Genetic Learning of Fuzzy Rules based on Low Quality Data

Luciano Sánchez [a,*], Inés Couso [b], Jorge Casillas [c]

[a] *Computer Science Department of the Oviedo University, Spain*

[b] *Statistics Department, Oviedo University, Spain.*

[c] *Computer Science Department, Granada University, Spain.*

Abstract

Genetic Fuzzy Systems (GFS) are based on the use of Genetic Algorithms for designing fuzzy systems, and for providing them with learning and adaptation capabilities. In this context, fuzzy sets represent linguistic granules of information, contained in the antecedents and consequents of the rules, whereas the data used in the genetic learning is assumed to be crisp. GFS seldom deal with fuzzy-valued data.

In this paper we address this problem, and propose a set of techniques that can be incorporated to different GFS in order to learn a Knowledge Base (KB) from interval and fuzzy data for regression problems. Details will be given about the representation of non-standard data with fuzzy sets, about the needed changes in the reasoning method of the Fuzzy Rule Based System, and also about a new generalization of the mean squared error to vague data. In addition, we will show that the learning process requires a genetic algorithm that must be capable of optimizing a multicriteria fitness function, containing both crisp and interval-valued criteria.

Lastly, we benchmark our procedures with some machine learning related datasets and a real-world problem of marketing, and the techniques proposed here are shown to improve the generalization properties of other KBs obtained from crisp training data.

Key words: Genetic Fuzzy Systems, Fuzzy Rule Based Systems, Vague Data

[*] Corresponding author.
Email addresses: luciano@uniovi.es (Luciano Sánchez), couso@uniovi.es (Inés Couso), casillas@decsai.ugr.es (Jorge Casillas).

1 Introduction

In most real world problems, data have a certain degree of imprecision. Sometimes, this imprecision is small enough so that it can be safely ignored. On other occasions, the uncertainty of the data can be modeled by a probability distribution (e.g. additive random noise). Lastly, there is a third kind of problems where the imprecision is significant, and a probability distribution is not a natural model. For example, in [23], up to eight sources of information appropriately characterized by intervals were studied: plus-or-minus reports, significant digits, intermittent measurement, non-detects, censoring, data binning, missing data and gross ignorance. As a further matter, there are ongoing researches about the use of fuzzy models for modeling the interaction between variability and imprecision [4,16].

In spite of this widespread use of interval and fuzzy data, the common practice in Genetic Fuzzy Systems (GFS) [29], is to associate fuzzy sets to words and use them to model vague linguistic assertions, but let the models that depend on these words to input crisp data, and output crisp results. In previous works [49–51] we have advocated the use of fuzzy data to learn and evaluate GFS, and raised the use of specific reasoning methods and fuzzy-valued fitness functions to formulate that kind of problems. In this work we give a comprehensive foundation of the changes that must be effected on GFSs in order to learn a Knowledge Base (KB) from imprecise data. The description includes some issues about the representation of low quality data with fuzzy sets, reasoning methods suitable for using vague data, bounds of the accuracy of a KB on vague data, and Multicriteria Genetic Algorithms capable of optimizing a mix of crisp and fuzzy objectives.

This work is organized as follows: In Section 2 we give a theoretical foundation of the use of vague data in our models. In Section 3 we will detail the reasoning method we need to apply in order to obtain an output which is coherent with our interpretation of a fuzzy set. In Section 4 we define how to obtain bounds of the mean squared error from vague data. In Section 5 we describe in detail how to use a GA to optimize the accuracy (known through the bounds defined in the preceding section,) in combination with a crisp measure of complexity. Lastly, in Section 6 we will show how to integrate the new concepts in a GFS, whose behavior is studied through some benchmarks and a practical example related to marketing. Section 7 concludes the paper.

2 Preliminary concepts and theoretical background

Fuzzy KBs contain "IF-THEN" rules, whose antecedents and consequents are composed of fuzzy logic statements. Fuzzy Rule Based Systems (FRBS) contain a KB and an inference engine module that, in turn, contains an inference system, fuzzification and defuzzification interfaces [10]. For the most part, these systems are designed to process crisp infomation. In this sense, it has been shown that FRBS can approximate any real valued function with arbitrary accuracy [9]. Compared to other universal approximators, FRBS have an additional value because they balance accuracy and linguistic interpretability [6], exploiting the potential of fuzzy logic as a representation scheme and calculus for uncertain or vague notions.

2.1 Low quality data

Complementing fuzzy logic, fuzzy statistics are intended to process vague or low quality data, or data where not all the attributes are precisely known. In this context, fuzzy memberships are understood as degrees of similarity, preference or uncertainty [21].

In this paper we will adopt the interpretation that was introduced first in [27,28]: a fuzzy set can be identified with the family of all random sets with the same one-point coverage function as the membership of the former. Furthermore, if we restrict ourselves to the class of all consonant random sets associated to certain one-point coverage function, the membership of the corresponding fuzzy set represents a degree of uncertainty: the membership of an element x to a fuzzy set $\widetilde{A}$ means the degree of possibility that an imprecisely known parameter x_0 has value x, when the available information about x_0 is that "x_0 is $\widetilde{A}$", as proposed by Zadeh [58]. Therefore, any normal possibility measure Π can be represented by a normalized fuzzy set $\widetilde{A}$. The key result, that makes this interpretation interesting for us, was demonstrated in [15] and, independently, in [3]. In these papers it was shown that the fuzzy set $\widetilde{A}$ carries the same information about the unknown parameter x_0 than a family of nested sets $A_1 \supseteq \ldots \supseteq A_n$ for which the probability that the unknown parameter x_0 belongs to A_i is greater or equal than $1 - \alpha_i$, where the strong α-cuts $\widetilde{A}_\alpha$ are the most precise sets that satisfy the inequalities $P(\widetilde{A}_\alpha) \geq 1 - \alpha,\ P \leq \Pi$. The main practical consequence of this is that we can obtain fuzzy memberships from crisp data. We will provide a real-world example of this in Section 6.3, where we compress multi-item data into fuzzy elements.

In short, given a sample of imprecise observations of an unknown parameter x_0, we can infer a fuzzy membership $\widetilde{A}$ that describes our incomplete information about this parameter. Interestly enough, the algorithm that arises was proposed before this last result has been proved (see [41]). These techniques have a great potential in practical situations, because this estimation can be done with samples of very small size, as shown in [40]. Such numerical procedures are illustrated in the example that follows.

Example 2.1 *Let us suppose that a parameter x_0 is described by the following set* *of observations:*

$$X = \{2, 1, 3, 3, 2, 2, 4\}. \tag{1}$$

We will assume that X is a simple random sample of a population whose mean is the unknown value x_0. Let $\widetilde{A}$ be the membership function of the fuzzy set that describes our knowledge about x_0. Then, the family of its cuts $\{\widetilde{A}^\alpha\}$ is a nested family of confidence intervals such that $P(x_0 \in A^\alpha) = P(A^\alpha) \geq 1 - \alpha$.

If we assume that the sample was drawn from a normal population, we obtain the membership function of $\widetilde{X}$ which is shown in Figure 1. Observe that we can approximate it by a triangular membership function without incurring large errors [41]. Note that other, different techniques for estimating the needed confidence intervals would also be possible. For instance, building the membership function from the quantiles of the bootstrap distribution of the sample mean or the median is a sensible choice. In the right part of Figure 1 we have plotted a bootstrap-based estimation of the membership function associated to the same data.

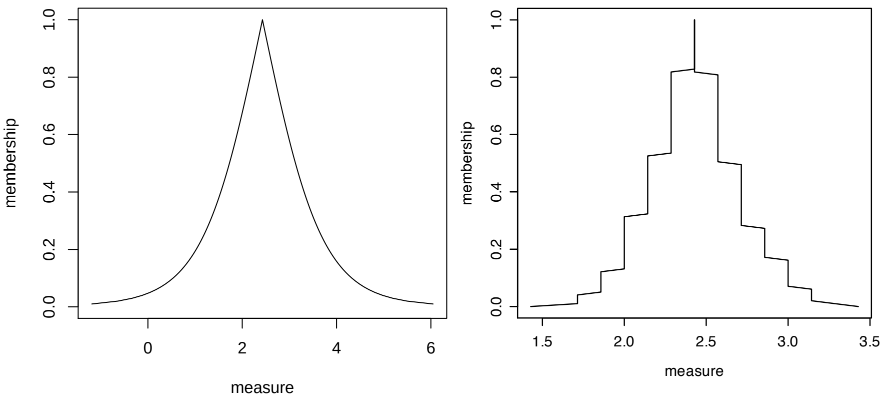


Fig. 1. Membership function of the set $\widetilde{X}$ that represents the sample X in Section 2. The left one was obtained under normality assumptions, and the right one is a basic bootstrap estimation.

3 Combining Fuzzy Inference and Low Quality Data

FRBSs, as we have mentioned, input a list of crisp numbers and generate an intermediate fuzzy set, which is in turn defuzzified to a crisp output. In this section we discuss how to extend this definition to the case where the input is an interval or a fuzzy set.

3.1 An Extension Principle-based Reasoning Method

Let X be the input space and Y the output space, and let $\{\widetilde{A}_i \rightarrow \widetilde{B}_i\}_{i=1,\ldots,M}$ be a set of M fuzzy rules. Given an input $x \in X$, the most common reasoning method for computing the output of a FRBS takes two stages [10]:

(1) An intermediate fuzzy set is composed:

$$\widetilde{\text{out}}(x)(y) = \max_{i=1,\ldots,M} \min\{\widetilde{A}_i(x), \widetilde{B}_i(y)\}. \tag{2}$$

(2) This intermediate fuzzy set is transformed in a crisp value $\text{defuz}(\widetilde{\text{out}}(x)) \in Y$, where $\text{defuz} : \mathcal{F}(\mathbb{R}) \rightarrow \mathbb{R}$ is the function that assigns to each fuzzy number its center of gravity,

$$\text{defuz}(\widetilde{B}) = \frac{\int_{\mathbb{R}} y\widetilde{B}(y)dy}{\int_{\mathbb{R}} \widetilde{B}(y)dy}. \tag{3}$$

Therefore, the value $\text{defuz}(\widetilde{\text{out}}(x)) \in Y$ is computed as follows:

$$\text{defuz}(\widetilde{\text{out}}(x)) = \frac{\int_{\mathbb{R}} y\,\widetilde{\text{out}}(x)(y)dy}{\int_{\mathbb{R}} \widetilde{\text{out}}(x)(y)dy}. \tag{4}$$

Given a set-valued input $A \subseteq X$ (that, in our context, means “all we know about the input is that it is in the set A”) we propose to operate as follows:

(1) We determine a family of intermediate fuzzy sets in the universe $\mathcal{F}(Y)$, $\widetilde{\text{out}}(A) \in \wp(\mathcal{F}(Y))$, defined as

$$\widetilde{\text{out}}(A) = \{\widetilde{\text{out}}(x) \text{ s. t. } x \in A\} \tag{5}$$

(2) An element of $\wp(Y)$ (that is to say, a set of crisp outputs $\text{defuz}(\widetilde{\text{out}}(A)) \in \wp(Y)$) is obtained, according to the following definition:

$$\text{defuz}(\widetilde{\text{out}}(A)) = \{\text{defuz}(\widetilde{\text{out}}(x)) \text{ s. t. } x \in A\}. \tag{6}$$

Lastly, given a fuzzy input $\widetilde{A} \in \mathcal{F}(X)$, we will assign it, according to the Extension Principle (which is compatible with the possibilistic interpretation of fuzzy sets) a fuzzy set computed as follows:

(1) We determine an intermediate fuzzy set on the universe $\mathcal{F}(Y)$, $\widetilde{\text{out}}(\widetilde{A}) \in \mathcal{F}(\mathcal{F}(Y))$, defined as

$$\widetilde{\text{out}}(\widetilde{A})(\widetilde{B}) = \sup\{\widetilde{A}(x) \text{ s. t. } \widetilde{\text{out}}(x) = \widetilde{B}\}, \quad \forall \widetilde{B} \in \mathcal{F}(Y) \tag{7}$$

(2) An element of $\mathcal{F}(Y)$ (that is to say, a fuzzy output) $\text{defuz}(\widetilde{\text{out}}(\widetilde{A})) \in \mathcal{F}(Y)$ is obtained as follows:

$$\text{defuz}(\widetilde{\text{out}}(\widetilde{A}))(y) = \sup\{\widetilde{A}(x) \text{ s. t. } \text{defuz}(\widetilde{\text{out}}(x)) = y\}, \quad \forall y \in Y. \tag{8}$$

Observe that the fuzzy set $\mathrm{defuz}(\widetilde{\mathrm{out}}(\widetilde{A}))$ is associated to the nested family of sets $\{\mathrm{defuz}(\widetilde{\mathrm{out}}(\widetilde{A}_\alpha))\}_{\alpha\in[0,1]}$, and that explains the possibilistic interpretation of this procedure.

The numerical aspects of this definition will be explained with the help of an example.

Example 3.1 *Let us study the following knowledge base:*

```
if X is small then Y is large
if X is large then Y is small
```

where the membership functions of the linguistic values are small$(t) = 1 - t$*, and large*$(t) = t$ *for* $t \in [0, 1]$.

