Expert Systems With Applications 234 (2023) 121011

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Enhancing soft computing techniques to actively address imbalanced regression problems

María Arteaga [a], María José Gacto [b,*], Marta Galende [c,d], Jesús Alcalá-Fdez [a], Rafael Alcalá [a]

[a] Department of Computer Science and Artificial Intelligence, University of Granada, E-18071, Granada, Spain
[b] Department of Software Engineering, University of Granada, E-18071, Granada, Spain
[c] CARTIF Centro Tecnológico, Parque Tecnológico de Boecillo, Boecillo 47151, Valladolid, Spain
[d] Department of Systems Engineering and Control, School of Industrial Engineering, University of Valladolid, Valladolid 47011, Spain

ARTICLE INFO

Keywords:
Imbalanced regression
Multi-objective evolutionary algorithms
Fuzzy rule-based systems

ABSTRACT

While research in the area of imbalance, which is understood as classes that are not equally represented, is mainly addressed in classification, it has hardly been studied in regression, where data maldistribution, or imbalance, can be defined as the existence of some specific subdomains of the output variable misrepresented in the training data set, resulting in low accuracy for new instances within these subdomains. The small amount of state-of-the-art techniques are "passive", meaning they are only applied in preprocessing. In this contribution, we propose two new specific evolutionary algorithms based on fuzzy rules to "actively" address imbalanced regression problems and improve the overall performance of the algorithms instead of just addressing the imbalance problem. The results obtained after applying statistical tests to 32 regression datasets that handle more than 3000 partitions show the effectiveness of the proposed methods when compared to the best previous proposal, a passive method called SMOGN. We can conclude: (1) we cannot affirm, since the equality hypotheses have not been rejected, that there are significant performance differences between using stratified and non-stratified data, thus we will use stratification to preserve a minimum representation of the minority set, (2) both fuzzy rule-based methods obtain better results in terms of the imbalance metric when using SMOGN, but in both methods this incurs a cost in accuracy (with confidence scores of over 99.0%), and (3) the proposed methods outperform those using SMOGN as they get slightly better results in the imbalance metric, with better average ranks in both proposals, and obtain significantly better results in global accuracy, that is, all the performance metrics studied improve statistically with a confidence score of over 99.0%, with the exception of one metric, which scores above 90.0%.

1. Introduction

Datasets available to solve real-world problems frequently present some level of imbalance in the obtained data due to the particularities of each application or problem, i.e., difficult measurement of some types of relevant cases, the existence of only a few numbers of cases (usually the most relevant ones), etc. Generally, the imbalance problem is present in real-world datasets in which a subdomain of the output variable is underrepresented. The existence of a relatively high level of imbalance in the data must be treated since, in most of the cases, it severely affects the predictive modeling ability, that is, we obtain high degradation in the performance of predictive models in general, and poor or non-existent model representation in particular underrepresented classes or data subdomains.

The problem of imbalanced domains has been widely addressed in supervised machine learning for classification problems, where the aim is to predict a class, i.e., a nominal variable (Juez-Gil et al., 2021; Singh & Purohit, 2015; Yan et al., 2022). In fact, this called *imbalanced classification* and is a particularly well-known type of application, (Wang et al., 2021). Therefore, we can find many application examples involving imbalanced classification, such as cancer detection (Ranjbarzadeh et al., 2021, 2023) (a few cases with cancer compared to many healthy ones), multi-oriented and curved text detection (Ranjbarzadeh et al., 2022) (character frequencies are quite different), etc. Data imbalance, however, has been rarely identified for regression problems, which is the other large research area within supervised machine learning, where the variable to predict is in a continuous domain and there are only a few solutions (Branco et al., 2015, 2017). This is mainly due to

* Corresponding author.
E-mail addresses: m.arteagajover@gmail.com (M. Arteaga), mjgacto@ugr.es (M.J. Gacto), margal@cartif.es (M. Galende), jalcala@decsai.ugr.es (J. Alcalá-Fdez), alcala@decsai.ugr.es (R. Alcalá).

https://doi.org/10.1016/j.eswa.2023.121011
Received 27 November 2022; Received in revised form 7 July 2023; Accepted 16 July 2023
Available online 20 July 2023

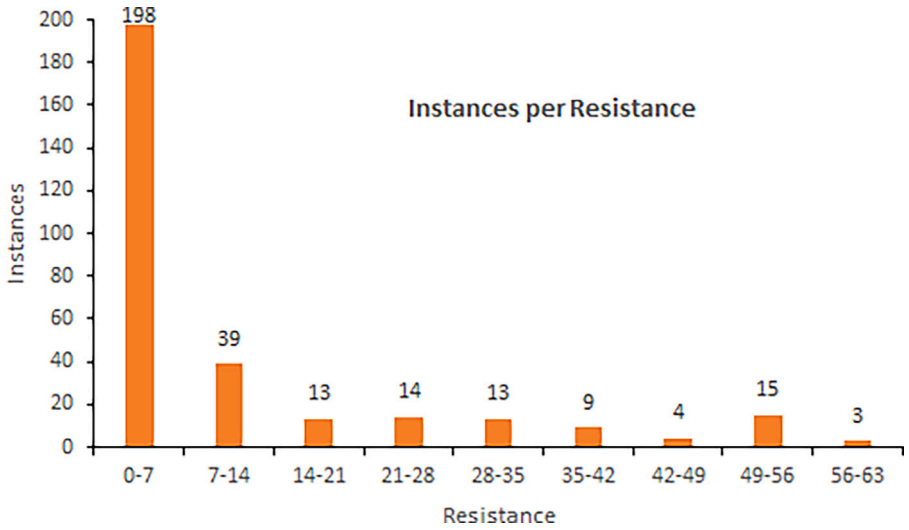

Fig. 1. Yacht Hydrodynamics dataset data distribution.

the fact that in the imbalance problem for regression tasks we find the following three differential factors:

- It is easy to identify minority groups (minority classes in classification) when they are present, but as nobody has ever considered the possibility of the problem of a complete representation of a continuous output variable domain, it has rarely been considered.
- User preferences in the variable domain, where we can find subdomains of the output variable that are more relevant than others, are not uniform.
- There is poor representation of the user relevant instances in the domain of the variable.

We can also find, however, very interesting regression problems involving some kind of imbalance in the output data distribution such as, for example, building energy load prediction (Zhang et al., 2021), where there may be much less data available for some conditions than for other conditions, or statistical downscaling of precipitation (Steininger et al., 2021), since in most locations there are far fewer rainy days than dry days. Let us focus on analyzing the Yacht Hydrodynamics dataset (Gerritsma et al., 2013) as a real-life representative problem. In this example, the experts try to predict the residuary resistance per unit displacement weight of sailing yachts based on various hydrodynamic characteristics. Predicting the resistance of yachts during the initial design stage holds significant importance in evaluating ship performance and estimating the necessary propulsive power in order to avoid wasting time and money. The main problem is that low resistance values with small non-relevant differences are easily measured and do not represent a problem for the production line, whereas a small subset of yachts was available to include high resistance values, just which we would like to predict before production. In this case, the designers are interested in high resistance values, since they are an early indication that the yacht should be redesigned and not moved to a later production phase. This represents a clear example of imbalanced regression, where user-relevant data are present in underrepresented sub-domains of the continuous output variable. See Fig. 1 for a graphical representation of the instances of this problem (Gerritsma et al., 2013) per resistance values.

Classical prediction algorithms, as well as classical evaluation measures, focus on the set of instances with the greatest representation, which leads to a degradation of the subset of user-relevant data that are underrepresented. In the example described above, these would be sailing yachts with bad or high resistance values, resulting in poor design and a waste of time and money.

The Abalone dataset (Nash et al., 1994) is also an interesting example dataset since it usually is prone to overfitting (taking into account our previous experience (Gacto et al., 2019) with this dataset). The objective of the Abalone dataset is to predict the age of abalones (i.e., number of rings in the shell[1]) using physical measurements. This problem has been approached from either a regression or classification point of view; the latter involves grouping the data into three classes ("young" < 6 rings, "adult" 6 to 13 rings, "old" > 13 rings). An analysis of the data shows that there is a large imbalance in the output variable (Rings), since there are actually a few number of values for the extreme cases (low or high number of rings). This makes it difficult to correctly predict the extreme values, which are underrepresented but still very important. For a graphical representation of the distribution of the instances of the Abalone problem (Nash et al., 1994) per number of rings, see Fig. 2.

Both types of datasets (with poor representation of relevant instances in the domain of the target variable) should be analyzed through imbalanced regression techniques. Otherwise, the models obtained from general-purpose regression techniques will only or mainly focus on minimizing error in high-density (overrepresented) domain regions, and this could lead to completely overlooking low-density (underrepresented) regions. Of course, just as in the case of classification, where minority classes representing more than the 20 or 25% of the total data are not a problem, in the case of regression we are also talking about problems involving underrepresented data with a significant imbalance ratio, so that even the most accurate learning approaches will favor significantly better modeling of high-density regions with respect to the low-density ones.

Having defined this, most of the limited number of current proposals for imbalanced regression tasks are mainly designed for solutions focused on data preprocessing techniques (Branco et al., 2015, 2017; Murphey et al., 2004; Torgo et al., 2013), which could be considered "passive" techniques from the learning process point of view. These techniques can be considered to be passive since they directly affect the data distribution and do not directly affect the learning process. These passive preprocessing proposals have been validated in the specialized literature via an analysis of their effect in the efficiency of different learning algorithms as *Support Vector Machines* (SVMs), *Random Forest*, *Multivariate Adaptive Regression Splines* and *Neural Networks* (NNs) (Murphey et al., 2004); where, although SVMs, Random Forest and the

[1] The age of the abalone is determined by cutting the shell through the cone, staining it, counting the number of rings through a microscope and adding +1.5.

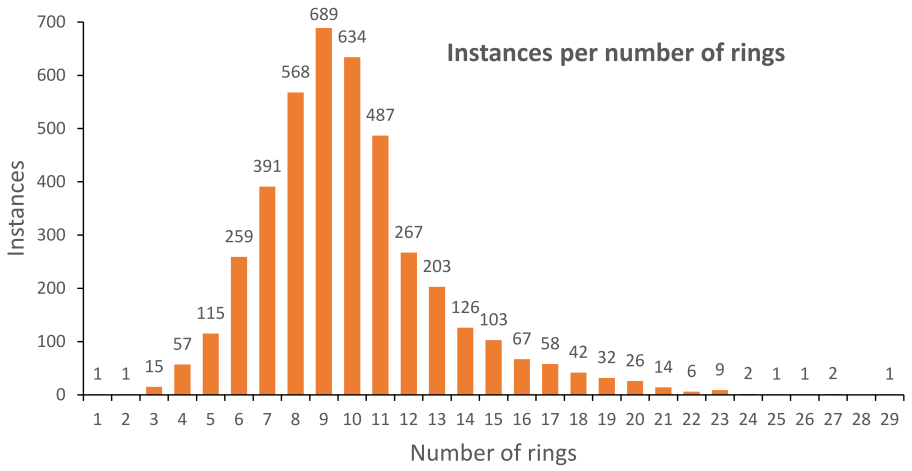

Fig. 2. Abalone dataset data distribution.

advanced NNs have shown excellent performance in many regression applications (see the experimental study on 164 regression algorithms in Gacto et al., 2019), they can further improve their performance in the imbalanced regions as shown in Branco et al. (2017) and Murphey et al. (2004).

In contrast to the use of data preprocessing techniques, the so-called learning models for a specific purpose or algorithmic-based techniques can be considered "active" from a learning process point of view, since the learning algorithm itself is designed to address the imbalance problem. With respect to these types of techniques for imbalanced regression, there is only one somewhat "related" proposal (Torgo & Ribeiro, 2007) based on a utility metric that considers any *ad-hoc* user-defined relevance function. Since this proposal does not consider the well-defined metrics specifically designed for imbalanced regression (Torgo & Ribeiro, 2009), imbalanced regression is not automatically addressed, and it relies on the user's ability for each particular dataset. It was only tested in one dataset and applied to three methods available in the R packages *rpart*, *nnet* and *e1071* (SVMs). In fact, and although there are many proposals in the classification framework, as stated in a recent review of the state-of-the-art in imbalanced scenarios (Krawczyk, 2016): "The research community has only taken its first step into the problem of imbalanced regression and further works on this topic are of vital importance" since "so far little attention has been paid to it" in the regression framework.

For the first-time, this paper will "actively" address the problem of imbalanced regression by focusing on fuzzy rule-based models that can be interpreted by humans (Barredo Arrieta et al., 2020) and therefore could be very useful in real-world applications. Rule-based learners and fuzzy logic-based models are considered interpretable approaches by the research community since they are supposed to be "transparent machine learning models that can be understood by themselves" (Barredo Arrieta et al., 2020). These rule-based predictive models allow for an easy and natural representation of the results (a necessary and even essential characteristic in certain fields such as medicine Kaieski et al., 2020, market prediction Goli et al., 2021, finance Ghorbani & Korzeniowski, 2020; Korzeniowski & Ghorbani, 2021; Pena et al., 2021, economy Ahmadi, 2021, etc.) Considering that the imposition of a rule structure implies fewer degrees of freedom than in the case of black boxes with obviously better predictive performance, we think that these models will be even more sensitive to the imbalance problem. Therefore, designing solutions to the imbalance problem for this type of descriptive modeling is of special interest to us, although the design of active techniques for other types (as SVMs, Random Forest, NNs, etc.) is still an open issue in the regression framework.

The main points that will be covered in this paper are:

- First, we will analyze the evolution of published metrics and techniques currently published to address the problem of prediction in imbalanced regression, which can model the importance of some of the data.
- We will perform a selection of datasets to consider only those that will allow us to reliably evaluate the performance of methods in imbalanced regression. We will then carry out a study on widely recognized datasets, evaluating the imbalance (in percentage) present, and selecting those that will allow the methods to be analyzed in a wide range of cases (dimensionality, size...).
- We will analyze the study of the behavior of the current algorithms based on fuzzy rules. To this end, we will carry out a new experimental and statistical study in which we will analyze and compare the behavior of "passive" techniques that currently exist in the literature for problems of imbalanced regression (Torgo et al., 2013) in order to discover which techniques best adapt to imbalanced regression problems.
- Finally, we will present the first two "active" proposals and make a comparison with the current literature. We will define two new versions of evolutionary algorithms based on fuzzy rules, so that they focus on obtaining good results for imbalanced regression problems. In order to do this, as we are dealing with genetic algorithms, we have proposed an adaptation that provides us with a way to tackle the imbalance problem and, consequently, be able to improve the evaluation metrics in imbalanced continuous domains (Torgo & Ribeiro, 2009) as well as the global Mean Square Error (MSE), i.e., always seeking to maintain a balance between performance in the relevant subset of data and performance in the total set.

We have statistically tested the first two "active" algorithms proposed for imbalanced regression on 32 regression datasets with different complexities (from 2 to 58 variables and 159 to 22784 instances) and with different imbalance percentages (from 4.12% to 20.6% imbalance). The results obtained show the effectiveness of the proposed methods on accuracy (MSE, R^2, MAE-Mean Absolute Error and $MAPE$-Mean Absolute Percentage Error) and $F1$ (imbalance metric) by applying Wilcoxon's tests (Sheskin, 2007; Wilcoxon, 1945) compared to the best previous proposal, the "passive" technique for preprocessing SMOGN. Both proposals get better rankings in the imbalance metric, but also significant improvements in the global accuracy, that is, all the performance metrics are statistically improved.

As a result, one of the main conclusions of this contribution is that active imbalanced regression techniques become a promising new line of research with respect to passive ones. Since both methods proposed in this contribution represent the first attempts at active imbalanced

regression techniques, future work could focus on designing more advanced active techniques.

This contribution is organized as follows. The state of the art in the imbalanced regression problem is shown in Section 2. The main concepts and the particular evaluation metrics for imbalanced regression are formulated in Section 3. Section 4 introduces two fuzzy evolutionary algorithms that will form the basis of the proposed active techniques in the imbalanced regression problem. Section 5 presents two specific "active" algorithms for fuzzy imbalance regression problems, trying to improve the overall performance of the algorithms beyond simply addressing the imbalance. In Section 6, the structure of the experimental study considered in this contribution is detailed. Section 7 includes a statistical initial study in which we will show the behavior of the currently existent passive techniques for problems of imbalanced regression and a comparison with our new active proposals to treat the problem of imbalance. Finally, in Section 8 some conclusions are drawn.

2. State-of-the-art in imbalanced regression

In this section, we will delve into the existing proposals for dealing with imbalanced regression problems in the current literature, analyzing the different types of existing solutions and evaluating these techniques so as to obtain a general overview. Currently, as explained in Krawczyk (2016), only a few techniques have been developed in the field of regression. They can be categorized into two main types as used in the field of classification research (Krawczyk, 2016):

- *Preprocessing*: These could be considered as "passive" techniques from a learning point of view.
- *Learning models for a specific purpose or algorithmic approaches*: These, on the other hand, could be considered "active" techniques since they actively address the imbalance problem within the learning process itself.

