# First study of the behaviour of genetic fuzzy classifier based on low quality data respect to the preprocessing of low quality imbalanced datasets

Ana M. Palacios, Luciano Sánchez and Inés Couso

*Abstract*— **There are real-world dataset where we can found classes with a very different percentage of patterns between them, that is to say we have classes represented by many examples (high percentage of patterns) and classes represented by few examples (low percentage of patterns). These kind of datasets receive the name of "imbalanced datasets". In the field of classification problems the imbalanced dataset are a focus of study both in preprocessing mechanisms and in classification systems. In this paper we study the behaviour of genetic fuzzy system (GFS) respect to imbalanced datasets, where this GFS is able to support low quality data. We will analyse the different preprocessing mechanisms of imbalanced datasets and will show the necessity of extending theses preprocessing mechanisms a "low quality imbalanced datasets". In addition, we include a comprehensive description of the new algorithm to able to preprocessing low quality imbalanced datasets. Severals real-world, low quality imbalanced datasets, are used to evaluate the results obtained with the GFS after using the new algorithm proposed in this paper.**

## I. Introduction

GFSs depend on fuzzy rule-based systems (FRBS), that deal with fuzzy logic and "IF-THEN" rules [9]. These FRBSs use fuzzy sets to describe subjective xknowledge about a classifier or a regression model, which otherwise accept crisp inputs and produce crisp outputs. However, in our prior works we extended the Genetic Fuzzy Classifiers to able to use low quality data [27], [28], [26]. In the classification problems is commom found classes with a different percentage of patterns between them, that is to say classes represented by many instances (known as negative classes) and classes represented by few instances (known as positive classes). These problem receive the name classification problems with imbalanced datasets.

Most classifiers that works with imbalanced datasets have a poor perform because they are designed to minimize the global error rate. They usually have tendency toward the majority classes trying to maximize the accuracy. There are studies that show that most classification methods lose their classification ability when dealing with imbalanced data [17], [30].

In this paper we study the behaviour of GFSs in the field of imbalanced datasets where these datasets contain low quality data both in the input as in the output. We are interested in the preprocessing of these low quality imbalanced dataset and the effect caused in the GFS once balanced the data. In the literature, we can found few works that study the use of fuzzy classifiers for the imbalanced dataset problem [10], [33], [35], [36] and the E-Algorithm [40] that uses a linguistic approach. Others works [12], [13], [14], [15] employ a preprocessing step in order to balance the training data before the training because these preprocessing methods are vey useful when dealing with imbalanced dataset problems[1]. In [12] we can find a study about the effect of imbalance between the classes in the framework of FRBS and also show the necessity to apply a re-sampling procedure, specifically, the "Synthetic Minority Oversampling Technique" (SMOTE) [2] that obtains a very good behaviour.

So the aim of this paper is obtain a new algorithm able to balancing the low quality imbalanced datasets from SMOTE. These dataset are imbalanced due to theirs percentage of examples in the different classes or by the number of imprecise output. Therefore, the percentage of examples in the different classes will be imprecise if the dataset contain imprecise output. To extend the SMOTE we have take into account the fuzzy arithmetic operators reviewed in [4] and [8] and the ranking of fuzzy numbers. Many authors have investigated various ranking methods since that in [18], [19] employed the concept of maximizing set to order the fuzzy numbers. The decision process to ranking of fuzzy numbers has a important consequent in the minimum risk problem. Ranking or comparison of fuzzy numbers is not an easy task and in this paper we will focus in the centroid index rakning method [11], [6], [7], [23], [32], [37] which is the most commoly used techniques in the application of ranking numbers [31].

Finally, after obtaining the new algorithm to balacing low quality dataset, we will analyse the behaviour of the GFS proposed in [26] using preprocessing of low quality imbalanced dataset before the learning phase. For this, we will compare the results obtained in severals real-world about the diagnosis of dyslexic [28] and the perform in a competitiom of athletics[26].

The structure of this paper is as follows: in the next section, Section 2, we introduce the problem of imbalanced dataset and some preprocesing techniques for imbalanced datasets, highlighting the SMOTE algorithm [2]. In Section 3 we present the new algorithm to balancing low quality imbalanced datasets take into account the imprecise output of the dataset. In Section 4 we show the results obtained in

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

the GFS able to use low quality data applying the algorithm proposd here. Also we will compare these results with the results obtained by the GFS with the original low quality dataset. The paper finishes with the conclusions and future works, in Section 5.

