International Journal of Approximate Reasoning 49 (2008) 607–622

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Mutual information-based feature selection and partition design in fuzzy rule-based classifiers from vague data

Luciano Sánchez [a,*], M. Rosario Suárez [a], J.R. Villar [a], Inés Couso [b]

[a] Computer Science Department, University Oviedo, 33071 Gijón, Asturias, Spain
[b] Statistics Department, University Oviedo, 33071 Oviedo, Asturias, Spain

ARTICLE INFO

Article history:
Received 27 November 2007
Received in revised form 13 June 2008
Accepted 17 June 2008
Available online 12 July 2008

Keywords:
Genetic fuzzy systems
Feature selection
Vague data
Fuzzy fitness

ABSTRACT

Algorithms for preprocessing databases with incomplete and imprecise data are seldom studied. For the most part, we lack numerical tools to quantify the mutual information between fuzzy random variables. Therefore, these algorithms (discretization, instance selection, feature selection, etc.) have to use crisp estimations of the interdependency between continuous variables, whose application to vague datasets is arguable.

In particular, when we select features for being used in fuzzy rule-based classifiers, we often use a mutual information-based ranking of the relevance of inputs. But, either with crisp or fuzzy data, fuzzy rule-based systems route the input through a fuzzification interface. The fuzzification process may alter this ranking, as the partition of the input data does not need to be optimal. In our opinion, to discover the most important variables for a fuzzy rule-based system, we want to compute the mutual information between the fuzzified variables, and we should not assume that the ranking between the crisp variables is the best one.

In this paper we address these problems, and propose an extended definition of the mutual information between two fuzzified continuous variables. We also introduce a numerical algorithm for estimating the mutual information from a sample of vague data. We will show that this estimation can be included in a feature selection algorithm, and also that, in combination with a genetic optimization, the same definition can be used to obtain the most informative fuzzy partition for the data. Both applications will be exemplified with the help of some benchmark problems.

1. Introduction

Although fuzzy rule-based systems are intended for using vague data, most learning algorithms can only use precise information. Extracting fuzzy rules from imprecise examples is an open problem [11,12]. Initial works were based in a random sets-based representation, where each piece of data was described by a crisp value and a confidence interval defining its tolerance [23]. Conversely, recent works in fuzzy random variables prop up using a fuzzy representation when the data is known through more than one confidence interval [4]. In this respect, for quantifying the fitting between a model and data described in this manner, a fuzzy-valued measure of accuracy arises in a natural way [30].

Recent fuzzy rule learning algorithms also balance accuracy and linguistic quality [7]. When the data is vague, the linguistic quality can still be quantified by a real number [1], but the accuracy is fuzzy-valued. Therefore, the learning algorithm involves the joint optimization of a mix of crisp and fuzzy objectives. This last problem can be solved by means of multicriteria genetic algorithms [27] or metaheuristics [31]. In turn, both approaches are related to previous studies about the use of fuzzy fitness functions [18] and precedence operators between imprecise values [19,35].

* Corresponding author.
E-mail address: luciano@uniovi.es (L. Sánchez).

0888-613X/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2008.06.005

This use of a fuzzy-valued fitness function to obtain fuzzy rules from vague data defines a new branch of genetic fuzzy systems (GFS), as proposed in [30]. According to [12], there are four paradigms of GFS (Pittsburgh, Michigan, iterative rule learning (IRL) and genetic cooperative-competitive learning (GCCL)), and all these types are susceptible of being extended to vague data. A Pittsburgh approach for extracting fuzzy rules from interval and fuzzy-valued data in classification was proposed in [25,27] and applied to regression problems in [28]. In [26], backfitting and boosting (in the context of the IRL approach [15,24]) were extended to vague data, and in [32] a GCCL-type algorithm was introduced.

1.1. Preprocessing of vague data

However, the *preprocessing* of imprecise databases is seldom studied. There are many recent works dealing with feature selection procedures that use fuzzy techniques [14,34,36,37] or are designed to be used in combination with fuzzy systems [10,33,38,39], but we are not aware of any feature selection algorithms that can be applied to interval-valued or fuzzy data. In particular, to our knowledge, although there are some studies regarding the information a fuzzy model carries about crisp data [23,33], a definition of the mutual information between fuzzy random variables has not been explicitly proposed yet. In this work we will propose such a definition, based in a procedure defined by [16]. We have applied a similar method before [4,5], to analyze other properties of a fuzzy random variable.

Said definition would also solve a secondary problem: there is a loss of information in the linguistic discretization of the variables. Depending on the membership functions of the linguistic variables, this loss will be different for each input. Since we lack measures of mutual information between fuzzy data, the interdependency between variables is estimated before they are fuzzified, without taking into account the shape of the membership functions in the antecedents of the rules. But, it might happen that an apparently informative variable is rendered useless when it is rewritten in linguistical terms. We want to measure the amount of information that a variable carries *after* it passes through the fuzzification interface. In this paper we will show that the set of variables that a feature selection algorithm produces depends on these membership functions, and also that taking this factor into account causes significant improvements in the accuracy of fuzzy classifiers.

In particular, we will address a rather common situation in real world problems. We are faced with a mix of crisp and vague data, being represented as a fuzzy subset of a finite set of linguistic labels, which in turn are associated to a Ruspini fuzzy partition [22]. Let us illustrate this situation with the help of a numerical example. A fuzzification stage converts a crisp value of 45° into a fuzzy subset $\{0.0/\mathrm{COLD} + 0.2/\mathrm{WARM} + 0.8/\mathrm{HOT}\}$, say. Being elements a Ruspini's partition, the sum of the memberships of a crisp measurement is 1. Nonetheless, a vague measurement of the temperature could be represented by a fuzzy subset $\{0.1/\mathrm{COLD} + 0.3/\mathrm{WARM} + 0.9/\mathrm{HOT}\}$. A missing value, by the set $\{1/\mathrm{COLD} + 1/\mathrm{WARM} + 1/\mathrm{HOT}\}$. Note that the last two fuzzy subsets do not match any crisp value. We want to define a method that can process the three cases, which are of practical interest, but not often homogeneously studied in learning fuzzy classifiers.

Summarizing, in this work we will propose a new definition of the mutual information between fuzzy random variables, and a numerical algorithm for computing it from vague or fuzzified data. In addition, we will show that

- this mutual information can be optimized by means of a multi-objective genetic algorithm, and be used to find the fuzzy partition that carries the most information about the class of the object, thus providing us with the best linguistic partition of the data for a given number of terms, and
- it can be included in a filter type feature selection procedure, so it can take into account the shapes of the membership functions in the linguistic variables for choosing the most relevant features.

This paper is organized as follows: in the second section, we give a short introduction to the learning of fuzzy rules from vague data, and present the state of the art in the topic. In the third section, we introduce our definition of mutual information and detail how to estimate it from vague data. In the fourth section we will give some details about the genetic optimization of the mutual information, and the fifth section introduces an MIFS-like algorithm [2] that uses the new definition to select the most relevant features. The paper finishes with concluding remarks and future work.

2. Mutual information between a random variable and a fuzzy random variable

A fuzzy random variable can be regarded (see [3]) as a nested family of random sets, $(\Lambda_\alpha)_{\alpha\in(0,1)}$, each one associated to a confidence level $1-\alpha$. A random set is a mapping where the images of the outcomes of the random experiment are crisp sets. A random variable X is a selection of a random set Γ when the image of any outcome by X is contained in the image of the same outcome by Γ. This is to say, for a random variable $X:\Omega\to\mathbf{R}$ and a random set $\Gamma:\Omega\to\mathscr{P}(\mathbf{R})$, X is a selection of Γ (and we write $X\in S(\Gamma)$) when

$$X(\omega)\in\Gamma(\omega)\quad\text{for all }\omega\in\Omega. \tag{1}$$

In turn, a random set can be viewed as a family of random variables (its selections).
In previous works [29] we have defined the mutual information between a random variable X and a random set Γ as the set of all the values of mutual information between the variable X and each one of the selections of Γ:

$$\mathrm{MI}(X,\Gamma)=\{\mathrm{MI}(X,T)|T\in S(\Gamma)\}. \tag{2}$$

Generalizing this concept to fuzzy random variables is immediate, according to a general procedure proposed in [16]. We define the mutual information between a random variable X and a fuzzy random variable Λ as the fuzzy set defined by the membership function:

$$\widetilde{\mathrm{MI}}(X,\Lambda)(t) = \sup\{\alpha | t \in \mathrm{MI}(X,\Lambda_\alpha)\}. \tag{3}$$

2.1. Computer algorithm

In this section we show, by means of an example, how to estimate the mutual information between a fuzzy random variable and a crisp random variable.