In this contribution, we introduce this particular consideration (passive vs. active) in order to highlight how both types of techniques contribute to the performance of the model obtained. We will look at each of them in detail in the following subsections.

2.1. Preprocessing or passive techniques

The main purpose of the existing preprocessing techniques for imbalanced data is to change its distribution, so that standard regression models focus on the instances that are most relevant to the user. These techniques are considered to be passive proposals, since they are applied only and exclusively during preprocessing, i.e. they do not directly affect the learning process. The main advantage of this strategy is that it can be applied to any existing learning model in regression problems without the need for any specific changes. However, transforming the distribution of the original dataset to a new distribution that emphasizes user preferences is not an easy task, and is dependent on the problem.

The main type of preprocessing strategy and the only one where imbalanced regression techniques have been applied is *Re-sampling*, which mainly changes the distribution of data by forcing the future model to focus on underrepresented data. In the *re-sampling* techniques for regression, the problem of the definition of the concept of the "minority class" belonging to the scope of classification arises once again. This regression concept consists, as previously mentioned, of a specific subdomain of the output variable.

Within these techniques, we found a proposal called *SMOTE for regression* (Torgo et al., 2013), which is based on the *SMOTE* method (Chawla et al., 2002) for classification and consists of the generation of synthetic instances combining over-sampling techniques in the "minority class" and under-sampling in the "majority class". To apply this proposal to regression, a relevance function was implemented and a t_r threshold was set on the relevance values (defined by the user) to discern between the relevant data. The algorithm performs over-sampling on the subset D_r of relevant data (i.e. those instances with a relevance value above t_r). Under-sampling, however, will be performed on the subset D_n of non-relevant data (those with relevance values less than or equal to t_r). The main problem with this technique is that it performs under-sampling and over-sampling processes randomly.

Subsequently, a new proposal arises when trying to solve this problem by means of an algorithm with more information obtained from the domain called SMOGN (Branco et al., 2017). This method is based on the combination of the above-mentioned technique of *SMOTE for regression* with the introduction of Gaussian noise. The under-sampling process does not change with respect to the method *SMOTE for regression*. New synthetic instances are produced for each case within the D_r set (called "seed") in the over-sampling process. *SMOGN* will use *SMOTE for regression* or Gaussian noise as a function of the distance between the "seed" and k chosen nearby neighbor. The concept of a "safe zone" is defined, where the selected neighbor is in a safe zone if it is at a considerable distance to perform the interpolation using *SMOTE for regression*. If it is not within this zone, this means that it is far away enough to perform an interpolation, and thus it would be better to generate the synthetic instance by introducing Gaussian noise. This threshold depends on the distance between the "seed" and the k closest neighbors, so the threshold is defined as the median of the distances between both of them. The objective of this method is to limit the risks of *SMOTE for regression* by performing interpolations with neighbors that are very far away from each other, and to increase the diversity of the new synthetic instances generated by over-sampling.

2.2. Specific purpose or active learning techniques

There are other types of solutions, called specific purpose learning techniques, that actively address learning in imbalanced scenarios. They can also be used for specific purposes as they actively affect the learning model (they are direct modifications of the learning process), thus making them more powerful. These solutions modify the existing prediction algorithms for regression so that they fit better in imbalanced domains. The main difficulty of these types of solutions is the specific knowledge of the data domain and the specific algorithm to be modified, since the weak points of the algorithm must be identified when dealing with imbalanced problems.

Very few studies have been carried out regarding these types of solutions for regression problems, and currently there is only one related proposal based solely on density distributions, i.e., it does not consider the well-defined metrics specifically designed for imbalanced regression (Torgo & Ribeiro, 2009). It is called Utility Based Regression (Torgo & Ribeiro, 2007), and is a proposal that tries to obtain models guided by a utility metric considering any *ad-hoc* user-defined relevance functions. Therefore, imbalanced regression data are not automatically addressed and their consideration would depend on the user ability and particular deep knowledge of the data. It was only tested on one dataset and applied to three methods available in the R software environment in the packages rpart, nnet and e1071 (svm).

3. Formal definition of the main concepts and metrics of the problem

In this section, we formally define the main concepts of the imbalanced regression problem. Then, we present the particular evaluation metrics that are typically applied in imbalanced continuous domains. The prediction task applied to datasets for regression problems (Branco et al., 2015, 2017) consists in giving an unknown function $f(X_i, X_2, \ldots, X_p)$, (p being the number of predictor variables) providing an approximation that is as accurate and close as possible to the function that defines the output variable. In the case of supervised

learning for regression problems, we have a set of training data from which we will obtain our approximate function F. This training set is defined as $D = \{\langle x_i, y_i\rangle\}_{i=1}^{n}$, where x are the input variables, y is the output variable and n is the size of the dataset.

3.1. Main concepts: Relevance distribution function, relevant data and non-relevant data

As we have seen, the main problem of regression with imbalanced distribution is that the user assigns a greater importance to the performance of a given subset of the output variable. The distinction between this subset of our data is represented (Branco et al., 2015) by a relevance function $\Phi(Y)$ (Torgo & Ribeiro, 2007), which defines the importance range of a certain value of the output variable (Y), where $\mathcal{Y}$ is the domain of Y, 0 implies no relevance and 1 the maximum relevance:

$$\Phi(Y) : \mathcal{Y} \rightarrow [0, 1] \tag{1}$$

Once we have defined a function to model user relevance in the domain of the output variable, our dataset is gradually divided into a dataset that is relevant to the user, and other datasets that are not. This is where the concept of a relevance threshold (t_r) appears, which is a set evaluated according to user distinction that allows us to subdivide our dataset by discriminating between relevant and non-relevant data.

Then, we define $D_r \in D$ as a data subset whose relevance function that is applied to the output variable is greater than t_r, shown as follows:

$$D_r = \{\langle x_i, y_y\rangle \in D : \Phi(y_i) > t_r\} \tag{2}$$

We define the subset of non-relevant or normal data denoted D_n as the difference between D and D_r, $D_n = D \setminus D_r$. It is formulated as follows:

$$D_n = \{\langle x_i, y_y\rangle \in D : \Phi(y_i) \leq t_r\} \tag{3}$$

Once we have a formal definition of the user distinction in imbalanced problems, we can formally define the existing problems in the previously mentioned imbalanced regression as follows:

- $\Phi(Y)$ is not uniform in $D(Y)$.
- The number of instances in the D_r dataset is much less than in D_n.

Once these concepts have been defined, we must take into account that the standard evaluation measures assume that the relevance distribution given by the $\Phi(Y)$ function is uniform, thus they are not sensitive to the imbalance defined by $\Phi(Y)$. Therefore, we need to consequently define appropriate evaluation metrics so that they are sensitive to this imbalance. In the following, we present the current status of the existing evaluation measures for continuous output domains, i.e., for regression.

3.2. Evaluation metrics

As already mentioned, standard evaluation measures focus on the most common instances, so it is necessary to develop new evaluation metrics in imbalanced domains in order to compare models according to user preferences (Krawczyk, 2016).

In addition, such metrics cannot only be used to evaluate the models, but can also be used to guide the learning of these models so that solutions can be generated that are *cost sensitive* and capable of adapting a penalty for the degree of relevance assigned to the D_r set.

In the specialized literature, scarce attention has been paid to the development of evaluation metrics for regression in imbalanced domains. The most common measure used for the evaluation of regression models is MSE:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

where y is the known desired output, $\hat{y}$ is the output obtained, and n is the size of the dataset.

This measure is not adequate because it assumes a uniform distribution of relevance in the output variable. The initial approximations proposed to solve this problem consist in the introduction of weights in the prediction, which establish a cost, such as the lost function LIN-LIN (Granger, 1969):

$$L(y_i - \hat{y}_i) = \begin{cases} a * |y_i - \hat{y}_i| & if \quad (y_i - \hat{y}_i) > 0 \\ b * |y_i - \hat{y}_i| & if \quad (y_i - \hat{y}_i) \leq 0 \\ 0 & if \quad y_i = \hat{y}_i \end{cases} \tag{5}$$

where a and b are parameters selected by the user, i.e. if $\frac{a}{b} = 2$ that means that the loss associated with a positive error is twice the loss associated with a negative error of the same magnitude. The main problem with this type of solution is that it can only distinguish between over- and under-predictions, while the imbalance problem occurs in a specific range of continuous values.

Further proposals that try to alleviate this problem have emerged, such as the adaptation of the notion of the *ROC* curves that are based on asymmetric loss present in regression problems, where the overestimates and underestimates have different costs. One of these proposals is *ROC space for regression* (*RROC*) (Hernández-Orallo, 2013), where this curve is defined as the representation of the total overestimation on the X-axis, and the total underestimation on the Y-axis. From this new proposition the metric of the area under the *RROC* curve emerges, which is equivalent to the error variance (but only distinguishes between over- and under-predictions). In order to solve this problem within the ROC metrics, new metrics have arisen such as *Regression Error Characteristic* (*REC curve*) (Bi & Bennett, 2003), which is mainly based on the use of the cumulative distribution function of the error obtained. In this way, the error obtained in the prediction is modeled by a lost function $L(\hat{y}, y)$, and the relevance by the function $\Phi(Y)$.

Leaving aside the metrics based on the idea of ROC curves, new propositions based on the concept of utility have arisen that try to establish a relationship between the error obtained by the prediction, and the concept of relevance, which is non-uniformly distributed over the domain of the output variable. Within this previously defined framework, the *Recall*, *Precision* and *F-measure* approach can be applied for some metrics (Torgo & Ribeiro, 2009).

Torgo et al. in Torgo and Ribeiro (2009) defines the function of relevance $\Phi(Y)$, based on the concept of extreme Y values. This relevance function is established in the following sigmoid function:

$$\Phi(Y) = \frac{1}{1 + e^{-s*(Y-c)}} \tag{6}$$

where c is the center of the sigmoid, that is, the value where $\Phi(Y) = 0.5$ and s is the shape of the sigmoid.

In addition, they differentiate between low extreme values and high extreme values, so that if both exist, $\Phi(Y)$ is defined with two different sigmoid functions.

In this way, they establish the measurements of *Recall* and *Precision* as:

$$Recall = \frac{\sum_{\Phi(y_i)\geq t_r} \alpha(\hat{y}_i, y_i) * \Phi(y_i)}{\sum_{\Phi(y_i)\geq t_r} \Phi(y_i)} \tag{7}$$

where $\alpha(\hat{y}_i, y_i)$ is defined as an indicator function $I()$, given as 1 if its argument is true and 0 otherwise, as it is based on checking the precision of the error of a prediction with a lost function (L), where, in order for it to be considered an admissible error, they focus on a threshold t_L that is dependent on the domain of the output variable: where I() is the indicator function given as 1 if its argument is true and 0 otherwise,

$$\alpha(\hat{y}_i, y_i) = I(L(\hat{y}_i, y_i) \leq t_L) \tag{8}$$

Thus, like the concept of *Recall* in classification ($\frac{TP}{TP+FN}$ where TP is *True Positives* and *FN* is *False Negatives*), they establish a connection

between those examples that have an acceptable error (well classified in classification problems) and their relevance so that they further penalize those relevant bad predictions.

Similar to the concept of Precision in classification ($\frac{TP}{TP+FP}$ where FP are *False Positives*), *Precision* focuses on the relevance of the prediction. Therefore, the concept of *Precision* that they propose in Torgo and Ribeiro (2009) is as follows:

$$Precision = \frac{\sum_{\Phi(\hat{y}_i)\geq t_r} \alpha(\hat{y}_i, y_i) * \Phi(\hat{y}_i)}{\sum_{\Phi(\hat{y}_i)\geq t_r} \Phi(\hat{y}_i)} \tag{9}$$

Once these two metrics have been defined, the researchers (Torgo & Ribeiro, 2009) propose the metric *F-Measure* to be the harmonic mean between *Precision* and *Recall* as follows:

$$F = \frac{(\beta^2 + 1) * Precision * Recall}{\beta^2 * Precision + Recall} \tag{10}$$

where $0 \leq \beta \leq 1$ controls the relative importance assigned by the user to *Recall* and *Precision*, so a value of $\beta = 1$ means equal importance.

This combination of metrics allows models to be compared in a single score. In this contribution, one of the performance measures considered is the *F1* metric, i.e. F-measure with $\beta = 1$, in order to generate and/or evaluate the models learned in this paper considering the data imbalance.

4. Algorithms applied to the problem

Once the possible evaluation metrics for the treatment of regression problems in imbalanced continuous domains have been defined, new specific strategies need to be established to solve this problem. In this section, we briefly describe two state-of-the-art evolutionary algorithms based on fuzzy rules, as they are tested in combination with preprocessing techniques and then further extended for active imbalanced regression. The two algorithms are as follows: a linguistic algorithm called FSMOGFS^e+TUN^e (Alcalá et al., 2011) and an approximate algorithm called METSK-HD^e (Gacto et al., 2014). Both are used to learn fuzzy rules, ranked from more to less descriptive (linguistic vs. free fuzzy variables).

As of yet, no study has evaluated the effectiveness of these (not purely approximate) methods with preprocessing solutions such as *SMOGN* (Branco et al., 2017). It has only been studied in purely approximate algorithms such as *Multivariate Adaptive Regression Splines*, *Support Vector Machines*, *Random Forests* and *Neural Networks* (Murphey et al., 2004). For this reason, we think it could be of interest to study the behavior of the aforementioned preprocessing method (SMOGN) toward these types of techniques in order to learn fuzzy rules, which unlike the methods already studied, would enable us to achieve greater readability. In addition, we will also compare the effectiveness of *SMOGN* when applied to the two rule-based algorithms, versus the effectiveness obtained when only using stratified partitioning of the data, in order to see if we really achieve a significant improvement when *SMOGN* is applied to these methods.

The reason that this contribution has been conducted is that in many predictive modeling domains, e.g. medicine, there is a need not only to understand the predictions, but also the reasons behind them and the deductions that have led to them. Applying FRBSs to imbalanced regression problems is new in the literature and will allow users who face these problems to be able to check/read the reasoning of the prediction made and understand its effectiveness. A current real-life example of the importance of the possibility to understand/read the model is the rise of the Internet of Things (IoT) and artificial intelligence. In many scientific fields, understanding the behavior of the obtained predictors is vital, and as explained in the scientific journal *Nature* (Castelvecchi, 2016), in many of today's models, knowledge is integrated into the model, rather than into us.

Now, let us very briefly describe the algorithms we will use to study the effectiveness of the imbalance treatment techniques for regression problems.

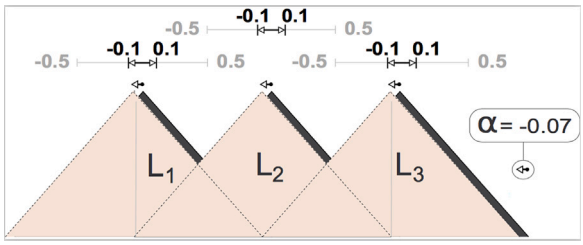

Fig. 3. Lateral displacement in $[-0.1, 0.1]$ of the whole linguistic partition $S = \{s_0, s_1, s_2, s_3, s_4\}$ (Alcalá et al., 2007).

4.1. FSMOGFS^e+TUN^e

This method is a fuzzy evolutionary algorithm based on linguistic rules designed to deal with regression problems when the number of variables and/or instances becomes too high. The method, called *Fast and Scalable Multi-Objective Genetic Fuzzy System* (FSMOGFS^e+TUN^e) (Alcalá et al., 2011), consists of two well-defined separate stages: A first stage of learning, followed by a second stage for tuning the membership function parameters and rule selection.

4.1.1. First stage

Let us define the concepts of Data Base (DB), Rule Base (RB) and Knowledge Base (KB) in the context of FRBSs. A DB is composed of the sematic concepts (only for linguistic models), and the corresponding parameters defining the membership functions giving meaning to them (both linguistic and approximate models). An RB is composed of a set of rules, i.e., the implication relationships between input variables and the outcome considering the existing membership functions. Finally, a KB is composed of a DB together with the corresponding RB, i.e., the whole FRBS model.

This first stage learns the embedded genetic DB and uses the lateral tuning (Alcalá et al., 2007) of the fuzzy partitions (Fig. 3); i.e. a reduced lateral displacement of the fuzzy partitions that helps reduce the search space. This form of representation allows us to consider a single parameter per label (slight displacements to the left/right of the original membership functions) instead of the three parameters that are usually considered per label in classical tuning.

It is based on an improved Multi-Objective Evolutionary Algorithm (MOEA) (enhanced version of SPEA2 (Zitzler et al., 2001)) that was designed to efficiently select the relevant variables while learning the appropriate granularity (number of linguistic labels) and their associated displacement parameters. The well-known *Wang and Mendel* (Wang & Mendel, 1992) algorithm is applied to obtain the complete KB, which includes a rule cropping mechanism to avoid unnecessary computing time when the obtained rule base is too large for each of the DBs provided by MOEA. The two objectives to be minimized by the MOEA are the mean square error divided by 2 ($MSE_{/2}$) and the number of rules.

