2011 IEEE International Conference on Fuzzy Systems
June 27-30, 2011, Taipei, Taiwan

Using The Adaboost Algorithm For Extracting Fuzzy Rules From Low Quality Data: Some Preliminary Results

Ana M. Palacios
Departamento de Informática
Universidad de Oviedo, Spain
email: palaciosana@uniovi.es.

Luciano Sánchez
Departamento de Informática
Universidad de Oviedo, Spain
email: luciano@uniovi.es

Inés Couso
Departamento de Estadística e I.O. y D.M
Universidad de Oviedo, Spain
email: couso@uniovi.es

***Abstract*—When the Adaboost algorithm is used for extracting fuzzy rules from data, each rule is regarded as a weak learner, and knowledge bases as assimilated to ensembles. In this paper we propose an extension of this framework for obtaining fuzzy rule-based classifiers from imprecise data. In the new approach, the mentioned search of the best rule at each iteration is carried out by a genetic algorithm with a fuzzy fitness function. The instances will be assigned fuzzy weights, however each fuzzy rule will be associated to a crisp number of votes.**

***Index Terms*—Genetic Fuzzy Systems, Low Quality Data, Boosting**

I. Introduction

Boosting algorithms combine different classifiers for obtaining an ensemble that performs better than any of its components [1]. These algorithms repeatedly invoke a learning algorithm to successively generate a committee of simple, inaccurate classifiers. Each time a new simple classifier is added to this ensemble, the examples in the training set are re-weighted so that future classifiers will focus on the most difficult examples, and a voting strength is assigned to the classifier. The number of votes a classifier is given depends on the confidence in its classification accuracy, as measured on the training set.

Fuzzy Rule Based Classification Systems (FRBCS) can also be regarded as ensembles, where each fuzzy rule is matched to one of the mentioned simple classifiers. When an appropriate reasoning scheme is used, it has been shown that the mechanism used for obtaining the output of a FRBCS, given an input value, can be assimilated to a voting process [2], and the voting strength of each simple classifier is the product of the compatibility between the antecedent of the corresponding fuzzy rule and the degree of certainty of its consequent. This interpretation has been used before, and the Adaboost algorithm has been applied to the learning of fuzzy rules from data [3][4][5], in combination with a Genetic Algorithm (GA). The resulting learning algorithm is a Genetic Fuzzy System (GFS) that shares certain properties with Iterative Rule Learning approaches [6]. Furthermore, after the work of Friedman [7], this process has been regarded as a forward stepwise estimation of the statistical parameters defining a logit transform of a Generalized Additive Model, giving rise to the LogitBoost algorithm [7] and its genetic extension for learning FRBCS from data [8]. The use of boosting techniques for learning fuzzy rules from data is therefore a well known technique, whose main virtues are a fast learning and accurate results, but this use is not free from problems. The first drawback of the employment of Adaboost for learning FRBCS is in the use of a voting-based inference, that makes the linguistic interpretation of the Knowledge Base (KB) harder. This problem was addressed in [9], where a new algorithm was proposed for which the degrees of confidence in the consequents of the rules were iteratively adjusted until the results produced by a "winner-takes-all" type of inference were similar to that of voting. Nevertheless, there are other difficulties that are still open. A second issue with the application of Adaboost to the learning of fuzzy rules lies in the requirements imposed to the data, that must be precisely perceived: being based in Generalized Additive Models, Adaboost or Logitboost can cope with random uncertainty, but they are not able to process data when a combination of epistemic and random uncertainties is present [10]. Because of these reasons, this paper contains a new proposal for applying the Adaboost algorithm for learning FRBCS from low quality data [11], including both problems with imprecisely perceived features and problems with a partial lack of knowledge about the class labels attributed to some instances.

The organization of this work is as follows: the following section, briefly introduces the Adaboost algorithm, describes the type of fuzzy classifiers that are boosted and recall a method for boosting descriptive fuzzy rule bases [12]. In Section 3, we propose an extension of this algorithm for interval and fuzzy data. In Section 4 some properties of the proposed algorithms are evaluated on a number of data sets and compared to that of other GFS for low quality data. In Section 5 some concluding remarks and future works are discussed.

II. Adaboost and fuzzy rule based classifiers

At this point we introduce the basic notation employed throughout the paper. Let $\mathbf{X}$ be the feature space, and let $\mathbf{x}$ be a feature vector $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbf{X}$. Let p be the number of classes. The training set is a sample of m classified examples

978-1-4244-7317-5/11/$26.00 ©2011 IEEE

$(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i \in \mathbf{X}$, $1 \leq y_i \leq p$, $1 \leq i \leq m$.

The KB of the FRBCS comprises N rules. The antecedents of the rules are logical combinations of fuzzy logic asserts, whose degrees of truth are modeled by N fuzzy subsets $A^j \in \mathcal{F}(\mathbf{X})$, forming a fuzzy partition $\mathcal{A} = \{A^j\}_{j=1\ldots N}$ of the feature space. A fuzzy rule based classifier is therefore defined by means of a fuzzy relationship defined on $\mathcal{A} \times \{1, \ldots, p\}$. Values of this relationship describe the degrees of compatibility among the fuzzy subsets of the feature space collected in $\mathcal{A}$, and each of the classes. In other words, for every antecedent A^j there are p numbers between 0 and 1 that represent the confidence in the assertion "All elements in the fuzzy set A^j belong to class number k". Values close to 1 indicate "high confidence," and values close to 0 denote "absence of knowledge about the assertion."

In this paper, we will translate the former fuzzy relationship into linguistic statements by combining p terms

$$\text{compatibility}(A^j, c_k) = s_k \quad k = 1, \ldots p,$$

into a single sentence, as follows:

$$\text{if } \mathbf{x} \text{ is } A^j \text{ then truth}(c_1) = s_1^j \text{ and } \cdots \text{ and truth}(c_p) = s_p^j.$$

The antecedents A^j are decomposed in a Cartesian product of fuzzy sets defined over each feature, $A^j = A_1^j \times A_2^j \times \ldots \times A_n^j$, thus the rules are

$$\begin{array}{c}\text{if } x_1 \text{ is } A_1^j \text{ and } \ldots \text{ and } x_n \text{ is } A_n^j \\ \text{then truth}(c_1) = s_1^j \text{ and } \cdots \text{ and truth}(c_p) = s_p^j.\end{array}$$

An instance $\mathbf{x}$ is assigned to the class

$$\arg\max_{k=1,\ldots,p} \bigvee_{j=1}^{N} A^j(\mathbf{x}) \wedge s_k^j \tag{1}$$

where "$\wedge$" is the product and "$\vee$" is the arithmetic sum, so called "maximum voting scheme" [13].

