Advocating the Use of Imprecisely Observed Data in Genetic Fuzzy Systems

Luciano Sánchez, *Member, IEEE*, and Inés Couso

***Abstract*—In our opinion, and in accordance with current literature, the precise contribution of genetic fuzzy systems to the corpus of the machine learning theory has not been clearly stated yet. In particular, we question the existence of a set of problems for which the use of fuzzy rules, in combination with genetic algorithms, produces more robust models, or classifiers that are inherently better than those arising from the Bayesian point of view.**

We will show that this set of problems actually exists, and comprises interval and fuzzy valued datasets, but it is not being exploited. Current genetic fuzzy classifiers deal with crisp classification problems, where the role of fuzzy sets is reduced to give a parametric definition of a set of discriminant functions, with a convenient linguistic interpretation. Provided that the customary use of fuzzy sets in statistics is vague data, we propose to test genetic fuzzy classifiers over imprecisely measured data and design experiments well suited to these problems. The same can be said about genetic fuzzy models: the use of a scalar fitness function assumes crisp data, where fuzzy models, *a priori*, do not have advantages over statistical regression.

***Index Terms*—Fuzzy fitness function, fuzzy rule-based classifiers, fuzzy rule-based models, genetic fuzzy systems, vague data.**

I. Introduction

STATISTICS and machine learning are closely intertwined. The ambit of application of statistics ranges from experimental designs [32] to theoretical studies about the generalization properties of algorithms, or computational learning theory [24]. Genetic fuzzy systems (GFSs) are likewise influenced by this trend, and statistical tests are a standard tool when the performances of genetic classifiers or models are compared [8].

Apart from the machine learning field, fuzzy statistics [3] is an active research area, and there have been advances in the fuzzy counterparts of most of the aforementioned techniques. But, as far as we know, there are scarce connections between fuzzy statistics and GFSs. Contrary to this, we think that the nature of GFSs makes the introduction of some elements of fuzzy statistics desirable. In this paper we will study the changes in the definitions of the fitness function that are needed when interval and fuzzy extensions of either the classification or the regression problems are introduced and discuss the impact of these extensions on the design of new genetic fuzzy learning algorithms, and over the experimental design of GFSs.

Manuscript received November 6, 2005; revised May 21, 2006 and June 1, 2006. This work was supported in part by the Spanish Ministry of Education and Science under Grants TIC2002-04036-C05-05, TIN2005-08036-C05-5, and MTM2004-01269.

L. Sánchez is with the Computer Science Department, University of Oviedo, Oviedo, Spain (e-mail: luciano@uniovi.es).

I. Couso is with the Statistics Department, University of Oviedo, Oviedo, Spain (e-mail: couso@uniovi.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TFUZZ.2007.895942

This work is structured as follows. In Sections II and III, the statistical definitions of classification and modeling under stochastic noise are introduced, and it is explained that current GFSs are designed to solve the stochastic, crisp problem and not the fuzzy one. In Sections IV and V, extensions to these definitions are introduced. These extensions deal with both stochastic noise and observation errors and are based upon the relations between random and fuzzy sets. In Section VI, it is shown, by means of examples, how the fuzzy valued fitness functions of the respective extended GFS are evaluated. In Section VII, the results of the applications of these concepts to practical problems are reviewed. This paper finishes with the concluding remarks and a discussion about the newly opened research lines.

II. Statistical Classifiers and GFSs

A. Classification Systems Based on Discriminant Functions

Let us suppose we have a set Ω that contains objects ω, and let us admit also that each one of them is assigned to a class C_i, $i = 1 \dots N_c$. We are given a set of measurements $X(\omega) = (X_1(\omega), \dots, X_K(\omega))$ over every object. We will say that a classification system is a decision rule that maps every element of $X(\Omega)$ to a class C_i, whose main objective is to produce a low number of errors.

For example, let Ω be a set of fruits: apples (C_1), pears (C_2), or bananas (C_3). We observe the weight and the color of a randomly selected fruit, for example, $X(\omega) = (\texttt{yellow}, 150)$. Our classification system relates the pair (yellow, 150) to the class C_3, and we wish this relation to be true on most occasions (i.e., most of the yellow fruits that weigh 150 g are bananas).

Since we do not assume that $\omega_1 \neq \omega_2 \Rightarrow X(\omega_1) \neq X(\omega_2)$ (i.e., we admit that there can exist a yellow pear weighing 150 g) perhaps a decision rule that never fails cannot be defined for this problem. But an optimum classifier can be defined with respect to the average number of errors.

To define the concept "average number of errors," we need to assume that the mapping X fulfills all the necessary conditions to be a random variable. Let us also define a new random variable that quantifies the cost of assigning the class C_i to an object when it belongs to class C_j, $\text{cost}(i, j)$. If we choose $\text{cost}(i, j) = 1$ when $i \neq j$ and zero else, the expectation of the cost function is the mean number of errors. This rule is called "minimum error Bayes rule."

For the problem stated, if the classifier is a decision rule $D(X)$ and $\text{class}(\omega)$ is the class of the object ω, then the merit value of a classifier can be numerically quantified as

$$\text{err}(D) = \int_{\Omega} \text{cost}(D(X(\omega)), \text{class}(\omega))\, \text{dP} \qquad (1)$$

where the error function is integrated with respect to a probability measure P defined over Ω. It is well known that this error is optimized by a classifier defined as follows [23]:

$$D(\mathbf{x}) = \arg \max_{i=1,\ldots,N_c} P(\text{class}(\omega) = C_i | X = \mathbf{x}). \quad (2)$$

In practical designs, a monotonic transformation $M(\cdot)$ of the conditional probabilities $P(C_i|X)$ does not alter the classification but simplifies computations. We define N_c functions $g_i(X) = M(P(C_i|X))$, taking as the decision rule that X belongs to class i for which $g_i(X)$ is maximum. gs are called "discriminant functions," and the optimal classifier is written

$$D(\mathbf{x}) = \arg \max_{i=1,\ldots,N_c} g_i(\mathbf{x}). \quad (3)$$

B. Genetic Fuzzy Classifiers Should be Learnt and Evaluated With Fuzzy Data

It is important to note that the latest approach is followed, to our knowledge, by all genetic fuzzy classifiers [9]. The random nature of the problem is clearly assumed by all genetic fuzzy classifiers' authors, because current standard experimental designs (leave one out, cross validation, etc.) are unbiased estimations of the classification error over the whole population Ω [see (1)], and therefore the optimal classifier, no matter the learning technique, is defined by (2), or by one of its transformations, as defined in (3). Moreover, when an input is applied to a fuzzy rule base, the inference process eventually computes N_c truth values [19] or N_c number of votes [29] for the set of assertions "the input matches class C_i" and the defuzzification, in classification problems, consists in choosing the class maximizing the corresponding set of votes. This process is not different from that depicted in (3).

As a consequence of this, the term "fuzzy" does not mean in genetic fuzzy classifiers that a classification problem different from the crisp one is being solved. "Fuzzy" means here that the parameterizing of the discriminant functions has a linguistic interpretation compatible with the fuzzy logic postulates. This does not mean that a fuzzy classifier cannot be fed with fuzzy data; obviously, it can. We mean that neither learning algorithms nor statistical tests take into account the fuzzy nature of the output of the classifier. For example: we know that a random piece of fruit is yellow and weighs "about 150 g" Now imagine that we want to compare two classifiers, A and B. Classifier A outputs "pear" with confidence 0.1 and "apple" with confidence 0.2. Classifier B outputs confidences 0.8 and 0.9. Which one is better? To our knowledge, for all the statistical tests used by GFS researchers, the two are assigned the same error because they will both eventually classify the fruit as being an apple.

The experimental designs of GFSs that are focused on imprecisely observed data are not being actively studied by the GFS community. Contrary to this, and according to the fuzzy statistics community, the customary use of fuzzy sets in classification and regression problems is the treatment of vague data [14], [21]. We think that this last point should not be understressed. If we admit that the classification problem being solved by GFSs is not different from the crisp one, the following may follow.

- GFSs are not a different machine learning technique than Bayesian classifiers. They are a numerical method able to obtain discriminant functions with intuitive meanings.
- There are no reasons different from linguistic interpretability that favor fuzzy rule based classifiers. Therefore, the usefulness of approximate classifiers (fuzzy classifiers with scatter partitions) [1], where linguistic concerns are secondary, is compromised.

On the other hand, if imprecise data sets were used to train and test fuzzy classifiers, and specific statistical tests were devised to compare classifiers with fuzzy outputs, we would be able to compare different fuzzy classifiers over the set of problems where we expect that fuzzy classifiers make a difference over crisp classifiers, namely, data sets with interval or fuzzy valued data.

III. Statistical Models and GFSs

Unlike the classification case, where the Bayesian framework provided us with a widely accepted definition of the optimal classifier, there exist many definitions of statistical regression (a survey of many of them can be found in [4]). We have decided to evaluate the fuzzy extension of the standard case first, and leave robust regression techniques to be carried out in future works. Therefore, as we will show in the next section, in this paper we will define the output of the model as the conditional expectation of the output random variable, given a certain input.

