is still firing. The case of μ_i remaining constant is trivial, and if μ_i is decreasing, the firing degree of the whole rule is decreasing. In no case the firing degree μ_R of the rule R is increasing.

Because of x_i being a relevant or semi-relevant variable, varying p_i has to lead into $\mathcal{C}^+$ or $\mathcal{C}^-$. Without loss of generality we consider $\mathcal{C}^+$. We can reach $\mathcal{C}^+$ iff

1. Λ_i^+ is increasing, and if we have that $\Lambda_i^- \not\equiv 0$ is also increasing, then Λ_i^+ has to increase faster than Λ_i^-.

2. Every rule R firing for $\mathcal{C}^-$ with maximum degree has x_i as a dominating variable, and in every rule $\mu_R^{(x_i)}$ is decreasing, so that the total membership degree $\mu_{\mathcal{C}^-}^{(\mathcal{R})}$ for $\mathcal{C}^-$ is decreasing. If all the rules firing for $\mathcal{C}^+$ with maximum degree have x_i as a dominating variable, too, then $\mu_{\mathcal{C}^-}^{(\mathcal{R})}$ has to decrease faster than $\mu_{\mathcal{C}^+}^{(\mathcal{R})}$.

In any case there are at least two variables x_i and x_j with (1.) being satisfied for x_i for reaching $\mathcal{C}^+$ and for x_j for reaching $\mathcal{C}^-$ or with (2.) being satisfied for x_i and $\mathcal{C}^+$ and for x_j and $\mathcal{C}^-$, because of the following:
Suppose that the variables x_k, $k \in A \subseteq \{1, \cdots, n\}$, being relevant or semi-relevant for (without loss of generality) $\mathcal{C}^+$ are satisfying (1.), when p_k is moved by δ_k. This means that $\Lambda_i^+ \not\equiv 0$ holds for these variables in the direction of the movement. Furthermore, suppose that the variables x_l, $l \in B \subseteq \{1, \cdots, n\}$, being relevant or semi-relevant for $\mathcal{C}^-$ are satisfying (2.), when p_l is moved by δ_l. This means that all these variables are dominating in every rule firing for $\mathcal{C}^+$. If we have more than two relevant or semi-relevant variables, this is a contradiction, because there is at least one rule for $\mathcal{C}^+$ having only one dominating variable x_i, so that the other variables x_l, $l \neq i$, cannot satisfy (2.).
Now we have to consider the two cases that are left:

1. At least for two variables (without loss of generality x_i and x_j) there are rules with only this variable dominating and with $\mu_{\mathcal{C}^+}$ increasing when varying p_i and $\mu_{\mathcal{C}^-}$ increasing when varying p_j. Then the procedure is the same as described in section 4.1.

2. Without loss of generality every rule firing for $\mathcal{C}^-$ has x_i as one dominating variable, and in each such rule $\mu_R^{(i)}$ is decreasing when p_i is varied into the right direction. The same is the case for x_j and $\mathcal{C}^+$.

   We vary p_i by $\delta_i \neq 0$, so that this leads into $\mathcal{C}^+$, because $\mu_R^{(i)}$ is decreasing in every rule R firing for $\mathcal{C}^-$ with maximum degree and with it $\mu_{\mathcal{C}^-}^{(\mathcal{R})}$.

   When there is at least one rule R firing for $\mathcal{C}^+$ that does not have x_i as a dominating variable or with $\mu_R^{(i)}$ not decreasing, then $\mu_R^{\mathcal{C}^+}$ is not changing. Otherwise $\mu_R^{(i)}$ does decrease more slowly, so that we still have $\mu_{\mathcal{C}^+}^{(\mathcal{R})}(x+\delta_i^{(i)}) > \mu_{\mathcal{C}^-}^{(\mathcal{R})}(x+\delta_i^{(i)})$, because x_i is relevant or semi-relevant for $\mathcal{C}^+$.

   After this the variation of p_j by δ_j leads back to $\mathcal{D}$, because for $\mathcal{C}^+$ every membership degree $\mu_R^{(j)}$ is decreasing until

$$\mu_{\mathcal{C}^-}^{(\mathcal{R})}(p+\delta_i^{(i)}+\delta_j^{(j)}) = \mu_{\mathcal{C}^+}^{(\mathcal{R})}(p+\delta_i^{(i)}+\delta_j^{(j)}).$$

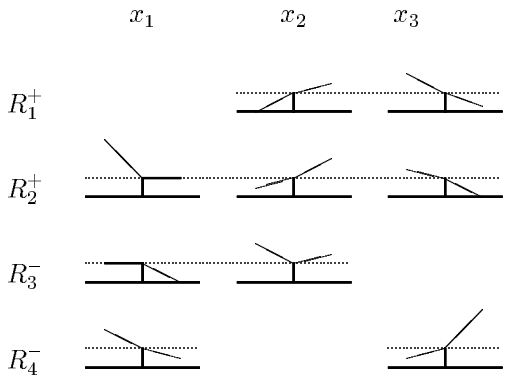


Figure 7: An example for three semi-relevant variables with two rules firing for $\mathcal{C}^+$ and two firing for $\mathcal{C}^-$, $x_i = x_1$ and $x_j = x_2$.

