16th World Congress of the International Fuzzy Systems Association (IFSA)
9th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT)

Insights into Interpretability of Neuro-Fuzzy Systems

Marco Pota, Massimo Esposito

Institute for High Performance Computing and Networking (ICAR-CNR), Napoli, Italy

Abstract

Neuro-fuzzy networks revealed their proficiency in learning from data, while offering a transparent and somehow interpretable rule-based model. Recent research focused either on the interpretability of the chosen model or on the system performance. Regarding the interpretability, here an index to control the trade-off between complexity and performance, some insights into fuzzy partitions properties, an ideal fuzzy sets shape, and an evaluation of rules are proposed. All the evaluations are made taking into account the required output and performance. A discussion on results of a system built using the Wisconsin Breast Cancer Dataset is performed as a proof of concept.

Keywords: Neuro-fuzzy systems, Semantic Interpretability, Complexity, Fuzzy sets shape, Rule weights.

1. Introduction

In the field of data-driven knowledge extraction, neuro-fuzzy networks are gaining increasing importance. On the one hand, fuzzy rule-based systems offer transparent models, which can be more or less simple to be interpreted by a human being. On the other hand, the neural network representation of the system allows performing a well-established learning process based on gradient descent technique.

Recent research focused either on the understandability of the chosen model, usually [1,2] divided into complexity and semantic interpretability, or on the system performance. In this work, some proposals for improving the state-of-the-art of neuro-fuzzy systems are presented, which take into account both perspectives. In more detail, the fuzzy partitioning and rule base construction are considered, and in particular the choices made by the developer for the system design are evaluated, independently from the optimization procedures employed. Some of these choices are discussed, apart from interpretability and performance issues, also with respect to the possibility of employing particular optimization procedures like neural networks.

Firstly, the fuzzy partitions are considered. With respect to its complexity, an index for the trade-off between system performance and partition cardinality is proposed. Regarding the semantic interpretability, some controversial properties regarding each fuzzy set and the entire fuzzy partition [1,2] were discussed. Moreover, the fuzzy sets shape is particularly considered. Most of the previous works used a fixed fuzzy sets shape [3,4], while very a few made a comparison among them [5]. Triangular and trapezoidal membership functions (MFs) have been considered for their simplicity [4], but Gaussian MFs are often used for their differentiability and the little number of parameters, in order to use gradient descent optimization with neural networks. In this work, an alternative fuzzy sets shape, already proposed by authors [6], which can overcome all the problems associated with the most used ones, will be evaluated by considering interpretability issues at different levels, differentiability, and the granularity of information that can be encoded with respect to the number of associated parameters.

Secondly, different representation schemas of fuzzy rules are considered. In the ambit of fuzzy systems, different types of inferences were performed. . In particular, interpretability evaluations were made, both at the single rule and at the rule base levels, by considering either Mamdani [7] or Takagi-Sugeno (TS) [1-3] fuzzy systems. Clearly, the fuzzy system to choose should depend also on the amount and granularity of information to be modeled, e.g. a classification model is expected to be simpler to understand with respect to a regression model. However, interpretability issues have rarely [3] been correlated to the type of the required system output until now. Moreover, weights assigned to each rule [5,7] and to rule antecedents [5], and different types of rule consequents [5] are rarely considered from the interpretability viewpoint. In this work, a semantic evaluation of fuzzy rules will be presented. In particular, different levels of information will be ordered, to correlate the level of interpretability of the model and the level of granularity of the required output information. The associated considerations will be made about interpretability of the single rule when different types of consequents and/or antecedent weights are used, and of the rule base when rule weights are used. Thus, the proper rules design depending on the system output will be suggested.

The proposals will be discussed using a well-known dataset, i.e. the Wisconsin Breast Cancer Dataset, as a proof of concept, and the choices to design a classifier are discussed taking into account interpretability and performance at the same time.

The remaining part of the paper is made as follows. In Section 2, a general fuzzy system is described, while in Section 3 the state of the art about system interpretability is resumed. In Section 4, proposals regarding fuzzy partitions are presented, while in Section 5 the sight is moved to the rules. Experimental results are presented in Section 6, and Section 7 concludes the work.

2. A fuzzy inference system

A fuzzy system is basically made of the fuzzy partitions of the variables of interest (once these have been selected), and of a rule base.

Each fuzzy partition is made of a collection of fuzzy sets, representing the terms of the associated linguistic variable. Suppose that n variables are considered. Then, suppose that the range of the j-th variable is partitioned into M_j fuzzy sets $\hat{A}_{m\_j}^{(j)}$, with $m_j=1,...,M_j$. Each of the N data samples $\mathbf{x}_i=\{x_i^{(1)},...,x_i^{(n)}\}$, $i=1,...,N$, belongs to the fuzzy set $\hat{A}_{m\_j}^{(j)}$ with the membership grade $\mu_{m\_j}^{(j)}(x_i^{(j)})$.

A rule base can be written as a set of R rules r_ρ:

$$if\ x^{(1)}\ is\ \hat{A}_{(\rho)}^{(1)}\ and\ ...\ and\ x^{(n)}\ is\ \hat{A}_{(\rho)}^{(n)}\ then\ \hat{C}_{(\rho)}\ , \quad (1)$$

with $\rho=1,...,R$. The antecedents correspond to the fuzzy sets of the partitions $\hat{A}_{(\rho)}^{(j)} \in \{\hat{A}_1^{(j)},...,\hat{A}_{M\_j}^{(j)}\}$, and the consequents $\hat{C}_{(\rho)}$ are fuzzy sets [8,9]. In case of a classification, these fuzzy sets are defined on the set of nominal values which the output variable can assume; in case of a regression [9], consequents are defined in the continuous domain of the output variable Y.

