# Towards an Axiomatic Treatment of "Interpretability"

Ulrich Bodenhofer* and Peter Bauer†

*Software Competence Center Hagenberg GmbH
A-4232 Hagenberg, Austria
Tel. +43 7236 3343 832, Fax +43 7236 3343 888, E-Mail ulrich.bodenhofer@scch.at

†Fuzzy Logic Laboratorium Linz-Hagenberg, Johannes Kepler Universität
A-4040 Linz, Austria
Tel. +43 7236 3343 432, Fax +43 7236 3343 434, E-Mail pete@flll.uni-linz.ac.at

Key Words: computing with words, interpretability, linguistic variable, tuning.

***Abstract***— The aim of this contribution is to point out the importance of interpretability of fuzzy rule-based systems. We try to approach this key property axiomatically, as a starting point, restricting to interpretability of linguistic variables (in Zadeh's sense). The proposed approach is underlined with an intuitive example. Finally, the benefits of considering interpretability in design and tuning of fuzzy systems are discussed.

## 1 Introduction

The brilliant idea of L. A. Zadeh's early work was to utilize what he called "fuzzy sets" as mathematical models of linguistic expressions which cannot be represented in the framework of classical binary logic and set theory in a natural way. In the introduction of his epoch-making article on fuzzy sets [17], he writes:

> *"More often than not, the classes of objects encountered in the real physical world do not have precisely defined criteria of membership. [ ... ] Yet, the fact remains that such imprecisely defined "classes" play an important role in human thinking, particularly in the domains of pattern recognition, communication of information, and abstraction."*

Fuzzy systems, which take advantage of this concept, became a tremendously successful paradigm—a remarkable triumph which started with well-selling applications in consumer goods implemented by Japanese engineers. The reasons for this development are manifold; however, one is usually confronted with the following "classical" arguments:

(1) The main difference between fuzzy systems and other control or decision support systems is that they are parametrized in an interpretable way—by means of rules consisting of linguistic expressions. Fuzzy systems, therefore, allow rapid prototyping as well as easy maintenance and adaptation.

(2) Fuzzy systems offer completely new opportunities to deal with processes for which only a linguistic description is available. They allow to achieve a robust, secure, and reproducible automation of such tasks.

(3) Even if conventional strategies can be employed, reformulating a system's actions by means of linguistic rules can lead to a deeper understanding of its behavior.

We would like to raise the question whether fuzzy systems, as they appear in daily practice, really reflect these—undoubtedly nice—advantages. One may observe that the *possibility to estimate the system's behavior by reading and understanding the rule base only* is a basic requirement for the validity of the above points. If we adopt the usual wide understanding of fuzzy systems (rule-based systems incorporating vague linguistic expressions), we can see, however, that this property—let us call it *interpretability*—is not guaranteed by definition.

In our opinion, interpretability should be *the* key property of fuzzy systems. If it is neglected, one ends up in nothing else than black-box descriptions of input-output relationships for which, without any doubt, other methods which are computationally less expensive could be employed (e.g. neural networks, classical interpolation, statistical methods, etc.).

To summarize the above arguments, we state that fuzzy systems do not offer "white-box" descriptions of input-output relationships by definition. Beside this key statement, it is the purpose of this paper to give a constructive answer how interpretability can be defined in a way which is intuitive and mathematically exact. As a first approach, we leave aspects of fuzzy inference aside and concentrate on the interpretability of verbal values of a linguistic variable. Finally, the practical relevance of these considerations for design and tuning procedures will be explained.

## 2 Formal Definition

Since it has more or less become standard and offers much freedom, in particular with respect to integration of linguistic modifiers and connectives, we start from Zadeh's original definition of linguistic variables [18, 19, 20].

**Definition 1.** A *linguistic variable* is a quintuple of the form

$$(A, T(A), U, G, M),$$

where $A$, $T(A)$, $U$, $G$, and $M$ are defined as follows:

(1) $A$ is the name of the linguistic variable.
(2) $T(A)$ is the set of verbal (linguistic) values of $A$ (the so-called *term set*).
(3) $U$ is the universe of discourse of variable $A$.
(4) $G$ is the definition of the underlying grammar of the term set.
(5) $M$ is a $T(A) \to \mathcal{F}(U)$ mapping which assigns a fuzzy subset of $U$ to each verbal value from $T(A)$.

In the following, unless indicated otherwise, let us assume that the symbols $A$, $T(A)$, $U$, $G$, and $M$ have the meaning as defined above.

