Modeling Vague Data with Genetic Fuzzy Systems under a Combination of Crisp and Imprecise Criteria

Luciano Sánchez, Inés Couso and Jorge Casillas

***Abstract*— Multicriteria genetic algorithms can produce fuzzy models with a good balance between their precision and their complexity. The accuracy of a model is usually measured by the mean squared error of its residual. When vague training data is used, the residual becomes a fuzzy number, and it is needed to optimize a combination of crisp and fuzzy objetives in order to learn balanced models. In this paper, we will extend the NSGA-II algorithm to this last case, and test it over a practical problem of causal modeling in marketing. Different setups of this algorithm are compared, and it is shown that the algorithm proposed here is able to improve the generalization properties of those models obtained from the defuzzified training data.**

I. Introduction

There are many practical problems requiring to learn models from uncertain data. The most studied problem is that of the additive random noise, but many scenarios are well known not to match a stochastic model. Think in coarse-grained digital data (for example, weighing small objects in a low resolution scale) or in data items comprising both a numerical measure and a confidence interval defining its imprecision (the position given by a GPS sensor, say.) In these last cases, there is an unknown difference between the true measure and the observed one, but assuming that this difference is stochastic noise is an oversimplification. The scale would be a good candidate for a random-sets based solution, while the GPS output matches well with a model based on fuzzy random variables (frv), for instance.

In both the random sets and the frv-based models we assume the observation error follows certain probability distribution, but there is not information enough to infer this distribution from the observed data. Our knowledge allows us to determine a family of probability distributions that contain it, though. This means, in practice, that the output of so called *imprecise statistics* related models are intervals or fuzzy numbers that contain the true outcome. This vision complements the common practice in Genetic Fuzzy Systems, where fuzzy sets are associated to words and used to model vague linguistic assertions, but the models that depend on these words are fed with crisp data and produce crisp results anyway. In previous works [1] we advocated the use of fuzzy data to learn and evaluate Genetic Fuzzy Systems, and raised the use of fuzzy-valued fitness functions to formulate that kind of problems.

L. Sánchez is with the Computer Science Department of the Oviedo University, Spain. luciano@uniovi.es. Inés Couso is with Statistics Department, Oviedo University, Spain. couso@uniovi.es. Jorge Casillas is with the Computer Science Department, Granada University, Spain. casillas@decsai.ugr.es.

The use of a fuzzy-valued fitness function poses some numerical problems, some of which have been already addressed [2][3][4][5]. Currently, we are interested in extending the use of fuzzy data to learn rule based models with balanced accuracy and linguistic understandability. On the one hand, when crisp data is used, this kind of models can be obtained with multicriteria genetic algorithms [6]. On the other hand, when using imprecise data, the accuracies of the models become fuzzy numbers, thus it is needed to optimize a combination of crisp and fuzzy objetives in order to learn them. In this paper, we will extend the NSGA-II algorithm [7] to this case, and evaluate it over a practical problem of causal modeling in marketing.

This work is organized as follows: In Section II, the imprecise fitness function we need to optimize is described. In Section III we will detail the differences between the NSGA-II algorithm and our own extension of it, and in Section IV the practical problem that we will use to assess the method is described. In Section V the experimental results are discussed, and Section VI concludes the paper.

II. Measuring the quality of a model

In the first place, we will define the squared error of a model for fuzzy training data. Let Ω be a population of objects $\omega \in \Omega$, and let also $\tilde{X}(\omega)$, $\tilde{Y}(\omega)$ be fuzzy observations of certain atttributes of these objects. The training data will be a list of pairs $\{(\tilde{X}(\omega_1),\tilde{Y}(\omega_1)),\ldots,(\tilde{X}(\omega_n),\tilde{Y}(\omega_n))\}$ with $\omega_i \in \Omega$. We are given a rule-based model $\tilde{M}$, whose output $\tilde{M}(\tilde{X}(\omega))$ is computed by means of certain fuzzy logic-based inference algorithm. $\tilde{M}(\tilde{X}(\omega))$ is intended to approximate $\tilde{Y}(\omega)$.

Before defining the quality of this approximation, it is remarked that the fuzzy logic-based inference used to obtain $\tilde{M}(\tilde{X}(\omega))$ must fulfill some properties. For the output of the model to be a family of confidence intervals [1], it must happen that (1) $\tilde{M}(\mathbf{x})$ is crisp when $\mathbf{x}$ is a crisp vector, and (2), the α-cuts of the output are

$$[\tilde{M}(\tilde{X})]_\alpha = \{\tilde{M}(\mathbf{x}) \mid \mathbf{x} \in \tilde{X}_\alpha\}. \quad (1)$$

Observe that the cylindric extension of the input, followed by the intersection with the fuzzy graph of the model, projection over the output space and defuzzification of the result fulfills the first but not necessarily the second condition.

A. Definition of the Fuzzy Mean Squared Error (FMSE)

Let us define a new variable $\tilde{D} = \tilde{Y} - \tilde{M}(\tilde{X})$, the fuzzy residual of the model. The expectation of the squared residual $E(\tilde{D}^2)$ will be named Fuzzy Mean Squared Error (FMSE)

and generalizes the Mean Squared Error. We will use the FMSE to measure the accuracy of a model.