A. The Adaboost algorithm

Let us define a set $\{g^1, g^2, \ldots, g^N\}$ of simple, but possibly unreliable binary classifiers. Boosting consists in combining these low quality classifiers (so called "weak hypotheses" in the boosting literature) with a voting scheme to generate an overall classifier that performs better than any of its individual constituents alone. It has been shown that a fuzzy rule can be regarded as a particular case of a weak hypothesis, and a fuzzy rule base can be interpreted as a weighted combination of weak hypotheses.

Weak hypotheses take feature values as input and produce both a class number as well as a degree of confidence in the given classification. In two-classes problems, these two outputs are encoded by a single real number, $g^j(\mathbf{x}) \in \mathbf{R}$, whose sign is interpreted as the label of $\mathbf{x}$ and whose absolute value is interpreted as the confidence in the classification. The higher this value the more confidence is given to the classification. AdaBoost is intended to produce a linear threshold of all hypotheses:

$$\text{sign}\left(\sum_{j=1}^{N} \alpha^j g^j(\mathbf{x})\right). \tag{2}$$

Observe that Adaboost can operate with any learning algorithm that generates a confidence rated classifier on a given weighted data set. There are different algorithms for assigning the number of votes to a weak hypothesis, and for adjusting the weights of the examples. For example, in confidence-rated Adaboost [15] the number of votes of the weak hypothesis g^h is given by the value α^h that minimizes the following function:

$$Z(\alpha) = \sum_{i=1}^{m} w_i \exp(-\alpha y_i g^h(\mathbf{x}_i)) \tag{3}$$

and the weights w_i of the examples are updated according to

$$w_i \leftarrow w_i \exp(-\alpha^h y_i g^h(\mathbf{x}_i))/v \tag{4}$$

where v is a normalization factor such that $\sum w_i = 1$. There are analytical approximations and heuristics that may replace this formula in specific problems.

B. Boosting fuzzy rules in binary problems

For the sake of simplicity, we restrict the discussion for the time being to two-classes problems $y_i \in \{-1, 1\}$. The space of weak hypotheses is, in this case, the product space of the fuzzy partition and the class labels, $\mathcal{A} \times \{-1, 1\}$. Given the results in the preceding subsection, the fitness of a fuzzy rule is

$$\text{fitness}(\text{ if } \mathbf{x} \text{ is } A^j \text{ then } c^j) = \sum_i w_i \exp(-y_i c^j A^j(\mathbf{x}_i)) \tag{5}$$

and the number of votes of this rule is the value of α minimizing the following expression [14]:

$$Z(\alpha) = \sum_i w_i \exp(-\alpha y_i A^j(\mathbf{x}_i)). \tag{6}$$

$Z(\alpha)$ is convex except for the case in which the rule antecedent does not cover any negative examples; for avoiding this and other numerical instabilities, a term that penalizes large values of α is added

$$Z(\alpha) = \sum_i w_i \exp(-\alpha y_i A^j(\mathbf{x}_i)) + \sum_{i:A^j(\mathbf{x}_i)=0} w_i \exp(|\alpha\epsilon|) \tag{7}$$

where ϵ is determined by hand.

The weights w_i of all the instances are recalculated each time a new weak learner is added to the ensemble. The weights of correctly classified instances are lowered, and these of misclassified examples are increased, using the expression that follows:

$$w_i \leftarrow \frac{w_i \exp(-\alpha^j y_i A^j(\mathbf{x}_i))}{\sum_i w_i \exp(-\alpha^j y_i A^j(\mathbf{x}_i))} \tag{8}$$

Notice, that an instance is never completely removed unless $\alpha^j \to \infty$, which is prevented by the penalty term in eq. (7).

C. Boosting fuzzy rules in multiclass problems

We have mentioned that the space of weak hypothesis in two-classes problems is $\mathcal{A} \times \{c_1, c_2\}$. That is to say, the j-th iteration of the learning algorithm searches for a pair (A^j, c^j), which are parameters of the rule

$$\text{if } \mathbf{x} \text{ is } A^j \text{ then } c^j, \tag{9}$$

for which the fitness value is maximum. This value was defined in Eq. (5), which we rewrite here as follows:

$$\begin{aligned}\text{fitness}(\text{if } \mathbf{x} \text{ is } A^j \text{ then } c^j) = \\ \sum_{i:y_i=c^j} w_i \exp(A^j(\mathbf{x}_i)) \\ +\sum_{i:y_i\neq c^j} w_i \exp(-A^j(\mathbf{x}_i)).\end{aligned} \tag{10}$$

Once the pair (A^j, c^j) maximizing this fitness function is found, the rule is assigned a number of votes s^j (see Eq. (6)), thus its final linguistic writing is:

$$\text{if } \mathbf{x} \text{ is } A^j \text{ then truth}(c^j) = s^j. \tag{11}$$

For extending this scheme to multiclass problems [3] the initial problem is transformed into p binary problems, where the instances of the k-th binary problem are re-labelled:

$$y_i^{(p)} = \begin{cases} 1 & y_i = p \\ -1 & y_i \neq p. \end{cases} \tag{12}$$

The solution of the k-th problem comprises fuzzy rules with this form:

$$\begin{aligned}&\text{if } x_1 \text{ is } A_1^j \text{ and } \ldots \text{ and } x_n \text{ is } A_n^j \\ &\text{then tr(class} = c_k) = s_1^{jk} \text{ and tr(class} \neq c_k) = s_{-1}^{jk}.\end{aligned} \tag{13}$$

These rules can be replaced by their equivalent p-classes consequents, by substituting the "class $\neq c_k$" label with the corresponding set of class labels of the multi-class problem, i.e.

$$\begin{aligned}&\text{if } x_1 \text{ is } A_1^j \text{ and } \ldots \text{ and } x_n \text{ is } A_n^j \text{ then} \\ &\text{tr}(c_1, \ldots, c_k, \ldots, c_p) = (s_{-1}^{jk}, \ldots, s_{-1}^{jk}, s_1^{jk}, s_{-1}^{jk}, \ldots, s_{-1}^{jk}).\end{aligned} \tag{14}$$

III. An Extension of the Adaboost Algorithm to Learn Fuzzy Rule Based Classifiers from Low Quality Data

We will represent the epistemic uncertainty in the data by means a nested set of confidence intervals, that provide us with the same information about the unknown value as a possibility distribution for which the α-cuts of its associated fuzzy membership function are confidence intervals with level $1-\alpha$, as explained in [11]. Representing the imprecision by means of sets of possible values implies that the output of the FRBCS might not be completely determined and generally speaking, it will be a fuzzy subset of the class labels. From the foregoing it can be deduced that the fitness of a rule should also be a set, however extending Eq. 5 is not trivial. Different decisions can be taken that influence the accuracy of the method and also its computational cost.