Let us recall first that the prevalent fuzzy logic-based reasoning method is not coherent with our possibilistic interpretation of a fuzzy set. This method consists in the intersection of the cylindric extension of the input with the fuzzy graph of the model, followed by a projection on the output space. The fuzzy graph of this system is

$$\widetilde{out}(x)(y) = \max\{\min\{x, 1-y\}, \min\{1-x, y\}\}. \tag{9}$$

Therefore, given a fuzzy input $\tilde{X}$*, we would obtain a fuzzy set*

$$\widetilde{out}^*(\widetilde{A})(y) = \sup_x \min\{\widetilde{out}(x)(y), \widetilde{A}(x)\}, \tag{10}$$

which does not encode a possibility distribution (it is not normal). To make our point clearer, in the left part of Figure 2 we have plotted the membership function of the set defined in eq. (2), for an arbitrary crisp input $x_0 = 0.25$*. In the right part of the same figure we have plotted the defuzzified output of the FRBS (eq. (4)) of for all values of* x_0 *between 0 and 1. Lastly, in the left part of Figure 3, the inferred output of the knowledge base, when the input is the triangular fuzzy number* $(0.15; 0.25; 0.35)$*, has been computed with eq. (10).*

In the right part of Figure 3, we have computed defuz$(\widetilde{out}(\widetilde{A}))$ *according to our own definition in eq. (8). Observe that the* α*-cuts of defuz*$(\widetilde{out}(\widetilde{A}))$ *are the sets* $\{\mathit{defuz}(\widetilde{out}(x)) \mid \widetilde{A}(x) \geq \alpha\}$*. The modal point of this last set is also the output of the knowledge base for the modal point of the fuzzy input* $\widetilde{A}$*.*

The computation of the membership of $\mathrm{defuz}(\widetilde{\mathrm{out}}(\widetilde{A}))$ can be numerically solved with the fuzzy profile method [19]. In our experimementation, we have approximated all the fuzzy sets by means of piecewise linear continuous membership functions, assuming that the supports of the input data are small enough so that

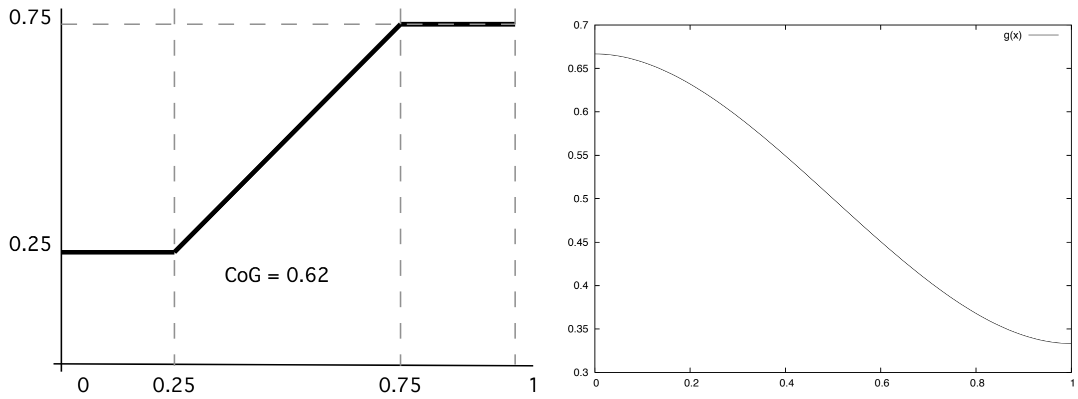


Fig. 2. Left: Fuzzy output of the knowledge base in the example when the input is the crisp number 0.25. Right: Defuzzified output of the same base, for crisp inputs ranging between 0 and 1.

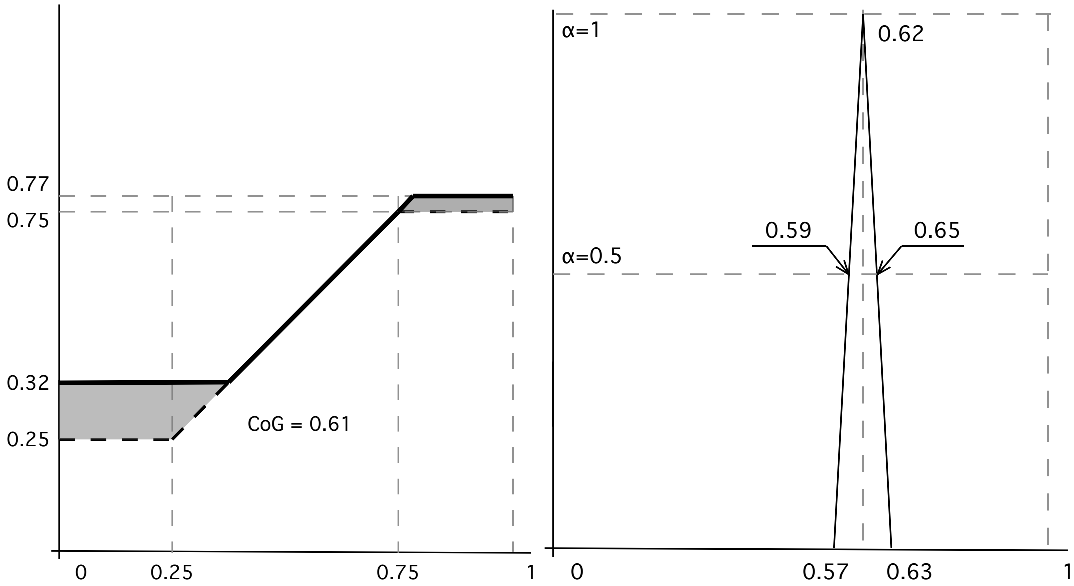


Fig. 3. Left: Output of the knowledge base in the example, by means of eq. 10, when the input is the triangular fuzzy number $(0.15; 0.25; 0.35)$. The grey areas are the difference between this output and that shown in the left part of Figure 2, when the input was the crisp number 0.25. Right: Fuzzy output of the same base when the inference is based on eq. 8.

our model is locally monotonic with regard to each argument, thus we only need to evaluate it in the vertexes of the membership functions.

4 Bounding the accuracy of a FRBS on vague data

In this section we will study how to evaluate the accuracy of a FRBS on vague data. We will generalize the mean squared error of a model f on a sample of N

input-output pairs (x_i, y_i), that is to say:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(f(x_i) - y_i)^2. \tag{11}$$

If some (of all) of the elements of the sample are imprecisely observed, the preceding expression must be extended. Roughly speaking, if we are given a sample of fuzzy pairs $(\widetilde{X}_i, \widetilde{Y}_i)$, we can think of two different extensions of eq. (11): distance-based and fuzzy arithmetic-based. A distance-based generalization would be as follows:

$$\text{FMSE}_d = \frac{1}{N}\sum_{i=1}^{N} d(f(\widetilde{X}_i), \widetilde{Y}_i)^2, \tag{12}$$

where d is a suitable distance between fuzzy sets. Or else, a fuzzy arithmetic-based extension, which would have an expression similar to

$$\text{FMSE}_a = \frac{1}{N}\bigoplus_{i=1}^{N}(f(\widetilde{X}_i) \ominus \widetilde{Y}_i)^2. \tag{13}$$

Arguably, the distance-based extension will produce a crisp value, while the fuzzy arithmetic-based expression is a fuzzy set. We will discuss both in the remainder of this section.

Furthermore, we have decided to derive our definition of MSE in the context of fuzzy random variables (FRV). This makes sense for us, because the concept of variance of a FRV has been studied before [13,14,20,35], and the definition of the MSE of a regression model is similar to that of the variance, thus there is no need to introduce a new definition. Therefore, in the following parts of this section we will review some different interpretations of fuzzy random variable, along with their associated definitions of variance, and choose the most adequate for our purposes.

4.1 Fuzzy Random Variables

A FRV is a function $\widetilde{X} : \Omega \to \mathcal{F}(I\!R)$ that maps each outcome ω of a random experiment to a fuzzy subset $\widetilde{X}(\omega)$ of the real line. In our context, we will assume that both the input and the output data are instances of two random variables $\widetilde{X}$ and $\widetilde{Y}$, thus $(\widetilde{X}_i, \widetilde{Y}_i) = (\widetilde{X}(\omega_i), \widetilde{Y}(\omega_i))$, and these two FRV model our imprecise knowledge about the true features of the object.

There are many different definitions of FRV, that differ in the measurability conditions imposed on the random variables [18,31,32,34,46], but in this work we are not interested in the formal aspects of these definitions, but in the different meanings that can be assigned to the concept of probability measure induced by the FRV. In the context of imprecise probabilities, and following [16], one can think of three different interpretations of the mapping $\widetilde{X}$:

(1) Some authors [46] consider that a FRV is a measurable function, in the classical sense, between certain σ-algebra of events in the original space and a σ-algebra defined over a class of fuzzy subsets of $I\!R$. Each fuzzy set $\widetilde{X}(\omega)$ can be assigned a linguistic label, and therefore probability values can be assigned to these labels. For example: the result is "large" with a probability of 0.5, "medium" with a probability of 0.25 and "small" with a probability of 0.25, where "large", "medium" and "small" are linguistic labels associated to fuzzy subsets of $I\!R$.

(2) A second interpretation, which is based in [33], states that a FRV represents imprecise or vague knowledge about an unknown random variable, which is called "original random variable." Therefore, the membership degree of a point x to the fuzzy set $\tilde{X}(\omega)$ will represent the possibility degree of the assertion "The image of ω is x." For each random variable, $X : \Omega \to I\!R$, they define their "acceptability degree" as the value: $\operatorname{acc}(X) = \inf_{\omega \in \Omega} \tilde{X}(\omega)(X(\omega))$. The acceptability $\operatorname{acc}(X)$ represents the possibility degree of X being the true random variable that models the studied experiment. Therefore, the information provided by the FRV about the probability of a (crisp) outcome is defined by a fuzzy set in $[0, 1]$.

(3) A third interpretation is based in two different sub-experiments, whose sample spaces are Ω and $I\!R$ [2,42]. We assume that the probability distribution that models the first sub-experiment is completely determined, and we consider a family of conditional possibility measures $\{\Pi(\cdot \mid \omega)\}_{\omega \in \Omega}$, each one of them determined by the fuzzy set $\tilde{X}(\omega)$. The fuzzy set $\tilde{X}(\omega)$ is a fuzzy restriction of the possible outcomes of the second experiment, given that the outcome of the first experiment has been ω. The combination, using natural extension techniques [56], of both sources of information, allows us to describe the available information about the probability distribution that rules the second sub-experiment by means of an upper probability. The information provided by the FRV about the probability of a (crisp) outcome is defined by a crisp subset of $[0, 1]$.

These three interpretations are illustrated in the example that follows:

Example 4.1 *Let the sample space be* $\Omega = \{\omega_1, \omega_2, \omega_3\}$, *and* $p(\omega_1) = p(\omega_2) = p(\omega_3) = \frac{1}{3}$. *The random variable* $Y : \Omega \to \{0, 1\}$, *defined by*

$$Y(\omega_1) = 0 \quad Y(\omega_2) = 1 \quad Y(\omega_3) = 1. \tag{14}$$

measures certain feature of an object. The induced probability on the output space *is, therefore,*

$$P_Y(\{0\}) = \tfrac{1}{3} \quad P_Y(\{1\}) = \tfrac{2}{3}. \tag{15}$$

Let us assume that we have incomplete information about Y, and this information is given by the FRV $\widetilde{X}$ (see Figure 4), defined as follows:

$$\widetilde{X}(\omega_1) = \{1/0 + 0.1/1\} \quad \widetilde{X}(\omega_2) = \{1\} \quad \widetilde{X}(\omega_3) = \{1\}. \tag{16}$$

We want to compute the induced probabilities of the sets $\{0\}$ and $\{1\}$ for this FRV, under the three different interpretations mentioned before, and compare this probabilities with those induced by Y.

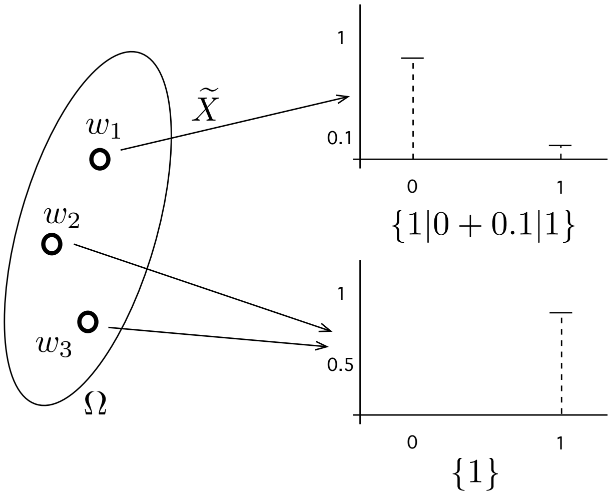


Fig. 4. Fuzzy Random Variable used in the example 4.1

(1) *Since $\{0\}$ is not an image of $\widetilde{X}$, it is immediate that*

$$P_{\widetilde{X}}(\{0\}) = 0, \quad P_{\widetilde{X}}(\{1\}) = \frac{2}{3} \tag{17}$$

(2) *Eight different crisp random variables $X_1, \ldots, X_8$ can be defined between Ω and $\{0,1\}$. These variables and their acceptabilities are shown in the table that follows:*

	Images of the Crisp Random Variables							
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
ω_1	*0*	*1*	*0*	*1*	*0*	*1*	*0*	*1*
ω_2	*0*	*0*	*1*	*1*	*0*	*0*	*1*	*1*
ω_3	*0*	*0*	*0*	*0*	*1*	*1*	*1*	*1*
acceptability	*0*	*0*	*0*	*0*	*0*	*0*	*1*	*0.1*

The probabilities induced by the variables with non-null acceptability are $P_{X_7}(\{0\}) = \frac{1}{3}$, $P_{X_7}(\{1\}) = \frac{2}{3}$, $P_{X_8}(\{0\}) = 0$, $P_{X_8}(\{1\}) = 1$. *Therefore,*

$$P_{\widetilde{X}}(\{0\}) = \{0.1/0 + 1/\tfrac{1}{3}\} \quad P_{\widetilde{X}}(\{1\}) = \{1/\tfrac{2}{3} + 0.1/1\} \tag{18}$$

Observe that both probabilities are fuzzy subsets of $[0, 1]$*, imprecise descriptions of the values* $P_Y(\{0\}) = \frac{1}{3}$ *and* $P_Y(\{1\}) = \frac{2}{3}$.

(3) Let $\pi_\omega(x) = \Pi(\{x\} \mid \omega)$. *If the result of the first sub-experiment is*

$$\begin{cases} \omega_1 & \textit{then } \pi_{\omega_1}(0) = 1,\ \pi_{\omega_1}(1) = 0.1 \\ \omega_2, \omega_3 & \textit{then } \pi_{\omega_2}(0) = \pi_{\omega_3}(0) = 0,\ \pi_{\omega_2}(1) = \pi_{\omega_3}(1) = 1. \end{cases}$$

and

$$\begin{aligned} P(\{0\}) &\leq \tfrac{1}{3}\pi_{\omega_1}(0) + \tfrac{1}{3}\pi_{\omega_2}(0) + \tfrac{1}{3}\pi_{\omega_3}(0) = \tfrac{1}{3} \\ P(\{1\}) &\leq \tfrac{1}{3}\pi_{\omega_1}(1) + \tfrac{1}{3}\pi_{\omega_2}(1) + \tfrac{1}{3}\pi_{\omega_3}(1) = \tfrac{21}{30} \end{aligned} \tag{19}$$

therefore

$$\frac{9}{30} \leq P(\{0\}) \leq \frac{1}{3}, \quad \frac{2}{3} \leq P(\{1\}) \leq \frac{21}{30} \tag{20}$$

Observe that both probabilities are intervals of $[0, 1]$*, that contain the values* $P_Y(\{0\}) = \frac{1}{3}$ *and* $P_Y(\{1\}) = \frac{2}{3}$.