In our point of view, the ability to interpret the meaning of a rule base qualitatively relies deeply upon the reader's intuitive understanding of the involved linguistic expressions which, of course, requires knowledge about inherent relationships between the involved linguistic expressions. Therefore, if qualitative estimations are desired, these relationships need to transfer to the underlying semantics, i.e. the fuzzy sets modeling the labels.

In other words, interpretability is strongly connected to the preservation of inherent relationships by the mapping $M$ (according to Definition 1).

The following definition gives an exact mathematical formulation of this property.

**Definition 2.** Consider a linguistic variable $A$. Let $R = (R_i)_{i\in I}$ be a family of relations on the set of verbal values $T(A)$. Provided that every relation $R_i$ has finite arity $a_i$ and a counterpart $Q_i$ on the fuzzy powerset $\mathcal{F}(U)$ with the same arity, the linguistic variable $A$ is called *$R$-$Q$-interpretable* if and only if the following holds:

$$\forall i \in I\ \forall x_1, \dots, x_{a_i} \in T(A): \quad R_i\big(x_1, \dots, x_{a_i}\big) \implies Q_i\big(M(x_1), \dots, M(x_{a_i})\big) \quad (1)$$

For convenience, let us denote the family $(Q_i)_{i\in I}$ simply with $Q$.

**Remark 3.** In case that this is necessary, the generalization of Definition 2 to fuzzy relations is straightforward. If we admit fuzziness of the relations $R_i$ and $Q_i$, the implication in Eq. (1) has to be replaced by

$$R_i\big(x_1, \dots, x_{a_i}\big) \le Q_i\big(M(x_1), \dots, M(x_{a_i})\big).$$

Table 1: Grammar $G$ of linguistic variable $D$.

| | | |
|---|---|---|
| S | := | ⟨exp⟩ ; |
| ⟨exp⟩ | := | ⟨adjective⟩ \| ⟨adverb⟩ ⟨adjective⟩ ; |
| ⟨adjective⟩ | := | "small" \| "medium" \| "large" ; |
| ⟨adverb⟩ | := | "at least" \| "at most" ; |

## 3 A Detailed Study by Means of a Practical Example

In almost all fuzzy control applications, the domains of the system variables are divided into a certain number of fuzzy sets by means of the underlying ordering—a fact which is typically reflected in expressions like "small", "medium", or "large". We will now discuss a simple example involving orderings to illustrate the concrete meaning of Definition 2.

Let us consider a linguistic variable named $D$ (e.g. distance) with a universe of discourse $U = [0, 100]$. The BNF grammar definition [11] shown in Table 1 determines the term set of $D$ as follows:

$$\begin{aligned} T(D) = \{&\text{"small"}, \text{"medium"}, \text{"large"},\\ &\text{"at least small"}, \text{"at least medium"},\\ &\text{"at least large"}, \text{"at most small"},\\ &\text{"at most medium"}, \text{"at most large"}\} \end{aligned}$$

Taking the "background" or "context" (in this example, distance) of the variable into account, almost every human has an intuitive understanding of the qualitative meaning of each of the above linguistic expressions, even if absolutely nothing about the quanititative meaning, i.e. the corresponding fuzzy sets, is known. This understanding, to a major part, can be attributed to elementary relationships between the linguistic values. According to Definition 2, these inherent relationships correspond to the family of relations $R = (R_i)_{i\in I}$.

In our opinion, the most obvious relationships in the example term set $T(D)$ are orderings and inclusions, so let us consider the following two binary relations (for convenience, since both relations are binary, we can switch to infix notations here):

$$R = (\preceq, \sqsubseteq) \quad (2)$$

As an example, a human would intuitively expect an ordering of the adjectives like

$$\text{"small"} \preceq \text{"medium"} \preceq \text{"large"}.$$

Moreover, the following monoticities seem obvious for all adjectives $A, B \in \{\text{"small"}, \text{"medium"}, \text{"large"}\}$:

$$A \sqsubseteq \text{"at least"}\, A$$
$$B \sqsubseteq \text{"at most"}\, B$$

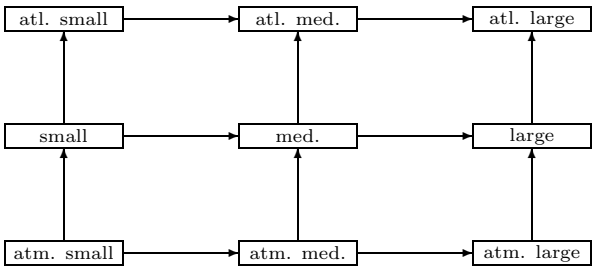


Figure 1: Hasse diagram of ordering relation $\preceq$.

