Linguistic Cost-Sensitive Learning of Genetic Fuzzy Classifiers for Imprecise Data

Ana M. Palacios[1] Luciano Sánchez[1] Inés Couso[2]

1. Departamento de Informática, Universidad de Oviedo, Gijón, Asturias, Spain
2. Departamento de Estadística e I.O. y D.M, Universidad de Oviedo, Gijón, Asturias Spain
Email: palaciosana@uniovi.es, luciano@uniovi.es, couso@uniovi.es

Abstract

Cost-sensitive classification is based on a set of weights defining the expected cost of misclassifying an object. In this paper, a Genetic Fuzzy Classifier, which is able to extract fuzzy rules from interval or fuzzy valued data, is extended to this type of classification. This extension consists in enclosing the estimation of the expected misclassification risk of a classifier, when assessed on low quality data, in an interval or a fuzzy number. A cooperative-competitive genetic algorithm searches for the knowledge base whose fitness is primal with respect to a precedence relation between the values of this interval or fuzzy valued risk. In addition to this, the numerical estimation of this risk depends on the entrywise product of cost and confusion matrices. These have been, in turn, generalized to vague data. The flexible assignment of values to the cost function is also tackled, owing to the fact that the use of linguistic terms in the definition of the misclassification cost is allowed.

1 Introduction

There are circumstances where the cost associated to a misclassification depends on the class of the individual [22]. The paradigmatic example of this situation is a prescreening test for a serious disease, where the cost of a false positive (making a second diagnosis) is much lower than the opposite case (not detecting the problem) [37, 43, 51].

Following [62], there are two categories of cost-sensitive algorithms. According to their assumptions about the cost function, these are:

1. Class-dependent costs, defined by a matrix of expected risks of misclassification between classes [8, 23, 24, 62, 66, 67].

2. Example-dependent costs [1, 41, 42, 64, 65], where different examples may have different misclassification costs even though they belong to the same class and are also misclassified with the same class.

Notwithstanding these well known foundations, the particular problem of learning fuzzy rule-based classifiers from the perspective of a minimum risk problem has been seldom addressed, except for the particular case of "imbalanced learning" [10], which has been thoroughly studied in the context of Genetic Fuzzy Systems (GFSs) [25]. Nonetheless, some authors have dealt with the concept of "false positives" [53, 58] or

taken into account the confusion matrix in the fitness function [56]. There are also publications related to fuzzy ordered classifiers [32, 33, 57], where an ordering of the class labels defines, in a certain sense, a risk function different than the training error. However, up to our knowledge, the matrix of expected misclassification costs has not been an integral part of the fitness function of a GFS yet. In this paper we will address this issue, and propose a new algorithm for obtaining fuzzy rule-based classifiers from imprecise data with genetic algorithms, extending our own previous works in the subject [47, 48, 49] to problems with class-dependent costs or, in other words, to those cases whose statistical formulation matches the "minimum risk" Bayes classification problem, and the best classifier is defined by the maximum of the conditional risk of each class, given the input [6].

The cost-based GFS that we introduce in this paper is based on a fitness function which is computed by combining the confusion matrix with the expected misclassification cost matrix. It is remarked that we allow that both matrices are interval or fuzzy-valued, and therefore the proposed algorithm can be applied to fuzzy data, the misclassification costs can be fuzzy numbers, or both.

The problem of the flexible assignment of values to the cost function will also be addressed; since the cost matrix can be fuzzy-valued, the use of linguistic terms in the definition of the misclassification cost is allowed. This is useful for solving problems akin to that situation where an expert considers, for instance, that the cost of not detecting certain disease is "very high", while a false positive has a "low" cost. We aim to produce a rule base without asking first the expert to convert his/her quantification into numerical values. In this regard, we are aware of previous published results about the definition of a cost matrix comprising linguistic values, that have been recently introduced in certain decision problems [40, 63]; however, to the best of our knowledge there are not preceding works related to cost-sensitive classification where the cost matrix is not numeric.

The structure of this paper is as follows: in Section 2 we introduce a fuzzy extension of the minimum risk classification problem. In Section 3, we describe a GFS able to extract fuzzy rules from imprecise data, minimizing this extended risk. In Section 4, we have evaluated different aspects of the performance of the new algorithm. The paper finishes with some concluding remarks, in Section 5.

2 A fuzzy extension of the minimum risk classification problem

This section begins reviewing the basics of statistical decision theory, and then interval and fuzzy extensions to this definition are proposed.

2.1 Statistical decision theory

In the following, we will use a bold face, lower case character, such as $\mathbf{x}$, to denote a random variable (or a vector random variable) and lower case roman letters to denote scalar numbers or real vectors. Calligraphic upper case letters are crisp sets.

Let $(\mathbf{x}, \mathbf{c})$ be a random pair taking values in $\mathbb{R}^d \times \mathcal{C}$, where the continuous random vector $\mathbf{x}$ is the feature or input vector, comprising d real values, and the discrete variable $\mathbf{c} \in \mathcal{C} = \{c_1, c_2, \ldots, c_C\}$ is the class. Let $f(x)$ be the density function of the random vector $\mathbf{x}$, and $f(x|c)$ the density function of this vector, conditioned on the class $\mathbf{c} = c$.

$P(c_i)$ is the *a priori* probability of class c_i, $i = 1, \ldots, C$. $P(c_i|x)$ is the a posteriori probability of c_i, given that $\mathbf{x} = x$.

A classifier Φ is a mapping $\Phi : \mathbb{R}^d \to \mathcal{C}$, where $\Phi(x) \in \mathcal{C}$ denotes the class that an object is assigned when it is perceived through the feature vector x. A classifier defines so many decision regions $\mathcal{D}_i$ as classes,

$$\mathcal{D}_i = \{x \in \mathbb{R}^d \mid \Phi(x) = c_i\}, \quad i = 1, 2, \ldots C. \tag{1}$$

Let us define a matrix $B = [b_{ij}] \in \mathcal{M}_{C\times C}$, where $b_{ij} = \text{cost}(c_i, c_j)$ is the cost of deciding that an object is of class c_i when its actual class is c_j. The performance of a classifier can be measured by the average misclassification risk

$$R(\Phi) = \sum_{i=1}^{C} \int_{\mathcal{D}_i} \sum_{j=1}^{C} b_{ij} P(c_j|x) f(x) \mathrm{d}x. \tag{2}$$

Let the conditional risk be

$$R(c_i|x) = \sum_{j=1}^{C} b_{ij} P(c_j|x). \tag{3}$$

The decision rule minimizing the average misclassification risk in Eq. (2) is

$$\Phi_B(x) = \arg\min_{c\in\mathcal{C}} R(c|x), \tag{4}$$

so called "minimum risk Bayes rule" [6]. Observe that setting $b_{ij} = 1$ for $i \neq j$ and $b_{ii} = 0$ causes that Eq. (2) is proportional to the expected fraction of misclassifications of the classifier Φ, and

$$R(c_i|x) = \sum_{\substack{j\in\{1,\ldots,C\} \\ i\neq j}} P(c_j|x) = 1 - P(c_i|x), \tag{5}$$

thus the best decision rule is

$$\Phi_B(x) = \arg\min_{c\in\mathcal{C}} R(c|x) = \arg\max_{c\in\mathcal{C}} P(c|x), \tag{6}$$

the so called "minimum error Bayes rule" [6].

Generally speaking, the conditional probabilities $P(c_j|x)$ are unknown and thus the minimum error and minimum risk Bayes rules cannot be directly applied. Instead, in this work we will discuss how to make estimates of Eq. (2) from data, and search for the knowledge base whose estimated risk is minimum. In the first place, we will suggest how to define this estimator for crisp, interval and fuzzy data.

2.2 Estimation of the expected risk with crisp data and crisp costs

Let us consider a random sample or *dataset* $\mathcal{D}$ comprising N objects, where each object is perceived through a pair comprising a vector and a number; the features of the k-th object form the vector x_k and the class of the same k-th object is c_{y_k}:

$$\mathcal{D} = \{(x_k, y_k)\}_{k=1}^{N}. \tag{7}$$

Let also N_i be the number of objects of class c_i,

$$\sum_{i=1}^{C} N_i = N. \tag{8}$$

We will compute an approximated value of the expected risk of the classifier, on the basis of the mentioned dataset. Let us assume first that there are not duplicate elements in the sample; in this case, we can define a (crisp) partition $\{\mathcal{V}_k\}_{k=1}^N$ of the input space such that each feature vector x_k is in a set $\mathcal{V}_k$. Our approximation consists in admitting that all densities are simple functions, attaining constant values in the elements of this partition.

Let $I(x) \in \{1, \ldots, N\}$ denote the index of the set in the partition $\{\mathcal{V}_k\}_{k=1}^N$ that contains the element x, thus $x \in \mathcal{V}_{I(x)}$ and $I(x_k) = k$. We will approximate $f(x)$ by

$$\hat{f}(x) = \frac{1}{N||\mathcal{V}_{I(x)}||} \tag{9}$$

(where the modulus operator means Lebesgue measure, or volume) and

$$\hat{f}(x|c_i) = \frac{\delta_{i,y_{I(x)}}}{N_i||\mathcal{V}_{I(x)}||} \tag{10}$$

where the symbol δ is Dirichlet's delta. The risk of the classifier reduces to the expression that follows:

$$\begin{aligned}
\widehat{R}(\Phi, \mathcal{D}) &= \sum_{i=1}^{C} \int_{\mathcal{D}_i} \sum_{j=1}^{C} b_{ij} \hat{f}(x|c_j) P(c_j) \mathrm{d}x \\
&= \sum_{i=1}^{C} \int_{\mathcal{D}_i} \sum_{j=1}^{C} b_{ij} \frac{\delta_{j,y_{I(x)}}}{N_j||\mathcal{V}_{I(x)}||} \frac{N_j}{N} \mathrm{d}x \\
&= \sum_{i=1}^{C} \sum_{\{k|\Phi(x_k)=c_i\}} ||\mathcal{V}_{I(x_k)}|| \sum_{j=1}^{C} b_{ij} \frac{\delta_{j,y_{I(x_k)}}}{||\mathcal{V}_{I(x_k)}||} \frac{1}{N} \\
&= \sum_{i=1}^{C} \sum_{\{k|\Phi(x_k)=c_i\}} \sum_{j=1}^{C} \frac{1}{N} b_{ij} \delta_{j,y_k}.
\end{aligned} \tag{11}$$

Eq. (11) can be expressed in terms of the confusion matrix of the classifier and the cost matrix. Let $S(\Phi, \mathcal{D}) = [s_{ij}]$ be the confusion matrix of the classifier Φ on the dataset $\mathcal{D}$. s_{ij} is the number of elements in the sample for which the output $\Phi(x_k)$ of the classifier is c_i and the class of the element is c_{y_k}. Let us express this as follows:

$$s_{ij} = \sum_{k=1}^{N} \delta_{c_i,\Phi(x_k)} \delta_{j,y_k}, \tag{12}$$

where we have use Kronecker's delta both for natural numbers and elements of $\mathcal{C}$. Lastly, let

$$M(\Phi, \mathcal{D}) = \frac{1}{N} B \circ S(\Phi, \mathcal{D}) \tag{13}$$

where $B \circ S = [b_{ij} s_{ij}] = [m_{ij}]$ is the Hadamard product of the cost matrix and the confusion matrix. Then

$$\widehat{R}(\Phi, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{C} \sum_{j=1}^{C} m_{ij}. \tag{14}$$

Observe that, in this crisp case, Eq. (14) can also be written as

$$\widehat{R}(\Phi, \mathcal{D}) = \frac{1}{N} \sum_{k=1}^{N} \text{cost}(\Phi(x_k), c_{y_k}). \tag{15}$$