4.2 Proposed definition of Fuzzy Mean Squared Error

In this section we will formalize the estimation of the accuracy of a FRBS on a vague sample of data. We will assume that the residual of the model (that is to say, the difference between the desired output of the model and the actual output) is a FRV, whose second moment about the origin we will call Fuzzy Mean Squared Error. To the best of our knowledge, there are not previous works about the definition of the MSE of a FRV, but the variance of a FRV (the second moment about the mean) has been previously studied in the context of the three interpretations seen before: the variance of the FRV has been defined as a crisp number [35], a fuzzy number [13,14,33] or an interval [14].

We can understand the word "difference" in the definition of residual as "distance" between the fuzzy desired output and the actual output of the model, or else as a fuzzy restriction on the values on the unknown, actual error. Let us clarify this with an example: imagine that one point in the training sample comprises two intervals $([1, 2], [2, 4])$, and we want to evaluate in this point the residual of the model $y = x^2$, whose output is $f([1, 2]) = \{x^2 \mid x \in [1, 2]\} = [2, 4]$. We can think of two different interpretations:

(1) $d([2,4],[2,4]) = 0$, because the actual output and the desired output are the same.
(2) $[2,4] \ominus [2,4] = [-2,2]$ because the actual output is some unknown value in $[2,4]$, the desired output is other, possibly different, point in $[2,4]$ and the most we can say about its difference is that it is in $[-2,2]$.

Arguably, the first interpretation not compatible with the possibilistic interpretation of a fuzzy set (but, conversely, that would have been our choice if our fuzzy sets represented degrees of compatibility between numerical and linguistic values). This prevents the use of the definition in proposed in [35]. Furthermore, from a purely computational point of view, the third definition is advantageous, because the accuracy of a regression model can be measured with one interval, which can be represented with less parameters than the membership of a fuzzy set and therefore allows for faster calculations. For that reason, we will use this last definition in the remaning of the paper.

4.2.1 Theoretical definition

Let $\widetilde{D}(\omega) = \widetilde{Y}(\omega) \ominus \widetilde{f}(\widetilde{X}(\omega))$ be the residual of f in the object ω. We want to define the second moment of $\widetilde{D}$ about the origin, according to the third interpretation seen before.

On the one hand, the probability measure P models the first sub-experiment (that is, obtaining an instance ω of the train set). On the other hand, the fuzzy sets $\widetilde{D}(\omega)$ represent conditional possibility measures $\Pi(\cdot|\omega)$. The residual of the FRBS in the object ω is in a subset of the output space $A \in \beta_{I\!R}$ with probability

$$P(A \mid \omega) \leq \Pi(A \mid \omega) = \Pi_{\widetilde{D}(\omega)}(A) = \sup_{x \in A} \widetilde{D}(\omega)(x), \; \forall\, A \in \beta_{I\!R}, \; \forall\, \omega. \qquad (21)$$

All we know about the probability distribution that models this second experiment is that it is in the set:

$$\begin{aligned} &\{Q_2 \mid Q_2 \text{ marginal of } P \text{ and } Q(\cdot|\cdot),\; Q(\cdot|\cdot) \in \mathcal{C}\} \\ &\text{where } \mathcal{C} = \{Q(\cdot|\cdot) \text{ transition prob.} \mid Q(A|\omega) \leq \Pi(A|\omega) \\ &\qquad\qquad\qquad\qquad\qquad\qquad \forall\, A \in \beta_{I\!R}, \omega \in \Omega\}. \end{aligned} \qquad (22)$$

Therefore, in the proposed imprecise probabilities model, all we know about the second moment about the origin of the residual, or *fuzzy mean squared error (FMSE),* is that it is in the interval:

$$\mathrm{FMSE}(\tilde{D}) = \left\{ \int_{I\!R} x^2 dQ_2 \mid Q_2 \text{ marginal of } P \text{ and } \; Q(\cdot|\cdot),\; Q(\cdot|\cdot) \in \mathcal{C} \right\} \qquad (23)$$

or, alternatively [14]:

$$\mathrm{FMSE}(\tilde{D}) = \left\{ \int\limits_{I\!R} d^2 dQ \mid Q(A) \leq \overline{P}_{\tilde{D}}(A), \, \forall \, A \in \beta_{I\!R} \right\} \tag{24}$$

where $\overline{P}_{\tilde{D}}$ is the sub-additive set-valued function given by

$$\overline{P}_{\tilde{D}}(A) = \int\limits_0^1 P^*_{\tilde{D}_\alpha}(A) \, d\alpha, \forall \, A \in \beta_{I\!R} \tag{25}$$

and, for every $\alpha \in (0, 1]$, $P^*_{\tilde{D}_\alpha}$ is Dempster's upper probability of the random set $\tilde{D}_\alpha$.

In the next section we will detail how to numerically approximate the FMSE of a FRBS from a sample of fuzzy residuals.

4.2.2 Computer algorithm for approximating the FMSE

We will use the procedure described in [20] to approximate the FMSE of an FRV:

(1) For each of the membership functions $\widetilde{D}_i = \widetilde{Y}_i \ominus f(\widetilde{X}_i)$, we have to compute the fuzzy set $\widetilde{D}^2_i$, whose α-cuts are $\widetilde{D}^2_\alpha = \{x^2 \mid x \in \widetilde{D}_\alpha\}$.

Since the function x^2 is not locally monotonic, to compute this cuts we must divide the area under the membership functions in zones separated by the changes in the slope of this function. This is graphically illustrated in Figure 5. If the membership of $\widetilde{D}$ does not cut the line $x = 0$, the number of vertices is preserved. Otherwise, the left part of the profile is replaced by a vertical segment, and the new right profile is the maximum of the squares of the former left and right parts.

(2) We compute the FMSE of each $\widetilde{D}^2(\mathbf{x}_i)$. Let M^-_i be the left profile of $\widetilde{D}^2(\mathbf{x}_i)$, and M^+_i the right one. Then,

$$\mathrm{FMSE}_i = \left[\int_0^1 M^-_i d\alpha, \int_0^1 M^+_i d\alpha \right] \tag{26}$$

(3) The FMSE of the whole model is contained in the interval

$$\mathrm{FMSE} = \frac{1}{n} \left(\mathrm{FMSE}_1 \oplus \ldots \oplus \mathrm{FMSE}_n \right). \tag{27}$$

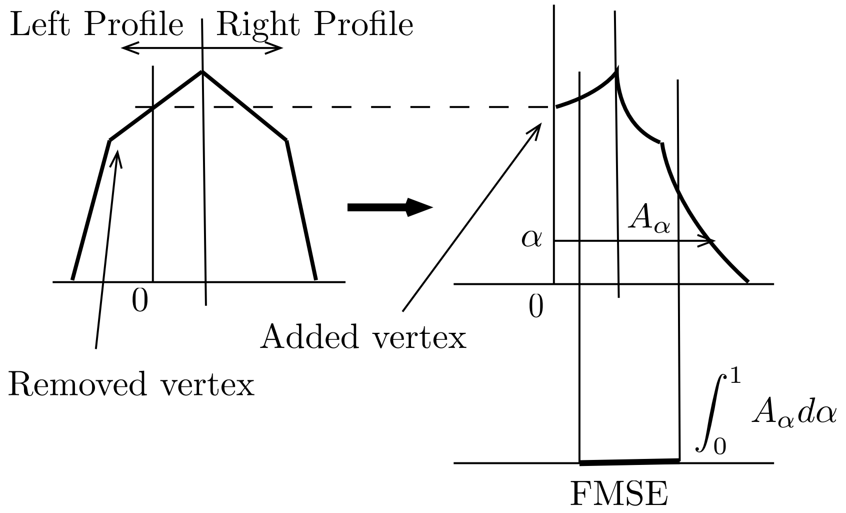


Fig. 5. The same extension to nonmonotonic functions of the profile method that is used to compute the sample variance of an FRV can be applied to obtain its FMSE.

5 Using a GA for learning a fuzzy KB from vague data

In this section, the NSGA-II algorithm [17] will be extended so that it can find a set of nondominated solutions for a two-objectives problem: 1) the mean squared error of the FRBS, as defined in the last section, and 2) the complexity of the KB. The first objective is an interval, and the second one is crisp.

For the most part, there are two modules in the NSGA-II algorithm where the fitness function is assumed to be a vector of crisp numbers: the precedence (dominance) operator, and the crowding distance. Therefore, our extension consists in alternate definitions for these modules.

5.1 Precedence in imprecise fitness-based Genetic Algorithms

Let θ_1, θ_2 be the fitnesses of two KBs (i.e. the MSEs of these two KBs on the training set). In this section, we will assume that θ_1 and θ_2 are unknown, but we know two intervals FMSE_1 and FMSE_2, computed as explained in the preceding section, that contain them. We want to determine whether one individual precedes the other, thus we need a procedure that estimates whether the probability of $\theta_1 < \theta_2$ is greater than that of $\theta_1 \geq \theta_2$ (thus $\text{FMSE}_1 \prec \text{FMSE}_2$) or not. We also want to find those cases where there is no statistical evidence in FMSE_1 and FMSE_2 that makes us prefer one of them (thus $\text{FMSE}_{x_1} \parallel \text{FMSE}_{x_2}$). Our approach can be regarded as a PQI interval order [37], where there is a zone of hesitation between strict difference and strict similarity.

5.1.1 Strong Dominance

Without further assumptions, when FMSE_1 and FMSE_2 are non-disjoint intervals, we do not have evidence to prefer one of them. Otherwise, the decision is trivial. This criterion has been called *strong dominance* in [36].

5.1.2 Probabilistic Prior

The inability to distinguish between intervals with a non-empty intersection is a major problem. We can improve the situation by introducing prior knowledge about the probability distribution of the fitness. If a joint probability $P(\theta_1, \theta_2)$ was known, comparing two individuals would be a statistical decision problem. For instance, we can decide that $\text{FMSE}_1 \prec \text{FMSE}_2$ when

$$\frac{P(\{(\theta_1, \theta_2) : \theta_1 < \theta_2\}}{P(\{(\theta_1, \theta_2) : \theta_1 \geq \theta_2\}} > 1. \tag{28}$$

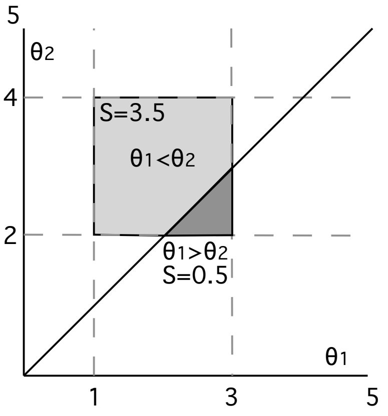


Fig. 6. Graphical representation of the example 5.1.

For instance, in [55] it was assumed that $P(\theta_1, \theta_2)$ was uniform. This use of a uniform prior will be made clear in the examples below.

Example 5.1 *Let FMSE*$_1 = [1, 3]$ *and FMSE*$_2 = [2, 4]$ *two non-disjoint intervals. If we assume that* $P(\theta_1, \theta_2)$ *is uniform in* $[1, 3] \times [2, 4]$ *(see Figure 6) we obtain*

$$\frac{P(\{(\theta_1, \theta_2) : \theta_1 < \theta_2\}}{P(\{(\theta_1, \theta_2) : \theta_1 \geq \theta_2\}} = \frac{3.5/4}{0.5/4} > 1 \tag{29}$$

thus we can state that FMSE$_1 \prec$ *FMSE*$_2$*.*

Example 5.2 *Let FMSE*$_1 = [1, 5]$ *and FMSE*$_2 = [1.9, 4]$ *two non-disjoint intervals. The application of the same principle produces*

$$\frac{P(\{(\theta_1, \theta_2) : \theta_1 < \theta_2\}}{P(\{(\theta_1, \theta_2) : \theta_1 \geq \theta_2\}} = \frac{4.095}{4.305} < 1 \tag{30}$$

therefore FMSE$_2 \prec$ *FMSE*$_1$*.*

The uniform prior defines a total order in the population, since every pair of intervals is comparable. We may question the consistency of this order, though. In the last example, there might be situations where a fitness $[1, 5]$ could be prefered to $[1.9, 4]$, and it is also reasonable to state that these two intervals cannot be compared.

5.1.3 Imprecise Probabilities-based Prior

Assuming a probabilistic prior is numerically the same as normalizing the fuzzy output of the model (so that the sum of all the memberships is 1) and then assume that these normalized memberships are the unknown probabilities of the different observations of the output. Therefore, we propose to assume a less strong prior knowledge, something intermediate between the strong dominance and the uniform prior, that allow us to compare some non-disjoint intervals but that also detects that some intervals are almost the same. In this section, we will combine an upper bound of the posterior probability with an imprecisely known prior.