Let us first suppose that we are given two paired samples $(X_1,X_2,\ldots,X_N)$ and $(Y_1,Y_2,\ldots,Y_N)$ from two (standard) random variables X and Y. We will assume that both universes of discourse are finite. Let $p_1,p_2,\ldots,p_n$ and $q_1,q_2,\ldots,q_m$ are the relative frequencies of the values of the samples of X and Y, respectively, and let $r_1,r_2,\ldots,r_s$ be the frequencies of the values of the joint sample $X\times Y$. The mutual information between the variables X and Y is estimated as follows:

$$\mathrm{MI}((X_1,\ldots,X_N),(Y_1,\ldots,Y_N)) = -\sum_{i=1}^{n} p_i \log p_i - \sum_{i=1}^{m} q_i \log q_i + \sum_{i=1}^{s} r_i \log r_i. \tag{4}$$

Let us now suppose that we are given two paired samples $(X_1,X_2,\ldots,X_N)$ and $(\Lambda_1,\Lambda_2,\ldots,\Lambda_N)$ of a crisp random variable X and a fuzzy random variable Λ.

We will estimate the mutual information between X and Λ by the fuzzy set

$$\widehat{\mathrm{MI}}((X_1,\ldots,X_N),(\Lambda_1,\ldots,\Lambda_N))(t) = \sup\{\alpha | t \in \{\mathrm{MI}((X_1,\ldots,X_N),(Y_1,\ldots,Y_N)) | (Y_1,\ldots,Y_N) \in S((\Lambda_1,\ldots,\Lambda_N)_\alpha,)\}\}. \tag{5}$$

Example. Consider the following samples of size 3 of the variables Λ and X:

Λ	X
$\{0.0/\mathrm{COLD}+0.2/\mathrm{WARM}+0.9/\mathrm{HOT}\}$	A
$\{0.4/\mathrm{COLD}+0.6/\mathrm{WARM}+0.0/\mathrm{HOT}\}$	B
$\{1.0/\mathrm{COLD}+0.0/\mathrm{WARM}+0.0/\mathrm{HOT}\}$	A

We want to estimate the mutual information between X and Λ. Firstly, we generate the set of samples $Y_1,\ldots,Y_4$ with non-null membership, which is computed as follows:

Y_1	X
WARM	A
COLD	B
COLD	A
$\mathrm{Membership}=\min\{0.2,0.4,1\}=0.2$	

Y_2	X
WARM	A
WARM	B
COLD	A
$\mathrm{Membership}=\min\{0.2,0.6,1\}=0.2$	

Y_3	X
HOT	A
COLD	B
COLD	A
Membership=$\min\{0.9, 0.4, 1\} = 0.4$	

Y_4	X
HOT	A
WARM	B
COLD	A
Membership=$\min\{0.9, 0.6, 1\} = 0.6$	

Now, we compute the estimates $\mathrm{MI}(Y_1, X), \ldots, \mathrm{MI}(Y_4, X)$:

MI	Membership
$\mathrm{MI}(Y_1, X) = 0.5441$	0.2
$\mathrm{MI}(Y_2, X) = 0.5441$	0.2
$\mathrm{MI}(Y_3, X) = 0.5441$	0.4
$\mathrm{MI}(Y_4, X) = 1.2108$	0.6

Lastly, we estimate the mutual information between Λ and X as the fuzzy set

$$\widehat{\mathrm{MI}} = 0.4/0.5441 + 0.6/1.2108,$$

defined by assigning to each value of MI its maximum membership.

Notice that the number of samples Y with non-null membership grows *i*th the number of labels raised to the volume of the sample. Enumerating all of them only is feasible in very small problems, therefore this definition only has theoretical interest. In the following sections we propose an alternative definition that is better suited for an approximate algorithm, that will be introduced later (see Section 2.4.)

2.2. Alternative interpretation of a fuzzy membership

The fuzzy representation we mentioned in the introduction can also be interpreted as a set of bounds for the probability of the result of the experiment [8]. For example, the fuzzy set $\{0.0/\mathrm{COLD} + 0.2/\mathrm{WARM} + 0.9/\mathrm{HOT}\}$ means that the probability of the temperature being 'COLD' is 0, the probability of 'WARM' is not greater than 0.2 and the probability of 'HOT' is not greater than 0.9.

The corresponding lower bounds are implicit. For instance, $p(\mathrm{WARM}) \geqslant 1 - (p^*(\mathrm{COLD}) + p^*(\mathrm{HOT})) = 0.1$. Observe that, with this interpretation, the set $\{1/\mathrm{COLD} + 1/\mathrm{WARM} + 1/\mathrm{HOT}\}$, mentioned in the introduction, represents the total absence of knowledge about the input value.

We can also use sets as $\{0.5/\mathrm{COLD} + 0.5/\mathrm{WARM} + 0.5/\mathrm{HOT}\}$, that does not signal a preference for either of the linguistic values, but states that their probabilities are not higher than 0.5. Observe also that the fuzzy set $\{0.0/\mathrm{COLD} + 0.2/\mathrm{WARM} + 0.8/\mathrm{HOT}\}$ provides us with precise information about the probability distribution, because $0.0 + 0.2 + 0.8 = 1$. This kind of fuzzy sets occur when a precise numerical value is passed though a fuzzification interface based on a Ruspini partition. Lastly, observe that a set like $\{0.0/\mathrm{COLD} + 0.2/\mathrm{WARM} + 0.4/\mathrm{HOT}\}$ (where $0.0 + 0.2 + 0.4 < 1$) can not be used with this interpretation.

2.3. Alternative definition of mutual information

Let us interpret the acceptability of a fuzzy random variable [16] as an upper bound of an otherwise unknown probability distribution p_Λ defined on the class of the random variables from Ω to **R**:

$$p^*_\Lambda(Y) = \sup\{\alpha | Y \in \Lambda_\alpha\}. \tag{6}$$

p_Λ induces a probability distribution on the values of the mutual information:

$$p(\mathrm{MI}(X,\Lambda) = t) = \sum_{Y|\mathrm{MI}(X,Y)=t} p_\Lambda(Y). \tag{7}$$

We can estimate upper and lower bounds of $p(\mathrm{MI}(X,\Lambda))$ from estimations of the bounds $p^*_\Lambda(Y)$ and $p_{\Lambda*}(Y)$, and estimate in turn the expected value of MI, as shown in the next subsection.

2.4. Computer algorithm for the alternative definition

Let us suppose that we are given two paired samples of X and Λ, as in the first algorithm in this section.

The probability of a sample of any crisp random variable Y is the product of all the probabilities of the asserts "Y_i is the true image of the experiment," under the model given by Λ_i:

$$p_\Lambda((Y_1, Y_2, \ldots, Y_N)) = \prod_{i=1}^{N} p_{\Lambda_i}(Y_i), \tag{8}$$

and the estimation of the mutual information is defined by the probability distribution

$$p(\widehat{\mathrm{MI}}((X_1,\ldots,X_N),(\Lambda_1,\ldots,\Lambda_N)) = t) = \sum\nolimits_{\mathrm{MI}((X_1,\ldots,X_N),(Y_1,\ldots,Y_N))=t} \mathrm{p}_\Lambda((Y_1,\ldots,Y_N))\}. \tag{9}$$

We can compute approximate bounds for this probability and for the expectation of MI, as shown in the next example.

