Preprocessing Vague Imbalanced Datasets And Its Use In Genetic Fuzzy Classifiers

Ana M. Palacios, Luciano Sánchez and Inés Couso

Abstract— **When there is a substantial difference between the number of cases of the majority and minority classes, minimum error-based classification systems tend to overlook these last instances. This can be corrected either by preprocessing the dataset or by altering the objective function of the classifier. In this paper we analyze the first approach, in the context of genetic fuzzy systems (GFS), and in particular of those that can operate with imprecisely observed and low quality data. We will analyze the different preprocessing mechanisms of imbalanced datasets and will show the necessity of extending these for solving those problems where the data is both imprecise and imbalanced. In addition, we include a comprehensive description of a new algorithm, able to preprocess imprecise imbalanced datasets. Several real-world datasets are used to evaluate the proposal.**

I. Introduction

Most classifiers designed for minimizing the global error rate perform poorly in imbalanced datasets [17], [31]. This is because the minimum error Bayes rule does not equalize the misclassification rate for the different classes unless their sizes are similar. Conversely, the method of choice for these last problems is based on the minimum risk Bayes rule. Alternatively, we can preprocess the dataset and lower the differences between the majority and minority classes.

Genetic Fuzzy Systems (GFSs) are not an exception to this behavior. For using a GFS with an imbalanced dataset, either we can alter the fitness function by including a cost matrix [27] or we can preprocess the data. Both techniques have already been studied in the context of GFSs: there are works that deal with the use of fuzzy classifiers for the imbalanced dataset problem [10], [34], [36], [37], [41], and others that employ a preprocessing step in order to balance the training data before the training, which has been shown to solve the problem [1], [12], [13], [14], [15]. In particular, there is an study in [12] about the combination of imbalanced classes in the framework of FRBS and the application of a re-sampling procedure named "Synthetic Minority Oversampling Technique" or SMOTE [2].

Notwithstanding, the learning of Fuzzy Rule-based Systems (FRBSs) from datasets that are both imprecisely perceived and imbalanced has not yet been addressed from the perspective of the preprocessing of the training data. Therefore, we are chiefly interested in mechanisms for preprocessing these low quality imbalanced dataset and the effect caused in the GFS once the data is balanced. It is remarked that the GFSs that we will evaluate are based on our prior work, where we extended the Genetic Fuzzy Classifiers to the use of low quality data [26], [28], [29]. We want also to make clear that there are certain difficulties in this extension. A crisp dataset is imbalanced when the sizes of its classes are much different, and we can estimate these sizes from the data in the training set. On the contrary, a fuzzy dataset may have imprecision in the perception of the classes of the objects, thus these percentages are not completely known. Furthermore, a low specificity in the output variable easily leads to an imbalanced dataset, as we will show in the sections that follow.

We have chosen to base our preprocessing stage on the aforementioned SMOTE algorithm. In turn, for extending SMOTE to fuzzy data we have taken into account the fuzzy arithmetic operators reviewed in [4] and [8], and different rankings of fuzzy numbers. Ranking methods play a crucial role in this work; beginning with [18], [19], where the concept was first introduced for ordering fuzzy numbers, ranking or comparing fuzzy numbers has now many different interpretations; in this paper we will focus on the centroid index ranking method [6], [7], [11], [23], [33], [38] which arguably is a commonly used technique for ranking numbers [32]. Finally, after describing a new algorithm for balancing low quality datasets, we will analyze the behaviour of the GFS proposed in [26], preprocessing low quality imbalanced datasets before the learning phase, and compare the results obtained in several real-world problems about the diagnosis of dyslexic children [29] and the future performance of athletes in a competition [26].

The structure of this paper is as follows: in Section II we introduce the problem of imbalanced datasets and some preprocesing techniques for imbalanced datasets, focusing the SMOTE algorithm [2]. In Section III we present the new algorithm for balancing low quality imbalanced datasets, taking into account the possibly imprecise outputs. In Section IV we show the results obtained with those GFSs able to use low quality data, after applying the algorithm proposed here. We will also compare these results with the results obtained by the same GFSs in the original low quality dataset. The paper finishes with the conclusions and future works, in Section V.

II. Imbalanced datasets in classification

In this section we introduce the imbalanced dataset problem and we will show some preprocessing methods that are

Ana M. Palacios is with the Departamento de Informática, Universidad de Oviedo, Gijón, Asturias, Spain; email: palaciosana@uniovi.es.