Let us assume for the time being that we do not know $P(\theta_1, \theta_2)$ but a pair of lower-upper probabilities

$$P_*(\theta_1, \theta_2) \leq P(\theta_1, \theta_2) \leq P^*(\theta_1, \theta_2) \tag{31}$$

thus $\mathrm{FMSE}_1 \prec \mathrm{FMSE}_2$ when

$$\frac{P_*(\{(\theta_1, \theta_2) : \theta_1 < \theta_2\})}{P^*(\{(\theta_1, \theta_2) : \theta_1 \geq \theta_2\})} > 1. \tag{32}$$

It is remarked that eq. 32 makes it possible that $\mathrm{FMSE}_1 \not\prec \mathrm{FMSE}_2$ and, at the same time, $\mathrm{FMSE}_2 \not\prec \mathrm{FMSE}_1$. In this case, we will state that $\mathrm{FMSE}_1 \parallel \mathrm{FMSE}_2$. Also observe that, since $P^*(A) = 1 - P_*(A^c)$, eq. 32 reduces to

$$P_*(\{(\theta_1, \theta_2) : \theta_1 \geq \theta_2\}) \geq 1/2 \tag{33}$$

The set of priors defined by the bounds P^* and P_* can range from the completely uninformative $P^*(\theta_1, \theta_2) = 1$ and $P_*(\theta_1, \theta_2) = 0$ (and then this criterion reduces to the strong dominance) to a family where $P_* = P^*$, that contains only one probability distribution (and then it reduces to the probabilistic prior of the preceding section). We want to use a family of priors that is restrictive enough so we can assign precedences to intervals with a non-empty intersection, but not so restrictive as the uniform prior. Many different families can be defined, depending on the practical problem in hand and the interest of the researcher. In the following, we will use the family of priors that is defined in the example that follows:

Example 5.3 *Let us assume that the MSE is between 0 and 5, and let FMSE*$_1 = [1, 3]$ *and FMSE*$_2 = [2, 4]$ *two non-disjoint intervals. We want to decide whether FMSE*$_1 \prec$ *FMSE*$_2$*, assuming the family* $\mathcal{P}$ *of priors that includes all the probability distributions whose density is*

$$f(x) = \epsilon + \frac{1}{m} - \frac{2\epsilon}{m}x \tag{34}$$

for $\epsilon \in [0, 1/m]$*, where* m *is an upper bound of the fitness of an individual (5, in this example). A graphical representation of this family is shown in Figure 7. This family models our incomplete knowledge about the density of individuals in the genetic population. It states that the probability of an individual having a low fitness is higher than the opposite, but we do not know how much higher, thus any* *distribution between the uniform ($\epsilon = 0$) and the linear density with the maximum slope ($\epsilon = 1/m$) is reasonable.*

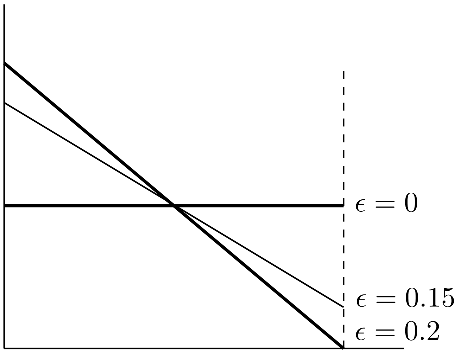


Fig. 7. Prior familiy of the example 5.3

The calculations are as follows:

$$\begin{aligned} P_*(\{(\theta_1,\theta_2):\theta_1<\theta_2\}) &= 1 - P^*(\{(\theta_1,\theta_2):\theta_1\geq\theta_2\}) \\ &= 1 - \sup_{P\in\mathcal{P}} \frac{\int_{\theta_1\geq\theta_2} I_{[1,3]}(\theta_1) I_{[2,4]}(\theta_2) dP(\theta_1,\theta_2)}{\int I_{[1,3]}(\theta_1) I_{[2,4]}(\theta_2) dP(\theta_1,\theta_2)} \\ &= 1 - \sup_{\epsilon\in[0,0.2]} \frac{\int_2^3\int_2^x (\epsilon+\frac{1}{5}-\frac{2\epsilon}{5}x)(\epsilon+\frac{1}{5}-\frac{2\epsilon}{5}y)\,\mathrm{d}x\mathrm{d}y}{\int_1^3\int_2^4 (\epsilon+\frac{1}{5}-\frac{2\epsilon}{5}x)(\epsilon+\frac{1}{5}-\frac{2\epsilon}{5}y)\,\mathrm{d}x\mathrm{d}y} \\ &= 0.869 \end{aligned} \tag{35}$$

and $P_*(\{(\theta_1,\theta_2):\theta_1\geq\theta_2\}) = 0.125$, therefore we state that FMSE$_1 \prec$ FMSE$_2$.

Example 5.4 *Let FMSE$_1 = [1,5]$ and FMSE$_2 = [1.9,4]$. We want to decide whether FMSE$_1 \prec$ FMSE$_2$, assuming that the family of priors of the preceding example is used.*

In this case, the calculations produce the values

$$P_*(\{(\theta_1,\theta_2):\theta_1<\theta_2\}) = 0.332 \tag{36}$$

$$P_*(\{(\theta_1,\theta_2):\theta_1\geq\theta_2\}) = 0.488 \tag{37}$$

thus FMSE$_1 \parallel$ FMSE$_2$.

5.2 Crowding distance

The last module that needs to be extended is the crowding distance. The crowding distance is aimed to uniformly sample the front, making the individuals in the most dense areas less likely to be selected. In the crisp case, if the s individuals in the population are sorted such that $\theta_i < \theta_{i+1}$, thus the local density at the i-th individual is approximately

$$\rho_i = \frac{3}{s \cdot (\theta_{i+1} - \theta_{i-1})} \tag{38}$$

because the number of points lying in the volume $[\theta_{i-1}, \theta_{i+1}]$ is three. In other words, the crowding distance is inversely proportional to the density of individuals in the fitness space, based on a 2-neighbours criterion:

$$d_i = \frac{3}{s\rho_i} \tag{39}$$

where ρ_i is the local density at the i-th individual. To extend this definition to the interval case, let us suppose the s individuals in the population have a fitness $\theta_i \in I_i$. The local density is bounded by

$$\rho_i \in \left[\frac{3}{sV_i^{\max}}, \frac{3}{sV_i^{\min}} \right] \tag{40}$$

where $V_i^{\max}$ is the smallest interval that completely contains the fitness of I_i and two other individuals,

$$I_i \subseteq V_i^{\max}, \quad \#\{j : I_j \subseteq V_i^{\max}\} = 3 \tag{41}$$

and $V_i^{\min}$ is the smallest individual that has a non-empty intersection with the fitness of three individuals, I_1 being one of them,

$$I_i \cap V_i^{\max} \neq \emptyset, \quad \#\{j : I_j \cap V_i^{\max} \neq \emptyset\} = 3. \tag{42}$$

Therefore, the crowding distance associated to the i-th individual is (see Figure 8)

$$d_i \in \left[||V_i^{\min}||, ||V_i^{\max}|| \right]. \tag{43}$$

Unfortunately, this generalization does not produce good results, because the upper bound of the crowding depends too much on the uncertainty of the fitness being compared. An individual surrounded by two identical copies of itself can be assigned a high upper crowding distance if these individuals are uncertain.

Therefore, we have decided to use a crisp metric between the imprecise values of the fitness. Many different metrics between intervals can be can be chosen: euclidean, like Bertoluzza's, L_2 or Wasserstein, or non-euclidean, like Hausdorff, L_1, L_∞ (see, for instance, the survey in [30]). Possibly, the most common are [5]:

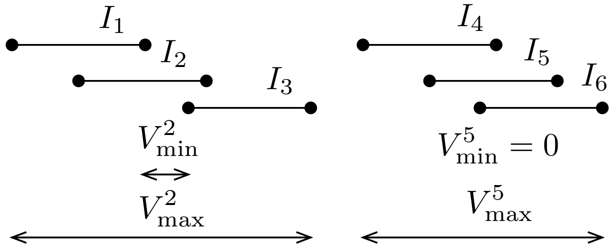


Fig. 8. Minimum and maximum crowding distances between interval-valued fitness functions. The maximum distance is the volume of the smallest interval that contains the fitness of three individuals, while the minimum distance is the volume of the interval that has non-null intersection with three individuals.

(1) The Hausdorff distance, that for two intervals $A = [a_1, a_2]$ and $B = [b_1, b_2]$ is defined as

$$d_H(A, B) = \max\{|a_1 - b_1|, |a_2 - b_2|\}. \tag{44}$$

(2) A combination of the two differences $|a_1 - b_1|$ and $|a_2 - b_2|$. For instance,

$$d_1^*(A, B) = \frac{1}{2}[|a_1 - b_1| + |a_2 - b_2|] \tag{45}$$

or the class of distances introduced in [5]:

$$d^2(A, B) = \int\limits_0^1 [t \cdot |a_1 - b_1| + (1 - t)|a_2 - b_2|]^2 dg(t) \tag{46}$$

where g is a probability measure.

We wish the crowding distance between two intervals $x \pm \epsilon$ and $y \pm \epsilon$ to lay between the bounds mentioned before, not to depend on ϵ, and also to be compatible with the crisp definition, i.e. $d(x \pm \epsilon, y \pm \epsilon) = |y - x|$. In other words, we want a metric between imprecise values of fitness that is not influenced by the nonspecificity of the measures. We have decided to use the Hausdorff metric, which fulfills all the needed properties, but many other distances could serve as well [25,47].

The Hausdorff distance produces a selection probability inversely proportional to a local density of individuals defined by

$$\rho_i = \frac{3}{V_H(I_i)} \tag{47}$$

where $V_H(I_i)$ is the volume of a sphere that contains the fitness of I_i and with the lowest radius needed to also contain the fitness of two other individuals in the population. The crowding distance is therefore, defined as the distance between the nearest (as defined by the Hausdorff metric) individual preceding I_i and the nearest individual following I_i. The first and the last individuals are assigned a high crowding distance. The meaning of 'precede', 'follow', 'first' and 'last' is given by an ordering depending on the precedence operation.

5.3 Precedence between fitness values comprising both an interval objective and a crisp objective

In Section 5.1.3 we have explained how to implement the precedence between interval-valued fitness functions. Heterogeneous pairs, comprising an interval and an integer value, must be compared eventually. The precedence between these pairs is as follows: for any two compound fitness values (n_1, FMSE_1) and (n_2, FMSE_2), we say that

$$(n_1, \text{FMSE}_1) \preceq (n_2, \text{FMSE}_2) \text{ if } \begin{cases} n_1 < n_2 \text{ and } \text{FMSE}_1 \preceq \text{FMSE}_2 \\ n_1 \leq n_2 \text{ and } \text{FMSE}_1 \prec \text{FMSE}_2 \end{cases} \tag{48}$$

where $\text{FMSE}_1 \preceq \text{FMSE}_2$ is defined as $(\text{FMSE}_1 \prec \text{FMSE}_2) \vee (\text{FMSE}_1 \parallel \text{FMSE}_2)$.

6 Experimental results

In this section we detail how to combine all the concepts introduced in this paper and build a GFS able to learn a KB from vague data. In the first place, we will integrate the new representation, inference and interval-valued fitness function into a new proposal of GFS. Next, we benchmark the new algorithm with some common datasets, and also with some synthetic datasets of our own, where the stochastic uncertainty and the imprecision in the observation are mixed in different proportions. Lastly, a second GFS is applied to a practical problem of linguistic modeling in marketing, and the results compared to previous works in the same subject.

6.1 Proposed Genetic Fuzzy System

Apart from the inference operators and the fitness, that have been discussed in 3 and Sections 4, the definition of a GFS also includes the fuzzy rule structure, the coding scheme, the evolutionary scheme and the genetic operators. The only change with respect to an standard GFS is in the evolutionary scheme, which has to optimize an interval valued fitness function or a combination of an interval and a real number.

6.1.1 *Fuzzy Rule Structure*

We have used a compact description based on the disjunctive normal form (DNF) [26]. This kind of fuzzy rule structure has the following form:

$$\textbf{IF } X_1 \text{ is } \widetilde{A_1} \text{ and } \ldots \text{ and } X_n \text{ is } \widetilde{A_n} \textbf{ THEN } Y \text{ is } B$$

where each input variable X_i takes as a value a set of linguistic terms $\widetilde{A_i} = \{A_{i1} \vee \ldots \vee A_{il_i}\}$, whose members are joined by a disjunctive (T-conorm) operator, whilst the output variable remains a usual linguistic variable with a single label associated. We use the *bounded sum* ($min\{1, a + b\}$) as T-conorm. The structure is a natural support to allow the absence of some input variables in each rule (simply making $\widetilde{A_i}$ be the whole set of linguistic terms available).

6.1.2 *Coding scheme*

Each individual of the population represents a set of fuzzy rules (i.e., Pittsburgh style). Each chromosome consists of the concatenation of a number of rules. Each rule (part of the chromosome) is encoded by a binary string for the antecedent part and an integer coding scheme for the consequent part. The antecedent part has a size equal to the sum of the number of linguistic terms used in each input variable. The allele ‘1’ means that the corresponding linguistic term is used in the corresponding variable. The consequent part has a size equal to the number of output variables. In that part, each gene contains the index of the linguistic term used for the corresponding output variable.

For example, assuming we have three linguistic terms (S, M, and L) for each input/output variable, the fuzzy rule [IF X_1 is S and X_2 is {M or L} THEN Y is M] is encoded as [100|011||2]. Therefore, a chromosome would be the concatenation of a number of these fuzzy rules, e.g., [100|011||2 010|111||1 001|101||3] for a set of three rules.

6.1.3 *Evolutionary Scheme*

A generational approach with the multiobjective NSGA-II replacement strategy [17] is considered. Crowding distance in the objective function space is used; in each front, this measure is normalized by the minimum and maximum values for each objective in that front. Binary tournament selection based on the non-domination rank (or the crowding distance when both solutions belong to the same front) is applied.

Non-standard extensions of the NSGA-II algorithms were used, as explained in Section 5, for implementing the non-dominated sorting and the crowding distance. The three types of precedence introduced in that section (strong dominance, probabilistic prior and imprecise prior) were compared, as we will show in the numerical results.

6.1.4 Genetic Operators

The *crossover* operator randomly chooses a cross point between two fuzzy rules at each chromosome and exchanges the right string of them. Therefore, the crossover only exchanges complete rules, but it does not create new ones since it respects rule boundaries on chromosomes representing the individual rule base. In the case that inconsistent rules appear after crossover, the ones whose antecedent is logically subsumed by the antecedent of a more general rule are removed. Redundant rules are also removed.

The *mutation* operator randomly selects an input or output variable of a specific rule. If an input variable is selected, one of the three following possibilities is applied: *expansion*, which flips to '1' a gene of the selected variable; *contraction*, which flips to '0' a gene of the selected variable; or *shift*, which flips to '0' a gene of the variable and flips to '1' the gene immediately before or after it. The selection of one of these mechanisms is made randomly among the available choices (e.g., contraction cannot be applied if only a gene of the selected variable has the allele '1').

If an output variable is selected, the mutation operator is different between descriptive and scatter knowledge bases [10]. For descriptive rules, the mutation operator simply increases or decreases the integer value. In the same way, specific rules which appear after mutation are subsumed by the most general ones and redundant rules are removed. For scatter rules, the consequent is a real number, which is determined by applying the SVD selection described in [53] to the rules codified in the chromosome.