Example. Suppose we are given samples of size 3 of the variables Λ and X:

Λ	X
$\{0.0/\mathrm{COLD} + 0.2/\mathrm{WARM} + 0.9/\mathrm{HOT}\}$	A
$\{0.4/\mathrm{COLD} + 0.6/\mathrm{WARM} + 0.0/\mathrm{HOT}\}$	B
$\{1.0/\mathrm{COLD} + 0.0/\mathrm{WARM} + 0.0/\mathrm{HOT}\}$	A

We wish to estimate the mutual information between X and Λ. Firstly, we enumerate the set of samples whose probability is not null, and compute bounds of these probabilities. Let $Y_1, \ldots, Y_4$ be these samples:

Y_1	X
WARM	A
COLD	B
COLD	A
$\mathrm{Probability}=[0.1, 0.2] \otimes 0.4 \otimes 1 = [0.04, 0.08]$	

Y_2	X
WARM	A
WARM	B
COLD	A
$\mathrm{Probability}=[0.1, 0.2] \otimes 0.6 \otimes 1 = [0.06, 0.12]$	

Y_3	X
HOT	A
COLD	B
COLD	A
Probability=$[0.8, 0.9] \otimes 0.4 \otimes 1 = [0.32, 0.36]$	

Y_4	X
HOT	A
WARM	B
COLD	A
Probability=$[0.8, 0.9] \otimes 0.6 \otimes 1 = [0.48, 0.54]$	

In the second step, we compute the mutual information $\mathrm{MI}(X, Y_1), \ldots, \mathrm{MI}(X, Y_4)$ of these samples:

MI	probability
$\mathrm{MI}(X, Y_1) = 0.5441$	[0.04,0.08]
$\mathrm{MI}(X, Y_2) = 0.5441$	[0.06,0.12]
$\mathrm{MI}(X, Y_3) = 0.5441$	[0.32,0.36]
$\mathrm{MI}(X, Y_4) = 1.2108$	[0.48,0.54].

In the last step, we estimate the mean value of the MI between Λ and X, which is the range of values of the expression:

$$\mathrm{E}(\widehat{\mathrm{MI}}) = p_1 * 0.5441 + p_2 * 1.2108,$$

subject to the constraints $p_1 + p_2 = 1$, $0.42 \leqslant p_1 \leqslant 0.56$, $0.48 \leqslant p_2 \leqslant 0.54$, therefore

$$\mathrm{E}(\widehat{\mathrm{MI}}) = [0.87, 0.89].$$

Since the number of samples with non-null probability is the same as the number of samples of non-null membership in Section 2.1, this algorithm still can not be applied to practical problems, but now we can select a small subsample and obtain an approximate solution. Let us suppose that our subsample comprises two elements:

MI	probability
$\mathrm{MI}(X, Y_2) = 0.5441$	[0.32,0.36]
$\mathrm{MI}(X, Y_4) = 1.2108$	[0.48,0.54].

The expectation of MI is the range of

$$\mathrm{E}(\widehat{\mathrm{MI}}) = \frac{q_1 * 0.5441 + q_2 * 1.2108}{q_1 + q_2}$$

constrained by $0.32 \leqslant q_1 \leqslant 0.36$, $0.48 \leqslant q_2 \leqslant 0.54$. This problem of non-linear optimization can be, in turn, too hard to be solved in a short time, thus we propose the following approximate solution:

(1) Firstly, we approximate the unknown mean with the centers of the probability intervals:

$$\mathrm{E}_1(\widehat{\mathrm{MI}}) = \frac{0.5441 * 0.34 + 1.2108 * 0.51}{0.34 + 0.51} = 0.9441.$$

(2) The upper bound of the probability is computed by assigning the upper probability to each sample whose MI is greater than the approximate mean, and the lower probability to the remaining ones:

$$E^*(\widehat{MI}) = \frac{0.5441 * 0.32 + 1.2108 * 0.54}{0.32 + 0.54} = 0.9627.$$

(3) The lower bound is computed with the reciprocal values:

$$E_*(\widehat{MI}) = \frac{0.5441 * 0.36 + 1.2108 * 0.48}{0.36 + 0.48} = 0.9251.$$

Therefore, the approximated value is

$$E(\widehat{MI}) = [0.9251, 0.9627].$$

It is worth mentioning that, following the interpretation described in Section 2.2, when all the fuzzified inputs originate from crisp numerical values, the algorithm in this section produces a crisp value. On the other hand, if there are some imprecise examples, this algorithm produces an interval.

3. Application I: estimation of the most informative fuzzy partition

The best fuzzy discretization of an input variable in a fuzzy rule-based system, from the point of view of the mutual information, is the one that maximizes the dependence between the fuzzified input and the output variable, i.e., the partition that loses the least information in the discretization. It is assumed that a rule learning algorithm that uses such a partition will produce the most accurate knowledge bases, as we will show later.

As an example, in Fig. 1 we have plotted the decision surfaces obtained by the fuzzy Adaboost algorithm [15], for the "Gauss" dataset (that will be described in Section 3.2) and different fuzzy partitions. The values produced by our estimation of the MI are displayed for each partition. Observe that, even though the error rate of the classifier is not being optimized by our procedure, the best classifier was indeed obtained by the most informative fuzzy partition.

Finding the fuzzy partition that loses the least information in crisp problems is a problem that can be solved with many numerical optimization algorithms, because the MI is also a crisp value. Yet, when the input data is vague or there are missing values in the dataset, the MI is an interval, as we have mentioned. In this case, obtaining the best fuzzy partition involves finding the minimum of an interval-valued function, which is not intuitive. We propose to use genetic algorithms to solve this problem, as we detail in the next subsection.

3.1. Genetic search of the most informative fuzzy partition for vague data

Finding the minimum of an interval-valued function is not feasible but in particular cases. For the most part, we can only compare intervals that do not intersect. For instance, imagine that we are given the list of values of MI {[1,3], [2,4], [5,7], [6,8]}. We know that the values [1,3] and [2,4] are not the best, but we can not decide whether [5,7] is better than [6,8].

In previous works [25], we have proposed that finding the minimum of an interval-valued function is a problem that can be assimilated to those addressed by multi-objective genetic algorithms (MOGA). A MOGA solves a problem that is similar to that of finding the set of minimal elements in a partial order. There are few changes that must be effected to a MOGA in order to solve those problems where the objective function is not defined by an array of real numbers, but a generic object like an interval or a fuzzy set. In this paper, we will be using an extension of the NSGA-II algorithm, as described in the preceding references. Our implementation is based on alternate precedence operators, and suitable algorithms for performing the non-dominated sorting and also for computing the crowding distance [27].

3.1.1. Coding scheme and genetic operators

We are interested in fuzzification interfaces defined by Ruspini's partitions, as mentioned. Furthermore, we will restrict ourselves to triangular membership functions and fixed-size linguistic partitions. In this particular case, a fuzzy partition comprising N linguistic terms can be codified with an array of N real numbers. A chromosome comprises so many partitions as input variables. For simplifying the genetic operators, each partition is represented by the minimum value of the variable, together with a list of positive values. These values are the distances between the coordinates of the modal points of the fuzzy sets, and those of their predecessors (see Fig. 2). We have used real coding and arithmetic crossover and mutation [20].

3.1.2. Fitness function

The training data is fuzzified according to the partition represented in the chromosome, and the mutual information between the class and the joint input is computed, using the numerical approximation of the MI defined in Section 2.4. A sensible subsample size is selected for limiting the execution time (between 1000 and 10,000 terms, depending on the problem at hand).