The chromosome includes a double coding scheme; the granularity C_G (the number of labels of each variable) together with the displacements C_L of the corresponding linguistic partition, both jointly encoding a whole DB.

The crossover operator depends on the part of the chromosome to which it is applied. The Parent Centric BLX (PCBLX) crossover operator (Lozano et al., 2004), which is based on BLX-α, is applied to the C_L part (real coding). A crossover point is randomly generated and the classical crossover operator is applied to this point for the C_G part (integer coding). The mutation operator is applied with a certain probability and consists of either decreasing the granularity in a gene g selected at random by 1, or randomly determining a higher granularity with the same probability in C_G. The same gene g is also changed at random in C_L.

This algorithm also includes an incest prevention mechanism and a restarting mechanism in order to promote a better balance between exploration and exploitation (Eshelman, 1991).

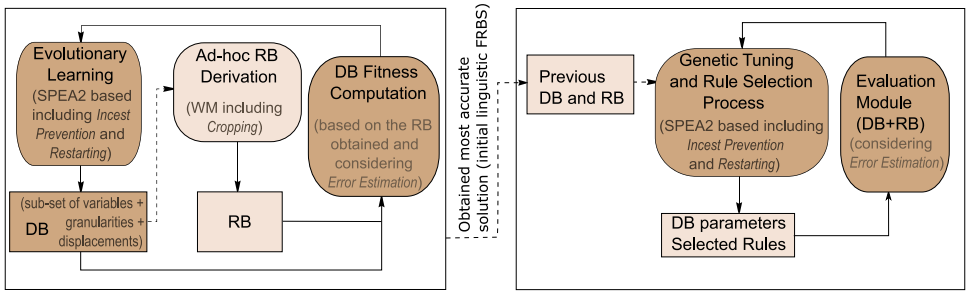

Fig. 4. FSMOGFS^e+TUN^e algorithm scheme.

4.1.2. Second stage

The second post-processing stage consists of an evolutionary process that is applied to the final KB obtained in the first stage (the most accurate one). It performs a classical tuning of the three points of the DB's membership functions together with rule selection, which present positive synergy when considered within the same process.

This stage is based on $SPEA2_{E/E}$ (Zitzler et al., 2001), which is an efficient MOEA designed particularly for the learning/tuning of systems based on fuzzy rules. Similar to the evolutionary process carried out in the first phase, incest prevention is also included along with a modified parent selection scheme that progressively focuses on the most accurate Pareto front solutions, so that it combines exploration with periods of exploitation from each restart point to the next point. The $MSE_{/2}$ and the number of rules are used as objectives in this second phase for the evaluation of the chromosomes.

The diagram in Fig. 4 summarizes the processes performed during the first stage of learning and the second stage of tuning and rule selection. As we can see, the method's first phase is used for learning, as well as implementing the mechanisms mentioned above to make learning faster. Once the RB and the DB are obtained, they are used as input for the second tuning and rule selection phase. Please, see Alcalá et al. (2011) for a more detailed explanation of the complete FSMOGFS^e+TUN^e method, as well as the appropriate justification of its main components.

4.2. METSK-HD^e

This algorithm called METSK-HD^e (Gacto et al., 2014) consists of a two-stage method for accurate fuzzy modeling in high-dimensionality regression problems to learn accurate Takagi–Sugeno–Kang (TSK) fuzzy systems (Sugeno & Kang, 1988; Takagi & Sugeno, 1985). In the first stage, a learning of the evolutionary DB together with an embedded learning of the RB within the same process is carried out. The second stage is a post-processing stage, in which rule selection and tuning of the membership functions are performed to a further refinement of the solutions learned. In addition, the second stage incorporates an efficient Kalman filter (Kalman, 1960) to learn the coefficients of the consequent polynomial function of the TSK rules. Both stages include mechanisms that significantly improve the model accuracy and ensure a rapid convergence in large-scale and high-dimensional regression datasets. In the following subsections, we will briefly introduce the main parts of the algorithm.

4.2.1. First stage

This first stage consists of using an MOEA to learn the initial KB. This method is based on the integrated genetic learning of the DB (including variables, granularities and slight lateral displacements of fuzzy partitions) which allows the algorithm to learn quickly upfront, reducing the dimensionality and making use of effective mechanisms that ensure rapid convergence in domains of regression with high dimensionality.

The integrated genetic learning of the DB is based on an evolutionary process that encodes and evolves different DBs, which are evaluated by applying a fast inductive rule generation method and calculating the system error based on the fuzzy rules thus obtained. The components required to implement this stage of the algorithm are explained in depth below:

1. A double coding scheme (C_G y C_L) to represent both the granularity and lateral displacement parameters (two linguistic tuples Alcalá et al., 2007 in Fig. 3).
2. The three objectives of this algorithm are:
   - Minimize the $MSE_{/2}$ (Main objective).
   - Minimize the number of rules.
   - Maximize the coverage degree of the examples or instances.
3. Initial population: The initial population is divided into two subsets of individuals. In the first set, each chromosome has the same number of labels for all the input variables in the system. To provide diversity in the set of labels, these solutions have been generated by considering all possible combinations of the input variables, i.e., from 2 labels to 7 labels. In addition, for each of these combinations, two copies with different values are included in the lateral displacements part. Finally, in the second subset, the population is filled with random solutions.
4. Crossover and mutation operators are the same as those used for the FSMOGFS^e+TUN^e method and are explained in Section 4.1.1.
5. Finally, in order to prevent incest, the corresponding concept from CHC (Eshelman, 1991) is used to maintain population diversity and avoid premature convergence. With respect to the stop condition, the algorithm ends when the maximum number of evaluations is reached.

4.2.2. Second stage

Once the KB obtained in the first stage has been generated, a post-processing stage is carried out. An MOEA is applied to perform tuning of the membership functions and rule selection, which will help to significantly improve the accuracy of the model. This stage of the algorithm considers the same three objectives presented above. However, since they are performing a rule selection, to ensure the full coverage of the training examples they apply a penalty to the $MSE_{/2}$ value if any rule does not cover all examples in the training set. In this case, once they calculate the $MSE_{/2}$ associated with this unwanted solution, they add the $MSE_{/2}$ from the initial solution as a penalty. This ensures that the most accurate solution obtained through evolution always covers all training examples.

The diagram in Fig. 5 summarizes the processes performed during the first stage of learning and the second stage of tuning and rule selection. As we can see, the method's first phase is used for learning, as well as implementing the mechanisms mentioned above to make learning

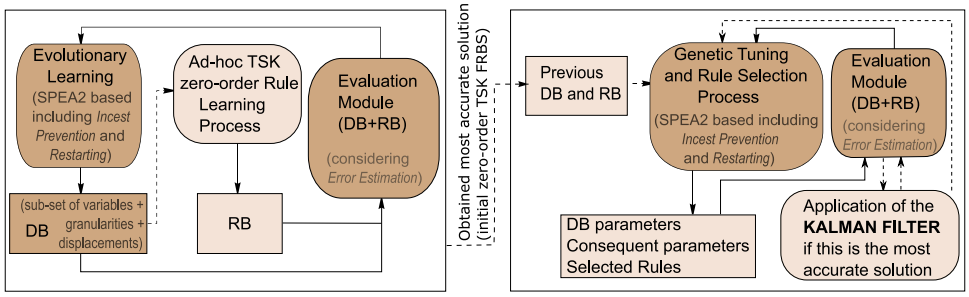

Fig. 5. METSK-HD^e algorithm scheme.

faster. Once the RB and the DB are obtained, they are used as input for the second tuning and rule selection phase, where the corresponding genetic algorithm together with the *Kalman* filter are applied. Please, see Gacto et al. (2014) for a more detailed explanation of the complete METSK-HD^e method, as well as the appropriate justification of its main components.

5. Fuzzy rule-based systems for imbalanced regression

The main problem our model has with classic prediction algorithms for regression is that, when working on imbalanced sets, it does not represent knowledge concerning the importance that the user gives to the set of relevant values D_r, since, as this set is a minority, the algorithm is focused on optimizing predictions for the examples that are mostly found in D_n.

On the other hand, we have already seen that the current effective measures in the state of the art consist of passive preprocessing measures, which generally improve the performance of the models in imbalanced problems. However, they do not directly attack the behavior of the models in the learning phase performed on the dataset, and thus only the training set is modified without affecting the learning model. This, therefore, will always mean that performance will be worse than when defining specific learning models that are optimized to work in imbalanced regression problems.

In this section, we are going to propose two active algorithms based on the adaptation of METSK-HD^e and FSMOGFS^e+TUN^e to treat imbalanced regression. As we have seen, both proposals are evolutionary algorithms, with two clearly defined phases: a first phase of learning a KB followed by a second phase of tuning and rule selection to improve performance. Both evolutionary algorithms are guided by objective functions in each phase. In the case of the linguistic algorithm (FSMOGFS^e+TUN^e) there are two objectives ($MSE_{/2}$ and number of rules) and in the case of the approximate algorithm (METSK-HD^e), the three objectives ($MSE_{/2}$, number of rules and maximize the degree of coverage of the examples). We refer these two new methods of Evolutionary Fuzzy Rule-Based Systems for imbalanced regression problems as *Linguistic-IR* (Linguistic model for Imbalanced Regression), which is based on FSMOGFS^e+TUN^e (Alcalá et al., 2011), and *TSK-IR* (Approximate TSK model for Imbalanced Regression), which is based on METSK-HD^e (Gacto et al., 2014).

Unlike other techniques, the flexibility of evolutionary algorithms enables them to adapt to specific problems in a versatile manner (e.g., monotonic regression Alcalá-Fdez et al., 2017). Based on this feature, we propose a solution that allows algorithms to be guided toward results that take imbalance into account by means of a more specific evaluation function with more information.

Formally, the representation of imbalance in a regression dataset uses the concept of relevance (Eq. (1)). Taking this concept into account, the idea is to define a multi-objective function that establishes a balance between performance in the subsets D_r and D_n respectively, fulfilling the following conditions:

1. The main objective: maximize the performance obtained in the set of relevant data D_r (guided by the relevance function), without unbalancing the performance that we obtain in the rest of the data D_n (guided by $MSE_{/2}$).
2. Minimize the number of rules required for the construction of the model.
3. Maximize the coverage degree of the examples (in the case of the TSK-IR method).

5.1. Metric used to guide the algorithm

In order to solve the described problem, and respect the conditions that balance the performance, a new metric has been established that enables the algorithms to be guided more efficiently: Weighted Mean Square Error ($MSE^W_{/2}$), according to the relevance of the individual. To do this, we will divide the weighted quadratic error by the sum of the number of non-relevant elements $|D_n|$ plus the number of relevant elements weighted by their relevance, and add the constant a (experimentally this value is set at 5, since it is the one that obtained the best results), defining the denominator (d_w) for the weighted average as follows:

$$d_w = \left(\left(\sum_{i=1}^{|D_r|} a + \Phi(y_i) \right) + |D_n| \right) \tag{11}$$

As such, once the denominator d_w is defined for the weighted average, $MSE^W_{/2}$ for a given individual/model j is defined by a function in parts, where, depending on the relevance of the instances, MSE or relevance-weighted MSE will be applied:

$$MSE^W_{/2\ j} = \frac{1}{d_w * 2} * \begin{cases} \sum_{i=1}^{|D_r|} (F(x^i) - y^i)^2 + (F(x^i) - y^i)^2 * (a + \Phi(y_i)) \\ \quad \text{if } \Phi(y_i) > t_r \\ \sum_{i=1}^{|D_n|} (F(x^i) - y^i)^2 \\ \quad \text{if } \Phi(y_i) \leq t_r \end{cases} \tag{12}$$

In this equation, we multiply d_w by 2 following the same philosophy as the $MSE_{/2}$ equation. In a more intelligible way, we can define $MSE^W_{/2}$ based on the membership of the instance $\langle x_i, y_i \rangle$ to D_r and D_n sets, applying an error weighting by means of its relevance in the first case and MSE in the second:

$$MSE^W_{/2\ j} = \frac{1}{d_w * 2} * \begin{cases} MSE + MSE * (a + \Phi(y_i)) & \text{if } \langle x_i, y_i \rangle \in D_r \\ MSE & \text{if } \langle x_i, y_i \rangle \in D_n \end{cases} \tag{13}$$

Our evolutionary algorithms are implemented using the defined metrics.

5.2. Linguistic-IR

This section presents the modifications made to the learning model of our algorithm so that it adapts effectively to imbalanced problems. This proposal is based on the FSMOGFS^e+TUN^e (Alcalá et al., 2011) algorithm (explained previously in Section 4.1) and it is structured in two different stages: a first stage where the KB (variables, granularities and number of rules) is learnt and a second phase of tuning and rule selection that uses a multiobjective evolutionary process that seeks to minimize the error and the number of rules obtained.

In this active proposal, the objective of the evolutionary algorithm in the second phase has been modified in order to minimize the new metric proposed in the previous section ($MSE^W_{/2}$). As a result, we get an algorithm that is based on two phases, and whose objectives differ from the objectives of the FSMOGFS^e+TUN^e algorithm:

1. The goal of the first phase is to classically optimize the $MSE_{/2}$ without taking the dataset imbalance into account.
2. In the second phase of tuning and rule selection, once the KB is obtained, the rules are optimized to adjust the $MSE^W_{/2}$ as much as possible using the relevance as measure of information for the objective function, so that performance is increased and focused on the data belonging to D_r.

The use of the two measures $MSE_{/2}$ and $MSE^W_{/2}$ in the two phases, respectively, enables the model to increase the diversity of the solutions obtained in such a way that the learning is not solely focused on the relevant data, but rather the search space is balanced by first obtaining a more general KB and then by being adjusted to the performance in the relevant dataset in the second phase.

5.3. TSK-IR

The second algorithm developed is *TSK-IR*, which is based on METSK-HD^e (Gacto et al., 2014). As we show in Section 4.2, it is a two-stage algorithm for fuzzy modeling in high dimensionality regression problems using Approximate TSK Fuzzy Rule-Based Systems. The structure of TSK-IR is also divided into two clearly defined stages, both employing evolutionary models: a first stage of learning the KB, and a second stage of processing in order to refine the learned solutions by incorporating a Kalman filter, which also includes rule selection and the genetic tuning of the DB.

Similar to that used in Linguistic-IR, the idea is to modify the evaluation function in this second stage of tuning and rule selection so that it is carried out by weighting those predictions considered relevant, or, belonging to D_r.

1. The first stage, where genetic learning is mainly guided by $MSE_{/2}$, in addition to minimizing the number of rules and maximizing the degree of coverage.
2. The second stage of tuning and rule selection is modified so that the objective is to get the number of rules, the degree of coverage and, the $MSE^W_{/2}$ metric (instead of $MSE_{/2}$) to focus on refinement in the relevant set, thus improving its performance.

As such, and similarly to *Linguistic-IR*, we diversified the search space to obtain a balance between $MSE_{/2}$ and $F1$ (defined in 3.2), that is, first we performed an exploration phase, which was followed by an exploitation phase in order to focus on performance in D_r.

6. Experimental study

Once the proposals for the treatment of imbalance in regression problems and the corresponding evaluation metrics have been analyzed, our main objective is to carry out an experimental study (supported by statistical tests) of the effectiveness of the current methods for the treatment of imbalance in regression problems on the two fuzzy evolutionary algorithms based on competitive fuzzy rules. This experiment, framed within the imbalanced regression, involves new difficulties that should be taken into account in order to ensure the effectiveness and validity of our results:

- **Datasets**: The choice of the dataset is not trivial, as it is necessary to analyze each dataset independently in order to guarantee the presence of imbalance in the output variable to be predicted.
- **Selection of metrics**: It is important to select evaluation metrics that enable the performance of the models on the imbalanced data to be evaluated. It is also important to use aggregated classical measures to observe the evolution at both a specific level (imbalanced metrics) and at a global level (classic metrics).
- **Models used**: Explanation of the methods used in the experiment, presentation of the stratification process and the parameters used by the SMOGN (Branco et al., 2017) method.
- **Evaluation**: We need to establish a way to evaluate the results of the performance metrics on the datasets and ensure that the partitions that the model will train and test are independent. Once we obtain the results, we will carry out a statistical study to support our conclusions.

In this section, we will deal with these points describing each one of them in detail. After this, we will draw conclusions from the results analysis obtained, and observe their effectiveness. We will also present our reasoning and make comparisons of the values obtained in those algorithms that rely on statistical analysis.

6.1. Datasets

To evaluate the effectiveness of the imbalance treatment strategies on the algorithms, an experimental framework was designed consisting of 32 regression datasets from different domains. In order to select the datasets, a study must to be carried out on each one independently in order to see whether they contain imbalance or not. Furthermore, the selection must guarantee the diversity of datasets in both size and dimensionality. Focusing on the imbalance, we are going to use the previously mentioned concept of relevance, so that we can identify the two existing subsets in each dataset: the relevant set (D_r in Eq. (2)), and the non-relevant or "normal" set (D_n in Eq. (3)).