The membership grades of antecedents are usually combined by T-norms to get the firing strength of each rule. Each rule represents an implication between antecedents and consequents which can be modeled by a T-norm (or an S-norm), and different rules can be aggregated by S-norms (or T-norms). Final defuzzification can be performed, depending on the required output.

3. Interpretability issues

In order to briefly describe the criteria to evaluate the interpretability of fuzzy inference systems, a taxonomy proposed in [1] and refined in [2] is used here. In particular, two types of difficulties associated to the process of understanding the system can be individuated, which are *semantic interpretability* [1,2] and *complexity* [1] (named "readability" in [2]). On the other hand, four levels of the system whose interpretation is needed can be individuated [2], which are the *single fuzzy sets*, the *fuzzy partitions*, the *single rules*, and the whole *rule base*.

The complexity of the system should be minimized, and it increases with the *number of features* used by the model, with the *number of fuzzy sets* for each partition (cardinality), the *number of antecedents* in each rule and the total *number of rules* [2].

Semantic interpretability can be taken into account by imposing constraints or optimizing some measures, at all the abstraction levels. Authors agree on a long list of evaluation criteria, reported as follows. Regarding each fuzzy set, *one-dimensionality*, *normality*, *continuity*, *convexity*, and *unimodality* are required constraints for semantic interpretability as terms of a linguistic variable [1-3]. *Differentiability* is required if a gradient descent method is used for optimization [4,10]. All these characteristics will be considered in this work. On the contrary, the opportunity of using the *symmetry* condition for each fuzzy set [11] will be discussed. Different fuzzy sets of a fuzzy partition should satisfy *distinguishability*, *coverage* and *proper ordering* [1-3]. In particular, for distinguishability and coverage some measures to minimize [12] will be considered:

$$J_D = \sum_x \left(\left(SUM_p(x) - 1 \right)^2 \cdot H\left(SUM_p(x) - 1 \right) \right) \quad (2)$$

$$J_C = \sum_x \left(\left(SUM_p(x) - \varepsilon \right)^2 \cdot H\left(\varepsilon - SUM_p(x) \right) \right) \quad (3)$$

where

$$SUM_p(x) = \left(\sum_{m=1}^{M} \mu_m^p(x) \right)^{1/p} , \quad (4)$$

ε is a threshold fixed for coverage, $H(\cdot)$ is the Heaviside step function, M is the cardinality, and p is a positive integer (in [12], p=1). Proper ordering will be imposed by considering the strong condition of relation preservation defined in [3], since the corresponding weak condition [3] seems not enough. Moreover, the presence of *leftmost and rightmost* fuzzy sets [3] should be ensured. The usefulness of complementarity among fuzzy sets:

$$\forall x, S-norm\left[\mu_1(x),...,\mu_C(x) \right] = 1 \quad (5)$$

or of their *orthogonality* (when the S-norm is the sum) is controversial [3,13]. Some works also considered the presence of *prototypes* on special elements (among them, the zero element if any) [3], and *uniform granulation* [3]. The last three conditions will be discussed.

The semantic interpretability of each rule depends on the *type of antecedents*, *type of consequent* and on the presence of *antecedent weights*. In the following sections, some observations will be made in particular, regarding the comparison of consequents made of a single class, different classes with respective probabilities and 0-order TS type. To compare the first two consequent types, a measure of rule clearness [7] will also be used:

$$J_{RC} = \prod_\rho \max_k \left[p(c_k \mid \rho) \right] . \quad (6)$$

The opportunity of using antecedent weights will be evaluated as well, considering rule interpretability and the output performance. If the whole rule base is considered, syntactic interpretability is evaluated by checking *Modus Ponens*, *consistency*, *locality* and *completeness* of the rule base [1-3]. Moreover, other qualitative settings can be accounted, as different rule interpretations (associated with different models for implication and aggregation operators) [3], different types of input variables (selected from those given in the dataset or extracted from them) [3], in case of multiple output the difference between multi-input-single-output (MISO) and muti-input-multi-output (MIMO) systems [3], and so on. In the following, all these characteristics will be considered as established, without checking them. In particular, MISO systems with conjunctive implication are considered. Instead, the average *number of firing rules* [1-3] will be considered to assess the understandability of the inference process, and the opportunity of using *rule weights* [4,5,7,14] will be evaluated, since they reduce the system interpretability [7,14] but can greatly improve system performances [4,5].

4. Proposals on fuzzy partitions

Fuzzy partitioning aims at modeling terms of linguistic variables associated to continuous variables with interpretable fuzzy sets. Firstly, let us distinguish between *magnitude-related* linguistic terms [13] like "high temperature", and *value-related* linguistic terms like

"around 37 °C". In the following sections, an index to control the trade-off between performance and complexity while imposing the partition cardinality is proposed. Moreover, after some observations regarding interpretability, differentiability, performance and number of representative parameters, a fuzzy sets shape is described which satisfies many usual requirements.

4.1. Performance/complexity trade-off

In the fuzzy partitioning, each Universe Of Discourse $UOD^{(j)}$ of the continuous selected variable $X^{(j)}$, i.e. the interval $[\min_i[x_i^{(j)}],\max_i[x_i^{(j)}]]$, is partitioned into a certain number M_j of fuzzy sets, i.e. the cardinality.

It is straightforward that if M_j is increased, the highest granularity of the input allows to better approximate the output, thus improving the system performances. On the other hand, the system is undoubtedly complicated by the presence of an increasing number of linguistic terms for each variable, which increase complexity. A trade-off between these opposite trends should be addressed. In past works [6], authors used to accept the addition of a further fuzzy set of the partition if it is associated with an error decrease greater than 10%.