The definition of the squared error $E(\tilde{D}^2)$ resembles that of the variance $E(\tilde{D}-E(\tilde{D}))^2$ of a frv. This is a well studied case. There are three different definitions of the variance of a frv [8] that could be adapted to our purposes. The first type states that the variance of a frv is a crisp number [9][10]. The second definition [11] produces a fuzzy number, and the third one [8] is an interval. We have stated in [2] that the first type is not compatible with our nested confidence intervals interpretation, and also that we prefer the second option for the sake of simplicity.

This last definition of variance (or squared error) is based on a possibilistic interpretation of a fuzzy model, which has been shown to be the same interpretation as the nested confidence intervals [12]. Let us consider that the membership of the output of the model at a point t, $\tilde{M}(\tilde{X}(\omega))(t)$, is the conditional possibility of t being the true output, given $\tilde{X}$:

$$\tilde{M}(\tilde{X}(\omega))(t) = \Pi_{\tilde{M}(\omega)}(t) = \Pi(t|\omega). \tag{2}$$

The probability distribution of the residual is then dominated by the possibility distribution

$$\Pi_{\tilde{D}(\omega)}(t) = (\tilde{M}(\tilde{X}(\omega)) \ominus \tilde{Y}(\omega))(t), \tag{3}$$

and the expectation of the square of $\tilde{D}(\omega)$ is the set of all the expectations arising from the probabilities that are compatible with eq. (3):

$$\begin{aligned} \text{FMSE } (\tilde{M}) = \{ & \textstyle\sum t^2 P(t) : \\ & P \text{ is a probability distribution on } T \\ & \text{and } P(t) \leq \textstyle\sum_{\omega\in\Omega} P(\omega)\Pi_{\tilde{D}(\omega)}(t) \\ & \text{for all } t \in T\}. \end{aligned} \tag{4}$$

where T is the set of all the possible residual values of the model.

B. Numerical estimation of the FMSE

The same procedure described in [13] to estimate the variance of a frv can be applied to estimate the FMSE from the training set. For each of the membership functions $\tilde{D}(\omega_1), \tilde{D}(\omega_2), \ldots . \tilde{D}(\omega_n)$, we compute the fuzzy set $\tilde{D}^2$, whose α-cuts are $\tilde{D}^2_\alpha = \{x^2 | x \in \tilde{D}_\alpha\}$. Since the function x^2 is not locally monotonic, to evaluate the image of a fuzzy set 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 1. If the membership of $\tilde{D}$ 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. After we have computed all of the $\tilde{D}^2(\omega_i)$, we compute the FMSE of each one of them. Let L_i be the left profile of $\tilde{D}^2(\omega_i)$, and R_i the right one. Then,

$$\text{FMSE}_i = \left[\int_0^1 L_i d\alpha, \int_0^1 R_i d\alpha \right] \tag{5}$$

and the FMSE of the whole model is

$$\text{FMSE} = \frac{1}{n} (\text{FMSE}_1 \oplus \ldots \oplus \text{FMSE}_n) \tag{6}$$

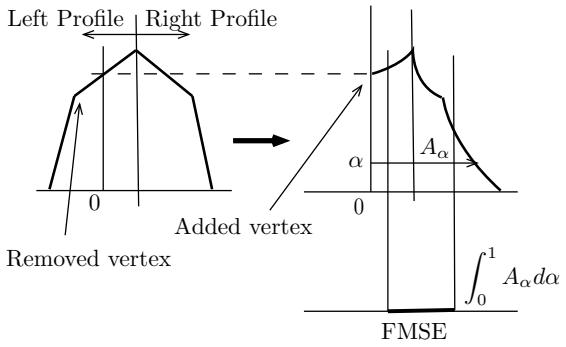

Fig. 1. The same extension to nonmonotonic functions of the profile method that is used to compute the sample variance of a frv can be applied to obtain its FMSE.

III. An extension of the NSGA-II algorithm

In this section, the NSGA-II algorithm [7] will be extended so that it can find a set of nondominated solutions for a two-objective problem, where one of the objectives is crisp (the complexity of the fuzzy rule base) and the other one is the FMSE of the candidate model, as defined in the preceding section, which is an interval. We want all of our extensions to reduce to the original formulation when the input data is crisp, and that they are also compatible with the possibilistic interpretation of the output of the model introduced before.

There are only three modules in the NSGA-II algorithm where the fitness function is assumed to be a vector of crisp numbers: the precedence (dominance) operator, the non-dominated sorting of the individuals, and the crowding distance. Therefore, our extension consists in alternate definitions for these modules.