Luciano Sánchez is with the Departamento de Informática, Universidad de Oviedo, Gijón, Asturias, Spain; email: luciano@uniovi.es

Inés Couso is with the Departamento de Estadística e I.O. y D.M, Universidad de Oviedo, Gijón, Asturias Spain; email: couso@uniovi.es

commonly applied in imbalanced datasets, highlighting the SMOTE algorithm.

A. The problem of imbalanced datasets

The problem of imbalanced datasets in classification occurs when the number of instances of one class is much lower than that of the other classes. Specifically, when the dataset has only two classes, this happens when one class is represented by a high number of examples, while the other is represented by only a few [3]. Some authors have named this problem "datasets with rare classes" [39].

TABLE I
CONFUSION MATRIX FOR A PROBLEM OF TWO CLASSES

	Positive Prediction	Negative Prediction
Positive class	True Positive (TP)	False Negative (FN)
Negative class	False Positive (FP)	True Negative (TN)

Usually the minority class represents the concept of interest, especially in medical applications [20], [24], [30]. We will study one these problems later, that of diagnosing dyslexia in children. The so called "others" class represents the counterpart of the concept, for example children without dyslexia. The evaluation of the performance of classifier, traditionally, is based on the confusion matrix (see Table I for a typical confusion matrix for a problem of two classes). From this table the average classification error is defined as the total number of misclassified examples divided by the total number of available examples (1). The accuracy is given by eq. (2):

$$\text{Err} = \frac{FP + FN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} = 1 - \text{Err} \tag{2}$$

It is clear that minimizing the global error has a bias towards the majority class. In other words, the instances that belong to the minority class are misclassified more often than the other classes.

B. Preprocessing imbalanced datasets

To deal with the imbalanced dataset problem we can alter the objective function of the classifier, making it to depend on a cost matrix. We can also leave the classification system as is and process the data in order to diminish the effect caused by their class imbalance. It has been proved that applying a preprocessing method to balance the classes is a satisfactory solution to the problem of imbalanced datasets, see for instance [1]. In [1], [12] different methods of preprocessing where studied. These methods are classified in three categories:

- Under-sampling methods: Obtain a subset of the original dataset by eliminating some of the examples of the majority class. This category comprises the Condensed Nearest Neighbour rule (CNN) [16], Tomek links [35], One-sided selection (OSS) [21], Neighbourhood cleaning rule (NCL) [22], Wilson's Edited Nearest Neighbour (ENN) [40] and the random under-sampling.
- Over-sampling methods: Obtain a superset of the original dataset by replicating some of the examples of the minority class or creating new ones from the original minority class instances. These methods are Synthetic minority over-sampling technique (SMOTE) [2] and random over-sampling.
- Hybrid methods: These combine over-sampling and under-sampling, and obtain a set by combining the two previous methods. For instance, SMOTE+Tomek Link and SMOTE+ENN.

In [12] these preprocesing methods were compared in the context of FRBCSs, showing the good behaviour for the over-sampling methods, and in particular SMOTE.

C. SMOTE algorithm

In the SMOTE algorithm, the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any or all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen [2]. For example, if the implementation uses four nearest neighbors ($k = 4$) and the amount of over-sampling needed is 200%, only two neighbors from the four nearest neighbors are chosen and one sample is generated in the direction of each. In Figure 1 an example is shown where x_i is the selected point, x_{i1} to x_{i4} are some of the selected nearest neighbors and r_1 to r_2 are the synthetic data points created by the randomized interpolation.

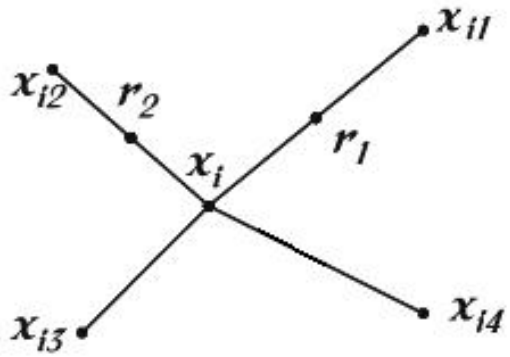

Fig. 1. Creation of synthetic data points in the SMOTE algorithm.

Synthetic samples are generated in the following way: Take the difference between the feature vector (sample) under consideration and its nearest neighbour. Multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration. This causes the selection of a random point along the line segment between

two specific features. This approach effectively forces the decision region of the minority class to become more general [2]. An example is detailed in Table II.