2.3 Estimation of the expected risk with interval-valued data and/or interval-valued costs

Suppose that the features and the classes of the objects in the dataset cannot be accurately perceived, but we are given sets (other than singletons, in general) that contain them:

$$\overline{\mathcal{D}} = \{(\mathcal{X}_k, \mathcal{Y}_k)\}_{k=1}^{N} \tag{16}$$

where $\mathcal{X}_k \subset \mathbb{R}^d$ and $\mathcal{Y}_k \subset \{1, \ldots, C\}$. The most precise output of the classifier Φ for a set-valued input $\mathcal{X}$ is

$$\Phi(\mathcal{X}) = \{\Phi(x) \mid x \in \mathcal{X}\}. \tag{17}$$

In this case, the elements of the confusion matrix $\overline{S}$ are also sets. Let us define, for simplicity in the notation, the set-valued function $\overline{\delta} : \mathcal{C} \times \mathcal{P}(\mathcal{C}) \to \mathcal{P}(\{0, 1\})$

$$\overline{\delta}_{a,\mathcal{A}} = \{\delta_{a,b} \,:\, b \in \mathcal{A}\} = \begin{cases} \{1\} & \{a\} = \mathcal{A} \\ \{0\} & a \notin \mathcal{A} \\ \{0, 1\} & \text{else.} \end{cases} \tag{18}$$

With the help of this function, the confusion matrix in the preceding subsection is generalized to an interval-valued matrix $\overline{S} = [\overline{s}_{ij}]$, as follows:

$$\overline{s}_{ij} = \sum_{k=1}^{N} \overline{\delta}_{c_i, \Phi(\mathcal{X}_k)} \overline{\delta}_{j, \mathcal{Y}_k}. \tag{19}$$

Observe that this last expression makes use of set-valued addition and multiplication,

$$\mathcal{A} + \mathcal{B} = \{a + b \mid a \in \mathcal{A}, b \in \mathcal{B}\} \tag{20}$$

$$\mathcal{A} \cdot \mathcal{B} = \{ab \mid a \in \mathcal{A}, b \in \mathcal{B}\}. \tag{21}$$

Given an interval-valued cost matrix $\overline{B}$, Eq. (13) is transformed into

$$\overline{M} = [\overline{m}_{ij}] = \frac{1}{N} \overline{B} \circ \overline{S} \tag{22}$$

and the set-valued risk is

$$\overline{R}(\Phi, \overline{\mathcal{D}}) = \frac{1}{N} \sum_{i=1}^{C} \sum_{j=1}^{C} \overline{m}_{ij}. \tag{23}$$

2.4 Estimation of the expected risk with fuzzy data and/or fuzzy costs

In this paper we will use a possibilistic semantic for vague data. This consists in regarding the noise in the data as random and assuming that our knowledge about the probability distribution of this noise is incomplete. In other words, a fuzzy set $\widetilde{\mathcal{X}}$ is

meta-knowledge about an imprecisely perceived value, and provides information about the probability distribution of an unknown random variable $\mathbf{x}$,

$$P(\mathbf{x} \in [\widetilde{\mathcal{X}}]_\alpha) \geq 1 - \alpha. \tag{24}$$

Observe that this definition extends the interval-valued problem mentioned before. In this context, intervals are a particular case of fuzzy sets because we can regard an interval $\mathcal{X}$ as an incomplete characterization of a random variable $\mathbf{x}$ for which our only knowledge is

$$P(\mathbf{x} \in \mathcal{X}) = 1. \tag{25}$$

From the foregoing it can be inferred that, when both the features and the classes are fuzzy, the dataset

$$\widetilde{\mathcal{D}} = \{(\widetilde{\mathcal{X}}_k, \widetilde{\mathcal{Y}}_k)\}_{k=1}^N, \tag{26}$$

where $\widetilde{\mathcal{X}}_k \in \mathcal{F}(\mathbb{R}^d)$ and $\widetilde{\mathcal{Y}}_k \in \mathcal{F}(\{1, \ldots, C\})$ is a generalization of the interval dataset seen in the preceding section. Regarding fuzzy sets as families of α-cuts, it can be defined

$$[\widetilde{R}(\Phi, \widetilde{\mathcal{D}})]_\alpha = \overline{R}(\Phi, [\widetilde{\mathcal{D}}]_\alpha). \tag{27}$$

Nonetheless, from a computational point of view it is convenient to express this result in a different form. In the first place, let us define the output of the classifier Φ for a fuzzy input $\widetilde{\mathcal{X}}$ as the fuzzy set

$$\Phi(\widetilde{\mathcal{X}})(c) = \sup\{\alpha \mid \Phi(x) = c,\, x \in [\widetilde{\mathcal{X}}]_\alpha\}. \tag{28}$$

Second, let us define the fuzzy function $\widetilde{\delta} : \mathcal{C} \times \mathcal{F}(\mathcal{C}) \to \mathcal{F}(\{0, 1\})$ as

$$\begin{array}{ll} \widetilde{\delta}_{a,\widetilde{\mathcal{A}}}(0) & = \sup\{\widetilde{\mathcal{A}}(b) \,:\, \delta_{a,b} = 0\} = \max\{\widetilde{\mathcal{A}}(c) \mid c \in \mathcal{C}, c \neq a\} \\ \widetilde{\delta}_{a,\widetilde{\mathcal{A}}}(1) & = \sup\{\widetilde{\mathcal{A}}(b) \,:\, \delta_{a,b} = 1\} = \widetilde{\mathcal{A}}(a), \end{array} \tag{29}$$

where we have used the extension principle for extending δ from $\mathcal{C} \times \mathcal{C}$ to $\mathcal{C} \times \mathcal{F}(\mathcal{C})$. With the help of this function, we define the confusion matrix $\widetilde{S}(\Phi, \widetilde{D}) = [\widetilde{s}_{ij}]$ of a classifier Φ for a fuzzy dataset $\widetilde{\mathcal{D}}$ as

$$\widetilde{s}_{ij} = \bigoplus_{k=1}^N \widetilde{\delta}_{c_i, \Phi(\widetilde{\mathcal{X}}_k)} \odot \widetilde{\delta}_{j, \widetilde{\mathcal{Y}}_k}, \tag{30}$$

where

$$(\widetilde{\mathcal{A} \oplus \mathcal{B}})(x) = \sup\{\alpha \mid x = a + b, a \in [\widetilde{\mathcal{A}}]_\alpha, b \in [\widetilde{\mathcal{B}}]_\alpha\} \tag{31}$$

$$(\widetilde{\mathcal{A} \odot \mathcal{B}})(x) = \sup\{\alpha \mid x = ab, a \in [\widetilde{\mathcal{A}}]_\alpha, b \in [\widetilde{\mathcal{B}}]_\alpha\}. \tag{32}$$

Given a fuzzy cost matrix $\widetilde{B}$, the sum of the elements of the entrywise product of $\widetilde{S}$ and $\widetilde{B}$ is proportional to the expected risk of the classifier:

$$\widetilde{M} = [\widetilde{m}_{ij}] = \frac{1}{N} \widetilde{B} \circ \widetilde{S} \tag{33}$$

and

$$\widetilde{R}(\Phi, \widetilde{\mathcal{D}}) = \frac{1}{N} \bigoplus_{i=1}^C \bigoplus_{j=1}^C \widetilde{m}_{ij}. \tag{34}$$

3 A GFS for imprecise data and linguistic costs

In this section we will detail the computational steps needed for obtaining a classifier Φ from a dataset $\widetilde{\mathcal{D}}$. The classifier has to optimize the risk $\widetilde{R}(\Phi, \widetilde{\mathcal{D}})$, and satisfy the following properties:

- The classification system is based on a Knowledge Base (KB) comprising descriptive fuzzy rules [16, 17], and the linguistic terms in these rules are associated to fuzzy partitions of the input features. We will assume that these partitions do not change during the learning, to preserve their linguistic meaning. The inference mechanism defined in [48] will be used, as it fulfills Eq. (28).
- The expected risk is fuzzy-valued and thus conventional genetic algorithms cannot be applied without alterations. In this paper we will use a cooperative-competitive algorithm that searches for the set of rules whose combined fitness evolves toward the primal elements of certain order, defined by a precedence relation between interval or fuzzy values [48].

3.1 Fuzzy inference with vague data

Let us recall the extension of fuzzy inference to vague data introduced in [48], and rewrite it with the notation used in this paper. It is remarked that this inference cannot be applied to arbitrary fuzzy data. We will assume that we can attribute a possibilitic meaning to the vague information [18], thus all fuzzy sets are normal.

Let $x = (x_1, \ldots, x_d)$ be a vector of features. Consider a KB comprising M rules

$$\begin{array}{c} R_1 : \texttt{If } x \texttt{ is } \widetilde{\mathcal{A}}_1 \texttt{ then class is } \mathtt{c}_{\mathtt{q}_1} \\ \cdots \\ R_M : \texttt{If } x \texttt{ is } \widetilde{\mathcal{A}}_M \texttt{ then class is } \mathtt{c}_{\mathtt{q}_M}, \end{array} \tag{35}$$

where $\widetilde{\mathcal{A}}_r$ is a fuzzy subset of $\mathbb{R}^d$. Generally speaking, the expression "x is $\widetilde{\mathcal{A}}_r$" will be a combination of asserts of the form "x_p is $\widetilde{\mathcal{A}}_{rq}$" by means of different logical connectives, where the terms $\widetilde{\mathcal{A}}_{rq}$ are fuzzy subsets of $\mathbb{R}$ that have been assigned a linguistic meaning, and the membership function of $\widetilde{\mathcal{A}}_r$ models the degree of truth of this combination.

Given a precise observation x of the features of an object, the classification system assigns to this object the class given by the consequent of the winner rule R_w, where

$$w = \arg \max_{r=1\ldots M} \widetilde{\mathcal{A}}_r(x) \tag{36}$$

and the output of the classifier is $\Phi(x) = c_{q_w}$. If the input is the imprecise value $\widetilde{\mathcal{X}}$, there is a fuzzy set of winner rules,

$$\widetilde{\mathcal{W}}(\widetilde{\mathcal{X}})(r) = \sup\{\alpha \mid r = \arg \max_{r=1\ldots M} \widetilde{\mathcal{A}}_r(x),\ x \in [\widetilde{\mathcal{X}}]_\alpha\} \tag{37}$$

and the output of the classifier is a normal fuzzy subset of $\mathcal{C}$,

$$\widetilde{\Phi}(\widetilde{\mathcal{X}})(c) = \sup\{\alpha \mid c = c_{q_{\arg\max \widetilde{\mathcal{A}}_r(x)}},\ x \in [\widetilde{\mathcal{X}}]_\alpha\}. \tag{38}$$

3.2 Cooperative-competitive algorithm

The genetic algorithm that we will define in this section is inspired by [34], and generalizes to linguistic costs those Cooperative-Competitive Genetic Algorithms introduced in [47, 48] for error-based classification with imprecise data. Similar to this reference, each chromosome encodes the antecedent of a rule, and the individuals in the population cooperate to form a KB. Likewise, the consequents of the rules are not subject to evolution; a deterministic function of the antecedent is used instead. However, in [34], the distribution of the fitness among the rules consisted in assigning to each individual the number of instances in the dataset that are well classified by its associated rule: the one formed by the antecedent encoded in the chromosome and a consequent obtained, in turn, with the mentioned deterministic procedure. On the contrary, in this work the fitness of the KB is distributed among the individuals in such a way that the sum of the fitness of all the chromosomes in the population is a set that contains the expected risk of the classifier, and the fitness of an individual is an interval or a fuzzy set bounding the average risk of the corresponding rule. Finally, in both ref. [34] and this work, the competition is based on the survival of the fittest; those rules that cover a higher number of instances that are compatible with their consequents have better chances of being selected for recombination.

