

# Some Fundamental Interpretability Issues in Fuzzy Modeling

**Corrado Mencar**  
 Department of Informatics  
 University of Bari, Italy  
 mencar@di.uniba.it

**Giovanna Castellano**  
 Department of Informatics  
 University of Bari, Italy  
 castellano@di.uniba.it

**Anna M. Fanelli**  
 Department of Informatics  
 University of Bari, Italy  
 fanelli@di.uniba.it

## Abstract

Interpretability is a fundamental requirement for fuzzy models that has not been exhaustively addressed in literature. This paper rises some fundamental questions concerning interpretability with the aim of promoting deeper insights in the study and application of this property in fuzzy modeling.

**Keywords:** Interpretability, fuzzy modeling.

## 1 Introduction

In the last decade, a great emphasis has been laid to the role of Fuzzy Logic (in the broad sense) in representing knowledge embodied in Computational Intelligence models. Besides being a powerful mathematical tool for modeling techniques, Fuzzy Logic has been recognized as a robust formal underpinning for representing a kind of knowledge that appears close to that embodied in human minds [16].

In consequence of such viewpoint, many researchers arose the question of whether – and eventually how – the knowledge acquired by a fuzzy model<sup>1</sup> through learning techniques could be accessible by human users. In other words, those researchers came up with a number of *interpretability* issues in fuzzy modeling (besides other issues, such as accuracy), which have led to the development of a new flourishing research direction (see, e.g. [1] for some relevant contributions). An important driving force toward this direction

<sup>1</sup>The term “fuzzy model” is hereafter intended to cover all models based on Fuzzy Logic, in particular fuzzy rule-based models.

is due to emerging industrial needs for making results of massive computations understandable by the user. Knowledge Discovery from Data (KDD), and specifically Data Mining, is surely the most emblematic field where such needs are particularly felt [7].

Interestingly, the study of interpretability issues in fuzzy modeling runs in parallel with a still open-ended study for *comprehensibility* in Machine Learning (in the classical AI sense) [2]. Such study sheds light on some interesting issues for deeper insights on interpretability of fuzzy models, which are however rarely explored in fuzzy literature.

This paper is aimed at reconsidering the issues arisen for the “Comprehensibility Problem” in Machine Learning (ML) within the Fuzzy Logic paradigm, with the main purpose of promoting deeper insights on interpretability issues in fuzzy modeling, as well as to open new directions for scientific investigation.

## 2 Defining interpretability

Fuzzy models are usually acquired from data, through the application of various learning techniques. There are two main purposes for applying learning techniques to a dataset: (i) to achieve an expert system (in the broad sense) that can be used to generate appropriate results; and (ii) to better understand the process that produced the data. While the former objective is mainly concerned with the accuracy of the final model, the latter is more appropriately tackled by taking into account the “interpretability”, or “compre-

hensibility”<sup>2</sup> of the final model.

A first basic question thus concerns the characterization of the blurry notion of interpretability, which appears as a badly defined and hardly measurable concept. In agreement with [19], the notion of comprehensibility makes sense only in the task of communicating some piece of knowledge from one actor (the fuzzy model, in this case) to another (a human user). For this reason, it is convenient to outline a general framework to address comprehensibility. The framework (adapted from [18]) consists of the following elements:

- A model, denoted by  $\mathcal{F}$ , which maps inputs to outputs, designed by means of a language  $\mathcal{L}$  used for specifying the components of the model;
- A user-oriented language  $\mathcal{L}'$ , used to describe the mapping  $\mathcal{F}$  to the human user.  $\mathcal{L}'$  is usually the natural language, or a form close to it. The description of  $\mathcal{F}$  through  $\mathcal{L}'$  is denoted with  $\mathcal{F}'$ .