A. Least Squared Models and Conditional Expectation

Let us suppose, again, that we have a set Ω that contains objects ω, but now let us assign to each one of them a numerical value $Y(\omega)$ instead of a label "class(ω)." We are given a set of measurements $X(\omega) = (X_1(\omega), X_2(\omega), \ldots, X_K(\omega))$ over every object. We will say that a model is a mapping that associates every element of $X(\Omega)$ with a value $g(X)$, whose main objective is to minimize the differences between $Y(\omega)$ and $g(X(\omega))$ over Ω.

For example, let Ω be a set of people. We observe the height and the weight of a randomly selected person and want to know his expected body fat percentage. Suppose that someone measures and weighs $X(\omega) = (180, 82)$ and has $Y(\omega) = 20\%$ of fat. We wish that the difference between the value that our model assigns to him $g(180, 82) = 22$ and the true value $Y(\omega) = 20$ be as low as possible. If we admit that there can exist two different people that measure 180 cm and weigh 82 kg but have a different percentage of fat because of their different body constitution, the assignment $g(\mathbf{x}) = Y(X^{-1}(\mathbf{x}))$ cannot be defined. In the example at hand, this means that the model will assign the same value $g = 22$ to all people that measure 180 cm and weigh 82 kg; therefore the optimal model should be defined with respect to averaged weight differences.

Again, to define the concept "averaged differences," we need to assume that the mappings X and Y fulfill all necessary conditions to be random variables. Let us also define a new random variable that quantifies the cost of assigning the value y to an object when its true value is z, cost(y, z). For the problem stated, if the model is a mapping $g(\mathbf{x})$ and $Y(\omega)$ is the value associated to the object ω, then the merit value of the model can be numerically quantified as

$$\text{err}(g) = \int_{\Omega} \text{cost}(g(X(\omega)), Y(\omega))\, \text{dP} \quad (4)$$

where the error function is integrated with respect to a probability measure P defined over Ω.

If we choose $\text{cost}(y,z) = (y-z)^2$, the expectation of the cost function is the mean squared error, which gives rise to "least squares regression." Given that $E[(Y-g(X))^2] \geq E[(Y-E[Y|X])^2]$ for any function g, the conditional expectation $g(\mathbf{x}) = E[Y|X=\mathbf{x}]$ is then the optimal definition of model

$$g(\mathbf{x}) = \frac{\int_{\{\omega:X(\omega)=\mathbf{x}\}} y\, dP}{\int_{\{\omega:X(\omega)=\mathbf{x}\}} dP} = \int y\, f(y|\mathbf{x})\, dy \tag{5}$$

where $f(y|\mathbf{x})$ is the conditional density of the output variable conditioned to a given input variable.

B. Genetic Fuzzy Models and Fuzzy Data

There have been many interval and fuzzy valued extensions of the modeling problem, in both the statistics [25], [41] and the fuzzy rule learning fields [34], [35], [37], [36]. However, the most widely used genetic methods for learning fuzzy models are least squares based [9].

In least squares based learning methods, the genetic algorithm is designed to minimize the estimation of (4) over the population, using a standard experimental design (leave one out, cross validation, etc.). But, being the optimal classifier defined by (5), it is immediate that, whenever the quality of a fuzzy model is assessed by means of its mean squared error over a sample, the best models will be nonparametric estimators of the conditional expectation. Again, as was pointed out in the preceding section, from a statistical point of view, the crisp problem is being solved, and not the fuzzy one. Therefore, least squared based GFSs are not inherently better than statistical methods over crisp problems, whatever the complexity of the genetic search.

IV. An Extended Definition of the Classification Problem

In this and the following sections, we will generalize the definitions given in Sections II and III. We will study the case where we cannot precisely observe the values obtained by the set of measures X. This vagueness will be modeled by a fuzzy set that contains the true measurement with certain confidence; therefore we will end up with a fuzzy valued data set.

We will use the interpretation of a fuzzy random variable as a nested family of random sets, which in turn are defined as imprecise observations of an unknown random variable, called the *original* random variable [28]. Consequently, it will be considered that a fuzzy valued data set is a sample of a fuzzy random variable, as defined in [20], whose α-cuts are random sets. This interpretation allows us to extend the definition of the crisp problem to the interval case first, and then to extend it to the fuzzy case.

A. Interval Data

Remember (2): to succeed, a learning algorithm should be able to estimate the values $P(\text{class}(\mathbf{x})|\mathbf{x})$ from a sample of measures taken over a subset of Ω. To simplify the notation, in this section we assume that X takes values in $\mathbb{R}$. When X has absolutely continuous distribution, the standard technique consists in making a transformation

$$D(\mathbf{x}) = \arg \max_{i=1,\ldots,N_c} \frac{f(\mathbf{x}|C_i)P(C_i)}{f(\mathbf{x})} \tag{6}$$

where f is the density function induced by the random variable X. The denominator can be removed without affecting the result

$$D(\mathbf{x}) = \arg \max_{i=1,\ldots,N_c} f(\mathbf{x}|C_i)P(C_i) \tag{7}$$

and we obtain a well-known result: from a statistical point of view, learning a classifier is the same problem as estimating a density function from a sample of a random variable.

Now we are presented with a sample from a random set and need to know how can we estimate the density function of the underlying, imprecisely observed random variable (the aforementioned *original* random variable [28]). Rephrasing the problem, we need to generalize the concept of density function to the random set case. Our primary thought was to define an "upper" density function as

$$f^*(x) := \lim_{h \downarrow 0} \frac{P^*((x-h, x+h))}{h}$$

provided that this limit exists. For instance, if the random set $\Gamma : \Omega \to \mathcal{P}(\mathbb{R})$ is a random interval of the form $\Gamma(\omega) = [X(\omega) - \epsilon, X(\omega)]$, $\forall \omega$, where X is a random variable with absolutely continuous distribution, this limit exists almost everywhere, but it is ∞. Observe that

$$\begin{aligned} &P^*((x-h, x+h)) \\ &\quad = P(\{\omega \in \Omega \mid \Gamma(\omega) \cap (x-h, x+h) \neq \emptyset\}) \\ &\quad = P(\{\omega \in \Omega \mid X(\omega) - \epsilon < x+h,\ X(\omega) > x-h\}) \\ &\quad = P(\{\omega \in \Omega \mid x-h < X(\omega) < x+h+\epsilon\}). \end{aligned}$$

For the continuity of the probability distribution induced by X, this probability converges to $P_X([x, x+\epsilon])$ when h tends to zero. When this last one is not null, the limit of the quotient tends to infinite.

To solve this problem, we work directly with (2). We need to estimate the values $P(C_i|X=\mathbf{x})$ to choose in each case the i for which the corresponding value is maximum (where X represents, in this section, the original variable, whose imprecise observation is given by Γ). For each $h > 0$, we can try to give a couple of upper and lower bounds for the value $P(C_i|X \in (x-h, x+h))$. Following [2], the limit when h tends to zero of these quantities is the value we need, $P(C_i|X=x)$. Applying the definition of conditional probability, we have that $P(C_i|X \in (x-h, x+h))$ equals

$$\frac{P(C_i \cap \{\omega \in \Omega \mid X(\omega) \in (x-h, x+h)\})}{P(\{\omega \in \Omega \mid X(\omega) \in (x-h, x+h)\})}. \tag{8}$$

The denominator is, again, the same for all classes; therefore we only need to compare the numerators for the different classes and give lower and upper bounds for it. The bounds of $P(C_i \cap \{\omega \in \Omega \mid X(\omega) \in (x-h, x+h)\})$ are

$$\underline{P}_i(x,h) = P(C_i \cap \{\omega \in \Omega \mid \Gamma(\omega) \subseteq (x-h, x+h)\}) \tag{9}$$

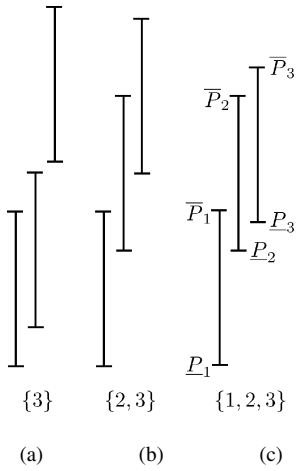


Fig. 1. Example of three possible situations when comparing interval-valued probabilities. (a) The lower bound of P_3 is higher than the upper bounds of P_2 and P_3, thus the object is assigned to class 3. (b) The lower bound of P_3 is contained in the range of values of P_2 but is higher than P_1, thus the object can be assigned to classes 2 or 3. (c) The intersection of all ranges is not empty, thus the object can be assigned to classes 1, 2, or 3.

and

$$\overline{P}_i(x,h) = P(C_i \cap \{\omega \in \Omega \mid \Gamma(\omega) \cap (x-h, x+h) \neq \emptyset\}). \quad (10)$$

Since we do not know the value of P_i but a set that contains it, it is clear that, unless the intervals $[\underline{P}_i, \overline{P}_i]$ do not overlap, we cannot know if $P_i > P_j$ for all pairs of classes; thus the decision rule D is not completely defined. This is graphically illustrated in Fig. 1: the decision rule D is no longer a point function but a set valued function, where

$$\tilde{D}(\mathbf{x}) = \{i \mid \nexists j \neq i \text{ with } \underline{P}_j > \overline{P}_i\} \quad (11)$$

and h is assigned a value small enough for the problem.

