Proceedings of Machine Learning Research 74:36–50, 2017 Full Paper - LIDTA 2017

SMOGN: a Pre-processing Approach for Imbalanced Regression

Paula Branco PAULA.BRANCO@DCC.FC.UP.PT
Luís Torgo LTORGO@DCC.FC.UP.PT
Rita P. Ribeiro RPRIBEIRO@DCC.FC.UP.PT
LIAAD-INESC TEC DCC-FCUP, University of Porto
Porto, Portugal

Editors: Luís Torgo, Bartosz Krawczyk, Paula Branco and Nuno Moniz.

Abstract

The problem of imbalanced domains, framed within predictive tasks, is relevant in many practical applications. When dealing with imbalanced domains a performance degradation is usually observed on the most rare and relevant cases for the user. This problem has been thoroughly studied within a classification setting where the target variable is nominal. The exploration of this problem in other contexts is more recent within the research community. For regression tasks, where the target variable is continuous, only a few solutions exist. Pre-processing strategies are among the most successful proposals for tackling this problem. In this paper we propose a new pre-processing approach for dealing with imbalanced regression. Our algorithm, SMOGN, incorporates two existing proposals trying to solve problems detected in both of them. We show that SMOGN has advantages in comparison to other approaches. We also show that our method has a different impact on the learners used, displaying more advantages for Random Forest and Multivariate Adaptive Regression Splines learners.

Keywords: Imbalanced domains, Regression, Pre-processing

1. Introduction

Imbalanced domains are a relevant problem that has been studied mostly in the context of classification tasks (He and Garcia, 2009; López et al., 2013). This is an important problem with applications in a diversity of real world domains. Several proposals were put forward for dealing with imbalanced classification tasks. However, imbalanced domains also occur in other predictive contexts, such as regression tasks, data streams or time series forecasting (Branco et al., 2016b; Krawczyk, 2016). Still, the exploration of new strategies suitable for these tasks is scarce. Imbalanced domains represent a problem due to the concurrence of two factors: i) the non-uniform preferences of the user across the target variable domain; and ii) the scarce representation, in the available data, of the most relevant cases to the user. The conjugation of these two factors hinders the learners predictive performance on the cases that are most important to the user.

The problem of imbalanced domains in regression presents an increased difficulty when compared to classification. In fact, in regression data sets the continuous nature of the target variable adds complexity to the task because there is a potentially infinite number of values to deal with, and the specification of the more/less relevant values of the target

is also not straightforward. For tackling imbalanced regression problems, which is the focus of this paper, only a few proposals were made. In this paper we propose a new method for addressing the problem of imbalanced regression. This method, which we named SMOGN, combines an under-sampling strategy with two over-sampling strategies. The main motivation for the use of two over-sampling procedures is to alleviate some of the problems inherent to those procedures.

This paper is organized as follows. Section 2 presents the definition of the problem of imbalanced regression. In Section 3 an overview of the existing related work is presented. Our SMOGN algorithm is described in Section 4 and the results of an extensive experimental evaluation are discussed in Section 5. Finally, Section 6 presents the main conclusions.

2. Problem Definition

The problem of imbalanced domains occurs in the context of predictive tasks. The main goal in these tasks is to obtain a model that approximates an unknown function $Y = f(\mathbf{x})$. In order to find this model, a training set $\mathcal{D} = \{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^{N}$ with N examples is used. The predictive task is named regression when the target variable Y is continuous, and is named classification when Y is nominal.

Imbalanced regression tasks are a particular class of regression problems that can be characterized by two properties: i) the user has **non-uniform preferences** across the target variable domain and ii) the most **important ranges are poorly represented**. This means that in imbalanced regression the user assigns more importance to the predictive performance achieved in some poorly represented ranges of the target variable in comparison with other more frequent ranges. The conjugation of these two factors cause a performance degradation on the most important cases for the user. If the user preferences are biased towards ranges of the target variable domain which are well represented, then the learning algorithms will not have difficulty in learning those cases and we do not have a problem of imbalanced domains. On the other hand, if there are ranges of the target variable domain poorly represented but the user is uniformly interested in all the domain values, then we also do not face a problem of imbalanced domains because all the cases are equally relevant to the user.