In classical problem-solvers – such as decision trees, expert systems, etc. – the two languages usually coincide, i.e.  $\mathcal{L} = \mathcal{L}'$ , but for other models the two languages could be very different. In particular, in fuzzy models the language  $\mathcal{L}$  describes the model in terms of compounding fuzzy sets, characterized by their membership functions and related parameters (e.g. Gaussian membership functions with related center and width).  $\mathcal{L}$  is hence related to the so-called *deep structure* of the model [20]. On the other hand, the language  $\mathcal{L}'$  (related to the *surface structure* of the model) describes the fuzzy models in terms of linguistic variables (e.g. “TEMPERATURE IS HOT”). As a consequence a matching mechanism is required to translate the model  $\mathcal{F}$  from the language  $\mathcal{L}$  to the user language  $\mathcal{L}'$ .

Within such general framework, the following “Comprehensibility Postulate”, due to Michalski [11] can be formulated:

The results of computer induction [i.e.  $\mathcal{F}'$ ] should be symbolic descriptions of

<sup>2</sup>Hereafter, the two terms will be treated indistinguishably, until differently specified.

given entities, semantically and structurally similar to those a human expert might produce observing the same entities. Components of these descriptions should be comprehensible as single “chunks” of information, directly interpretable in natural language, and should relate quantitative and qualitative concepts in an integrated fashion.

Based on such postulate, a model  $\mathcal{F}$  can be considered comprehensible (interpretable) if its description  $\mathcal{F}'$  transmitted by the computer under the language  $\mathcal{L}'$  can be computed by the user itself, with no additional means that its own computational facilities [18].

The comprehensibility postulate puts in evidence the key role of “information granules” (i.e. the “chunks” of information) and fuzzy information granules in particular, because they model the vagueness of concepts – typically processed by humans – better than classical crisp information structures. Michalski also highlights the need of representing both quantitative and qualitative concepts in an integrated fashion. In this sense, fuzzy information granules are advantageous as they are naturally conceived to deal with concepts of different nature. They can indeed represent crisp/fuzzy numbers (or vectors in the multi-dimensional case), crisp/fuzzy intervals (resp. hyper-boxes) but also qualitative properties (resp. qualitative relations), which are inherently fuzzy and can be symbolically represented by linguistic terms.

It should be pointed out, however, that the association of linguistic terms to fuzzy information granules is a very delicate task. Indeed, natural language terms implicitly bear a semantics that is shared among all speakers of that language. Let us call this implicit semantics “metaphor”. As an example, the metaphor of the linguistic term “TALL” (referred to the height of a human body) is commonly shared by all English-speaking people. The communication of such term from person  $A$  to person  $B$  immediately highlights in  $B$ ’s mind a (fuzzy) set of body heights that highly matches the (fuzzy) set of body heights in  $A$ ’s mind.

Fuzzy models –especially those acquired from

data— should adopt information granules with a semantics that is also shared by their users if interpretability has to be achieved. This necessary condition does not heavily restrict the flexibility of such models: their learning ability should be able to adjust the semantics of information granules to better adapt to data (i.e. to improve their accuracy). This is common in human beings: the semantics of the term “TALL” in person  $A$  is not required to be perfectly the same of that in  $B$ , but only highly matching, since both  $A$  and  $B$  might have matured the concept of tallness on the basis of a different experience.

### 3 Fundamental interpretability issues

The general framework depicted in the previous Section puts in evidence a number of fundamental issues concerning interpretability that are independent on the adopted learning technique to generate fuzzy models. Such issues should be taken into account when interpretability is the major concern of the modeling process.

#### Who needs interpretability?

The ultimate actor that needs interpretability is the user interacting with the fuzzy model. As a consequence, a fundamental issue concerns the definition of the class of such users, in order to choose the most appropriate language  $\mathcal{L}'$  to describe the model. Designers must then consider the class of users the model is intended for, otherwise efforts for achieving interpretability might be useless. If users do not belong to the modeling sphere, interpretable fuzzy models could be abandoned in favor of black-box models (i.e. models for which  $\mathcal{L}'$  is empty, such as neural networks) for which more powerful learning schemes are available.