Given that $\tilde{D}$ is a set valued function, the average error of the classifier is no longer known (or, alternatively, we could say that the average error is a set valued statistic). We can find upper and lower bounds for it [see (1)]. Let us define a pair of functions

$$\underline{\text{cost}}(\omega) = 0 \text{if class}(\omega) \in \tilde{D}(\Gamma(\omega)), 1 \text{ otherwise} \quad (12)$$
$$\overline{\text{cost}}(\omega) = 0 \text{if} \tilde{D}(\Gamma(\omega)) = \{\text{class}(\omega)\}, 1 \text{ otherwise}. \quad (13)$$

In words, $\underline{\text{cost}}$ is the optimistic estimation of the error, where we admit that an object is correctly classified if its class number is included in the output, and $\overline{\text{cost}}$ is a pessimistic estimation, where we suppose that an object is misclassified unless its class number is the only output of the classifier. Therefore, the average classification error is contained in the interval

$$\text{err}(\tilde{D}) = \left[\int_\Omega \underline{\text{cost}}(\omega)\,\text{dP}, \int_\Omega \overline{\text{cost}}(\omega)\,\text{dP}\right]. \quad (14)$$

B. Fuzzy Data

If we are given a fuzzy data set, both the output of the classifier and its expected error will be fuzzy sets, as we show in this section.

Fuzzy data sets can be regarded as samples of a fuzzy random variable $\widetilde{\Gamma}$. Every instance of the variable combines two types of noise: random noise, originated in the selection of the object ("we choose a piece of fruit at random"), and observation error, originated in an imprecise measure ("the weight of the fruit is high, where 'high' is one of the values of the linguistic variable 'weight'").

α-cuts of $\widetilde{\Gamma}$ are random sets (for example, the 0.5-cut of the value "high" can be the interval [100, 160]). Therefore, for every value of α, we can build an interval classifier, as shown in the preceding section, whose output is a discrete set of class labels ("if the weight is [100, 160], then the object is compatible with both pear and apple"). It is intuitive to conclude that the output of the classifier, if fed with a fuzzy input, will be a discrete fuzzy set defined over the set of class labels ("if the weight is high, then the object is 0.1/apple + 0.6 pear.") The same can be said about the average error of the fuzzy classifier; it will be a fuzzy set.

To obtain this last value, the best description we can make about the probability $P(C_i \cap \{\omega \in \Omega \mid X(\omega) \in (x-h, x+h)\})$, given that the original random variable X is contained in the fuzzy random variable $\widetilde{\Gamma}$, is a fuzzy set $\widetilde{P}_i$, whose α-cuts are intervals $[\underline{P}_i^\alpha, \overline{P}_i^\alpha]$ defined as follows:

$$\underline{P}_i^\alpha = P(C_i \cap \{\omega \in \Omega \mid [\widetilde{\Gamma}(\omega)]_\alpha \subseteq (x-h, x+h)\}) \quad (15)$$

and

$$\overline{P}_i^\alpha = P(C_i \cap \{\omega \in \Omega \mid [\widetilde{\Gamma}(\omega)]_\alpha \cap (x-h, x+h) \neq \emptyset\}). \quad (16)$$

Therefore, the fuzzy output of the classifier will be the set

$$[\widetilde{\tilde{D}}(\mathbf{x})]_\alpha = \{i \mid \nexists j \neq i \text{ with } [\underline{P}_j]^\alpha > [\overline{P}_i]^\alpha\} \quad (17)$$

and its average error is another fuzzy set

$$[\widetilde{\text{err}}(\tilde{D})]_\alpha = \left[\int_\Omega \underline{\text{cost}}_\alpha(\omega)\,\text{dP}, \int_\Omega \overline{\text{cost}}_\alpha(\omega)\,\text{dP}\right] \quad (18)$$

where

$$\underline{\text{cost}}_\alpha(\omega) = 0 \text{if class}(\omega) \in [\tilde{D}(\Gamma(\omega))]_\alpha, 1 \text{ otherwise} \quad (19)$$
$$\overline{\text{cost}}_\alpha(\omega) = 0 \text{if} [\tilde{D}(\Gamma(\omega))]_\alpha = \{\text{class}(\omega)\}, 1 \text{ otherwise}. \quad (20)$$

C. Computer-Friendly Definition

In the preceding section, we have stated that the average error of a classifier, when its input comprises fuzzy sets, should also be a fuzzy set. Therefore, the fitness functions in GFSs will return a fuzzy value. This value can be numerically estimated by means of (18)–(20). Since these equations are expressed in terms of a family of α-cuts, we give a rewriting of them that is easier to codify in a computer.

Let $\tilde{D}(x) = \sum_{i=1\ldots N_c} \mu_i/i$ be the fuzzy output of the classifier, with p the index of the class with maximum membership value in $\tilde{D}(x)$, $p = \arg\max_{i=1\ldots N_c}(\mu_i)$, and q the second maximum membership, $q = \arg\max_{i=1\ldots N_c, i\neq p}(\mu_i)$. Let the height of $\tilde{D}(x)$, $\mu_p = 1$. Then, the contribution of the object ω to the total error is

$$\widetilde{\text{cost}}(\omega) = \begin{cases} \frac{1}{0} + \frac{\mu_q}{1} & \text{if class}(\omega) = C_p \\ \frac{\mu_{\text{class}(\omega)}}{0} + \frac{1}{1} & \text{otherwise} \end{cases}. \quad (21)$$

For example, suppose that, in a problem with three classes, the output of the classifier is the fuzzy set $\{0.2/\text{pear}\} + 1/\text{apple} + 0.8/\text{banana}\}$. If the object were a pear, the accumulated error of the classifier would be increased by the fuzzy amount $\{0.2/0+1/1\}$. If it were an apple, the new error would be $\{1/0+0.8/1\}$ higher, or $\{0.8/0+1/1\}$ if it were a banana.

V. An Extended Definition of the Modeling Problem

As before, a fuzzy valued data set is a sample of a fuzzy random variable, as defined in [20], whose α-cuts are random sets. We will extend first the definition of the modeling problem to the interval case, and then apply the results to all cuts of the fuzzy random variable sample.

A. Interval Data

Remember (5): to succeed, a learning algorithm should be able to estimate the values $E[Y|X=\mathbf{x}]$ from a sample of measures taken from a subset of Ω. Suppose that we are given samples from two random sets Γ and Γ_Y that model the imprecise observations of X and Y

$$X(\omega) \in \Gamma(\omega) \quad \omega \in \Omega \tag{22}$$

$$Y(\omega) \in \Gamma_Y(\omega) \quad \omega \in \Omega \tag{23}$$

and need to define the conditional expectation $E[Y|X]$ of the underlying, imprecisely observed random variables.

Let us suppose that Y is a discrete random variable, $Y(\omega) \in \{y_1, y_2, \ldots, y_N\}$. Then

$$g(\mathbf{x}) = \int y\, f(y|\mathbf{x})\, dy = \sum_{i=1}^{N} y_i\, P(y_i|\mathbf{x}). \tag{24}$$

We need to estimate the values $P(y_i|X=\mathbf{x})$. For a given small value $h > 0$, we can try to give a couple of upper and lower bounds for the value $P(y_i|X \in (x-h, x+h))$. Following [2], the limit when h tends to zero of these quantities is the value we need, $P(y_i|X=x)$. Applying the definition of conditional probability, we have that $P(y_i|X \in (x-h, x+h))$ equals

$$\frac{P(\{\omega \in \Omega | Y(\omega) = y_i\} \cap \{\omega \in \Omega | X(\omega) \in (x-h, x+h)\})}{P(\{\omega \in \Omega | X(\omega) \in (x-h, x+h)\})}. \tag{25}$$

The bounds of $P(\{\omega \in \Omega \mid Y(\omega) = y_i\} \cap \{\omega \in \Omega \mid X(\omega) \in (x-h, x+h)\})$ are $\underline{P}_i(x,h)$ equals

$$P(\{\omega \in \Omega | Y(\omega) = y_i\} \cap \{\omega \in \Omega \mid \Gamma(\omega) \subseteq (x-h, x+h)\})$$

and $\overline{P}_i(x,h)$ equals

$$P(\{\omega \in \Omega | Y(\omega) = y_i\} \cap \{\omega \in \Omega \mid \Gamma(\omega) \cap (x-h, x+h) \neq \emptyset\})$$

and the bounds of $P(\{\omega \in \Omega \mid X(\omega) \in (x-h, x+h)\})$ are $\underline{P}(x,h)$ equals

$$P(\{\omega \in \Omega \mid \Gamma(\omega) \subseteq (x-h, x+h)\})$$

and $\overline{P}(x,h)$ equals

$$P(\{\omega \in \Omega \mid \Gamma(\omega) \cap (x-h, x+h) \neq \emptyset\}).$$

Thus we can know that the conditional expectation $E[Y|X]$ is contained in the interval defined as follows [the denominator of (25) does not depend on i]:

$$[\underline{g}, \overline{g}](\Gamma, h) = \frac{\bigoplus_{i=1}^{N} y_i\, [\underline{P}_i(x,h), \overline{P}_i(x,h)]}{[\underline{P}(x,h), \overline{P}(x,h)]} \tag{26}$$

where $[a,b] \oplus [c,d] = \{u+v \mid u \in [a,b], v \in [c,d]\}$, and the quotient must be understood as an interval valued operation $[a,b]/[c,d] = \{u/v \mid u \in [a,b], v \in [c,d]\}$.