## II. Imbalanced datasets in classification

In this section we introduce the imbalanced dataset problem and we will show some preprocessing methods that are commonly applied in the imbalanced dataset, highlighting the SMOTE algorithm.

### A. The problem of the imbalanced dataset

The problem of imbalanced datasets in classification occurs when the number of instances of one class is much lower than the instances of the other classes. Specifically when the dataset has only two classes because one class is represented by a high number of examples, while the other one is represented by only few examples [3]. Some authors have named this problem "datasets with rare classes" [38].

TABLE I
Confusion matrix for a problem of two class

| | 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 the medical applications [20], [24], [29], for example children with dyslexic. The others class represents the counterpart of the concept, for example children without dyslexic. The evaluation of the performance of classifier, tradicionally, is based on the confusion matrix. The Table I shows a confusion matrix for a problem of two class. From this table the average classification error is defined as the total number of misclassified example divided by the total number of available examples (1) (in (2) the accuracy).

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

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

Therefore the classifier algorithms have a bias towards the majority class. This implies that the instances that belong to the minority class are misclassified more often than the other classes.

### B. Preprocessing imbalanced dataset

To deal with the imbalanced dataset problem we can apply internal approaches that create new algorithms or modify existing ones taking into account this problem. Also we can apply external approaches that preprocess the data in order to diminish the effect caused by their class imbalance. Has been proved that applying a preprocessing method to balance the class is a positive solution to the problem of imbalanced datasets [1]. In [1], [12] studied different methods of preprocessing where these methods are classified in three kind:

- Under-sampling methods: Obtain a subset of the original dataset by eliminating some of the examples of the majority class. These methods are Condensed nearest neighbour rule (CNN) [16], Tomek links [34], One-sided selection (OSS) [21], Neighbourhood cleaning rule (NCL) [22], Wilson's edited nearest neighbour (ENN) [39] 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: over-sampling + under-sampling: Obtain a set by combining the two previous methods. These methods can be SMOTE+Tomek Link and SMOTE+ENN.

In [12] compared these preprocesing methods with FRBCSs, showing the good behaviour for the over-sampling methods, specially in the case of the 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/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 the Figure 1 is shown this example, where $x_i$ is the selected point, $x_{i1}$ to $x_{i4}$ are some selected nearest neighbour and $r_1$ to $r_2$ the synthetic data points created by the randomized interpolation.

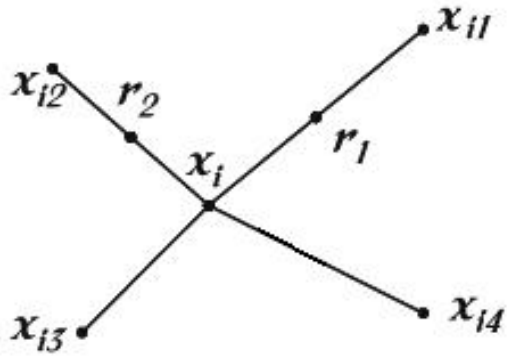


Fig. 1. Creation of synthetic data points in the SMOTE algorithm.

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 explain in II-A, the problem of imbalanced datasets in classification occurs when the number of instances of one class is much lower than the instances of the other classes. This also happens when the dataset contain low quality data, as interval-valued or fuzzy numbers. However, we must take into account that these dataset also can have imprecise output. For instance, if one instance is labeled as "classA,B" we are not saying that this example belongs to both categories at the same time (wich is not a imprecise output), where are saying that the output can be A or B and we do not know which one is the correct. Due to these imprecise output the percentage of instances that belong a one class will be defined by a imprecise value. As example in the Table III is shown the dataset "Long-4" [26] which contains imprecise input and output.

TABLE III
PERCENTAGE OF INSTANCES IN THE DIFFERENT CLASSES, IN THE DATASET "LONG-4".

| Dataset | Instances | Atributtes | Classes | Imprecises Output | %Classes |
|---|---|---|---|---|---|
| Long-4 | 25 | 4 | (0,1) | 7 | ([36,64],[36,64]) |

This factor, imprecise output, must be taken into account when we are preprocessing the low quality imbalanced dataset. To preprocess these kind of datasets we propose a new algorithm based in the SMOTE algorithm explained in the section II-C.

The algorithm proposed here has to consider three important aspect due to is working with datasets that contain fuzzy input and imprecises output. This is:

1) Selection of the minority class and the amount the synthetic examples
2) Computer the $k$ nearest neighbours from the example selected. The implementation applied in this work uses the euclidean distance to select the $k$ nearest neighbour and it implies the uses of fuzzy arithmetic operators and the ranking of fuzzy numbers.
3) Generation of the synthetic example of the minority class. For it, we use fuzzy arithmetic operators and we have to control the values out of range in the differents attributes.