#### Why and When interpretability is needed?

Interpretability is needed to: (i) easily and reliably verify the acquired knowledge and to relate it to user’s domain knowledge; (ii) facilitate debugging and improving the fuzzy model and the related learning algorithm; (iii) validate the system, for its maintenance, and for its evolution in view of changes in the external world; (iv) con-

vince the user that the model’s behavior is reliable [18]. Some of such requirements are in conflict with other performance factors, such as accuracy or learning speed, hence interpretability must be accurately balanced with those factors in a modeling context. For example, in fuzzy control, interpretability has a lower priority w.r.t. accuracy, unless the produced fuzzy model has to be used for explaining the behavior of a controlled plant. On the other hand, in decision making the interpretability of a fuzzy model for decision support is of prominent importance, as the user must be “convinced” on a decision suggested by the model. In this sense, fuzzy models have better chances to be adopted in applicative domains such as medical diagnosis or financial forecasting, provided that interpretability is fully taken into account in the design process.

#### What should be interpretable?

Interpretability is mainly required for the results of learning, i.e. for the knowledge base describing the data. Often, such knowledge base is defined in terms of fuzzy rules that combine two information granules, one for the antecedent and the other for the consequent. Interpretability of such information granules is essential, and most of research efforts on interpretable fuzzy models are focused on those objects [3]. However, the association between antecedent and consequent is sometimes disregarded. In fact, the classical “IF-THEN” rule structure is ambiguous, especially in fuzzy models, as it can be interpreted either in an implicative fashion (*if* the antecedent is true, *then* the consequent must be true; if the antecedent is false, the consequent could either true or false) or in a conjunctive fashion (the antecedent is true *and then* the consequent is also true) [4].

Conjunctive rules are often used in fuzzy modeling, because of the t-norm used to infer the consequent degree of truth and the t-conorm used for aggregating the consequents of all rules. However, they are disguised as implicative rules because of the adoption of the terms “IF” and “THEN”. This causes ambiguity and incomprehension of the knowledge base embodied in a fuzzy model. Furthermore, theoretical studies on rule-based fuzzy models that do not make any distinction

between these two types of interpretation might provide misleading results (e.g. inconsistency assessment [5], which does not apply for conjunctive rules). An approach to solve such ambiguity problem could be a standard formal notation for representing fuzzy rules, which would distinguish rules of different semantics. Unfortunately, such notation appears to be still missing in literature.

An interpretable knowledge base may not be sufficient to achieve interpretability of a fuzzy model. Indeed, the output of the fuzzy model must be also interpretable, as well as the procedure by which the output has been produced. Fuzzy outputs are highly desirable, as they convey vagueness information that is inherent in the model's knowledge base and could be well designated by linguistic terms. Unfortunately, standard fuzzy models (e.g. Mamdani-type Fuzzy Inference Systems) are not able to infer interpretable fuzzy output, as the fuzzy sets resulting from inference do not satisfy basic interpretability requirements (such as normality, convexity, etc.), and the association of linguistic terms is rather difficult.

To deal with this difficulty, often defuzzification techniques are used, in order to provide a single numerical value, which has a direct semantics even though it does not convey vagueness information. Nevertheless, the defuzzification procedure that yields the final numerical value should be itself interpretable, i.e. the user should be able to understand the meaning of the numerical output provided by the model. As an example, Center of Gravity is often used for defuzzification, but its interpretation is often unclear in many modeling situations [13]. This issue is rarely addressed and the interpretability of the outputs of fuzzy models is still an open question.