3.2.1 Genetic representation and procedure for choosing consequents

As we have mentioned before, chromosomes only contain the antecedents of the rules. Following [29], a linguistic term is represented with a chain of bits. There are as many bits in the chain as different terms in the corresponding linguistic partition. If a term appears in the rule, its bit has the value '1', or '0' otherwise. For example, let $\{$Low, Med, High$\}$ be the linguistic labels of all features in a problem involving three input variables. The antecedent of the rule

$$\texttt{If } x_1 \texttt{ is High and } x_2 \texttt{ is Med and } x_3 \texttt{ is Low then class is c},$$

is codified with the chain 001 010 100. This encoding can be used for representing rules for which not all variables appear in the antecedent, and also for 'OR' combinations of terms in the antecedent. For example, the rule

$$\texttt{If } x_1 \texttt{ is High and } x_3 \texttt{ is Low then class is c},$$

is codified with the chain 001 000 100, and the rule

$$\texttt{If } x_1 \texttt{ is } (\texttt{High or Med}) \texttt{ and } x_3 \texttt{ is Low then class is c},$$

will be assigned the chain 011 000 100.

With respect to the definition of the consequent, the alternative with lower risk is preferred. This generalizes the most common procedure, which is selecting the alternative with higher confidence. The expression of the confidence of the fuzzy rule

$$\texttt{If } x \texttt{ is } \widetilde{\mathcal{A}} \texttt{ then class is c},$$

on a crisp dataset $D = \{(x_k, y_k)\}_{k=1}^N$, is

$$\text{confidence}(\widetilde{\mathcal{A}}, c, \mathcal{D}) = \frac{\sum_k \delta_{cy_k} \widetilde{\mathcal{A}}(x_k)}{\sum_k \widetilde{\mathcal{A}}(x_k)}, \tag{39}$$

and thus given an antecedent $\widetilde{\mathcal{A}}$ the class c is chosen that fulfills

$$c = \arg \max_{i=1,\ldots,C} \text{confidence}(\widetilde{\mathcal{A}}, c_i, \mathcal{D}). \tag{40}$$

Observe that the denominator of Eq. (39) does not depend on c and it can be removed without changing the result of Eq. (40). Let us use the word "compat" for denoting the degree of compatibility between a rule and the dataset $\mathcal{D}$:

$$\text{compat}(\widetilde{\mathcal{A}}, c, \mathcal{D}) = \sum_k \delta_{cy_k} \widetilde{\mathcal{A}}(x_k), \tag{41}$$

$$\arg \max_{i=1,\ldots,C} \text{confidence}(\widetilde{\mathcal{A}}, c_i, \mathcal{D}) = \arg \max_{i=1,\ldots,C} \text{compat}(\widetilde{\mathcal{A}}, c_i, \mathcal{D}). \tag{42}$$

This simplification is useful for generalizing expression in Eq. (39) to imprecise data. Given our interpretation of a fuzzy membership, we assume that there exist unknown values x_k, y_k and our knowledge about them is given by the fuzzy sets $\widetilde{\mathcal{X}}_k$, $\widetilde{\mathcal{Y}}_k$ (see Eq. (24)), thus eq. (41) becomes

$$\begin{aligned} &\widetilde{\text{compat}}(\widetilde{\mathcal{A}}, c, \widetilde{\mathcal{D}})(t) = \\ &\qquad \max\{\alpha \mid t = \text{compat}(\widetilde{\mathcal{A}}, c, \mathcal{D}),\ x_k \in [\widetilde{\mathcal{X}}_k]_\alpha, y_k \in [\widetilde{\mathcal{Y}}_k]_\alpha\} \end{aligned} \tag{43}$$

where

$$\widetilde{\mathcal{A}}(\widetilde{\mathcal{X}})(t) = \sup\{\alpha \mid t = \widetilde{\mathcal{A}}(x),\ x \in [\widetilde{\mathcal{X}}]_\alpha\}. \tag{44}$$

We propose to similarly define the risk of the same fuzzy rule seen before, given a cost matrix $B = [b_{ij}]$, as

$$\text{risk}(\widetilde{\mathcal{A}}, c, \mathcal{D}, B) = \sum_k b_{cy_k} \widetilde{\mathcal{A}}(x_k). \tag{45}$$

thus the preferred consequent is

$$c = \arg \min_{i=1,\ldots,C} \text{risk}(\widetilde{\mathcal{A}}, c_i, \mathcal{D}, B). \tag{46}$$

The generalization of this expression to a fuzzy dataset $\widetilde{\mathcal{D}} = \{(\widetilde{\mathcal{X}}_k, \widetilde{\mathcal{Y}}_k)\}_{k=1}^N$ and a fuzzy cost matrix $\widetilde{B} = [\widetilde{b}_{ij}]$ is

$$\begin{aligned} \widetilde{\text{risk}}(\widetilde{\mathcal{A}}, c, \widetilde{\mathcal{D}}, \widetilde{B})(t) = \max\{\alpha \mid &\text{risk}(\widetilde{\mathcal{A}}, c, \mathcal{D}, B) = t, \\ &x_k \in [\widetilde{\mathcal{X}}_k]_\alpha, y_k \in [\widetilde{\mathcal{Y}}_k]_\alpha,\ b_{ij} \in [\widetilde{b}_{ij}]_\alpha \text{ for all } i, j, k\ \}, \end{aligned} \tag{47}$$

which is a fuzzy set. We want to find the alternative c with the lowest risk, but the meaning of "lowest risk" admits different interpretations in this context. If the specificity of the imprecise features is high, we can make the approximation that follows without incurring large deviations:

$$\text{approx.risk}(\widetilde{\mathcal{A}}, c, \widetilde{\mathcal{D}}, \widetilde{B}) = \bigoplus_k \widetilde{\mathcal{A}}(\widetilde{\mathcal{X}}_k) \odot \bigvee_{d \in \mathcal{C}} (\widetilde{b}_{cd} \wedge \widetilde{\mathcal{Y}}_k(d)), \tag{48}$$

where $\oplus$ and $\odot$ are the fuzzy arithmetic extensions of addition and multiplication. In this work we will sort the results of Eqs. (43) or (48) with the help of a precedence operator between fuzzy sets (this operator will be defined in this section) and select the value of 'c' associated to the primal element in the order that this operator induces.

3.2.2 Initial population

A fraction of the initial population is generated at random, with different probabilities for the symbols '1' and '0'. Provided that the higher the percentage of the symbol '1', the less specific are the rules, a high number of appearances of this symbol produce initial knowledge bases that are less likely to leave uncovered examples. We do not allow the presence of 'OR' combinations involving all the linguistic terms of a variable, which are replaced by zeroes, representing "do not care" terms.

The remaining instances are generated to cover randomly chosen elements in the dataset. Let $\mathcal{L}$ be the finite crisp set of all the possible antecedents (recall that we are using descriptive rules, without membership tuning). If an instance $(\widetilde{\mathcal{X}}_k, \widetilde{\mathcal{Y}}_k)$ is selected, then an individual is generated whose antecedent $\widetilde{\mathcal{K}} \in \mathcal{L}$ fullfills

$$\widetilde{\mathcal{K}}(\widetilde{\mathcal{X}}_k) \succeq \widetilde{\mathcal{A}}(\widetilde{\mathcal{X}}_k) \text{ for all } \widetilde{\mathcal{A}} \in \mathcal{L} \ , \tag{49}$$

and $\widetilde{\mathcal{A}}(\widetilde{\mathcal{X}}_k)$ was defined in Eq. (44).

3.2.3 Precedence operators

Many authors have proposed different operators for ranking fuzzy numbers, beginning with the seminal works in [35, 36]. Often [12, 13, 20, 38, 54, 60] the uncertainty is removed and the centroids of the membership functions are compared, but there is a wide range of alternative techniques [11, 14, 15, 61]. Generally speaking, no matter which of the mentioned rankings would serve for our purpose. Nevertheless, in this work it is given a possibilistic interpretation to the fuzzy information in the datasets, thus we will provide a ranking method which is based in a stochastic precedence. We want to remark that the criterion suggested here is still based on ad-hoc hypothesis about the distribution of the random variables encoded in the fuzzy memberships, and thus the order that it induces is not less arbitrary than any of the cited references. However, with this definition we will be aware of the hypothesis we are introducing, while many of the mentioned works are based on heuristic or epistemic foundations whose suitability cannot always be assessed for this application.

Let $\widetilde{\mathcal{A}}$, $\widetilde{\mathcal{B}}$ be two fuzzy values (which, in this context, are fuzzy restrictions of the misclassification risk of a fuzzy rule). We want to determine whether $\widetilde{\mathcal{A}} \preceq \widetilde{\mathcal{B}}$, $\widetilde{\mathcal{B}} \preceq \widetilde{\mathcal{A}}$, or $\widetilde{\mathcal{A}} \parallel \widetilde{\mathcal{B}}$. We have mentioned before that our possibilistic semantic for vague data consists in considering a stochastic behaviour whose characterization is incomplete, i.e. each fuzzy membership $\widetilde{\mathcal{A}}$ is meta-knowledge about an imprecisely perceived value: we admit that there exists a random variable $\mathbf{a}$, and the fuzzy set provides information about the probability distribution of this variable. This knowledge is

$$P(\mathbf{a} \in [\widetilde{\mathcal{A}}]_\alpha) \geq 1 - \alpha. \tag{50}$$

Furthermore, we will match the fuzzy precedence between $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{B}}$ with the stochastic precedence that follows:

$$\widetilde{\mathcal{A}} \preceq \widetilde{\mathcal{B}} \iff P(\mathbf{a} \leq \mathbf{b}) \geq P(\mathbf{b} < \mathbf{a}) \tag{51}$$

or

$$\widetilde{\mathcal{A}} \preceq \widetilde{\mathcal{B}} \iff P(\mathbf{a} - \mathbf{b} \leq 0) \geq 1/2, \tag{52}$$

thus in case the vector $(\mathbf{a}, \mathbf{b})$ is continuous this criteria is related to the sign of the median of the difference between the two unknown variables $\mathbf{a}$ and $\mathbf{b}$.

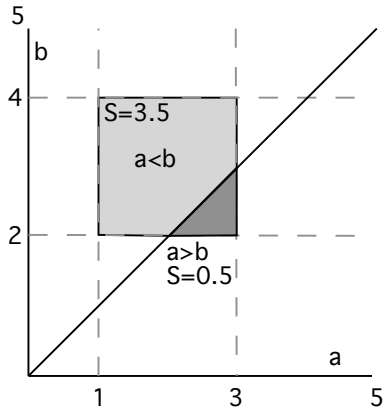

Figure 1: Graphical representation of the calculations needed for determining the precedence between the interval valued risks $[1, 3]$ and $[2, 4]$ in example 1.

Unless further assumptions are made, if the supports of $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{B}}$ are not disjoint then $\widetilde{\mathcal{A}} \parallel \widetilde{\mathcal{B}}$; the criterion that is obtained in this case is similar in concept to the *strong dominance* in [39]. In spite of this, there are many other criteria in the literature can also be regarded as particular cases of this stochastic precedence. For instance, if it is assumed that $\mathbf{a}$ and $\mathbf{b}$ are independent, and the joint distribution of the random vector $(\mathbf{a}, \mathbf{b})$ is uniform, we obtain the commonly used uniform precedence [55], which was originally defined for interval-valued data. This precedence is illustrated in Fig. 1 and in the examples that follow.

Example 1 *Let* $\mathcal{A} = [1, 3]$ *and* $\mathcal{B} = [2, 4]$. *If we assume that* $P(\mathbf{a}, \mathbf{b})$ *is uniform in* $[1, 3] \times [2, 4]$ *(see Figure 1) we obtain*

$$\frac{P(\{(\mathbf{a}, \mathbf{b}) : \mathbf{a} \leq \mathbf{b}\})}{P(\{(\mathbf{a}, \mathbf{b}) : \mathbf{a} > \mathbf{b}\})} = \frac{3.5/4}{0.5/4} > 1 \tag{53}$$

thus $\mathcal{A} \preceq \mathcal{B}$.