Let us use an illustrative example for introducing the problems that arise when we regard the computation of the output of an FRBCS for imprecise data as a voting-based committee. Suppose that we are given the KB that follows:

if $x < 1.5$ then truth(class is A)= 0.4 and truth(class is B)= 0.8
if $x \in [1, 2]$ then truth(class is A)= 0.8 and truth(class is B)= 0.1
if $x > 2$ then truth(class is A)= 0.2 and truth(class is B)= 0.6

and the input is

$x \in [1.2, 1.8]$.

If we determine first the set of compatibilities of each rule with the imprecise example and then add the sets of votes, the situation is as follows:

Rule	Votes for class 'A'	Votes for class 'B'
Rule # 1:	{0,0.4}	{0,0.8}
Rule # 2:	0.8	0.1
Rule # 3:	0	0
Total	{0.8, 1.2}	{0.1, 0.9}

In words, the output of the classifier is the set of classes {A,B}, because we cannot state that any element of {0.8, 1.2} is higher than any element of {0.1, 0.9} neither the opposite. However, this course of reasoning will not always produce the most specific answer. Observe that, if the actual value of x is between 1.2 and 1.5, rules 1 and 2 are true, thus the number of votes of classes 'A' and 'B' are 1.2 and 0.9. If the value of x is between 1.5 and 1.8, the number of votes are 0.8 and 0.1. In either case, the object should have been assigned the class 'A'.

With this example we have shown that regarding a KB as an ensemble is not immediate when the data are imprecise: the most specific output of the classifier, when the input is a crisp set, is the set of classes

$$\text{class}(\overline{X}) = \{\arg\max_{k=1\ldots p} \sum_{j=1}^{N} A^j(x) s_k^j \mid x \in \overline{X}\}. \tag{15}$$

However, if we regard the KB as an ensemble of weak classifiers, where each one of them is assigned a set-valued, number of votes, the output is the set

$$\text{bclass}(\overline{X}) = \operatorname*{notdom}_{k=1\ldots p} \left\{ \bigoplus_{j=1}^{N} s_k^j \odot \{A^j(x) \mid x \in \overline{X}\} \right\} \tag{16}$$

where the operator "notdom" means

$$\operatorname*{notdom}_{k=1\ldots p}\{V_k\} = \{q \mid V_q \preceq V_r, r = 1, \ldots, p\} \tag{17}$$

and the precedence between set-valued votes is

$$A \prec B \iff a < b \text{ for all } a \in A, b \in B \tag{18}$$

$$A \parallel B = \neg((A \prec B) \vee (B \prec A)) \tag{19}$$

$$A \preceq B = (A \prec B) \vee (A \parallel B). \tag{20}$$

Observe that

$$\text{class}(\overline{X}) \subseteq \text{bclass}(\overline{X}) \tag{21}$$

but the equality does not hold, in general.

If the data is fuzzy, the situation is similar. The most specific output of the classifier is the fuzzy set whose membership function is as follows:

$$\begin{aligned}\text{class}(\widetilde{X})(t) = \quad & \max\{\alpha \mid \\ & t = \arg\max_{k=1\ldots p} \textstyle\sum_{j=1}^{N} A^j(x) s_k^j \\ & \text{and } \widetilde{X}(x) \geq \alpha\}.\end{aligned} \tag{22}$$

In case we regard the KB as an ensemble of weak classifiers, where each one of them is assigned a fuzzy number of votes, the output is in turn the fuzzy set

$$\text{bclass}(\widetilde{X})(t) = \max\{\alpha \mid t \in \text{bclass}([\widetilde{X}]_\alpha)\}. \tag{23}$$

Summarizing, when the data is imprecise, the most voted option of the ensemble is not known but our information about it is a normal fuzzy set. If the KB is furthermore regarded as an ensemble, thus the number of votes of each rule are independently computed, the fuzzy set describing the result of the classification is a superset of the result given by Eq. (22). This loss of specificity in the view of a KB as an ensemble has to be taken into account in the definition of the fitness function.

A. Fitness of a rule with interval or fuzzy data

We propose to generalize the fitness function in Eq. (10) to interval-valued data with the set-valued function that follows, which is based on the bounds given in Eq. (16):

$$\begin{aligned}&\overline{\text{fitness}}_{\{(\overline{\mathbf{x}}_i,\overline{y}_i)\}}(\text{if } \mathbf{x} \text{ is } A^j \text{ then } c^j) = \\ &\oplus_{i:\overline{y}_i=c^j}\overline{w}_i \odot \{\exp(A^j(x)) \mid x \in \overline{\mathbf{x}}_i\} \oplus \\ &\oplus_{i:\overline{y}_i\cap c^j=\emptyset}\overline{w}_i \odot \{\exp(-A^j(x)) \mid x \in \overline{\mathbf{x}}_i\} \oplus \\ &\oplus_{i:\overline{y}_i\neq c^j,\,\overline{y}_i\cap \overline{c}^j\neq\emptyset}\overline{w}_i \odot \{\exp(\{A^j(x), -A^j(x)\}) \mid x \in \overline{\mathbf{x}}_i\}\end{aligned} \tag{24}$$

Observe that we allow the use of set valued weights $\overline{w}_i$. In turn, applying the extension principle, the same function can be extended to fuzzy data. The membership function of the fuzzy fitness of a rule is

$$\begin{aligned}&\widetilde{\text{fitness}}_{\{(\widetilde{\mathbf{x}}_i,\widetilde{y}_i)\}}(\text{if } \mathbf{x} \text{ is } A^j \text{ then } c^j)(t) = \\ &\max\{\alpha \mid t \in \overline{\text{fitness}}_{\{([\widetilde{\mathbf{x}}_i]_\alpha,[\widetilde{y}_i]_\alpha)\}}(\text{if } \mathbf{x} \text{ is } A^j \text{ then } c^j)\},\end{aligned} \tag{25}$$

where t is a real number.