Finally, comprehensibility issues could also be arisen for the learning technique used to generate the fuzzy model from data. Comprehensibility of the learning technique is necessary to understand how knowledge has been generated and what is its ultimate semantics. As an example, the well-known Fuzzy C-Means (FCM) [6] clustering algorithm has some interpretable facets and more obscure ones. Its objective function (1) is indeed interpretable (as it evaluates intra-cluster distances, which must be minimized, and inter-cluster dis-

tances, which must be maximized), but the probabilistic constraint (2) and the fuzzification parameter  $m$  do not have an immediate meaning. Possibilistic C-Means [8] –as well as other clustering algorithms– are alternative techniques that are aimed at weakening FCM assumptions to provide more comprehensible clustering schemes.

$$J = \sum_{i=1}^c \sum_{j=1}^N u_{ij}^m \|\mathbf{x}_j - \mathbf{v}_i\|^2 \quad (1)$$

$$\forall j : \sum_{i=1}^c u_{ij} = 1 \quad (2)$$

$$\forall i : 0 < \sum_{j=1}^N u_{ij} < N \quad (3)$$

### How interpretability can be achieved and assessed?

To achieve interpretability, designers must develop a conceptual framework and an implementation method which satisfy the requirement of human comprehensibility with the goal of producing symbolic descriptions expressed in high-level, human oriented terms and forms. This rises some cognitive and epistemic questions: How does an user understand? Which cognitive processes and structures are involved in the process of understanding? How this process can be facilitated?

As a general guideline that follows the Comprehensibility Postulate, description components must be apprehended as single chunks of information (i.e. information granules), so as to allow the creation of mental models of the information they convey. In this sense, Fuzzy Logic appears to be as a step forward the development of interpretable systems because it allows the representation of knowledge without crisp boundaries, similarly to concepts conceived by human beings. But Fuzzy Logic is also a sophisticated mathematical theory that can be used without taking into account the interpretability of the involved objects. As a consequence, when interpretable fuzzy models have to be designed, Fuzzy Logic must be enriched with constraints that somehow capture the blurry notion of interpretability [17].

In this context, psychological and cognitive studies are fundamental for developing such inter-

pretability constraints. The work of Miller [12] concerning the optimal number of chunks of information simultaneously held in human short-term memory is emblematic to show how the analysis of interpretability could benefit from interdisciplinary studies. Other works, such as those trying to extract the meaning of linguistic terms from human subjects [14] are also fundamental for interpretability analysis, and there is still room for scientific investigation.

To achieve interpretability, the architecture of the fuzzy model takes a fundamental importance. Roughly speaking, two classes of architectures can be distinguished: single and multiple representation architectures [18]. Single representation architectures are based on a unique representation of the knowledge base, usually in a rule-based form, which is used both for determining and explaining the model behavior. Models of this type are easy to build but might offer a “flat” explanation of the knowledge base. Furthermore, in order to preserve interpretability, accuracy of the model can be seriously hampered.

A different approach provides for a multiple representation of knowledge, where one representation (not necessary interpretable) is used to generate an accurate behavior of the model, while the other is used to explain it in an interpretable form. Such dual representation has evidence in some cognitive studies, which show that different areas of the human brain are devoted to perception, action performing and natural language communication. Multiple representation architecture are more common in classical AI, but there are also systems developed within the Computational Intelligence paradigm that are worthwhile to mention, like that described in [15]. Another approach, presented in [9], provides for a multiple representation of fuzzy information granules, which are basic building blocks for designing accurate yet interpretable fuzzy models. Multiple representation architecture appears as a promising approach to reach a proper balance between accuracy and interpretability. For this reason, deeper investigation on it would be highly beneficial for interpretable fuzzy modeling.

## 4 Final remarks

Interpretability of fuzzy models is a requirement that calls for deeper insights than those usually appearing in literature. This paper rises only just few fundamental issues, which promote more profound interdisciplinary investigations. Furthermore, such fundamental issues may arise secondary questions that have not been tackled here but are yet important for fuzzy model design. Secondary issues include the prevention and remedy of incomprehensibility of acquired knowledge, the granularity of interpretable descriptions, etc.