To do so, we will use the method proposed by *Ribeiro* (Torgo & Ribeiro, 2007). In this method, the quartiles and the interquartile range of the distribution of the output variable in the dataset are used to assign greater relevance to the extreme high and low values of the output variable. Therefore, the considered datasets will have either one extreme (in the high or low values of the output variable) or two extremes (high and low extremes of the output variable).

It is also necessary to establish the relevance threshold t_r, therefore, the dataset is bisected into a relevant and non-relevant set, using instances that have a relevance threshold value higher or lower than t_r, respectively. Once this is established, we can now effectively measure the imbalance in our datasets, so that, given the sizes of the relevant set versus the total number of sets, we can also calculate the imbalance percentage of the dataset by dividing the number of instances of the D_r subset by the total number of instances.

The selected datasets come from "*Irvine Machine Learning Repository*" (UCI) (Dua & Graff, 2017), "*Knowledge Extraction based on Evolutionary Learning*" (KEEL) (Triguero et al., 2017), "*Dataset Collections of Weka*" (WEKA) (Witten et al., 2016), "*Delve Datasets*" (DELVE) (Akujuobi & Zhang, 2017), "*Luis Torgo Repository*" (LTR) (Torgo, 2023) and from "*Journal of Statistics Education Data Archive*" (JSE) (JSE, 2023). These repositories are high quality, certified and supported by many other studies. Table 1 summarizes the main characteristics of the datasets, where *Name* is the short name, *Acro* is the acronym, *Var* is the number of input variables, *Instances* is the number of examples or instances, and *% Imb* is the imbalance percentage, which is calculated

Table 1
Datasets considered in this contribution: First, those with less than 5000 instances and, second, those with more than 5000.

Name	Acro	Var	Instances	% Imb
Abalone	ABA	8	4177	16.25
Airfoil Self-Noise	AIR	5	1503	4.12
Anacalt	ANA	7	4052	20.60
Baseball	BAS	16	337	4.15
Boston housing	BOS	13	506	12.84
Concrete Compressive Strength	CON	8	1030	5.33
Machine CPU	CPU	6	209	16.26
Electrical Length	ELE1	2	495	8.08
Electrical Maintenance	ELE2	4	1056	10.41
Facebook Measures	FAC	17	495	12.32
Forest Fires	FOR	12	517	15.28
Laser generated	LAS	4	993	8.55
Mortgage	MOR	15	1049	10.10
AutoPrice	PRI	15	159	8.17
Quake	QUA	3	2178	5.41
Servo	SER	4	167	20.35
Strikes	STR	6	625	12.96
Treasury	TRE	15	1049	10.39
Triazines	TRI	58	186	10.75
Yacht Hydrodynamics	YH	6	308	16.88
Add10	ADD	10	9792	4.47
Ailerons	AIL	40	13750	6.95
Bank32	BK32	32	8192	14.03
Bank8	BK8	8	8192	6.40
Computer activity	CA	21	8192	8.70
California	CAL	8	20640	8.82
Cpusmall	CPS	12	8192	8.70
Deltaail	DAIL	5	7129	6.31
Deltaelv	DELV	6	9517	11.65
House16	H16	16	22784	12.61
House8	H8	8	22784	12.61
Puma32	PUM	32	8192	10.46

using the following equation: 100 × (number_of_instances with $\Phi(Y) > 0.8$)/number_of_instances. As we can see, the selected datasets vary from a minimum of 4.12% imbalance to a maximum of 20.6%. In addition, the dimensionality and size of the datasets have also been selected so that we can see the effectiveness of the methods with both a small and large number of instances, and with higher and lower dimensionality.

In addition, we will carry out the evaluation in the 32 datasets described independently: first, without any preprocessing to treat the imbalance; then, with a developed stratification strategy; and after that, we will apply SMOGN to treat the imbalance. As such, the method is compared with a stratification strategy in order to see if it performs better than simpler methods. Moreover, we can see how the application of SMOGN as a preprocessing technique affects the process and whether the results improve.

6.2. Selection of metrics

To correctly carry out the study, it is necessary to establish what metrics need to be obtained in order to efficiently evaluate the two algorithms. To do so, we will use a metric that allows us to evaluate the treatment of imbalance using our method: **F1**, and on the other hand, a classical one such as **MSE** that allows us to maintain an equilibrium and observe if the algorithm does not focus solely and exclusively on the relevant subset of data D_r with a total contempt of D_n. Furthermore, as we have seen in the *Preliminaries* section, METSK-HD^e and FSMOGFS^e+TUN^e are methods that deal with high-dimensionality problems correctly, so we will also take the number of rules and variables that the model uses into account. The selected performance metrics are:

- **F1**: We will use the most recent metrics proposed by Torgo (Torgo & Ribeiro, 2009) for the evaluation of the models in the imbalanced set. We will use $F-\beta$ (Eq. (10)), which is based on the concept of classification, using the harmonic mean of *Precision* and *Recall* that has been adapted for regression. In our case, we will use $\beta = 1$, thus we must weigh the measures (*Recall* and *Precision*) in the same way. In the package implemented in *R* (Branco, 2019), specific metrics are developed for imbalanced datasets (Torgo & Ribeiro, 2009) as well as for the SMOGN method (Branco et al., 2017). We will use it to calculate the $F1$ values for our algorithms.
- **Mean Square Error**: MSE is used to avoid focusing solely on the imbalance so that the behavior of the complete dataset can be observed.
- **Number of rules**: Both algorithms used in the study try to minimize the number of rules, so it would also be useful to know the number of rules obtained.

These metrics have been selected with the aim of observing the evolution and performance of the models in the most relevant subsets (D_r); we can also obtain a global view of the values in general by means of classic measures, such as MSE. This is required because it is possible that our model might only focus on the D_r subset, while neglecting the rest of the set. As such, we would get a very small error in F1 but a very high MSE, which would not solve any of the problems. Besides, the number of rules enables us to quickly see the complexity of the model obtained proportionally, that is, the more rules obtained, the greater the complexity, while the smaller the number, the less complexity we obtain.

Additionally, we consider two well-known and strong global accuracy metrics to compare our two active proposals with the previous passive state-of-the-art, SMOGN. They are:

- **R**^{2}: The coefficient of determination helps to assess the accuracy of a model. It is an absolute value that takes a value of 1 when the model gets to perfectly explains all the data, and 0 when none of the data variability is explained by the model. Thus, the higher R^2, the better the accuracy.
- **Mean Absolute Error**: MAE is the average of the absolute differences between the actual output values and their corresponding predicted ones. Since it represents an error, the lower MAE, the better accuracy.

6.3. Models used

To carry out the experimentation it is necessary to create a work ecosystem where we have previously presented algorithms, metrics and techniques to treat the imbalance. In order to do this, we will evaluate the performance of the current best preprocessing proposal, SMOGN (Branco et al., 2017) on the selected datasets. In addition, we will perform the comparison to evaluate performance in three different cases:

1. Not employing any imbalance processing, i.e., executing our two algorithms on our datasets without any specific preprocessing.
2. Using stratification in the dataset at the time of validation, so that in each partition we have equitable samples of all ranges of our dataset. In this way, we ensure that the original diversity of our dataset is maintained when evaluating.
3. Using the SMOGN imbalance treatment technique in our dataset before the model learns.

Once the experiments to be carried out have been defined, the parameters of each algorithm are shown in the following subsections.

6.3.1. Stratification process

Although the process of stratification in regression problems is widely known, there are some details, like how it is used and applied in certain problems, that are missing. Specifically, the stratification that will be used in the evaluation of the models when carrying out the corresponding data partitions consists of a discretization of the output or objective variable in a total of c cuts, c being the result of dividing the total number of examples of our dataset ($|D|$) by the number of desired partitions (n_p):

$$c = \frac{|D|}{n_p} \tag{14}$$

Once discretized, the instances belonging to each cut are randomly and proportionally distributed among the desired partitions. This is an attempt to minimize the loss of diversity incurred when partitioning the dataset during the evaluation.

6.3.2. SMOGN parameters

SMOGN makes use of the necessary parameters to set the relevance of each attribute so that it can make the distinction between D_r and D_n. The method considers three issues: The first is the parameters needed to establish the relevance function. If the user does not set the values manually, it can be calculated using the "extremes" method, based on the density function of the output variable: when using the Boxplot, a matrix is derived with the interpolation points required for the relevance function. The second is concerned with the binary loss function used in the relevance function as seen above in Eq. (5). The third is that it is necessary to set the relevance threshold starting at the point at which the separation of D_r and D_n is performed. For our case we will use a value of 0.8, so that higher values will be considered relevant, and the lower values, normal. We have chosen this value after experimentally and empirically testing different values, and 0.8 was the one with the best results.

As we can see, the issues related to the relevance function are domain-dependent, because they are mainly based on the density distribution of the output variable. SMOGN, at the time of evaluation, will be executed only on the training dataset of our two models, so the output is what the METSK-HD^e and FSMOGFS^e+TUN^e methods learn the test dataset. For this reason, the relevance issues must come only from our training set, since if we used those taken directly from the total dataset (training + test) we would be providing information to SMOGN that does not belong exclusively to the training set, and it would not be valid in the evaluation.

6.4. Evaluation

Once the datasets and the metrics have been selected and the imbalance methods have been defined, a reliable and representative method to evaluate the models needs to be defined.

Error estimation tends to be quite variable, depending on which data from the set are in the learning partition (where our models will be trained) and which are in the test partition (where our models will predict). For this reason, in order to alleviate this error overestimation as much as possible, we will use a Cross Validation (CV) mechanism with 10 partitions. As such, we will take the original data from the dataset and create ten partitions with two separate sets in each partition: a first training set, and a second validation set (independent in the 10 partitions). In addition, we will repeat the process twice in order to further reduce random bias.

As previously mentioned, in the partitioning process we will alternate between two partitioning processes: a simple random sampling (setting the random seed to 3), and using the stratification technique described above to carry out the partitioning. Furthermore, our study will be carried out twice, first without applying the imbalance treatment technique (SMOGN), and then applying it the second time.

This will give us a total of 5 files per partition (*fold*). First, we will carry out a simple cross validation using random sampling:

- *Train*: Training partition obtained with random sampling.
- *Test*: Test partition obtained with random sampling.

Second, another cross validation will be carried out, partitioning the data using the stratification strategy described above. As such, we would get three other files:

- *Train Strat.*: Training partition obtained with stratified sampling.
- *Test Strat.*: Test partition obtained with stratified sampling.
- *Train Strat. with SMOGN*: Partition *"Train Strat."* after applying *SMOGN*.

In total, there will be a total of 5 files x 10 partitions x 2 CV x 32 datasets, i.e. 3200 partition files for our experimental study. As such, we will observe the behavior of the models according to the partitioning performed, and see how it affects the complete imbalanced regression problem.

As we have already seen, F1 requires both relevance and loss function arguments, and these arguments are totally dependent on our dataset's output variable domain. As we are evaluating the models, the arguments of relevance and loss function will be calculated on all of the set without partitioning, since our model does not have all the information at the time of learning, and in order for us to be able to evaluate its performance in F1, all the real information from our problem must be used (Figs. 6 and 7). In this way, we would get the training error (TR_i) and test error (TS_i) of the corresponding partition, and validate the imbalance treatment by using F1 on the whole dataset (DS).

Fig. 6 presents the scheme for normal validations without SMOGN. In this case, the algorithms are executed directly on TR_i and TS_i, without any preprocessing, and F1 is evaluated using the whole dataset in the same way.

Fig. 7 shows the scheme for validations with SMOGN. In this case, the regression algorithms are executed on the preprocessed partition and F1 is evaluated using the whole dataset.

7. Results and discussion

In this section, we will evaluate the results obtained in our experimental framework and compare the results obtained by the two algorithms: FSMOGFS^e+TUN^e (Alcalá et al., 2011) (**Linguistic**) and METSK-HD^e (Gacto et al., 2014) (**TSK**), and the results obtained with our two new proposed techniques: Linguistic-IR and TSK-IR. In order to evaluate the behavior of the methods, we have performed statistical studies by means of Wilcoxon tests (Sheskin, 2007; Wilcoxon, 1945), establishing the equality of the methods in the null hypothesis, and we will see if the analyzed methods provide any statistical evidence.

We will obtain results from the $F1$ and MSE performance measures from the measures corresponding to the number of rules used by the algorithm (since they are rules-based algorithms) and the number of variables used by the algorithm. Additionally, for the comparison with our two new proposals we also consider R^2 and MAE as performance measures (plus $MAPE$ for the last statistical comparison).

This section is organized as follows:

- Section 7.1 compares algorithms with and without stratification.
- Section 7.2 shows the comparison made between methods considering SMOGN.
- Section 7.3 presents the comparison made with our new proposals.

7.1. Comparison between algorithms with and without stratification

In this section, we will first begin the study by comparing the use of the designed stratification explained above in our algorithms in imbalanced datasets. We execute the two algorithms by first performing 2x10CV on the original dataset without stratifying, and then on the stratified one, obtaining the following results shown in Table 2.

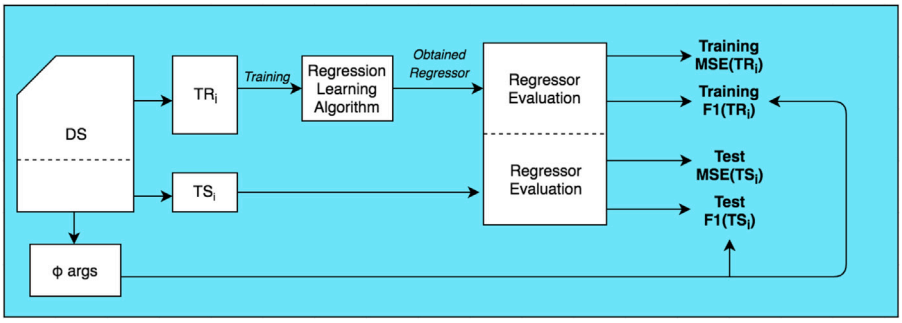

Fig. 6. Scheme for normal validations without preprocessing, using the original TR_i to build the regressor.

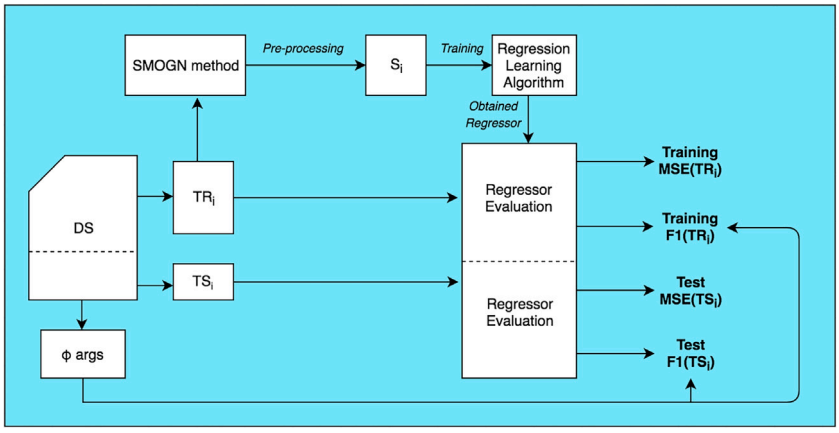

Fig. 7. Application of the SMOGN method to obtain a balanced set S_i that will be used consequently to build the regressor.

Table 2
Results comparing methods with vs without stratification.