The problem of imbalanced regression implies an increased level of difficulty in comparison to the class imbalance problem because the target variable has a potentially infinite number of values. Therefore, the definition of the important/unimportant values of the target variable is an issue that must be considered. To address this issue, Torgo and Ribeiro (2007) and Ribeiro (2011) proposed the definition of a **relevance function**. The **relevance function**, $\phi : \mathcal{Y} \rightarrow [0, 1]$, maps the target variable domain into a scale of relevance, where 1 corresponds to the maximal relevance and 0 to the minimum relevance. Still, the task of deriving this relevance can be hard in regression. Moreover, this information is domain dependent and ideally should be provided by domain experts. To overcome this issue, Ribeiro (2011) proposed an automatic way for estimating the relevance function, $\phi(y)$, from the target variable sample distribution. The method proposed to obtain this estimate assumes the frequent setting where the rare and most extreme cases are the most relevant to the user. With a relevance function defined we can determine the sets of normal and rare values. To achieve this, the user is required to set a threshold t_R on the relevance values.

Given this threshold we can formally define the set of rare and relevant cases, $\mathcal{D}_R$, and the set of normal and uninteresting cases, $\mathcal{D}_N$, as follows: $\mathcal{D}_R = \{\langle \mathbf{x}, y \rangle \in \mathcal{D} : \phi(y) \geq t_R\}$ and $\mathcal{D}_N = \{\langle \mathbf{x}, y \rangle \in \mathcal{D} : \phi(y) < t_R\}$.

To deal with imbalanced regression problems we need to take into account both the performance assessment issue and the problem of biasing the learning algorithm towards the relevant cases. Regarding the performance assessment issue, the use of standard evaluation measures is not a suitable option (Ribeiro, 2011; Branco et al., 2016b). Special purpose evaluation measures are required in this context. A suitable framework was proposed by Torgo and Ribeiro (2009) and Ribeiro (2011) for obtaining precision and recall for imbalanced regression tasks. In this paper we use the F_1-measure (F_1^{ϕ}) proposed by Branco (2014) that is based on the mentioned framework. In this paper we are focused on the second issue and propose a new pre-processing solution for improving the learners capability in imbalanced regression tasks.

3. Related Work

The problem of imbalanced domains has been addressed mainly in a classification context. Therefore, a diversity of strategies exist to tackle this problem when the predictive tasks involve a nominal target variable. However, other predictive tasks that also suffer from the problem of imbalanced domains still remain scarcely studied (Branco et al., 2016b). This is the case of regression tasks, where the target variable is numeric.

The approaches for dealing with imbalanced domains may be clustered according to the moment where an intervention is made in the learning process. These approaches can be categorized as: i) pre-processing; ii) special purpose algorithm; iii) post-processing; or iv) hybrid. In this paper we focus on the first of these approaches. Pre-processing solutions change the original data distribution before the learning algorithm is applied. The goal is to change the target variable distribution to force the learning algorithm to focus on the rare and interesting cases. These solutions are among the most commonly used due to their flexibility regarding the use of any learning algorithm and their simplicity because they only involve manipulating the original data set distribution. However, we highlight that the efficiency of these methods is dependent on how the change in the data distribution is carried out, which is still an open issue.

An extensive set of proposals exist for tackling class imbalance problems. Regarding imbalanced regression tasks, only a few pre-processing methods were proposed. We will briefly describe the three following strategies: random under-sampling (Torgo et al., 2013, 2015), SmoteR (Torgo et al., 2013) and introduction of Gaussian Noise (Branco et al., 2016a). These methods were initially proposed for dealing with class imbalance and were later adapted to a regression context. In all these methods, to achieve this adaptation, the user must provide both a relevance function and a threshold on the relevance that are used to determine the D_R and D_N sets.

Random under-sampling is a straightforward strategy that randomly removes examples belonging to the normal and less interesting ranges of the target variable. This allows to achieve a better balance between the interesting/rare an uninteresting/normal cases. The user is also required to set the amount of reduction to be carried out in the normal cases. SmoteR (Torgo et al., 2013) is an adaption for regression of the well-known Smote

(Chawla et al., 2002) algorithm. This proposal applies random under-sampling in the normal cases and generates new synthetic "smoted" examples from the rare cases. The synthetic cases are generated through an interpolation strategy. This interpolation is carried out using two rare cases (one is a seed case and the other is randomly selected from the k-nearest neighbours of the seed). The features of the two cases are interpolated, and the new target variable value is determined as a weighted average of the target variable values of the two rare cases used. All rare cases are used in turn as seed examples. The user is also required to define the percentage of over and under-sampling to be carried out[1]. The introduction of Gaussian Noise (Branco et al., 2016a) is an adaptation to regression of the method proposed by Lee (1999, 2000) for classification tasks. This method also combines random under-sampling of the normal cases with the generation of synthetic rare examples. However, the new cases are generated using the addition of normally distributed noise to existing rare cases. The user is required to set the amount of over-/under-sampling to be carried out and the amount of noise that can be used in the synthetic cases generation.

