Some Results about Mutual Information-based Feature Selection and Fuzzy Discretization of Vague Data

Luciano Sánchez, *Member, IEEE,* M. Rosario Suárez, J. R. Villar and Inés Couso

Abstract— **Algorithms for preprocessing databases with incomplete and imprecise data are seldom studied, partly because we lack numerical tools to quantify the interdependency between fuzzy random variables. In particular, many filter-type feature selection algorithms rely on crisp discretizations for estimating the mutual information between continuous variables, effectively preventing the use of vague data.**

Fuzzy rule based systems pass continuous input variables, in turn, through their own fuzzification interface. In the context of feature selection, should we rank the relevance of the inputs by means of their mutual information, it might happen that an apparently informative variable is useless after having been codified as a fuzzy subset of our catalog of linguistic terms.

In this paper we propose to address both problems by estimating the mutual information with the same set of fuzzy partitions that will be used to codify the antecedents of the fuzzy rules. That is to say, we introduce a numerical algorithm for estimating the mutual information between two fuzzified continuous variables. This algorithm can be included in certain feature selection algorithms, and can also be used to obtain the most informative fuzzy partition for the data. The use of our definition will be exemplified with the help of some benchmark problems.

I. Introduction

Although fuzzy rule-based systems are intended for using vague data, most learning algorithms can only use precise information. Those learning algorithms that can extract rules from imprecise examples are a current research area and, in particular, since recently we can use genetic algorithms to extract fuzzy rules from interval and fuzzy valued data in classification [13], [14] and regression problems [15].

On the contrary, the *preprocessing* of imprecise databases is seldom studied. For instance, there are many recent works dealing with feature selection procedures that use fuzzy techniques [7], [17], [16], [18] or are designed to be used in combination with fuzzy systems [20], [5], [19], but we are not aware of any feature selection algorithms that can be applied to interval-valued or fuzzy data. In particular, up to our knowledge, the definition of the mutual information between fuzzy random variables has not been studied.

That definition would also solve a secondary problem, common to the use of both precise and vague data in fuzzy rule-based systems. Currently, the mutual information is being estimated with the help of of an intermediate crisp partition or approximated by a smooth estimator of the density function of the inputs [1]. None of these approaches takes into account the shape of the membership functions in the fuzzification interface. 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 particular, we will address a rather common situation in Genetic Fuzzy Systems (GFS,) that of numerical variables that are transformed into fuzzy subsets of the set of linguistic labels by means of a Ruspini's fuzzy partition [11]. For example, the fuzzification stage can convert a numerical value of 45 degrees into a fuzzy subset like $\{0.0/\mathrm{COLD} + 0.2/\mathrm{WARM} + 0.8/\mathrm{HOT}\}$. Rule based-systems could also manage subsets like $\{0.1/\mathrm{COLD}{+}0.3/\mathrm{WARM}{+}0.9/\mathrm{HOT}\}$ or $\{0.5/\mathrm{COLD} + 0.5/\mathrm{WARM} + 0.5/\mathrm{HOT}\}$, that do not match any numerical value. Those subsets represent a vague measure and the absence of knowledge about the temperature, respectively, which are cases of practical interest, but again seldom studied in GFS.

In this work, we will propose a new definition of the mutual information between fuzzy random variables, that can be used for measuring the dependence between the variables in either case. In addition, we will show that

- this information measure can be optimized by means of a multiobjective genetic algorithm, and be used to find the fuzzy partition that carries the most information about the class of the object, and
- it can be included in a filter type feature selection procedure that takes into account the shapes of the membership functions in the fuzzification interface.

This paper is organized as follows: in the second section, we introduce our definition of mutual information [12] and detail how to estimate it from vague data. In the third section we will give some details about the genetic optimization of the mutual information, and the fourth section introduces a preliminary MIFS-like algorithm [1] that uses the new measure to select the most relevant features. The fifth section contains numerical results of both approaches. The paper finishes with the concluding remarks and the future work.

II. 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 of them 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

Luciano Sánchez, José R. Villar and M. Rosario Suárez are with the Computer Science Department, University of Oviedo, Campus de Viesques, 33203 Gijón, Asturias, Spain (email: [luciano,mrsuarez,villarjose]@uniovi.es).

Inés Couso is with the Statistics Department, University of Oviedo, Facultad de Ciencias, 33071 Oviedo, Asturias, Spain (email: couso@uniovi.es).