Here, an index is proposed which increases when the accuracy increases and when complexity decreases, therefore it should be maximized:

$$Q_M = \frac{1-E}{M + 1/q_M - 1}, \tag{7}$$

where E is the error, M is the cardinality of each partition M_j or the sum of cardinalities, and q_M is a constant, $0<q_M\leq 1$, which controls the desired trade-off: if it is near 0, the performance is considered mostly, while if it increases to 1, greatest consideration is given to the complexity.

4.2. Fuzzy partitions interpretability properties

In this section, some insights into interpretability properties of fuzzy partitions are given. In particular, some evaluations are proposed about the symmetry properties, the presence of prototypes, the orthogonality, and the uniform granulation, which are often considered for fuzzy partitions design, but whose usefulness is controversial.

The usual symmetry condition (S1) [11] states that the increasing and the decreasing sections of *the same* MF should be symmetrical with respect to the line $x=x_c$, where x_c is the center of the core of the fuzzy set. However, this condition seems not always necessary. As a reliable alternative, here it is proposed to impose the symmetry condition (S2) on the increasing and decreasing sections of *two adjacent* MFs, which should be symmetrical with respect to their intersection point. This way, each linguistic term is not symmetric, but the uncertainty between two linguistic terms is. The condition S1 is at the fuzzy set level, and in the opinion of the authors, it should be imposed for value-related linguistic terms only. On the contrary, the alternative condition S2 is at the fuzzy partition level and, in association with a proper ordering condition, should be imposed on all the couples of adjacent linguistic terms.

Some prototypes can be expressly required [3] in correspondence of special elements x_{SE} like 0. In this case, they should be associated with value-related linguistic terms, which are symmetric (S1) and have x_{SE} as the core. A linguistic term "*normal X*" can be required, and in this case it should be magnitude-related and properly associated with corresponding consequents of the rule base. In the authors' opinion, if not expressly required, the use of prototypes is not necessary. Moreover, all the other linguistic terms should be magnitude-related, in order to improve both interpretability and complexity. Finally, the prototypes should be counted as part of the fuzzy partition satisfy all the interpretability properties, comprising proper ordering and condition S2.

Even if orthogonality requirement is controversial [3,13], in authors' opinion it is very useful, since ensures distinguishability and coverage at the same time. If, e.g., p=1 is fixed in Eqs. (2) and (3), then for orthogonal partitions $J_C=J_D=0$.

The condition of uniform granulation [3], in authors' opinion, should not be imposed to any fuzzy partition. Indeed, this can be roughly traduced in conditions like "tall people, medium height people and short people are in equal number", which can be untrue, do not improve semantic interpretability, and is based on unsupervised information which do not increase system performance.

4.3. Fuzzy sets shape

Only a few works have discussed about the opportunity of using a certain shape for MFs [5,15,16]. However, performance seems to be application-dependent, and interpretability of fuzzy sets with different shapes has rarely been compared. In the following, some considerations concerning the interpretability and performance of fuzzy sets of different shapes are given. After that, our proposal and its peculiarities are described.

4.3.1 Fuzzy numbers vs. fuzzy intervals

Here fuzzy numbers and fuzzy intervals are compared. Their core is respectively a point or an interval.

Fuzzy numbers can be viewed as particular cases of fuzzy intervals. Even if they are used very often, their advantage with respect to fuzzy intervals can be ascribed only to a lower number of parameters to represent each of them, which reflects the difference between crisp numbers and intervals.

Most effective advantages of fuzzy logic [17] substituting crisp intervals stands in the fact that *in the zone around the separation* of adjacent crisp intervals, fuzzy intervals can model the uncertainty about the membership of values of this zone to an interval rather than the other one. If fuzzy numbers are used, then *the whole UOD* is modeled as uncertain, which could be untrue.

As consequences of the previous two observations, on the one hand, a whole partition made of the same number M_j of fuzzy sets is defined by less parameters in case of fuzzy numbers, , with respect to fuzzy intervals. On the other hand, more fuzzy numbers than fuzzy intervals could be needed to model the same information. In order to better show these opposite trends, an example is given, which is referred to linear MFs. In this case, fuzzy numbers have triangular MFs, while fuzzy

intervals have trapezoidal MFs. Suppose that the output variable Y to model is a triangular function of the variable X as reported in Fig. 1(a). If orthogonal partitions are optimized, then the partition made of triangular MFs (Fig. 1(b)) requires 1 parameter, while the same behavior is modeled by trapezoidal MFs (Fig. 1(c)) using 4 parameters. Therefore, in this case, the same information is modeled by the same number of fuzzy sets, but trapezoidal MFs require more parameters. On the other hand, suppose, as more generally happens, that the output variable Y is a trapezoidal function of the variable X as reported in Fig. 1(d). If orthogonal partitions are optimized, then the partition made of triangular MFs (Fig. 1(e)) requires 4 parameters and 6 fuzzy sets, while the same behavior is modeled by trapezoidal MFs (Fig. 1(f)) using always 4 parameters, but only 3 fuzzy sets. In this case, the use of fuzzy intervals reduces the number of fuzzy sets, using the same total number of parameters. These limiting cases can be generalized as follows: *i)* in order to model U uncertainty zones, U+1 fuzzy set are required, which are modeled by U–1 parameters for orthogonal triangular MFs or 2U for trapezoidal ones; *ii)* in order to model P plateaus between uncertainty zones, P further triangular fuzzy sets are required, with P more parameters, while no further fuzzy set nor parameter are required using trapezoidal fuzzy sets. Therefore, in Fig. 1(a), a situation of U=2 and P=0, thus $U>P-1$, is favorable to triangular MFs, while in Fig. 1(d), a situation of U=2 and P=3, thus $U=P-1$, is favorable to trapezoids.