A. Precedence in imprecise fitness-based Genetic Algorithms

Let us admit that a fuzzy model depends on a vector x of parameters and has a fitness value θ. θ is unknown except for a possibility distribution given by the fuzzy fitness $\tilde{F}_x$ (in this paper, $\tilde{F}_x$ will be the FMSE of the model), thus we know that

$$\Pi(\theta|\mathbf{x}) = \tilde{F}_x(\theta). \tag{7}$$

To determine whether one individual precedes another, it is needed to set up a procedure that, given two imprecise observations $\tilde{F}_{x_1}$ and $\tilde{F}_{x_2}$ of two unknown fitness values θ_1 and θ_2, estimates whether the probability of $\theta_1 < \theta_2$ is greater than that of $\theta_1 \geq \theta_2$, thus $\tilde{F}_{x_1} \prec \tilde{F}_{x_2}$. In this sense, the criteria that we pursue can be regarded as a special case of fuzzy ranking. However, we also want to find those cases where there is not statistical evidence in $\tilde{F}_{x_1}$ and $\tilde{F}_{x_2}$ that makes us to prefer one of them (thus $\tilde{F}_{x_1} \parallel \tilde{F}_{x_2}$.)

If a joint probability $P((\theta_1, \theta_2)|(x_1, x_2))$ were known, comparing two individuals would be an statistical decision problem. For instance, we could use

$$\frac{P(\{(\theta_1, \theta_2) : \theta_1 < \theta_2\}}{P(\{(\theta_1, \theta_2) : \theta_1 \geq \theta_2\}} <> 1. \tag{8}$$

Unfortunately, as we have mentioned, the imprecise fitness provides us with less information than this probability distri-

bution, because it is only an upper probability, that dominates the posterior probability of the crisp fitness.

In other words, given an individual x, the information we have about its fitness takes a value θ is limited to

$$\widetilde{F}_x(\theta) = P^*(\theta|x) \geq P(\theta|x) \tag{9}$$

thus the decision rule becomes

$$\frac{P_*(\{(\theta_1,\theta_2) : \theta_1 < \theta_2\}|x)}{P^*(\{(\theta_1,\theta_2) : \theta_1 \geq \theta_2\}|x)} > 1. \tag{10}$$

We also know that $P^*(\cdot|x)$ is a possibility for all x, thus we can state that

$$P^*(A|x) = \max\{P^*(\theta|x) : \theta \in A\}. \tag{11}$$

Since $P^*(A) = 1 - P_*(A^c)$, the expression 10 is reduced to

$$P^*(\{(\theta_1,\theta_2) : \theta_1 \geq \theta_2\}|x) < 1/2 \tag{12}$$

thus to decide whether $\tilde{F}_{x_1} \prec \tilde{F}_{x_2}$ we check that

$$\begin{aligned} &\max\{P^*(\theta|x_1) \cdot P^*(\theta|x_2) : \theta_1 < \theta_2\} \geq 1/2 \\ &\max\{P^*(\theta|x_1) \cdot P^*(\theta|x_2) : \theta_1 \geq \theta_2\} < 1/2 \end{aligned} \tag{13}$$

with $P^*(\theta|x_1) = \widetilde{F}_{x_1}$ and $\widetilde{P}^*(\theta|x_2) = F_{x_2}$.

It is remarked that, in this last case, rejecting that $\theta_1 < \theta_2$ does not imply that $\theta_1 \geq \theta_2$, because it may happen that eq. (12) is higher than 1/2 and also that $P^*(\{(\theta_1,\theta_2) : \theta_1 < \theta_2\}|x) \geq 1/2$, and then we must conclude that the fuzzy memberships $\tilde{F}_{x_1}$ and $\tilde{F}_{x_2}$ do not contain information enough to appreciate significant differences between them. In particular, this will always happen when $\tilde{F}_{x_1}$ and $\tilde{F}_{x_2}$ are non-disjoint intervals. It is remarked that the application of eq. (10) for interval-valued fitness is numerically the same as the so called *strong dominance* proposed in [5].

B. Introduction of a probabilistic prior

The inability to distinguish between intervals with non empty intersection is a major problem. We can improve the situation by introducing prior knowledge about the probability distribution of the fitness. Should we admit the profile likelihood rule [14], i.e. that the normalized likelihood function is

$$L(\theta) = \frac{P(x|\theta)}{\max\{P(x|\theta)\}} = P^*(\theta|x) \tag{14}$$

and also an uniform prior distribution, applying the Bayes rule we obtain that

$$P(A|x) = \frac{\sum_{\theta\in A} \widetilde{F}_x(\theta)}{\sum_{\theta\in\Theta} \widetilde{F}_x(\theta)}, \tag{15}$$

which in turn produces the decision rule

$$\frac{\sum_{\theta_1<\theta_2} \widetilde{F}_{x_1}(\theta_1)\widetilde{F}_{x_2}(\theta_2)}{\sum_{\theta_1,\theta_2\in\Theta} \widetilde{F}_{x_1}(\theta_1)\widetilde{F}_{x_2}(\theta_2)} > 1. \tag{16}$$

Example 1: Let $\Theta = [0,5]$, and let $F_{x_1} = [1,3]$ and $F_{x_2} = [2,4]$ two non disjoint intervals. Applying the rule 16 we obtain

$$\frac{7/8}{1/8} > 1$$

thus we can state that $F_{x_1} \prec F_{x_2}$.

Example 2: Let $\Theta = [0,5]$, and let $F_{x_1} = [1,5]$ y $F_{x_2} = [1.9,4]$ two non disjoint intervals. The application of 16 produces

$$\frac{0.4875}{0.5125} < 1$$

therefore $F_{x_2} \prec F_{x_1}$.