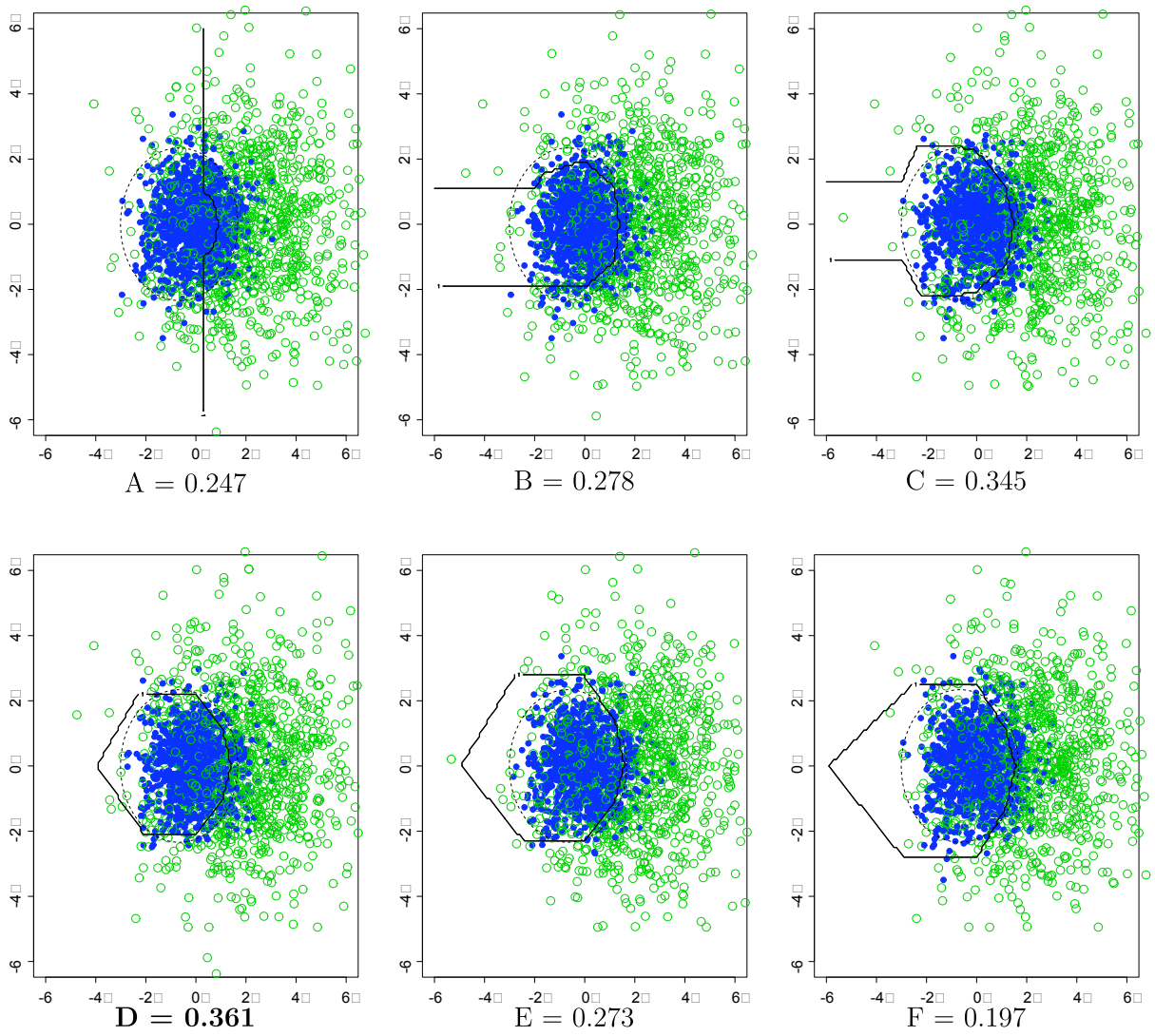

Fig. 1. Decision surfaces obtained by a learning algorithm (fuzzy Adaboost [15]) and expected values of MI for the same number of fuzzy rules and linguistic terms, but different, non-homogeneous, fuzzy partitions. The partition which keeps the highest amount of information, according to our estimator, is labeled 'D'. The dotted ellipse is the optimal Bayesian classifier for this problem.

3.1.3. Genetic scheme

As described in [25,27], we have used a generational approach with the multi-objective NSGA-II [9] replacement strategy, binary tournament selection based on rank and crowding distance, and a precedence operator that assumes a uniform prior. The non-dominated sorting depends on the product of the so-obtained probabilities of precedence. Lastly, the crowding is based on the Hausdorff distance.

3.2. Numerical results

Thirteen different fuzzy rule learning algorithms have been considered, both heuristic and genetic algorithms-based. In all cases, the number of linguistic terms in each partition is set beforehand, and not optimized by the learning algorithm. The experiments have been repeated ten times for different permutations of the datasets (10cv experimental setup). We have decided not to include the p-values of the statistical tests assessing the differences between the classifiers, but a selection of boxplots that show the relevance of the differences more visually.

The heuristic classifiers, according to the definitions in [13], use weighted fuzzy rules: the antecedent is always a conjunction of linguistic terms, and the consequent is the class mark. The weights of the rules are assigned as follows: always 1 (HEU1), same weight as the confidence (HEU2), differences between the confidences (HEU3, HEU4, HEU5), weights tuned by reward-punishment (REWP) and analytical learning (ANAL). Four of the genetic fuzzy classifiers are defined in the same

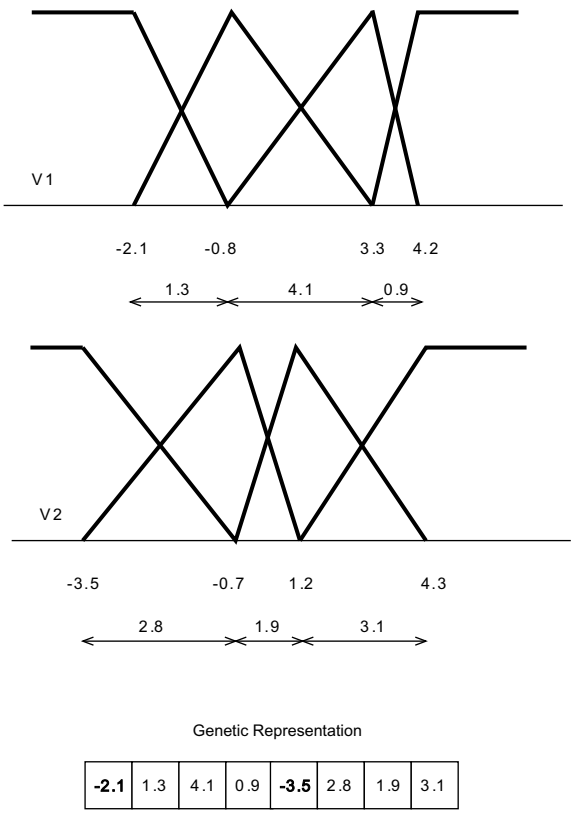

Fig. 2. Example of genetic representation of the fuzzy partitioning of the input space in a fuzzy classification problem. For a two-input problem with two linguistic terms for each variable, the chromosome length is 8. The leftmost point of each partition, plus the distances between the successive modal points, are represented.

reference [13]: genetic selection of rules taken from HEU3 (GENS), Michigan learning (MICH) – with population size 25 and 1000 generations, – Pittsburgh learning (PITT) – with population size 50, 25 rules each individual and 50 generations, – and hybrid learning (HYBR) – same parameters as PITT, macromutation with probability 0.8. These are based on the same kind of rules as the heuristic classifiers. To complete the study, two GFS of the Iterative Rule Learning type are added: fuzzy ababoost (ADAB) – less than 25 rules with a single consequent, fuzzy inference by sum of votes – [15] and Fuzzy Logitboost (LOGI) –l ess of 10 rules with multiple consequents, fuzzy inference by sum of votes – [21].

Eight crisp datasets taken from the UCI repository, and four imprecise datasets of our own have been used to assess the definition of the estimator and its use in the design of fuzzy partitions (see Tables 1 and 2). The imprecise datasets were designed for this paper, because we have not found similar problems in the literature. In the first place, we defined a fuzzy knowledge base comprising the following nine rules:

```
if X is SMALL and Y is SMALL then CLASS is A
if X is SMALL and Y is MEDIUM then CLASS is B
if X is SMALL and Y is LARGE then CLASS is A
if X is MEDIUM and Y is SMALL then CLASS is B
if X is MEDIUM and Y is MEDIUM then CLASS is A
if X is MEDIUM and Y is LARGE then CLASS is B
if X is LARGE and Y is SMALL then CLASS is A
if X is LARGE and Y is MEDIUM then CLASS is B
if X is LARGE and Y is LARGE then CLASS is A
```

The fuzzy partitions defining the meaning of the terms "SMALL", "MEDIUM" and "LARGE" are of the same type depicted in Fig. 2, with modal points $\{0.1, 0.2, 0.9\}$ and $\{0.1, 0.8, 0.9\}$ for the input variables X and Y, respectively. Secondly, we generated 1000 pairs of random values between 0 and 1, and built a noiseless dataset by joining each pair with its corresponding class, according to the fuzzy knowledge base defined before. Lastly, we corrupted the noiseless dataset, simulating the measurement of the input values with a digital scale that rounds to the first decimal, in different conditions. These conditions are:

Table 1
Test error of different fuzzy rule-based classifiers over uniform partitions and MI-optimized partitions