5 Conclusions

We have shown that fuzzy max-min classification systems assign data to a class locally (mainly) on the basis of two attributes. The set of points for which this property is satisfied, is a dense set within the the class boundaries. Although this sounds like a negative result, it has also positive aspects. First of all, the result holds only locally so that the classification system can still take all attributes into account, when we consider it from a global point of view. And although the local reduction to two variables seems to be very restrictive, it is positive in terms of interpretability. Since we usually want interpretable fuzzy rules, this property definitely helps to understand the rules – especially when we take into account that humans usually do not consider a larger number of attributes simultaneously. Our result can also be applied to examine a fuzzy max-min classification system, i.e. which attributes are relevant in which region.
It should also be noted that we can at least approximate any kind of (continuous) class boundaries by fuzzy max-min classification system and that we replace the maximum or minimum by another t-conorm or t-norm, the situation is completely different [6].

6 Appendix

When two functions f_1 and f_2 are given that have a Taylor expansion in x_0, we want to use the Taylor expansions to know which function has the greater values, when going from x_0 into one direction.
Therefore, we take the first term of the Taylor expansion that is different for the two functions. Without loss of generality we have $f_1^{(n)}(x_0) > f_2^{(n)}(x_0)$ (and with this the n^{th} coefficient of the Taylor expansion of f_1 is greater than that one of

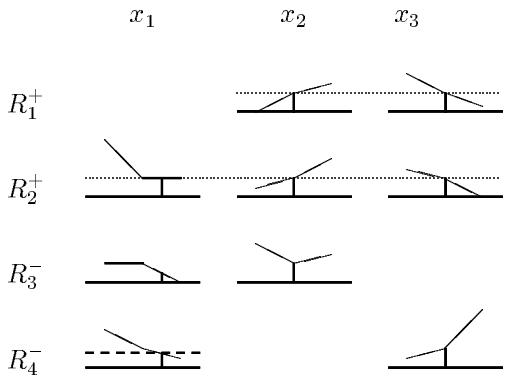
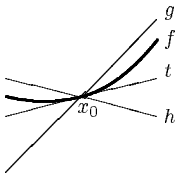


Figure 8: The example after having pushed x_1. R_1^+ and R_2^+ did not change, the firing degree of R_4^- decreased with x_3 not being dominating any more, and because of $\mu_{R_3^-} < \mu_{R_4^-}$ R_3^- is not firing any more.

f_2), while for $i \in \{1, \cdots, n\}$ we have $f_1^{(i)}(x_0) = f_2^{(i)}(x_0)$. Then the values of f_1 are greater than those for f_2, when $x > x_0$, and the other way round for $x < x_0$ in a neighbourhood $\mathcal{N}_\varepsilon(x_0)$ of x_0, as the following argumentation will show.

Lemma 3 *Assume $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable in a neighbourhood of x_0 and let t be the tangent to f at point x_0. t has the slope $m_t = f'(x_0)$. Let g and h be straight lines with $g(x_0) = h(x_0) = f(x_0)$ and with g having slope $m_g > m_t$ and h having slope $m_h < m_t$.*
Then there is an $\varepsilon > 0$ so that in $\mathcal{N}_\varepsilon(x_0)$ f lies between g and h. This means:

$$\forall x \in \mathcal{N}_\varepsilon(x_0), x < x_0 : \quad g(x) < f(x) < h(x) \quad \textit{and}$$
$$\forall x \in \mathcal{N}_\varepsilon(x_0), x > x_0 : \quad g(x) > f(x) > h(x).$$



Proof: We can write $t(x) = f(x_0) + f'(x_0) \cdot (x - x_0)$ and with this

$$f(x) = t(x) + r(x)(x - x_0)^2$$

with $r(x_0) = 0$ and r being a continuous function. r being continuous means

$$\forall\, \delta > 0\, \exists\, \varepsilon > 0 : (|x - x_0| < \varepsilon \Rightarrow |r(x) - r(x_0)| = |r(x)| < \delta).$$

Considering $x > x_0$ with $x - x_0 \leq 1$ we obtain

$$g(x) - f(x) = \underbrace{\underbrace{(\underbrace{m_g - f'(x_0)}_{=:\delta>0} - \underbrace{r(x) \cdot (x - x_0)}_{<\delta \cdot 1})}_{>0} \cdot \underbrace{(x - x_0)}_{>0}}_{} > 0,$$