### A. Selection of the minority class

In [2] the number of minority class samples (T) and the amount of synthetic examples (N) are inputs of the SMOTE algorithm. Now the percentage of example in one class is defined by a imprecise value therefore, the own algorithm will determine the amount of synthetic example for each class. Also we have the option to indicate in the algorith the minority classes and the amount of synthetic examples (N) in each minority class. In the Figure 4 in the lines 1 to 13, we observe as the algorithm obtain the amount of sysnthetic examples (N) for each class. All classes, until the majority, will obtain synthetic examples and the this way the examples with imprecise output will have less relevance in the classification.

### B. Computer the $k$ nearest neighbours

Before selecting the $k$ nearest neighbours we collect all the examples that before to the minority class although, these examples have imprecise output, see Figure 4 lines 15 to 20. Then we obtain the $k$ nearest neighbours of the example selected by the euclidean distance (lines 21 to 25). It implies the uses of fuzzy arithmetic operators and the ranking of fuzzy numbers to order these distances. The euclidean distance with fuzzy arithmetic operators [4], [8] is shown in (3):

$$D(i,j) = \left[\bigoplus_{m=1}^{n} (abs(\tilde{x}_{im} \ominus \tilde{x}_{jm}))^2\right]^{\frac{1}{2}} \tag{3}$$

where

$$abs\left(\tilde{A}\right) = \left|\tilde{A}\right| * 1$$

where

$$\left|\tilde{A}\right| = \begin{cases} A.a = A.a * -1 & A.a < 0 \\ A.b = A.b * -1 & A.b < 0 \\ A.c = A.c * -1 & A.c < 0 \\ A.d = A.d * -1 & A.d < 0 \end{cases}$$

and $\tilde{A}$ is a trapezoidal fuzzy number as $\tilde{A} = (a, b, c, d)$, see Figure 2.

In the Figure 4 in the line 26 , we have to use the operation "ranking" for determing the $k$ nearest neighbours of the example selected. In [31] is showing as no single ranking method in the centroid concept is superior to all other methods in ranking fuzzy numbers since each method appears to have some advantages as well as disadvantages. Therefore, in this porposal we are going to use the ranking method defined in [37] that is improved the [11]. The ranking

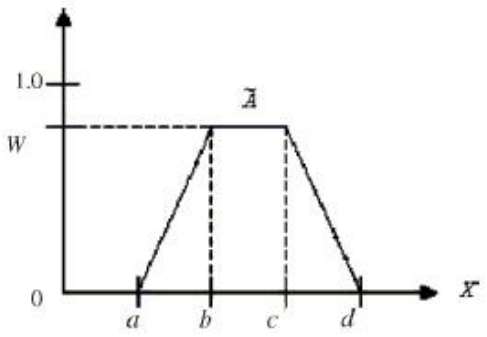


Fig. 2. A generalized trapezoidal fuzzy number.

method is based in $\widetilde{x}(A)$ value and in $\widetilde{y}(A)$ value but however $\widetilde{x}(A)$ and $\widetilde{y}(A)$ have different degrees of importance on a fuzzy number A. For example, in Figure 3, $\widetilde{x}(A)$ indicates the representative location of fuzzy number A, and $\widetilde{y}(A)$ presents the average height of the fuzzy number. To rank fuzzy numbers, we know that the importance of the degree of representative location is higher than average height [37]. Therefore, for any two fuzzy numbers A and B, we have following situations [37]:

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

where $\widetilde{x}(A)$ and $\widetilde{y}(A)$ for a fuzzy number A is defined as [5], [25]:

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

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

and 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.

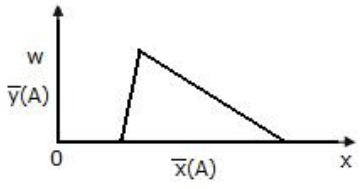


Fig. 3. The $\widetilde{x}(A)$ and $\widetilde{y}(A)$ of fuzzy number A.

### C. Generation of the synthetic example

The generation of the synthetic examples, as in [2], consist in take the difference between the feature vector (sample) under consideration and its nearest neighbor. Multiply this difference by a random number between 0 and 1, and add it to the feature of the synthetic example. For it, we use fuzzy arithmetic operators, see Figure 4 lines 31 to 34, and we control the values out of range in the differents attributes, line 34.