Dataset	Linguistic								TSK							
	–			Stratification					–			Stratification				
	#R(#V)	$MSE_{f2\ tst}$	$F1_{tst}$	#R(#V)	$MSE_{f2\ tst}$	$F1_{tst}$	$DIFF_{MSE}$	$DIFF_{F1}$	#R(#V)	$MSE_{f2\ tst}$	$F1_{tst}$	#R(#V)	$MSE_{f2\ tst}$	$F1_{tst}$	$DIFF_{MSE}$	$DIFF_{F1}$
ABA	8.05(3.35)	2.480	**0.715**	9.5(3.1)	**2.472**	**0.715**	0.003	0.000	20(3.45)	**2.410**	**0.718**	21.85(3.4)	2.413	0.714	−0.001	−0.004
AIR	24.45(3.85)	**7.118**	0.165	23.8(3.85)	7.128	**0.191**	−0.001	0.026	47.9(4.75)	5.178	**0.156**	45.75(4.65)	**5.160**	0.129	0.003	−0.027
ANA	8.95(2.65)	**3.11E−03**	**0.530**	8.9(2.7)	3.30E−03	0.513	−0.056	−0.017	12.9(4.75)	1.29E−01	**0.480**	12.45(4.65)	**1.17E−01**	0.478	0.096	−0.001
BAS	14.7(6.15)	**216887**	0.280	16.3(5.9)	243270	**0.346**	−0.108	0.066	59.05(6.95)	**317448**	**0.388**	57.3(7.10)	322466	0.373	−0.016	−0.015
BOS	22.8(4.35)	9.259	0.865	23.9(4.3)	**8.404**	**0.885**	0.092	0.019	47.2(5.8)	**9.269**	0.852	49.35 (5.85)	9.307	**0.890**	−0.004	0.038
CO	16.85(3.55)	35.389	**0.777**	15.4(3.45)	**34.528**	0.773	0.024	−0.004	51.4(4.2)	**22.104**	**0.893**	49.4 (4.3)	25.363	0.859	−0.129	−0.034
CPU	16.3(4)	**1914**	0.703	15.6(3.90)	1928	**0.731**	−0.008	0.027	31.45(4.8)	**3020**	**0.688**	33.3 (4.85)	3697	0.615	−0.183	−0.073
ELE1	8.85(2)	**211214**	0.739	8.4(2)	214443	**0.746**	−0.015	0.007	14.3(2)	228610	0.723	13.55 (2)	**211702**	**0.743**	0.074	0.020
ELE2	8.3(2)	**11047**	0.978	9.4(2)	11265	**0.979**	−0.019	0.001	41.25(3.95)	4882	0.274	42.2 (3.95)	**4329**	**0.343**	0.113	0.068
FAC	4.6(2.2)	22426	0.907	4.4(2.25)	**17743**	**0.920**	0.209	0.012	5.05(2.35)	**28758**	0.932	5.05 (2.25)	29085	**0.938**	−0.011	0.007
FOR	12.58(3.6)	**2427**	**0.345**	13.3(3.65)	2736	0.332	−0.113	−0.014	33.55(4.65)	**4336**	**0.347**	37.35 (4.7)	4898	0.333	−0.115	−0.014
LAS	19.8(3.05)	**42.580**	**0.956**	18.8(3)	43.156	0.951	−0.013	−0.005	56(4)	40.281	**0.964**	55.55 (4)	**35.641**	0.963	0.115	−0.001
MOR	6.35(2.05)	**0.017**	**0.985**	6(2.2)	0.018	**0.985**	−0.037	0.000	13.05(2.25)	**0.014**	0.9867	12.65 (2.1)	0.016	**0.9872**	−0.144	0.001
PRI	23.6(5.35)	**3007102**	**0.513**	24.9(4.85)	3179709	0.446	−0.054	−0.068	48.8(8.05)	5269901	0.380	46.65 (8.35)	**4754469**	**0.436**	0.098	0.056
QUA	3.3(1.4)	1.79E−02	**0.000**	3.3(1.15)	**1.78E−02**	**0.000**	0.002	0.000	28.9(3)	**5.92E−02**	**0.056**	28.35 (2.95)	6.61E−02	0.000	−0.104	−0.056
SER	15.9(3.2)	0.191	**0.789**	17.6(3.05)	**0.168**	0.774	0.120	−0.015	65.95(4)	**0.217**	0.637	64.35 (4)	0.277	**0.646**	−0.217	0.009
STR	16.1(3.55)	**171386**	0.524	18.2(3.7)	182971	**0.556**	−0.063	0.032	45.1(5)	184951	0.546	44.5(5)	**183881**	**0.553**	0.006	0.008
TRE	7.1(2.6)	**0.041**	**0.977**	7.5(2.75)	0.047	**0.977**	−0.131	0.000	11.95(2.45)	0.043	0.974	12.75 (2.55)	**0.036**	**0.979**	0.169	0.005
TRI	27.65(9.8)	**0.012**	**0.145**	28.6(10.1)	0.014	0.116	−0.172	−0.029	76(11.3)	0.014	**0.128**	77.8 (11.8)	**0.013**	0.115	0.135	−0.013
YH	12.15(2.1)	**1.055**	**0.957**	13.6(2.05)	1.160	0.950	−0.090	−0.006	13.1(2.75)	**1.301**	0.818	12.75 (2.85)	1.302	**0.841**	−0.001	0.023
ADD	26.65(3.35)	2.715	0.155	28.5(3.45)	**2.627**	**0.321**	0.033	0.165	51.9(3.85)	1.934	0.308	55.65 (3.9)	**1.863**	**0.435**	0.037	0.127
AIL	17.9(3.75)	2.031E−08	**0.115**	17.6(3.85)	**2.029E−08**	0.109	0.001	−0.006	39.05(4.7)	1.46E−08	0.121	38.95 (5.05)	**1.43E−08**	**0.123**	0.027	0.002
BK32	4.35(3)	3.88E−03	0.688	4.3(3)	**3.85E−03**	**0.690**	0.008	0.002	15.15(3.3)	3.79E−03	0.693	21.7 (3.2)	**3.76E−03**	**0.697**	0.008	0.003
BK8	4.95(2)	7.50E−04	0.944	4.35(2.05)	**7.41E−04**	**0.947**	0.012	0.003	8.95(2)	**6.95E−04**	**0.952**	11.25 (2)	7.06E−04	0.951	−0.016	−0.001
CA	13.25(5.05)	5.349	0.582	12.1(4.9)	**5.222**	**0.590**	0.024	0.008	30.6(5)	**5.323**	0.608	27.95 (4.65)	5.391	**0.618**	−0.012	0.010
CAL	7.2(2.8)	3.02E+09	0.799	8.6(3.25)	**2.98E+09**	**0.840**	0.011	0.042	43.75(4.6)	2.79E+09	0.812	43.85 (4.95)	**2.68E+09**	**0.856**	0.039	0.044
CPS	10.35(4.6)	**6.043**	0.582	12.8(4.75)	6.075	**0.592**	−0.005	0.010	30.55(5.35)	**5.742**	0.619	33.9 (5.4)	5.781	**0.644**	−0.007	0.026
DAIL	5.5(2.55)	**1.54E−08**	0.729	5.1(2.3)	1.55E−08	**0.739**	−0.005	0.010	36.3(5)	1.36E−08	0.740	36.9 (4.9)	**1.35E−08**	**0.748**	0.008	0.009
DELV	7.4(2.6)	1.09E−06	0.634	7(2.75)	**1.08E−06**	**0.708**	0.007	0.074	39.45(3.9)	**1.02E−06**	0.641	30.55 (3.55)	1.04E−06	**0.722**	−0.017	0.081
H16	10(4.6)	9.42E+08	0.592	10.9(5.05)	**9.31E+08**	**0.595**	0.012	0.003	30.6(4.3)	**8.77E+08**	**0.611**	30.4 (4.1)	8.82E+08	0.607	−0.006	−0.004
H8	13.95(3.25)	**6.16E+08**	**0.651**	14(3.85)	6.19E+08	0.650	−0.005	−0.001	32.55(4.45)	**5.46E+08**	**0.670**	35.25 (4.5)	5.57E+08	0.666	−0.020	−0.004
PUM	15.4(2)	**2.90E−05**	0.947	14.1(2)	**2.90E−05**	**0.948**	0.000	0.001	29.5(2)	**2.90E−05**	**0.946**	28.5 (2)	**2.90E−05**	0.945	0.000	−0.001
Average	12.98 (3.45)		0.633	13.31 (3.47)		0.644			34.73 (4.34)		0.614	34.93 (4.36)		0.623		

Table 3
Wilcoxon test in order to compare methods without (R^+) and with stratification (R^-) in $MSE_{/2\ tst}$ and $F1_{tst}$.

Methods analyzed	Comparison	R^+	R^-	Hypothesis ($\alpha = 0.1$)	p-value
Linguistic without Strat. vs. Linguistic with Strat.	$MSE_{/2\ tst}$	281	184	Accepted	0.324
Linguistic without Strat. vs. Linguistic with Strat.	$F1_{tst}$	169	327	Accepted	0.124
TSK without Strat. vs. TSK with Strat.	$MSE_{/2\ tst}$	263	233	Accepted	0.776
TSK without Strat. vs. TSK with Strat.	$F1_{tst}$	222	274	Accepted	0.617

In Table 2 algorithms with and without stratification are grouped in columns, and the average of the results obtained by each algorithm in all the studied datasets are shown. For each algorithm, the first column shows the average number of both the number of rules (#R) and the used variables (#V). The second and third columns show the average $MSE_{/2}$ and F1 in the test data (Tst.) The fourth and fifth columns show the differences (*DIFF*) between the metrics obtained by the two methods analyzed (in this case with and without stratification). Finally, the last row of the table shows the global average values (Average). Moreover, the best result (in both F1 and $MSE_{/2}$) for each dataset is shown in boldface.

As we can see in this table, the mean results of F1, as well as the differences (*DIFF*) between metrics obtained without using stratification (random sampling) and those obtained using the designed stratification, are quite similar. The results stay the same for $MSE_{/2}$. We performed a statistical test to verify if there was any statistical evidence that was different or not. Table 3 presents the results of the Wilcoxon test (Sheskin, 2007; Wilcoxon, 1945) for the method with and without stratification. The results show that both methods with and without stratification are almost equivalent, or slightly better only in F1 for the stratification with the linguistic approach.

The application of stratification in the partitioning of our dataset is necessary, since, as in classification, it is possible that when partitioning random data for evaluation, and taking into account that the set D_r is small compared to D_n, D_r might not be represented in the training set or in the test set, and this would make the evaluation of our methods unrepresentative and dependent on randomness. Therefore, the use of stratification for evaluation in imbalanced problems is almost mandatory in order to preserve a minimum representation of the minority set, and is relevant to the user in the partitions. Furthermore, there is also statistical evidence that they are almost equivalent; from now on, we will consider only stratified methods in the rest of the study.

This preliminary study helps us to demonstrate that the need for a simple stratification of the dataset at the time of making the partitioning to evaluate does not improve the performance of the algorithms in imbalanced problems, suggesting that imbalance treatment techniques are required.

7.2. Comparison between methods considering SMOGN

Once the study of the behavior of a stratified model was carried out, and in view of the need to use stratification when evaluating the two algorithms, we then studied, for the first time ever, the behavior of the most current passive technique *SMOGN* (Branco et al., 2017) on two evolutionary algorithms based on fuzzy rules, *Linguistic* and *TSK*. As such, we have studied how effective it is to apply this passive preprocessing technique to rule-based algorithms so that users can choose the correct technique for treating imbalance.

Table 4 uses the same terminology as Table 2, but in this case, the comparison is made between methods using stratification with and without SMOGN. As we can see in this table, the results obtained for both algorithms using SMOGN (a passive imbalance preprocessing technique) are better for $F1$, so the *SMOGN* proposal is ranked higher than the *Linguistic* and *TSK* algorithms without imbalance treatment.

These results show that using the technique improves the performance of the algorithms in the relevant set D_r, which means that it is not neglecting the user-relevant values. On the other hand, if we look at the results obtained for $MSE_{/2}$, we see that the use of *SMOGN* produces a cost in this metric, which makes it worse than not using preprocessing.

A statistical analysis obtained using the Wilcoxon test for the method with and without SMOGN is shown in Table 5. On the one hand, looking at the F1 metric, the both methods reject the equality hypothesis (with confidence of 99.0%). It shows that algorithms that use SMOGN as a passive imbalance treatment strategy outperform the original algorithms in F1, meaning that this method focuses on the relevant set. On the other hand, we can observe that $MSE_{/2}$ worsens when using SMOGN. Both the *TSK* and *Linguistic* methods reject the equality hypothesis with a p-value much lower than 0.01. This means that with a confidence of 99.0%, we can affirm that there is statistical evidence that shows that using this passive technique to treat imbalance substantially worsens $MSE_{/2}$.

7.3. Comparison with our new proposals

Previously, in Section 7.2, we conducted a comprehensive and in-depth study of SMOGN's behavior in our two fuzzy rule-based methods, which is state of the art and has never been performed before on this type of algorithm. The final study carried out consists of the evaluation of our first active proposal, which aims to become state of the art, as compared to the current passive SMOGN proposal. Table 6 (which uses the same terminology as the previous tables but includes R^2 and MAE) shows the results obtained when using SMOGN as a preprocessing technique before using *Linguistic* and *TSK*, and the results obtained with our proposals *Linguistic-IR* and *TSK-IR* (presented in Sections 5.2 and 5.3, respectively) without using SMOGN as a preprocessing technique. The partitions are stratified in both cases.

As we can see in Figs. 8 and 9, the results obtained in F1 by our proposals without using SMOGN are slightly better to those obtained when we use SMOGN preprocessing. That is, our active proposals *Linguistic-IR* and *TSK-IR* obtain similar results without preprocessing, and even higher than average results (Linguistic-IR method) than those obtained using SMOGN (the best current passive preprocessing technique); therefore, our two active proposals are efficient in treating imbalanced regression. In addition, we can observe that our active proposals obtain much better results in global accuracy than those obtained when using SMOGN. Both proposals perform much better in MSE, obtaining a balance in performance on the relevant dataset D_r and the remaining data D_n. Similar behavior can be observed in Table 6 for R^2 and MAE, where both proposals outperform their respective proposals when SMOGN is used in most of the datasets.

Table 7 shows the results of the Wilcoxon statistical test to support the following statements. Here, we also consider the $MAPE$ metric (Mean Absolute Percentage Error, i.e., the mean absolute percentage difference between the actual and the predicted value). MAPE is considered only with statistical comparative purposes in order to keep the previous table readable.

In $F1$, the null hypothesis is accepted, which allows us to affirm that there is statistical evidence to prove that they perform similarly, as both algorithms-IR obtain better ranking (R^+, R^-) in both cases with respect to the methods with SMOGN. In fact, in the case of Linguistic-IR, we have a confidence of almost 85% to reject the equality hypothesis in our favor, so that we could even say that we outperform the use of SMOGN in F1 with the linguistic approach. On the other hand, we

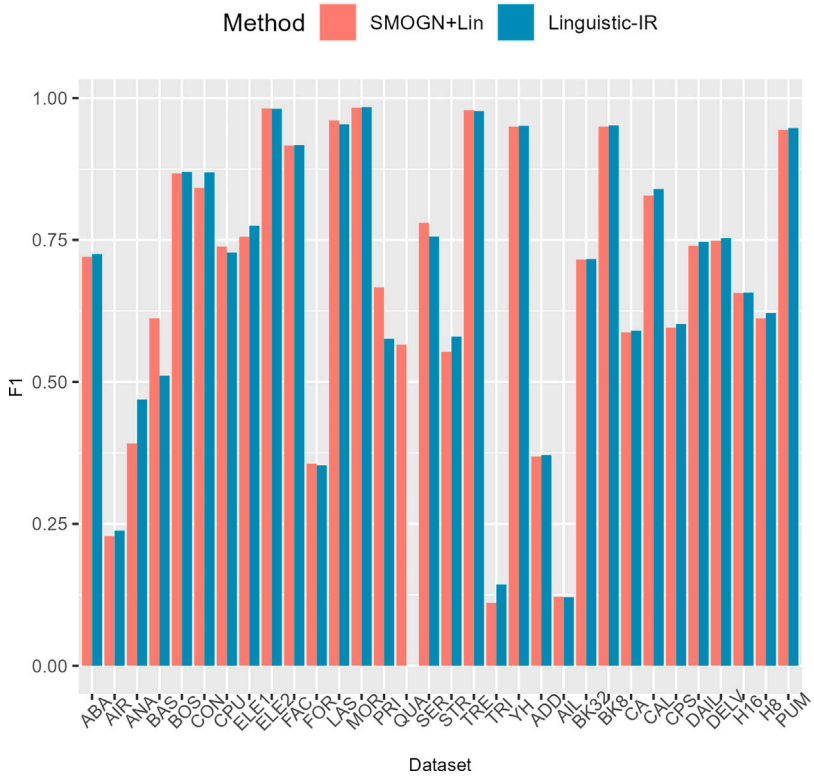

Fig. 8. SMOGN+Linguistic vs. Linguistic-IR F1.

Table 4
Results comparing methods on the stratified 10-folds with vs without SMOGN.