	HEU1	HEU2	HEU3	HEU4	HEU5	REWP	ANAL	GENS	MICH	PITT	HYBR	ADAB	LOGI
Iris uniform	**0.027**	**0.033**	0.060	0.067	0.067	0.047	**0.033**	0.067	0.047	0.060	**0.047**	0.047	**0.040**
Iris MI	0.040	0.040	**0.040**	**0.040**	**0.040**	**0.040**	0.040	**0.060**	**0.040**	**0.047**	0.060	**0.040**	0.047
Pima uniform	0.28	0.27	**0.25**	**0.25**	**0.25**	**0.26**	0.28	**0.26**	**0.35**	**0.28**	**0.27**	**0.25**	**0.23**
Pima MI	**0.26**	**0.25**	**0.25**	**0.25**	**0.25**	0.27	**0.27**	**0.26**	**0.35**	**0.28**	0.28	0.26	0.24
Gauss uniform	0.45	0.43	0.27	0.27	0.27	0.30	**0.20**	**0.21**	0.31	0.31	0.27	**0.21**	**0.20**
Gauss MI	**0.22**	**0.22**	**0.22**	**0.22**	**0.22**	**0.22**	0.23	0.22	**0.22**	**0.22**	**0.22**	0.23	0.22
Gauss-5 uniform	0.55	0.52	0.49	0.45	0.39	0.44	**0.31**	0.41	0.57	0.54	0.52	**0.32**	**0.32**
Gauss-5 MI	**0.33**	**0.33**	**0.33**	**0.33**	**0.32**	**0.33**	**0.31**	**0.32**	**0.32**	**0.32**	**0.32**	**0.32**	**0.32**
Glass uniform	0.38	0.37	0.37	0.36	0.35	0.37	0.37	0.36	0.49	0.37	0.43	0.34	**0.32**
Glass MI	**0.36**	**0.33**	**0.34**	**0.34**	**0.33**	**0.33**	**0.34**	**0.34**	**0.42**	**0.34**	**0.39**	**0.33**	0.36
Cancer uniform	0.040	0.039	0.037	0.037	0.037	0.087	0.081	0.046	**0.043**	0.077	**0.036**	0.205	0.033
Cancer MI	**0.030**	**0.031**	**0.031**	**0.031**	**0.031**	**0.039**	**0.040**	**0.029**	0.062	**0.037**	0.039	**0.102**	**0.027**
Skulls uniform	0.85	0.86	0.84	0.83	0.81	0.86	0.81	0.81	**0.83**	0.81	**0.81**	0.75	0.75
Skulls MI	**0.79**	**0.79**	**0.79**	**0.78**	**0.73**	**0.79**	**0.75**	**0.71**	0.84	**0.77**	0.84	**0.74**	**0.71**

Crisp data.

Table 2
Test error of different fuzzy rule-based classifiers over uniform partitions and MI-optimized partitions

	HEU1	HEU2	HEU3	HEU4	HEU5	REWP	ANAL	GENS	MICH	PITT	HYBR	ADAB	LOGI
Weight-c uniform	0.47	0.45	0.36	0.36	0.36	0.31	**0.29**	**0.29**	0.48	0.43	0.43	**0.12**	**0.20**
Weight-c MI	**0.30**	**0.30**	**0.30**	**0.30**	**0.30**	**0.29**	**0.29**	0.32	**0.30**	**0.30**	**0.30**	0.35	0.28
Weight-uc uniform	0.45	0.45	0.39	0.39	0.39	0.38	0.33	**0.29**	0.46	0.41	0.42	**0.24**	**0.23**
Weight-uc MI	**0.29**	**0.29**	**0.29**	**0.29**	**0.29**	**0.29**	**0.29**	0.30	**0.29**	**0.29**	**0.29**	0.39	0.29
Weight-2uc uniform	0.47	0.47	0.36	0.36	0.36	**0.31**	**0.34**	**0.26**	0.46	0.42	0.43	**0.21**	**0.20**
Weight-2uc MI	**0.34**	**0.34**	**0.34**	**0.34**	**0.34**	0.35	**0.34**	0.37	**0.36**	**0.37**	**0.37**	0.38	0.31
Weight-mv uniform	0.45	0.43	0.34	0.34	0.34	0.36	0.27	0.32	0.48	0.43	0.44	**0.31**	0.23
Weight-mv MI	**0.24**	**0.24**	**0.24**	**0.24**	**0.24**	**0.24**	**0.24**	**0.30**	**0.25**	**0.25**	**0.26**	**0.31**	**0.20**

Interval data.

(1) A well calibrated scale (dataset "weight-c": values between $x - 0.05$ and $x + 0.05$ are mapped to the value x).
(2) An uncalibrated scale ("weight-uc": values between $x - 0.01$ and $x + 0.09$ are mapped to the value x).
(3) A random selection between the preceding two scales ("weight-2uc").
(4) Precise inputs, but 5% of missing values at either coordinate are missing ("weight-mv").

The results in Tables 1 and 2 show the error rate of different fuzzy classifiers, when the fuzzy partition is uniform of size 3, and when the fuzzy partition is the result of a genetic optimization guided by our definition of Mutual Information. The improvements are almost universal, as expected. For crisp datasets, the theoretical problems Gauss and Gauss-5 were the most benefited from the optimized partitions. All classifiers, heuristic and GFS, are very near the optimum. Real-world problems achieved less unequivocal results, but the coherence of the measure of information is clear. In vague datasets, the gain is also very noticeable, but the true fuzzy partitions were not always found. It is remarked the positive influence of the optimization of the partition in the dataset "weight-mv", with missing values.

4. Application II: an MIFS-like feature selection algorithm for fuzzy rule learning algorithms

As we have mentioned in the introduction, the use of fuzzified data has theoretical advantages when selecting features to be used in fuzzy rule-based systems. An example is shown in Fig. 3. If we compute the mutual information between the class (black or white) and the input variables X and Y, we will obtain that the most informative variable is X. But, if our linguistic variables have the values 'LOW' and 'HIGH' with the membership functions shown in Fig. 3, most of the information about the class is lost for the variable $\widetilde{X}$. That is to say, the mutual information between the linguistic variable $\widetilde{X}$ and the class is lower than the mutual information between the linguistic variable $\widetilde{Y}$ and the class. The opposite happens with the crisp variables X and Y. In this example, an estimation of the mutual information that does not take the memberships of the linguistic terms 'LOW' and 'HIGH' into account would produce incorrect results.

When the input data is crisp, our estimator of the mutual information can be used in combination with any filter-type feature selection algorithm which is based on the mutual information, because our mutual information will also take crisp values. Otherwise (vague data or missing values) our estimation produces an interval and some modifications are needed. As an example, the pseudocode of the MIFS algorithm [2] is adapted as follows, so that it can use the interval-valued mutual information:

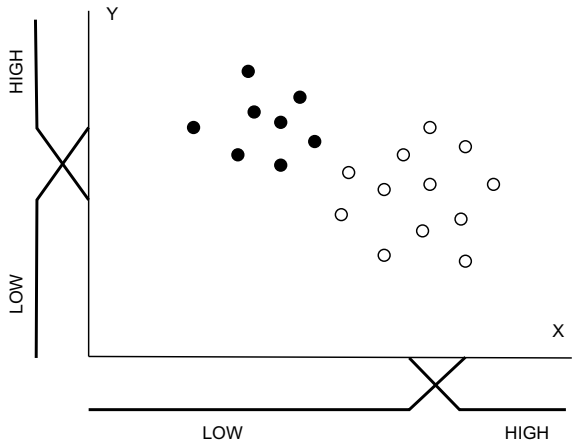

Fig. 3. Example of the theoretical advantages of the proposed estimator in the design of fuzzy rule-based systems. The mutual information between the variable X and the class (black or white) is higher than that of Y. However, choosing the variable X is the worst decision when designing a fuzzy rule-based classification system depending on the fuzzy variables $\tilde{X}$ and $\tilde{Y}$, which take the linguistic values "LOW" and "HIGH," whose memberships are shown in the figure. The estimator of the mutual information defined in this paper assigns a higher value to the variable $\tilde{Y}$, as desired.