In words, when the model was fed with a real input $\mathbf{x}$, its output was $g(\mathbf{x})$. Now we have fed the model with an interval Γ, and we knew that $\mathbf{x}$ was contained in Γ. Its output has been the interval $[\underline{g}, \overline{g}](\mathbf{x})$, which has been constructed to contain $g(\mathbf{x})$.

B. Fuzzy Data

If we are given a fuzzy data set, both the output of the model and its expected error will be fuzzy sets, as we show in this section.

Fuzzy data sets can be regarded as samples of a fuzzy random variable $\tilde{X} \times \tilde{Y}$. Every instance of the variable combines two types of noise: random noise, originated in the selection of the object ("we choose a person at random"), and observation error, originated in an imprecise measure ("the weight of the person is high, where 'high' is one of the values of the linguistic variable 'weight'").

The α-cuts $\tilde{X}_\alpha$ and $\tilde{Y}_\alpha$ are random sets (for example, the 0.5-cut of the value "high" can be the interval [80, 110]). Therefore, for every value of α we can build an interval model, as shown in the preceding section, whose output is an interval of values ("if the weight is [80, 110], then the percentage of body fat is between 20 and 30"). It is intuitive to conclude that the output of the model, if presented a fuzzy input, will be a fuzzy set defined over the set of outputs ("if the weight is high, and height is low, then the body fat is high.") In fact, this is the usual structure of a fuzzy rule.

To obtain this last value, the best description we can make about the probability $P(\{\omega \in \Omega \mid Y(\omega) = y_i\} \cap \{\omega \in \Omega \mid X(\omega) \in (x-h, x+h)\})$, given that the original random variable X is contained in the fuzzy random variable $\tilde{X}$, is a fuzzy set $\widetilde{P}_i$, whose α-cuts are intervals $[\underline{P}_i^\alpha, \overline{P}_i^\alpha]$, where $\underline{P}_i^\alpha$ equals

$$= P(\{\omega \in \Omega | Y(\omega) = y_i\} \cap \{\omega \in \Omega | [\tilde{X}(\omega)]_\alpha \subseteq (x-h, x+h)\})$$

and $\overline{P}_i^\alpha$, $\underline{P}^\alpha$, and $\overline{P}^\alpha$ are defined similarly, as we did in the preceding section. Therefore, the fuzzy output of the model will be the set $\widetilde{g}(\tilde{X}, h)$, defined by its α-cuts

$$[\widetilde{g}(\tilde{X}, h)]_\alpha = \frac{\bigoplus_{i=1}^{N} y_i\, [\underline{P}_i^\alpha(x,h), \overline{P}_i^\alpha(x,h)]}{[\underline{P}^\alpha(x,h), \overline{P}^\alpha(x,h)]}. \tag{27}$$

The fuzzy-arithmetic based quotient in (27) may produce rather conservative estimations of g; i.e., we can expect the nonspecificity of $\widetilde{g}$ to be large. However, we will seldom build a fuzzy model using (27). It is easier to define a parametric family of functions (for instance, that family could be the set of all fuzzy rule-based models that depend on certain linguistic

partitions of the variables) and then to choose the element of that family that minimizes the squared error. The difficulty here is that the output of the model is no longer known but a fuzzy set that contains it, and therefore we can at most determine a fuzzy interval for the squared error of a model. This problem will be studied in the next section.

C. Mean Squared Error of a Fuzzy Model and the Variance of a Fuzzy Random Variable

Since we cannot know the precise output but fuzzy sets that describe them, we cannot compute a number that measures the error of a candidate model over our train data, but we can provide a fuzzy interval for it. Let $\tilde{D} = \tilde{Y} - g(\tilde{X})$ be the residual of the model. We are interested in computing $E(\tilde{D}^2)$. Observe that the definition of this value is very near to that of the variance of a fuzzy random variable (FRV), which is a well studied problem. The best known definitions of the variance of an FRV are two, which we will name "classical" and "imprecise." We will derive a definition of the mean squared error from either one of them. There are also some recent new definitions, not so widespread, that we will not cover here. A review of them can be found in [13].

1) Classical Variance: Let us consider a probability space $(\Omega, \mathcal{A}, P)$ and a metric d defined over the class of the fuzzy subsets of $\mathbb{R}$ (or over a subclass), and let us suppose that $\tilde{X} : \Omega \to \tilde{\mathcal{P}}(\mathbb{R})$ is a function $\mathcal{A}-\beta(d)$-measurable (here, $\beta(d)$ represents the Borel σ-algebra induced by d). The *classical variance* of $\tilde{X}$ is the quantity

$$\mathrm{Var}_{\mathrm{Cl}}(\tilde{X}) = \int_{\Omega} d(X, E(\tilde{X}))^2 \, dP$$

and we will call the *classical mean squared error* (CMSE) of a model with residual $\tilde{D}$ to the quantity

$$\mathrm{CMSE}(\tilde{D}) = \int_{\Omega} d(\tilde{D}, 0)^2 \, dP.$$

The different definitions of variance in the literature that fit this formulation differ in the metric used and in the definition of the expectation of a fuzzy random variable [27], [30].

This definition is very convenient from a numerical point of view because the error of a fuzzy model is reduced to a crisp number that could be easily optimized. Unfortunately, it is not compatible with our semantic interpretation and its use would not produce meaningful results, as we point out in the example that follows.

2) Example 1: Let us suppose that we have a sample of size two of the residual of the model and that this residual is the FRV $\tilde{D}$, whose images are the triangular fuzzy sets $\tilde{D}(\mathbf{x}_1) = (-K, 0, K)$ and $\tilde{D}(\mathbf{x}_2) = (-K, 0, K)$ (see Fig. 2.) This FRV, if regarded as a classical measurable function, has null CMSE, since it is a constant function, and therefore the model is assigned a null error. However, in our context, it is not coherent to state that we cannot know the precise data that the model must fit but only certain fuzzy sets that contains them, and simultaneously be able to affirm that the model matches these unknown data without error.

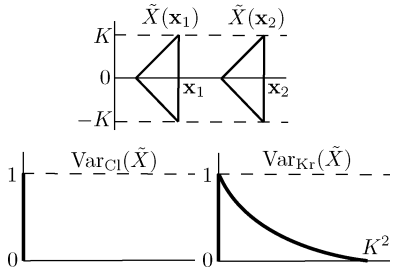


Fig. 2. Both definitions of variance explained in the text are summarized here for the data in Example 1. The classical squared error of this model is the crisp number zero. The imprecise squared error produces a fuzzy interval that contains the true, unknown error of the model.

3) Imprecise Variance: In [31], Kruse defines the variance of a multivalued mapping, $\Gamma : \Omega \to \mathcal{P}(\mathbb{R})$, as the set

$$\mathrm{Var}_{\mathrm{Kr}}(\Gamma) = \{\mathrm{Var}(X) \mid X \in S(\Gamma)\}$$

where $S(\Gamma)$ represents the set of all measurable selections of the multivalued mapping. The preceding definition can be easily extended to the case of fuzzy random variables as follows: Let us call *Kruse's variance* of the fuzzy random variable $\tilde{X} : \Omega \to \tilde{\mathcal{P}}(\mathbb{R})$ to the only fuzzy set determined by the nested family of sets

$$[\mathrm{Var}_{\mathrm{Kr}}(\tilde{X})]_{\alpha} := \mathrm{Var}_{\mathrm{Kr}}(\tilde{X}_{\alpha}), \forall\, \alpha$$

where $\tilde{X}_{\alpha}$ is the multivalued mapping α-cut of $\tilde{X}$. (The variance of Example 1 is plotted in the bottom part of Fig. 2.)

Therefore, we define the *second-order mean squared error* (SMSE) as the fuzzy set

$$[\mathrm{SMSE}(\tilde{D})]_{\alpha} := \mathrm{Sq}_{\mathrm{Kr}}(\tilde{D}_{\alpha}), \forall\, \alpha$$

where

$$\mathrm{Sq}_{\mathrm{Kr}}(\Gamma) = \{\mathrm{E}(D^2) \mid D \in S(\tilde{D})\}.$$

We can easily check that the membership function of this fuzzy set is given by the expression

$$\mathrm{SMSE}(\tilde{D})(x) = \sup\{\mathrm{acc}(D) \mid \mathrm{E}(D^2) = d\},\ \forall\, d \in \mathbb{R}.$$

The membership degree of a value x to the fuzzy set $\mathrm{SMSE}(\tilde{D})$ represents the possibility degree of the original random variable's being one of those whose squared error is equal to x. We propose using the SMSE as the fitness of a fuzzy model when applied to fuzzy data.