Definitively, as a result of these observations, in order to optimally model information, trapezoidal MFs should be used instead of triangular ones because they ensure to model the same amount of information, with a smaller number of fuzzy sets (U+1, with respect to $U+P$+1 of triangular MFs), which are easier to be understood since minimize complexity. On the other hand, in order to optimally employ a certain number of fuzzy sets, the use of trapezoidal fuzzy sets requires more parameters, but ensures to better model information about the separation between the uncertainty zones and the plateaus, thus achieving better performance. Therefore, the two approaches can work equivalently if the desired system output is a simple class, while in case the desired system output is a number, or a set of possible classes with respective confidences, trapezoidal MFs are preferable both from complexity and performance points of view, with the only drawback of more parameters to optimize (2U, with respect to $U+P-1$ of triangular MFs) in particular cases when $U>P-1$.

The observations made above about the difference between triangular and trapezoidal MFs are still valid in case of nonlinear MFs. Therefore, fuzzy intervals should be preferred to fuzzy numbers, in order to better model the system output and keep low the fuzzy partitions complexity.

Regarding semantic interpretability, fuzzy intervals are generally more appropriate to model magnitude-related linguistic terms, while fuzzy numbers can only model value-related linguistic terms, therefore they should be used only to fuzzify special elements.

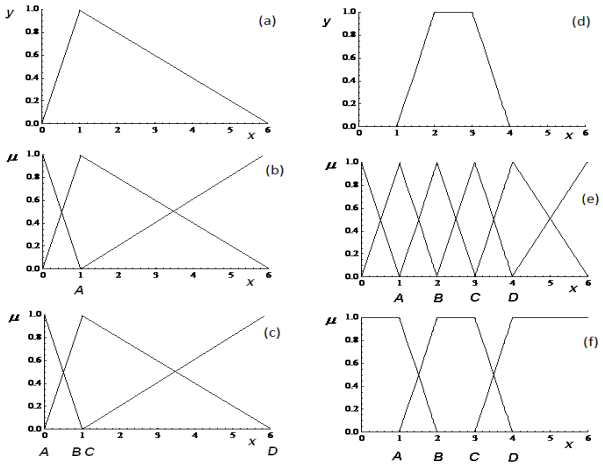

Figure 1: (a) Triangular and (d) trapezoidal output functions, optimally modeled by (b and e) triangular and (c and f) trapezoidal partitions. A, B, C, and D are the parameters.

Moreover, if orthogonal partitions made of fuzzy intervals instead of fuzzy numbers are used, then the number of rules fired at the same time is minimized, therefore the understandability of the inference process is improved.

Nonlinear MFs are required if differentiable functions should be used, in order to employ gradient descent optimization procedures like neural networks.

The widespread use of Gaussian MFs [10] is based on different reasons, like their differentiability, the little number of required parameters (the same of symmetric triangular MFs), and the easily interpretable shape of each fuzzy set. However, the Gaussian MFs are fuzzy numbers, not intervals, which are generally preferable. Moreover, Gaussians cannot be orthogonal, thus distinguishability and coverage have to be checked. Finally, proper ordering condition is violated [3]: e.g. if two Gaussian MFs are considered, with core values x_A and x_B with $x_A<x_B$, and standard deviations σ_A and σ_B with $\sigma_A>\sigma_B$, then there are values $x>x_B$ where $\mu_A>\mu_B$, which violates the proper ordering. A weak ordering condition has been stated [3] to "save" this type of MFs. However, if only the weak ordering condition is imposed, the whole fuzzy partitions often results not well interpretable, as in [10].

4.3.2 An ideal fuzzy sets shape: Sigmoidal MFs

As a consequence of the previous observations, the ideal shape of MFs is such that they generally represent fuzzy intervals. They should satisfy all the properties to have interpretable fuzzy sets and partitions, plus orthogonality and symmetry condition S2. Moreover, each MF should be able to satisfy symmetry condition S1 if required. Finally, differentiable functions enabling the use of gradient descent optimization methods are required for neuro-fuzzy systems.

The MFs suggested here are in line with all the previous observations. They are constructed by using sigmoid functions S:

$$S(x;a,b)=\frac{1}{1+e^{\left(t\cdot\frac{b+a}{b-a}-2\cdot t\cdot\frac{x}{b-a}\right)}}\quad . \tag{8}$$

Among the MFs $\mu_m(x;a,b,c,d)$ of the partition, the leftmost MF μ_1, internal MFs μ_m, and the rightmost MF μ_M are respectively:

$$\begin{cases} \mu_1(x;a_1,b_1) = 1 - S(x;a_1,b_1) \\ \mu_m(x;a_{m-1},b_{m-1},a_m,b_m) = S(x;a_{m-1},b_{m-1}) - S(x;a_m,b_m) \\ \mu_M(x;a_{M-1},b_{M-1}) = S(x;a_{M-1},b_{M-1}) \end{cases}. \quad (9)$$

with the condition:

$$a_1 < b_1 \le ... \le a_m < b_m \le ... \le a_{M-1} < b_{M-1} \quad (10)$$

and with $t=\mathrm{Log}(1/\varepsilon-1)$, where $\varepsilon<<1$ is a fixed positive constant.