$F =$ initial set of n features; $S = \{\emptyset\}$
For each feature $f \in F$ compute $\mathrm{MI}(f, C)$
Perform a non-dominated sorting of the values of MI
Select the first element and set $F = F \setminus \{f\}$, $S = S \cup \{f\}$
Repeat until $|S| = k$
  For all couples of values (f, s) with $f \in F$
    and $s \in S$, compute $\mathrm{MI}(f, s)$
  Perform a non-dominated sorting of the values
    $\mathrm{MI}(f, C) \ominus \beta \oplus_{s \in S} \mathrm{MI}(f, s)$
  Select the first element and set $F = F \setminus \{f\}$, $S = S \cup \{f\}$ Output the set S

This algorithm will be called 'FMIFS' in the remainder of the paper. The non-dominated sorting of the interval-valued estimation of the Mutual Information can be performed by any of the methods proposed in reference [27]. In particular, in this paper we have assumed an uniform prior distribution on the intervals $\mathrm{MI}(f, C) \ominus \beta \oplus_{s \in S} \mathrm{MI}(f, s)$, and sorted them according to their probability of containing the highest value. That is to say, for sorting a set of intervals we will define first how to compare two of them: Let two intervals be $\mathrm{I}_1 = [a, b]$ and $\mathrm{I}_2 = [c, d]$. If we assume that there are two random variables $X \to \mathscr{U}(a, b)$ and $Y \to \mathscr{U}(c, d)$, then we can define

$$p(I_i > I_2) = p(X > Y) = \int \int_{(a,b)\times(c,d)\cap\{(x,y):x>y\}} \frac{\mathrm{d}x\mathrm{d}y}{(b-a)(d-c)}. \tag{10}$$

Hence, if we are given n intervals $I_1, \ldots, I_n$, the probability of I_i containing the maximum value is

$$p_i = \prod_{\substack{i \neq j \\ 1 \leqslant i \leqslant n}} p(I_i > I_j), \tag{11}$$

and lastly we can sort the values of I_i in the same order as the values p_i.

4.1. Numerical results

We have selected, with different methods, the five most relevant features in five crisp datasets from UCI (German, Ion, Pima, Sonar and Wine), and used the reduced datasets to train the 13 fuzzy rule-based classifiers introduced in Section 3.2. None of the classifiers evolves the definition of the partitions, which are uniform and of size 3. The use of a coarse fuzzy partition is intended to show the advantages of our approach. Finer partitions are less prone to loss information, and FMIFS would tend to be the same as MIFS.

According to Table 3, the algorithm FMIFS was not different from the best one in 47 of the 65 cases. SSGA [6] was the best choice in 30, RELIEF [17] in 8 and the crisp version of MIFS was the best in six. Boxplots with the dispersion of the test error for all the problems are shown in Fig. 4. It can be seen that there are relevant differences for both genetic fuzzy systems and heuristic methods, and also that the improvement depends on the dataset. The gain is more evident in datasets as SONAR, with a high number of input variables. In datasets with a high dimension, there are potentially many subsets of variables

Table 3
10-Fold cross validation-based test error of different fuzzy rule-based classifiers after a feature selection was performed

	HEU1	HEU2	HEU3	HEU4	HEU5	REWP	ANAL	GENS	MICH	PITT	HYBR	ADAB	LOGI	best
GERMAN – RELIEF	0.295	0.285	0.275	0.275	0.275	0.280	0.275	0.270	**0.295**	0.285	0.295	0.290	0.260	1
GERMAN – SSGA	0.265	**0.255**	**0.250**	**0.255**	**0.255**	**0.250**	0.260	0.255	**0.295**	**0.275**	**0.255**	**0.260**	0.255	9
GERMAN – MIFS	0.280	0.265	0.265	0.265	0.265	0.265	0.260	0.265	**0.295**	**0.275**	0.285	0.265	**0.250**	3
GERMAN – FMIFS	**0.255**	**0.255**	0.255	**0.255**	**0.255**	0.260	**0.245**	**0.250**	0.305	**0.275**	**0.255**	0.265	0.270	8
ION – RELIEF	0.328	0.314	0.285	0.285	0.285	0.200	0.257	0.157	0.428	0.228	0.214	**0.114**	0.142	1
ION – SSGA	0.200	0.185	0.157	0.157	0.157	0.142	**0.157**	0.128	0.328	**0.114**	**0.114**	0.514	0.100	3
ION – MIFS	0.200	0.200	0.200	0.200	0.200	0.185	0.185	0.185	0.357	0.157	0.142	0.514	0.171	0
ION – FMIFS	**0.185**	**0.142**	**0.128**	**0.128**	**0.128**	**0.128**	0.171	**0.100**	**0.200**	**0.114**	0.128	0.514	**0.085**	10
PIMA – RELIEF	0.289	0.289	0.276	0.276	0.276	0.269	0.269	0.263	**0.355**	**0.230**	0.256	0.243	0.250	2
PIMA – SSGA	0.302	0.289	**0.263**	**0.263**	**0.263**	**0.263**	**0.263**	0.263	**0.355**	0.243	**0.243**	**0.217**	**0.217**	9
PIMA – MIFS	**0.276**	**0.276**	0.276	0.276	0.276	0.276	0.276	0.269	**0.355**	0.256	0.276	0.223	0.243	3
PIMA – FMIFS	0.302	0.289	**0.263**	**0.263**	**0.263**	**0.263**	**0.263**	**0.243**	**0.355**	0.250	0.276	**0.217**	**0.217**	9
SONAR – RELIEF	0.300	0.275	0.250	0.250	0.250	0.275	0.375	0.300	**0.300**	**0.275**	0.325	0.300	0.250	2
SONAR – SSGA	0.300	0.325	0.250	0.250	0.250	0.300	0.325	0.250	**0.300**	0.300	0.250	0.250	0.250	1
SONAR – MIFS	0.350	0.325	0.300	0.300	0.300	0.350	0.350	0.250	0.350	0.325	0.350	0.350	0.325	0
SONAR – FMIFS	**0.225**	**0.200**	**0.175**	**0.175**	**0.175**	**0.200**	**0.225**	**0.175**	**0.300**	**0.275**	**0.225**	**0.150**	**0.200**	13
WINE – RELIEF	0.500	0.411	0.235	0.205	0.176	0.088	0.235	**0.029**	0.647	0.205	**0.029**	0.058	0.058	2
WINE – SSGA	**0.176**	0.176	0.147	0.235	**0.147**	**0.058**	**0.088**	**0.147**	0.147	0.058	**0.029**	**0.000**	**0.029**	8
WINE – MIFS	0.323	0.323	0.264	0.205	0.176	0.117	0.235	0.176	0.617	0.058	0.176	0.058	0.058	0
WINE – FMIFS	**0.176**	**0.147**	**0.117**	**0.176**	**0.147**	**0.058**	0.147	0.117	0.176	**0.029**	0.088	0.058	0.058	7

The number of times each algorithm was the best is shown in the last column. FMIFS, SSGA, RELIEF and MIFS were the best 47, 30, 8 and 6 times, respectively. Five input variables and three linguistic labels were used for each variable.

with similar information about the class. Using information about the fuzzy partition allows us to further distinguish between them, and discard those that are more affected by a coarse partition.

We can conclude that the best overall selection of variables was produced by FMIFS, and also that the worst algorithm is MIFS, which incidentally uses the same algorithm as FMIFS, as mentioned. This result makes us to believe that generalizing a wrapper algorithm, like SSGA, so that it can use fuzzy data, is likely to improve the results shown here. It is worth pointing that we do not claim that the use of the fuzzy mutual information universally improves the performance of fuzzy classifiers, since it depends on the linguistic partition. We claim that there exist cases where the linguistic partition has to be taken into account, and that those cases are not pathological: we have used uniform partitions, which are the most common in practical situations. In case the fuzzy partition is optimal (for instance, if we used a partition optimized as shown in the first application described in this paper) the gain is no longer relevant.

To prove that the improvement is sound not only in trivial partitions, but also in the average case, we have conducted a new study on how often our method will produce significant improvements. We have generated 100 random fuzzy partitions of size three. For each of them, we have learned a heuristic fuzzy classifier (type HEUR2) and estimated its test error with 10-fold cross validation, for the following four scenarios:

(1) Feature selection with RELIEF.
(2) Feature selection with SSGA.
(3) Feature selection with MIFS.
(4) Feature selection with FMIFS.

The first three scenarios use the same set of features (since these algorithms do not depend on the linguistic partition) and FMIFS uses a custom set of variables for each partition. In Fig. 5 and Table 4 the histograms of the test error and the mean values are, respectively displayed. The FMIFS algorithm was significantly better in three of five datasets, not different than the best in one case and worse than the best in one case. The histograms in Fig. 5 make it clear that the mode of the distribution is skewed to the left in all cases but one, showing that the FMIFS algorithm and therefore our definition of fuzzy mutual information captures better the dependency between the variables than the crisp version.