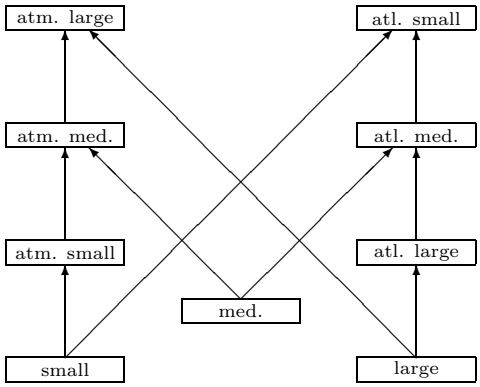


Figure 2: Hasse diagram of inclusion relation $\sqsubseteq$.

$$A \preceq B \Longrightarrow \text{“at least”}A \preceq \text{“at least”}B$$
$$A \preceq B \Longrightarrow \text{“at most”}A \preceq \text{“at most”}B$$

$$A \preceq B \Longrightarrow \text{“at least”}A \sqsupseteq \text{“at least”}B$$
$$A \preceq B \Longrightarrow \text{“at most”}A \sqsubseteq \text{“at most”}B$$

Figures 1 and 2 show Hasse diagrams which fully describe the two relations $\preceq$ and $\sqsubseteq$ (for the sake of simplicity, reflexivity is not explicitly indicated there).

Now we have to fix meaningful fuzzy counterparts of the relations in $R$ on $\mathcal{F}(U)$. We start with the usual inclusion of fuzzy sets according to Zadeh [17].

**Definition 4.** Consider two fuzzy subsets of $X$ denoted $A$ and $B$. $A$ is called *subset* of $B$, short $A \subseteq B$, if and only if, for all $x \in X$,

$$\mu_A(x) \leq \mu_B(x).$$

Consequently, $B$ is called *superset* of $A$.

Concerning orderings, we adopt a simple variant of Bodenhofer's general framework for ordering fuzzy sets [2, 3], which includes well-known orderings of fuzzy numbers based on the extension principle [6, 7].

**Definition 5.** Suppose that a universe $X$ is equipped with a crisp linear ordering $\leq$. Then a preordering $\lesssim$ of

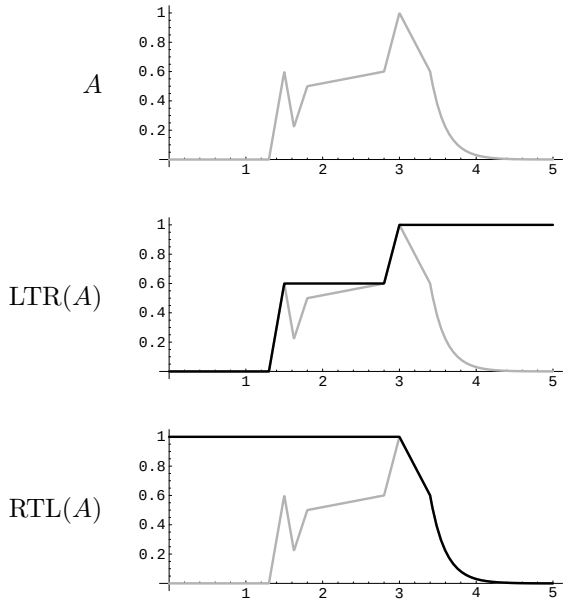


Figure 3: A fuzzy set $A \in \mathcal{F}(\mathbb{R})$ and the results which are obtained when applying the operators LTR and RTL.

fuzzy sets can be defined by

$$A \lesssim B \iff \big(\mathrm{LTR}(A) \supseteq \mathrm{LTR}(B) \wedge \mathrm{RTL}(A) \subseteq \mathrm{RTL}(B)\big),$$

where the operators LTR and RTL are defined as follows:

$$\mu_{\mathrm{LTR}(A)}(x) = \sup\{\mu_A(y) \mid y \leq x\}$$
$$\mu_{\mathrm{RTL}(A)}(x) = \sup\{\mu_A(y) \mid y \geq x\}$$

Figure 3 shows an example what the operators LTR and RTL give for a non-trivial fuzzy set. It may be easy to see that LTR always yields the smallest superset with non-decreasing membership function, while RTL yields the smallest superset with non-increasing membership function [1, 2].