Such MFs are approximately normal, since

$$\forall m \exists \{x_{0m}, x_{1m}\} : \mu_m(x_{0m}) < \varepsilon, \mu_m(x_{1m}) > 1-\varepsilon \ . \quad (11)$$

Moreover, one-dimensionality, continuity and unimodality are straightforward, while convexity is ensured by (10). Therefore, each fuzzy set is perfectly interpretable as a magnitude-related linguistic term. The symmetry condition S1 can be easily imposed by $b_{m-1}-a_{m-1}=b_m-a_m$, if a value-related linguistic term is required on a special element $x_{SE}=b_{m-1}=a_m$.

At the fuzzy partition level, it is easy to check that the proposed partition is made of orthogonal fuzzy sets, therefore maximal distinguishability and coverage are ensured. Moreover, there are the leftmost and rightmost fuzzy sets, the symmetry condition S2 is satisfied, and (10) ensures proper ordering. Therefore, a perfectly interpretable fuzzy partition is obtained.

The core is generally an interval, and the number of parameters is the same of trapezoidal MFs. Therefore, as explained before, the best compromise between capacity of encoding information and number of parameters is achieved. If the required system result is only categorical, the number of parameters can be reduced, letting cores collapse in points by imposing

$$\begin{cases} a_1 = \min_i \left[x_i^{(j)} \right] \\ b_{m-1} = a_m, m = 2, ..., M-1 \\ b_{M-1} = \max_i \left[x_i^{(j)} \right] \end{cases}. \quad (12)$$

Finally, the differentiability of MFs enables to use them for gradient descent optimization methods.

In Fig. 2, different types of partitions are shown, all with a number of terms $M=3$. Triangular (a) and Gaussian (b) fuzzy numbers fuzzify the whole UOD, while trapezoidal (c) and sigmoidal (d) MFs fuzzify only some uncertainty zones. MFs (c) and (d) encode more information, with respect to (a) and (b). MFs (d) are similar to (c), but they are differentiable.

In the following application, sigmoidal MFs will be extracted from the dataset, their number will be optimized by (7), and their interpretability will be shown. A comparison will be made with trapezoidal and Gaussian MFs [10].

5. Proposals on the rule base

In this section, some observations are given regarding the rule design, considering in particular the impact of different types of rule antecedents and consequents, and of antecedent and rule weights, on both interpretability and performance.

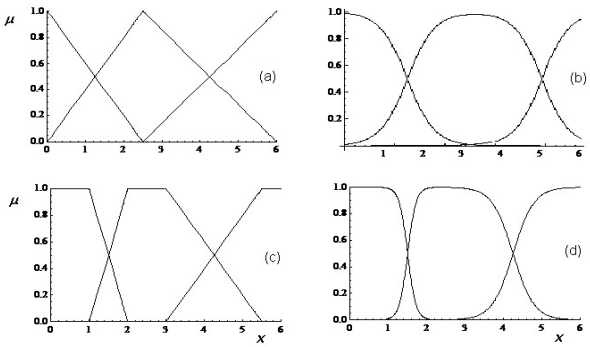

Figure 2: (a) Triangular, (b) Gaussian, (c) trapezoidal, and (d) sigmoidal fuzzy partitions.

Some works tried to use optimized antecedent weights and rule weights, in order to maximize performance [4,5]. Some works also evaluated the improvement of the classification performance gained if single class consequents are substituted by fuzzy consequents [5]. Here, these choices are correlated to the output required to the system and to interpretability issues.

5.1. Information levels

A kind of order of *information levels* is proposed, as shown in Fig. 3. The same order can be associated to system inputs, output, rule antecedents, and rule consequents. Clearly, the difficulty to understand these entities as well as the quantity of encoded information increase going from level *I) simple classes* (Fig. 3(a)), to level *II) classes with associated membership grade* (Fig. 3(b)), to level *III) ordered classes with associated membership grade* (Fig. 3(c)), to level *IV) fuzzy sets interpretable as linguistic terms* (Fig. 3(d)), to level *V) numbers* (Fig. 3(e)), to level *VI) functions* (Fig. 3(f)).

Input data x can be of level I (e.g. *male*/*female*), level II (e.g. known probabilities of categories), level III (e.g. a known midway between ordered classes), level IV (e.g. "*large*"/"*small*"), and very often level V. The use of fuzzy systems is particularly devoted to decrease the difficulty level of input variables, transforming the knowledge about their continuous values (level V) to knowledge encoded in linguistic terms (level IV).

The output of the system can be of level I, if a simple class is required to be assigned to the sample, or level II or III, if information about class uncertainty is required, level IV if the result is a fuzzy set comprising uncertainty information about the numerical output, level V if the numerical output should be given without uncertainty information. Here, level IV is higher than level V, since the result of a fuzzy inference is often a fuzzy set which is not interpretable.

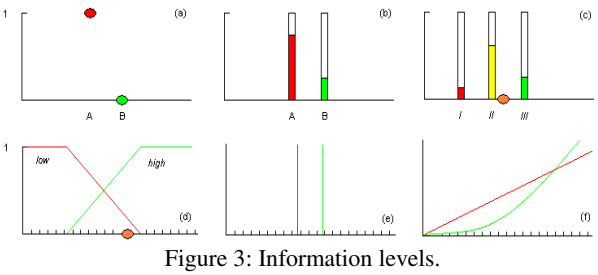

Figure 3: Information levels.

5.2. Rules semantic evaluation

Each rule of the MISO rule base is made of a number of antecedents and one consequent.