1-4244-1210-2/07/$25.00 ©2007 IEEE.

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 $X : \Omega \rightarrow \mathbf{R}$ and a random set $\Gamma : \Omega \rightarrow \mathcal{P}(\mathbf{R})$, 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.)

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

In previous works [12] 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:

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

Generalizing this concept to fuzzy random variables is immediate, according to a general procedure proposed in [9]. 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

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

A. 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:

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

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

$$\begin{array}{l}\widehat{\text{MI}}((X_1, \ldots, X_N), (\Lambda_1, \ldots, \Lambda_N))(t) = \\ \sup\{\alpha \mid t \in \{\text{MI}((X_1, \ldots, X_N), (Y_1, \ldots, Y_N)) \mid \\ \quad (Y_1, \ldots, Y_N) \in S((\Lambda_1, \ldots, \Lambda_N)_\alpha,)\} \\ \}\end{array} \tag{5}$$

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

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

We want to estimate the mutual information between X and Λ. In the first place, 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
Membership=min$\{0.2, 0.4, 1\} = 0.2$	

Y_2	X
WARM	A
WARM	B
COLD	A
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 $\text{MI}(Y_1, X), \ldots, \text{MI}(Y_4, X)$:

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

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

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

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

It is remarked that the number of samples Y with non-null membership grows with the number of labels raised to the volume of the sample. Enumerating all of them is not feasible but in very small problems, therefore this definition has only 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 III-D.)

B. 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 [4]. For example, the fuzzy set $\{0.0/\text{COLD} + 0.2/\text{WARM} + 0.9/\text{HOT}\}$ means that the probability of the temperature is '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, $p(\text{WARM}) \geq 1 - (p^*(\text{COLD}) + p^*(\text{HOT})) = 0.1$. Observe that, with this interpretation, the set $\{1/\text{COLD} + 1/\text{WARM} + 1/\text{HOT}\}$ represents the total absence of knowledge about the input value but the set $\{0.5/\text{COLD} + 0.5/\text{WARM} + 0.5/\text{HOT}\}$

mentioned in the introduction, while does not signal a preference to neither of the linguistic values, states that the probability of any of them is not higher than 0.5, which is more restrictive. Observe also that the fuzzy set $\{0.0/\text{COLD} + 0.2/\text{WARM} + 0.8/\text{HOT}\}$ provides us with a precise information about the probability distribution, because $0.0+0.2+0.8 = 1$. This kind of fuzzy sets arise when a precise numerical value is passed though a fuzzification interface based on a Ruspini's partition. Lastly, observe that a set like $\{0.0/\text{COLD} + 0.2/\text{WARM} + 0.4/\text{HOT}\}$ (where $0.0+0.2+0.4 < 1$) can not be used with this interpretation.

C. Alternative definition of Mutual Information

Let us interpret the acceptability of a fuzzy random variable [9] as an upper bound of an otherwise unknown probability distribution p_Λ defined on the class of the random variables from Ω to $\mathbf{R}$:

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

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

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

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

D. Computer algorithm for the alternative definition

Let us suppose that we are given two paired samples of X and Λ, as we did 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:

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

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

$$p(\widehat{\text{MI}}((X_1,\ldots,X_N),(\Lambda_1,\ldots,\Lambda_N)) = t\ ) = \sum\nolimits_{\text{MI}((X_1,\ldots,X_N),(Y_1,\ldots,Y_N))=t} \text{p}_\Lambda((Y_1,Y_2,\ldots,Y_N))\}. \tag{9}$$

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

We wish to estimate the mutual information between X and Λ. In the first place, 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
Probability=$[0.1, 0.2] \otimes 0.4 \otimes 1 = [0.04, 0.08]$	

Y_2	X
WARM	A
WARM	B
COLD	A
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 $\text{MI}(X, Y_1), \ldots, \text{MI}(X, Y_4)$ of these samples:

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

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

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

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

Since the number of samples with non-null probability is the same as the number of samples of non-null membership in Section III-A, 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
$\text{MI}(X, Y_2) = 0.5441$	[0.32,0.36]
$\text{MI}(X, Y_4) = 1.2108$	[0.48,0.54].