TABLE II
EXAMPLE OF THE SMOTE METHOD.

Consider a sample (6,4) and let (4,3) be its nearest neighbor.
(6,4) is the sample for which k-nearest neighbors are being identified.
(4,3) is one of its k-nearest neighbors.
Let:
f1_1 = 6 f2_1 = 4 f2_1 - f1_1 = -2
f1_2 = 4 f2_2 = 3 f2_2 - f1_2 = -1
The new samples will be generated as
(f1',f2') = (6,4) + rand(0-1) * (-2,-1)
rand(0-1) generates a random number between 0 and 1.

III. PREPROCESSING OF LOW QUALITY IMBALANCED DATASETS

As we have explained in Section II-A, the problem of imbalanced datasets in classification occurs when the number of instances of one class is much lower than that of the other classes. This also happens when the dataset contains low quality data, interval-valued or fuzzy numbers, in the output variables. For instance, if one instance is labeled as "class {A, B}" or, in words, if we do not know whether the true class of the instance is A or B, then the percentage of instances that belong to each class is also an imprecise value. If the specificity of these imprecise values is low, the dataset is possibly imbalanced. For instance, imagine a problem with three classes where, after computing the ranges of the relative frequencies of the classes, we obtain that $f_1 \in [0.05, 0.25]$, $f_2 \in [0.05, 0.35]$ and $f_3 \in [0.4, 0.9]$. This means that the actual frequencies might be 0.25, 0.35 and 0.4, which is not, strictly speaking, an unbalanced problem, but it is also possible that they are 0.05, 0.05 and 0.9. In this case, a classification system will not perform well on classes A and B unless we preprocess the dataset.

To preprocess these kind of datasets we propose a new algorithm based on SMOTE (explained in Section II-C). There are three aspects in our generalization that deserve a detailed study:

1) Selection of the minority class and the amount of synthetic examples.
2) Computation of the k nearest neighbours of any example. The implementation applied in this work uses the euclidean distance to select the k nearest neighbors and it uses fuzzy arithmetic operators and a fuzzy ranking, as we will explain later.
3) Generation of synthetic examples from the minority class. We will use fuzzy arithmetic operators, and control the values that may be out of range for the different attributes.

A. Selection of the minority class

The inputs to the SMOTE algorithm in [2] are the number of minority class samples (T) and the amount of synthetic examples (N). In our generalization, the fraction of examples in each class is defined by an imprecise value and the algorithm has to determine the amount of synthetic examples for each class. In Figure 3, lines 1 to 13, we detail how this last value (N) is determined for each class. All classes but the majority will be assigned synthetic examples and the examples with an imprecise output will have less relevance in the classification.

B. Computation of the k nearest neighbors

In a first step we collect all the examples that possibly belong to the minority class (that includes those whose class we know and those whose class we cannot affirm is different than the minority). This is outlined in Figure 3, lines 15 to 20.

The second step consists in obtaining the k nearest neighbors of the example, where the meaning of "nearest" is given by a generalized euclidean distance (lines 21 to 25) and a certain method for ranking these distances. That is to say, the euclidean distance between two vectors of fuzzy numbers $(\widetilde{A}_{i1}, \ldots, \widetilde{A}_{in})$ and $(\widetilde{B}_{j1}, \ldots, \widetilde{B}_{jn})$ is generalized as follows:

$$\widetilde{D}_{ij} = \left[\bigoplus_{m=1}^{n} (\tilde{A}_{im} \ominus \tilde{B}_{jm})^2 \right]^{\frac{1}{2}} \tag{3}$$

where all fuzzy numbers are trapezoidal, $\tilde{A} = (a, b, c, d)$ (see Figure 2) and all the arithmetic operators are also fuzzy (see references [4], [8]). We will consider that $\widetilde{D}_{ij}$ is a generalized trapezoidal fuzzy number.

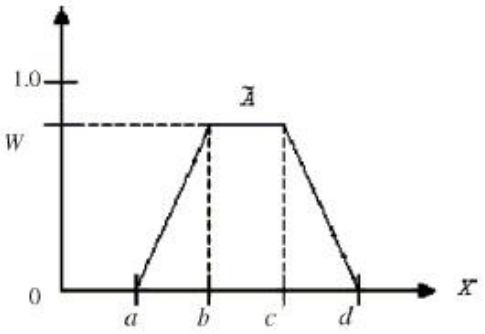

Fig. 2. A trapezoidal fuzzy number.