B. Issues with the fitness of a rule in multiclass problems

When extending the procedure seen in Section II-C for solving multi-class problems, the intermediate binary sets of data might contain unlabeled instances. Let us explain this with an illustrative example.

Consider the imprecise dataset that follows, comprising three crisp examples with set-valued labels. It is remarked that, in this context, an instance labelled "$\{c_1, c_2\}$" means that we are sure that the true class of the object is not c_3, but this knowledge cannot be further precised:

$(x_1, y_1) = (1, \{c_1, c_2\})$
$(x_2, y_3) = (2, \{c_1, c_3\})$
$(x_3, y_3) = (3, \{c_2, c_3\})$.

Following the procedure described in Section II-C, this dataset will be decomposed in three binary problems. Let us generalize that procedure to imprecise instances by using the most specific set of labels that is compatible with the data, as follows:

Problem # 1
$(1, \{1, -1\})$
$(2, \{1, -1\})$
$(3, -1)$

Problem # 2
$(1, \{1, -1\})$
$(2, -1)$
$(3, \{1, -1\})$

Problem # 3
$(1, -1)$
$(2, \{1, -1\})$
$(3, \{1, -1\})$

Each one of these three datasets has two elements whose classes are $\{-1, 1\}$ and therefore the fitness does not depend on how they are labelled: for instance, for Problem # 1, no matter which rule between

if x=1 then t(class = 1)=$s_1^{1,1}$

and

if x=1 then t(class $\neq$ 1)=$s_{-1}^{1,1}$

is chosen, the same fitness is obtained (see Eq. 24).

In this case, a learning algorithm which is only guided by the optimization of the fitness function can produce any KBs formed by selecting one rule from each line that follows. We have grayed out the rules that do not appear in an arbitrary selection, whose merging will be studied later.

Problem # 1
if x=1 then t(class = 1)=$s_1^{1,1}$ $\approx$ if x=1 then t(class $\neq$ 1)=$s_{-1}^{1,1}$
if x=2 then t(class = 1)=$s_1^{2,1}$ $\approx$ if x=2 then t(class $\neq$ 1)=$s_{-1}^{2,1}$
if x=3 then t(class $\neq$ 1)=$s_{-1}^{3,1}$

Problem # 2
if x=1 then t(class = 2)=$s_1^{1,2}$ $\approx$ if x=1 then t(class $\neq$ 2)=$s_{-1}^{1,2}$
if x=2 then t(class $\neq$ 2)=$s_{-1}^{2,2}$
if x=3 then t(class = 2)=$s_1^{3,2}$ $\approx$ if x=3 then t(class $\neq$ 2)=$s_{-1}^{3,2}$

Problem # 3
if x=1 then t(class $\neq$ 3)=$s_{-1}^{1,3}$
if x=2 then t(class = 3)=$s_1^{2,3}$ $\approx$ if x=2 then t(class $\neq$ 3)=$s_{-1}^{2,3}$
if x=3 then t(class = 3)=$s_1^{3,3}$ $\approx$ if x=3 then t(class $\neq$ 3)=$s_{-1}^{3,3}$

The merging of the selected rules is

if x=1 then t(1,2,3)=($s_1^{1,1}$+$s_{-1}^{1,3}$,$s_1^{1,2}$+$s_{-1}^{1,3}$,0)
if x=2 then t(1,2,3)=($s_1^{2,1}$+$s_{-1}^{2,2}$,0,$s_{-1}^{2,2}$+$s_1^{2,3}$)
if x=3 then t(1,2,3)=($s_{-1}^{3,3}$,$s_{-1}^{3,1}$+$s_1^{3,2}$+$s_{-1}^{3,3}$,$s_{-1}^{3,1}$).

Observe that the result of this arbitrary selection has assigned a non-null degree confidence to the class c_1 in the third rule, which obviously is not the best possible KB for this problem. We realize that there are other selections that achieve the objective, but our point is showing that there is a chance that the committee does not contain the proper rules unless the unlabeled instances in the intermediate problems are removed, as follows:

Problem # 1
if x=3 then t(class $\neq$ 1)=$s_{-1}^{3,1}$

Problem # 2
if x=2 then t(class $\neq$ 2)=$s_{-1}^{2,2}$

Problem # 3
if x=1 then t(class $\neq$ 3)=$s_{-1}^{1,3}$

whose merging is:

if x=1 then t(1,2,3)=($s_{-1}^{1,3}$,$s_{-1}^{1,3}$,0)
if x=2 then t(1,2,3)=($s_{-1}^{2,2}$,0,$s_{-1}^{2,2}$)
if x=3 then t(1,2,3)=(0,$s_{-1}^{3,1}$,$s_{-1}^{3,1}$).

Notwithstanding, if the unlabeled instances are removed then additional problems with the decomposition in binary problems will appear. Since the pruned individual problems do not longer contain information about where the removed instances were located, it may happen that the corresponding areas of the feature space are covered at the same time by different rules that negatively interact between themselves. This problem also happens with Iterative Rule Learning (IRL) algorithms [12], where it is solved by a simplification stage that it is not a part of the boosting algorithm. Therefore, we have decided to simplify the search and not to generate the intermediate datasets when working with multiclass problems, and propose instead to define the fitness of a rule with multiple consequents as a vector of values, and to define a lexicographic ordering between them. The fitness function we propose is as follows:

$$\begin{aligned}
&\overline{\text{fitness}}_{\{(\overline{\mathbf{x}}_i,\overline{y}_i)\}}(\text{if } \mathbf{x} \text{ is } A^j \text{ then } (c_1,\ldots,c_p)) = \\
&\Big(\bigoplus_{\overline{y_i}=c_1} \overline{w}_i \odot \exp(\{A^j(x) \mid x \in \overline{\mathbf{x}}_i\}) \oplus \\
&\quad \bigoplus_{c_1 \in \overline{y}_i, c_1 \neq \overline{y}_i} \overline{w}_i \odot \\
&\qquad \exp(\{-A^j(x), A^j(x) \mid x \in \overline{\mathbf{x}}_i\}) \oplus \\
&\quad \bigoplus_{c_1 \notin \overline{y}_i} \overline{w}_i \odot \exp(\{-A^j(x) \mid x \in \overline{\mathbf{x}}_i\}), \\
&\quad \vdots \\
&\quad \bigoplus_{\overline{y_i}=c_p} \overline{w}_i \odot \exp(\{A^j(x) \mid x \in \overline{\mathbf{x}}_i\}) \oplus \\
&\quad \bigoplus_{c_p \in \overline{y}_i, c_p \neq \overline{y}_i} \overline{w}_i \odot \\
&\qquad \exp(\{-A^j(x), A^j(x) \mid x \in \overline{\mathbf{x}}_i\}) \oplus \\
&\quad \bigoplus_{c_p \notin \overline{y}_i} \overline{w}_i \odot \exp(\{-A^j(x) \mid x \in \overline{\mathbf{x}}_i\}\Big).
\end{aligned} \tag{26}$$