Dataset	Linguistic with stratification							TSK with stratification						
	Linguistic			SMOGN+Linguistic				TSK			SMOGN+TSK			
	#R(#V)	$MSE_{/2\ tst}$	$F1_{tst}$	#R(#V)	$MSE_{/2\ tst}$	$F1_{tst}$	$DIFF_{MSE}$	$DIFF_{F1}$	#R(#V)	$MSE_{/2\ tst}$	$F1_{tst}$	#R(#V)	$MSE_{/2\ tst}$	$F1_{tst}$	$DIFF_{MSE}$	$DIFF_{F1}$
ABA	9.5 (3.1)	**2.472**	0.715	8 (2.35)	3.750	**0.720**	−0.341	0.005	21.85 (3.4)	**2.413**	0.714	22.85 (3.25)	3.633	**0.727**	−0.336	0.013
AIR	23.8 (3.85)	**7.128**	0.191	24.05 (3.45)	8.043	**0.229**	−0.114	0.038	45.75 (4.65)	**5.160**	0.129	44.75 (4.70)	7.393	**0.270**	−0.302	0.141
ANA	8.9 (2.7)	**3.30E−03**	**0.513**	11.4 (2.8)	3.79E−03	0.392	−0.128	−0.121	12.45 (4.65)	**1.17E−01**	**0.478**	27.10 (4.65)	2.00E−01	0.363	−0.419	−0.116
BAS	16.3 (5.9)	**243270**	0.346	20.75 (5.7)	343971	**0.612**	−0.293	0.267	57.3 (7.10)	**322466**	0.373	20.75 (5.70)	343971	**0.542**	−0.063	0.170
BOS	23.9 (4.3)	**8.404**	**0.885**	21.25 (4.1)	11.170	0.867	−0.248	−0.017	49.35 (5.85)	9.307	**0.890**	46.40 (5.20)	**9.267**	0.887	0.004	−0.003
CO	15.4 (3.45)	**34.528**	0.773	17.9 (3.7)	66.001	**0.842**	−0.477	0.069	49.4 (4.3)	**25.363**	0.859	43.35 (4.45)	50.536	**0.865**	−0.498	0.006
CPU	15.6 (3.90)	**1928**	0.731	17.15 (3.75)	1957	**0.739**	−0.015	0.008	33.3 (4.85)	3697	0.615	35.90 (4.85)	**3354**	**0.625**	0.093	0.009
ELE1	8.4 (2)	**214443**	0.746	12.15 (2)	301952	**0.756**	−0.290	0.010	13.55 (2)	**211702**	0.743	16.20 (2.00)	275725	**0.748**	−0.232	0.005
ELE2	9.4 (2)	11265	0.979	11.75 (2)	**11045**	**0.982**	0.020	0.003	42.2 (3.95)	**4329**	0.343	42.10 (3.90)	7567	**0.936**	−0.428	0.594
FAC	4.4 (2.25)	**17743**	**0.920**	5 (2.3)	23523	0.916	−0.246	−0.004	5.05 (2.25)	**29085**	**0.938**	5.15 (2.25)	33235	0.929	−0.125	−0.009
FOR	13.3 (3.65)	**2736**	0.332	12.95 (3.95)	4129	**0.356**	−0.338	0.024	37.35 (4.7)	**4898**	**0.333**	40.95 (4.75)	7192	0.309	−0.319	−0.024
LAS	18.8 (3)	**43.156**	0.951	21.3 (3.05)	56.675	**0.961**	−0.239	0.010	55.55 (4)	**35.641**	**0.963**	52.40 (4.00)	63.409	0.961	−0.438	−0.002
MOR	6 (2.2)	**0.018**	**0.985**	6.65 (2.3)	0.022	0.983	−0.187	−0.001	12.65 (2.1)	0.016	**0.987**	12.95 (2.15)	**0.015**	**0.987**	0.096	0.000
PRI	24.9 (4.85)	**3179709**	0.446	23.75 (5.55)	3496691	**0.667**	−0.091	0.221	46.65 (8.35)	**4754469**	0.436	33.30 (6.70)	5501181	**0.437**	−0.136	0.001
QUA	3.3 (1.15)	**1.78E−02**	0.000	3.45 (1.45)	4.47E−02	**0.566**	−0.601	0.566	28.35 (2.95)	**6.61E−02**	0.000	41.50 (3.00)	6.87E−02	**0.587**	−0.037	0.587
SER	17.6 (3.05)	**0.168**	0.774	15 (3)	0.240	**0.780**	−0.298	0.007	64.35 (4)	**0.277**	**0.646**	76.65 (4.00)	0.574	0.586	−0.517	−0.059
STR	18.2 (3.7)	**182971**	**0.556**	21.1 (3.55)	292993	0.553	−0.376	−0.003	44.5 (5)	**183881**	0.553	49.00 (4.85)	352234	**0.583**	−0.478	0.030
TRE	7.5 (2.75)	0.047	0.977	7.1 (2.95)	**0.042**	**0.979**	0.098	0.002	12.75 (2.55)	**0.036**	**0.979**	11.85 (2.20)	0.041	0.972	−0.121	−0.007
TRI	28.6 (10.1)	**0.014**	**0.116**	29.40 (10.05)	0.015	0.111	−0.057	−0.005	77.8 (11.8)	**0.013**	0.115	60.50 (11.85)	0.020	**0.166**	−0.360	0.051
YH	13.6 (2.05)	**1.160**	**0.950**	7.85 (2.15)	1.257	**0.950**	−0.078	0.000	12.75 (2.85)	1.302	**0.841**	16.35 (3.00)	**1.143**	0.833	0.122	−0.007
ADD	28.5 (3.45)	**2.627**	0.321	22 (3.6)	5.536	**0.369**	−0.526	0.048	55.65 (3.9)	**1.863**	0.435	49.7 (3)	4.058	**0.479**	−0.541	0.044
AIL	17.6 (3.85)	**2.029E−08**	0.109	11.89 (4.85)	2.72E−08	**0.122**	−0.255	0.013	38.95 (5.05)	**1.43E−08**	0.123	36.8 (5)	2.13E−08	**0.126**	−0.331	0.003
BK32	4.3 (3)	**3.85E−03**	0.690	4.7 (3.1)	5.53E−03	**0.716**	−0.304	0.026	21.7 (3.2)	**3.76E−03**	0.697	16.8 (3.15)	5.47E−03	**0.713**	−0.312	0.016
BK8	4.35 (2.05)	**7.41E−04**	0.947	11.25 (2)	7.93E−04	**0.950**	−0.066	0.003	11.25 (2)	**7.06E−04**	**0.951**	21.05 (2.05)	7.32E−04	0.948	−0.036	−0.002
CA	12.1 (4.9)	**5.222**	**0.590**	13.15 (5.2)	6.036	0.588	−0.135	−0.002	27.95 (4.65)	**5.391**	**0.618**	30.6 (5.35)	5.666	0.613	−0.049	−0.005
CAL	8.6 (3.25)	**2.98E+09**	**0.840**	8.65 (2.6)	4.72E+09	0.828	−0.368	−0.012	43.85 (4.95)	**2.68E+09**	**0.856**	26.75 (4.2)	4.25E+09	0.840	−0.368	−0.016
CPS	12.8 (4.75)	**6.075**	0.592	12.6 (5.2)	7.008	**0.596**	−0.133	0.003	33.9 (5.4)	**5.781**	**0.644**	40.65 (4.8)	6.581	0.610	−0.122	−0.035
DAIL	5.1 (2.3)	**1.55E−08**	0.739	12.95 (2.9)	3.52E−08	**0.740**	−0.560	0.001	36.9 (4.9)	**1.35E−08**	0.748	31.95 (4.6)	2.90E−08	**0.750**	−0.533	0.002
DELV	7 (2.75)	**1.08E−06**	0.708	12.9 (2.9)	1.91E−06	**0.749**	−0.433	0.040	30.55 (3.55)	**1.04E−06**	0.722	31.6 (3)	1.86E−06	**0.753**	−0.439	0.030
H16	10.9 (5.05)	**9.31E+08**	0.595	13.2 (4.75)	1.42E+09	**0.612**	−0.342	0.017	30.4 (4.1)	**8.82E+08**	0.607	13.2 (4.75)	1.42E+09	**0.612**	−0.376	0.005
H8	14 (3.85)	**6.19E+08**	0.650	14.2 (3.5)	7.84E+08	**0.657**	−0.210	0.007	35.25 (4.5)	**5.57E+08**	0.666	29.05 (4.35)	7.01E+08	**0.669**	−0.206	0.003
PUM	14.1 (2)	**2.90E−05**	**0.948**	14.15 (2.05)	3.60E−05	0.944	−0.194	−0.004	28.5 (2)	**2.90E−05**	**0.945**	29.6 (2.1)	3.30E−05	0.944	−0.121	0.000
Average	13.31 (3.47)		0.644	14.05 (3.53)		0.682			34.93 (4.36)		0.623	33.05 (4.18)		0.668		

Table 5
Wilcoxon test to compare methods with stratification without (R^+) and with SMOGN (R^-) in $MSE_{/2\ tst}$ and $F1_{tst}$.

Methods analyzed	Comparison	R^+	R^-	Hypothesis ($\alpha = 0.1$)	p-value
Linguistic vs. SMOGN+Linguistic	$MSE_{/2\ tst}$	520	8	Rejected	1.16e−08
Linguistic vs. SMOGN+Linguistic	$F1_{tst}$	119	409	Rejected	5,71E−03
TSK vs. SMOGN+TSK	$MSE_{/2\ tst}$	506	22	Rejected	2,50E−07
TSK vs. SMOGN+TSK	$F1_{tst}$	52	476	Rejected	1,80E−05

Table 6
Results comparing methods with SMOGN vs new proposals (without using SMOGN) on the stratified 10-folds.

Dataset	Linguistic with stratification													
	SMOGN+Linguistic					Linguistic-IR								
	#R(#V)	$MSE_{/2\ tst}$	$F1_{tst}$	R^2	MAE	#R(#V)	$MSE_{/2\ tst}$	$F1_{tst}$	R^2	MAE	$DIFF_{MSE}$	$DIFF_{F1}$	$DIFF_{R^2}$	$DIFF_{MAE}$
ABA	8 (2.35)	3.750	0.720	0.274	2.182	9.45 (3.1)	**3.507**	**0.725**	**0.323**	**2.100**	0.065	0.005	0.048	0.038
AIR	24.05 (3.45)	8.043	0.229	0.660	3.062	23.5 (3.85)	**7.470**	**0.238**	**0.684**	**2.964**	0.071	0.009	0.025	0.032
ANA	11.4 (2.8)	**3.79E−03**	0.392	**0.974**	**0.029**	7.85 (2.7)	3.96E−03	**0.469**	0.973	0.035	−0.044	0.077	−0.001	−0.185
BAS	20.75 (5.7)	343971	**0.612**	0.553	532.20	15.9 (5.9)	**268351**	0.511	**0.644**	**475.628**	0.220	−0.101	0.091	0.106
BOS	21.25 (4.1)	11.170	0.867	0.722	3.192	21.95 (4.3)	**10.245**	**0.870**	**0.749**	**3.003**	0.083	0.003	0.027	0.059
CO	17.9 (3.7)	66.001	0.842	0.523	8.874	13.75 (3.45)	**39.990**	**0.869**	**0.709**	**6.974**	0.394	0.027	0.187	0.214
CPU	17.15 (3.75)	**1957**	**0.739**	0.760	31.974	15.3 (3.9)	2103	0.728	**0.762**	**31.908**	−0.070	−0.011	0.002	0.002
ELE1	12.15 (2)	301952	0.756	0.522	519.22	8.05 (2)	**251168**	**0.775**	**0.597**	**472**	0.168	0.019	0.075	0.091
ELE2	11.75 (2)	**11045**	**0.982**	**0.9934**	**109.70**	9.45 (2)	11168	0.981	0.9933	110	−0.011	−0.001	0.000	−0.006
FAC	5 (2.3)	23523	0.916	0.902	25.347	4.3 (2.25)	**18367**	**0.917**	**0.941**	**23.970**	0.219	0.001	0.038	0.054
FOR	12.95 (3.95)	4129	**0.356**	**0.001**	38.584	11.25 (3.65)	**3035**	0.353	0.000	**35.988**	0.265	−0.003	−0.001	0.067
LAS	21.3 (3.05)	56.675	**0.961**	0.948	6.075	18.25 (3)	**49.687**	0.954	**0.954**	**5.989**	0.123	−0.007	0.006	0.014
MOR	6.65 (2.3)	0.022	0.983	0.995	0.146	5.85 (2.2)	**0.019**	**0.984**	**0.996**	**0.142**	0.108	0.001	0.000	0.029
PRI	23.75 (5.55)	3496691	**0.667**	0.754	1825	22.85 (4.85)	**3021332**	0.576	**0.808**	**1555**	0.136	−0.091	0.054	0.148
QUA	3.45 (1.45)	4.47E−02	**0.566**	**0.000**	0.267	2.7 (1.15)	**2.23E−02**	0.000	**0.000**	**0.183**	0.500	−0.566	0.000	0.316
SER	15 (3)	0.240	**0.780**	0.768	0.350	14.4 (3.05)	**0.188**	0.756	**0.828**	**0.342**	0.217	−0.024	0.060	0.023
STR	21.1 (3.55)	292993	0.553	**0.012**	450	15.65 (3.7)	**246721**	**0.580**	0.010	**420**	0.158	0.027	−0.002	0.068
TRE	7.1 (2.95)	**0.042**	**0.979**	**0.993**	**0.183**	7.95 (2.75)	0.045	0.977	0.992	0.188	−0.063	−0.002	0.000	−0.025
TRI	29.40 (10.05)	**0.015**	0.111	**0.147**	**0.126**	26.1 (10.1)	0.016	**0.143**	0.069	0.131	−0.069	0.032	−0.078	−0.037
YH	7.85 (2.15)	1.257	0.950	0.9882	1.005	13.7 (2.05)	**1.252**	**0.951**	**0.9884**	**0.998**	0.004	0.001	0.000	0.007
ADD	22 (3.6)	5.536	0.369	0.551	2.614	26.7 (3.45)	**2.868**	**0.371**	**0.767**	**1.920**	0.482	0.002	0.216	0.266
AIL	11.89 (4.85)	2.72E−08	**0.122**	0.673	1.73E−04	17.25 (3.85)	**2.25E−08**	0.121	**0.730**	**1.59E−04**	0.175	−0.001	0.056	0.081
BK32	4.7 (3.1)	5.53E−03	0.7158	0.250	7.96E−02	4.1 (3)	**5.49E−03**	**0.7164**	**0.254**	**7.95E−02**	0.006	0.001	0.004	0.001
BK8	11.25 (2)	7.93E−04	0.950	0.931	0.0291	4.75 (2.05)	**7.82E−04**	**0.952**	**0.932**	**0.0290**	0.014	0.002	0.001	0.002
CA	13.15 (5.2)	6.036	0.588	0.964	2.545	12.55 (4.9)	**5.545**	**0.590**	**0.967**	**2.460**	0.081	0.003	0.003	0.033
CAL	8.65 (2.6)	4.72E+09	0.828	0.291	78530	7.65 (3.25)	**3.76E+09**	**0.840**	**0.434**	**69210**	0.203	0.011	0.144	0.119
CPS	12.6 (5.2)	7.008	0.596	0.958	2.777	11.8 (4.75)	**6.858**	**0.602**	**0.959**	**2.717**	0.021	0.006	0.001	0.022
DAIL	12.95 (2.9)	3.52E−08	0.740	0.228	2.10E−04	5.15 (2.3)	**2.00E−08**	**0.747**	**0.563**	**1.49E−04**	0.433	0.007	0.335	0.290
DELV	12.9 (2.9)	1.91E−06	0.749	0.321	1.59E−03	6.85 (2.75)	**1.40E−06**	**0.753**	**0.501**	**1.32E−03**	0.265	0.004	0.181	0.168
H16	13.2 (4.75)	1.42E+09	0.612	0.038	37721	9.35 (5.05)	**1.26E+09**	**0.622**	**0.100**	**35172**	0.109	0.009	0.062	0.068
H8	14.2 (3.5)	7.84E+08	0.6567	0.437	**26466**	13.95 (3.85)	**7.69E+08**	**0.6572**	**0.446**	26578	0.018	0.000	0.008	−0.004
PUM	14.15 (2.05)	3.60E−05	0.944	0.922	6.69E−03	14.2 (2)	**3.00E−05**	**0.947**	**0.935**	**6.15E−03**	0.167	0.003	0.013	0.081
Average	14.05 (3.53)		0.682	0.595		12.58 (3.47)		0.665	0.644					
Dataset	TSK with stratification													
	SMOGN+TSK					TSK-IR								
ABA	22.85 (3.25)	3.633	0.727	0.313	2.113	19.6 (3.35)	**3.219**	**0.730**	**0.378**	**1.977**	0.114	0.003	0.066	0.064
AIR	44.75 (4.7)	7.393	**0.270**	**0.901**	2.573	53.6 (4.95)	**5.149**	0.148	0.782	**2.427**	0.303	−0.122	−0.119	0.057
ANA	27.1 (4.65)	0.200	0.363	0.019	0.571	24.1 (3.2)	**0.004**	**0.739**	**0.971**	**0.030**	0.979	0.376	0.952	0.947
BAS	20.75 (5.7)	343971	**0.542**	**0.553**	532	57.05 (6.95)	**367365**	0.193	0.530	**525**	−0.064	−0.349	−0.022	0.013
BOS	46.4 (5.2)	9.267	**0.887**	**0.787**	2.876	53.5 (5.6)	**9.155**	0.837	0.774	**2.738**	0.012	−0.051	−0.013	0.048
CO	43.35 (4.45)	50.536	**0.865**	0.654	7.194	47.2 (4)	**26.743**	0.402	**0.807**	**5.505**	0.471	−0.463	0.153	0.235
CPU	35.90 (4.85)	**3354**	**0.625**	**0.662**	**34.670**	31.9 (4.95)	3474	0.561	0.615	36.083	−0.035	−0.064	−0.047	−0.039
ELE1	16.2 (2)	275725	**0.748**	0.570	485	13.35 (2)	**249739**	0.720	**0.603**	**462**	0.094	−0.028	0.033	0.048
ELE2	42.1 (3.9)	7567	0.936	0.996	87.526	33.55 (3.7)	**4363**	**0.943**	**0.997**	**69.591**	0.424	0.007	0.002	0.205
FAC	5.15 (2.25)	33235	0.929	0.845	29.462	4.9 (2.2)	**29719**	**0.931**	**0.905**	**23.056**	0.106	0.002	0.060	0.217
FOR	40.95 (4.75)	7192	0.309	0.000	44.534	41.8 (4.45)	**5589**	**0.341**	**0.009**	**35.414**	0.223	0.032	0.009	0.205
LAS	52.40 (4)	63.409	**0.961**	0.942	4.261	56.4 (4)	**44.888**	0.913	**0.960**	**3.890**	0.292	−0.048	0.019	0.087
MOR	12.95 (2.15)	0.015	**0.987**	0.9970	0.127	16.25 (2.3)	**0.013**	0.986	**0.9972**	**0.116**	0.120	−0.001	0.000	0.090
PRI	33.3 (6.7)	**5501181**	**0.437**	**0.693**	**1996**	75.4 (8.35)	7750072	0.330	0.531	2212	−0.290	−0.106	−0.162	−0.098
QUA	41.5 (3)	0.069	**0.587**	**0.309**	0.263	11.6 (2.75)	**0.022**	0.147	0.000	**0.179**	0.675	−0.439	−0.309	0.320
SER	76.65 (4)	0.574	**0.586**	0.515	0.546	72.65 (4)	**0.428**	0.566	**0.613**	**0.437**	0.255	−0.020	0.098	0.199
STR	49 (4.85)	352234	0.583	0.014	381	42.2 (4)	**233178**	**0.586**	**0.029**	**343**	0.338	0.003	0.014	0.099
TRE	11.85 (2.2)	0.041	0.972	0.993	0.180	10.3 (2)	**0.035**	**0.979**	**0.994**	**0.169**	0.134	0.007	0.001	0.064
TRI	60.5 (11.85)	0.020	0.166	**0.198**	0.147	84.85 (11.5)	**0.012**	**0.218**	0.143	**0.110**	0.389	0.052	−0.055	0.247
YH	16.35 (3)	**1.143**	0.833	**0.989**	**0.784**	77 (4.9)	8.423	**0.872**	0.904	1.788	−0.864	0.039	−0.085	−0.561
ADD	49.7 (3)	4.058	0.479	0.672	2.285	55.3 (3.7)	**2.103**	**0.499**	**0.830**	**1.658**	0.482	0.020	0.157	0.274
AIL	36.8 (5)	2.13E−08	0.126	0.744	1.53E−04	46.65 (4.45)	**1.64E−08**	**0.130**	**0.802**	**1.33E−04**	0.233	0.005	0.058	0.131
BK32	16.8 (3.15)	5.47E−03	**0.713**	0.258	0.077	24.25 (3.3)	**5.36E−03**	0.712	**0.273**	**0.075**	0.020	−0.001	0.015	0.019
BK8	21.05 (2.05)	7.32E−04	0.948	0.936	**0.027**	9.35 (2.05)	**7.31E−04**	**0.952**	**0.937**	0.028	0.001	0.003	0.000	−0.021
CA	30.6 (5.35)	**5.666**	0.613	**0.967**	**2.386**	30.55 (4.85)	5.799	**0.623**	**0.967**	2.401	−0.023	0.009	0.000	−0.006
CAL	26.75 (4.2)	4.25E+09	0.840	0.362	73023	49.3 (4.6)	**2.30E+09**	**0.866**	**0.654**	**50182**	0.458	0.026	0.292	0.313
CPS	40.65 (4.8)	**6.581**	0.610	**0.962**	**2.589**	27.7 (4.75)	7.045	**0.617**	0.960	2.650	−0.066	0.008	−0.002	−0.023
DAIL	31.95 (4.6)	2.90E−08	0.750	0.364	1.81E−04	29.05 (4.9)	**1.69E−08**	**0.763**	**0.631**	**1.30E−04**	0.417	0.013	0.267	0.282
DELV	31.6 (3)	1.86E−06	0.753	0.339	0.002	32.75 (3.75)	**1.33E−06**	**0.758**	**0.528**	**0.001**	0.285	0.005	0.189	0.180
H16	13.2 (4.75)	1.42E+09	0.612	0.038	37721	27.25 (3.9)	**1.23E+09**	**0.622**	**0.130**	**33576**	0.133	0.009	0.092	0.110
H8	29.05 (4.35)	7.01E+08	0.669	0.495	24040	33.55 (4.55)	**6.86E+08**	**0.678**	**0.505**	**23837**	0.022	0.008	0.011	0.008
PUM	29.6 (2.1)	3.30E−05	0.944	0.928	6.44E−03	26.45 (2)	**2.90E−05**	**0.950**	**0.937**	**6.03E−03**	0.121	0.005	0.009	0.062
Average	33.05 (4.18)		0.668	0.594		38.1 (4.25)		0.635	0.647					