6.2 Numerical results

In this section we will benchmark the results of the proposed GFS in three different scenarios:

(1) Combination of stochastic noise and imprecision
(2) Synthetic datasets with stochastic noise
(3) Machine Learning benchmarks

The first type of problems is intended to show the advantages of the use of the interval-valued fitness function (from now on, we will use the acronym IVFF for naming the GFS proposed in this paper) in problems where the amount of stochastic noise and the imprecision in the observation are known. The second category shows

Name	Inputs	Examples
f_1	2	675
$f_1 - 10$	2	675
$f_1 - 20$	2	675
$f_1 - 50$	2	675
f_2	2	675
$f_2 - 10$	2	675
$f_2 - 20$	2	675
$f_2 - 50$	2	675
elec	2	490
Friedman	5	1200
machine-CPU	6	209
daily-elec	6	365
building	14	4208

Table 1
Properties of the datasets used in the numerical analysis. f_1, f_2 and Friedman are synthetic datasets. $f_1 - *$ and $f_2 - *$ contain different amounts of stochastic noise. The remaining problems contain real-world data.

the behavior of the algorithm in datasets for which the theoretical error is known, and the third family of problems are used to compare this and other statistical and artificial intelligence algorithms over datasets in the public domain.

Nine laboratory problems and four real world problems have been used to benchmark the algorithms proposed in this paper (see Table 1). The laboratory problems are:

- f_1: high slope function $f_1(x, y) = x^2 + y^2$, no noise added [48]. $f_1 - 10$, $f_1 - 20$, $f_1 - 50$ are the same function, with 10%, 20% and 50% of gaussian noise.
- f_2: low slope function $f_2(x, y) = 10(x - xy)/(x - 2xy + y)$, no noise added [48]. $f_2 - 10$, $f_2 - 20$, $f_2 - 50$ are the same function, with 10%, 20% and 50% of gaussian noise.
- *Friedman*: Synthetic benchmark dataset proposed in [24] $f(x, y, z, t, u) = 10 \sin(\pi xy) + 20(z - 0.5)^2 + 10t + 5u$, with added gaussian noise.

The real world problems are:

- *elec*: Relationship between the population and the radius of a village and the length of its electrical line [12].
- *Machine-CPU*: Relative performance of CPUs based on other computer charac-

teristics [22].

- *daily-elec*: Daily price of electrical energy in Spain, in 2003, as a function of the technology mix [1,39].
- *building*: problem taken from the benchmark [45], with 8 binary and 6 continuous input variables.

In all experiments, 50% of the points were used to train the models, which were tested against the remaining 50%. The roles of training and test sets were interchanged and the process repeated, and this was replicated 5 times for different permutations of the dataset, which gives 10 repetitions of the learning algorithm for each dataset. The results shown are the mean test values of 10 repetitions of the experiments.

6.2.1 *Performance in imprecise datasets*

In this section we compare the algorithm IVFF and a state-of-the-art crisp GFS (an Iterative Rule Learning algorithm [52]) in imprecisely observed data, crafted by us to show a case when the observation error has a high influence in the performance of the FRBS. The datasets have been constructed from the functions f_1 and f_2.

	1%	1%	5%	5%	10%	10%
	MSE	FMSE	MSE	FMSE	MSE	FMSE
f_1	0.87	**0.37**	6.99	**6.34**	25.92	**25.02**
f_1-10	2.14	**1.78**	8.55	**8.10**	29.56	**28.56**
f_2	0.32	**0.26**	0.56	**0.52**	1.29	**1.28**
f_2-10	0.47	**0.40**	0.79	**0.78**	1.89	**1.86**
elec	458	**436**	604	**576**	1060	**1027**

Table 2
Comparison of crisp and fuzzy GFSs in datasets with 1%, 5% and 10% of interval-valued imprecision in the outputs. The columns labeled “FMSE” show the average test MSE of FRBSs obtained with the method in this paper and the augmented interval-valued dataset explained in the text. The columns labeled “MSE” display the average test error of a FRBS that minimizes the train MSE on the non-augmented crisp dataset. The FMSE-based FRBS scored better in all the tests. This shows that there exist problems where the observation error is not best processed assuming a zero mean probability distribution of the noise.

We have assumed that our perception of the data is biased: the output variable was added a fixed amount of 1%, 5% and 10% of the range of the variable. We have also added, in the rows marked with “-10”, random noise with Gaussian distribution, null mean and variance equal to 10% of that of the variable. The crisp GFS has been trained with this contaminated data, and IVFF was trained with intervals, centered in the contaminated data, and the same width as the imprecision in the observation; that is to say, intervals that always contain the sum of the original variable and the stochastic noise. For instance, if the true value of the variable is 1 and the imprecision in the observation is ± 0.1, the crisp GFS is trained with the value 1.1 and IVFF is trained with the interval $[1.0, 1.2]$. Notice that we have only used datasets for which the stochastic error is lower than the observation error.

The test results on the original data are shown in Table 2. As expected, IVFF improved the results in all the tests, but the statistical significance is lower if the imprecision error is not the main source of uncertainty. The boxplots in Figure 9 show the statistical relevance of the differences expressed in Table 2. These results show that there exist problems where the observation error is not best processed assuming a zero mean probability distribution of the noise.

6.2.2 Performance in synthetic datasets with stochastic noise

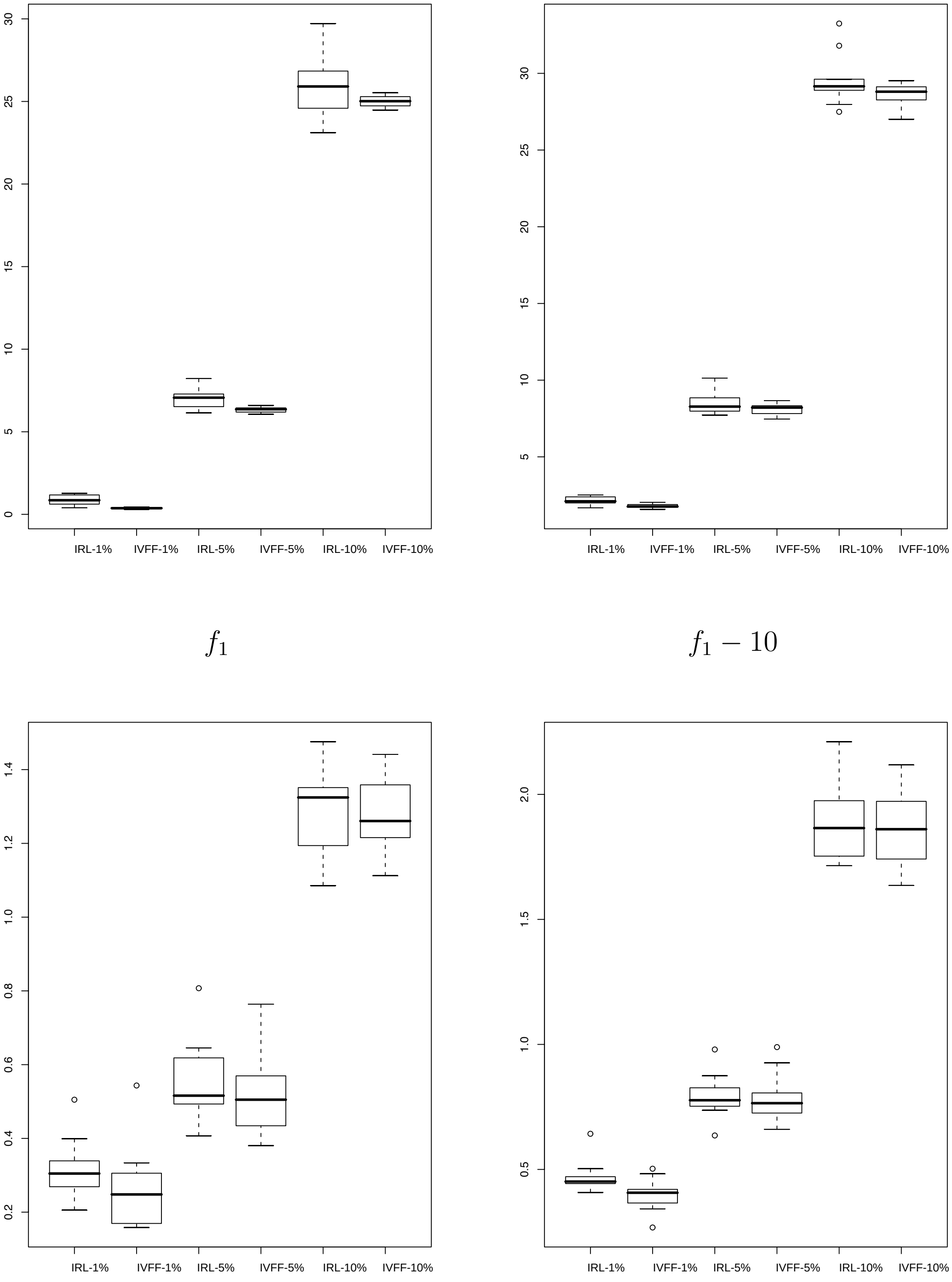
The second benchmark compares IVFF with other approaches in synthetic datasets, whose degree of contamination with stochastic noise is known. The algorithms to which we compare are: Linear Regression (LIN), Quadratic Regression (QUA) a Conjugate-Gradient trained Multilayer Perceptron (NEU), and the fuzzy rule learning algorithms that follow: Wang and Mendel [57] with importance degrees 'maximum', 'mean' and 'product of maximum and mean' (WM1, WM2 and WM3, respectively) and the same three versions of Cordón and Herrera's method [11] (CH1, CH2, CH3). Nozaki, Ishibuchi and Tanaka's fuzzy rule learning [44] (NIT), TSK rules [54] optimized with Weighted Least Squares, and Iterative Rule Learning [52]. A second GFS based on vague data, NMIC [53] has also been included in the comparison.

The best overall result and the most accurate fuzzy rule base are boldfaced in Table 3, which contains the median of the test error. We have also included a selection of boxplots in Figure 10. These provide a graphical insight of the relevance of the differences. Observe that the performances of the heuristic algorithms are always worse than those of the GFS.

These results show that there are not, in general, significant differences in accuracy between GFS, statistical nonlinear regression and neural networks. IVFF is not worse than standard GFSs in crisp data but, in contrast with the results of the preceding section, the advantages over IRL are not statistically sound (95% level).

6.2.3 Performance in machine learning datasets

The fourth benchmark uses standard Machine Learning benchmarks to assess the accuracy but also the compactness of the IVFF algorithm, in problems with different sizes and number of inputs. In order to keep the discussion simple, we have chosen a selection of the algorithms in the preceding section. CH1, CH2 and CH3 were discarded because they produced a much higher number of rules than WM,



f_1

$f_1 - 10$

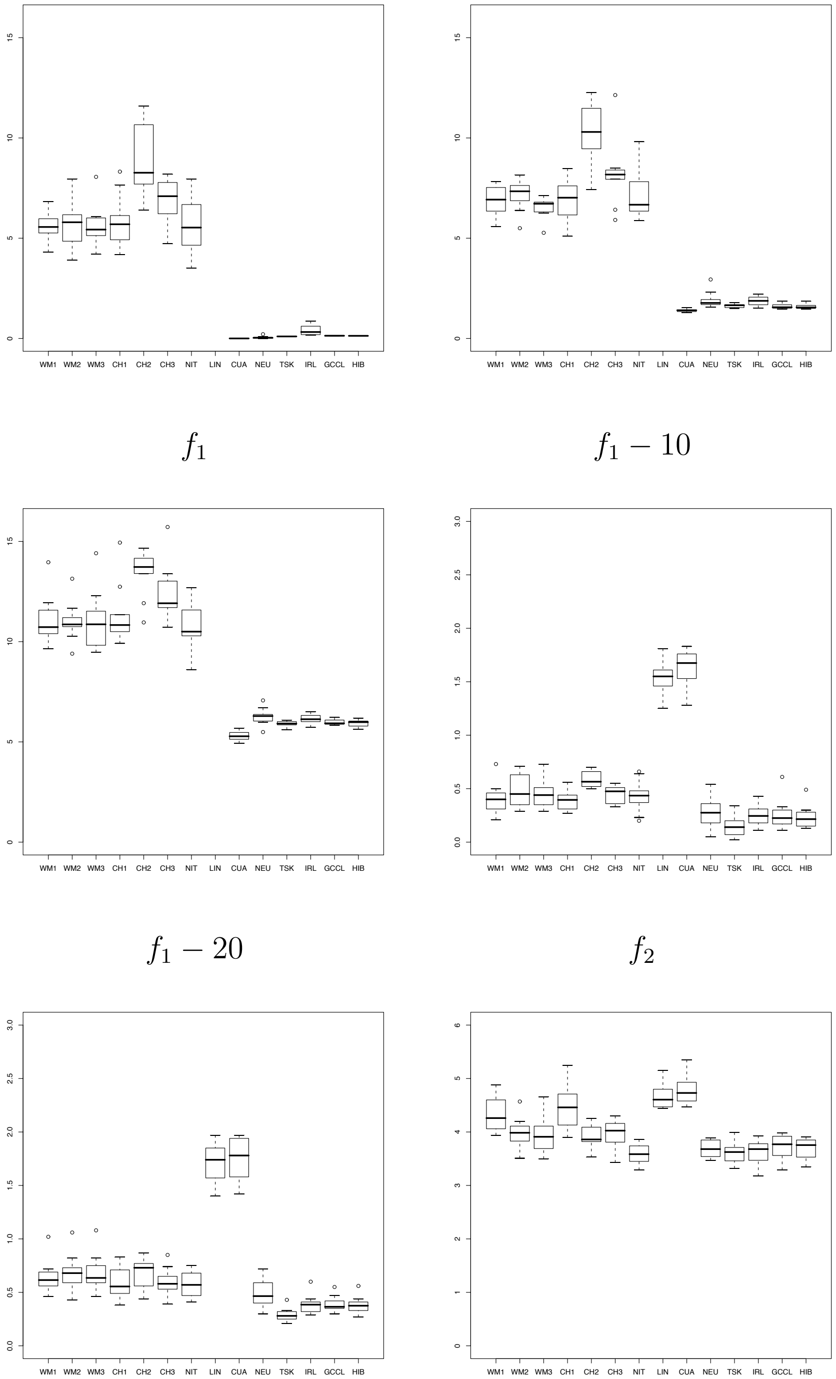
f_2

$f_2 - 10$

Fig. 9. Statistical relevance of the differences between IVFF and IRL, in the datasets f_1 and f_2. The boxplots show the dispersion of the results in Table 2. Left: no stochastic noise added. Right: 10% of stochastic noise. The labels of the boxplots show the amount of imprecision in the observation of the output variable.