In line 26 of Figure 3 we have used the operation "ranking" for determining the k nearest neighbours of a given example. It is well known (see [32]) that no single ranking method is superior to all other methods; each ranking appears to have some advantages as well as disadvantages. In this proposal we use a method that was defined in [38] and improved in [11]. Given two fuzzy numbers A and B, this method is based on four crisp values $\overline{x}(A)$, $\overline{y}(A)$, $\overline{x}(B)$, $\overline{y}(B)$. $\overline{x}(A)$ indicates the representative location of fuzzy number A, and $\overline{y}(A)$ presents the average height of the fuzzy number. For a generalized trapezoidal fuzzy number $A = (a, b, c, d)$, these values are defined as follows [5], [25]:

$$\overline{x}(A) = \frac{\int_a^b (x f_A^L) dx + \int_b^c x dx + \int_c^d (x f_A^R) dx}{\int_b^a (f_A^L) dx + \int_b^c dx + \int_c^d (f_A^R) dx} \tag{4}$$

$$\overline{y}(A) = \frac{\int_0^w (y g_A^L) dy + \int_0^w (y g_A^R) dy}{\int_0^w (g_A^L) dy + \int_0^w (g_A^R) dy} \quad (5)$$

where f_A^L and f_A^R are the left and right membership functions of fuzzy number A, respectively. g_A^L and g_A^R are the inverse functions of f_A^L and f_A^R, respectively. This method, assumes that the importance of the degree of representative location is higher than average height. Lastly, observe for any two fuzzy numbers A and B, we have three different situations, whose associated orderings are [38]:

1) If $\overline{x}(A) > \overline{x}(B)$, then A $>$ B.
2) If $\overline{x}(A) < \overline{x}(B)$, then A $<$ B.
3) If $\overline{x}(A) = \overline{x}(B)$, then
   - If $\overline{y}(A) > \overline{y}(B)$, then A $>$ B.
   - If $\overline{y}(A) < \overline{y}(B)$, then A $<$ B.
   - If $\overline{y}(A) = \overline{y}(B)$, then A $=$ B.

C. Generation of the synthetic examples

The generation of the synthetic examples, as in [2], consists in taking the difference between the feature vector (sample) under consideration and its nearest neighbor. This difference is multiplied by a random number between 0 and 1, and added to the feature of the synthetic example. These operations involve fuzzy arithmetic, as described in Figure 3, lines 31 to 34. We control the values that are out of range in the different attributes in line 34.

IV. Numerical Results

Imbalanced datasets appear often in practice, and they are particularly relevant for medical applications, as mentioned in the introduction (see also references [20], [24], [30]). In this section we will study several real-world problems. Some of them are related to medical diagnosis (diagnosing dyslexia in children [29]), and the future performance of athletes in certain tests is studied in the others [26]. These datasets are summarized as follows:

- "**Athlete**" datasets: This set comprises 8 datasets that are used to predict whether an athlete will improve certain threshold in the long jump, 100 meters and 200 meters, given some relevant indicators of each event. All the features are interval-valued except in "B100ml-P", "B100ml-I", "B200ml-P" and "B100ml-I", where there are mixed interval-valued and fuzzy-valued data, obtained by reconciling different measurements taken by three different observers.
- "**Dyslexic**" datasets: This set is composed by 3 datasets that are used to diagnose whether one child is dyslexic or not. All the datasets contain mixed interval-valued and crisp data.

In all cases we have a certain degree of imbalance and vagueness in the perception of the features and the class. This experimentation is intended to assess the performance of GFSs designed for being used with low quality data, when applied to both unprocessed and preprocessed datasets.

```
Algorithm LowQuality_Imbalanced(Dataset,Minority,N,k)
1   if (Minority == ∅ and N == ∅) then
2     Minority[] = 0
3     N[] = 0
4     for example in {1,...,N}
5       if ({class(example)}.size == 1) then
6         Minority[class(example)]= Minority[class(example)]+1
7       end if
8     end for example
9     order(Minority)
10    for class in {1,...,Majority}
11      N[class] = (int) Minority[Majority] / Minority[class]
12    end for class
13  end if
14  for Minority in {1,...,Majority}
15    Sample = ∅
16    for example in {1,...,N}
17      if (Minority ⊂ {class(example)}) then
18        Sample = Sample ∪ example
19      end if
20    end for example
21    euclidean[] = 0
22    for Sample_i in {1,...,N}
23      for Sample_j in {1,...,N}
24        euclidean[j] = distance(i,j)
25      end for Sample_j
26      ranking(euclidean)
27      for N in {1,...,N[Minority]}
28        neighbour = random (1,k)
29        synthetic = ∅
30        for Attribute in {1,...,M}
31          dif= Attribute(Sample[neighbour]) ⊖
            Attribute(Sample_i)
32          gap = random (0,1)
33          Sum = Attribute(Sample_i) ⊕ (dif ⊗ gap)
34          synthetic = synthetic ∪ range(Sum)
35        end for Attribute
36        Dataset = Dataset ∪ synthetic
37      end for N
38    end for Sample_i
39  end for Minority
return Dataset
```