Finally, we can write down the set of counterpart relations $Q$ according to Definition 2:

$$Q = (\lesssim, \subseteq) \tag{3}$$

The only remaining component of the linguistic variable $D$ is the mapping $M$ which provides the semantics of the expressions in $T(D)$, i.e. which assigns a fuzzy set to each linguistic value. Concerning interpretability, this mapping plays the crucial role—$R$-$Q$-interpretability now exactly means that the obvious orderings and inclusions of linguistic terms (as shown in Figures 1 and 2) must not be violated by the corresponding fuzzy sets.

The first basic requirement which must be fulfilled by the mapping $M$ is that the fuzzy sets associated with the three adjectives must be in proper order, i.e.

$$M(\text{“small”}) \lesssim M(\text{“medium”}) \lesssim M(\text{“large”}). \tag{4}$$

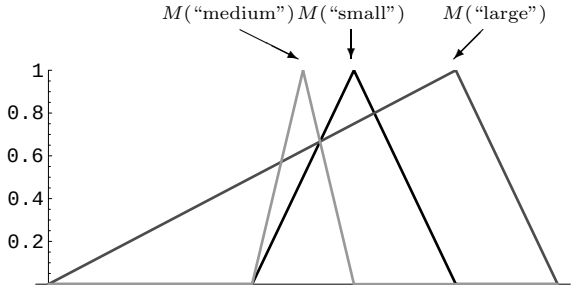


Figure 4: A non-interpretable setting.

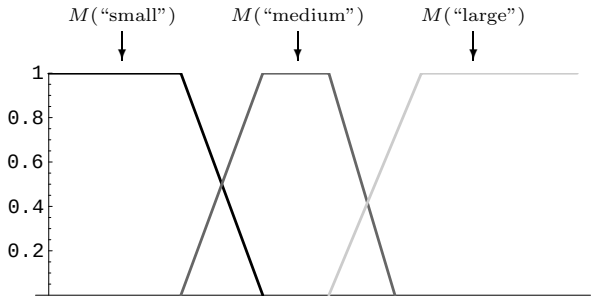


Figure 5: An example of an interpretable setting.

It is easy to observe that this basic ordering is violated in Figure 4, while it is fulfilled by the fuzzy sets in Figure 5 (for more details on the preordering $\lesssim$, see [2, 3]).

Before we can check $R$-$Q$-interpretability of $D$, the semantics of linguistic expressions containing an adverb ("at least" or "at most") have to be defined. Two different ways are possible: One variant is to define a separate fuzzy set for each expression, the second variant uses fuzzy modifiers to define the semantics of such expressions, i.e. the semantics of an adverb is modeled by a $\mathcal{F}(U) \to \mathcal{F}(U)$ function. Since it is, by far, simpler and easier to handle with respect to interpretability, we strongly suggest the second variant. In this example, it is straightforward to use the fuzzy modifiers introduced in Definition 5:

$$M(\text{"at least"}\,A) = \mathrm{LTR}\big(M(A)\big)$$
$$M(\text{"at most"}\,A) = \mathrm{RTL}\big(M(A)\big)$$

**Proposition 6.** *Let us consider the linguistic variable $D$ as defined above. If the relation families $R$ and $Q$ are defined as in Eq. (2) and (3), respectively, and the fuzzy sets associated with the adjectives "small", "medium", and "large" are normalized and fulfill the ordering condition (4), then $D$ is $R$-$Q$-interpretable.*

*Proof.* The following basic properties hold for all normalized fuzzy sets $A, B \in \mathcal{F}(U)$ [2]:

$$A \subseteq \mathrm{LTR}(A) \tag{5}$$
$$A \subseteq \mathrm{RTL}(A) \tag{6}$$
$$\mathrm{LTR}\big(\mathrm{LTR}(A)\big) = \mathrm{LTR}(A) \tag{7}$$
$$\mathrm{RTL}\big(\mathrm{RTL}(A)\big) = \mathrm{RTL}(A) \tag{8}$$
$$\mathrm{LTR}\big(\mathrm{RTL}(A)\big) = \mathrm{RTL}\big(\mathrm{LTR}(A)\big) = U \tag{9}$$
$$A \subseteq B \implies \mathrm{LTR}(A) \subseteq \mathrm{LTR}(A) \tag{10}$$
$$A \subseteq B \implies \mathrm{RTL}(A) \subseteq \mathrm{RTL}(A) \tag{11}$$

Since the relations $\subseteq$ and $\lesssim$ are reflexive and transitive [2], it is sufficient to prove the relations indicated by arrows in the two Hasse diagrams (see Figures 1 and 2).