Table 7
Wilcoxon test to compare our new proposals (R^+) vs methods with SMOGN (R^-) in $MSE_{/2\ tst}$, $F1_{tst}$, R^2, MAE and $MAPE$.

Methods analyzed	Comparison	R^+	R^-	Hypothesis ($\alpha = 0.1$)	p-value
Linguistic-IR vs. SMOGN+Linguistic	$MSE_{/2\ tst}$	486	42	Rejected	5.25E−06
Linguistic-IR vs. SMOGN+Linguistic	$F1_{tst}$	341	187	Accepted	0.154
Linguistic-IR vs. SMOGN+Linguistic	R^2	445	51	Rejected	1.18E−04
Linguistic-IR vs. SMOGN+Linguistic	MAE	465	63	Rejected	6.03E−05
Linguistic-IR vs. SMOGN+Linguistic	$MAPE$	422	106	Rejected	2.39E−03
TSK-IR vs. SMOGN+TSK	$MSE_{/2\ tst}$	446	82	Rejected	3.65E−04
TSK-IR vs. SMOGN+TSK	$F1_{tst}$	268	260	Accepted	0.949
TSK-IR vs. SMOGN+TSK	R^2	339	157	Rejected	0.076
TSK-IR vs. SMOGN+TSK	MAE	461	67	Rejected	9.05E−05
TSK-IR vs. SMOGN+TSK	$MAPE$	423	105	Rejected	2.23E−03

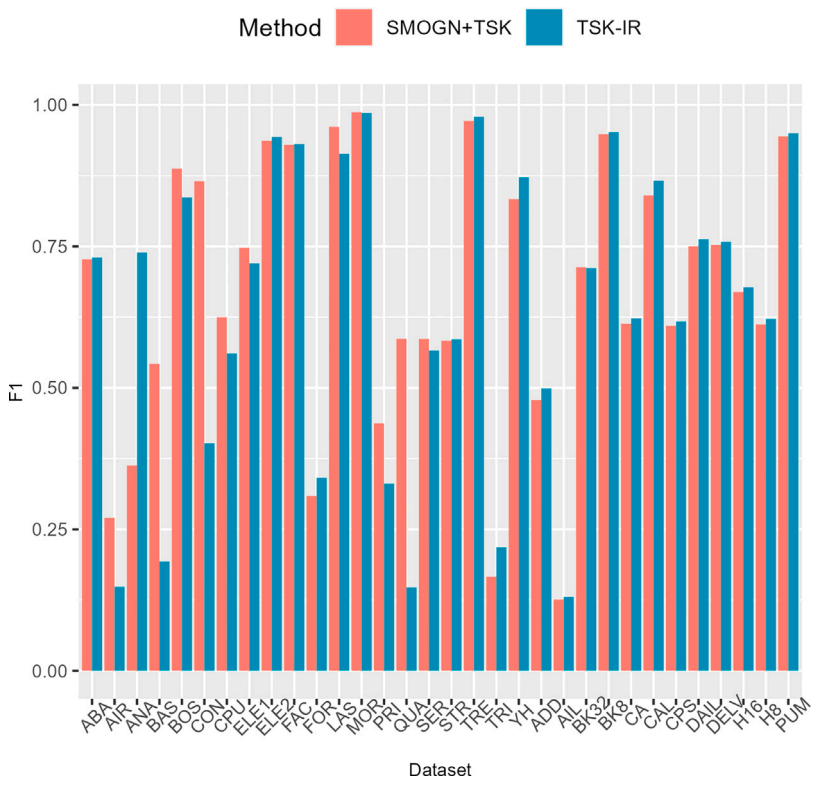

Fig. 9. SMOGN+TSK vs. TSK-IR F1.

can see that in MSE, R^2, MAE and $MAPE$ the null hypothesis is rejected in both cases (Linguistic and TSK) with 99.99% confidence (but in one with 90.0%), which means that there is clear statistical evidences showing that our proposals greatly improve the methods with SMOGN in all the accuracy metrics, without preprocessing, and perform equally (or slightly better) with statistical evidence in $F1$ (our proposals outperform the methods with SMOGN in the ranking). This means that *Linguistic-IR* and *TSK-IR* perform well with the imbalance metric, demonstrating that our algorithms are actively adapted to work in imbalanced sets, regardless of their origin. In addition, unlike SMOGN, they obtain better results in MSE, R^2, MAE and $MAPE$ as compared to the current state of the art; they obtain a balance between the performance in D_n and D_r, so that a much better MSE, R^2, MAE and $MAPE$ are achieved than when using SMOGN. This is also important, because although our main interest is that our algorithms perform well in the minority and relevant data subset for the user D_r, obtaining a balance in the predictions outside that subset is also of importance, since focusing only and exclusively on the relevant data is deemed useless as it disregards the performance of the non-relevant data. Graphically, Figs. 10 and 11 show how MSE is significantly improved in both proposals as compared to methods with SMOGN.

Our active proposals *Linguistic-IR* and *TSK-IR* improve the best current passive proposal: SMOGN (Branco et al., 2017) (evolution of *SMOTEReg* Torgo et al., 2013). The proposed methods produce good results in the relevant values while obtaining a balance in the performance in the rest of the dataset.

8. Conclusions and future work

This contribution is focused on the importance of addressing an issue scarcely tackled to date, imbalance in regression, in order to obtain good performance on the relevant and less represented data without impairing the overall performance of the obtained models.

In this contribution we have verified that traditional performance evaluation measures for regression problems do not accurately represent their performance in imbalanced regression problems since they focus on the most frequent instances and therefore, underestimate the underrepresented/relevant subset. Consequently, to evaluate models in imbalanced domains and compare them in accordance with user preferences, specific metrics must be used. We propose using *F1* as an imbalance metric, which enables the evaluation of the overall performance and prioritizes the predictions made by the model according to the relevance of the instance to be predicted.

We have seen that the use of *F1* in combination with classic metrics, for instance *MSE*, R^2, MAE and $MAPE$, provides us a balanced evaluation of imbalance sets, since *F1* reports on performance while considering the relevant predictions and *MSE* provides a more global view of performance. That is why improving both *F1* and *MSE* together allows us to focus on the relevant instances without compromising performance in the remaining instances.

By utilizing existing techniques, we have been able to evaluate the effectiveness of current proposals for treating imbalance in fuzzy rule-based algorithms for regression problems, both linguistically and approximately. The main conclusions are:

- There is no significant difference in the performance when using stratified and non-stratified data (*F1* and *MSE* are equivalent). Nevertheless, stratification is necessary to preserve a minimum representation of these ranges of values of output values. Otherwise, when partitioning random data for evaluation, that minority set might not be represented in the training or in the test set.
- Recent previous proposals (passive in any case) seem to statistically improve the performance on the relevant data subset, but worsen the error produced for the rest of the data. While these data are frequently represented and less significant from the point of view of imbalanced regression, they are still part of the dataset problem. In particular, the current best passive approach (SMOGN) applied to the two current fuzzy rule-based competitive algorithms improves performance on the imbalanced relevant data subsets (*F1* improves) at the cost of producing worse estimates for the remaining problematic data (*MSE's* performance deteriorates for both).

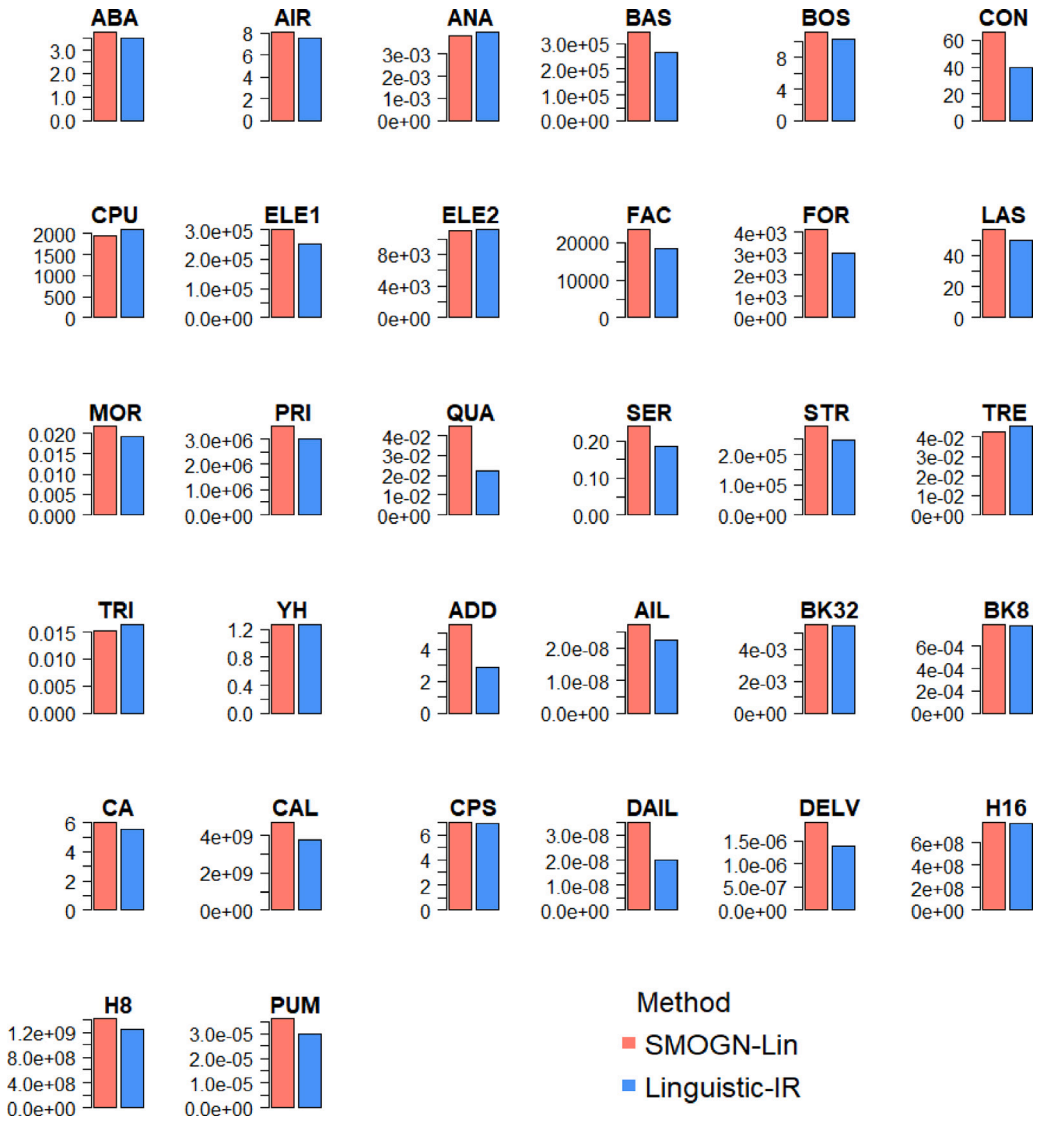

Fig. 10. Linguistic-IR vs. SMOGN+Linguistic MSE.

- We have developed two proposals, Linguistic-IR and TSK-IR, to tackle the challenge of enhancing overall performance while maintaining performance in the imbalanced dataset. Compared to the available state-of-the-art passive proposals, SMOGN, it has been demonstrated that our active proposals achieve superior performance since they statistically perform similarly, or even slightly better, in *F1* (as both proposals obtain better rankings with respect to the methods with SMOGN) and statistically outperform in *MSE*, R^2, MAE and $MAPE$.

In conclusion, our proposals were intended to address imbalance in regression and, in fact, outperform the application of the best state-of-the-art passive technique, SMOGN. These new active proposals (i.e. they do not imply preprocessing) have proven that it is not necessary to sacrifice overall performance to improve the performance of the model in the relevant data subset, since they do not only statistically equalize performance on underrepresented data, but also statistically improve overall model performance on imbalanced regression datasets. Thus, they represent a remarkable alternative in the scarcely tackled imbalanced regression problem.