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 can not be compared.

It is also remarked that the Bayes rule with an uniform prior produces, in the particular case that $\tilde{F}$ is an interval, the same criteria proposed in [4]. Recently [15], a fuzzy ranking also similar to this one has been proposed, although with a different theoretical foundation.

C. Introduction of an imprecise prior

The use of a probabilistic prior is not compatible with the possibilistic interpretation. In the preceding subsection, we have recovered the posterior probability from an upper bound of itself and a probabilistic prior. This is 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 memberships are probabilities.

On the contrary, a combination of an upper bound of the posterior probability and an *imprecise* prior is meaningful. Following the proposal of [16], we can assume that the prior distribution is in certain family $\mathcal{P}$ of probabilities, and interpret that $\tilde{F}$ is an upper envelope of the set of posterior probability distributions that arise when the likelihood function is combined with every prior in the family:

$$P^*(A|x) = \sup_{P\in\mathcal{P}} \frac{\sum_{\theta\in A} L(\theta)P(\theta)}{\sum_{\theta\in\Theta} L(\theta)P(\theta)} \tag{17}$$

and $\tilde{F}_{x_1} \prec \tilde{F}_{x_2}$ when

$$P_*(\{(\theta_1,\theta_2) : \theta_1 < \theta_2\}|x) > 1/2. \tag{18}$$

The measure (17) is a plausibility. We only obtain a possibility in certain particular cases. According to [14], the most informative possibility distribution containing (17) and verifying some regularity conditions is

$$\pi_0(\theta) = \frac{\sum_{\varphi:L(\varphi)\leq L(\theta)} L(\varphi)}{\sum_{\varphi\in\Theta} L(\varphi)}. \tag{19}$$

We can make $\widetilde{F}_x(\theta) = \pi_0(\theta)$ and solve for L, and then substitute L and the prior in eq. (17). In case we use the FMSE this is immediate, because, when $\widetilde{F}_x$ is an interval, $L(\theta) = \pi_0(\theta) = \widetilde{F}_x(\theta)$.

The set of priors can range from the completely uninformative $P^*(\theta) = 1$ and $P_*(\theta) = 0$ (in this case this criterion reduces again to the strong dominance) 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

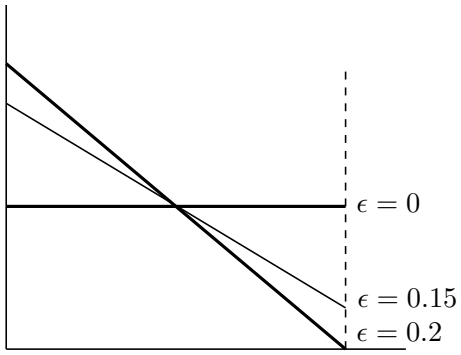

Fig. 2. Prior familiy of the example 3

to intervals with nonempty intersection, but not so restrictive as the uniform prior. In particular, we propose to use the family of priors that will be used in the example that follows:

Example 3: Let $\Theta = [0, 5]$ and let $F_{x_1} = [1, 3]$ and $F_{x_2} = [2, 4]$ two non disjoint intervals. We want to decide whether $F_{x_1} \prec F_{x_2}$, assuming the family of priors that includes all the probability distributions whose density is

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

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 in Figure 2. 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 the opposite, but we do not know how higher, thus any distribution between the uniform ($\epsilon = 0$) and the linear density with the maximum slope ($\epsilon = 1/m$) is reasonable.

The likelihoods are the same as the indicator functions,

$$L_1 = I_{[1,3]}, \quad L_2 = I_{[2,4]}$$

thus the equation 17 becomes

$$P_*(\theta_1 < \theta_2|x) = 1 - P^*(\theta_1 \geq \theta_2|x)$$

$$= 1 - \sup_{P\in\mathcal{P}} \frac{\int_{\theta_1\geq\theta_2} I_{[1,3]}(\theta_1)I_{[2,4]}(\theta_2)dP(\theta_1,\theta_2)}{\int I_{[1,3]}(\theta_1)I_{[2,4]}(\theta_2)dP(\theta_1,\theta_2)} =$$

$$= 1 - \sup_{\epsilon\in[0,0.2]} \frac{\int_2^3\int_2^x(\epsilon+\frac{1}{5}-\frac{2\epsilon}{5}x)(\epsilon+\frac{1}{5}-\frac{2\epsilon}{5}y)\,\mathrm{d}x\mathrm{d}y}{\int_1^3\int_2^4(\epsilon+\frac{1}{5}-\frac{2\epsilon}{5}x)(\epsilon+\frac{1}{5}-\frac{2\epsilon}{5}y)\,\mathrm{d}x\mathrm{d}y}$$

$$= 0.869$$

$$\text{and } P_*(\theta_1 \geq \theta_2|x) = 0.125$$

therefore we state that $\tilde{F}_{x_1} \prec \tilde{F}_{x_2}$.

Example 4: Let $\Theta = [0, 5]$, and let $F_{x_1} = [1, 5]$ and $F_{x_2} = [1.9, 4]$. We want to decide whether $\tilde{F}_{x_1} \prec \tilde{F}_{x_2}$, assuming that the family of priors of the preceding example is used.