Let us start with the ordering relation. The validity of the relations in the middle row is an assumption which we need not prove. The relations in the two other rows follow directly from the following two relationships which can be proved using Eq. (7), (8), and (9):

$$A \lesssim B \implies \mathrm{LTR}(A) \lesssim \mathrm{LTR}(B)$$
$$A \lesssim B \implies \mathrm{RTL}(A) \lesssim \mathrm{RTL}(B)$$

The three vertical relationships in Figure 1 follow directly from

$$\mathrm{RTL}(A) \lesssim A \lesssim \mathrm{LTR}(A)$$

which can be shown using (5), (6), (10), and (11).

The relations in the Hasse diagram in Figure 2 follow from Eq. (5), (6), and the definition of the preordering $\lesssim$ (cf. Definition 5). □

Obviously, the notion of interpretability of $D$ (with respect to the families $R$ and $Q$) does not necessarily correspond to human intuition. It is a rather weak necessary condition which is by far not sufficient for interpretability in a stricter sense. The intention of this example was to illustrate the concrete meaning and practical relevance of Definition 2. However, if more advanced relationships are desired, these have to be included in the relation families $R$ and $Q$. In order to illustrate the richness of the proposed concept, we briefly mention a few examples of fundamental relationships which can be expressed in this framework:

- It is possible to make elementary assumptions about properties of the fuzzy sets by including unary relations, i.e. predicates (e.g. convexity, normality, etc.).
- If we add symbols for conjunction, disjunction, the whole universe, and the empty set, it is even possible to force partition constraints (disjointness and coverage properties) by means of the inclusion relation. Table 2 shows an example of a grammar containing such constructs.

Table 2: Grammar $G$ of linguistic variable $D'$.

| | | |
|---|---|---|
| S | := | ⟨exp⟩ ; |
| ⟨exp⟩ | := | ⟨adjective⟩ \| ⟨adverb⟩ ⟨adjective⟩ \| "(" ⟨exp⟩ ")" ⟨binary⟩ "(" ⟨exp⟩ ")"; |
| ⟨adjective⟩ | := | "small" \| "medium" \| "large" \| "anything" \| "undefined" ; |
| ⟨adverb⟩ | := | "at least" \| "at most" ; |
| ⟨binary⟩ | := | "and" \| "or" ; |

# 4 Applications

## 4.1 Tuning

Automatic design and tuning of fuzzy systems has become a central issue in machine learning and data analysis. In the last years, a vast number of scientific publications dealt with this problem. Most of them, however, disregard the importance of interpretability—leading to results which are, actually, black-box functions (typical pictures like in Figure 4 can be found in several papers [9, 14, 15, 16]) that do not provide any meaningful linguistic information. More recent research [12, 13] slowly starts to take the interpretability aspect into account, although still lacking a general theoretical foundation.

One may argue that proper input-output behavior is the central goal of automatic tuning. Recalling Point (3) from Section 1, however, clearly demonstrates that considering interpretability is indispensable.

## 4.2 Design of Complex Fuzzy Systems

As long as the top-down construction of small fuzzy systems (e.g. two-input single-output fuzzy controllers) is concerned, interpretability is usually not such an important issue, since the system is simple enough that a conscious user will refrain from making settings which contradict his/her intuition.

However, in the design of complex fuzzy systems with a large number of variables and rules, interpretability is a crucial point. Integrating tools which guide the user through the design of a large fuzzy system by preventing the him/her from making non-interpretable settings accidentally could be extremely helpful. As a matter of fact, debugging of large fuzzy systems becomes a tedious task if it is not guaranteed that the intuitive meanings of the labels used in the rule base are reflected in their corresponding semantics.

To be more precise, the goal is not to bother the user with additional theoretical aspects. Instead, the idea is to integrate these aspects into software tools for fuzzy systems design, but not necessarily transparent for the user, with the aim that he/she can build interpretable fuzzy systems in an even easier way than with today's software tools.

## 4.3 Rule Base Simplification

Let us assume that the families $R$ and $Q$ contain equality relations (implicitly contained in inclusion relations). If a long linguistic expression is equal to a shorter one, such an equality relationship can be understood as a *simplification rule*, e.g.

$$(\text{“at least medium”} \wedge \text{“at most medium”}) \equiv \text{“medium”}.$$

Then interpretability in the sense of Definition 2 automatically guarantees that these simplification rules also hold on the semantic level. As a consequence, under the assumption of $R$-$Q$-interpretability, rule simplification can be done on the syntactical level (i.e. the linguistic labels) only.