4. SMOGN Algorithm

In this section we describe our proposal for dealing with imbalanced regression problems where the most important cases to the user are poorly represented in the available data. The algorithm we present is framed within the pre-processing approaches for tackling imbalanced domains which act before the learning process stage.

Our method is named SMOGN and combines random under-sampling with two over-sampling techniques: SmoteR and introduction of Gaussian Noise. The key idea of SMOGN algorithm is to combine both strategies for generating synthetic examples with the goal of simultaneously limiting the risks that SmoteR can incur into by using the more conservative strategy of introducing Gaussian Noise, and allow an increase of the diversity in examples generation, which is not possible to achieve using only the introduction of Gaussian Noise. SMOGN will generate new synthetic examples with SmoteR only when the seed example and the k-nearest neighbour selected are "close enough" and will use the introduction of Gaussian Noise when the two examples are "more distant". SMOGN is motivated by: i) the limitation of the risks incurred when using SmoteR because it will not use the most distant examples in the interpolation process; and ii) allowing the expansion of the decision boundaries for the rare cases increasing the generalization capability, which is more difficult to achieve with the introduction of Gaussian Noise because it is a more conservative approach.

Algorithm 1 describes our proposed SMOGN strategy. SMOGN algorithm begins by building data partitions containing consecutive examples considering the target variable value. These partitions are clustered into two types: $Bins_R$ - the rare and important partitions, and $Bins_N$ - the normal and less important partitions. This means that the data partitions in $Bins_R$ contain the higher relevance examples, i.e., examples with relevance above a pre-defined threshold, while the partitions included in $Bins_N$ have examples less interesting to the user because they have a lower relevance, i.e., the relevance score of the examples target variable value is below the threshold set. To the partitions included in $Bins_N$ a random under-sampling procedure is applied. On the other hand, the partitions in

1. Further details regarding SmoteR algorithm can be obtained in Torgo et al. (2013).

$Bins_R$ will be targeted with an over-sampling procedure. For each case (the seed example) in a partition belonging to $Bins_R$ a number of synthetic cases is generated. The over-sampling will use either SmoteR or the introduction of Gaussian Noise strategy to generate new cases depending on the distance between the seed example and the selected k-nearest neighbour. The main idea is that if the selected neighbour is "safe" then he is in a distance considered to be suitable to perform interpolation through the SmoteR strategy. On the other hand, if the selected neighbour is not in the safe range, then he is too far away to be used to perform interpolation which means that, in this case, it is better to generate a new example by introducing Gaussian Noise on the seed case. The threshold that is used to decide if the neighbour is at a safe or unsafe distance depends on the distance between the seed example and all the remaining cases in the partition under consideration. We used half of the median of the distances between the seed example and the other examples in the same partition.

Figure 1 shows a synthetic example with the 5-nearest neighbours of a seed case, where some are within a safe distance and others are at an unsafe distance. Examples marked with bullets are from a relevant bin, while examples marked with crosses are from a normal bin. In this example we show thatexamples belonging to the normal bin are more likely to overlap with the examples of the relevant bin at an unsafe distance.