The expectation of MI is the range of

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

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

1) In the first place, we approximate the unknown mean with the centers of the probability intervals:

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

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:

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

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

$$\mathrm{E}_*(\widehat{\mathrm{MI}}) = \frac{0.5441 * 0.36 + 1.2108 * 0.48}{0.36 + 0.48} = 0.9251$$

Therefore, the approximated value is

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

Interesting enough to mention, following the interpretation described in section III-B, when all the fuzzified inputs originate from crisp numerical values, the algorithm in this section produces a crisp value. On the contrary, if there are some imprecise examples, this algorithm produces an interval.

IV. 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 less information in the discretization. It is assumed that a rule learning that uses such a partition will produce the most accurate knowledge bases, as we experimentally show in the next section.

In case the input data is vague or there are missing values in the dataset, the MI is an interval, as we have mentioned. Seeking the minimum of an interval-valued function is a problem that can be solved with certain multicriteria genetic algorithms [13]. We have used a generational approach with the multiobjective NSGA-II replacement strategy, binary tournament selection based on rank and crowding distance, and a precedence operator that assumes an uniform prior [14]. The nondominated sorting depends on the product of the probabilities of precedence and the crowding is based on the Hausdorff distance.

In this paper we are interested in fuzzification interfaces defined by Ruspini's partitions, as mentioned. We will restrict ourselves to triangular membership functions. In this particular case, a fuzzy partition comprising N linguistic terms can be codified with an array of N numbers: the minimum value of the variable and the distances between all the points of membership 1 and their predecessors. Arithmetic crossover and mutation, and real coding were used.

V. Application II: A 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 Figure 1: in this case an estimation of the mutual information that does not take the memberships of the linguistic terms into account might conclude that certain variable is informative, when it is not.

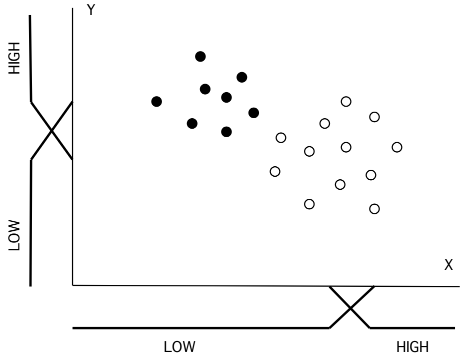

Fig. 1. Example of the theoretical advantages of the proposed estimator in the design of fuzzy rule-based systems. The mutual information between the variable X and the class (white or black) is higher than that of Y. However, choosing the variable X is the worst decision when designing a fuzzy rule-based classification system depending on the fuzzy variables $\widetilde{X}$ and $\widetilde{Y}$, 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 $\widetilde{Y}$, as desired.

```
$F$=initial set of $n$ features; S={$\emptyset$}
For each feature $f \in F$ compute MI$(f, C)$
Perform a nondominated 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 MI$(f, s)$
        Perform a nondominated sorting of the values
          MI$(f, C) \ominus \beta \bigoplus_{s \in S}$ MI$(f, s)$
        Select the first element and set $F = F \setminus \{f\}$,
          $S = S \cup \{f\}$
Output the set $S$
```

Fig. 2. Pseudocode of the MIFS algorithm adapted for its use with an interval-valued estimation of the Mutual Information. The nondominated sorting of the interval-valued estimation of the Mutual Information is performed as explained in reference [14].

In case the 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. Otherwise (vague data or missing values) our estimation produces an interval and some modifications are needed. As an example, in Figure 2 the MIFS algorithm [1] is adapted so that it can use the interval-valued mutual information.

VI. Numerical analysis

The algorithms described in sections IV and V are evaluated and the results are discussed in this section. Thirteen different fuzzy rule learning algorithms have been considered, both heuristic and genetic algorithms-based. The heuristic classifiers are described in [6]: no weights (HEU1), same weight as the confidence (HEU2), differences between the confidences (HEU3, HEU4, HEU5), weights tuned by

	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**

TABLE I

TEST ERROR OF DIFFERENT FUZZY RULE-BASED CLASSIFIERS OVER UNIFORM PARTITIONS AND MI-OPTIMIZED PARTITIONS. CRISP DATA.