Example 2 *Let* $\mathcal{A} = [1, 5]$ *and* $\mathcal{B} = [1.9, 4]$. *The application of the same principle produces*

$$\frac{P(\{(\mathbf{a}, \mathbf{b}) : \mathbf{a} \leq \mathbf{b}\})}{P(\{(\mathbf{a}, \mathbf{b}) : \mathbf{a} > \mathbf{b}\})} = \frac{4.095}{4.305} < 1 \tag{54}$$

therefore $\mathcal{B} \preceq \mathcal{A}$.

Depending on the shape of the membership functions of $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{B}}$ the hypothesis about the uniform distribution of $(\mathbf{a}, \mathbf{b})$ still makes sense for fuzzy data. In this paper we have assumed that

$$(\mathbf{a}, \mathbf{b}) \to \mathcal{U}\left(\left[l(\widetilde{\mathcal{A}}), r(\widetilde{\mathcal{A}})\right] \times \left[l(\widetilde{\mathcal{B}}), r(\widetilde{\mathcal{B}})\right]\right) \tag{55}$$

where $l(\widetilde{\mathcal{A}})$, $r(\widetilde{\mathcal{A}})$ and the corresponding values for $\widetilde{\mathcal{B}}$ are the bounds of the expectation of the fuzzy number, as defined in [21].

3.2.4 Fitness function

For crisp data, when the k-th instance is presented to the classifier, the fitness of the winner rule w is penalized with a value that matches the risk of classifying this object,

$$\mathrm{fit}(w,k) = b_{q_r y_k}. \tag{56}$$

It is remarked that the objective of this learning is to minimize the fitness (minimize the risk), contrary to the usual practice in this kind of algorithms, where the objective is maximizing the fitness (the number of well classified instances).

Before extending this expression to interval data, let us rewrite Eq. (56) as follows:

$$\mathrm{fit}(w,k) = \sum_{j=1}^{C} \delta_{j,y_k} b_{q_r j}. \tag{57}$$

For interval data, each rule R_r in the winner set is penalized with an interval-valued risk, because the true class of the k-th object can be perceived as a set of elements of $\mathcal{C}$. In this case, our knowledge about the fitness value is given by an extension of Eq. (57):

$$\overline{\mathrm{fit}}(r,k) = \{\sum_{j=1}^{C} \gamma_j \overline{b}_{q_r j} | \gamma_j \in \overline{\delta}_{j,Y_k} \text{ and } \sum_{j=1}^{C} \gamma_j = 1\} \tag{58}$$

where $\overline{\delta}$ is a set-valued generalization of Dirichlet's delta, that was defined in Eq. (18). Since the computation of the preceding expression is costly, we will enclose it in the set

$$\overline{\mathrm{fit}}(r,k) = \sum_{j=1}^{C} \overline{\delta}_{j,Y_k} \overline{b}_{q_r j}. \tag{59}$$

Let us clarify the meaning of this expression with a numerical example. Let q_r = c_2, Y_k=$\{c_1, c_3\}$, $C = \{c_1, c_2, c_3\}$, and let the matrix $\overline{B}$=$[\overline{b}_{ij}]$ be

$$\overline{B} = \begin{vmatrix} 0 & [0.8,1] & [0.7,0.9] \\ [0.1,0.3] & [0.1,0.15] & [0.3,0.6] \\ 1 & [0.6,0.85] & 0 \end{vmatrix}$$

The fitness of the r-th ruke will be:

$$\overline{\mathrm{fit}}(r,k) = \{\{0,1\}[0.1,0.3] + \{0\}[0.1,0.15] + \{0,1\}[0.3,0.6]\} = \{0.1, 0.3, 0.6\}.$$

For fuzzy data, each rule R_r in the support of the winner set $\widetilde{\mathcal{W}}$ is penalized with the risk of their classification, which in turn might be a fuzzy set, if the true class of the k-th object is partially unknown:

$$\widetilde{\mathrm{fit}}(r,k) = \bigoplus_{j=1}^{C} \widetilde{\delta}_{j,\widetilde{y}_k} \odot \widetilde{b}_{q_r j}, \tag{60}$$

where $\widetilde{\delta}$ was defined in Eq. (29).

3.2.5 Generational scheme and genetic operators

This GFS operates by selecting two parents with the help of a double binary tournament, where the order between the fuzzy valued fitness function depends on the sign of the median of the difference of those random variables we have assumed implicit in the fuzzy memberships, with hypothesis of independence and uniform distribution, as explained in the Subsection 3.2.3. These two parents are recombined and mutated with standard two-point crossover [45] and uniform mutation [44], respectively. After the application of crossover or mutation we search the individuals for the occurrence of chains where there exist 'OR' combinations involving all the linguistic terms of a variable. As we have mentioned, these chains are replaced by chains of zeroes, representing "do not care" terms.

The consequent with a lower risk is determined for each element of the offspring, according to the procedure in Subsection 3.2.1, and inserted into a secondary population, whose size is smaller than that of the primary population. The worse individuals of the primary population (again, according to the same precedence operator) are replaced by those in the secondary population at each generation.

Once these individuals have been replaced, the fitness assignment begins. Each rule keeps a fuzzy counter, which is zeroed first. The second step consists in determining the set of winner rules defined in Subsection 3.1, for each instance k in the dataset. The counters of these winner rules are incremented the amount defined in Subsection 3.2.4. After one pass through the training set, the values stored at these counters are the fitness values of the rules. Duplicate rules are assigned a high risk. The algorithm ends when the number of generations reaches a limit or there are not changes in the global risk in certain number of generations. A detailed pseudocode of the generational scheme has been included in Appendix A.

4 Numerical Results

The experimental validation comprises nine datasets, originated in two problems related to linguistic classification systems with imprecise data (diagnosis of dyslexia [49] and future performance of athletes [47]). We have asked the experts that helped us with these problems to express their preferences about the classification results either with intervals or linguistic values. We intend to show that the algorithm proposed in this paper is able to exploit the subjective costs given by the human experts and produce a fuzzy rule based classification system according to their preferences.

These datasets contain imprecision in both the input and the output variables. Regarding the imprecision in the output, those instances with uncertainties in the class label can be regarded as multi-label data [9]. Nevertheless, observe that we do not intend to predict the crisp or fuzzy sets of classes assigned to those instances; we interpret a multi-label instance as an individual whose category was not clear to the expert, but he/she knows for sure a set of classes that this instance does not belong to. For instance, when diagnosing dyslexia, there were cases where the psychologist could not decide whether a child had dyslexia or an attention disorder. This does not mean that we should label the child as having both problems; on the contrary, the most precise fact we can attest about this child is that he should not be classified as "not dyslexic".

We have also observed that, in some cases, the use of a cost matrix produces rule bases that improve the results obtained with the same algorithm and a zero-one loss. We attribute this interesting result to the fact that the use of costs modifies the default

exploratory behavior of the genetic algorithm, making that some regions of the input space with a low density of examples are able to source rules that are still competitive in the latter stages of the learning. This effect will be further studied later.

The structure of this section is as follows: in the first place, we will describe an experiment illustrating the differences between numerical, interval-valued and linguistic (fuzzy) costs from the point of view of the human expert. Second, the datasets are described, and the experimental setting introduced, including the cost matrices, the metrics used for evaluating the results and those mechanisms we have used for removing the uncertainty in the data (needed for comparing this algorithm to other classification systems that cannot use imprecise data). The compared results between the new GFS and other alternatives are included at the end of this part.

4.1 Illustrative example

We have carried a small experiment for assessing the coherence of a subjective assignment of costs in classification problems. Our experts were asked to provide either a numerical cost, or a range of numbers or a linguistic term for each type of misclassification, according to their own preferences.

Our catalog of linguistic terms comprises eleven labels, described in Table 1, where their semantics are defined by means of trapezoidal fuzzy intervals, described in turn by four parameters (lowest element of the support, lowest element of the mode, highest element of the mode, highest element of the support). The left and rightmost terms "Absolutely low" and "Unacceptable" are crisp labels, following a requirement of one of the experts. Apart from this, experts were not explained this semantic; their choice was guided by the linguistic meaning they attributed to each label by themselves.

Linguistic term	**Fuzzy membership**
Absolutely-low	(0,0,0,0)
Insignificant	(0,0.052,0.105,0.157)
Very Low	(0.105,0.157,0.210,0.263)
Low	(0.210,0.263,0.315,0.368)
Fairly-low	(0.315,0.368,0.421,0.473)
Medium	(0.421,0.473,0.526,0.578)
Medium-high	(0.526,0.578,0.631,0.684)
Fairly-high	(0.631,0.684,0.736,0.789)
High	(0.736,0.789,0.842,0.894)
Very-high	(0.842,0.894,0.947,1)
Unacceptable	(1,1,1,1)

Table 1: Linguistic terms and parameters defining their membership functions.

The experts we are working with, that is to say both the expert in athletism and the expert in dyslexia, found natural to use the linguistic terms. When they asked to use numbers or intervals they made a conversion table and used their prior linguistic selection to find an equivalent numerical score, to which they assigned an amplitude reflecting their uncertainty about the number. Generally speaking, there were large overlappings between their intervals. For example, an expert had not conflicts choosing the linguistic cost of misclassification "Fairly-high" between the eleven alternatives, but assigned to the same subjective cost the interval $[0.55, 0.85]$. There was also consensus assigning the highest cost to those cases where the result of a misclassification had

undesired consequences. Interestingly enough, if the experts are asked to use a scale different than $[0, 1]$ (between 1 and 1000, for instance) their judgement was different. As an example, in the Table 2 we have collected the responses of an expert that was asked, at different times, to assign a cost to certain misclassification:

Scale	Interval	Number	Linguistic term
[0, 1]	[0.8, 1]	1	High
[1, 1000]	[700, 850]	800	High

Table 2: Answers of an expert when asked to assign a cost to certain misclassification.

In this example, the expert was consistent in the selection of a linguistic value, not so when selecting a numerical value: the first time he was asked, he chose the highest numerical cost (1) for a decision he did not associate the highest linguistic cost to. Furthermore, when the scale was changed, the numerical cost was different too, and in this last case this cost was similar to the corresponding trapezoidal fuzzy set in Table 1. Generally speaking, we can conclude that the linguistic assignment of cost was preferred to the numerical assignment, and that a subjective assignment of numbers or ranges to costs produces less coherent results than linguistic values. This result will be illustrated later with numerical experiments: we will show that the classification systems obtained when the expert builds a linguistic cost matrix have a confusion matrix that is preferable to that of the rule base arising from the interval-valued cost matrix.

4.2 Description of the datasets

The datasets "Diagnosis of the Dyslexic" and "Athletics at the Oviedo University", have been introduced in [49] and [47], respectively, and are available in the data set repository of keel-dataset (http://www.keel.es/datasets.php) [3, 4]. Their description is reproduced here for the convenience of the reader.

Dyslexia can be defined as a learning disability in people with normal intellectual coefficient, and without further physical or psychological problems that can explain such disability. The dataset "Diagnosis of the Dyslexic" is based on the early diagnosis (ages between 6 and 8) of schoolchildren of Asturias (Spain), where this disorder is not rare. All schoolchildren at Asturias are routinely examined by a psychologist that can diagnose dyslexia (in Table 3 there is a list of the tests that are applied in Spanish schools for detecting this problem). It has been estimated that between 4% and 5% of these schoolchildren have dyslexia. The average number of children in a Spanish classroom is 25, therefore there are cases at most classrooms [2]. Notwithstanding the widespread presence of dyslexic children, detecting the problem at this stage is a complex process, that depends on many different indicators, mainly intended to detect whether reading, writing and calculus skills are being acquired at the proper rate. Moreover, there are disorders different than dyslexia that share some of their symptoms and therefore the tests not only have to detect abnormal values of the mentioned indicators; in addition, they must also separate those children which actually suffer dyslexia from those where the problem can be related to other causes (inattention, hyperactivity, etc.).