5. Concluding remarks and future work

There are hardly any references to the preprocessing of databases with imprecise data in the literature. In this paper we have proposed a numerical algorithm to compute the degree of dependence between two fuzzy variables, and have shown how to apply it to the design of the fuzzification interface of a rule-based system and also to select the most relevant features when the input data is vague.

The results shown in the field of feature selection are preliminary, but promising. We have shown that there exist problems where we obtain a consistent improvement for the whole catalog of fuzzy systems that were tested, but we have also found problems for which the new algorithm produces similar results to the crisp version. Intuitively, the method proposed here should be applied in those situations exemplified in Fig. 3, but further work is needed to characterize this family of

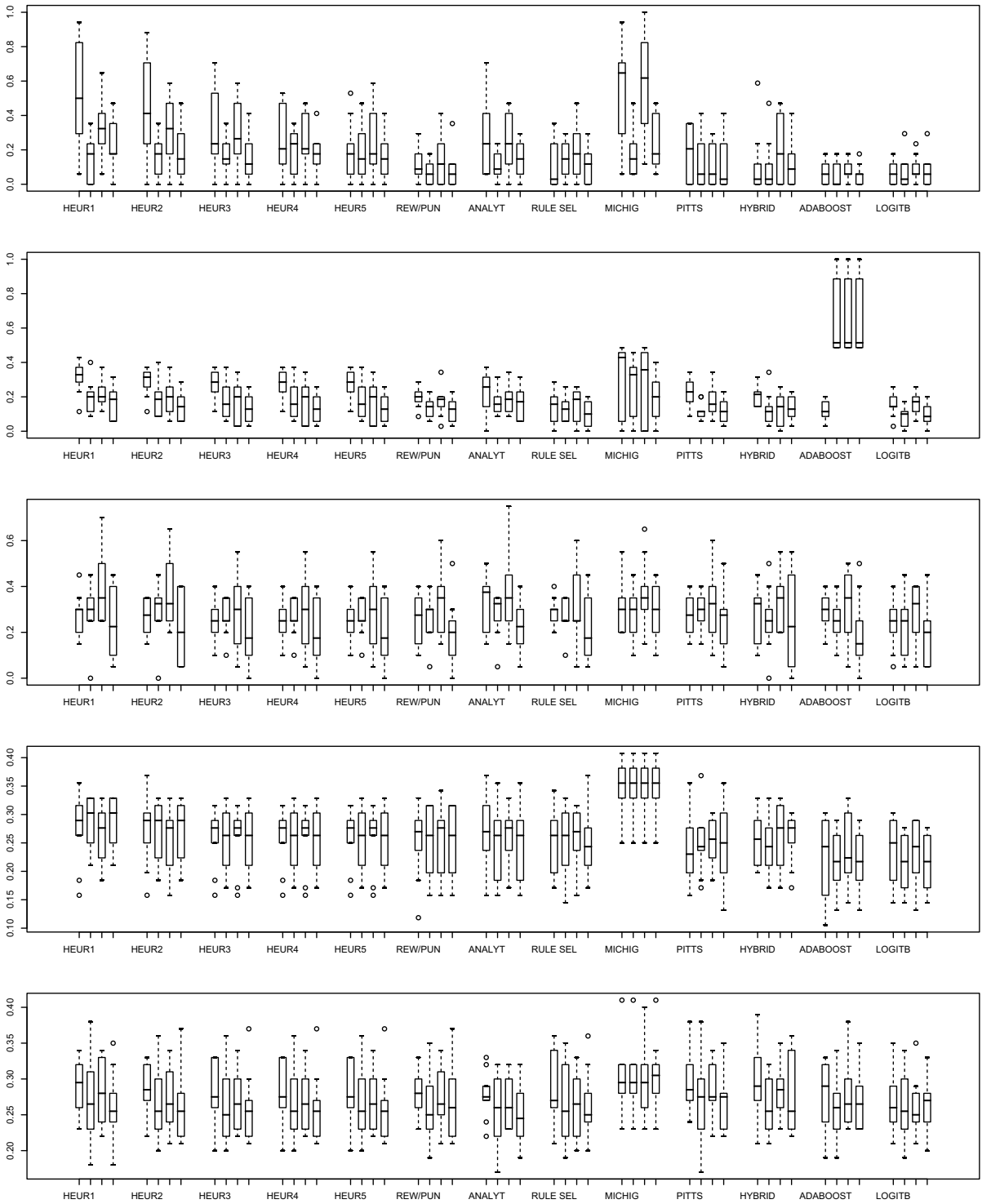

Fig. 4. Boxplots of the test errors of different fuzzy rule-based classifiers with the original MIFS algorithm and the modified version proposed in this paper. From top to bottom: WINE, ION, SONAR, PIMA, GERMAN datasets. The columns are displayed in the order RELIEF, SSGA, MIFS, FMIFS.

problems. Lastly, much work remains to be done to perform feature selection with vague data. A set of benchmark problems that include vague data is needed, and also some criteria to compare the efficiency of the new algorithms with that of the crisp ones over the new set of problems.

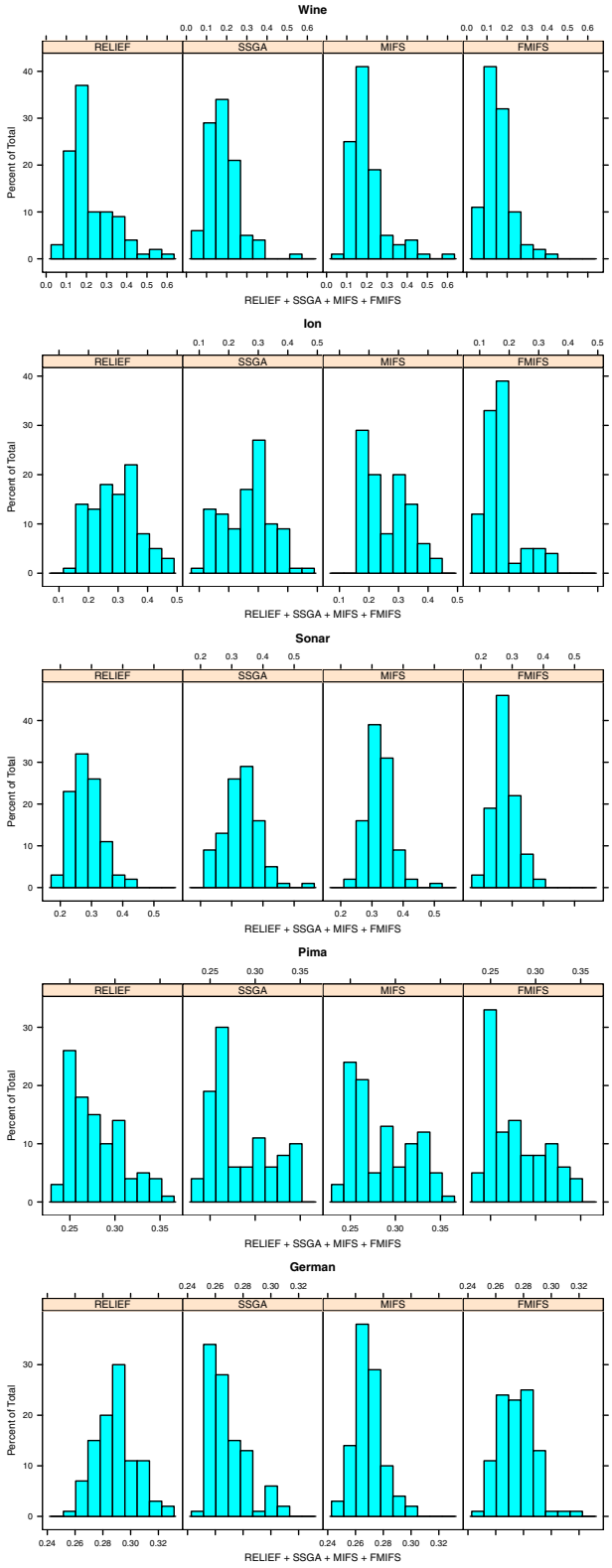

Fig. 5. Histogram of the test error, 1000 runs of HEUR2 (each sample is the mean test error in a 10-fold cross validation, which was repeated 100 times, with random Ruspini partitions of the input variables.) From top to bottom: WINE, ION, SONAR, PIMA and GERMAN datasets. Observe that there exist datasets (e.g. ION) for which the density function of the test error in FMIFS is skewed, showing a correlation between the ranking of the features and the membership functions of the input variables.