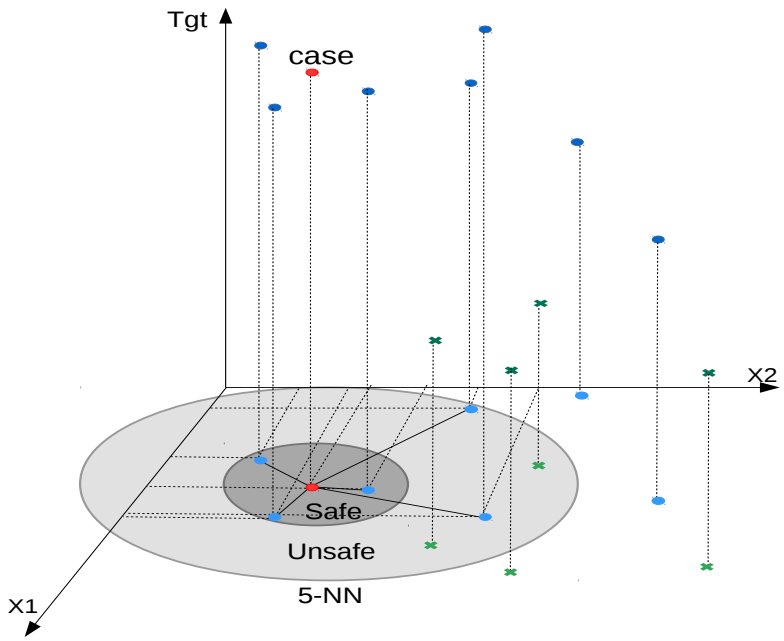

Figure 1: Synthetic example of the application of SMOGN algorithm.

5. Experimental Evaluation

We designed an experimental setup with the goal of assessing the effectiveness of SMOGN strategy in the context of imbalanced regression tasks. For this purpose we selected 20

Algorithm 1: SMOGN Algorithm.

Input: $\mathcal{D}$ - data set with target continuous variable Y
t_R - threshold for relevance on Y values
$\%u$ - percentage of under-sampling
$\%o$ - percentage of over-sampling
k - number of nearest neighbours
$dist$ - distance metric
Output: $newD$ - a new modified data set
$OrdD \leftarrow$ order $\mathcal{D}$ by ascending value of Y
$\phi() \leftarrow$ relevance function obtained from Y distribution
$Bins_N \leftarrow$ partitions of consecutive examples $\langle \mathbf{x}_i, y_i \rangle \in OrdD$, such that $\phi(y_i) < t_R$
$Bins_R \leftarrow$ partitions of consecutive examples $\langle \mathbf{x}_i, y_i \rangle \in OrdD$, such that $\phi(y_i) \geq t_R$
$newD \leftarrow Bins_R$
foreach $B \in Bins_N$ **do** `// random under-sampling procedure`
  $selNormCases \leftarrow$ randomly sample $\%u \times |B|$ cases from B
  $newD \leftarrow newD \bigcup selNormCases$
end
foreach $B \in Bins_R$ **do** `// over-sampling procedure`
  $ng \leftarrow \%o \times |B|$ `// nr of synthetic cases to generate for each case in` B
  foreach $case \in B$ **do** `// generate synthetic examples`
    $nns \leftarrow kNN(k, case, B, dist)$ `// k-Nearest Neighbours of case`
    $DistM \leftarrow$ distances between the case and the examples in B
    $maxD \leftarrow median(DistM)/2$
    for $i \leftarrow 1$ ***to*** ng **do**
      $x \leftarrow$ randomly choose one of the nns
      if $DistM(x) < maxD$ **then** `// safe kNN selected`
        $new \leftarrow$ use SmoteR to interpolate x and $case$
      else `// non-safe kNN selected`
        $pert \leftarrow min(maxD, 0.02)$
        $new \leftarrow$ introduce Gaussian Noise in $case$ with a perturbation $pert$
      end
      $newD \leftarrow newD \bigcup \{new\}$ `// add synthetic case to newD`
    end
  end
end
return $newD$

regression data sets from different imbalanced domains. Table 1 shows the main characteristics of the used data sets. We obtained a relevance function for each data set through the automatic method proposed by Ribeiro (2011). In this method the quartiles and inter-quartile range of the target variable distribution are used for assigning a higher relevance to both high and low extreme values of the target variable[2]. Therefore, the considered data sets will have either one extreme (on the high or low values of the target variable) or two

2. Further details available in Ribeiro (2011).

Table 1: Data sets information by descending order of rare cases percentage. (N: nr of cases; $p.total$: nr predictors; $p.nom$: nr nominal predictors; $p.num$: nr numeric predictors; $nRare$: nr. cases with $\phi(y) > 0.8$; $\%Rare$: $100 \times nRare/N$).