	WM1	WM2	WM3	CH1	CH2	CH3	NIT	LIN	QUA	NEU	TSK	IRL	NMIC	IVFF
f_1	5.56	5.80	5.43	5.70	8.27	7.1	5.53	132.1	**0.00**	0.03	0.10	0.32	0.14	**0.13**
f_1-10	6.92	7.34	6.72	7.02	10.30	8.18	6.67	135.37	**1.41**	1.77	1.66	1.88	1.56	**1.55**
f_1-20	10.72	10.86	10.86	10.83	13.72	11.91	10.50	134.04	**5.28**	6.29	5.91	6.13	**5.92**	5.98
f_1-50	52.63	46.27	48.08	52.96	48.77	49.09	40.09	167.98	**33.39**	39.89	36.69	**38.94**	39.26	39.14
f_2	0.40	0.45	0.44	0.39	0.56	0.47	0.43	1.55	1.67	0.27	**0.14**	0.24	0.22	**0.21**
f_2-10	0.61	0.68	0.63	0.55	0.73	0.58	0.57	1.74	1.78	0.46	**0.28**	0.38	**0.36**	0.37
f_2-20	1.34	1.21	1.19	1.31	1.21	1.19	0.98	2.03	2.18	0.86	**0.73**	0.85	0.86	**0.84**
f_2-50	4.26	3.98	3.91	4.46	3.86	4.02	**3.58**	4.60	4.73	3.68	3.62	3.68	3.77	3.75

Table 3. Median of the test error of IVFF and a selection of algorithms for the synthetic datasets f_1 and f_2, after 10 repetitions. The algorithms include: heuristics (WM, CH), statistical regression (LIN, QUA), neural networks (NEU), rules with a real-valued consequent (NIT), TSK rules (TSK), iterative rule learning (IRL) and other vague data-based GFS (NMIC). IRL, NMIC and IVFF runs were limited to 25 fuzzy rules. The best of WM, CH, NIT, IRL, NMIC and IVFF, plus the best overall model, were highlighted for every dataset.



f_1

$f_1 - 10$

$f_1 - 20$

f_2

$f_2 - 10$

$f_2 - 20$

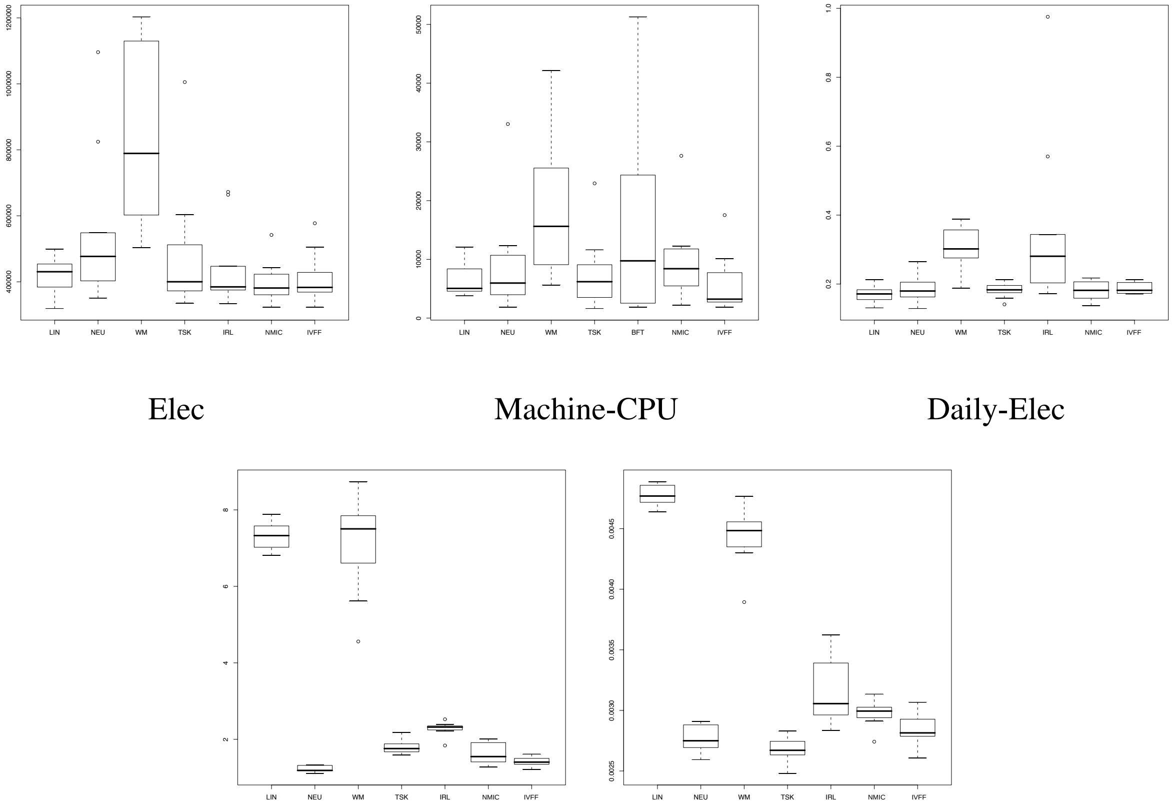
Fig. 10. Statistical relevance of the differences between IVFF and the selection of statistical and fuzzy rule-based regression algorithms mentioned in the text, in the datasets f_1 and f_2. The performance of the heuristic algorithms is worse than that of the GFS. In turn, there are not, in general, significant differences in accuracy between GFS, statistical nonlinear regression and neural networks.

	Statistical		TSK rules	Fuzzy Rule Learning			
	LIN	NEU	WLS	WM	IRL	NMIC	IVFF
elec $\cdot 10^{-3}$	431	477	400	789	385	**381**	383
machine-CPU	5057	5963	6196	15619	9748	8408	**3226**
daily-elec	0.171	0.180	0.183	0.302	0.281	**0.182**	**0.182**
Friedman	7.33	1.19	1.76	7.50	2.32	1.55	**1.41**
building $\cdot 10^{3}$	4.77	2.75	2.67	4.48	3.06	2.99	**2.81**

Table 4
Performance (MSE of test error) of a selection of statistical models, heuristic rule learning and GFS in some benchmarks. The two last algorithms, NMIC and IVFF, have been trained with fuzzy datasets. These have been obtained by augmenting the crisp output variable into a triangular fuzzy set, with small support, centered in the desired output value.

		TSK rules	Fuzzy Rule Learning		
	Labels	WLS	WM	IRL	IVFF
elec	3	8	7	10	**4**
machine-CPU	3	91	20	25	**4**
daily-elec	3	427	64	25	**5**
Friedman	3	242	192	25	**10**
building	3/2*	896*	789	30	**20**

Table 5
IVFF obtains similar accuracy than state-of-the-art GFSs, and at the same time allows using a smaller knowledge base. The table displays the number of rules that different algorithms produced, for the problems shown in the first column and the number of linguistic labels by variable shown in the second. In particular, IVFF improves grid-based learning algorithms by one order of magnitude in the number of parameters. We have used three linguistic labels in all the continuous variables. (*) In the dataset "building", the grid-based algorithm depends on partitions of size 2. Otherwise, the number of rules is too high for practical purposes.



Elec

Machine-CPU

Daily-Elec

Friedman

Building

Fig. 11. Compared accuracy (dispersion of test MSE) between Linear Regression, Neural Networks, Wang and Mendel, TSK rules, Iterative Rule Learning, NMIC and IVFF. The use of a fuzzy, augmented dataset, makes the accuracy of IVFF to be better, in general, than that of IRL.

without a noticeable increment of the accuracy. Also, since we are measuring the linguistic quality by the number of rules, without taking into account the complexities of the consequent part of the rules, we discard NIT, because it is less accurate than weighted linear squares-based TSK, for the same number of fuzzy rules. The quadratic regression is also removed, because most of times the corresponding polynomial has more terms than the multilayer perceptron needed for obtaining an equivalent accuracy.

Summarizing, linear regression and the neural network are left for reference purposes. Other than this, we have compared: one example-guided heuristic method (Wang and Mendel algorithm), one TSK grid-based algorithm (weighted least squares fitting of a plane to each set of elements covered by one cell of the fuzzy grid; the weights are the memberships of the elements to the cell), an IRL method (genetic backfitting), and two vague data-based GFSs, NMIC and IVFF. The crisp output variables in the training sets of these two last algorithms were transformed into triangular fuzzy sets, centered in the crisp value, and with a support of 1% of the range of the corresponding feature. The results are given in Tables 4, 5, and Figure 11.

As expected, the example-guided heuristic uses a moderately high number of rules, and it is not very precise (equivalent or worse than linear regression). The weighted linear squares obtention of TSK rules is the most precise method, similar to that of the neural network. Backfitting and IVFF both perform well and offer a good compromise between compacteness and accuracy. IVFF was significantly better than backfitting in four of five tests, where the fuzzy knowledge bases it generated were both more compact and more precise.

6.3 Practical application: Causal Modeling in Marketing

In this section we will apply all the concepts defined in the preceding sections of the paper to a practical problem of causal modeling in marketing. In this problem, both input and output variables are composed of sets of parameters (items). Each item provides only partial information to describe the variable. The information provided by different items may be in conflict.

The conversion of a set of items into a compound value that can be fed to the model has been solved in different ways. The classical solution [7] consists in preprocessing these sets of values, then replacing each one of them by a characteristic value. This solution might not be the most suitable, because the model should know not only about the characteristic values of the variable, but also about the degree of imprecision with which these values are known. In the next section we will show how to model each set of items by means of a fuzzy number, both for the input and the output variables.

The learning problem that arises involves optimizing a combination of fuzzy and crisp functions, as mentioned in Section 5. The accuracy of the model is measured by the FMSE between the fuzzy representation of the output variables, and the images of the input values provided by the candidate model, which are also fuzzy [49]. Its complexity depends, as we will show, on the number of rules, and the number of linguistic terms each rule uses. As done in previous works, we will jointly optimize the precision of the model and its complexity, by means of multicriteria genetic algorithms. The purpose of the learning is to find a model both accurate and linguistically understandable.

6.3.1 Representing multi-item data

Data were obtained by means of a questionnaire. We have considered the measurement model depicted in Figure 12, where the latent variables are depicted by circles. Since these latent variables are unmeasurable, we indirectly measure them by means of observable variables (items), depicted by rectangles in the figure.

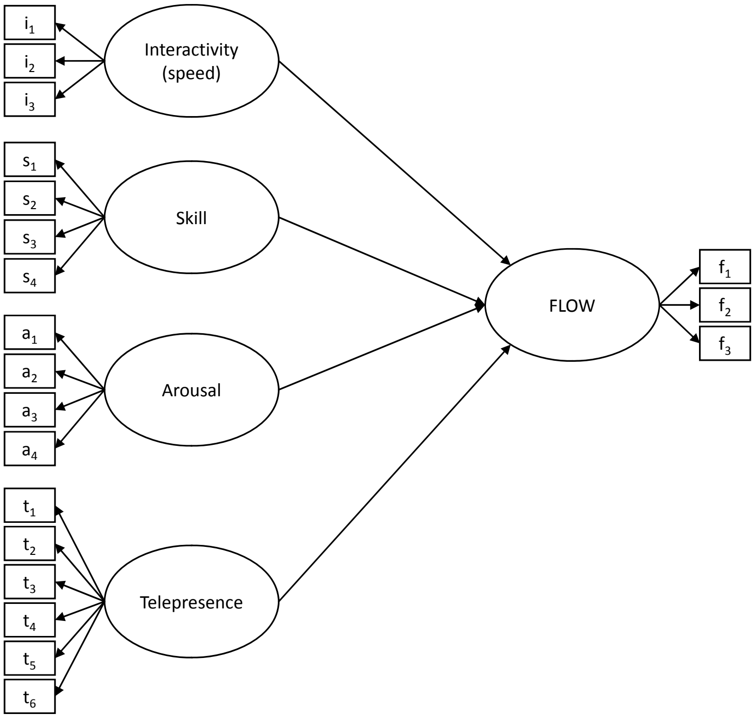


Fig. 12. Example of a simple measurement (structural) model (extracted from [38])

Likewise, with regard to the measurement scales, the constructs have been measured by means of several nine-points Likert scales ranging from 1: strongly disagree to 9: strongly agree. Specifically, in Table 6 we show a hypothetical example of the set of items that could have been used for measuring each one, while Table 7 shows an example of data available for this problem. It can be seen that the input and output data comprises multi-item values.

Interactivity (speed)	
i_1 :	When I use the Web there is very little waiting time between my actions and the computer’s response.
i_2 :	Interacting with the Web is slow and tedious. (R)
i_3 :	Pages on the Web sites I visit usually load quickly.
Telepresence	
t_1 :	Using the Web often makes me forget where I am.
t_2 :	After using the Web, I feel like I come back to the ”real world” after a journey.
t_3 :	Using the Web creates a new world for me, and this world suddenly disappears when I stop browsing.
t_4 :	When I use the Web, I feel I am in a world created by the Web sites I visit.
t_5 :	When I use the Web, my body is in the room, but my mind is inside the world created by the Web sites I visit.
t_6 :	When I use the Web, the world generated by the sites I visit is more real for me than the “real world.”
Arousal	
a_1 :	stimulated / relaxed
a_2 :	calm/excited (R)
a_3 :	frenzied / sluggish
a_4 :	unaroused/aroused (R)
Skill	
s_1 :	I am extremely skilled at using the Web.
s_2 :	I consider myself knowledgeable about good search techniques on the Web.
s_3 :	I know somewhat less about using the Web than most users. (R)
s_4 :	I know how to find what I am looking for on the Web.
Flow	
f_1 :	Do you think you have ever experienced “flow” on the Web?
f_2 :	In general, how frequently would you say you have experienced “flow” when you use the Web?
f_3 :	Most of the time I use the Web I feel that I am in “flow.”

Table 6
Questionnaire associated to the measurement model shown in Figure 12 (extracted from [45]). In a real questionnaire, the names of the latent variables are hidden to the consumer and the order of the questions is usually changed. The text (R), which indicates the item was reverse-scaled, is also omitted.

Interactivity			**Telepresence**						**Arousal**				**Skill**				**Flow**		
(speed)																			
2	3	3	7	7	9	7	2	3	3	3	3	2	2	3	2	1	7	6	3
8	4	3	8	8	8	8	7	6	5	5	3	4	2	2	3	2	5	5	5
4	6	6	5	5	5	7	6	5	6	6	7	6	7	6	6	4	6	6	5
4	6	7	7	3	7	5	5	3	5	2	2	3	1	2	2	3	3	2	2
5	5	5	5	5	5	5	2	2	5	5	5	5	5	5	5	5	5	5	5

Table 7
Example of the obtained data set that contains five consumers' responses about the items shown in Table 6. The values of the reverse-scaled items are already reversed

In previous works, different aggregation functions were proposed [8]. These functions ultimately lead to the assignment of a certain truth value to the assertion "the value of the variable is V" for certain label V of a suitable linguistic variable. Here we will be taking a different path. We assume that there exists a true value for the multi-item variable, but also that this value is unknown and we can know, at best, a set that contains it. Then, as we have explained in Section 2, we will find a fuzzy set such that their α-cuts are confidence intervals with degree $1 - \alpha$ of the expected value of the observation error. Each list of items will be replaced by its corresponding fuzzy set. This interpretation allows us to aggregate data, losing less information than central tendency measures.