Fig. 3. Algorithm to preprocess low quality imbalanced data.

A. Settings

All the datasets used in this section have been introduced in [26] and [29]. It is remarked that all of them have imprecise inputs and outputs. A brief description of this data is provided in Table IV, where it is shown for all of them the name, the number of examples (Ex), the number of attributes (Atts) and the number of classes. We have also computed the fraction of patterns in each class. The percentage of instances assigned to each class is deduced from the numbers of instances that belong to this class and from the instances with an imprecise output that contain this class.

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 in "athlete" datasets and 4 in "dyslexia" dataset. All

TABLE III

MEANS OF 100 REPETITIONS OF THE GFS FROM THE LOW QUALITY "ATHLETE" DATASETS WITH 5 LABELS/VARIABLE WITH ORIGINAL AND PREPROCESSED DATASETS.

	GFS Low Quality		**GFS Low Quality with preprocessing**	
Dataset	Error Train	Error Test	Error Train	Error Test
Long-4	[0.003,0.288]	[**0.323,0.592**]	[0.097,0.210]	[**0.245,0.514**]
BLong-4	[0.006,0.276]	[**0.326,0.625**]	[0.110,0.201]	[**0.254,0.554**]
100ml-4-I	[0.070,0.273]	[0.176,0.378]	[0.166,0.282]	[0.174,0.375]
100ml-4-P	[0.066,0.280]	[0.176,0.355]	[0.122,0.260]	[0.168,0.347]
B100ml-I	[0.075,0.281]	[0.172,0.369]	[0.191,0.277]	[0.169,0.367]
B100ml-P	[0.066,0.275]	[0.160,0.349]	[0.146,0.255]	[0.161,0.350]
B200ml-I	[0.011,0.264]	[**0.232,0.476**]	[0.270,0.364]	[**0.125,0.370**]
B200ml-P	[0.002,0.273]	[**0.262,0.480**]	[0.119,0.207]	[**0.261,0.479**]

TABLE IV

SUMMARY DESCRIPTIONS OF THE DATASETS.

Dataset	Ex.	Atts.	Classes	%Classes
Long-4	25	4	(0,1)	([36,64],[36,64])
BLong-4	25	4	(0,1)	([36,64],[36,64])
100ml-4-I	52	4	(0,1)	([0.44,0.63],[0.36,0.55])
100ml-4-P	52	4	(0,1)	([0.44,0.63],[0.36,0.55])
B100ml-I	52	4	(0,1)	([0.44,0.63],[0.36,0.55])
B100ml-P	52	4	(0,1)	([0.44,0.63],[0.36,0.55])
B200ml-I	19	4	(0,1)	([0.47,0.73],[0.26,0.52])
B200ml-P	19	5	(0,1)	([0.47,0.73],[0.26,0.52])
Dyslexic-12	65	12	(0,1,2,4)	([0.32,0.43],[0.07,0.16], [0.24,0.35],[0.12,0.35])
Dyslexic-12-01	65	12	(0,1,2)	([0.44,0.53],[0.24,0.35], [0.12,0.30])
Dyslexic-12-12	65	12	(0,1,2)	([0.32,0.43],[0.32,0.52] [0.12,0.30])

the imprecise experiments were repeated 100 times with bootstrapped resamples of the training set. The preprocessing method applied in this work uses the three nearest neighbors and balances all the classes taking into account the imprecise outputs, where the parameter "N" is estimated by the algorithm, unless when specified otherwise. The method is applied for preprocessing the 100 bootstrapped resamples of the training set.