In turn, the lexicographic precedence between two crisp fitness vectors $A = (a_1, \ldots, a_p)$ and $B = (b_1, \ldots, b_p)$ is as follows: let us first define two permutations of the set $\{1, \ldots, p\}$, denoted $(k_1^a, \ldots, k_p^a)$ and $(k_1^b, \ldots, k_p^b)$, such that $a_{k_1^a} \geq \ldots \geq a_{k_p^a}$ and $b_{k_1^b} \geq \ldots \geq b_{k_p^b}$. Then,

$$\begin{aligned}
&A \prec B \iff \exists m \in \{1, \ldots, p\} \mid \\
&\quad (a_{k_q^a} = b_{k_q^b})\ \forall q \in \{1, \ldots, m-1\} \wedge (a_{k_m^a} > b_{k_m^b}).
\end{aligned} \tag{27}$$

The extension of both the fitness function and the lexicographic ordering to fuzzy values results from applying the extension principle and from replacing the comparisons between real numbers with a suitable fuzzy ranking. The expression of the fitness function, in the general case, is as follows:

$$\begin{aligned}
&\widetilde{\text{fitness}}_{\{(\widetilde{\mathbf{x}}_i,\widetilde{y}_i)\}}(\text{if } \mathbf{x} \text{ is } A^j \text{ then } (c_1,\ldots,c_p))(t) = \\
&\quad \max\{\alpha \mid t \in \\
&\quad \overline{\text{fitness}}_{\{([\widetilde{\mathbf{x}}_i]_\alpha,[\widetilde{y}_i]_\alpha)\}}(\text{if } \mathbf{x} \text{ is } A^j \text{ then } (c_1,\ldots,c_p))\},
\end{aligned} \tag{28}$$

where t denotes a p-dimensional vector of real values.

C. Assignment of weights to the consequent part

The extension of Eq. (7) to set-valued data and multiclass problems consists in assigning to the k-th class in the consequent of the rule a confidence equal to the value of α_k minimizing the set-valued function Z^k defined as follows:

$$\begin{aligned}
&\overline{Z}^k_{\{(\overline{\mathbf{x}}_i,\overline{y}_i)\}}(\alpha) = \\
&\quad \{ \oplus_{i:c_k=\overline{y}_i} \overline{w}_i \odot \exp(-\alpha A^j(x)) \oplus \\
&\quad\ \oplus_{i:c_k \notin \overline{y}_i} \overline{w}_i \odot \exp(\alpha A^j(x)) \oplus \\
&\quad\ \oplus_{i:c_k \neq \overline{y}_i,\, c_k \in \overline{y}_i} \overline{w}_i \odot \exp(\{-\alpha A^j(x), \alpha A^j(x)\}) \\
&\quad \mid x \in \overline{\mathbf{x}}_i\}.
\end{aligned} \tag{29}$$

A numerical stabilization term

$$\bigoplus_{i:A^j(x)=0\ \forall x \in \overline{\mathbf{x}}_i} \overline{w}_i \exp(|\alpha\epsilon|) \tag{30}$$

with a suitable value of ϵ can also be added, if needed.

The extension of Eq. (29) to fuzzy data is as follows:

$$\widetilde{Z}^k_{\{(\widetilde{\mathbf{x}}_i,\widetilde{y}_i)\}}(\gamma)(t) = \max\{\alpha \mid t \in \overline{Z}^k_{\{([\widetilde{\mathbf{x}}_i]_\alpha,[\widetilde{y}_i]_\alpha)\}}(\gamma)\}. \tag{31}$$

In both the set-valued and fuzzy cases, the values of α_k can be found with a greedy algorithm that uses a precedence operator between fuzzy sets. However, this optimization is not as efficient as the Brent search used in the crisp version of the algorithm. Since the optimization must be launched each time a fitness value is computed, in this paper we have decided to approximate the value of α_k by the center of the set

$$\overline{\alpha}_k = \log(1 - E_k) - \log(E_k) \tag{32}$$

where E_k is a normalized weighted sum of the compatibilities of the antecedent of the j-th rule with the elements of the dataset whose class does not match the k-th term in the consequent:

$$\begin{aligned}
E_k = K \odot \Big(&\bigoplus_{\{i:c_k \notin \overline{y}_i\}} \overline{w}_i \odot \{A^j(x) \mid x \in \overline{\mathbf{x}}_i\} \oplus \\
&\bigoplus_{\{i:c_k \neq \overline{y}_i \wedge c_k \in \overline{y}_i\}} \overline{w}_i \odot \{\{0, A^j(x)\} \mid x \in \overline{\mathbf{x}}_i\}\Big)
\end{aligned} \tag{33}$$

The normalization factor K is the inverse of the upper bound of the normalized weighted sum of the compatibilities of the antecedent of the j-th rule with all the elements of the dataset:

$$K = (\sum_i \max\{\overline{w}_i\} \cdot \max\{A^j(x) \mid x \in \overline{\mathbf{x}}_i\})^{-1}. \tag{34}$$

D. Assignment of weights to the examples

After the j-th rule is added, the weights of the instances are recomputed as follows:

$$\overline{w}'_i = \bigcup_{x \in \overline{\mathbf{x}}_i} w'_i(x) \tag{35}$$

where $w'_i(x) =$

$$K' \odot \overline{w}_i \odot \begin{cases} c_k = \overline{y}_i & \exp(-\alpha_k A^j(x)) \\ c_k \notin \overline{y}_i & \exp(\alpha_k A^j(x)) \\ \text{else} & \exp(\alpha_k A^j(x) \odot \{-1, 1\}) \end{cases} \tag{36}$$

and K' is a crisp normalization factor such that $\max \oplus_i \overline{w}'_i = 1$.