**Algorithm** LowQuality_Imbalanced(Dataset,Minority,N,k)
1 **if** (Minority == $\emptyset$ **and** N == $\emptyset$) **then**
2 Minority[] = 0
3 N[] = 0
4 **for** example in $\{1, \ldots, \mathrm{N}\}$
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 $\{1, \ldots, \mathrm{Majority}\}$
11 N[class] = (int) Minority[Majority] / Minority[class]
12 **end for** class
13 **end if**
14 **for** Minority in $\{1, \ldots, \mathrm{Majority}\}$
15 Sample = $\emptyset$
16 **for** example in $\{1, \ldots, \mathrm{N}\}$
17 **if** (Minority $\subset$ {class(example)}) **then**
18 Sample = Sample $\cup$ example
19 **end if**
20 **end for** example
21 $\widetilde{euclidean}$[] = 0
22 **for** Sample_i in $\{1, \ldots, \mathrm{N}\}$
23 **for** Sample_j in $\{1, \ldots, \mathrm{N}\}$
24 $\widetilde{euclidean}$[j] = distance(i,j)
25 **end for** Sample_j
26 ranking($\widetilde{euclidean}$)
27 **for** N in $\{1, \ldots, \mathrm{N[Minority]}\}$
28 neighbour = random (1,k)
29 $\widetilde{synthetic} = \emptyset$
30 **for** Attribute in $\{1, \ldots, \mathrm{M}\}$
31 $\widetilde{dif}$= $\widetilde{Attribute}$(Sample[neighbour]) $\ominus$ $\widetilde{Attribute}$(Sample_i)
32 gap = random (0,1)
33 $\widetilde{Sum}$ = $\widetilde{Attribute}$(Sample_i) $\oplus$ ($\widetilde{dif}$ $\otimes$ gap)
34 $\widetilde{synthetic}$ = $\widetilde{synthetic}$ $\cup$ range($\widetilde{Sum}$)
35 **end for** Attribute
36 Dataset = Dataset $\cup$ $\widetilde{synthetic}$
37 **end for** N
38 **end for** Sample_i
39 **end for** Minority
**return** Dataset

Fig. 4. Algorithm to preprocess low quality imbalanced data.

## IV. Numerical Results

The problem of imbalanced datasets is extremely significant because it is implicit in most real world applications, particularly in medical applications [20], [24], [29]. In this section we have several real-world about the medical applications, the diagnostic of dylexic [28], with low quality imbalanced dataset. So, we have compared the results of the GFS with low quality dataset, applying the preprocessing method proposed here and the results obtained with the

original dataset. Moreover, we will study the behaviour of the GFS respect low quality imbalanced dataset of athletics [26].

### A. Settings

All the datasets use in this section have been introduced in [26] and [28] and all have imprecise input and output. A brief descripcion is provided in Table IV showing for each dataset the name, the number of examples (Ex), number of attributes (Atts), the classes and the percentage of patterns the each class.

TABLE IV
DATASETS SUMMARY DESCRIPTIONS.

| 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]) |
| B00ml-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]) |

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 to athletes's datasets and 4 to datasets of dyslexic. All the imprecise experiments were repeated 100 times with bootstrapped resamples of the training set. The preprocessing method applied in this work uses three nearest neighbour and balances all the classes taking into account the imprecises output where "N" is estimated by the algorithm, except when we specify otherwise. It is apply a all low quality imbalanced dataset, that is to say we preproces the 100 bootstrapped resamples of the training set.

### B. Compared results

The behaviour of the GFS able to use low quality data respect to the preprocessing method proposed from Athletics's dataset is shown in the Table V.

We observe that applying the preprocessing mechanism proprosed, the GFS improves its behaviour respect to the low quality dataset where theirs imbalance can be considered "medium". This kind of datasets can be "Long-4", "BLong-4", "B200ml-I" and "B200ml-P". However, in "B200ml-P" we apreciate that the behaviour of the GFS is similar due to this low quality dataset has five attributes where the last one is the knowlegde of the trainer and it seems that has not a relation between with the others four attributes.

As we expected the datasets "100ml-4-I", "100ml-4-P", "B100ml-4-I" and "B100ml-4-P" obtain similar results because these datasets are considered with a imbalance "low" or not imbalances due to the numbers of imprecises output is not very high 19% and the percentage of example in each class is very close (54%,45%), see Table IV . However in "Long-4" and "BLong-4" the numbers of imprecises output is 28% and in "B200ml-I" and "B200ml-P' the 26%. Addition, although we apply a preprocessing method we obtain better results when we are using the knowlegde of the coach, except in "B200ml-P", as in [26].