The problem "Athletics at the Oviedo University" comprises eight different datasets, whose descriptions are as follows:

Category	Test	Description
Verbal comprehension	BAPAE	Vocabulary
	BADIG	Verbal orders
	BOEHM	Basic concepts
Logic reasoning	RAVEN	Color
	BADIG	Figures
	ABC	Actions and details
Memory	Digit WISC-R	Verbal-additive memory
	BADIG	Visual memory
	ABC	Auditive memory
Level of maturation	ABC	Combination of different tests
Sensory-motor skills	BENDER	visual-motor coordination
	BADIG	Perception of shapes
	BAPAE	Spatial relations, Shapes, Orientation
	STAMBACK	Auditive perception, Rhythm
	HARRIS/HPL	Laterality, Pronunciation
	ABC	Pronunciation
	GOODENOUGHT	Spatial orientation, Body scheme
Attention	Toulose	Attention and fatigability
	ABC	Attention and fatigability
Reading-Writing	TALE	Analysis of reading and writing

Table 3: Categories of the tests currently applied in Spanish schools for detecting dyslexia when an expert evaluates the children.

1. Dataset “B200ml-I”: This dataset is used to predict whether an athlete will improve certain threshold in 200 meters. All the indicators or inputs are fuzzy-valued and the outputs are sets.

2. Dataset “B200mlP”: Same dataset as “B200mlI”, with an extra feature: the subjective grade that the trainer has assigned to each athlete. All the indicator are fuzzy-valued and the outputs are sets.

3. Dataset “Long”: This dataset is used to predict whether an athlete will improve certain threshold in the long jump. All the features are interval-valued and the outputs are sets. The coach has introduced his personal knowledge.

4. Dataset “BLong”: Same dataset as “Long”, but now the measurements or inputs are defined by fuzzy-valued data, obtained by reconciling different measurements taken by three different observers.

5. Dataset “100ml”: Used for predicting whether a threshold in the 100 metres sprint race is being achieved. Each measurement was repeated by three observers. The input variables are intervals and outputs are sets.

6. Dataset “100mlP”: Same dataset as “100mlI”, but the measurements have been replaced by the subjective grade the trainer has assigned to each indicator (i.e.“reaction time is low” instead of “reaction time is 0.1 seg”).

7. Dataset “B100mlI”: Same dataset as “100mlI”, but now the measurements are defined by fuzzy-valued data.

8. Dataset“B100mlP”: Same dataset as “100mlP”, but now the measurements are defined by fuzzy-valued data.

A brief summary of the statistics of these problems is provided in Table 4. The name, the number of examples (Ex.), number of attributes (Atts.), the classes (Classes) and the fraction of patterns of each class (%Inst_classes) of each dataset are displayed. Observe that these fractions are intervals, because the class labels of some instances are imprecise, and can be used for computing a range of imbalance ratios.

Dataset	**Ex.**	**Atts.**	**Classes**	**%Inst_classes**
B200mlI	19	4	2	([0.47,0.73],[0.26,0.52])
B200mlP	19	5	2	([0.47,0.73],[0.26,0.52])
Long	25	4	2	([36,64],[36,64])
BLong	25	4	2	([36,64],[36,64])
100mlI	52	4	2	([0.44,0.63],[0.36,0.55])
100mlP	52	4	2	([0.44,0.63],[0.36,0.55])
B100mlI	52	4	2	([0.44,0.63],[0.36,0.55])
B100mlP	52	4	2	([0.44,0.63],[0.36,0.55])
Dyslexic-12	65	12	4	([0.32,0.43],[0.07,0.16], [0.24,0.35],[0.12,0.35])

Table 4: Summary descriptions of the datasets used in this study.

4.3 Experimental settings

All the experiments have been run with a population size of 100, probabilities of crossover and mutation of 0.9 and 0.1, respectively, and limited to 100 generations. The fuzzy partitions of the labels are uniform and their size is 5. All the imprecise experiments were repeated 100 times with bootstrapped resamples of the training set and where each partition of test contains 1000 tests.

For those experiments involving preprocessed data, the GFS proposed in [50] is used, with three nearest neighbors. This algorithm balances all the classes taking into account the imprecise outputs. This method of preprocessing is also applied to the 100 bootstrapped resamples of the training set.

4.3.1 Matrix of misclassification costs

The cost matrices used in the different datasets of Athletics [47] are shown in Tables 5 and 6. In both tables the expert preferred to discard a potentially good athlete (class 1) over accepting someone who is not scoring good marks (class 0). The actual costs depend on the event, as shown in Tables 5 (intervals) and 6 (linguistic terms). Observe that in Table 5 the costs are defined either by interval or crisp values; we commanded the experts to define the costs by means of numerical values, and to use intervals when they could not precise the numbers.

	Jump			**100-200m**			**B100-B200m**	
	Estimated labels			Estimated labels			Estimated labels	
True class	0	1	True class	0	1	True class	0	1
0	0	[0.6,0.9]	0	0	0.8	0	0	[0.8,0.94]
1	0.5	0	1	0.4	0	1	[0.15,0.24]	0

Table 5: Interval cost matrices designed by a human expert in Athletics datasets.

	Jump			**100-200m and B100-B200m**	
	Estimated labels			Estimated labels	
True class	0	1	True class	0	1
0	Absolutely-low	Fairly-high	0	Absolutely-low	High
1	Low	Absolutely-low	1	Very low	Absolutely-low

Table 6: Linguistic cost matrices designed by a human expert in Athletics datasets.

The Dyslexic's dataset is more complex and the expert decided by herself that her numerical assignments were not reliable, recommending us a design based on her linguistic matrix instead. The initial design was intended to separate dyslexic children ("class 2") from those in need of "control and review" ("class 1") and those without the problem. This is akin to an imbalanced problem, albeit there were some problems derived from this initial assignment of costs. For instance, in the case that a child is not dyslexic ("class 0") and the classifier indicates that he has a learning problem different than dyslexia ("class 4"), the misclassificacion cost was "Absolutely-low", because the expert was understanding that the classifier would indicate that the child is not dyslexic. However, the expert did not take into account that, in this case, this child

would be subjected to psychological treatment, which could potentially cause him certain disorders. The same situation happened when the child has an attention disorder and the classifier indicates that the child is dyslexic. In this case, the misclassification cost was "Very-high", according to the idea that is was more important to mark off the dyslexic children than leaving a dyslexia case undetected. Again, the consequences can be negative for the misclassified child, thus a finer distinction is needed. The second design takes into account this possibilities, and is shown in the Table 7.

Dyslexic-12				
	Classifier			
True class	0	1	2	4
0	Absolutely-low	Medium	Very-high	Unacceptable
1	Fairly-low	Absolutely-low	Low	Very-high
2	Unacceptable	Medium	Absolutely-low	High
4	High	High	Low	Absolutely-low

Table 7: Linguistic cost matrix designed by a human expert in Dyslexic's dataset.

4.3.2 Metrics for evaluating the results

The classification cost, when a zero-one loss is used, is the fraction of misclassified instances. For instance, regarding the confusion matrix of a two-classes problem,

	Negative Prediction	Positive Prediction
Negative class	TN	FP
Positive class	FN	TP

this cost is

$$\text{loss}_{0-1} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{61}$$

For evaluating this error with imprecise data we will use the same expressions introduced in Section 2 for the minimum risk problem (Eqs. 23 and 34) with a cost matrix $B = [1 - \delta_{ij}]$. Using this binary cost matrix we can also generalize the zero-one loss to multiclass problems either with crisp, interval or imprecise data.

For different cost matrices we will compare algorithms on the basis of the value $\text{loss}_{MR} = \widetilde{R}$, as defined in Eqs. 23 and 34. Other commonly used metrics, like the Area Under the ROC Curve (AUC) [7, 52] have not been used in this work because a suitable generalization to multi-class imprecise problems has not yet been proposed.

4.3.3 Heuristics for the removal of meta-information

For those comparisons involving statistical or intelligent classifiers unable to accept imprecise data, a procedure for removing the meta-information in the data is needed. The rules that will be used in this paper are as follows:

- If the meta-information is in the input, each interval is replaced by its midpoint. In case the data is fuzzy, the midpoint of its modal interval is chosen instead.
- If the imprecision is in the class label, each sample is replicated for the different alternatives. For instance, an example (x=2, c=A,B) is converted in two examples (x=2, c=A) and (x=2, c=B).

Observe that each time an example is replicated the remaining instances have to be repeated the number of times needed for preserving the statistical significance of each object. The drawback of this procedure is that problems which seem to be simple by the standards of crisp classification systems become complex datasets when the uncertainty is removed. For example, Dyslexic-12, with 65 instances and a high degree of imprecision, is transformed into a crisp dataset with thousands of instances.

4.4 Compared results

In this section we will compare the performance of the different alternatives in the design of the new GFS, and the results of this new GFS to those of different classifiers. The experiments are organized as follows:

1. GFS with linguistic cost matrices vs. GFS with interval-valued costs.
2. GFS with zero-one loss vs. GFS with minimum risk-based loss.
3. A selection of crisp classifiers vs. GFS with minimum risk-based loss.
4. GFS for imbalanced data vs. GFS with risk-based loss.

4.4.1 Interval and fuzzy costs

With this experiment we compare the behaviors of the GFSs depending on numerical costs (interval-valued costs) to those depending on linguistic costs. We will study the confusion matrix of the classifiers obtained with interval-valued and fuzzy risks, using the matrices that the experts provided for each case. For computing a numerical confusion matrix we have applied the procedure described in Section 4.3.3 and extended the test set by duplicating the imprecise instances.

In the athletics problems, the coach prefers to label an athlete as not relevant ("class 0") when he/she is relevant ("class 1") than the opposite, thus the misclassification ("label a C_1 case as if it was C_0") is preferred over (C_0 as C_1). In Table 8 we show that the percentage of misclassifications "C_1 as C_0" achieved with the linguistic cost matrix is higher (82,52%) than that obtained with interval-valued costs (78,15%). Observe that we do not claim with this experiment that there is not a numerical or interval-valued set of costs that produces a classifier improving, in turn, this result: our point is that a linguistic description of weights models better the subjective preferences of the expert, and our system was able to exploit this linguistic description for evolving a rule base that follows the preferences of the user.

As mentioned in Section 4.3.1, the expert in the field of dyslexia decided that her numerical assigments were not reliable. For comparing the results obtained after her selection of a numerical cost matrix (comprising intervals and real numbers) with those obtained with the corresponding linguistic cost matrix we have built Table 9. Each row "C_p as C_q" shows the number of children for which the output of the classifier was C_p when the value should have been C_q. Observe that there are improvements for all the combinations but "C_2 as C_0", and the global number of misclassifications is also reduced.

4.4.2 Comparison between GFSs using Loss_{0-1} and Loss_{MR}

In this section the minimum error-based extended cooperative-competitive algorithm defined in [48] (labelled "GFS") will be compared to the minimum risk-based GFS in

	Interval-values		Linguistic terms	
Dataset	C_0 **as** C_1 **(FP)**	C_1 **as** C_0 **(FN)**	C_0 **as** C_1 **(FP)**	C_1 **as** C_0**(FN)**
100mlI	1752	5363	1038	6228
100mlP	1609	4795	970	6078
B100mlI	975	6268	991	6258
B100mlP	1085	5590	1103	5558
Long	1785	3289	1638	3385
BLong	1891	3418	1609	3762
B200mlI	182	2511	182	2479
B200mlP	165	2570	164	2577
%	21.85%	78.15%	17.48%	82.52%

Table 8: Misclassifications in the Athletics datasets from MR_GFS with a cost matrix defined by interval-valued and linguistic costs.