These types of active techniques are particularly suitable for regression problems with imbalanced data, since we have easily demonstrated that we can even improve the overall performance of the algorithms beyond simply dealing with imbalance. Future work will be directed toward finding new, more advanced active techniques, specifically designed to address the problem of imbalanced regression. In fact, although the proposed algorithms have been shown to be highly effective, they represent only two relatively simple initial attempts at active techniques for imbalanced regression. Consequently, we believe that there is still much room for improvement, making it a promising new line of research. Three promising initial are suggested in the following:

- We have proposed $MSE^W_{/2}$ as MSE or relevance weighted MSE depending on the relevance of the data, thus integrating both concepts (global error and imbalance penalization) in a single index. Even though that MOEAs suffer as more objectives are considered, it would be very interesting to consider MSE and $F1$ separately to find different trade-offs among both desired and (somehow) contradictory properties. It would probably allow obtaining a set of Pareto-optimal solutions to be obtained that users could check for the most interesting trade-off.
- Another problem to be faced is the computational time that $F1$ or the interest function takes. Reducing the time required to compute $MSE^W_{/2}$ or $F1$ by proposing other metrics or simpler estimations would allow for much faster algorithms that could be easily applied to problems with a much larger number of instances.

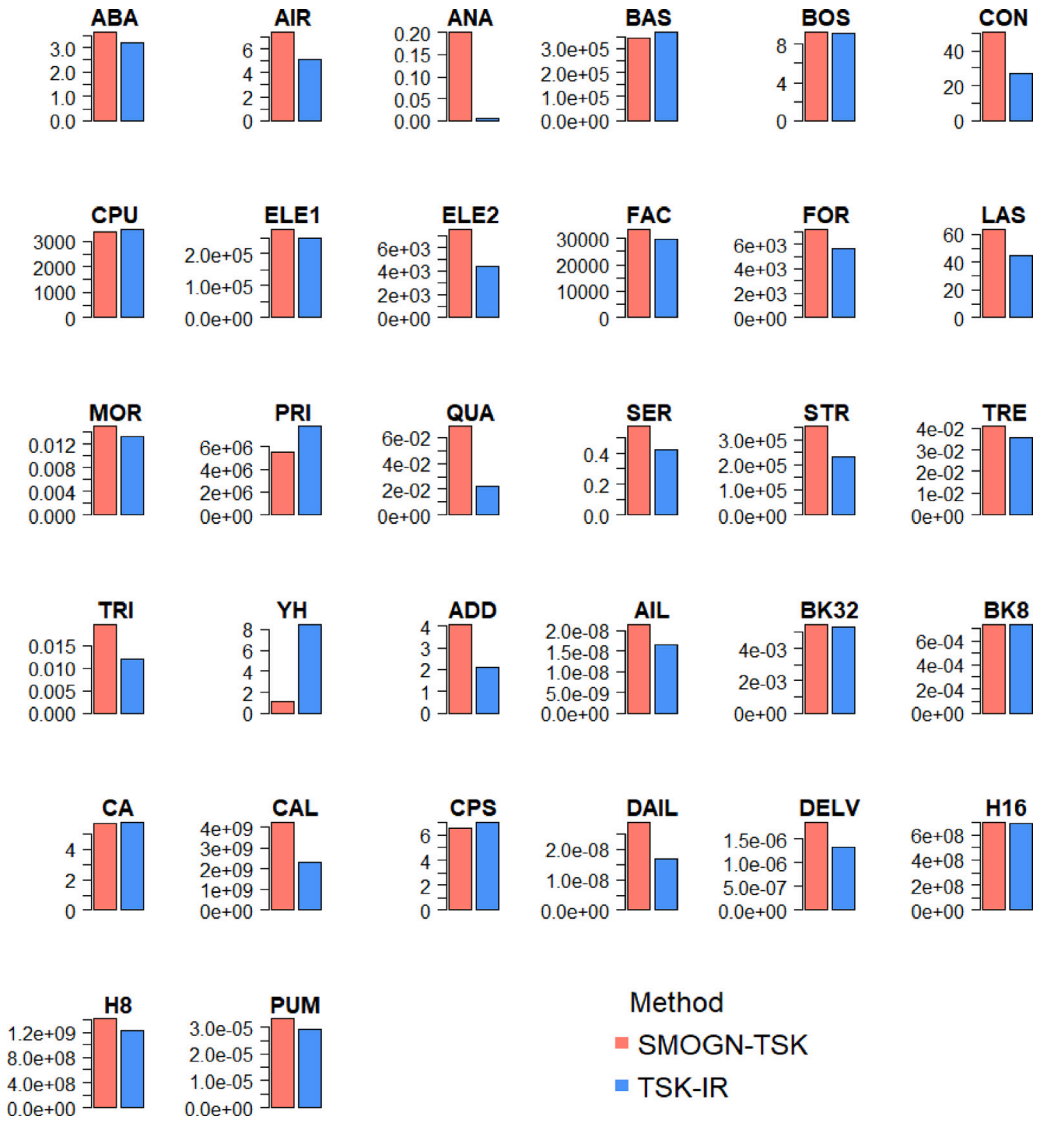

Fig. 11. TSK-IR vs. SMOGN+TSK MSE.

- Integrating linguistic interpretability metrics to the optimization process would be of interest in order to actually learn interpretable models. Some metrics and/or algorithms, such as the one in Biedma-Rdguez et al. (2022), could be considered together with data imbalance to obtain more reliable and robust models.

CRediT authorship contribution statement

María Arteaga: Investigation, Software, Formal analysis, Data curation, Validation, Writing – original draft, Writing – review & editing. **María José Gacto:** Conceptualization, Methodology, Formal analysis, Investigation, Software, Supervision, Resources, Writing – original draft, Writing – review & editing. **Marta Galende:** Resources, Writing – review & editing, Validation, Visualization, Software. **Jesús Alcalá-Fdez:** Methodology, Formal analysis, Writing – review & editing, Project administration, Funding acquisition. **Rafael Alcalá:** Conceptualization, Methodology, Formal analysis, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Datasets are available at public repositories.

Acknowledgments

This paper has been supported in part by the ERDF A way of making Europe/Health Institute Carlos III/Spanish Ministry of Science, Innovation and Universities (grant number PI20/00711), by the ERDF A way of making Europe/Regional Government of Andalusia/Ministry of Economic Transformation, Industry, Knowledge and Universities (grant numbers P18-RT-2248 and B-CTS-536-UGR20) and by the MCIN/AEI/10.13039/50110001103 (grant numbers PID2019-107793GB-I00 and PID2020-119478GB-I00). Funding for open access charge: Universidad de Granada / CBUA.

References

Ahmadi, M. (2021). A computational approach to uncovering economic growth factors. *Computational Economics*, *58*(4), 1051–1076.

Akujuobi, U., & Zhang, X. (2017). Delve: A dataset-driven scholarly search and analysis system. *SIGKDD Explorations Newsletters*, *19*(2), 36–46, http://www.cs.toronto.edu/~delve/data/datasets.html.

Alcalá, R., Alcalá-Fdez, J., & Herrera, F. (2007). A proposal for the genetic lateral tuning of linguistic fuzzy systems and its interaction with rule selection. *IEEE Transactions on Fuzzy Systems*, *15*(4), 616–635.

Alcalá, R., Gacto, M. J., & Herrera, F. (2011). A fast and scalable multi-objective genetic fuzzy system for linguistic fuzzy modeling in high-dimensional regression problems. *IEEE Transactions on Fuzzy Systems*, *19*(4), 666–681.

Alcalá-Fdez, J., Alcalá, R., González, S., Nojima, Y., & García, S. (2017). Evolutionary fuzzy rule-based methods for monotonic classification. *IEEE Transactions on Fuzzy Systems*, *25*(6), 1376–1390.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.

Bi, J., & Bennett, K. P. (2003). Regression error characteristic curves. In *Proceedings of the twentieth international conference on international conference on machine learning* (pp. 43–50). AAAI Press.

Biedma-Rdguez, C., Gacto, M. J., Anguita-Ruiz, A., Alcalá-Fdez, J., & Alcalá, R. (2022). Transparent but accurate evolutionary regression combining new linguistic fuzzy grammar and a novel interpretable linear extension. *International Journal of Fuzzy Systems*, *24*(7), 3082–3103.

Branco, P. (2019). An r package for utility-based learning. https://github.com/paobranco/UBL.

Branco, P., Torgo, L., & Ribeiro, R. P. (2015). A survey of predictive modelling under imbalanced distributions, CoRR. arXiv:1505.01658.

Branco, P., Torgo, L., & Ribeiro, R. P. (2017). SMOGN: a pre-processing approach for imbalanced regression. In L. s Torgo, B. Krawczyk, P. Branco, & N. Moniz (Eds.), *Proceedings of machine learning research: vol.74*, *Proceedings of the first international workshop on learning with imbalanced domains: theory and applications* (pp. 36–50). ECML-PKDD, Skopje, Macedonia: PMLR.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature*, *538*, 20–23.

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research (JAIR)*, *16*, 321–357.

Dua, D., & Graff, C. (2017). *UCI machine learning repository*. University of California, Irvine, School of Information and Computer Sciences, http://archive.ics.uci.edu/ml.

Eshelman, L. J. (1991). The CHC adaptive search algorithm: how to have safe search when engaging in nontraditional genetic recombination. In G. J. Rawlins (Ed.), *Foundations of genetic algorithms: vol. 1*, *FOGA* (pp. 265–283). Elsevier.

Gacto, M., Galende, M., Alcalá, R., & Herrera, F. (2014). METSK-HDe: A multiobjective evolutionary algorithm to learn accurate TSK-fuzzy systems in high-dimensional and large-scale regression problems. *Information Sciences*, *276*, 63–79.

Gacto, M. J., Soto-Hidalgo, J. M., Alcalá-Fdez, J., & Alcalá, R. (2019). Experimental study on 164 algorithms available in software tools for solving standard non-linear regression problems. *IEEE Access*, *7*, 108916–108939.

Gerritsma, J., Onnink, R., & Versluis, A. (2013). Yacht hydrodynamics. uci machine learning repository. http://dx.doi.org/10.24432/C5XG7R.

Ghorbani, N., & Korzeniowski, A. (2020). Adaptive risk hedging for call options under cox-ingersoll-ross interest rates. *Journal of Mathematical Finance*, *10*(4), 697–704.

Goli, A., Khademi-Zare, H., Tavakkoli-Moghaddam, R., Sadeghieh, A., Sasanian, M., & Kordestanizadeh, R. M. (2021). An integrated approach based on artificial intelligence and novel meta-heuristic algorithms to predict demand for dairy products: a case study. *Network. Computation in Neural Systems*, *32*(1), 1–35, PMID: 33390063.

Granger, C. W. J. (1969). Prediction with a generalized cost of error function. *Operational Research Quarterly*, *20*(2), 199–207.

Hernández-Orallo, J. (2013). ROC curves for regression. *Pattern Recognition*, *46*(12), 3395–3411.

JSE (2023). *Journal of Statistics Education Data Archive*, https://jse.amstat.org/jse_data_archive.htm. (Last Accessed 7 July 2023).

Juez-Gil, M., Arnaiz-González, Á., Rodríguez, J. J., & García-Osorio, C. (2021). Experimental evaluation of ensemble classifiers for imbalance in big data. *Applied Soft Computing*, *108*, Article 107447, URL https://www.sciencedirect.com/science/article/pii/S1568494621003707.

Kaieski, N., da Costa, C. A., da Rosa Righi, R., Lora, P. S., & Eskofier, B. (2020). Application of artificial intelligence methods in vital signs analysis of hospitalized patients: A systematic literature review. *Applied Soft Computing*, *96*, Article 106612, URL https://www.sciencedirect.com/science/article/pii/S1568494620305500.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, *82*(1), 35–45.

Korzeniowski, A., & Ghorbani, N. (2021). Put options with linear investment for hull-white interest rates. *Journal of Mathematical Finance*, *11*(1), 152–162.

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, *5*, 221–232.

Lozano, M., Herrera, F., Krasnogor, N., & Molina, D. (2004). Real-coded memetic algorithms with crossover hill-climbing. *Evolutionary Computation*, *12*(3), 273–302.

Murphey, Y., Guo, H., & Feldkamp, L. (2004). Neural learning from unbalanced data: Special issue: Engineering intelligent systems (guest editor: László monostori). *Applied Intelligence*, *21*, 117–128.

Nash, W., Sellers, T., Talbot, S., Cawthorn, A., & Ford, W. (1994). The population biology of abalone (haliotis species) in tasmania. i. blacklip abalone (h. rubra) from the north coast and islands of bass strait. *Sea fisheries division, Technical report No. 48*.

Pena, A., Patino, A., Chiclana, F., Caraffini, F., Gongora, M., Gonzalez-Ruiz, J. D., & Duque-Grisales, E. (2021). Fuzzy convolutional deep-learning model to estimate the operational risk capital using multi-source risk events. *Applied Soft Computing*, *107*, Article 107381, URL https://www.sciencedirect.com/science/article/pii/S1568494621003045.

Ranjbarzadeh, R., Bagherian Kasgari, A., Jafarzadeh Ghoushchi, S., Anari, S., Naseri, M., & Bendechache, M. (2021). Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images. *Scientific Reports*, *11*(1), 10930.

Ranjbarzadeh, R., Caputo, A., Tirkolaee, E. B., Jafarzadeh Ghoushchi, S., & Bendechache, M. (2023). Brain tumor segmentation of MRI images: A comprehensive review on the application of artificial intelligence tools. *Computers in Biology and Medicine*, *152*, Article 106405, URL https://www.sciencedirect.com/science/article/pii/S0010482522011131.

Ranjbarzadeh, R., Jafarzadeh Ghoushchi, S., Anari, S., Safavi, S., Tataei Sarshar, N., Babaee Tirkolaee, E., & Bendechache, M. (2022). A deep learning approach for robust, multi-oriented, and curved text detection. *Cognitive Computation*.

Sheskin, D. J. (2007). *Handbook of parametric and nonparametric statistical procedures* (4th ed.). Chapman & Hall/CRC.

Singh, A., & Purohit, A. (2015). A survey on methods for solving data imbalance problem for classification. *International Journal of Computer Applications*, *127*(15), 37–41.

Steininger, M., Kobs, K., Davidson, P., Krause, A., & Hotho, A. (2021). Density-based weighting for imbalanced regression. *Machine Learning*, *110*(8), 2187–2211.

Sugeno, M., & Kang, G. (1988). Structure identification of fuzzy model. *Fuzzy Sets and Systems*, *28*(1), 15–33.

Takagi, T., & Sugeno, M. (1985). Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-15*(1), 116–132.

Torgo, L. (2023). Luís torgo repository. https://www.dcc.fc.up.pt/~ltorgo/Regression/DataSets.html. (Last Accessed 7 July 2023).

Torgo, L., & Ribeiro, R. (2007). Utility-based regression. In J. N. Kok, J. Koronacki, R. Lopez de Mantaras, S. Matwin, D. Mladenič, & A. Skowron (Eds.), *Knowledge discovery in databases: PKDD 2007* (pp. 597–604). Berlin, Heidelberg: Springer Berlin Heidelberg.

Torgo, L., & Ribeiro, R. (2009). Precision and recall for regression. In J. a. Gama, V. S. Costa, A. M. Jorge, & P. B. Brazdil (Eds.), *Discovery science* (pp. 332–346). Berlin, Heidelberg: Springer Berlin Heidelberg.

Torgo, L., Ribeiro, R. P., Pfahringer, B., & Branco, P. (2013). SMOTE for regression. In L. Correia, L. P. Reis, & J. Cascalho (Eds.), *Progress in artificial intelligence* (pp. 378–389). Berlin, Heidelberg: Springer Berlin Heidelberg.

Triguero, I., Gonzalez, S., Moyano, J. M., García, S., Alcalá-Fdez, J., Luengo, J., Fernández, A., del Jesús, M., Sánchez, L., & Herrera, F. (2017). KEEL 3.0: An open source software for multi-stage analysis in data mining. *International Journal of Computational Intelligence Systems*, *10*, 1238–1249, http://www.keel.es.

Wang, L., Han, M., Li, X., Zhang, N., & Cheng, H. (2021). Review of classification methods on unbalanced data sets. *IEEE Access*, *9*, 64606–64628.

Wang, L., & Mendel, J. M. (1992). Generating fuzzy rules by learning from examples. *IEEE Transactions on Systems, Man, and Cybernetics*, *22*(6), 1414–1427.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining, fourth edition: practical machine learning tools and techniques* (4th ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., http://www.cs.waikato.ac.nz/ml/weka/.

Yan, Y., Jiang, Y., Zheng, Z., Yu, C., Zhang, Y., & Zhang, Y. (2022). LDAS: Local density-based adaptive sampling for imbalanced data classification. *Expert Systems with Applications*, *191*, Article 116213.

Zhang, C., Li, J., Zhao, Y., Li, T., Chen, Q., Zhang, X., & Qiu, W. (2021). Problem of data imbalance in building energy load prediction: Concept, influence, and solution. *Applied Energy*, *297*, Article 117139, URL https://www.sciencedirect.com/science/article/pii/S0306261921005791.

Zitzler, E., Laumanns, M., & Thiele, L. (2001). SPEA2: Improving the strength Pareto evolutionary algorithm for multiobjective optimization. In *Proc. evolutionary methods for design, optimization and control with app. to industrial problems* (pp. 95–100). Barcelona, Spain.