The knowledge about data is encoded in a different form by rule antecedents, which transform input values x into other values $\mu(x)$. Antecedents can be of level I, e.g., "*if x is A*", (x categorical, $\mu(x)\in\{0,1\}$), level II or III ($0\le x\le 1$, $\mu(x)$ the same), level IV, e.g., "*if x is low*", (x numerical, $0\le\mu(x)\le 1$), level V (x numerical, $\mu(x)$ the same) or level VI ($\mu(x)$ is a function of x). Level I antecedents have high interpretability, while trivially keep the minimum part of input information. Levels V and VI can be used for not interpretable regression systems. Level IV antecedents are a reliable solution to encode high level input information in an interpretable way, thus they represent the foundation of fuzzy systems.

Antecedent weights can be useful to improve performance even in case of classification problems [4]. However, a fuzzy system with antecedent weights has very low interpretability, comparable to that of a regression multi-linear model, which can be positioned between levels V and VI. Therefore, fuzzy models using antecedent weights should be used only when the output of the system is of level V (regression problems), and interpretability is not a main objective.

Rule consequents can be of level I (e.g. "*then Y is A*"), level II or III (e.g. "*then Y is {0.8 A, 0.2 B}*"), level IV (e.g. "*then Y is high*"), level V (e.g. "*then Y is 3*"), level VI (e.g. "*then Y is* x^2"). Levels I to IV individuate Mamdani-type systems, while TS systems are at level V (0-order) and VI.

The right choice of rule consequents type depends on the required system output.

In case of regression (output level ≥IV), each consequent of (1) can be a proper fuzzy set [8] (level IV), a number (or a fuzzy singleton, i.e. a particular fuzzy set whose support is a single value with a membership grade of 1, level V) or a function [9] (level VI). From the interpretability viewpoint, the use of level IV is recommended [3].

In case of a classifier (output level ≤III), each consequent can be a singleton whose support is a class [8] (level I), or a fuzzy set defined on different classes [8] (level II/III). However, some works also use 0-order TS systems [10] (level V). It is straightforward that low-level consequents are more interpretable, however, the right choice is different if the output level is I or II/III.

Suppose that there are K possible classes $c_1,...,c_K$. In a simple model, each consequent is a singleton whose support is a class (level I):

$$\hat{C}_{(\rho)} = \frac{c_{(\rho)}}{1} \ , \tag{13}$$

where $c_{(\rho)}\in\{c_1,...,c_K\}$. Therefore, the outfit is of the type:

$$then\ Y\ is\ c_{(\rho)} \ . \tag{14}$$

This is the most interpretable type of rule consequent. However, if output level is >I, then a system with these rule consequents is not able to model the uncertainty of the output [5], therefore it should be used only for level I output.

If the output of the system is required to be a set of classes associated with respective confidences (level II/III), different non-zero membership grades in correspondence of different classes should be assigned to the rule consequent, which can be regarded in this case as a proper fuzzy set (level II/III consequents):

$$\hat{C}_{(\rho)} = \frac{c_1}{p_{\rho-1}} + ... + \frac{c_K}{p_{\rho-K}} \ . \tag{15}$$

Therefore the outfit is of the type:

$$then\ Y\ is \begin{cases} \left(p_{\rho-1}\right)c_1 \\ ... \\ \left(p_{\rho-K}\right)c_K \end{cases} . \tag{16}$$

The interpretability of rules with this type of consequents is high. In fact, if the constraint is satisfied that

$$p_{\rho-1} + ... + p_{\rho-K} = 1 \ , \tag{17}$$

then the rule can be interpreted as follows. In case of output level II: "in the space restriction modeled by antecedents of the rule r_ρ, different classes have respective probabilities $p_{\rho-1}$, ..., $p_{\rho-K}$". In case of output level III: "in the space restriction of the rule r_ρ, the output is a midpoint between different classes represented by the fractions $p_{\rho-1}$, ..., $p_{\rho-K}$". If the rule base is consistent, this also ensures that the Modus Ponens is satisfied. The loss of interpretability due to the use of consequents as in (16) instead of (14) can be computed by J_{RC} (6), taking into account that the best result is obtained for (14) where $J_{RC}=1$.

If (17) does not stand, then a multiple 0-order TS system is used (level V consequents) of the type:

$$then\ Y\ is \begin{cases} \left(w_{\rho-1}\right)c_1 \\ ... \\ \left(w_{\rho-K}\right)c_K \end{cases} . \tag{18}$$

This type of consequent can encode more information than (16), but the rule becomes very less interpretable. E.g., a rule with consequent "*then Y is A with degree 0.00 and B with degree 0.54*" [10] is difficult to interpret. In this case, J_{RC} (6) cannot be used. Therefore, in order to model an output of level II/III, it is preferable from the interpretability perspective to use level II/III rule consequents (16).

In order to keep the same amount of information of (18), and have interpretable consequents as (16), a simple method was already proposed [6] to transform rule consequents from (18) to (16), by substituting:

$$p_{\rho-k} = \frac{w_{\rho-k}}{\sum_{\kappa=1}^{K} w_{\rho-\kappa}} \ . \tag{19}$$

At the same time, weights should be associated to different rules to keep results of aggregation:

$$W'_\rho = \sum_{\kappa=1}^{K} w_{\rho-\kappa} \ . \tag{20}$$

From the performance perspective, this procedure should not change the result of the inference, if it is correctly modeled. Therefore, a system with level V con-

sequents is equivalent to a system with level II/III consequents and rule weights.

From the interpretability perspective, two opposite effects are obtained: on the one hand, each rule with probabilities (19) in the consequent (level II/III consequent) is much more interpretable than the same rule in the form (18) (level V consequent); on the other hand, the whole rule base is much less interpretable [7,14] if weights (20) are applied.