Data Set	N	p.total	p.nom	p.num	nRare	% Rare
servo	167	4	2	2	34	20.4
a6	198	11	3	8	33	16.7
Abalone	4177	8	1	7	679	16.3
machineCpu	209	6	0	6	34	16.3
a3	198	11	3	8	32	16.2
a4	198	11	3	8	31	15.7
a1	198	11	3	8	28	14.1
a7	198	11	3	8	27	13.6
boston	506	13	0	13	65	12.8
a2	198	11	3	8	22	11.1
a5	198	11	3	8	21	10.6
fuelCons	1764	38	12	26	164	9.3
availPwr	1802	16	7	9	157	8.7
cpuSm	8192	13	0	13	713	8.7
maxTorq	1802	33	13	20	129	7.2
bank8FM	4499	9	0	9	288	6.4
dAiler	7129	5	0	5	450	6.3
ConcrStr	1030	8	0	8	55	5.3
Accel	1732	15	3	12	89	5.1
airfoild	1503	5	0	5	62	4.1

Learner	Parameter Variants	R package
MARS	$nk = \{10, 17\}, degree = \{1, 2\}, thresh = \{0.01, 0.001\}$	**earth** (Milborrow, 2012)
SVM	$cost = \{10, 150, 300\}, gamma = \{0.01, 0.001\}$	**e1071** (Dimitriadou et al., 2011)
RF	$mtry = \{5, 7\}, ntree = \{500, 750, 1500\}$	**randomForest** (Liaw and Wiener, 2002)
NNET	$size = \{1, 2, 5, 10\}, decay = \{0, 0.01\}$	**nnet** (Venables and Ripley, 2002)

Table 2: Regression algorithms, parameter variants, and respective R packages used.

extremes (high and low extremes of the target variable). We considered a threshold of 0.8 on the relevance values in all data sets to obtain the set of rare/important cases, $\mathcal{D}_R$ and the set of normal/unimportant cases, $\mathcal{D}_N$. We can observe on Table 1 that this method allows us to obtain different percentages of rarity on the 20 used data sets, with values ranging between 4% and 20%. To ensure the reproducibility of our results, all code, data sets and results obtained are available in https://github.com/paobranco/SMOGN-LIDTA17.

All our experiments were carried out in the R environment. To ensure the diversity of the learning algorithms, we selected the four following types: Multivariate Adaptive Regression Splines (MARS), Support Vector Machines (SVM), Random Forests (RF) and Neural Networks (NNET). The learning algorithms, respective R packages and the used parameter variants are displayed in Table 2.

We applied each of the 28 learning approaches (8 MARS variants + 6 SVM variants + 6 RF variants + 8 NNET variants) to each of the 20 regression problems using 5 different resampling strategies. The resampling strategies that we tested were as follows: i) carrying out no sampling, i.e., using the original imbalanced data set (**None**); ii) random under-

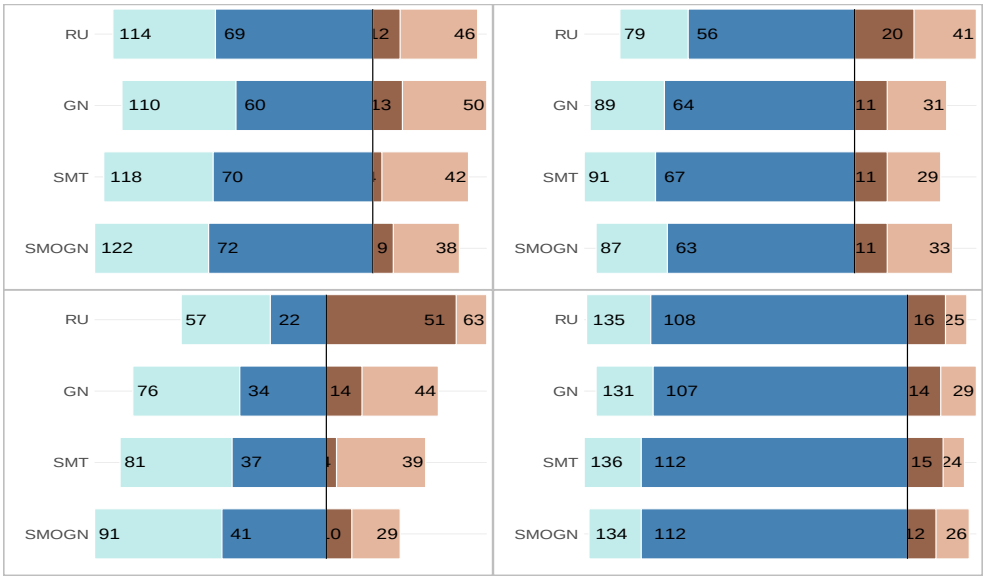

Figure 2: Wins (left) and losses (right) of each learner (top left: MARS, top right: SVM, bottom left: RF and bottom right: NNET) against **None**, i.e., the baseline of using the original imbalanced data sets.

sampling (**RU**); iii) SmoteR method (**SMT**); iv) introduction of Gaussian Noise (**GN**); and v) SMOGN algorithm (**SMOGN**). All the resampling strategies were applied with the goal of balancing the number of rare and normal cases and roughly maintain the same total number of examples in each data set, with the exception of random under-sampling strategy which is only able to reduce the data set size. Overall, we tested 2800 combinations ($28 \times 20 \times 5$).