VI. Examples of Fuzzy Fitness Evaluations

In this section, we will numerically evaluate the two proposed fitness functions (for fuzzy classifiers and for fuzzy models) over toy problems to clarify the computational procedures.

A. Example of Fitness Evaluation in the Extended Classifier

Let us suppose that we have to discriminate between three classes (apple, pear, banana), given the weight of a piece of fruit. To design the classifier, we are given a sample comprising five pieces, whose weights and classes are given in Table I.

TABLE I
DATA SET FOR THE EXAMPLE PROBLEM "FRUIT"

	crisp weight	fuzzy weight	class
1	111	(102,111,116)	pear
2	96	(88,96,112)	apple
3	116	(104,116,128)	pear
4	89	(83,89,90)	banana
5	101	(91,101,118)	apple

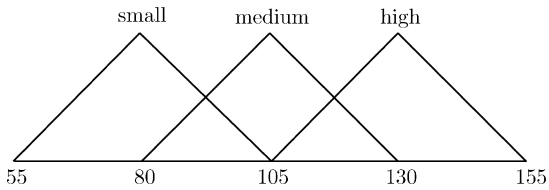


Fig. 3. Definition of the linguistic variable "weight," as used in the example problem "fruit."

TABLE II
OUTPUT OF THE CLASSIFIER IN THE EXAMPLE WHEN FED WITH CRISP DATA (COLUMN "CRISP WEIGHT" IN TABLE I)

Crisp Input	Output	Cost
111	apple	1
96	apple	0
116	apple	1
89	banana	0
101	apple	0

Weights are triangular fuzzy numbers, designated by three numbers: leftmost, center, and rightmost values.

Let us also suppose that the GFS has to evaluate the fitness of the rule base that follows:

```
if weight is small then banana
if weight is medium then apple
if weight is high then pear
```

where the linguistic variable "weight" takes the values shown in Fig. 3. We wish to assign a fitness value to this rule base, given the mentioned data set.

Observe that the fitness value assigned to this rule base measures the classifier error as defined in (18). This, on the one hand, assesses the degree to which these rules approximate the usual Bayes criterion in (7) but, on the other hand, also takes into account how sensitive these rules are to measurement errors. The nonspecificity of the fuzzy fitness value is higher when bases are less robust. The ranking of the fitness values should take into account more information than that given, for example, in the center of gravity of the set.

Let us evaluate first this classifier over the crisp dataset given by the column "crisp weight" in Table I. The output of the classifier, using the winner rule inference mechanism, is shown in Table II. The cost of this classifier is two (in other words, we estimate that it is wrong 40% of the time).

If we apply an interval input to the same classifier (the support of the fuzzy examples), its output is a crisp subset of the class labels. For example, the interval [88, 112] has associated with it the crisp subset $\tilde{D}(\Gamma(\mathbf{x})) = \tilde{D}([88{,}112]) = \{\text{banana, apple}\}$ because, if we classify all the points in [88, 112], we observe that points in [88, 92.5] are assigned the class "banana" and points in (92.5, 112] are assigned the class "apple." To calculate the cost of the classifier, we operate as shown in Table III. We find that the cost is contained in the interval [1, 4], i.e., when data are precisely measured, we estimate that the classification is wrong 40% of the time; when data are interval-valued, all we can say without assuming a random distribution of the observation error is that it is wrong between 20% and 80% of the time.

TABLE III
COST OF THE CLASSIFIER IN THE EXAMPLE WHEN IT IS FED WITH INTERVAL DATA (THE SUPPORT OF THE FUZZY EXAMPLES IN THE COLUMN "FUZZY WEIGHT" IN TABLE I)

Interval Input	Output	Cost
$[102, 116]$	{ apple }	$\{1\}$
$[88, 112]$	{ banana, apple }	$\{0, 1\}$
$[104, 128]$	{ apple, pear }	$\{0, 1\}$
$[83, 90]$	{ banana }	$\{0\}$
$[91, 118]$	{ banana, apple, pear }	$\{0, 1\}$

TABLE IV
OUTPUT OF THE EXAMPLE CLASSIFIER WHEN THE INPUT IS A FUZZY SET

Fuzzy Input	Output	Cost
$(102, 111, 116)$	{ 1/apple }	$\{1/1\}$
$(88, 96, 112)$	{ 0.5625/banana+1/apple }	$\{1/0 + 0.5625/1\}$
$(104, 116, 128)$	{ 0.875/apple+1/pear }	$\{0.875/0 + 1/1\}$
$(83, 89, 90)$	{ 1/banana }	$\{1/0\}$
$(91, 101, 118)$	{ 0.15/banana+1/apple+0.0294/pear }	$\{1/0 + 0.15/1\}$

Finally, if the classifier is applied a fuzzy input, its outputs and costs are as shown in Table IV. The inputs are fuzzy triangular numbers and the data (x, y, z) are left, center, and right point. The cost of the classifier is

$$\left\{\frac{1}{1}\right\} \oplus \left\{\frac{1}{0} + \frac{0.5625}{1}\right\} \oplus \left\{\frac{0.875}{0} + \frac{1}{1}\right\} \oplus \left\{\frac{1}{0}\right\} \oplus \left\{\frac{1}{0} + \frac{0.15}{1}\right\} = \left\{\frac{0.875}{1} + \frac{1}{2} + \frac{0.5625}{3} + \frac{0.15}{4}\right\}$$

or, in words, the error of the classifier is 20% with confidence 0.875, 40% with confidence 1, 60% with confidence 0.5625 and 80%, with confidence 0.15. The error is still between 20% and 80%, but a genetic algorithm could prefer this result over a different classifier that has, say, $\{0.875/1{+}1/2{+}0.5625/3{+}0.25/4\}$, even if the punctual estimations of the classification error of either are the same, because the differences in the fuzzy errors state that the former classifier is less affected by imprecision in the input data (it is less likely to obtain an 80% error.) Observe also that, if input data are triangular fuzzy sets, the punctual error of the classifier is given by the value with membership 1 in the fuzzy cost, in this example two (or 40% of errors).

It is remarked that the algorithm used here to calculate the output, and the error of the classifier, given a fuzzy input, does *not* produce the same results that we would have obtained by means of the direct use of fuzzy inference. For example, if we apply max-min inference to compute the output of the rule base when its input is the fuzzy set (91, 101, 118), we obtain $\{0.4/\text{banana} + 0.9048/\text{ apple} + 0.3095/\text{pear}\}$. However, the procedure proposed in this paper produces the set $\{0.15/\text{banana} + 1/\text{apple} + 0.0294/\text{pear}\}$. In other words, we have proposed using fuzzy logic to assign a class to a crisp input but a fuzzy statistics-based interpretation of the observation error to extend the classification to imprecise data.

TABLE V
DATA SET FOR THE EXAMPLE PROBLEM "BODY FAT"

	weight	fuzzy weight	fat	fuzzy fat
1	75	(74,75,76)	17	(12,17,19)
2	88	(87,88,89)	24	(21,24,27)
3	82	(81,82,83)	20	(18,20,22)
4	80	(79,80,81)	21	(18,21,22)
5	72	(71,72,73)	12	(11,12,13)

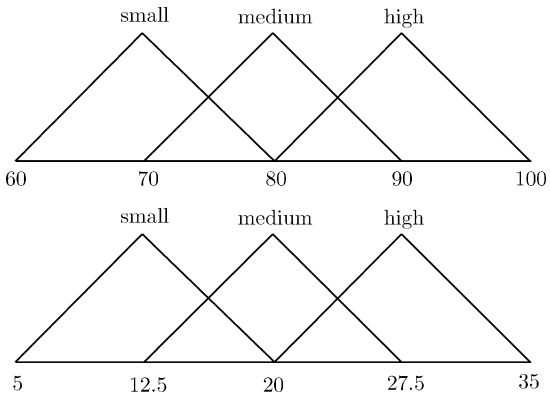


Fig. 4. Definition of the linguistic variables (top) "weight" and (bottom) "body fat" as used in the second example (extended models.).

TABLE VI
RESULTS OF EVALUATING THE EXAMPLE MODEL OVER THE CRISP DATA SET (COLUMN "WEIGHT" IN TABLE V)

X	Y	g	Cost
75	17	16.25	0.56
88	24	25.45	2.12
82	20	22.05	4.18
80	21	20	1
72	12	14.55	6.48

B. Example of Fitness Evaluation in the Extended Model

In this second example, let us suppose that we have to guess the percentage of body fat, given the weight of a person. To design the model, we are given a sample comprising five people, whose weights and percentages are given in Table V.

Let us also suppose that the GFS has to evaluate the fitness of the rule base that follows:

```
if weight is small then fat is small
if weight is medium then fat is medium
if weight is high then fat is high
```

where the linguistic variables "weight" and "body weight" take the values shown in Fig. 4.