In this case, the calculations produce the values

$$P_*(\theta_1 < \theta_2|x) = 0.332$$

$$P_*(\theta_1 \geq \theta_2|x) = 0.488$$

thus $\tilde{F}_{x_1} \parallel \tilde{F}_{x_2}$.

D. Non Dominated Sorting

The second module in the NSGA-II algorithm that must be replaced is the fast non dominated sorting. Observe that sorting the population with respect to an objective is the same as finding the best individual for this criterion; trivially, once this individual is found, we can remove it and recursively repeat the search to obtain an ordered population.

Finding the best individual is a procedure of statistical decision that generalizes the rules shown in the preceding section. Let us extend the rule in section III-C. Observe that we can bound the lower probability of the assert "the i-th individual has the best fitness in the population" by

$$M_i = P_*(\theta_i \text{ is the minimum}) = \prod_{j=1}^{s} P_*(\theta_i < \theta_j) \quad i \neq j.$$

We propose to admit that the best individual is that minimizing M_i. To sort the population, this best individual is removed in the first place and the remaining values of M_i are recalculated. Then we obtain the next element as the new minimum of M_i. The procedure is repeated until all the individuals have been extracted.

E. 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 $\theta_i < \theta_{i+1}$, the local density at the i-th individual is approximately

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

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:

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

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

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

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

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

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

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

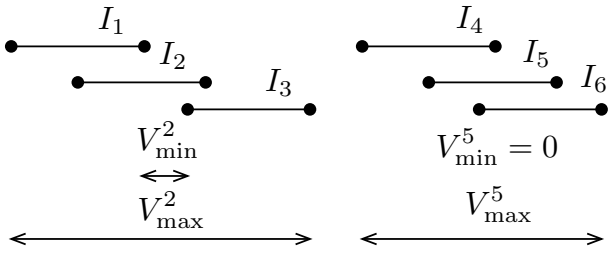

Fig. 3. Minimum and maximum crowding distances between interval-valued fitness functions. The maximum distance is the volume of the smallest interval that contains the fitness of three individuals, while the minimum distance is the volume of the interval that has non-null intersection with three individuals.

Therefore, the crowding distance associated to the i-th individual is (see Figure 3)

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

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.

We have solved the problem by introducing a metric between imprecise values of fitness that is not influenced by the nonspecificity of the measures. We wish that the crowding distance between two intervals $x \pm \epsilon$ and $y \pm \epsilon$ lays between the bounds mentioned before, does not depend on ϵ, and it is compatible with the crisp definition, i.e.

$$d(x \pm \epsilon, y \pm \epsilon) = |y - x|.$$

Many different metrics fulfilling these properties can be defined. We have chosen the Hausdorff distance. Given two sets A and B, this distance is defined as

$$d_H(A, B) = \max_{a \in A} \min_{b \in B} d(a, b)$$

which, when $A = [a_1, a_2]$ and $B = [b_1, b_2]$ are intervals reduces to

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

The crowding distance is 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 the order defined in the section III-D.

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

In section III-C we have explained how to implement the precedence between interval-valued fitness functions. Heterogeneous pairs, comprising an interval and an integer value each, must be compared eventually. The precedence between these pairs is as follows: for any two compound fitness values $(n_1, \tilde{F}_{x_1})$ and $(n_2, \tilde{F}_{x_1})$, we say that $(n_1, \tilde{F}_{x_1}) \preceq (n_2, \tilde{F}_{x_2})$ when $n_1 < n_2$ and $\tilde{F}_{x_1} \preceq \tilde{F}_{x_2}$ or $n_1 \leq n_2$ and $\tilde{F}_{x_1} \prec \tilde{F}_{x_2}$. $\tilde{F}_{x_1} \preceq \tilde{F}_{x_2}$ is defined as $(\tilde{F}_{x_1} \prec \tilde{F}_{x_2}) \vee (\tilde{F}_{x_1} \parallel \tilde{F}_{x_2})$.

IV. Practical Application: Causal Modeling in Marketing

In this section we briefly detail how the proposal in this paper can be implemented in a practical problem of causal modeling in marketing. In this respect, we focus this section on the modeling estimation techniques by introducing a knowledge extraction method that provides more quantity of qualitative information than preceding estimation techniques used in this field [17].

A. Acquisition of data

Data are obtained by means of a questionnaire. Specifically, in Table I we show a hypothetical example of the set of items that could have been used for measuring each one, while Table II shows an example of data available for this problem.