	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**

TABLE II

TEST ERROR OF DIFFERENT FUZZY RULE-BASED CLASSIFIERS OVER UNIFORM PARTITIONS AND MI-OPTIMIZED PARTITIONS. INTERVAL DATA.

reward-punishment (REWP) and analytical learning (ANAL). The genetic classifiers are: Selection of rules (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 than PITT, macromutation with probability 0.8– [6]. Lastly, two iterative rule learning algorithms are studied: Fuzzy Ababoost (ADAB) –25 rules of type I, fuzzy inference by sum of votes– [8] and Fuzzy Logitboost (LOGI) –10 rules of type III, fuzzy inference by sum of votes– [10]. All the experiments have been repeated ten times for different permutations of the datasets (10cv experimental setup).

A. *Design of the most informative partition*

Eight crisp datasets and four imprecise datasets have been used to assess the definition of the estimator and its use in the design of fuzzy partitions (see Tables I and II.) The imprecise datasets were designed for this paper, since we have not found similar problems in the literature. We built synthetical realistic problems, simulating the use of a digital scale that rounds the decimal part, in different conditions that include a well calibrated scale (dataset "weight-c": values between $x - 0.5$ and $x + 0.5$ are mapped to the integer value x,) an uncalibrated scale ("weight-uc": values between $x - 0.1$ and $x+0.9$ are mapped to the integer value x,) a random selection between the preceding two scales ("weight-2uc") and 5% of missing values at either coordinate are missing ("weight-mv.") The improvement of the results is almost universal, as expected, proving that our definition of mutual information produces coherent results in both crisp and vague datasets.

B. *Feature selection*

The results of the feature selection algorithm are preliminary. Our first experiments show relevant gains in some datasets, and in those cases that the gain is not significant, the set of features produced by the modified MIFS algorithm is similar to that produced by the classical estimation of the mutual information. In Table III we have compared the results of the new algorithm for some crisp datasets to those of the original MIFS algorithm. In one case the gain is not clear (PIMA), but significant improvements were obtained in the SONAR and WINE datasets. A boxplot with the dispersion of the test error for the WINE problem is shown in Figure 3.

VII. CONCLUDING REMARKS AND FUTURE WORK

The preprocessing of databases with imprecise data is hardly found 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 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

	HEU1	HEU2	HEU3	HEU4	HEU5	REWP	ANAL	GENS	MICH	PITT	HYBR	ADAB	LOGI
PIMA - 4 - Shannon MI	0.269	0.263	**0.263**	**0.263**	**0.263**	**0.256**	0.276	0.256	0.355	0.276	0.263	0.276	**0.230**
PIMA - 4 - Interval MI	**0.263**	**0.256**	0.269	0.269	0.269	0.269	0.**256**	**0.236**	0.355	**0.236**	**0.236**	**0.236**	0.243
SONAR - 5 - Shannon MI	0.300	0.325	0.300	0.300	0.300	**0.275**	0.300	**0.250**	0.350	0.275	0.325	0.300	0.275
SONAR - 5 - Interval MI	**0.275**	**0.300**	0.300	0.300	0.300	0.300	0.300	0.275	**0.250**	**0.250**	**0.300**	0.300	**0.225**
WINE - 5 - Shannon MI	0.323	0.323	0.264	0.205	0.176	0.117	0.235	0.205	0.617	0.205	0.176	0.058	**0.058**
WINE - 5 - Interval MI	**0.176**	**0.117**	**0.147**	**0.176**	0.176	**0.088**	**0.117**	**0.088**	**0.176**	**0.088**	**0.205**	**0.029**	0.088

TABLE III

TEST ERROR OF DIFFERENT FUZZY RULE-BASED CLASSIFIERS AFTER PERFORMING A FEATURE SELECTION, WITH THE ORIGINAL MIFS ALGORITHM AND WITH THE MODIFIED VERSION PROPOSED IN THIS PAPER. THE NUMBER OF FEATURES SELECTED IS SHOWN IN THE FIRST COLUMN.

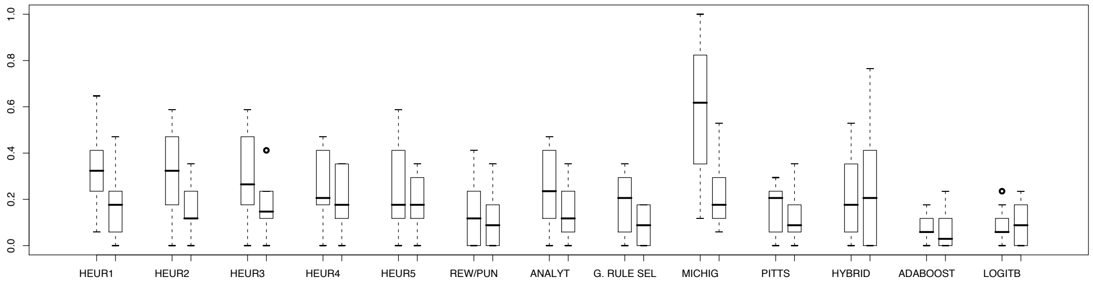

Fig. 3. Boxplots of the test errors of different fuzzy rule-based classifiers with the original MIFS algorithm and the modified version proposed in this paper, WINE dataset.