E. Some details of the genetic algorithm

Adaboost depends on a procedure that fits a weak learner to the weighted training set. If the vector-valued fitness we have proposed in this paper is to be used, learning a weak classifier reduces to finding the antecedent A^j that optimizes the fitness function explained in section III-B, with respect to the lexicographic ordering defined in the same section. Since all possible values of the fitness can be compared between themselves, implementing this criterion amounts to redefining the meaning of the operator "less than" ($<$) in an ordinary GA; there is no need to use multicriteria techniques [16], thus we have used instead a standard generational scheme with a tournament-based selection.

Details of this algorithm can be found in [3]. Let us recall for the convenience of the reader the only part in this GA besides the fitness function (whose explanation has occupied most of this paper) which departs from a standard implementation: the coding of the fuzzy memberships.

Descriptive fuzzy rules are coded by an integer, which is the index j of the antecedent A^j in $\mathcal{A}$. This integer is encoded in turn, by a sequence of n numbers that refer to labels of the linguistic terms in the underlying fuzzy partitions. In addition to linguistic labels describing subsets of the domain of each variable, such as "LOW" and "HIGH", each linguistic variable includes a wild card label "ANY VALUE", with a membership degree of 1 across the entire universe of discourse.

The linguistic expression of a fuzzy rule that contains a wild card can be simplified, as illustrated by the following example: assume a classification problem with two features (weight,height), where height={low,high} and weight={light, heavy}. The two linguistic variables are extended with a wild card term such that height={low, high, anyvalue} and weight={light, heavy, anyvalue}. If the rule antecedent is described by "anyvalue×heavy", the linguistic expression

if height is anyvalue and weight is heavy then t(1,2)=(0.4,0.8)

is simplified to

if weight is heavy then t(1,2)=(0.4,0.8).

This property of the genetic representation allows it to code general rules that only refer to a subset of all possible features, thus enabling the boosting algorithm to take advantage of feature selection.

The last rule is coded by the sequence (3,2), as "anyvalue" is the third linguistic term in the first linguistic variable, and "heavy" is the second term in the second linguistic variable. Observe that the class labels in the rule consequent are not part of the genetic codification, neither the confidences degrees in the consequents are.

IV. Numerical results

In this section we include the first results of the implementation of Adaboost for learning fuzzy rules from low quality data, applied to the imprecise datasets "Diagnosis of the Dyslexic" [17] and "Athletics at Oviedo University" [18]. We include first a brief description of these datasets and then we discuss the compared results of the application of the new method to these problems.

A. Description of the datasets and experimental settings

The datasets "Diagnosis of the Dyslexic" and "Athletics at the Oviedo University", have been introduced in [19] and [20], respectively, and are available in the data set repository of keel-dataset (http://www.keel.es/datasets.php) [21], [22]. The names, the number of examples (Ex.), number of attributes (Atts.), the classes (Classes) and the fraction of patterns of each class (%Classes) for each dataset are displayed in Table I. Observe that the proportions of the different patterns are intervals, because the class labels of some instances are imprecise.

TABLE I
SUMMARY DESCRIPTIONS OF DATASETS WITH META-INFORMATION.

Dataset	Ex.	Atts.	Classes	%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-12-12	65	12	3	([0.32,0.43],[0.32,0.52], [0.12,0.30])
Dyslexic-11-01	65	11	3	([0.43,0.53],[0.23,0.35], [0.12,0.30])
Dyslexic-11-12	65	11	3	([0.32,0.43],[0.32,0.52], [0.12,0.30])

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 150 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. Each partition of test contains 1000 tests.

B. Differences in accuracy

The compared accuracies between the proposed algorithm (Boosting_LQD) and other GFS capable of extracting rules from low quality data (GFS [23] and MR_GFS [24]) data are shown in Table II. The results are expressed by means of intervals, that are our best bounds about the mean values of the test error. These intervals are defined by the expression

$$\overline{\text{error}} = \left\{ \frac{1}{m} \sum_{i=1}^{m} e_i \mid e_i \in \overline{e}_i \right\} \tag{37}$$

where

$$\overline{e}_i = \begin{cases} 0 & \text{bclass}(\overline{x}_i) = \overline{y}_i \text{ and } \#(\overline{y}_i) = 1, \\ 1 & \text{bclass}(\overline{x}_i) \cap \overline{y}_i = \emptyset, \\ \{0,1\} & \text{else.} \end{cases} \tag{38}$$

The method "Boosting_LQD" shows a better performance than the remaining classifiers, in both binary problems (Athletics datasets) and multiclass (Dyslexic datasets). Observe that MR_GFS could not be applied to the multiclass datasets in this paper because a suitable matrix of costs was not given for

TABLE II
BEHAVIOUR OF "GFS", "MR_GFS" AND "BOOSTING_LQD" IN SEVERAL DATASETS OF ATHLETICS AND DYSLEXIA.

Dataset	GFS	MR_GFS	Boosting_LQD
100mlI	[0.176,0.378]	[0.178,0.380]	[0.170,0.376]
100mlP	[0.176,0.360]	[0.188,0.367]	[0.180,0.358]
Long	[0.321,0.590]	[0.288,0.557]	[0.231,0.499]
BLong	[0.326,0.625]	[0.286,0.586]	[0.219,0.519]
B100mlI	[0.172,0.369]	[0.188,0.385]	[0.158,0.356]
B100mlP	[0.160,0.349]	[0.161,0.350]	[0.161,0.350]
B200mlI	[0.232,0.473]	[0.178,0.418]	[0.171,0.415]
B200mlP	[0.262,0.480]	[0.215,0.433]	[0.188,0.406]
Athletics mean	[0.228,0.453]	[0.210,0.434]	[0.184,0.409]
Dyslexic-12-12	[0.386,0.557]	N/A	[0.376,0.530]
Dyslexic-11-01	[0.445,0.573]	N/A	[0.447,0.567]
Dyslexic-11-12	[0.528,0.690]	N/A	[0.458,0.595]
Dyslexic mean	[0.453,0.606]	N/A	[0.427,0.564]
Global mean	[0.340,0.529]	N/A	[0.305,0.486]

these datasets. The good behavior of the algorithm in the binary problems "Long", "BLong", "B200mlI" and "B200mlP" was remarkable, as shown in the graphs of the dispersion of the results of the 100 bootstrap tests, that have been plotted in Figures 1 to 3. In these Figures, the horizontal coordinate measures the fraction of errors of the boosting algorithm, and the vertical coordinate measures the same parameter in the counterpart algorithm. Each filled circle represents a case where the boosting algorithm was the best choice, and the squares mean the opposite result. A blank circle or square means a tie (or a difference with is not significant) between both algorithms. Those figures where there is a high density of filled circles in the upper left part signal cases where the dispersion of the results is compatible with a statistically significant improvement of the bound, and those cases where the cloud is near the diagonal or in the lower, right part display situations where the differences, if exist, are not significant.