This could be particularly useful for simplifying complex rule bases or for grammar-based rule base optimization methods (e.g. decision tree induction [8, 10] and fuzzy genetic programming [4, 5]).

# 5 Concluding Remarks

This paper should be understood as a pleading for the importance of interpretability of fuzzy rule-based systems. In order to approach this key property in a systematic and mathematically exact way, we have proposed to make implicit relationships between the linguistic labels explicit by formulating them as (fuzzy) relations. Then interpretability corresponds to the preservation of this relationships by the associated meaning. This idea has been illustrated and demonstrated extensively by means of a simple example. Moreover, possible applications in fuzzy systems design and tuning have been discussed.

# Acknowledgements

Ulrich Bodenhofer is working in the framework of the *Kplus Competence Center Program* which is funded by the Austrian Government, the Province of Upper Austria, and the Chamber of Commerce of Upper Austria.

## References


[1] U. Bodenhofer. The construction of ordering-based modifiers. In G. Brewka, R. Der, S. Gottwald, and A. Schierwagen, editors, *Fuzzy-Neuro Systems '99*, pages 55–62. Leipziger Universitätsverlag, 1999.

[2] U. Bodenhofer. *A Similarity-Based Generalization of Fuzzy Orderings*, volume C 26 of *Schriftenreihe der Johannes-Kepler-Universität Linz*. Universitätsverlag Rudolf Trauner, 1999.

[3] U. Bodenhofer. A general framework for ordering fuzzy alternatives with respect to fuzzy orderings. In *Proc. IPMU 2000*, 2000. (to appear).

[4] A. Geyer-Schulz. *Fuzzy Rule-Based Expert Systems and Genetic Machine Learning*, volume 3 of *Studies in Fuzziness*. Physica Verlag, Heidelberg, 1995.

[5] A. Geyer-Schulz. The MIT beer distribution game revisited: Genetic machine learning and managerial behavior in a dynamic decision making experiment. In F. Herrera and J. L. Verdegay, editors, *Genetic Algorithms and Soft Computing*, pages 658–682. Physica Verlag, Heidelberg, 1996.

[6] E. E. Kerre, M. Mareš, and R. Mesiar. On the orderings of generated fuzzy quantities. In *Proc. IPMU'98*, volume 1, pages 250–253, 1998.

[7] L. T. Kóczy and K. Hirota. Ordering, distance and closeness of fuzzy sets. *Fuzzy Sets and Systems*, 59(3):281–293, 1993.

[8] R. S. Michalski, I. Bratko, and M. Kubat. *Machine Learning and Data Mining*. John Wiley & Sons, Chichester, 1998.

[9] K. Mitsubuchi, S. Isaka, and Z. Y. Zhao. A fuzzy rule generation system. In *Proc. IFSA'93*, volume I, pages 11–14, 1993.

[10] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[11] A. Ralston, E. D. Reilly, and D. Hemmendinger, editors. *Encyclopedia of Computer Science*. Groves Dictionaries, Williston, VT, 4th edition, 2000.

[12] M. Setnes and J. A. Roubos. GA-fuzzy modeling and classification: Complexity and performance. *IEEE Trans. Fuzzy Systems*. (to appear).

[13] M. Setnes, R. Babuška, and H. B. Verbruggen. Rule-based modeling: Precision and transparency. *IEEE Trans. Syst. Man Cybern., Part C: Applications & Reviews*, 28:165–169, 1998.

[14] K. Shimojima, T. Fukuda, and Y. Hasegawa. Self-tuning fuzzy modeling with adaptive membership function, rules, and hierarchical structure based on genetic algorithm. *Fuzzy Sets and Systems*, 71(3):295–309, 1995.

[15] H. Surmann, A. Kanstein, and K. Goser. Self-organizing and genetic algorithms for an automatic design of fuzzy control and decision systems. In *Proc. EUFIT'93*, volume II, pages 1097–1104, 1993.

[16] G. V. Tan and X. Hu. On designing fuzzy controllers using genetic algorithms. In *Proc. FUZZ-IEEE'96*, volume II, pages 905–911, 1996.

[17] L. A. Zadeh. Fuzzy sets. *Inf. Control*, 8:338–353, 1965.

[18] L. A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning I. *Inform. Sci.*, 8:199–250, 1975.

[19] L. A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning II. *Inform. Sci.*, 8:301–357, 1975.

[20] L. A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning III. *Inform. Sci.*, 9:43–80, 1975.