Dyslexic	Interval-values	Int. N.	Linguistic terms	Ling. N.
C_0 as C_4	[0.6,0.9]	195	Unacceptable	124
C_2 as C_0	1	62	Unacceptable	80
C_4 as C_1	0.75	87	High	44
C_2 as C_4	0.6	329	High	272
C_1 as C_4	0.6	27	Very-high	16
C_0 as C_2	0.7	5326	Very-high	5314
C_2 as C_1	[0.3,0.35]	202	Medium	112
C_0 as C_1	[0.2,0.4]	424	Medium	359
Total N.	-	6652	-	6321

Table 9: Misclassifications in the Dyslexic datasets from MR_GFS with a cost matrix defined by interval-valued and linguistic costs.

this paper (labelled "MR_GFS"). Each rule base will be evaluated twice on the same test sets, using both a minimum-risk based criterion (Loss_{MR}) and a zero-one loss (Loss_{0-1}). Observe that the zero-one loss is the fraction of misclassified examples, or in other words the minimum-error based criterion.

In the first place, let us compare the misclassification rate (Loss_{0-1}) of "GFS" and "MR_GFS". It was expected that the cost-based classifier obtained the worst results, since it has not been designed for optimizing the zero-one loss. Rather surprisingly, the first two columns of Table 10 (GFS $Loss_{0-1}$ and MR_GFS $Loss_{0-1}$) contain evidence that the use of the new algorithm has improved the absolute number of misclassifications with respect to its minimum error-based counterpart in most datasets.

The statistical relevance of these differences has been graphically displayed in Figures 2 and 3. Each point in these figures represents one of the experiments. The abscissa is Loss_{0-1} (i.e. the fraction of errors) of the first approach and the ordinate is same type of risk for the second procedure. That is to say, points over the diagonal (circles) are the cases where the minimum error-based classifier produced a better rule set. Since the risks are interval-valued in this example, the figures are divided in two parts. The left part contains the comparison between the lower bounds of the risk, and the right part displays the upper bounds.

In this particular experiment we have selected two representative cases: the datasets

Dataset	**GFS** $Loss_{0-1}$	sup. std dev	**MR_GFS** $Loss_{0-1}$	sup. std dev	**GFS** $Loss_{MR}$	sup. std dev	**MR_GFS** $Loss_{MR}$	sup. std dev
100mlI	[0.176,0.378]	0.266	[0.178,0.380]	0.267	[0.075,0.166]	0.141	[0.044,0.104]	0.091
100mlP	[0.176,0.355]	0.249	[0.188,0.367]	0.254	[0.081,0.163]	0.144	[0.046,0.099]	0.075
B100mlI	[0.172,0.369]	0.267	[0.188,0.385]	0.270	[0.073,0.155]	0.140	[0.048,0.104]	0.091
B100mlP	[0.160,0.349]	0.263	[0.161,0.350]	0.263	[0.075,0.162]	0.152	[0.043,0.100]	0.084
Long	[0.321,0.590]	0.379	[0.288,0.557]	0.399	[0.168,0.315]	0.236	[0.129,0.236]	0.170
BLong	[0.326,0.625]	0.405	[0.286,0.586]	0.397	[0.203,0.394]	0.299	[0.140,0.265]	0.181
B200mlI	[0.232,0.473]	0.378	[0.174,0.418]	0.366	[0.098,0.154]	0.163	[0.047,0.094]	0.087
B200mlP	[0.262,0.480]	0.363	[0.215,0.433]	0.331	[0.092,0.152]	0.191	[0.049,0.095]	0.078
Partial mean	[0.227,0.451]	0.321	**[0.210,0.435]**	0.318	[0.107,0.207]	0.183	**[0.068,0.137]**	0.107
Dyslexic-12	**[0.447,0.594]**	0.240	[0.502,0.613]	0.196	[0.309,0.418]	0.184	**[0.277,0.377]**	0.162
Global mean	[0.252,0.467]	0.280	**[0.243,0.455]**	0.257	[0.129,0.259]	0.183	**[0.091,0.184]**	0.134

Table 10: Behaviour of "GFS" and "MR_GFS" with respect to Loss_{0-1} and Loss_{MR}.

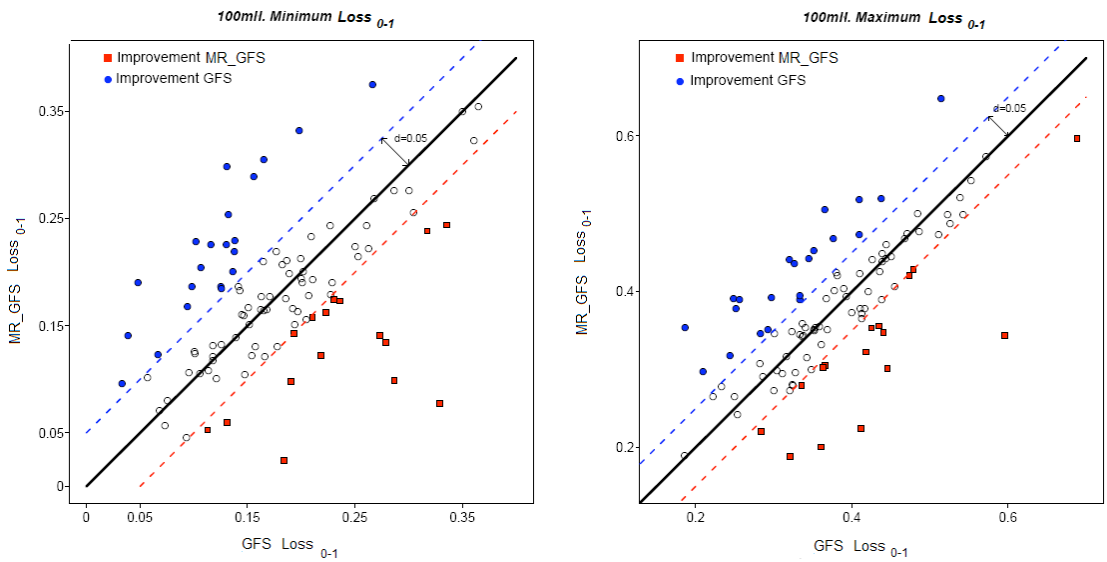

Figure 2: Behaviour of "GFS" and "MR_GFS" respect to $Loss_{0-1}$ in 100mlI. **Left figure:** Lower bounds. **Right figure:** Upper bounds.

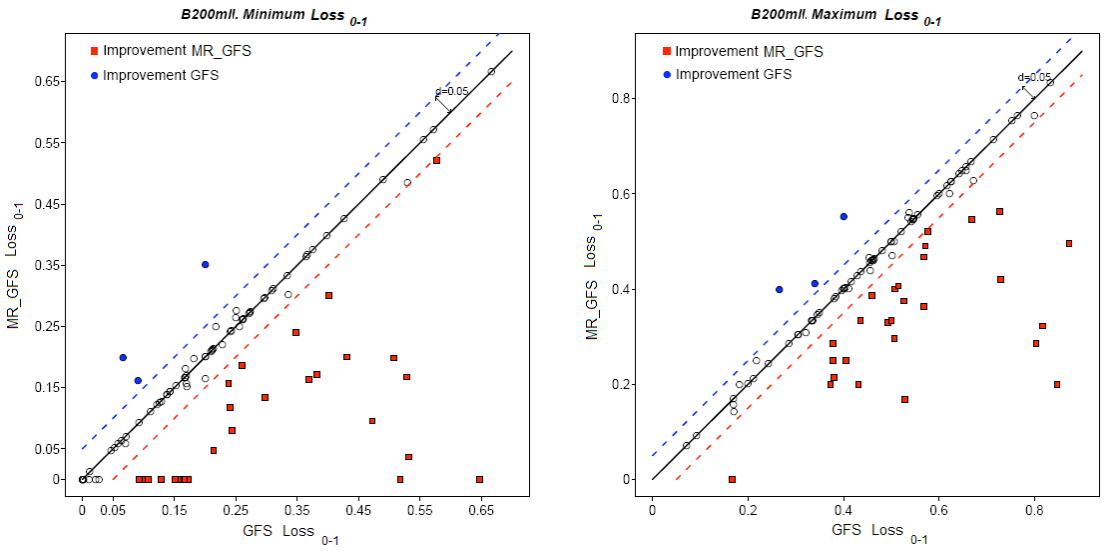

Figure 3: Behaviour of "GFS" and "MR_GFS" respect to $Loss_{0-1}$ in B200mII. **Left figure:** Lower bounds. **Right figure:** Upper bounds.

"100mII" and "B200mII". In Figure 2 we have shown the results of "100mII", where there is not a significant difference between either algorithm, thus the points are equally distributed above and below the diagonal. In Figure 3 we have included the results of "B200mII" where there is clear advantage of the new algorithm, in a problem where one may think that this difference should favor the minimum error-based procedure. We attribute this result to the fact that the use of costs modifies the default exploratory behavior of the genetic algorithm, making that some regions of the input space with a low density of examples are able to source rules that are still competitive in the latter stages of the learning.

On the contrary, the two last columns of Table 10 show the expected result: since the new GFS is optimizing the risk function, the risk of the minimum error-based algorithm is higher than the risk of the classifiers obtained with the approach in this paper. The graphical assessment of the relevance is displayed in Figure 4, for the dataset "100mII".

The improvements of "MR_GFS', in Athletics datasets are also shown in Table 11. Observe that the percentage of misclassifications "C_1 as C_0" is higher with "MR_GFS" (82.52%) than it is with "GFS" where the percentages are 56.94% and 43.96%, approaching 50% each, as expected in a minimum error-based problem.

The linguistic quality of the rules obtained by GFS and MR_GFS will be shown by means of an example. In Table 12 we have included two knowledge bases found by GFS and MR_GFS when the dataset 100mII is considered. Both bases are mostly similar, but some rules have different consequent parts. For instance, rule number 5 of the knowledge base produced by the algorithm "GFS" (see Table 12) is "IF ratio is High and reaction time is High and 20 meters speed is Medium and 40 meters speed is Medium then class is Relevant", while the same rule has the opposite consequent if the algorithm MR_GFS is used. This last rule is preferred, as the antecedent does not

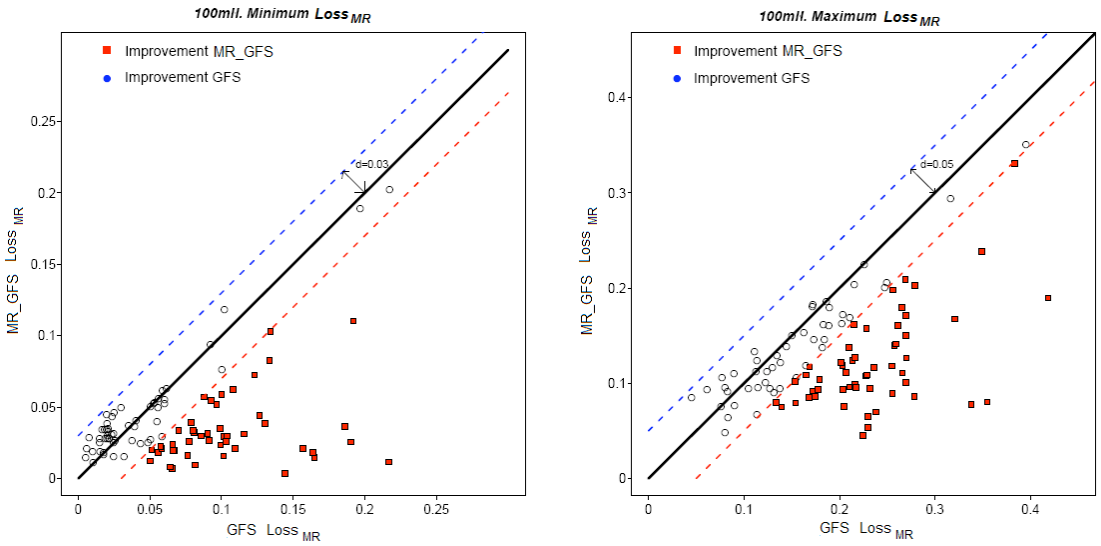

Figure 4: Behaviour of the "GFS" and "MR_GFS" with respect to $Loss_{MR}$ in 100mII. **Left figure:** Lower bounds. **Right figure:** Upper bounds.

clearly matches individuals that should be selected for taking part of the competition and the expert stated with his cost matrix that in case of doubt the decision should be "Not Relevant".