B. Compared results

The behaviour of that GFS which is able to use low quality data, when applied to both unprocessed and preprocessed "Athlete" datasets, is shown in Table III. This Table includes 3 columns. The first one, "Dataset", contains the names of the datasets. The second one, "GFS Low Quality", contains the errors obtained by the GFS for the unprocessed, original datasets. The last column, "GFS Low Quality with preprocessing", contains the errors produced by the GFS when the datasets have been preprocessed with the proposed mechanism. The interval-valued errors shown in this table represent the minimum and maximum error of all the obtained errors.

We have observed that, the application of the proposed preprocessing mechanism, causes the GFS to improve its behaviour with respect to the low quality dataset in those cases where the imbalance degree can be regarded as "medium". This kind of datasets comprises "Long-4", "BLong-4", "B200ml-I" and "B200ml-P". However, in "B200ml-P" we have not detected a noticeable improvement of the results. We think that this is due to the fact that this low quality dataset has an attribute (the expert knowledge of the trainer) that has not a definite relation with the other four attributes; this kind of uncorrelated features is known to have bad effects on the preprocessing stage used in this paper.

TABLE V

CONFUSION MATRIX FOR LOW QUALITY DASATES OF ATHLETICS.

	GFS Low Quality		GFS Low Quality Pre.	
Long-4				
	Class 0	Class 1	Class 0	Class 1
Class 0	2591	**3168**	3098	**2661**
Class 1	**2186**	3463	**1974**	3675
BLong-4				
	Class 0	Class 1	Class 0	Class 1
Class 0	2379	3720	3307	2792
Class 1	2005	3764	2245	3524
100ml-4-I				
	Class 0	Class 1	Class 0	Class 1
Class 0	9352	2867	8044	4175
Class 1	4346	6393	2986	7753
100ml-4-P				
	Class 0	Class 1	Class 0	Class 1
Class 0	9135	2974	9005	3104
Class 1	3863	6626	3549	6940
B100ml-4-I				
	Class 0	Class 1	Class 0	Class 1
Class 0	9286	2693	8009	3970
Class 1	4298	6261	2949	7610
B100ml-4-P				
	Class 0	Class 1	Class 0	Class 1
Class 0	9164	2945	8618	3491
Class 1	3754	6775	3195	7334
B200ml-I				
	Class 0	Class 1	Class 0	Class 1
Class 0	3983	696	4093	586
Class 1	2355	744	1745	1354
B200ml-P				
	Class 0	Class 1	Class 0	Class 1
Class 0	3973	686	2518	2141
Class 1	2368	571	855	2084

As we expected, we have obtained similar results for

TABLE VI

MEANS OF 100 REPETITIONS OF THE GFS FROM LOW QUALITY DATASETS OF TYPE "DYSLEXIC" WITH 4 LABELS/VARIABLE WITH THE ORIGINAL DATASET AND PREPROCESSED.

	GFS Low Quality		**GFS Low Quality with preprocessing**	
Dataset	Train	Exh.Test	Train	Exh.Test
Dyslexic-12				
M=∅ N=∅	[0.002,0.227]	[0.443,0.590]	[0.165,0.241]	[0.437,0.590]
M=[0,1,2,3] N=[1,2,2,1]	[0.002,0.227]	[**0.443,0.590**]	[0.121,0.216]	[**0.422,0.547**]
Dyslexic-12-01				
M=∅ N=∅	[0.004,0.188]	[0.344,0.476]	[0.131,0.199]	[0.375,0.520]
M=[0,1,2] N=[1,2,1]	[0.004,0.188]	**[0.344,0.476]**	[0.100,0.183]	[**0.337,0.450**]
Dyslexic-12-12				
Min.=∅ N=∅	[0.003,0.237]	[**0.386,0.557**]	[0.118,0.196]	[**0.362,0.540**]
M=[0,1,2] N=[2,1,2]	[0.003,0.237]	[**0.386,0.557**]	[0.100,0.193]	[**0.355,0.516**]

the datasets "100ml-4-I", "100ml-4-P", "B100ml-4-I" and "B100ml-4-P" because these datasets have a low or null imbalance; the number of instances with imprecise outputs is not very high (19%) and the percentages of examples in each class (without taking into account the instances with imprecise outputs) are homogeneous (54%,45%) as shown in Table IV. However in "Long-4" and "BLong-4" the number of instances with imprecise outputs is higher, up to 28% of the total of instances, and in "B200ml-I" and "B200ml-P' this percentage is of 26%. Relevant for the practitioners of the real world problem where this data has been taken from, although we have applied a preprocessing method, we still obtain better results when we are using the knowledge of the coach, except in "B200ml-P", as we had also found in [26].