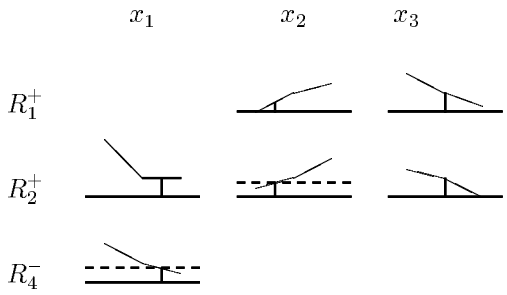


Figure 9: The example after having pushed x_2. The firing degree of R_1^+ and R_2^+ is given by x_2. Because of $\mu_{R_1^+} < \mu_{R_2^+}$ R_1^+ is nor firing any more. We have $\mu_{\mathcal{C}^+} = \mu_{R_2^+} = \mu_{R_4^-} = \mu_{\mathcal{C}^-}$.

because r is continuous. So we have $g(x) > f(x)$ for $x > x_0$, $x \in \mathcal{N}_\varepsilon(x_0)$. By the same argument we obtain $f(x) > h(x)$ for $x > x_0$ and $h(x) > f(x) > g(x)$ for $x < x_0$ within $\mathcal{N}_\varepsilon(x_0)$. □

Lemma 4 *Assume f_1 and f_2 are twice differentiable in a neighbourhood of x_0 and let $f_1(x_0) = f_2(x_0)$, but $f_1'(x_0) > f_2'(x_0)$. Then there are a straight line g with $g(x_0) = f_1(x_0) = f_2(x_0)$ and an $\varepsilon > 0$ so that g lies between f_1 and f_2 within a neighbourhood $\mathcal{N}_\varepsilon(x_0)$ of x_0. This means that*

$$\begin{array}{lll} f_1(x) < g(x) < f_2(x) & \text{for } x < x_0 & \text{and} \\ f_1(x) > g(x) > f_2(x) & \text{for } x > x_0. & \end{array}$$

Proof: Define g by

$$g(x) = f_1(x_0) + \frac{f_1'(x_0) + f_2'(x_0)}{2}(x - x_0)$$

with slope $m_g = \frac{f_1'(x_0)+f_2'(x_0)}{2}$. Adding two straight lines h_1 and h_2 with $h_1(x_0) = h_2(x_0) = g(x_0)$ with slopes $m_{h_1} > f_1'(x_0)$ and $m_{h_2} < f_2'(x_0)$ we can apply Lemma 3 to show that there is an ε so that g lies between f_1 and f_2 in $\mathcal{N}_\varepsilon(x_0)$. □

Lemma 5 *Assume, the functions f_1 and f_2 are $(n+2)$ times differentiable in a neighbourhood of x_0 and let $f_1^{(i)}(x_0) = f_2^{(i)}(x_0)$ for $i = 0, \cdots, n$, but $f_1^{(n+1)}(x_0) > f_2^{(n+1)}(x_0)$. Then there is an $\varepsilon > 0$ so that in $\mathcal{N}_\varepsilon(x_0)$ we have $f_1(x) < f_2(x)$ for $x < x_0$ and $f_1(x) > f_2(x)$ for $x > x_0$.*

Proof: We give a proof by induction:
Beginning of induction ($n = 0$): Let $f_1(x_0) = f_2(x_0)$ and $f_1'(x_0) > f_2'(x_0)$. Because of lemma 4 we can put a straight line between f_1 and f_2. So we have $f_1(x) > f_2(x)$ for $x > x_0$, $x \in \mathcal{N}_\varepsilon(x_0)$, and the other way round for $x < x_0$, $x \in \mathcal{N}_\varepsilon(x_0)$.

Induction hypothesis: When we have $\tilde{f}_1^{(i)}(x_0) = \tilde{f}_2^{(i)}(x_0)$ for $i = 0, ..., n-1$ and $\tilde{f}_1^{(n)}(x_0) > \tilde{f}_2^{(n)}(x_0)$, then there is an $\varepsilon > 0$ so that $\tilde{f}_1(x) > \tilde{f}_2(x)$ for $x > x_0$, $x \in \mathcal{N}_\varepsilon(x_0)$, and the other way round for $x < x_0$, $x \in \mathcal{N}_\varepsilon(x_0)$.
Induction step: We have $f_1^{(i)}(x_0) = f_2^{(i)}(x_0)$ for $i = 1, ..., n$ and $f_1^{(n+1)}(x_0) > f_2^{(n+1)}(x_0)$.When defining $\tilde{f}_1 := f_1'$ and $\tilde{f}_2 := f_2'$ we can use the hypothesis and calculate for $x = x_0 + \delta$, $0 < \delta < \varepsilon$:

$$\begin{aligned} f_1(x) - f_2(x) &= f_1(x_0+\delta) - f_2(x_0+\delta) \\ &= \int_0^\delta (f_1'(x_0+t) - f_2'(x_0+t))dt \\ &= \int_0^\delta \underbrace{(\tilde{f}_1(x_0+t) - \tilde{f}_2(x_0+t))}_{\geq 0} dt > 0, \end{aligned}$$

because we have $\tilde{f}_1(x_0 + t) > \tilde{f}_2(x_0 + t)$ for $0 < t < \varepsilon$. Therefore, we obtain $f_1(x) > f_2(x)$ for $x > x_0$, $x \in \mathcal{N}_\varepsilon(x_0)$. The same can be carried out for $x < x_0$, $x \in \mathcal{N}_\varepsilon(x_0)$. □

References

[1] P. Bauer, E.P. Klement, A. Leikermoser, B. Moser: Interpolation and approximation of real input-output functions using fuzzy rule bases. In: [10], 245-254

[2] J.L. Castro, E. Trillas, S. Cubillo: On consequence in approximate reasoning. Journal of Applied Non-Classical Logics **4** (1994), 91-103

[3] O. Cordón, M. José del Jesus, F. Herrera: Analysing the reasoning mechanism in fuzzy rule based classification systems. Mathware & Soft Computing 5 (1998), 321-332

[4] H. Genther, M. Glesner: Automatic generation of a fuzzy classification system using fuzzy clustering methods. Proc. ACM Symposium on Applied Computing (SAC'94), Phoenix (1994), 180-183

[5] H. Ishibuchi: A fuzzy classifier system that generates linguistic rules for pattern classification problems. In: Fuzzy Logic, Neural Networks, and Evolutionary Computation, Springer, Berlin (1996), 35-54

[6] F. Klawonn, E.P. Klement: Mathematical analysis of fuzzy classifiers. In: Mathematical analysis of fuzzy classifiers. In: X. Liu, P. Cohen, M. Berthold (eds.): Advances in intelligent data analysis. Springer, Berlin (1997), 359-370

[7] F. Klawonn, R. Kruse: Derivation of fuzzy classification rules from multidimensional data. In: G.E. Lasker, X. Liu (eds.): Advances in intelligent data analysis. The International Institute for Advanced Studies in Systems Research and Cybernetics, Windsor, Ontario (1995), 90-94

[8] B. Kosko: Fuzzy systems as universal approximators. Proc. IEEE International Conference on Fuzzy Systems 1992, San Diego (1992), 1153-1162

[9] R. Kruse, J. Gebhardt, F. Klawonn: Foundations of fuzzy systems. Wiley, Chichester (1994)

[10] R. Kruse, J. Gebhardt, R. Palm (eds.): Fuzzy systems in computer science. Vieweg, Braunschweig (1994)

[11] J. Lee, S. Chae: Analysis on function duplicating capabilities of fuzzy controllers. Fuzzy Sets and Systems **56** (1993), 127-143

[12] E.H. Mamdani, S. Assilian: An experiment in linguistic synthesis with a fuzzy logic controller. Intern. Journ. of Man Machine Studies 8 (1975), 1-13

[13] K.D. Meyer Gramann: Fuzzy classification: An overview. In: [10], 277-294

[14] B. Moser: Sugeno controllers with a bounded number of rules are nowhere dense. Fuzzy Sets and Systems 104 (1999), 269-277

[15] D. Nauck, R. Kruse: NEFCLASS – A neuro-fuzzy approach for the classification of data. In: K.M. George, J.H. Carrol, E. Deaton, D. Oppenheim, J. Hightower (eds.): Applied Computing 1995: Proc. of the 1995 ACM Symposium on Applied Computing. ACM Press, New York (1995), 461-465

[16] A. Nürnberger, A. Klose, R. Kruse: Discussing cluster shapes of fuzzy classifiers. Proc. Conf. of the North American Fuzzy Information Processing Society (NAFIPS'99) (to appear)

[17] H. Takagi, M. Sugeno: Fuzzy identification of systems and its application to modelling and control. IEEE Trans. on Systems, Man, and Cybernetics 15 (1985), 116-132

[18] L.X. Wang: Fuzzy systems are universal approximators. Proc. IEEE International Conference on Fuzzy Systems 1992, San Diego (1992), 1163-1169

[19] R. Weber: Fuzzy-ID3: A class of methods for automatic knowledge acquisition. In: Proc. 2nd International Conference on Fuzzy Logic and Neural Networks, Iizuka (1992), 265-268