In multiclass problems there is a substantial improvement in the datasets "Dyslexic-11-12" (Figure 4) and "Dyslexic-12-12" and a slight improvement in "Dyslexic-11-01" (Figure 5). By comparing the results obtained by the boosting algorithm in "Dyslexic-12-12" and "Dyslexic-11-12" we detect that the removal of the imprecision in the input values (dataset "Dyslexic-11-12") indeed lowers the performance of the algorithm, supporting a result already found in [17]. Moreover, the boosting algorithm show us that the imprecise outputs of that dataset are significant for determining the weights of the instances and the confidences in the consequents.

V. CONCLUDING REMARKS

In this paper we have proposed a new extension of the Adaboost algorithm for learning fuzzy rule based classifiers from interval-valued or fuzzy data. Fuzzy rules were regarded as weak learners and knowledge bases as ensembles. The number of votes of a weak learner was identified with the degree of confidence of its corresponding consequent. Both the objective function of the Adaboost algorithm and the weights of the instances were assigned interval or fuzzy values, however the number of votes each weak learning is assigned

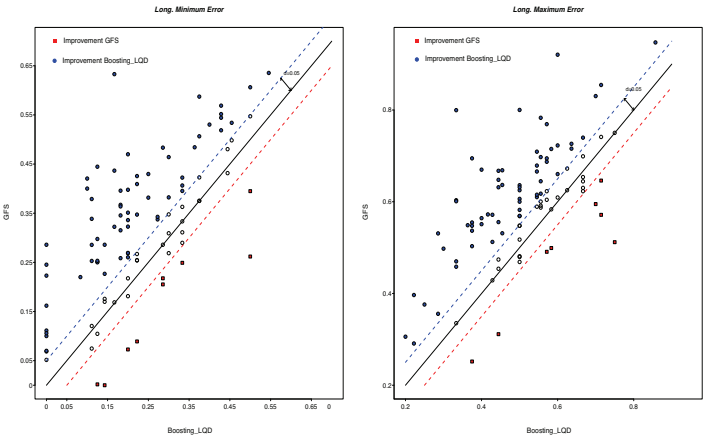

Fig. 1. Behaviour of "GFS" and "Boosting_LQD" respect to the dataset Long. **Left figure:** Lower bounds. **Right figure:** Upper bounds.

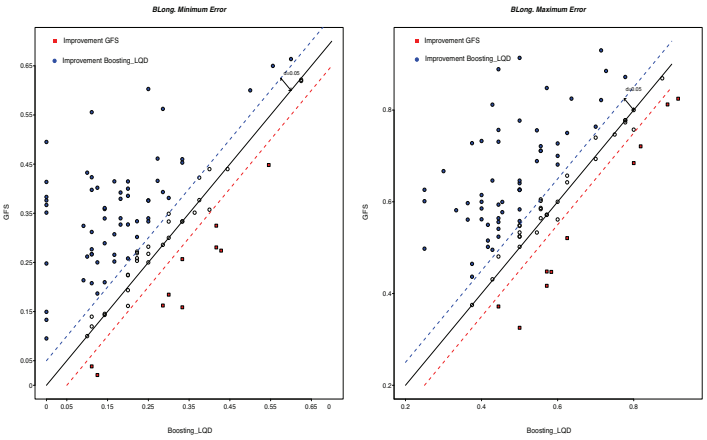

Fig. 2. Behaviour of "GFS" and "Boosting_LQD" respect to the dataset BLong. **Left figure:** Lower bounds. **Right figure:** Upper bounds.

has been designed to be a crisp number thus the classifier does not introduce uncertainty of its own.

The results of the new algorithm have been compared to that of previous genetic algorithms for low quality data. The results prove that this technique is fast, its accuracy is competitive and the number of rules in the knowledge base is not higher than that of the alternatives. On the other hand, it uses a voting-based inference, whose linguistic quality is not the best, and the performance gain is not highly relevant for multiclass problems.

ACKNOWLEDGEMENTS

This study has been supported by the Spanish Ministry of Science and Technology and by European Fund FEDER (project TIN2008-06681-C06-04) and by the Principado de Asturias, PCTI 2006-2009.

REFERENCES

[1] Freund, Y., Schapire, R. Experiments with a new boosting algorithm. In Machine Learning, Proc. 13th International Conference, 148-156 (1996)
[2] Kuncheva, L. I. Fuzzy Classifier Design. Springer-Verlag, NY, (2000).

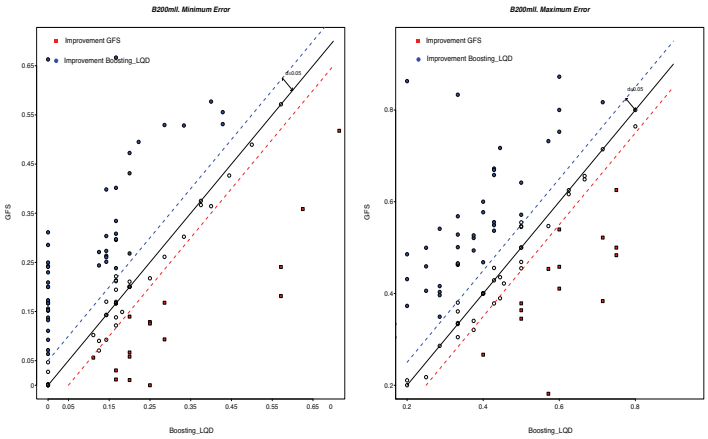

Fig. 3. Behaviour of "GFS" and "Boosting_LQD" respect to the dataset B200mII. **Left figure:** Lower bounds. **Right figure:** Upper bounds.