In Table V we have displayed the confusion matrix for "athlete" datasets. We can check how the FN and FP decrease in the datasets with a imbalance that we might label as "not low".

TABLE VII

CONFUSION MATRIX FOR THE LOW QUALITY DATASET "DYSLEXIC-12".

	GFS Low Quality			
Dyslexic-12				
	Class 0	Class 1	Class 2	Class 4
Class 0	6499	**222**	1612	495
Class 1	1910	178	942	**98**
Class 2	2242	**15**	3472	574
Class 4	3246	**37**	2479	88
	GFS Low Quality Preprocessing			
Dyslexic-12				
	Class 0	Class 1	Class 2	Class 4
Class 0	4666	**1753**	656	1753
Class 1	584	872	504	**1168**
Class 2	146	**1330**	2457	2504
Class 4	612	**1368**	1476	3193

The behaviour of that GFS able to use low quality data when applied to the "Dyslexic" datasets is shown in Table VI, where the error of this GFS when the datasets are either unprocessed or preprocessed, is shown. In addition, this table shows the behaviour of the GFS, when the parameter "N" is obtained with the preprocessing method proposed and also when we specify which classes are going to be balanced with the parameter "Minority" (M) and also their amount with the parameter "N". These two parameters have been obtained through the study performed with the confusion matrix obtained with the original datasets.

TABLE VIII

CONFUSION MATRIX FOR LOW QUALITY DATASETS "DYSLEXIC-12-01" AND "DYSLEXIC-12-12".

	GFS Low Quality			GFS Low Quality Pre.		
Dyslexic-12-01						
	M=∅ and N=∅			M=[0,1,2] and N=[1,2,1]		
	Class 0	Class 1	Class 2	Class 0	Class 1	Class2
Class 0	8902	902	104	7031	2460	418
Class 1	3264	3277	438	1399	5078	501
Class 2	3849	1731	838	1797	3611	1010
Dyslexic-12-12						
	M=∅ and N=∅			M=[0,1,2] and N=[2,1,2]		
	Class 0	Class 1	Class 2	Class 0	Class 1	Class2
Class 0	3911	4105	103	6628	972	518
Class 1	1836	7139	564	3153	3600	2786
Class 2	1158	4331	659	2391	1635	2122

Lastly, we can deduce from the information in Table VI that, when the assignments Minority $= \emptyset$ and N $= \emptyset$ are made, the preprocessing method does not influence the performance of the GFS. This is a consequence of the relationship that exists between the classes.

Apart from this, from the confusion matrix of "Dyslexic-12" (see Table VII) we can detect a relationship between the balanced classes, but for the most part this fact is only relevant for researchers working in the medical diagnosing of dyslexia. The GFS has a bias towards "class 1" and "class 4" (less frequent instances in the original dataset). This can be explained by the fact that, when one child is classified as "class 1',' is very probable that this child will be "class 0" in the next evaluation (and less often "class 2"). The same happens with "class 4" and "class 2-1" [29]. Therefore, "class 1" seems to be of little relevance and the results of "Dyslexic-12-01" and "Dyslexic-12-12" seem to confirm it. Otherwise, in Table VI, we observe that, if we study the confusion matrix from the original dataset, we can specify the parameter "Minority" and "N" and obtain improvements in the performance of the GFS.

V. Conclusions and Future Works

In this work we have considered the use of low quality imbalanced datasets in combination with certain GFSs that are able to use low quality data. We have studied different preprocessing methods for imbalanced datasets and used the SMOTE algorithm as a base to propose a new algorithm able to preprocess low quality imbalanced datasets.

The results have shown us how the behavior of a GFS is improved when using the preprocessing mechanism proposed here. In addition, we have observed that after applying the preprocessing method to a low quality dataset, with a low percentage of imprecise outputs or with a low degree of imbalance, the GFS has a similar behaviour to the original dataset, as we expected. Also, we have seen that, studying the confusion matrix obtained with the original dataset, we can estimate the parameters "Minority" and "N" needed in the preprocessing method.

In future works, we intend to incorporate information about the confusion matrix of the minimum error-based GFS into the preprocessing algorithm. This information can be used to fine tune the synthesis of instances in combination with a particular GFS. We have also observed that multiclass datasets might better suited for an internal approach that takes into account the cost of misclassification for each pair of classes (i.e. a minimum risk-based approach). In the last place, we think possible that, in those cases where the output variable is vague with high probability, and therefore we are not sure that the dataset is imbalanced, some techniques used in semi-supervised learning can be introduced in the preprocessing stage.