Let us also justify the use of graphical methods for assessing the statistical relevance of the differences. We have used these graphs because (up to our knowledge) a method for computing the p-values of a suitable statistical test for assessing the relevance of the differences between two imprecise samples has not been published yet (the closest reference –regarding bootstrap tests for imprecise data– is [19], where the computation of the p-value is shown for a generic case, but the selection of a test matching this application is not addressed). Nevertheless, in our opinion the relevance of the differences is clearly perceived in the mentioned graphs, that provide more insightful

	GFS		MR_GFS	
Dataset	C_0 **as** C_1 **(FP)**	C_1 **as** C_0 **(FN)**	C_0 **as** C_1 **(FP)**	C_1 **as** C_0**(FN)**
100mlI	2867	4346	1038	6228
100mlP	2974	3863	970	6078
B100mlI	2693	4298	991	6258
B100mlP	2945	3754	1103	5558
Long	3168	2186	1638	3385
BLong	3720	2005	1609	3762
B200mlI	696	2355	182	2479
B200mlP	686	2368	164	2577
%	43.96%	56.04%	17.48%	82.52%

Table 11: Misclassifications of Athletic's datasets obtained with "GFS" and "MR_GFS".

Id.	Antecedentet Rule	Consequent GFS	Consequent MR_GFS
1	IF ratio is Very-high and reaction time is Low and 20 meters speed is Medium and 40 meters speed is Low	Relevant	Relevant
2	IF ratio is Very-low and reaction time is Low and 20 meters speed is High and 40 meters speed is Medium	Not Relevant	Not Relevant
3	IF ratio is Medium and reaction time is Medium and 20 meters speed is High and 40 meters speed is Medium	Relevant	Not Relevant
4	IF ratio is High and reaction time is High and 20 meters speed is Very-high and 40 meters speed is Very-high	Not Relevant	Not Relevant
5	IF ratio is High and reaction time is High and 20 meters speed is Medium and 40 meters speed is Medium	Relevant	Not Relevant

Table 12: Knowledge bases obtained with GFS and MR_GFS, dataset 100mlI.

information than the bounds of the p-value.

4.4.3 Comparison with algorithms for imbalanced data

Given the results in Table 11 and the *a priori* probabilities of the different classes in the problems being studied (see Table 4), it makes sense to regard some of these imprecise datasets as imbalanced. Hence, in this section we compare the new cost-based algorithm with other, different techniques better suited for imbalanced vague data than the minimum error approach. For instance, the data can be preprocessed and new instances introduced before the learning phase [5, 25, 26, 27, 28]. Furthermore, it may be argued that a minimum error-based classifier would produce results similar to that obtained with the linguistic cost approach we are suggesting in this paper. To this we can answer that preprocessing for balancing data in two classes problems is indeed roughly equivalent to use a cost matrix whose diagonal is zero and the remaining elements are the inverses of the *a priori* probabilities from the preferences of the expert, but this implicit cost matrix might or might not reproduce the needs of the expert, while our linguistic approach is based on his/her preferences. A similar situation occurs in multi-class imbalanced problems, that again can be regarded as cost-based problems, albeit a more complex cost matrix would be needed in this case.

The experimental data supporting our preceding discussion is in Table 13, when we compare the classification error of the minimum risk-based algorithm (column "MR_GFS, Loss_{0-1}"), with the same algorithm over a preprocessed dataset (column "FS_MR_GFS, Loss_{0-1}"). The last three columns of this table contain the risks of the same rule bases and we have also added to them the measured risk of the minimum error-based algorithm over the preprocessed dataset (column "FS_GFS, Loss_{MR}"). Observe that the fraction of errors of "MR_GFS" tends to be lower when it is executed over the preprocessed data, however the risk is better if the data is not altered, therefore there is no reason for applying this stage, which in addition has an elevated computational cost. With respect to our initial question, that was comparing the minimum error-based classifier with preprocessed input data with the minimum risk approach, the latter is clearly better, as shown in Table 13 and also in Figure 5.

Dataset	**MR_GFS** $Loss_{0-1}$	**FS_MR_GFS** $Loss_{0-1}$	**MR_GFS** $Loss_{MR}$	**FS_MR_GFS** $Loss_{MR}$	**FS_GFS** $Loss_{MR}$
100mlI	[0.178,0.380]	[0.185,0.386]	[0.044,0.104]	[0.038,0.091]	[0.099,0.209]
100mlP	[0.188,0.367]	[0.201,0.380]	[0.046,0.099]	[0.043,0.091]	[0.084,0.167]
B100mlI	[0.188,0.385]	[0.201,0.398]	[0.048,0.104]	[0.040,0.089]	[0.095,0.199]
B100mlP	[0.161,0.350]	[0.169,0.358]	[0.043,0.100]	[0.039,0.089]	[0.087,0.177]
Long	[0.288,0.557]	[0.300,0.569]	[0.129,0.236]	[0.124,0.219]	[0.133,0.269]
BLong	[0.286,0.586]	[0.296,0.596]	[0.140,0.265]	[0.131,0.242]	[0.152,0.326]
B200mlI	[0.178,0.418]	[0.125,0.369]	[0.047,0.094]	[0.031,0.078]	[0.182,0.286]
B200mlP	[0.215,0.433]	[0.184,0.402]	[0.049,0.095]	[0.046,0.097]	[0.201,0.307]
Partial mean	[0.210,0.435]	[0.206,0.431]	**[0.068,0.137]**	[0.084,0.151]	[0.128,0.242]
Dyslexic-12	[0.502,0.613]	[0.504,0.615]	[0.277,0.377]	[0.279,0.379]	[0.309,0.418]
Global mean	[0.243,0.455]	[0.239,0.451]	**[0.091,0.184]**	[0.105,0.176]	[0.148,0.261]

Table 13: Behaviour of "MR_GFS", "FS_MR_GFS" and "FS_GFS" with respect to Loss_{0-1} and Loss_{MR}.

5 Concluding remarks

In this work we have defined a GFS that solves the minimum risk classification problem for imprecise data, where the cost matrix needs not to be precisely described, but it can be expressed with linguistic terms. We have extended first the concepts of confusion matrix and expected risk to interval valued and fuzzy data, and a Genetic Cooperative Competitive algorithm has been defined which can evolve a rule base that minimizes this extended risk, being understood that this minimization is done with respect to a certain precedence operator between fuzzy values.

The experimental results have evidenced that the use of linguistic terms is preferred to numerical costs or intervals, as the experts are able to express their preferences in a more consistent way. We have also shown that the new rule bases significantly improve the expected risk of former GFSs, and in certain cases the improvement in the risk is also accompanied by an enhanced misclassification rate. In the last place, we conclude that preprocessing the data for balancing the probabilities of the classes is not justified for this problem, as the implicit costs in a preprocessing stage will be in all likelihood different than the preferred costs of the expert.

Acknowledgements

This work was supported by the Spanish Ministry of Education and Science, under grant TIN2008-06681-C06-04, and by Principado de Asturias, PCTI 2006-2009.

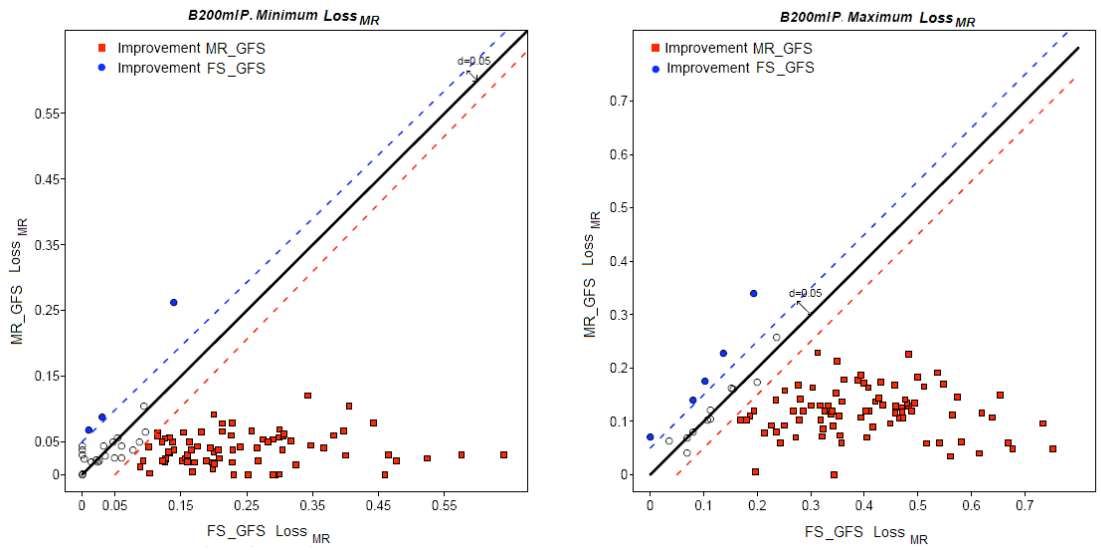

Figure 5: Behaviour of the "MR_GFS" and "FS_GFS" with respect to $Loss_{MR}$ in B200mlP. **Left figure:** Lower bounds. **Right figure:** Upper bounds.

A Pseudocode of the algorithm

The pseudocode of the genetic algorithm defined in this paper is included in this appendix. This algorithm depends on three modules. The first one defines the generational scheme and is as follows:

function GFS
Initialize population
for iter in $\{1, \ldots, \mathrm{Iterations}\}$ and equal_generations < 20
for sub in $\{1, \ldots, \mathrm{subPop}\}$
Select parents
Crossover and mutation
assignImpreciseConsequentt(offspring)
end for sub
Replace the worst subPop individuals
assignImpreciseFitnessApprox(population,dataset)
end for iter or equal_generations
Purge unused rules
return population

Observe that only the antecedent is represented in the genetic chain. The second module is used for determining the consequent that best matches a given antecedent, and is as follows:

function assignImpreciseConsequent(rule)
for c in $\{1, \ldots, \mathrm{C}\}$
grade = 0
compExample = 0
for k in $\{1, \ldots, \mathrm{D}\}$
m = fuzMembership(Antecedent,k,c)
for d in $\{1, \ldots, \mathrm{C}\}$
cost= cost $\oplus$ ($\widetilde{b}_{cd} \otimes \widetilde{Y}_k$(d))
end for d
grade = grade $\oplus$ ($m \oplus$ cost)
end for K
weight[c] = grade
end for c
mostFrequent = $\{1, \ldots, \mathrm{C}\}$
for c in $\{1, \ldots, \mathrm{C}\}$
for c_1 in $\{c+1, \ldots, \mathrm{C}\}$
if (weight[c] **dominates** weight[c_1]) **then**
mostFrequent = mostFrequent - { c_1 }
end if
end for c_1
end for c
Consequent = select(mostFrequent)
CF[rule] = computeConfidenceOfConsequent
return rule

In the last place, the third module is used for assigning the fitness values to the members of the population:

function assignImpreciseFitnessApprox(population,dataset)
for k in $\{1, \ldots, \mathrm{D}\}$
setWinnerRule = $\emptyset$

for r in $\{1, \ldots, \mathrm{M}\}$
dominated = FALSE
r.$\widetilde{\mathrm{m}}$ = fuzMembership(Antecedent[r],example)
for sRule in setWinnerRule
if (sRule **dominates** r) **then**
dominated = TRUE
end if
end for sRule
if (**not** dominated **and** r.$\widetilde{\mathrm{m}} > 0$) **then**
for sRule in setWinnerRule
if (r.$\widetilde{\mathrm{m}}$ **dominates** sRule) **then**
setWinnerRule = setWinnerRule $-\{$ sRule $\}$
end if
end for sRule
setWinnerRule = setWinnerRule $\cup\{$ r $\}$
end if
end for r
if (setWinnerRule == $\emptyset$) **then**
setWinnerRule = setWinnerRule $\cup\{$ rule_freq_class $\}$
for r in setWinnerRule
for d in C
$\widetilde{fit}$[r] = $\widetilde{fit}$[r] $\oplus$ $(\widetilde{\delta}_{d,\tilde{Y}_k} \otimes \widetilde{b}_{q_r d})$
end for d
end for r
end for k
return fitness

References

[1] N. Abe, B. Zadrozny, J. Langford. An iterative method for multi-class cost-sensitive learning. In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2004) 3-11.