TABLE I
EXAMPLE OF A QUESTIONNAIRE (EXTRACTED FROM [18])

Fashion consciousness	
f_1:	Fashion is an important means of self-expression
f_2:	I'm usually the first among my friends to learn about a new brand or product
Conservatism	
c_1:	I tend to achieve my goals one step at a time
c_2:	I'm the type to deliberate things
c_3:	I gather various information and study well when deciding to buy a specific item
Hedonism	
h_1:	I want to enjoy the present rather than think about the future
h_2:	I like to go out to night-time entertainment spots
h_3:	I want to lead a life with lots of ups and downs

TABLE II
EXAMPLE OF FOUR RESPONSES ABOUT THE ITEMS SHOWN IN TABLE I

Fashion consciousness		*Conservatism*			*Hedonism*		
f_1	f_2	c_1	c_2	c_3	h_1	h_2	h_3
2	3	7	6	5	2	3	3
6	6	2	3	3	8	7	7
8	7	2	1	2	7	8	9
5	5	2	2	2	7	7	7

To work with this unusual kind of data, one could think on reducing the items of a specific variable to a single value, but we have adopted a more sophisticated process that allows us to take profit from the original format without any pre-processing stage: the consideration of fuzzy numbers to describe each variable, as described in Section IV-B.

B. Semantics of a Fuzzy Set

Under the imprecise probabilities framework, it makes sense to understand a fuzzy set as a set of tolerances, each one of them is assigned a confidence degree, being the lower degree the narrower tolerance [19]. In particular, the α-cuts

of the fuzzy set can be regarded as confidence intervals with degree $1-\alpha$ [12].

This representation allows us to codify the information contained in a set of numbers by means of a fuzzy set. This will be made clear with the example that follows. Let us suppose that a variable X has associated the items valued

$$X = \{2, 1, 3, 3, 2, 2, 4\}. \tag{20}$$

The most immediate calculation of a summary value is the sample mean, which is 2.429. While this is a good compromise value, we are discarding information that might be relevant: there are some items as low as 1, and others as high as 4. To gain additional insight about the importance of the dispersion of the values, we will assume that the set of items X is a sample of a larger population, whose mean is unknown. Given the sample X, we can calculate confidence intervals for the value of this mean, at different degrees.

A graphical representation of the membership function of $\widetilde{X}$ is shown in Figure 4. Observe that we can approximate it by a triangular membership function without incurring large errors. The same procedure must be applied to all lists of input and output values, to obtain a fuzzy dataset, from which we want to learn a model. This will be discussed in the next section.

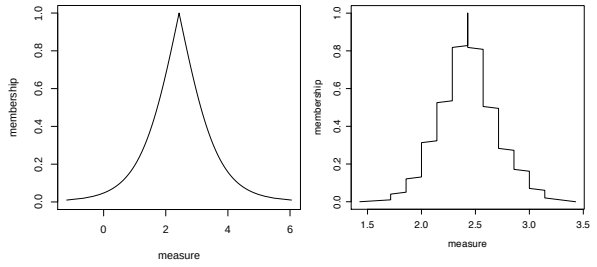

Fig. 4. Membership function of that set $\widetilde{X}$ that represents the sample X in Section IV-B. The left one was obtained under normality assumptions, and the right one is a basic bootstrap estimation.

C. Definition of the Genetic Fuzzy System

Once fixed the linguistic variables, a genetic fuzzy system is proposed in this section to automatically extract the knowledge existing in the considered fuzzy data. The obtained model should not be only accurate enough but also be easily legible, therefore we consider a multiobjective genetic fuzzy system, whose main components are described in the following sections.

1) Fuzzy Rule Structure: We opt by a compact description based on the disjunctive normal form (DNF) [20]:

IF X_1 is $\widehat{A_1}$ and ... and X_n is $\widehat{A_n}$ **THEN** Y is B

where each input variable X_i takes as a value a set of linguistic terms $\widehat{A_i} = \{A_{i1} \vee \ldots \vee A_{il_i}\}$, whose members are joined by a disjunctive (T-conorm) operator, whilst the output variable remains a usual linguistic variable with a single label associated. For instance, a fuzzy rule of the model given as example could be as follows:

IF *FashionConsciousness* is A_1 and *Conservatism* is A_2
THEN *Hedonism* is B.

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. The chromosome size is variable-length. 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. The "do not care" cases 111 and 000 are assigned membership 1 for all values.

3) Objective Functions: In addition to the FMSE, we also add an objective that intends to assess the linguistic complexity of the generated fuzzy rule set. We measure the number of rules of the fuzzy system $\mathcal{F}$ as $C_1(\mathcal{F})$. However, since each DNF-type fuzzy rule has also a complexity degree 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 fuzzy system $\mathcal{F}$, with l_{ri} being the number of linguistic terms used in the ith input variable of the rth DNF-type fuzzy rule (excluding all the "don't care" terms.) The joint objective is the product of both complexities.

4) Evolutionary Scheme: A generational approach with the multiobjective NSGA-II replacement strategy [7] is considered. Binary tournament selection based on the rank of each individual, depending on the Pareto-dominance relation defined in section III-F is used, with the crowding distance being used as a the secondary criterion for tiebreak. The precedence operator derives from the bayesian coherent inference with an imprecise prior, the dominated sorting is based on the product of the lower probabilities of precedence, and the crowding in based on the Hausdorff distance.

5) 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.