Acknowledgments

This work was supported by the Spanish Ministry of Education and Science, under grants TIN2008-06681-C06-04, TIN2007-67418-C03-03, and by Principado de Asturias, PCTI 2006-2009.

References

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

[2] Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P., SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligent Research 16, 321-357 (2002).

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

[4] Chen S. H, Operations on fuzzy numbers with function principal. Tamkang Journal of Management Sciences, 6(1), 13-25.(1985)

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

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

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

[8] Chen S. H., Ranking generalized fuzzy number with graded mean integration. In Proceedings of the eighth international fuzzy systems association world congress, Vol. 2. (pp. 899-902) (1999).

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

[10] Crockett K., Bandar Z., O'Shea J., On producing balanced fuzzy decision tree classifiers. IEEE Internat. Conf. on Fuzzy Systems 1756-1762, 2006.

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

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

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

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

[15] Fernández A., del Jesús M.J., Herrera F., On the 2-tuples based genetic tuning performance for fuzzy rule based classification systems in imbalanced data-sets. Information Sciences. DOI: 10.1016/j.ins.2009.12.014 (2010).

[16] Hart P., The condensed nearest neighbor rule. IEEE Trans. Inform. Theory 14, 515-516 (1968).

[17] Japkowicz N., Stephen S., The class imbalance problem: a systematic study. Intelligent Data Anal. 6 (5), 429-450, 2002.

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

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

[20] Kilic K., Uncu O., Türksen I.B., Comparison of different strategies of utilizing fuzzy clustering in structure identification. Information Sciences 177 (23), 5153-5162 (2007).

[21] Kubat M., Matwin S., Addressing the curse of imbalanced training sets: one-sided selection. Internat. Conf. Machine Learning, 170-186 (1997).

[22] Laurikkala J., Improving identification of difficult small classes by balancing class distribution. T.R. A-2001-2, University of Tampere (2001).

[23] Liang C., Wu J., Zhang J., 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)

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

[25] Murakami S., Maeda S., Imamura S., Fuzzy decision analysis on the development of centralized regional energy control system, IFAC Syrup. on Fuzzy Inform. Knowledge Representation and Decision Anal., 363-368 (1983).

[26] Palacios, A., Couso, I., Sánchez, L. Future performance modeling in athletism with low quality data-based GFSs. Accepted (2010)

[27] Palacios, A., Sánchez, L., Couso, I. A minimum-risk genetic fuzzy classifier based on low quality data. Lecture Notes in Computer Science 5572, Hibrid Artificial Intelligence Systems (HAIS) pp 654-661 (2009)

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

[29] Palacios, A., Sánchez, L., Couso, I. Diagnosis of dyslexia from vague data with Genetic Fuzzy System. IJAR. Accepted (2010)

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

[31] Phua C., Alahakoon D., Lee V., Minority report in fraud detection: classification of skewed data. SIGKDD Explorations Newsletter 6 (1), 50-59, 2004.

[32] Ramli N., Mohamad D., A comparative analysis of centroid methods in ranking fuzzy numbers. European Journal of Scientific Research, 28 (3): 492-501 (2009)

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

[34] Soler V., Cerquides J., Sabria J., Roig J., Prim M., Imbalanced datasets classification by fuzzy rule extraction and genetic algorithms. IEEE Internat. Conf. Data Mining –Workshops, 330-336, 2006.

[35] Tomek I., Two modifications of cnn. IEEE Trans. Systems Man Comm. 6, 769-772 (1976)

[36] Visa S., Ralescu A., Learning imbalanced and overlapping classes using fuzzy sets. Internat. Conf. Machine Learning –Workshop on Learning from Imbalanced Datasets II, 2003.

[37] Visa S., Ralescu A., The effect of imbalanced data class distribution on fuzzy classifiers–experimental study. IEEE Internat. Conf. on Fuzzy Systems, 749-754, 2005.

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

[39] Weiss G., Mining with rarity: a unifying framework. SIGKDD Explorations 6 (1), 7-19 (2004).

[40] Wilson D.R., Asymptotic properties of nearest neighbour rules using edited data. IEEE Trans. Systems Man Comm. 2(3), 408-421 (1972).

[41] Xu L., Chow M., Taylor L., Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification e-algorithm. IEEE Trans. Power Systems 22(1), 164-171, 2007.