As mentioned in Section 2, in imbalanced regression problems it is necessary to use a suitable evaluation measure. In all the experiments conducted we used the F_1^ϕ measure for regression (Branco, 2014). We used $\beta = 1$, which means that the same importance is given to both precision and recall scores. The F_1^ϕ values were estimated through a 2×10 - fold stratified cross validation process and the statistical significance of the observed paired differences was measured using the non-parametric Wilcoxon paired test for a significance level of 95%. The R packages used in our experiments were: *performanceEstimation* (Torgo, 2014) for the experimental infra-structure; *uba*[3] for obtaining the relevance function and F_1^ϕ metric evaluation; and *UBL* (Branco et al., 2016a) for the implementation of random under-sampling, SmoteR and introduction of Gaussian Noise resampling strategies.

The main results are summarized in Figures 2, 3 and 4. The detailed results, used code and data sets are provided in https://github.com/paobranco/SMOGN-LIDTA17. Figure 2 shows the total number of wins/losses and significant wins/losses obtained against the baseline of using the original data set through the Wilcoxon paired comparison for the F_1^ϕ measure. Darker bars indicate significant wins/losses while lighter bars represent wins/losses

3. Available at http://www.dcc.fc.up.pt/~rpribeiro/uba/.

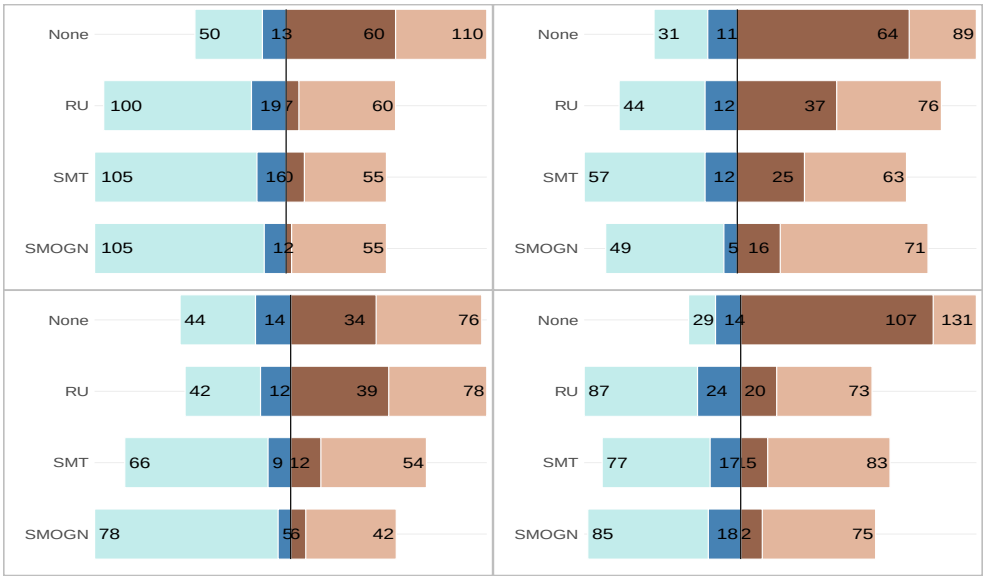

Figure 3: Wins (left) and losses (right) of each learner (top left: MARS, top right: SVM, bottom left: RF and bottom right: NNET) against the baseline of using the introduction of Gaussian Noise strategy.