TABLE V
MEANS OF 100 REPETITIONS OF THE GFS FROM THE LOW QUALITY ATHLETIC'S DATASETS WITH 5 LABELS/VARIABLE WITH THE ORIGINAL DATASET AND APPLYING PREPROCESSING.

| | GFS Low Quality | | GFS Low Quality Pre. | |
|---|---|---|---|---|
| Dataset | Train | Exh.Test | Train | Exh.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**] |

In the Table VI we have the confusion matrix the athletes's datasets and we can check as the FN and FP decrease in the datasets with a imbalance considered not "low".

TABLE VI
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 |

The behaviour of the GFS able to use low quality data respect to the preprocessing method proposed from Dyslexic's dataset is shown in the Table VII. This Table shows the behaviour of the GFS when the parameter "N" is obtained with the preprocessing method and when we specify which classes are going to be balanced with the parameter "Minority" (M) and theirs amount with the parameter "N". These two parameters have been obtained through of the study realized in the confusion matrix obtained with the original datasets.

TABLE VII
MEANS OF 100 REPETITIONS OF THE GFS FROM LOW QUALITY DATASETS OF DYSLEXIC WITH 4 LABELS/VARIABLE WITH THE ORIGINAL DATASET AND WITH PREPROCESSING.

| | GFS Low Quality | | GFS Low Quality Pre. | |
|---|---|---|---|---|
| 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**] |

We note, in Table VII that preprocessing method has not influences in the performace of the GFS. This is consecuence of the influence and relation that exist between the classes. In the confusion matrix of "Dyslecix-12", see Table VIII, we can see the relation between the balanced class. That is to say, the GFS has a bias towards the "class 1" and "class 4" (instances with a low percentage in the original dataset). This can be explained because when one child is classified as "class 1" is very probable that this child will be "class 0" in the next evaluation (and sometime "class 2" but in less percentage). The same happen with the "class 4" and "class 2-1" [28]. Therefore, the "class 1" seems to be little relevance and the results of "Dyslexic-12-01" and "Dyslexic-12-12" confirm it. Otherwise, in the Table VII, we observe that if we do a study of the confusion matrix from the original dataset, we can specify the parameter "Minority" and "N" and due to we obtain improvements in the performance of the GFS.

## V. CONCLUSIONS AND FUTURE WORKS

In this work we have considered the problem of low quality imbalanced dataset in the GFS able to use low quality data. We have to study different preprocessing methods of imbalanced dataset and we have used as base the SMOTE algorithm to propose a new algorithm able to preprocess low quality imbalanced dataset. The results have shown as the behaviour of the GFS able to use low quality data improve using preprocessing mechaninsm proposed here. Addition,

TABLE VIII
CONFUSION MATRIX FOR LOW QUALITY DATASET OF DYSLEXIC WITH 4 CLASES.

| | 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 |

TABLE IX
CONFUSION MATRIX FOR LOW QUALITY DATASET OF DYSLEXIC WITH 3 CLASES.

| | 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 |

we have observed that applying the preprocessing method a low quality dataset, with a low percentage of imprecise output or with a low imbalanced, the GFS has similar behaviour with the original dataset, as we expected. Also, we have seem that with a good study of the confusion matrix, obtained with the original dataset, we can give the parameters "Minority" and "N" in the preprocessing method.

Due to of the study in the confusion matrix in dataset with more the of class and the improvement the performance of GFS, we could consider in the preprocessing mechanism not only the percentage of classes and the imprecises output but also the matrix of confusion of the original dataset. Otherwise, we have observed that these dataset with more that two class maybe to deal with the low quality imbalanced dataset problem is better a internal approaches that modify the GFS taking into account the different classes and the costs of theses class (minimum risk). Moreover, we have observed that sometimes the low quality dataset is imbalanced due to of the imprecise output so would be desirable a algorithm able to preprocess the imprecise output.

## ACKNOWLEDGMENT

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. Submitted.

[27] 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. DOI: 10.1007/s12065-009-0024-1.

[28] Palacios, A., Sánchez, L., Couso, I. Diagnosis of dyslexia from vague data with Genetic Fuzzy System. Submitted.

[29] 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).

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

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

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

[33] 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.

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

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

[36] 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.

[37] 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)

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

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

[40] 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.