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 the Figure 1, but further work is needed to characterize this family of 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 to compare the efficiency of the new algorithms with that of the crisp ones over the new set of problems.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] Battiti, R. Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on Neural Networks, 5 (4), pp. 537-550. 1994.

[2] C. Baudrit, I. Couso and D. Dubois, Probabilities of Events Induced by Fuzzy Random Variables, Proc. of International Conference in Fuzzy Logic and Technology (EUSFLAT 05) Barcelona (Spain) 2005.

[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, 82-105. 2007.

[4] D. Dubois and H. Prade, When upper probabilities are possibility measures. Fuzzy Sets and Systems 49, pp. 65-74. 1992.

[5] Fernández-Riverola, F., Díaz, F., Corchado, J.M. Reducing the memory size of a Fuzzy case-based reasoning system applying rough set techniques IEEE Trans. SMC Part C, 37 (1), pp. 138-146. 2007.

[6] Ishibuchi, H., Nakashima, T., Nii, M. Classification and Modeling with Linguistic Information Granules. Springer. 2004.

[7] Jensen, R., Shen, Q. Fuzzy-rough sets assisted attribute selection IEEE Transactions on Fuzzy Systems, 15 (1), pp. 73-89. 2007.

[8] Jesus, M. J. del, Hoffmann F., Junco L., Sánchez L. Induction of Fuzzy Rule Based Classifiers with Evolutionary Boosting Algorithms. IEEE Transactions on Fuzzy Sets and Systems 12(3): 296-308, 2004.

[9] R. Kruse, K.D. Meyer. *Statistics with Vague Data* Vol. 33. Reidel, Dordrecht, 1987.

[10] Otero, J., Sánchez, L. Induction of descriptive fuzzy classifiers with the Logitboost algorithm. Soft Computing 10(9): 825-835, 2006

[11] Ruspini, E.H. A new approach to clustering. *Inf. Control* **15** pp. 22-32. 1969.

[12] Sanchez, L., Suárez, M. R., Couso, I. A fuzzy definition of Mutual Information with application to the Genetic Fuzzy Classifiers. International Conference on Machine Intelligence, Tozeur, Tunisia, November 5-7, 2005

[13] Sanchez, L.; Couso, I.; Casillas, J. A Multiobjective Genetic Fuzzy System with Imprecise Probability Fitness for Vague Data. Int. Symp. on Evolving Fuzzy Systems, pp. 131-136, 2006.

[14] Sánchez, L., Couso, I., Casillas, J. "Modelling vague data with genetic fuzzy systems under a combination of crisp and imprecise criteria" Proc. 2007 IEEE MCDM, Honolulu, USA,. 2007

[15] Sánchez, L., Otero, J., Villar, J. R. Learning fuzzy linguistic models from low quality data by genetic algorithms FUZZ-IEEE 2007, London. 2007.

[16] Sun, H.-J., Sun, M., Mei, Z. Feature selection via fuzzy clustering. Proc. 2006 Int. Conf. on Machine Learning and Cybernetics 2006, art. no. 4028283, pp. 1400-1405. 2006.

[17] Uncu, O., Türksen, I.B. A novel feature selection approach: Combining feature wrappers and filters Information Sciences, 177 (2), pp. 449-466. 2007.

[18] Xia, H., Hu, B.Q. Feature selection using fuzzy support vector machines Fuzzy Optimization and Decision Making, 5 (2), pp. 187-192. 2006.

[19] Xiong, N., Funk, P. Construction of fuzzy knowledge bases incorporating feature selection Soft Computing, 10 (9), pp. 796-804. 2006.

[20] Zhang, Y., Wu, X.-B., Xiang, Z.-R., Hu, W.-L. Design of high-dimensional fuzzy classification systems based on multi-objective evolutionary algorithm. Journal of System Simulation, 19 (1), pp. 210-215. 2007.