The idea of the authors on this issue is that if the sum of rule weights is set to 1 by substituting W'_ρ with W_ρ:

$$W_\rho = \frac{W'_\rho}{\sum_{\rho=1}^{R} W'_\rho}, \tag{21}$$

then the loss of interpretability associated with rule weights is compensated enough by the increased interpretability of each rule. Indeed, in this case, rule weights can be viewed as: "measures of the relative importance of different rules in the uncertainty zones", and can be modeled by weighted norms when rules are aggregated.

Summarizing, if the classifier output is required as a set of classes with respective probabilities (level II/III), then a system with level II/III consequents (16) should be modeled for interpretability. Rule weights (21) can be used if also the performance is a main objective.

It is straightforward that rule weights can enhance the system performance in the zones of uncertainty, also when the required output is of level I and level ≥IV.

An index computing the loss of interpretability due to rule weights has not been conceived yet. However, this loss is lower if a smaller number of rules are fired at the same time. In fact, the inference process should be interpretable, apart from its whole structure, also for each single prediction. This is more interpretable if a little number of rules are fired. Moreover, when only one rule is fired, weights are not considered, while if the number of rules increases, then the loss of interpretability due to the rule weights increases.

In the following application, an example of a rule base extracted from data for classification will be shown, and its interpretability and performances will be evaluated.

6. Application to real data

In this section, the Wisconsin Breast Cancer Dataset (WBCD) [18] is used as a proof of concept. Each one of the 699 samples is described by 9 variables, corresponding to the following samples features: Clump Thickness (CT), Uniformity of Cell Size (UCSi), Uniformity of Cell Shape (UCSh), Marginal Adhesion (MA), Single Epithelial Cell Size (SECS), Bare Nuclei (BN), Bland Chromatin (BC), Normal Nucleoli (NN), and Mitoses (Mi), and is classified into benign (c_B) or malignant (c_M).

A system able to predict these classes from input variables is requested to present an output of level I, or level II if a confidence measure is needed too.

A fuzzy inference system is used like (1), using the couple {UCSi,BN} as input variables, with fuzzy partitions like (9), rule consequents like (16), and rule weights like (21). The parameters of the fuzzy partitions and the rule weights and consequents are calculated as proposed in [6], by approximating likelihood functions with linear combinations of MFs. This method was named Likelihood-Fuzzy Analysis (LFA), and is also based on gradient descent optimization.

The obtained fuzzy partitions are shown in Fig. 4, while the rule base results as follows:

$$\begin{cases} (0.34)\ \textit{if}\ \text{UCSi}\ \textit{is}\ \text{low}\ \textit{and}\ \text{BN}\ \textit{is}\ \text{low}\ \textit{then}\ \ \text{benign} \\ (0.02)\ \textit{if}\ \text{UCSi}\ \textit{is}\ \text{low}\ \textit{and}\ \text{BN}\ \textit{is}\ \text{high}\ \textit{then}\ \ \text{benign} \\ (0.04)\ \textit{if}\ \text{UCSi}\ \textit{is}\ \text{high}\ \textit{and}\ \text{BN}\ \textit{is}\ \text{low}\ \textit{then} \begin{cases} (0.44)\,\text{benign} \\ (0.56)\,\text{malignant} \end{cases} \\ (0.60)\ \textit{if}\ \text{UCSi}\ \textit{is}\ \text{high}\ \textit{and}\ \text{BN}\ \textit{is}\ \text{high}\ \textit{then}\ \ \text{malignant} \end{cases} \tag{22}$$

Regarding the trade-off between fuzzy partitions complexity and performance, the number of 2 fuzzy sets for each partition was calculated in order to maximize Q_M (7), with q_M=0.5 and M equal to the sum of cardinalities, M=4. In this example, the same optimal number of fuzzy sets is obtained for a very large range of values of q_M; this means that the addition of a further fuzzy set to the partitions corresponds to a little accuracy improvement (around 0.1%, for M=6); on the other hand, if the value of q_M is set to very low values (below 0.001), then the little accuracy improvement is taken into account, and the optimal number of fuzzy sets corresponds to 3 for each partition.

Regarding semantic interpretability of fuzzy sets and partitions, all the discussed properties result satisfied by using sigmoidal MFs, while in general the partitions made of Gaussian MFs result much less interpretable. Indeed, sigmoidal MFs have an interval as a core, thus they are more interpretable as magnitude-related linguistic terms; moreover, distinguishability and coverage are the best, since J_C=J_D=0 in Eqs. (2) and (3); finally, since the symmetry condition S2 proposed here is satisfied instead of S1 [11], then the strong proper ordering condition [3] is satisfied instead of the corresponding weak condition [3] satisfied by Gaussian MFs.

Since the cores are made of intervals, the average number of rules fired at the same time is lower than the case of Gaussian MFs [10], therefore the inference process is more understandable.

With respect to trapezoidal and other not differentiable MFs, the chosen shape allows optimizing parameters by gradient descent methods, as required by neural networks or by LFA.

Each rule satisfies all the properties requested for semantic interpretability. No antecedent weights are used, since the output level is <V. However, the presence of fuzzy consequents (information level II) generates a loss of interpretability with respect to pure class consequents (level I), which can be evaluated by (6), resulting J_{RC}=0.56 instead of J_{RC}=1. Moreover, the presence of rule weights surely reduces the rule base interpretability.

On the other hand, the accuracy obtained by (22) is 96.1%, and similar results are gained if consequents are in the form (14), while accuracy decreases to 92.7% if rule weights are avoided. Therefore, the use of rule weights demonstrate a great power in improving per-

formance, for both level I and II output, while fuzzy consequents enable to refine the output of level II.