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

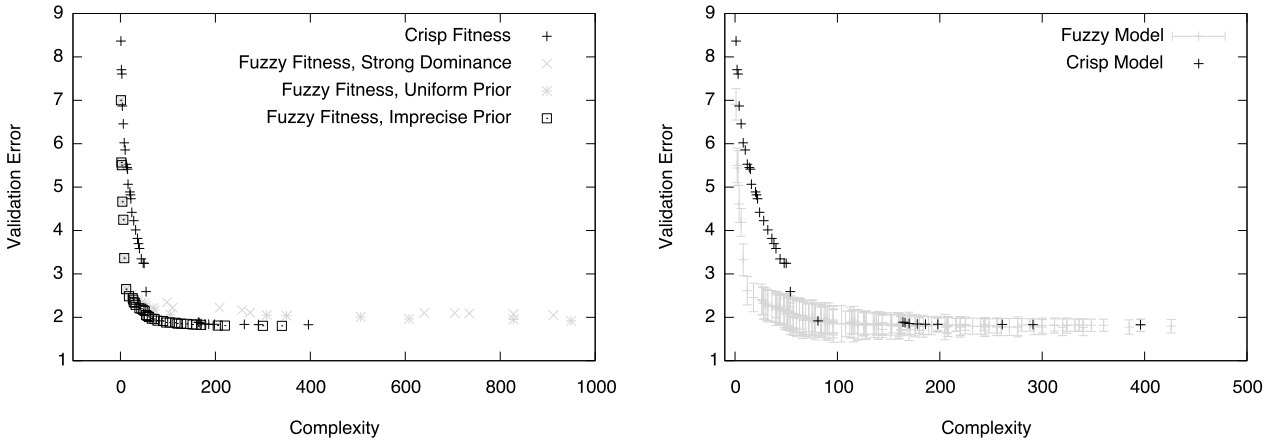

Fig. 5. Left part: validation error of the fuzzy models obtained with crisp and fuzzy fitness functions are shown. Three type of precedence operators have been evaluated in combination with our own definitions of dominated sorting and crowding distances: the strong dominance [5], the uniform prior [4] and the imprecise prior defined in this paper. Right part: FMSE and the validation errors of the crisp and the fuzzy model based on the imprecise prior.

of these mechanisms is made randomly among the available choices (e.g., contraction can not be applied if only a gene of the selected variable has the allele '1'). If an output variable is selected, the mutation operator simply increases or decreases the integer value.

V. Experimental Results

The consumer behavior model we have used for the experimentation is based on analyzing the consumer's *flow state* in interactive computer-mediated environments. Data have been obtained from the survey used in [21] to test a conceptual model previously presented by the same authors. We have adapted the original structural model proposed in that work by removing the least significant latent variable in each second-order variable. According to the partition performed by the authors, training data is composed by 1,154 examples (consumers' responses) and test data by 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.

We have run 10 times the proposed genetic fuzzy system, and compared its performance to that of a crisp error-based multiobjective approach in [22]. The resulting joint Pareto-fronts are displayed in Figure 5.

In the left part of the figure the validation error of the fuzzy models obtained with crisp and fuzzy fitness functions are shown. All these models have been validated over the crisp data used to learn the model in [22], but the fuzzy fitness-based have been trained over fuzzy data obtained with the bootstrap approximation mencioned in section IV-B. Three types of precedence operators have been evaluated along with our own definitions of dominated sorting and crowding distances: the strong dominance [5], the uniform prior [4] and the imprecise prior defined in this paper. Observe that the Pareto front obtained by the method proposed here dominates all other approaches. In the right part of the same figure, the FMSE and the validation errors of both the crisp and the fuzzy models are shown.

In the left part of Figure 6 a different measure of accuracy is used. The average of the differences between the output of these models and *all* the items in every output variable is computed, and the best, worse and mean test error in the ten repetitions is 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. In the right part of the figure, the comparison focuses in the three type of precedence operators evaluated in this paper. The differences between them begin to show in the latter generations, 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, but the imprecise prior gives more weight to the left part of the FMSEs, thus allowing the evolution to continue passed that 250th generation, where the uniform prior was stagnated. Observe that, while weighing more the left part of the fitness, the imprecise prior also causes that very similar FMSEs are indistinguishable, thus preventing the overfitting, as seen in the Pareto fronts in the preceding figure.

VI. Concluding Remarks

In this paper we have proposed an extension of the NSGA-II algorithm that is able to optimize a combination of crisp and interval-valued objectives. This extension was motivated in the need of optimizing the fitness function that arises when a fuzzy model is learned from vague data, and its accuracy is measured by mean of the FMSE. The FMSE is an interval of values arising from a possibilistic interpretation of the output of a fuzzy model.

To assess our algorithm, we have solved a marketing problem where the input data comprised multi-item examples. These multi-item examples were promoted to fuzzy sets by means of an interpretation of the membership function as a nested family of confidence intervals. We have shown, with