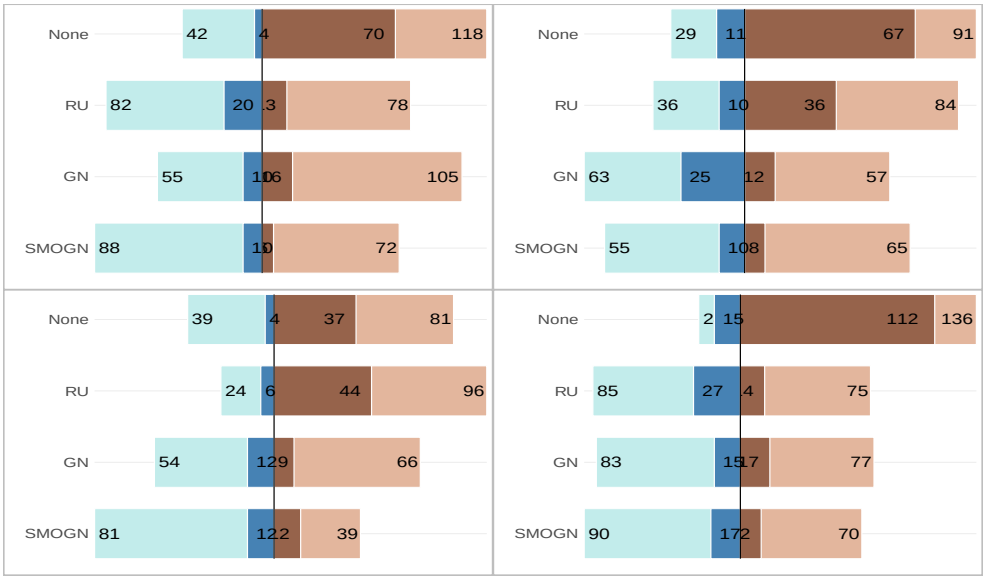

Figure 4: Wins (left) and losses (right) of each learner (top left: MARS, top right: SVM, bottom left: RF and bottom right: NNET) against the baseline of using the SmoteR strategy.

without statistical significance. The results are detailed by learning algorithm. Figures 3 and 4 display a similar comparison changing only the baseline against which the comparisons are made: introduction of Gaussian Noise is used on Figure 3 and SmoteR algorithm is used on Figure 4. A total of 120 (20 data sets $\times$ 6 learner variants) comparisons are made for the SVM and RF learners while 160 (20 data sets $\times$ 8 learner variants) comparisons are made for the MARS and NNET learners.

Generally, SMOGN algorithm has a performance comparable to SmoteR and to the introduction of Gaussian Noise. However, there is a clear difference in the results obtained for different learners. In fact, we observe that SMOGN has better results than the remaining algorithms for MARS and RF learners. The results are less favourable to SMOGN when an SVM is used, and the performance is similar to the SmoteR and introduction of Gaussian Noise strategies when NNET is applied. This means that our proposed method presents clear advantages when compared against None, introduction of Gaussian Noise and SmoteR when then learner used is MARS or RF.

Figures 5, 6 and 7 show the aggregated results obtained through the Wilcoxon paired comparison test for the resampling strategies tested against using respectively: the original data set (None), the SmoteR algorithm (SMT) and the introduction of Gaussian Noise strategy (GN). Each resampling strategy was compared against the baseline a total of 560 (20 data sets $\times$ 28 learners) times. The results presented show that globally the performance of SMOGN algorithm has advantages. For instance, the global number of wins of SMOGN is always larger than the alternative resampling strategies against all the baselines. Also for the global number of losses, SMOGN has always a lower number than the remaining strategies against all baselines. Moreover, all the strategies tested have more losses than wins against SmoteR strategy with the exception of SMOGN which displays more wins than losses. Against the introduction of Gaussian Noise, both SmoteR and SMOGN have more wins than losses.

The results show that our proposed algorithm has advantages when compared with the baseline of not using any resampling and also in comparison to the random under-sampling, SmoteR and introduction of Gaussian Noise resampling strategies. SMOGN algorithm achieves results close to the ones obtained through SmoteR and introduction of Gaussian Noise. We believe that this happens because the algorithm was build to deal with specific problems that can occur due to some data characteristics. When these problems are not present our method will have a behaviour similar to the SmoteR or the introduction of Gaussian Noise strategies. We believe that our method could stand out more when tested on data sets containing different regions of the features space with relevant cases. Due to space constraints, this exploration will be left for future work.

To provide a better understanding of the F_1^ϕ results, we show a brief analysis of the $prec^\phi$ and rec^ϕ metrics. This may be relevant because frequently, when dealing with the problem of imbalance domains in classification, the gains observed in terms of F_1 are achieved by considerably improving the recall while having some deterioration of performance on precision. Figures 8 and 9 show the Wilcoxon paired test results of $prec^\phi$ and rec^ϕ metrics in all learners, against the baseline of using the original imbalanced data sets. In this case, SMOGN method has a performance similar to the remaining methods.