6.3.2 Fitness function

We have used a fitness function with two components: the FMSE and a measure of the complexity of the knowledge base. This second objective intends to assess the linguistic complexity of the generated fuzzy rule set. Firstly, it is clear that the higher number of rules, the higher the complexity. Therefore, we measure the number of rules of the FRBS as $C_1(\mathcal{F})$. However, since each DNF-type fuzzy rule also has a degree of complexity itself, we should also consider this aspect. Then, let $C_2(\mathcal{F}) = \sum_{R_r \in \mathcal{F}} \prod_{i=1}^{n} l_{ri}$ be the complexity of the FRBS, with l_{ri} being the number of linguistic terms used in the ith input variable of the rth DNF-type fuzzy rule. The total number of available linguistic terms is computed when an input variable is not considered (i.e. "don't care"). We have decided that the joint objective is the product of both complexities.

6.3.3 Experiments

Data have been obtained from the survey used in [43] to test a conceptual model previously presented by the same authors. We have adapted the original structural model by removing the least significant latent variable in each second-order variable. According to the partition performed by the authors, training data is composed of 1,154 examples (consumers' responses) and test data of 500 examples. As an example, we focus the analysis on a specific relationship among the six relationships with a total of 12 variables available in the data set. Four constructs were used as input variables of the systems.

We have evaluated three configurations of the fuzzy fitness-based genetic fuzzy system, that have been validated over the data used to learn the model in [8]. This last model does not transform the multi-item data into a fuzzy set, but it takes into account the uncertainty in the data by means of an extension of the membership degree computation, the called multi-item fuzzification, which is based on a union of the partial information provided by each item.

Each configuration has been run 10 times. The resulting joint Pareto-fronts of the models are shown in Figure 13. Three types of precedence operators have been evaluated along with our own definitions of dominated sorting and crowding distances: the strong dominance [36], the uniform prior [55] and the imprecise prior defined in this paper. The fuzzy fitness-based algorithms have been trained over fuzzy data obtained with the bootstrap approximation mentioned in Section 6.3.1.

The fuzzy data-based algorithm produced better results for all the configurations. In particular, the use of a precedence based on an imprecise prior produces a better balance between complexity and precision, with a higher density of solutions in the center of the Pareto front. In Figure 14 a detail of the Pareto fronts of both the crisp and the imprecise prior-based approach, where it is clear that most of the rule bases found by the standard method are being dominated by this. In this figure the FMSE is also shown for the fuzzy model.

Average train and test errors of the individuals in the population during the learning phase are shown in Figure 15. Observe that, as expected, the use of fuzzy data improves the generalization error. The difference between the validation error in the train and test datasets is smaller, beginning in the initial generations. It is remarked that there is an overhead associated to the use of the reasoning method defined in Section 3, and the calculus of the FMSE. The algorithm [8] is between four and ten times faster than this.

In Figure 16, the average of the differences between the output of these models and all the items in every output variable is computed, and the best, worst and mean test error in the ten repetitions are plotted for every generation. Observe that the maximum value of test error in the fuzzy fitness is always better than the minimum value of the scalar fitness. The boxplot in Figure 16 also illustrates the dispersion of the test error in the ten repetitions.

In Figure 17, the comparison focuses on the three types of precedence operators evaluated in this paper. The differences between them emerge in the latter gener-

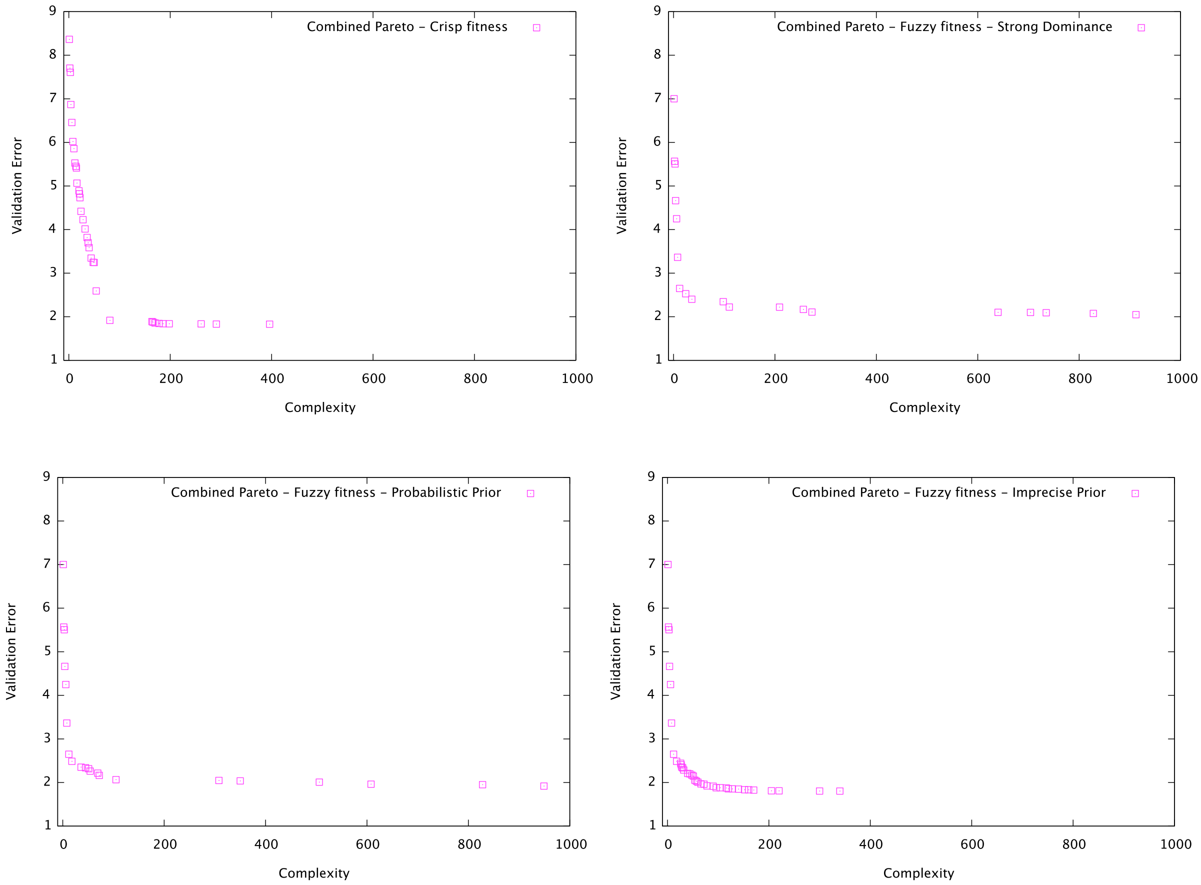


Fig. 13. Combined Pareto fronts in 10 repetitions of the GFS proposed in this paper and that proposed in [8]. Three different precedence operators have been evaluated in combination with our own definitions of nondominated sorting and crowding distances: the strong dominance [36], the probabilistic prior [55] and the imprecise prior defined in this paper.

ations, where the inability of the strong dominance to distinguish between overlapping FMSEs blocks further convergence after the 250th generation. The use of the uniform prior produced better average results in the first 200 generations, however the imprecise prior allows the evolution to continue past that 250th generation. Observe that, contrasting with the behavior of the uniform prior, the use of an imprecise prior causes that very similar FMSEs are indistinguishable. This reduces the overfitting, as was shown in the Pareto fronts in Figure 13.

7 Concluding remarks

In this paper we have given a comprehensive description of the use of certain kinds of ill-defined data in Genetic Fuzzy Systems. A new model of FRV has been used to justify our representation of vague data in terms of fuzzy sets. The definition of variance of an FRV, according to this model, has been adapted, so we could

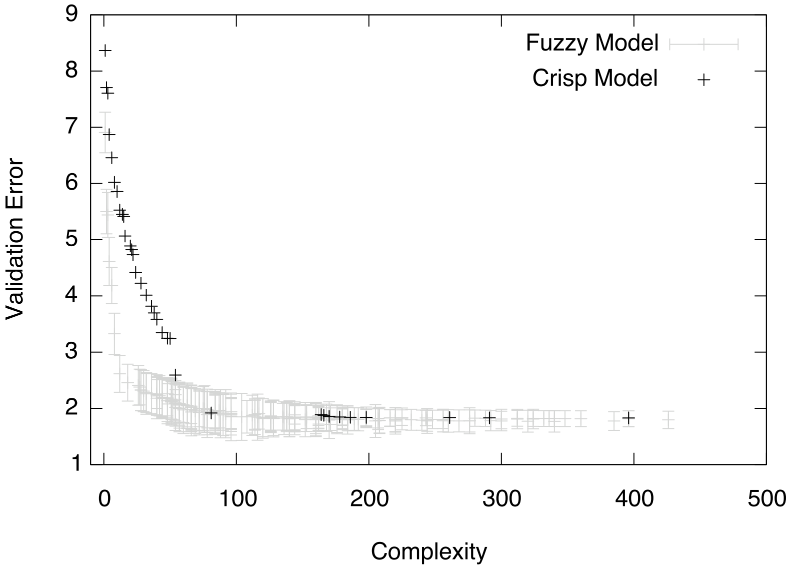


Fig. 14. FMSE and validation errors of the crisp and fuzzy models (precedence based on the imprecise prior).

bound the mean squared error of a rule base on a fuzzy dataset. Lastly, different precedence operators between values of the fitness function have been defined. All of these elements have been combined in an extension of the NSGA-II algorithm, that is able to optimize a combination of crisp and interval-valued objectives, and hence to learn fuzzy rules with balanced accuracy and complexity.

To assess our algorithm, we have benchmarked it with crisp, interval and fuzzy data, and also solved a marketing problem where the input data comprised multi-item examples. These multi-item examples were promoted to fuzzy sets by means of the mentioned interpretation of the membership function as a nested family of confidence intervals. We have shown, with the help of the experimentation, that the models obtained by minimizing the FMSE are more robust than standard genetic fuzzy models, and are able to capture the dependence between imprecise data without the need of aggregating them or removing their fuzziness.

References

[1] J. Alcalá et al. KEEL: A software tool to assess evolutionary algorithms for data mining problems. *Soft Computing*, 2008. In press.

[2] C. Baudrit, I. Couso, and D. Dubois. Joint propagation of probability and possibility in risk analysis: Towards a formal framework. *International Journal of Approximate Reasoning*, 45:82–105, 2007.

[3] C. Baudrit and D. Dubois. Practical representation of incomplete probabilistic information. *Comput. Statist. Data Anal.*, 51(1):86–108, 2006.

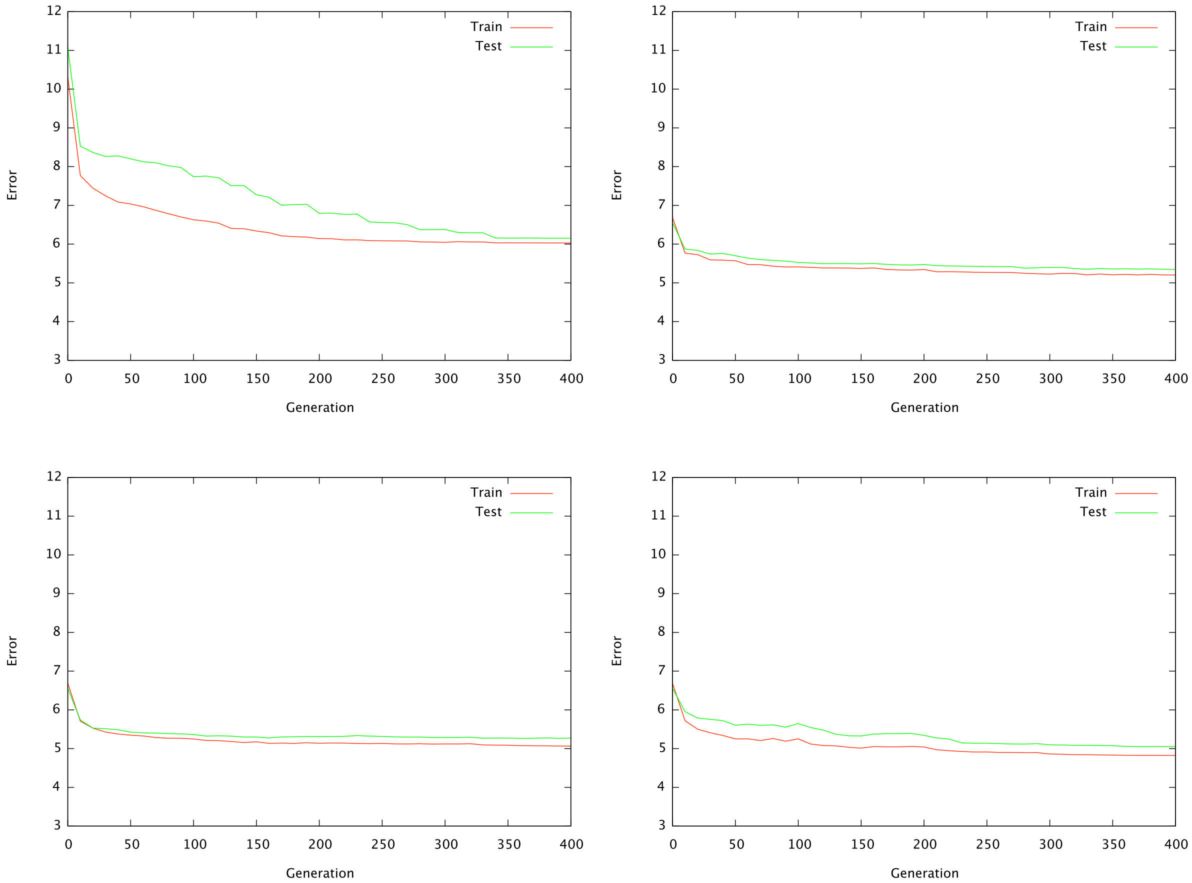


Fig. 15. Average train and test errors of the individuals in the population during the learning phase. Upper part, left: Crisp fitness. Right: Fuzzy fitness, strong dominance. Lower part, left: Fuzzy fitness, probabilistic prior. Right: Fuzzy fitness, imprecise prior. The differences between the crisp and the fuzzy versions are significant. The differences between the different precedences in the fuzzy fitness are less relevant. Best results were obtained with the imprecise prior.