[2] J. Ajuriaguerra, Manual de psiquiatría infantil, Toray-Masson (1976).

[3] J. Alcalá-Fdez, L. Sánchez, S. García, M.J. del Jesus, S. Ventura, J.M. Garrell, J. Otero, C. Romero, J. Bacardit, V.M. Rivas, J.C. Fernández, F. Herrera. KEEL: A Software Tool to Assess Evolutionary Algorithms to Data Mining Problems. Soft Computing 13(3) (2009) 307-318.

[4] J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. García, L. Sánchez, F. Herrera. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. Journal of Multiple-Valued Logic and Soft Computing, in press (2010).

[5] G. Batista, R. Prati, M. Monard. A study of the behaviour of several methods for balancing machine learning training data. SIGKDD Explorations 6(1) (2004) 20-29.

[6] J. Berger. Statistical decision theory and Bayesian Analysis. Springer-Verlag (1985).

[7] A.P. Bradley. The use of the are under the ROC cruve in the evaluation of machine learning algorithms. Pattern Recognition 30(7)(1997) 1145-1159.

[8] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone. Classification and Regression Trees. Belmont, CA: Wadsworth (1984).

[9] M.R. Boutell, J. Luo, X. Shen, C.M. Brown. Learning multi-label scene classification. Pattern Recognition 37 (2004) 1757-1771.

[10] N.V. Chawla, N. Japkowicz, A. Kolcz. Special issue on learning from imbalanced data sets. SIGKDD Explorations 6 (1) (2004) 1-6.

[11] C.H. Cheng. A new approach for ranking fuzzy numbers by distance method. Fuzzy Sets and Systems 95 (1998) 307-317.

[12] S.J. Chen, S.M. Chen. A new method for handling multicriteria fuzzy decision making problems using FN-IOWA operators. Cybernetics and Systems 34 (2003) 109-137.

[13] S.J Chen, S.M. Chen. Fuzzy risk analysis based on the ranking of generalized trapezoidal fuzzy numbers. Applied Intelligence 26(1) (2007) 1-11.

[14] L.H. Chen, H.W Lu. An approximate approach for ranking fuzzy numbers based on left and right dominance. Computers and Mathematics with Applications 41(12) (2001) 1589-1602.

[15] S.M Chen, C.H. Wang. Fuzzy risk analysis based on ranking fuzzy numbers using a-cuts, belief features and signal/noise ratios. Expert Systems with Applications 36 (2009) 5576-5581.

[16] O. Cordón, M.J. del Jesus, F. Herrera. A proposal on reasoning methods in fuzzy rule-based classification systems. International Journal of Approximate Reasoning 20(1) (1999) 21-45.

[17] O. Cordón, F. Herrera, F. Hoffmann, L. Magdalena. Genetic fuzzy systems. Evolutionary tuning and learning of fuzzy knowledge bases. World Scientific, Singapore (2001).

[18] I. Couso, L. Sánchez. Higher order models for fuzzy random variables. Fuzzy Sets and Systems 159 (2008) 237-258.

[19] I. Couso, L. Sánchez. Mark-recapture techniques in statistical tests for imprecise data. International Journal of Approximate Reasoning (2010) doi:10.1016/j.ijar.2010.07.009.

[20] T.C Chu, C.T. Tsao. Ranking fuzzy numbers with an area between the centroid point and original point. Computers and Mathematics with Applications 43 (2002) 111-117.

[21] D. Dubois, H. Prade. The mean value of a fuzzy number. Fuzzy Sets and Systems 24(3) (1987) 279-300.

[22] P. Domingos, M. Pazzani. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. Machine Learning 29 (1997) 103-130.

[23] P. Domingos. MetaCost: a general method for making classifiers cost-sensitive. International Conference on Knowledge Discovery and Data Mining (1999) 155-164.

[24] C. Elkan C. The Foundations of Cost-Sensitive Learning. In Proceedings of the 17th International Joint Conference on Artificial Intelligence (2001) 973-978.

[25] A. Fernández, S. Garcia, M.J del Jesús, F. Herrera. A study of the behaviour of linguistic fuzzy rule based classification system in the framework of imbalanced data-sets. Fuzzy Sets and Systems 159 (2008) 2378-2398.

[26] A. Fernández, M.J. del Jesús, F. Herrera. On the 2-tuples based genetic tuning performance for fuzzy rule based classification systems in imbalanced data-sets. Information Sciences 180 (2010) 1268-1291.

[27] A. Fernández A, M.J del Jesús, F. Herrera. On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced data-sets. Expert Systems with Applications 36 (2009) 9805-9812.

[28] A. Fernández A., M.J. del Jesús, F. Herrera F. Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets. International Journal of Approximate Reasoning 50 (2009) 561-577.

[29] A. Gonzales, R. Pérez. SLAVE: A genetic learning system based on an iterative approach. IEEE Transactions on Fuzzy 7(22) (2002) 176-191.

[30] D.J. Hand. Discrimination and Classification. Wiley Series in Probability and Mathematical Statistics, Chichester (1981).

[31] S. Haykin. Neural Networks: a comprehensive foundation, 2nd Edition. Prentice Hall (1999).

[32] E. Hüllermeier, J. Fürnkranz J. Ranking by Pairwise Comparison: A Note on Risk Minimization. IEEE International Conference on Fuzzy Systems (2004).

[33] E. Hüllermeier E, J. Fürnkranz J. Learning label preferences: Ranking error versus position error. Advances in Intelligent Data Analysis, LNCS 3646 (2005) 180-191.

[34] H. Ishibuchi, T. Nakashima, T. Murata. A fuzzy classifier system that generates fuzzy if-then rules for pattern classification problems. In Proc. of 2nd IEEE CEC (1995) 759-764.

[35] R. Jain. Decision-making in the presence of fuzzy variables. IEEE Trans. Systems Man and Cybernet. SMC- 6 (1976) 698-703.

[36] R. Jain R. A procedure for multi-aspect decision making using fuzzy sets, Internat. J. Systems Sci. 8 (1978) 1-7.

[37] K. Kilic, O. Uncu, I.B. Türksen. Comparison of different strategies of utilizing fuzzy clustering in structure identification. Information Sciences 177(23) (1007) 5153-5162.

[38] C. Liang C, J. Wu, J. Zhang. Ranking indices and rules for fuzzy numbers based on gravity center point. Paper presented at the 6th World Congress on Intelligent Control and Automation, Dalian, China (2006) 315-3163.

[39] P. Limbourg. Multi-objective optimization of problems with epistemic uncertainty. in EMO (2005) 413-427.

[40] L.Z. Lin, H.R. Yeh. Fuzzy linguistic decision to provide alternatives to market mechanism strategies. Expert Systems with Applications 37 (2010) 6986-6996.

[41] D. Margineantu. Class probability estimation and cost-sensitive classification decisions. In Proceedings of the 13th European Conference on Machine Learning (2002) 270-281.

[42] D. Margineantu. Methods for cost-sensitive learning. Technical report, Department of Computer Science, Oregon State University, Corvallis (2001).

[43] M. Mazurowski, P. Habas, J. Zurada, J. Lo, J. Baker, G. Tourassi. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural Networks 21 (2-3) (2008) 427-436.

[44] Z. Michalewicz Z, C. Janikow. Handling Constraints in Genetic Algoritms. Conference of Genetic Algorithms (1991)

[45] Z. Michalewicz. Genetic algorithms + data structures = Evolution Programs. Springer (1998).

[46] S.K. Pal, D.P. Mandal. Linguistic recognition system based on approximate reasoning. Information Sciences 61 (1992) 135-161.

[47] A. Palacios, I. Couso, L. Sánchez. Future performance modeling in athletism with low quality data-based GFSs. Journal of Multiple-Valued Logic and Soft Computing, in press (2010).

[48] A. Palacios, L. Sánchez, I. Couso. Extending a simple Genetic Cooperative-Competitive Learning Fuzzy Classifier to low quality datasets. Evolutionary Intelligence 1(2) (2009) 73-84.

[49] A. Palacios, L. Sánchez, I. Couso. Diagnosis of dyslexia from vague data with Genetic Fuzzy Systems. International Journal of Approximate Reasoning 51 (2010) 993-1009.

[50] A. Palacios, L. Sánchez, I. Couso. Preprocessing vague imbalanced datasets and its use in genetic fuzzy classifiers. Paper presented at the IEEE World Congress on Computational Intelligence, Barcelona, Spain (2010).

[51] X. Peng, I. King. Robust BMPM training based on second-order cone programming and its application in medical diagnosis, Neural Networks 21(2-3) (2008) 450-457.

[52] F. Provost, T. Fawcett. Robust classification systems for imprecise enviroments. In: Proc. AAAI (1998) 706-713.

[53] P. Pulkkinen, J. Hytönen, H. Koivisto. Developing a bioaerosol detector using hybrid genetic fuzzy systems. Engineering Applications of Artificial Intelligence 21(8) (2008) 1330-1346.

[54] B.S. Shieh. An approach to centroids of fuzzy numbers. International Journal of Fuzzy Systems 9(1) (2007) 51-54.

[55] J. Teich J. Pareto-front exploration with uncertain objectives. in EMO (2001) 314-328.

[56] A. Teredesai, V. Govindaraju. GP-based secondary classifiers. Pattern Recognition 38(4) (2005) 505-512.

[57] E. Broekhoven, V. Adriaenssens, B. De Baets. Interpretability-preserving genetic optimization of linguistic terms in fuzzy models for fuzzy ordered classification: An ecological case study. International Journal of Approximate Reasoning 44(1) (2007) 65-90.

[58] R. Verschae, J.R. Del Solar, M. Köppen, R.V. Garcia. Improvement of a face detection system by evolutionary multi-objective optimization Proc. HIS (2005) 361-366.

[59] L.X. Wang L.X, J.M. Mendel. Generating fuzzy rules by learning from examples. IEEE Transactions on Systems, Man, and Cybernetics 22(6) (1992) 1414-1427.

[60] Y.J. Wang, H.S. Lee. The revised method of ranking fuzzy numbers with an area between the centroid and original points. Computers and Mathematics with Applications 55 (2008) 2033-2042.

[61] Y.M. Wang, Y. Luo. Area ranking of fuzzy numbers based on positive and negative ideal points. Computers and Mathematics with Applications 58 (2009) 1769-1779.

[62] F. Xia, Y. Yang, L. Zhou, F. Li, M. Cai, D. Zeng. A closed-form reduction of multi-class cost-sensitive learning to weighted multi-class learning. Pattern Recognition 42 (2009) 1572-1581.

[63] Z. Yao, F. Zhiping. Method for multiple attribute decision making based on incomplete linguistic judgment matrix. Journal of Systems Engineering and Electronics 19(2) (2008) 298-303.

[64] B. Zadrozny, C. Elkan. Learning and making decisions when costs and probabilities are both unknown. In Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2001) 204-213.

[65] B. Zadrozny B. One-benefit learning: cost-sensitive learning with restricted cost information. In: Proceedings of the 1st International Workshop on Utility-based Data Mining (2005) 53-58.

[66] Z.H. Zhou, X.Y Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledgeand Data Engineering 18(1) (2006) 63-77.

[67] Z.H. Zhou, X.Y. Liu. On multi-class cost-sensitive learning. In: Proceedings of the 21st National Conference on Artificial Intelligence (2006) 567-572.