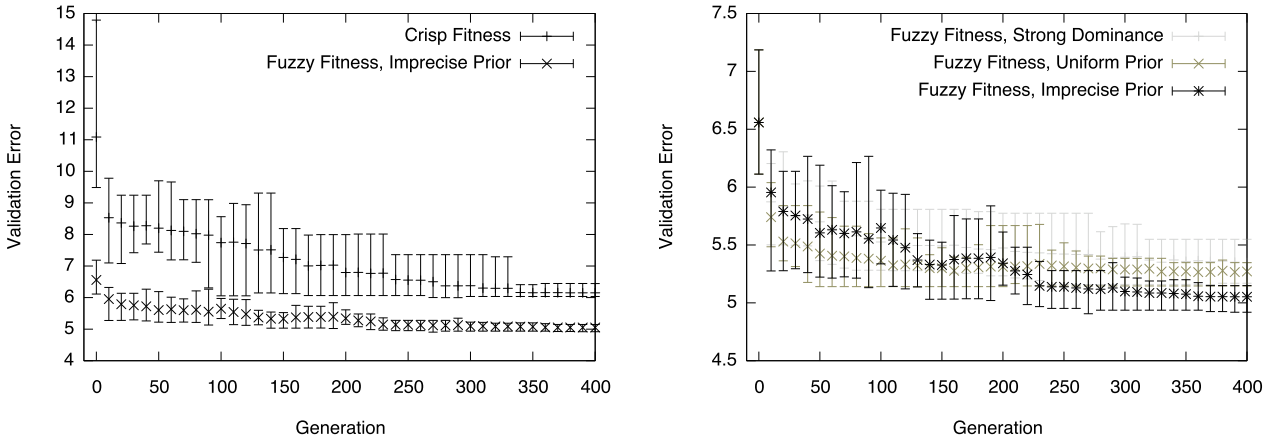

Fig. 6. Left part: The average of the differences between the output of crisp and fuzzy models and all the items in every output variable. Best, worse and mean test error are plotted. Right part: Comparison between the three type of precedence operators evaluated in this paper. The differences between them appear in the latter generations.

the help of a practical problem, that the models obtained by minimizing the FMSE are more robust than "state of the art" genetic fuzzy models, and are able to capture the dependence between imprecise data without the need of aggregating them or removing their fuzziness. From the multiobjective genetic algorithms point of view, we have also shown that the use of a coherent inference based precedence, depending on an imprecise prior, improves the convergence passed a point where both the strong dominance and the coherent inference with an uniform prior are stagnated. Work remains to be done in evaluating different test problems and other families of priors to model our knowledge about the probability distribution of the fitness function, along the evolution of the GA.

ACKNOWLEDGMENT

This work was supported by Spanish Ministry of Education and Science under grants TIN2005-08386-C05-01, TIN2005-08386-C05-05 and MTM2004-01269.

REFERENCES

[1] L. Sánchez and I. Couso, "Advocating the use of imprecisely observed data in genetic fuzzy systems," admitted for publication in IEEE Transactions on Fuzzy Systems.

[2] L. Sanchez, 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*, 2006, pp. 131–137.

[3] M. Koeppen, K. Franke, and B. Nickolay, "Fuzzy-Pareto-Dominance driven multiobjective genetic algorithm." in *Proc. 10th International Fuzzy Systems Assotiation World Congress (IFSA), Istanbul, Turkey*, 2003, pp. 450–453.

[4] J. Teich, "Pareto-front exploration with uncertain objectives." in *EMO*, 2001, pp. 314–328.

[5] P. Limbourg, "Multi-objective optimization of problems with epistemic uncertainty." in *EMO*, 2005, pp. 413–427.

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

[7] K. Deb, A. Pratap, S. Agarwal, and T. Meyarevian, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[8] I. Couso, S. Montes, and L. Sanchez, "Varianza de una variable aleatoria difusa. estudio de distintas definiciones," in *Proc XII Congreso Espanol sobre Tecnologias y Logica Fuzzy*, 2004.

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

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

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

[12] 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*, vol. 9, no. 2, pp. 249–262, 2001.

[13] 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*. IEEE, 2005, pp. 885–890.

[14] P. Walley and S. Moral, "Upper probabilities based only on the likelihood function," *J. Roy. Statist. Soc. Ser. B*, vol. 61, pp. 831–847, 1999.

[15] V. N. Huynh, Y. Nakamori, and J. Lawry, "Ranking Fuzzy Numbers Using Targets," in *Proc. of the 2006 International Conference in Information Processing and Management of Uncertainty in Knowledge-based Systems, Paris, France*, 2006, pp. 140–149.

[16] P. Walley, *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[17] G. V. Bruggen and B. Wierenga, "Broadening the perspective on marketing decision models," *International Journal of Research in Marketing*, vol. 17, pp. 159–168, 2000.

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

[19] I. R. Goodman, "Fuzzy sets as equivalence classes of possibility random sets," in *Fuzzy Sets and Possibility Theory: Recent Developments*, R. R. Yager, Ed. Pergamon, Oxford, 1982.

[20] A. González and R. Pérez, "Completeness and consistency conditions for learning fuzzy rules," *Fuzzy Sets and Systems*, vol. 96, no. 1, pp. 37–51, 1998.

[21] Y. Novak, D. Hoffman, and Y. Yung, "Measuring the customer experience in online environments: a structural modelling approach," *Marketing Science*, vol. 19, no. 1, pp. 22–42, 2000.

[22] J. Casillas, O. Delgado, and F. Martínez-López, "Predictive knowledge discovery by multiobjective genetic fuzzy systems for estimating consumer behavior models," in *Proceedings of the 4th European Conference in Fuzzy Logic and Technology*, Barcelona, Spain, 2005, pp. 272–278.