7. Conclusion

Fuzzy systems were considered, with particular regard to neuro-fuzzy systems. Some insights into their interpretability, taking into account the requested type of output and performance, were presented.

Regarding the fuzzy partition of the variables, an index was proposed to control the trade-off between the worsening of complexity due to an increasing number of fuzzy sets of the partitions, and the accuracy improvement gained at the same time. Moreover, evaluations about controversial interpretability properties was given, and a fuzzy sets shape is proposed, which results the best choice in terms of encoded information, number of parameters, and semantic interpretability at both the fuzzy set level and the partition level, and allows employing gradient descent optimization methods.

Regarding the rules design, a kind of order is proposed for different levels of information encoded by inputs, rule antecedents, rule consequents and output. On this basis, an evaluation is given about the opportunity of using antecedent weights, different types of rule antecedents and consequents, and rule weights.

The proposed system design was compared with the most used ones, and its strength was confirmed using the Wisconsin Breast Cancer Dataset.

References

[1] M.J.Gacto, R.Alcalá, F.Herrera, Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures, *Information Sciences*, 181: 4340–4360, 2011

[2] C. Mencar, Interpretability of fuzzy systems, *Fuzzy Logic and Applications Lecture Notes in Computer Science*, 8256: 22–35, 2013.

[3] C. Mencar, A. Fanelli, Interpretability constraints for fuzzy information granulation, *Information Sciences*, 178: 4585–4618, 2008.

[4] L. Rutkowski, Flexible neuro-fuzzy systems, *IEEE Transactions on neural networks*, 14: 554–574, 2003.

[5] M. Pota, M. Esposito, Degrees of Freedom and Advantages of Different Rule-Based Fuzzy Systems, in *Proc. of The 2014 International Conference on Pure Mathematics, Applied Mathematics and Computational Methods*, Santorini Island, Greece, July 17-19, 2014.

[6] M. Pota, M. Esposito, G. De Pietro, Best Fuzzy Partitions to Build Interpretable DSSs for Classification in Medicine, in *Hybrid Artificial Intelligent Systems, 8th International Conference, HAIS 2013, Salamanca, Spain, September 11-13, 2013. Proceedings*, series: *Lecture Notes in Computer Science*, 8073: 558–567, Springer Berlin Heidelberg Ed., 2013.

[7] R. Mikut, J. Jakel, L. Grall, Interpretability issues in data-based learning of fuzzy systems, *Fuzzy Sets and Systems*, 150: 179–197, 2005.

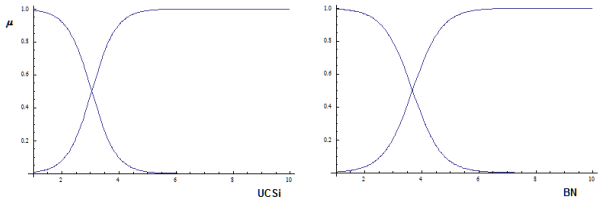

Figure 4: Fuzzy partitions extracted from WBCD dataset.

[8] E. H. Mamdani, S. Assilian, An experiment in linguistic synthesis with a fuzzy logic controller, *International Journal of Man-Machine Studies*, 7(1): 1–13, 1975.

[9] M. Sugeno, *Industrial applications of fuzzy control*, Elsevier Science Pub. Co., 1985.

[10] G. Castellano, A. M. Fanelli, C. Mencar, A neuro-fuzzy network to generate human-understandable knowledge from data, *Cognitive System Research*, 3: 125-144, Elsevier, 2002.

[11] P. Pulkkinen, H. Koivisto, A dynamically constrained multiobjective genetic fuzzy system for regression problems, *IEEE Transactions on Fuzzy Systems*,18: 61–177, 2010.

[12] P. Fazendeiro, J.V. de Oliveira, W. Pedrycz, A multiobjective design of a patient and anaesthetist-friendly neuromuscular blockade controller, *IEEE Transactions on Biomedical Engineering*, 54: 1667–1678, 2007.

[13] M. Pota, M. Esposito, G. De Pietro, Fuzzy Partitioning for Clinical DSSs Using Statistical Information Transformed into Possibility-Based Knowledge, *Knowledge-Based Systems*, 67: 1–15, 2014.

[14] A. Marquez, F. Márquez, A. Peregrin, A multiobjective evolutionary algorithm with an interpretability improvement mechanism for linguistic fuzzy systems with adaptive defuzzification, in: *IEEE World Congress on Computational Intelligence*, 2010.

[15] N. Gupta, S. K. Jain, Comparative analysis of fuzzy power system stabilizer using different membership functions, *International Journal of Computer and Electrical Engineering* 2(2): 262–267, 2010.

[16] J. G. Monicka, N. Sekhar, K. R. Kumar, Performance evaluation of membership functions on fuzzy logic controlled AC voltage controller for speed control of induction motor drive, *International Journal of Computer Applications*, 13(5): 8–12, 2011.

[17] J. M. Alonso, C. Castiello, M. Lucarelli, C. Mencar, Modelling interpretable fuzzy rule-based classifiers for Medical Decision Support, in *Medical Applications of Intelligent Data Analysis: Research advancements*, Magdalena, R., Soria, E., Guerrero, J., Gómez-Sanchis, J. and Serrano, A. J. Eds., IGI Global: 255-272, 2012.

[18] K. Bache, M. Lichman, UCI Machine Learning Repository, 2013. <http://archive.ics.uci.edu/ml>. Irvine, CA: University of California, School of Information and Computer Science.