On a greater extent, interpretability is only a facet of a more general requirement, invoked by Michie [10]: system humanization, which calls for a methodology for humanising the man-machine channel and is characterized by several facets – other than interpretability– such as mutual intelligibility, applicability, acceptability, interestingness, etc.. We are only at the beginning of this intriguing path.

## Acknowledgments

The authors wish to thank prof. Donato Malerba (University of Bari, Italy) and prof. Witold Pedrycz (University of Alberta, Canada) for their fruitful discussions on interpretability, which inspired the development of this paper.

## References

- [1] J. Casillas, O. Cordón, F. Herrera, L. Magdalena (eds.), *Interpretability Issues in Fuzzy Modeling*, Springer, 2003
- [2] A. Giboin, *ML Comprehensibility and KBS explanation: stating the problem collaboratively*, Proc. of IJCAI'95 Workshop on Machine Learning and Comprehensibility, 1995, pp. 1–11
- [3] S. Guillaume, *Designing fuzzy inference systems from data: An interpretability-oriented review*. IEEE Transactions on Fuzzy Systems, vol. 9, no. 3, 2001, pp.426–443.
- [4] P. Hájek, *Metamathematics of Fuzzy Logic*, Springer, 1998

- [5] Y. Jin, B. Sendhoff, Extracting interpretable fuzzy rules from RBF networks. *Neural Processing Letters*, vol. 17, 2003, pp. 149–164.
- [6] J. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum, 1981
- [7] Y. Kodratoff, Comprehensibility at the junction of Computer Science, Industry, and Cognitive Science, *Proc. of IJCAI'95 Workshop on Machine Learning and Comprehensibility*, 1995, p. 13
- [8] R. Krishnapuram, J. Keller, A possibilistic approach to clustering, *IEEE Transactions on Fuzzy Systems*, vol. 1, no. 2, 1993, pp.98–110
- [9] C. Mencar, A. Bargiela, G. Castellano, A.M. Fanelli, Interpretable Information Granules with Minkowski FCM, in *Proc. of the NAFIPS 2004*, pp. 456–461
- [10] D. Michie, *Machine Intelligence and Related Topics*. G&B Science Publishers, 1982
- [11] R.S. Michalski, A theory and methodology of inductive learning, *Artificial Intelligence*, vol. 20, 1983, pp. 111–161
- [12] G.A. Miller, The magical number seven, plus or minus two: some limits on our capacity for processing information, *Psychological Reviews*, vol. 63, 1956, pp. 81–97
- [13] S. Roychowdhury, An inquiry into the theory of defuzzification, in *Granular Computing: an introduction*, W. Pedrycz, Ed., Physica-Verlag, 2001, pp. 143–165
- [14] M.J. Smithson, Words about Uncertainty: Analogies and Contexts, in *Computing with Words in Information/Intelligent Systems 1 – foundations*, L.A. Zadeh, J. Kacprzyk, Eds., Physica-Verlag, 2003, pp. 119–135
- [15] R. Sun, E. Merrill, T. Peterson, From implicit skills to explicit knowledge: a bottom-up model of skill learning, *Cognitive Science*, vol. 25, no. 2, 2001, pp. 203–244
- [16] H. Toth, Fuzziness: From epistemic considerations to terminological clarification, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 5, 1997, pp. 481–503
- [17] J. Valente de Oliveira, Semantic constraints for membership function optimization. *IEEE Transactions on Systems, Man and Cybernetics*, part A, vol. 29, no. 1, 1999, pp. 128–138.
- [18] T. van de Merckt, C. Decaestecker, Multiple-Knowledge Representations in Concept Learning. *Lecture Notes in Artificial Intelligence*, vol. 914, 1995
- [19] M. van Someron, A perspective on Machine Learning and Comprehensibility from Knowledge Acquisition, *Proc. of IJCAI'95 Workshop on Machine Learning and Comprehensibility*, 1995, pp. 21–25
- [20] L.A. Zadeh, Soft computing and fuzzy logic, *IEEE Software*, vol. 11, no. 6, 1994, pp. 48–56