We wish to assign a fitness value to this rule base, given the mentioned data set. Let us evaluate first this model over the crisp data set given by the column "weight" in Table V. We have used weighted center of gravity defuzzification (the output of the model is computed as a weighted sum of the output of each rule, where the weights are the areas of the truncated memberships of the output). The results are displayed in Table VI, and the cost of this model is $(0.56 + 2.12 + 4.18 + 1 + 6.48)/5 = 2.86$.

If we apply an interval input to the same model (the support of the fuzzy examples), its output is an interval of values. Observe that, since the rule base in this example defines a monotonic continuous mapping, we just need to compute the output at the boundaries of the intervals, but this might not be true with a different rule base. The interval outputs and costs are displayed in Table VII, and the cost is contained in the interval [0.10, 20.11], i.e., when data are precisely measured, we estimate that the mean squared error was 2.86. With interval-valued data, all we can conclude, knowing (or assuming) the random distribution of the observation error, is that the error is contained in the interval [0.10, 20.11].

TABLE VII
INTERVAL OUTPUTS AND COSTS OF THE EXAMPLE MODEL OVER THE INTERVAL DATA SET (SUPPORTS OF THE FUZZY EXAMPLES IN THE COLUMN "FUZZY WEIGHT", TABLE V)

Sample Input	Sample Output	Model Output	Cost
$[74, 76]$	[12,19]	[15.74, 16.76]	[0,22.62]
$[87, 89]$	[21,27]	[24.81, 26.29]	[0,28.01]
$[81, 83]$	[18,22]	[21.21, 22.69]	[0,22.03]
$[79, 81]$	[18,22]	[18.79, 21.21]	[0,10.29]
$[71, 73]$	[11,13]	[13.71, 15.19]	[0.5,17.59]

TABLE VIII
OUTPUT OF THE EXAMPLE MODEL WHEN THE INPUT IS A FUZZY SET. THE TRIPLETS (a, b, c) REPRESENT THE LOWER LIMIT, MODE AND UPPER LIMIT OF THE CORRESPONDING FUZZY NUMBERS. THE ERROR OF THE MODEL IS (0.10, 2.86, 20.11), AND ITS MEMBERSHIP FUNCTION IS PLOTTED IN FIG. 5

Sample Input	Sample Output	Model Output	Cost
$(74, 75, 76)$	(12,17,19)	(15.74, 16.25, 16.76)	(0, 0.56, 22.62)
$(87, 88, 89)$	(21,24,27)	(24.81, 25.45, 26.29)	(0, 2.12, 28.01)
$(81, 82, 83)$	(18,20,22)	(21.21, 22.05, 22.69)	(0, 4.18, 22.03)
$(79, 80, 81)$	(18,21,22)	(18.79, 20, 21.21)	(0, 1, 10.29)
$(71, 72, 73)$	(11,12,13)	(13.71, 14.55, 15.19)	(0.5, 6.48, 17.59)

Finally, if the model is applied a fuzzy input, its outputs and costs are shown in Table VIII. The modal point is also 2.86. Observe that not the output, the costs, nor the average error are triangular fuzzy numbers; the fuzzy valued fitness of this rule base is plotted in Fig. 5. It is remarked that the numerical procedure described here is not different from those used to compute the variance of a sample of an FRV (see, for instance, [16]).

VII. REVIEW OF PRACTICAL APPLICATIONS

Beginning with the first publication of the ideas contained in this paper [38], we have developed some practical applications of GFSs that combine the use of a fuzzy-valued fitness function and vague data. In this section, we make a short review of these applications and reproduce their most relevant results.

Depending on the source of the fuzziness in the data, we can enumerate three categories, which we will detail below:

1) crisp data with hand-added fuzziness [40];
2) transformations of data based on semantic interpretations of fuzzy sets [7], [39];
3) inherently fuzzy data [33].

A. Crisp Data With Hand-Added Fuzziness

If a small amount of fuzziness is artificially added to each element of a crisp data set, the use of the fuzzy fitness function proposed in this paper might help to improve the robust-

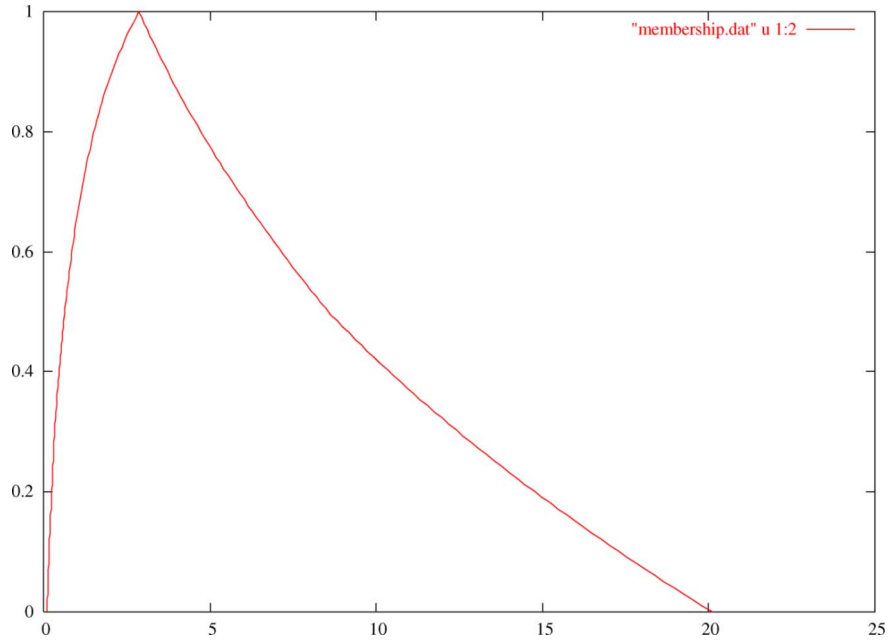


Fig. 5. SMSE of the example model. Observe that this fuzzy fitness carries information about the punctual error of the model (the mode) but also about the mean slope of the model: the higher the mean of the slopes of the model in the points of the sample, the less specific the SMSE is.

ness of both classifiers and models. An example of this technique is in [40], where a problem related to the learning of fuzzy rule-based models with backfitting algorithms was solved. The algorithm discussed in that reference incrementally generates fuzzy rules from data. Each new rule is chosen on the basis of the residual of the model in the preceding iteration; therefore some outliers might have rules assigned, and this is not desired. Regularization techniques help to mitigate the effect; if the maximum slope of the model is limited, an isolated point will not be assigned a rule whose consequent is too different from those in its neighborhood.

The regularization proposed in said paper consists of *fuzzifying* by hand the crisp data set and optimizing the subsequent fuzzy-valued fitness function with a multicriteria simulated annealing algorithm. Models with high slopes will be penalized because a small deviation in the input will mean they have a higher upper bound in the fuzzy error proposed here.

In Fig. 6, we have reproduced some of the results of this last work. The box-plots of the mean squared error over the test data are shown for (a) the original algorithm and (b) for the algorithm trained over data to which a triangular, symmetrical fuzzy set with a support of size 0.01 was added.

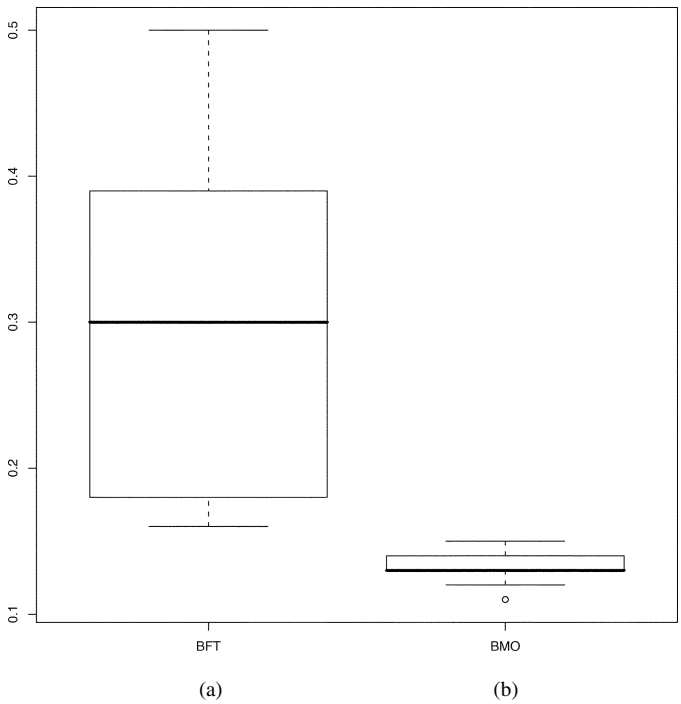


Fig. 6. (a) Test errors of the same algorithm trained over crisp data and data to which a triangular, symmetrical fuzzy set with a support of size 0.01 was added (reproduced from [40]).

B. Transformations Based on Semantic Interpretations of Fuzzy Sets

There exist certain problems where each pattern comprises a set of values. For instance, when questionnaires are designed, a factor can be evaluated by the answers to a set of different questions. All of these answers are averaged to obtain the level of that factor. They may have contradictions, though. In this case, averaging them discards potentially useful information. In [7], it is proposed to represent these lists of answers by means of a fuzzy set each. This set is defined by means of a semantic interpretation of a fuzzy set as a nested family of confidence intervals [12], [22].