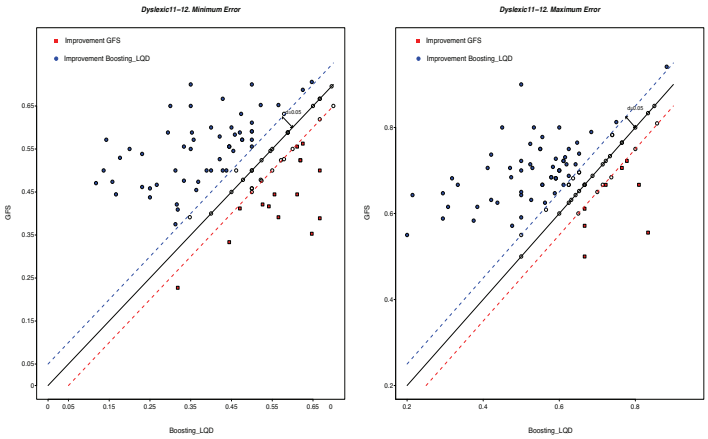

Fig. 4. Behaviour of "GFS" and "Boosting_LQD" respect to the dataset Dyslexic11-12. **Left figure:** Lower bounds. **Right figure:** Upper bounds.

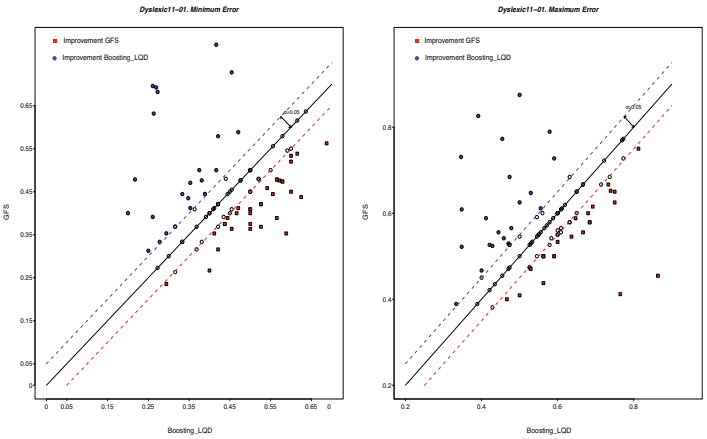

Fig. 5. Behaviour of "GFS" and "Boosting_LQD" respect to the dataset Dyslexic11-01. **Left figure:** Lower bounds. **Right figure:** Upper bounds.

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

[4] Hoffmann, F., Boosting a Genetic Fuzzy Classifier. in Proc. Joint 9th IFSA World Congress and 20th NAFIPS International Conference, vol. 3, (Vancouver, Canada), pp. 1564-1569 (2001)

[5] Junco, L., Sánchez, L. Using the Adaboost algorithm to induce fuzzy rules in classification problems, Proc. ESTYLF 2000, Sevilla, 297-301. (2000)

[6] González, A., Herrera, F., Multi-stage genetic fuzzy systems based on the iterative rule learning approach. Mathware and Soft Computing 4(3), 233-249 (1997).

[7] Friedman, J., Hastie, T., Tibshirani, R. Additive Logistic Regression: a Statistical View of Boosting, Annals of Statistics 28(2), 337-374. (2000).

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

[9] Sánchez, L. Otero. J., Boosting fuzzy rules in classification problems under single-winner inference. International Journal of Intelligent Systems 22(9) 1021-1035. (2007)

[10] Dubois, D., Guyonnet, D. Risk-informed decision-making in the presence of epistemic uncertainty. International Journal of General Systems 40 (2) 145-167 (2011)

[11] Sánchez, L., Couso, I., Advocating the use of Imprecisely Observed Data in Genetic Fuzzy Systems. IEEE Transactions on Fuzzy Systems 15 (4), 551-562 (2007)

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

[13] Ishibuchi, H., Nakashima, T. and Morisawa, T., Voting in fuzzy rule-based systems for pattern classification problems. Fuzzy Sets and Systems, vol 103, no. 2, 223-239, (1999).

[14] Schapire, R., Singer, Y. Improved Boosting Algorithms Using Confidence-rated Predictions. Machine Learning 37(3): 297-336. (1999)

[15] Schapire, R. E. Theoretical views of Boosting and Applications. Lecture Notes in Artificial Intelligence, Vol. 1720, pp 13 - 25. (1999).

[16] Sánchez, L., Couso,I., Casillas, J. Modeling vague data with genetic fuzzy systems under a combination of crisp and imprecise criteria Proceedings of the First IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making (MCDM 2007), 30-37 Honolulu, Hawaii, USA (2007)

[17] Palacios, A., Sánchez, L., Couso, I. Diagnosis of dyslexia with low quality data with genetic fuzzy systems. International Journal on Approximate Reasoning 51, 993-1009 (2010)

[18] Palacios, A. Sánchez, L., Couso, I. Future performance modelling in athletism with low quality data-based GFSs. Journal of Multivalued Logic and Soft Computing 17 (2-3) 207-228 (2011)

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

[20] Palacios, A. Sánchez, L. Couso, I. Future performance modeling in athletism with low quality data-based GFSs. Journal of Multiple-Valued Logic and Soft Computing, 17. 207-228. (2011)

[21] Alcalá, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., Herrera, F. KEEL Data-Mining software tool: data set repository, integration of algorithms and experimental analysis framework. Journal of Multivalued Logic and Soft Computing 17 (2-3) 255-287 (2011)

[22] Alcalá-Fdez, J., Sánchez, L., García, S., del Jesus, M.j., Ventura, S., Garrell, J. M., Otero, J., Romero, C., Bacardit, J., Rivas, V. M., Herrera, F. KEEL: A software tool to assess evolutionary algorithms to data mining problems Soft Computing 13 (3) 307-318 (2009)

[23] Palacios, A., Sánchez, L., Couso, I. Extending a simple cooperative-competitive learning fuzzy classifier to low quality datasets. Evolutionary Intelligence 2 (1-2), 73-84 (2010)

[24] Palacios, A., Sánchez, L., Couso, I. Linguistic Cost-Sensitive Learning of Genetic Fuzzy Classifiers for Imprecise Data. International Journal on Approximate Reasoning. In Press. (2011)