[4] C. Baudrit, D. Dubois, and N. Perror. Representing parametric probabilistic models tainted with imprecision. *Fuzzy Sets and Systems*, 15(1):1913–1928, 2008.

[5] C. Bertoluzza, N. Corral, and A. Salas. On a new class of distances between fuzzy numbers. *Mathware and Soft Computing*, 2:71–84, 1995.

[6] J. Casillas, O. Cordón, F. Herrera, and L. Magdalena, editors. *Interpretability issues in fuzzy modeling*. Springer, Heidelberg, Germany, 2003.

[7] J. Casillas, F. Martinez-Lopez, and F. Martinez. Fuzzy association rules for estimating consumer behaviour models and their application to explaining trust in internet shopping. *Fuzzy Economic Review*, IX(2):3–26, 2004.

[8] J. Casillas and F. Martínez-López. Mining uncertain data with multiobjective genetic fuzzy systems to be applied in consumer behaviour modelling. *Expert Systems with Applications*, 36(2):1645–1659, 2009.

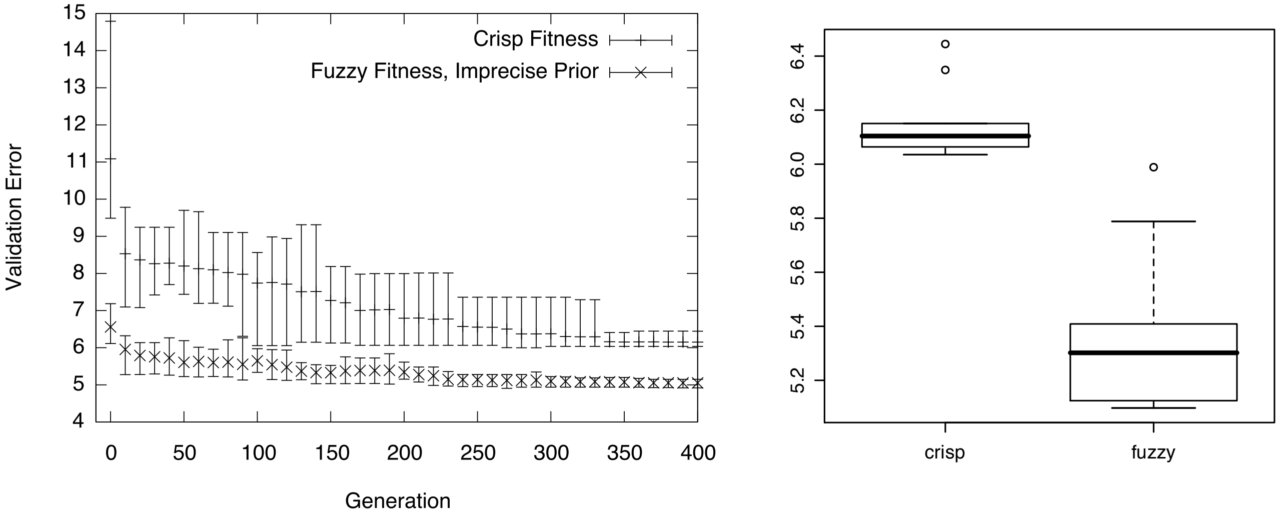


Fig. 16. Left: Best, worst and mean differences between the output of both crisp and fuzzy models, and all the items in every output variable. Right: Boxplot showing the differences between the item-wise errors of the crisp and fuzzy models after 400 generations and 10 runs of the experiments.

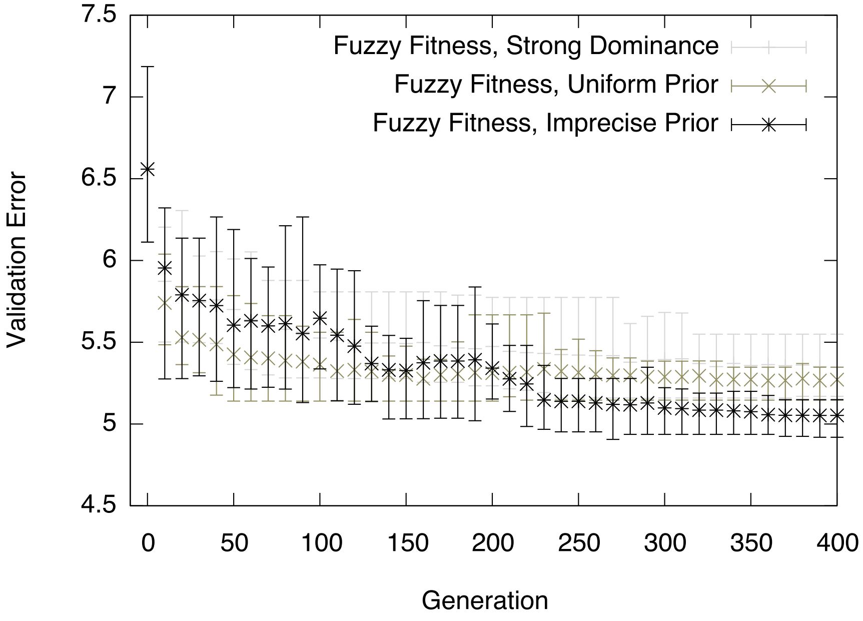


Fig. 17. Comparison between the three type of precedence operators evaluated in this paper. The differences between them appear in the latter generations.

[9] J. Castro and M. Delgado. Fuzzy systems with defuzzification are universal approximators. *IEEE Trans. on Syst., Man, and Cybern*, 26(1):149–152, 1996.

[10] O. Cordón, Herrera, F. Hoffmann, and L. Magdalena. *Genetic fuzzy systems: Evolutionary tuning and learning of fuzzy knowledge bases.* World Scientific Publishing Company, Singapore, 2001.

[11] O. Cordón and F. Herrera. A proposal for improving the accuracy of linguistic modeling. *IEEE Transactions on Fuzzy Systems*, 8(3):335–344, 2000.

[12] O. Cordón, F. Herrera, and L. Sánchez. Solving electrical distribution problems using hybrid evolutionary data analysis techniques. *Applied Intelligence*, 10(1):5–24, 1999.

[13] I. Couso and D. Dubois. On the variability of the concept of variance for fuzzy random variables. *IEEE Trans. Fuzzy Sets and Systems, submitted.*

[14] I. Couso, D. Dubois, S. Montes, and L. Sánchez. On various definitions of the variance of a fuzzy random variable. In *Proc. of Fifth International Symposium on Imprecise Probabilities: Theory and Applications (ISIPTA 07)*, pages 135–144, 2007.

[15] I. Couso, S. Montes, and P. Gil. The necessity of the strong alpha-cuts of a fuzzy set. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(2):249–262, 2001.

[16] I. Couso and L. Sánchez. Higher order models for fuzzy random variables. *Fuzzy Sets and Systems*, 159:237–258, 2008.

[17] K. Deb, A. Pratap, S. Agarwal, and T. Meyarevian. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002.

[18] P. Diamond and P. Kloeden. *Metric Spaces of Fuzzy Sets*. World Scientific, Singapore, 1994.

[19] D. Dubois, H. Fargier, and J. Fortin. A generalized vertex method for computing with fuzzy intervals . In *Proc. of the International Conference on Fuzzy Systems, Budapest, Hungary*, pages 541–546. IEEE, 2004.

[20] D. Dubois, H. Fargier, and J. Fortin. The empirical variance of a set of fuzzy intervals. In *Proc. of the 2005 IEEE International Conference on Fuzzy Systems, Reno, Nevada*, pages 885–890. IEEE, 2005.

[21] D. Dubois and H. Prade. The three semantics of fuzzy sets. *Fuzzy Sets and Systems*, 90:141–150, 1997.

[22] P. Ein-Dor and J. Feldmesser. Attributes of the performance of central processing units: A relative performance prediction model. *Communications of the ACM*, 30(4):308–317, 1987.

[23] S. Ferson, V. Kreinovich, J. Hajagos, W. Oberkampf, and L. Ginzburg. Experimental uncertainty estimation and statistics for data having interval uncertainty. Technical report, SAND2007-0939, Sandia National Laboratories, 2007.

[24] J. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19:1–141, 1991.

[25] J. Gómez and E. León. A fuzzy sets rule distance for evolving fuzzy anomaly detectors. In *IEEE World Congress on Computational Intelligence*, pages 2286–2292, 2006.

[26] A. Gonzalez and R. Perez. Completeness and consistency conditions for learning fuzzy rules. *Fuzzy Sets and Systems*, 96(1):37–51, 1998.

[27] I. R. Goodman. Fuzzy sets as equivalence classes of possibility random sets. In R. R. Yager, editor, *Fuzzy Sets and Possibility Theory: Recent Developments*, pages 327–343. Pergamon, Oxford, 1982.

[28] I. R. Goodman and H. T. Nguyen. *Uncertainty models for knowledge-based systems*. Elsevier, 1985.

[29] F. Herrera. Genetic fuzzy systems: Taxonomy and current research trends and prospects. *Evolutionary Intelligence*, 1:27–46, 2008.

[30] A. Irpino and R. Verde. Dynamic clustering of interval data using a wasserstein-based distance. *Pattern Recognition Letters*, 29(11):1648–1658, 2008.

[31] E. Klement, M. Puri, and D. Ralescu. Limit theorems for fuzzy random variables. *Proc. Roy. Soc. London A*, 407:171–182, 1986.

[32] V. Krätschmer. A unified approach to fuzzy random variables. *Fuzzy Sets and Systems*, 123:1–9, 2001.

[33] R. Kruse and K. Meyer. *Statistics with vague data*. D. Reidel Publishing Company, 1987.

[34] H. Kwakernaak. Fuzzy random variables. definition and theorems. *Inform. Sci.*, 15:1–29, 1989.

[35] R. Körner. On the variance of fuzzy random variables. *Fuzzy Sets and Systems*, 92:83–93, 1997.

[36] P. Limbourg. Multiobjective optimization of problems with epistemic uncertainty. In *Proc. Third International Conference on Evolutionary Multi-Criterion Optimization*, pages 413–427, 2005.

[37] A. T. M. Öztürk. Valued hesitation in intervals comparison. In *Proceedings of the SUM-07 conference, LNAI 4772*, pages 157–170. Springer-Verlag, 2007.

[38] S. MacLean and K. Gray. Structural equation modelling in market research. *Journal of the Australian Market Research Society*, 6:17–32, 1998.

[39] E. Marín and L. Sánchez. Supply estimation using coevolutionary genetic algorithms in the spanish electrical market. *Applied Intelligence*, 21(1):7–24, 2004.

[40] G. Mauris. Inferring a possibility distribution from very few measurements. In *Soft Methods for Handling Variability and Imprecision. (eds: D. Dubois, M. A. Lubiano, H. Prade, M. A. Gil; P. Grzegorzewski, O. Hyrniewicz) Springer, Heidelberg*, pages 92–99, 2008.

[41] G. Mauris, V. Lasserre, and L. Foulloy. Fuzzy modeling of measurement data acquired from physical sensors. *IEEE Trans. Instrum. Meas.*, 49(6):1201–1205, 2000.

[42] E. Miranda, G. de Cooman, and I. Couso. Imprecise probabilities induced by multi-valued mappings. *J. Stat. Plann. Inference.*, 133:173–197, 2005.

[43] Y. Novak, D. Hoffman, and Y. Yung. Measuring the customer experience in online environments: a structural modelling approach. *Marketing Science*, 19(1):22–42, 2000.

[44] K. Nozaki, H. Ishibuchi, and H. Tanaka. A simple but powerful heuristic method for generating fuzzy rules from numerical data. *Fuzzy Sets and Systems*, 86:251–270, 1997.

[45] L. Prechelt. Proben1 - a set of benchmarks and benchmarking rules for neural network training algorithms. tech. rep. 21/94. Technical report, Fakultat fur Informatik, Universitat Karlsruhe, 1994.

[46] M. Puri and D. Ralescu. Fuzzy random variables. *J. Math. Anal. Appl.*, 114:409–422, 1986.

[47] S. P. Riyaz Sikora. Efficient genetic algorithm based data mining using feature selection with hausdorff distance. *Information Technology and Management*, 6(5):315–331, 2005.

[48] L. Sánchez, J. Casillas, and O. Cordón. Some relationships between fuzzy and random set-based classifiers and models. *International Journal of Approximate Reasoning*, 29(2):175–213, 2002.

[49] L. Sánchez and I. Couso. Advocating the use of imprecisely observed data in genetic fuzzy systems. *IEEE Transactions on Fuzzy Systems*, 15:551–562, 2007.

[50] L. Sánchez, I. Couso, and J. Casillas. A multiobjective genetic fuzzy system with imprecise probability fitness for vague data. In *Proc. of the 2006 IEEE International Conference on Evolutionary Fuzzy Systems, Ambleside, UK*, pages 131–137, 2006.

[51] L. Sánchez, I. Couso, and J. Casillas. Modeling vague data with genetic fuzzy systems under a combination of crisp and imprecise criteria. In *Proc. of the 2007 IEEE Sysmposium on Computational Intelligence in Multicriteria Decision Making, Honolulu, USA*, pages 30–37, 2007.

[52] L. Sánchez and J. Otero. A fast genetic method for inducting descriptive fuzzy models. *Fuzzy Sets and Systems*, 141(1):33–46, 2004.

[53] L. Sánchez and J. Otero. Learning fuzzy linguistic models from low quality data by genetic algorithms. In *Fuzzy Systems Conference, 2007. FUZZ-IEEE 2007. IEEE International*, pages 1921–1926, 2007.

[54] T. Takagi and M. Sugeno. Fuzzy identification of systems and its application to modeling and control. *IEEE Transactions on System, Man and Cybernetics*, 15(1):116–132, 1985.

[55] J. Teich. Pareto-front exploration with uncertain objectives. In *Proc. First International Conference on Evolutionary Multi-Criterion Optimization*, pages 314–328, 2001.

[56] P. Walley. *Statistical reasoning with imprecise probabilities*. Chapman and Hall, London, 1991.

[57] L. X. Wang and J. Mendel. Generating fuzzy rules by learning from examples. *IEEE Trans. on Systems, Man and Cybernetics*, 25(2):353–361, 1992.

[58] L. Zadeh. Fuzzy sets as a basis for the theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.