In the aforementioned [7], a model of preferences in a marketing problem was solved with this procedure. The fuzzy-valued fitness function was optimized by means of a genetic algorithm that used a fuzzy ranking to select the best of any two fuzzy intervals. Later, in [39], the same data were

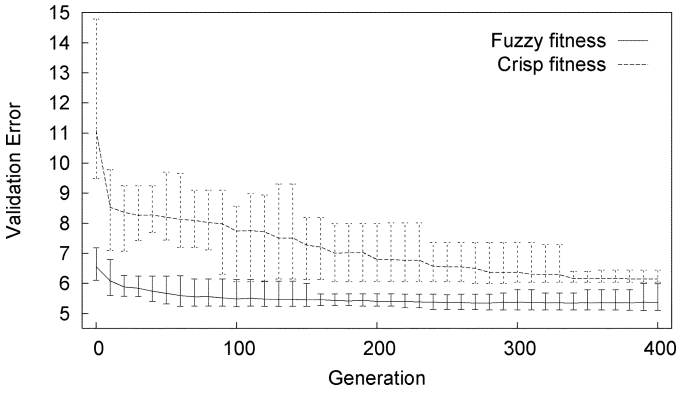


Fig. 7. Minimum, mean, and maximum values of the validation error of an algorithm that learns fuzzy rules from data, applied to a marketing problem (reproduced from [39]). The crisp problem is drawn with dashed lines and the fuzzy problem with solid lines.

optimized with a multiobjective genetic algorithm, derived from NSGA-2 [18], that does not use the fuzzy ranking but a true dominance relation based on the fuzzy fitness. Some of the results of this last [39] are reflected in Fig. 7, where it is shown that the validation error of the fuzzy fitness based model is lower than that of the original version of the algorithm, which was based on the average value of the answers.

C. Inherently Fuzzy Data

There are also certain practical problems where the data are inherently fuzzy; thus the use of a fuzzy valued fitness function is the natural choice. For instance, in [33], a novel industrial application, where taximeters are calibrated with the help of a Global Positioning System (GPS), is described. The output of a standard GPS receiver comprises a set of confidence intervals for the expected position of the vehicle, obtained at different significance levels; thus it matches the same semantic interpretation of a fuzzy set that was mentioned in the preceding section; it makes sense to state that a GPS gives fuzzy coordinates of the position of the vehicle (see Fig. 8). In the work reported in that paper, it was necessary to compute the lowest upper bound of all the trajectories compatible with a set of fuzzy coordinates. A modified NSGA-2 multiobjective genetic algorithm was used to search for the model that minimized the fuzzy error between that set of vague coordinates taken from the GPS and the model-based trajectory of the taxi.

VIII. Concluding Remarks and Open Problems

In both stochastic classifiers and models, when data are vague, it is necessary to introduce some hypotheses over the measurement errors. Fuzzy algorithms are less restrictive about these assumptions. Generally speaking, statistical models and classifiers assume a well-known probability distribution over the measurement errors, while fuzzy approaches only assume that we know a couple of lower and upper bounds for the probability of each error.

If data are "defuzzified" before they are fed to the learning algorithm, some information is lost. In this case, the optimal decisions are the Bayes classifier or the conditional expectation, thus we cannot expect GFS to outperform statistical methods, and the benefits of the fuzzy approach are restricted to the field of linguistic understandability. Nevertheless, when using vague data, fuzzy algorithms are able to draw conclusions under weaker assumptions than the stochastic ones. Therefore, we can state, in this sense, that fuzzy classifiers and models are better than their stochastic counterparts.

Other problems are still in the works. Apart from its obvious use (induce classifiers and models from fuzzy data), we think that the fuzzy-valued fitness functions proposed here might be useful in some crisp problems. For instance, observe the situation on the left-hand side of Fig. 9. The three decision surfaces

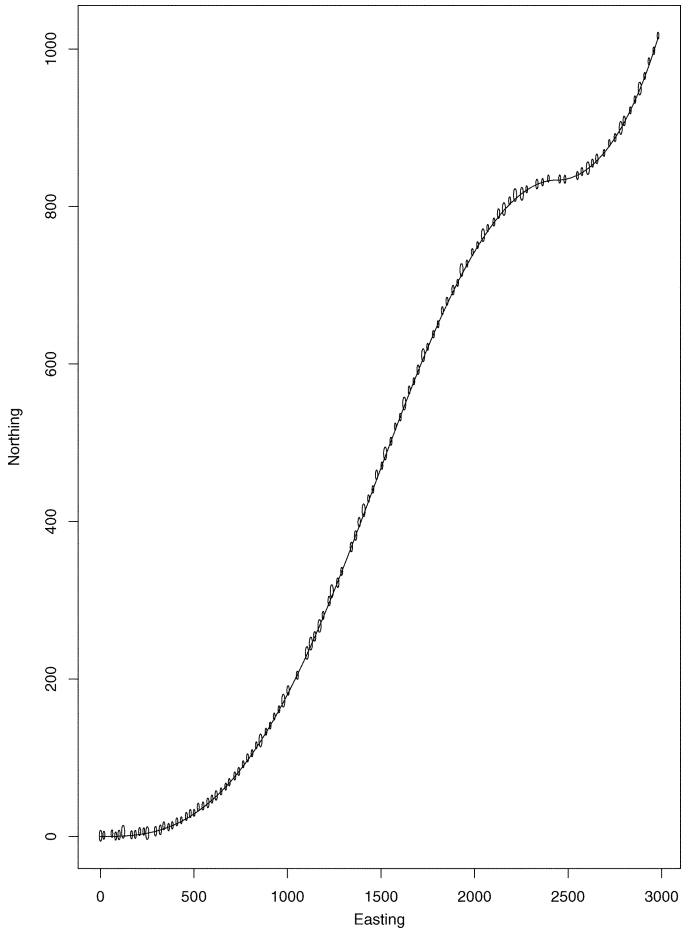


Fig. 8. A 0.9-cut of the fuzzy coordinates produced by a GPS and true trajectory of the vehicle. The distance between a model of the trajectory and the data of the GPS is inherently a fuzzy interval (reproduced from [33]).

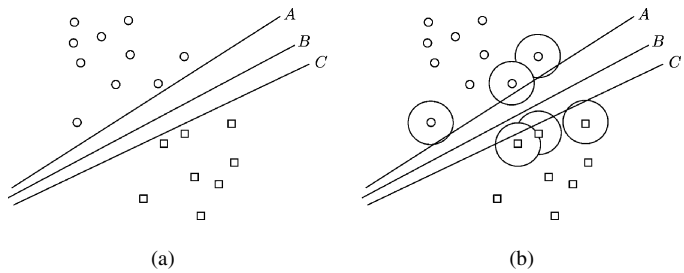


Fig. 9. (a) The three decision surfaces A, B, and C completely separate the two classes; thus they would be assigned the same *crisp* fitness. (b) If we enclose every example in a ball of size ϵ and proceed as if the examples were imprecise, our fuzzy fitness function biases B over A and C.

A, B, and C separate squares from circles without error. Despite the fact that the decision surface B is further from the examples of both classes, there is no information in the crisp fitness function that makes it preferable. Now let us artificially center an ϵ-sized neighborhood in each example and treat the sample as if it were imprecise, i.e., as if we only knew that the examples are contained in their corresponding neighborhoods. The three interval-valued fitness functions will contain the value 0 but the only surface whose error is null will be the surface B. This seems to suggest that connections between minimum margin classifiers [15] and genetic classifiers with interval and fuzzy valued fitness may be found.

Finally, the adoption of a fuzzy-valued fitness function poses some problems, both in the implementation of the genetic algorithm and in the experimental design.

1) To carry out fitness-based orderings of individuals in the GA, we must either use a fuzzy ranking [5], [7] to induce a total order over the fuzzy parts of $\mathbb{R}$ or induce only a partial ordering and use multiobjective genetic algorithms instead [10], [39]. The selection of the best fuzzy ranking or, from a more general point of view, the processing fuzzy values when evaluating the fitness function in genetic algorithms, is a problem that cannot yet be considered as generally solved [26].
2) New statistical tests have to be designed in order to judge the relevance of the differences of two fuzzy algorithms. There are some works in fuzzy statistical inference [6], [11], [17], but more work needs to be done in order to make practical comparisons between fuzzy valued algorithms. It is not clear yet how the comparison between fuzzy and crisp data should be made (and it is necessary, in order to compare extended GFSs to other algorithms), and this originates other problems: the definition of statistical tests about fuzzy-valued parameters in fuzzy random variables, the definition of parametric families of random sets or fuzzy random variables, and the design of their corresponding test statistics.
3) Common benchmarks used with GFSs include missing values that can be codified with fuzzy information, and linguistic data, that can also be assimilated to fuzzy sets, but we lack data sets of imprecisely measured data that allow us to compare the robustness of GFSs to that of stochastic methods in terms of the degree of imprecision in the data. This absence prevents us from optimizing GFSs towards the main objective of fuzzy techniques, as stated by Zadeh [42]: "Exploit the tolerance for imprecision [...] to achieve tractability, robustness, and low solution cost."