Table 4
Mean of the test error, 1000 runs of HEUR2 (10cv × 100 random partitions)

	RELIEF	SSGA	MIFS	FMIFS
GERMAN	0.290	**0.269**	**0.269**	0.275
ION	0.292	0.265	0.260	**0.171**
PIMA	0.279	0.283	0.284	**0.278**
WINE	0.217	0.180	0.199	**0.152**
SONAR	0.281	0.328	0.323	**0.276**

The FMIFS algorithm is significantly better in three of five datasets, not different than the best in one case and worse than the best in other case. We conclude that there is an overall benefit if the feature selection algorithm knows about the fuzzy partitions.

Acknowledgements

This work was funded by Spanish M. of Science and Technology and by FEDER funds, within the Grant Nos. TIN2005-08036-C05-05 and MTM2004-01269.

References

[1] J.M. Alonso, L. Magdalena, Equilibrio entre Interpretabilidad y Precisión en Sistemas Basados en Reglas Difusas: Nuevos retos, ESTYLF, in press (in Spanish).
[2] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Transactions on Neural Networks 5 (4) (1994) 537–550.
[3] C. Baudrit, I. Couso, D. Dubois, Joint propagation of probability and possibility in risk analysis: towards a formal framework, International Journal of Approximate Reasoning 45 (2007) 82–105.
[4] I. Couso, L. Sánchez, Higher order models for fuzzy random variables, Fuzzy Sets and Systems 159 (2008) 237–258.
[5] I. Couso, D. Dubois, S. Montes, and L. Sánchez, On various definitions of the variance of a fuzzy random variable, in: Fifth International Symposium on Imprecise Probabilities: Theory and Applications (ISIPTA 07) Prague (Czech Republic), 2007, pp. 135–144.
[6] J. Casillas, O. Cordón, M.J. del Jesus, F. Herrera, Genetic feature selection in a fuzzy rule-based classification system learning process, Information Sciences 136 (2001) 135–157.
[7] J. Casillas, O. Cordon, F. Herrera, L. Magdalena (Eds.), Interpretability Issues in Fuzzy Modeling, Springer, Heidelberg, Germany, 2003.
[8] D. Dubois, H. Prade, When upper probabilities are possibility measures, Fuzzy Sets and Systems 49 (1992) 65–74.
[9] K. Deb, A. Pratap, S. Agarwal, T. Meyarevian, A fast and elitist multi-objective genetic algorithm: NSGA-II, IEEE Transactions on Evolutionary Computation 6 (2) (2002) 182–197.
[10] F. Fernández-Riverola, F. Díaz, J.M. Corchado, Reducing the memory size of a fuzzy case-based reasoning system applying rough set techniques, IEEE Transactions SMC Part C 37 (1) (2007) 138–146.
[11] F. Herrera, Genetic fuzzy systems: status, critical considerations and future directions, International Journal of Computational Intelligence Research (2005) 59–67.
[12] F. Herrera, Genetic fuzzy systems: taxonomy, current research trends and prospects, Evolutionary Intelligence 1 (1) (2008) 27–46.
[13] H. Ishibuchi, T. Nakashima, M. Nii, Classification and Modeling with Linguistic Information Granules, Springer, 2004.
[14] R. Jensen, Q. Shen, Fuzzy-rough sets assisted attribute selection, IEEE Transactions on Fuzzy Systems 15 (1) (2007) 73–89.
[15] M.J.del. Jesus, F. Hoffmann, L. Junco, L. Sánchez, Induction of fuzzy rule-based classifiers with evolutionary boosting algorithms, IEEE Transactions on Fuzzy Sets and Systems 12 (3) (2004) 296–308.
[16] R. Kruse, K.D. Meyer, Statistics with Vague Data, vol. 33, Reidel, Dordrecht, 1987.
[17] K. Kira, L. Rendell, A practical approach to feature selection, in: D. Sleeman, P. Edwards (Eds.), Proceedings of the Ninth International Conference on Machine Learning (ICML-92), Morgan Kaufmannr, 1992, pp. 249–256.
[18] M. Koeppen, K. Franke, B. Nickolay, Fuzzy-pareto-dominance driven multi-objective genetic algorithm, in: Proceedings of the 10th International Fuzzy Systems Association World Congress (IFSA), Istanbul, Turkey, 2003, pp. 450–453.
[19] P. Limbourg, Multi-objective optimization of problems with epistemic 2005 uncertainty, in: EMO 2005, pp. 413–427.
[20] Z. Michalewicz, Genetic Algorithms + Data Structures = Evolution Programs, Springer-Verlag, 1999.
[21] J. Otero, L. Sánchez, Induction of descriptive fuzzy classifiers with the logitboost algorithm, Soft Computing 10 (9) (2006) 825–835.
[22] E.H. Ruspini, A new approach to clustering, Information Control 15 (1969) 22–32.
[23] L. Sánchez, A random sets-based method for identifying fuzzy models, Fuzzy Sets and Systems 98 (3) (1998) 343–354.
[24] L. Sanchez, J. Otero, A fast genetic method for inducting descriptive fuzzy models, Fuzzy Sets and Systems 141 (1) (2004) 33–46.
[25] L. Sánchez, I. Couso, J. Casillas, A multi-objective genetic fuzzy system with imprecise probability fitness for vague data, in: International Symposium on Evolving Fuzzy Systems (EFS 2006), 2006, pp. 131–136.
[26] L. Sanchez, J. Otero, J.R. Villar, Boosting of fuzzy models for high-dimensional imprecise datasets, in: Proceedings of the IPMU 2006, Paris, France, 2006, pp. 1965–1973.
[27] L. Sánchez, I. Couso, J. Casillas, Modelling vague data with genetic fuzzy systems under a combination of crisp and imprecise criteria, in: Proceedings of the 2007 IEEE MCDM, Honolulu, USA, 2007.
[28] L. Sanchez, J. Otero, J. Learning fuzzy linguistic models from low quality data by genetic algorithms, in: FUZZ-IEEE 2007, London, 2007, pp. 1–6.
[29] L. Sánchez, M.R. Suárez, J.R. Villar, I. Couso, Some results about mutual information-based feature selection and fuzzy discretization of vague data, in: FUZZ-IEEE 2007, London, 2007, pp. 1–6.
[30] L. Sanchez, I. Couso, Advocating the use of imprecisely observed data in genetic fuzzy systems, IEEE Transactions on Fuzzy Systems 15 (4) (2007) 551–562.
[31] L. Sanchez, J.R. Villar, Obtaining transparent models of chaotic systems with multi-objective simulated annealing algorithms, Information Sciences 178 (4) (2008) 952–970.
[32] L. Sanchez, J. Otero, I. Couso, Obtaining linguistic fuzzy rule-based regression models from imprecise data with multi-objective genetic algorithms. Soft Computing, in press.
[33] R. Silipo, M.R. Berthold, Input features' impact on fuzzy decision processes, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 30 (6) (2000) 821–834.
[34] H.-J. Sun, M. Sun, Z. Mei, Feature selection via fuzzy clustering, in: Proceedings of the 2006 International Conference on Machine Learning and Cybernetics, Art. no. 4028283, 2006, pp. 1400–1405.
[35] J. Teich, Pareto-front exploration with uncertain objectives, in: EMO, 2001, pp. 314–328.

[36] O. Uncu, I.B. Türksen, A novel feature selection approach, Combining Feature Wrappers and Filters Information Sciences 177 (2) (2007) 449–466.
[37] H. Xia, B.Q. Hu, Feature selection using fuzzy support vector machines, Fuzzy Optimization and Decision Making 5 (2) (2006) 187–192.
[38] N. Xiong, P. Funk, Construction of fuzzy knowledge bases incorporating feature selection, Soft Computing 10 (9) (2006) 796–804.
[39] Y. Zhang, X.-B. Wu, Z.-R. Xiang, W.-L. Hu, Design of high-dimensional fuzzy classification systems based on multi-objective evolutionary algorithm, Journal of System Simulation 19 (1) (2007) 210–215.