REFERENCES

[1] R. Alcalá, J. Casillas, O. Cordón, and F. Herrera, "Building fuzzy graphs: Features and taxonomy of learning non-grid-oriented fuzzy rule-based systems," *Int. J. Intell. Fuzzy Syst.*, vol. 11, pp. 99–119, 2001.

[2] R. B. Ash, *Basic Probability Theory*. New York: Wiley, 1970.

[3] C. Bertoluzza, M. A. Gil, and D. A. Ralescu, Eds., *Statistical Modeling, Analysis and Management of Fuzzy Data*, ser. Fuzziness and Soft Computing. Berlin, Germany: Springer, 2003, vol. 87.

[4] D. Birkes and Y. Dofge, *Alternative Methods of Regression*. New York: Wiley, 1993.

[5] L. Campos and A. González, "Further contributions to the study of the average value for ranking fuzzy numbers," *Int. J. Approx. Reason.*, vol. 10, no. 2, pp. 135–153, 1994.

[6] M. R. Casals, M. A. Gil, and P. Gil, "On the use of Zadeh's probabilistic definition for testing statistical hypotheses from fuzzy information," *Fuzzy Sets Syst.*, vol. 20, pp. 175–190, 1986.

[7] J. Casillas and L. Sánchez, "Knowledge extraction from fuzzy data for estimating consumer behavior models," in *Proc. Fuzzy IEEE*, Vancouver, BC, Canada, 2006.

[8] O. Cordón, F. A. C. Gomide, F. Herrera, F. Hoffmann, and L. Magdalena, "Genetic fuzzy systems. New developments," *Fuzzy Sets Syst.*, vol. 141, no. 1, pp. 1–3, 2004.

[9] O. Cordón, F. Herrera, F. Hoffmann, and L. Magdalena, *Genetic Fuzzy Systems, Evolutionary Tuning and Learning of Fuzzy Knowledge Bases*. Singapore: World Scientific, 2001.

[10] C. A. Coello, D. A. Van Veldhuizen, and G. B. Lamont, *Evolutionary Algorithms for Solving Multi-Objective Problems*. Norwell, MA: Kluwer Academic, 2002.

[11] N. Corral and M. A. Gil, "The minimum inaccuracy fuzzy estimation: An extension of the maximum likelihood principle," *Stochastica*, vol. 8, pp. 63–81, 1984.

[12] I. Couso, S. Montes, and P. Gil, "The necessity of the strong alpha-cuts of a fuzzy set," *Int. J. Uncertain., Fuzz. Knowl.-Based Syst.*, vol. 9, no. 2, pp. 249–262, 2001.

[13] I. Couso, S. Montes, and L. Sánchez, "Varianza de una variable aleatoria difusa: Estudio de distintas definiciones," in *Proc XII Congreso Espanol sobre Tecnologias y Logica Fuzzy* (in Spanish), Jaén, Spain, 2004, pp. 267–272.

[14] I. Couso, L. Sánchez, and P. Gil, "Imprecise distribution function associated to a random set," *Inf. Sci.*, vol. 159, no. 1-2, pp. 109–123, 2004.

[15] N. Cristiniani and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[16] D. Dubois, H. Fargier, and J. Fortin, "The empirical variance of a set of fuzzy intervals," in *Proc. Fuzzy IEEE 2005*, Reno, NV, 2005, pp. 885–890.

[17] M. A. Gil, N. Corral, and P. Gil, "The minimum inaccuracy estimates in χ^2 tests for goodness of fit with fuzzy observations," *J. Stat. Plan. Inf.*, vol. 19, pp. 95–115, 1985.

[18] K. Deb, S. Agrawal, A. Pratab, and T. Meyarivan, M. Schoenauer, K. Deb, G. Rudolph, X. Yao, E. Lutton, J. J. Merelo, and H.-P. Schwefel, Eds., "A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II," in *Proc. Parallel Problem Solving Nature VI Conf.*, 2000, pp. 849–858.

[19] D. Dubois and H. Prade, *Fuzzy Sets and Systems, Theory and Applications*. New York: Academic, 1980.

[20] F. Féron, "Ensembles aleatoires flous," *C.R. Acad. Sci. Paris Ser. A*, vol. 282, pp. 903–906, 1975.

[21] M. A. Gil, "Fuzziness and loss of information in statistical problems," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-17, pp. 1012–1025, 1987.

[22] I. R. Goodman and H. T. Nguyen, *Uncertainty Models for Knowledge-Based Systems*. Amsterdam, The Netherlands: North-Holland, 1985.

[23] D. J. Hand, *Discrimination and Classification*. New York: Wiley, 1981.

[24] R. Herbrich and T. Graepel, "Introduction to the special issue on learning theory," *JMLR (Special Issue on Learning Theory)*, pp. 755–757, 2003.

[25] R. Koenker and K. Hallock, "Quantile regression," *J. Econ. Perspect.*, vol. 15, pp. 143–156, 2001.

[26] M. Köppen, K. Franke, and B. Nickolay, "Fuzzy-Pareto-dominance driven multiobjective genetic algorithm," in *Proc. 10th Int. Fuzzy Syst. Assoc. World Congr. (IFSA)*, Istanbul, Turkey, 2003, pp. 450–453.

[27] R. Körner, "On the variance of fuzzy random variables," *Fuzzy Sets Syst.*, vol. 92, pp. 83–93, 1997.

[28] R. Kruse and K. D. Meyer, *Statistics with Vague Data*. Dordrecht, The Netherlands: Reidel, 1987, vol. 33.

[29] L. I. Kuncheva, *Fuzzy Classifier Design*. New York: Springer-Verlag, 2000.

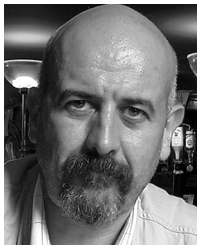
[30] M. A. Lubiano, M. A. Gil, M. López Díaz, and M. T. López, "The λ-mean squared dispersion associated with a fuzzy random variable," *Fuzzy Sets Syst.*, vol. 111, no. 3, pp. 307–317, 2000.

[31] K. D. Meyer and R. Krüse, *Statistics with Vague Data*. Dordrecht, The Netherlands: Reidel, 1987.

[32] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Machine Learn.*, vol. 52, pp. 239–281, 2003.

[33] A. Otero, J. Otero, L. Sánchez, and J. R. Villar, "Longest path estimation from inherently fuzzy data acquired with GPS using genetic algorithms," in *Proc. EFS*, Lancaster, U.K., 2006.

[34] L. Sánchez, "A random sets-based method for identifying fuzzy models," *Fuzzy Sets Syst.*, vol. 98, no. 3, pp. 343–354, 1998.
[35] L. Sánchez, "Interval-valued GA-P algorithms," *IEEE Trans. Evol. Comput.*, vol. 4, no. 1, pp. 64–72, 2000.
[36] L. Sánchez, J. Casillas, O. Cordón, and M. J. Del Jesus, "Some relationships between fuzzy and random classifiers and models," *Int. J. Approx. Reason.*, vol. 29, pp. 175–213, 2001.
[37] L. Sánchez and I. Couso, "Fuzzy random variables-based modeling with GA-P algorithms," in *Information, Uncertainty and Fusion*, R. Yager and B. Bouchon-Menier, Eds. Norwell, MA: Kluwer, 2000, pp. 245–256.
[38] L. Sánchez and I. Couso, "Advocating the use of imprecisely observed data in genetic fuzzy systems," in *Proc. GFS 2005*, Granada, Spain, 2005, pp. 124–129.
[39] L. Sánchez, I. Couso, and J. Casillas, "A multiobjective genetic fuzzy system with imprecise probability fitness for vague data," in *Proc. EFS*, Lancaster, U.K., 2006.
[40] L. Sánchez, J. Otero, and J. R. Villar, "Boosting of fuzzy models for high-dimensional imprecise datasets," in *Proc. IPMU 2006*, Paris, France, 2006.
[41] H. Tanaka, S. Uejima, and K. Asai, "Linear regression analysis with fuzzy model," *IEEE Trans. Syst., Man, Cybern.*, vol. 12, no. SMC-6, pp. 903–907, 1982.
[42] L. A. Zadeh, "Soft computing and fuzzy logic," *IEEE Software*, pp. 48–56, 1994.



Luciano Sanchez (M'07) received the electrical engineering degree and the Ph.D. degree from Oviedo University, Spain, in 1991 and 1994, respectively.

He is currently an Associate Professor of Computer Science at Oviedo University. In the summers of 1995 and 1996, he was a Visiting Professor at the University of California at Berkeley and at General Electric CRD, Schenectady, NY. His research interests include genetic fuzzy systems and the processing of imprecise data in machine learning problems.

Ines Couso received the Ms.C. degree in mathematics and the Ph.D. degree from Oviedo University, Spain, in 1995 and 1999, respectively.

She is currently an Associate Professor of statistics at Oviedo University. Her research interests include imprecise probabilities and fuzzy statistics. She is a member of the editorial board of *International Journal of Approximate Reasoning.*