Figures 10 and 11 show the results of Wilcoxon paired test against the baseline of introducing Gaussian Noise considering the $prec^\phi$ and rec^ϕ metrics. In this case results

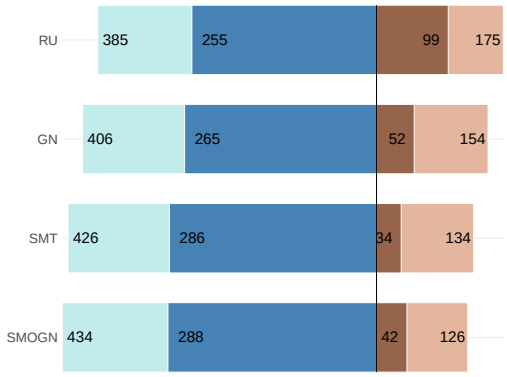

Figure 5: Overall wins (left) and losses (right) against the baseline of using the original imbalanced data sets.

show that random-undersampling method has a poor performance on $prec^{\phi}$. Still, we should note that all the methods with exception of the strategy that uses the original data sets show smaller significant wins. On the other hand, when considering rec^{ϕ} metric, all the pre-processing methods display higher wins and lower losses in comparison to the strategy of not applying any resampling. This confirms in the imbalanced regression context, what was already observed in imbalanced classification: pre-processing methods achieve an higher recall at the cost of worsening the precision results.

Figure 12 and 13 show the results of Wilcoxon paired test against the baseline of using SmoteR strategy. The same tendency is displayed in these figures. We highlight that random under-sampling method has generally less wins on $prec^{\phi}$ metric in comparison to the other methods when considering as baseline either the SmoteR or the introduction of Gaussian Noise method. Still, regarding the rec^{ϕ} this method has a similar performance. We also notice that SMOGN performance on both $prec^{\phi}$ and rec^{ϕ} stands out more when using the SmoteR as baseline. Further results of $prec^{\phi}$ and rec^{ϕ}, namely the metrics results by each learner, are available at https://github.com/paobranco/SMOGN-LIDTA17.

6. Conclusions

In this paper we presented a new pre-processing method, SMOGN, for tackling imbalanced regression problems. Being a pre-processing method, it has the advantage of being versatile because it allows the use of any standard learning algorithm. The method proposed tries to overcome difficulties in SmoteR strategy and in the introduction of Gaussian Noise strategy. It uses the interpolation method of SmoteR for interpolating examples that are closer. This way we tried to eliminate the risk of interpolating examples that, although being among the nearest neighbours of the seed example, are too distant. On the other hand, the use of SmoteR allows to expand the decision boundaries which is only achieved in a limited way with the more conservative method that generates synthetic cases by introducing Gaussian Noise.

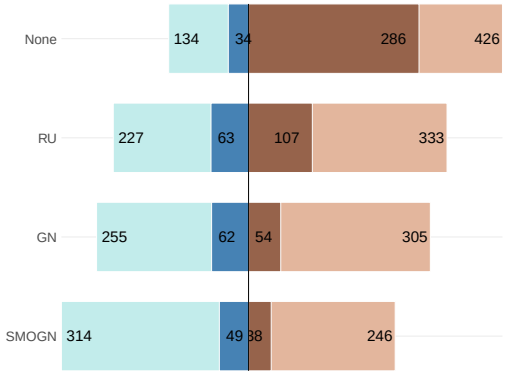

Figure 6: Overall wins (left) and losses (right) against the baseline of using **SMT** strategy.

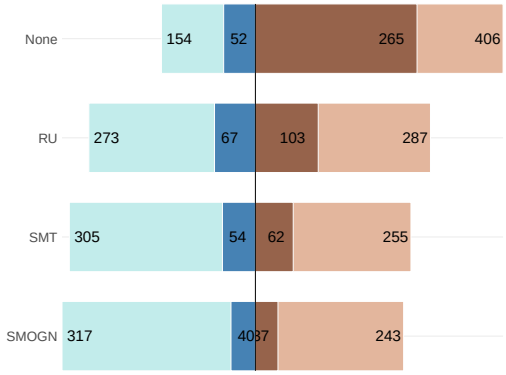

Figure 7: Overall wins (left) and losses (right) against the baseline of using the **GN** strategy.

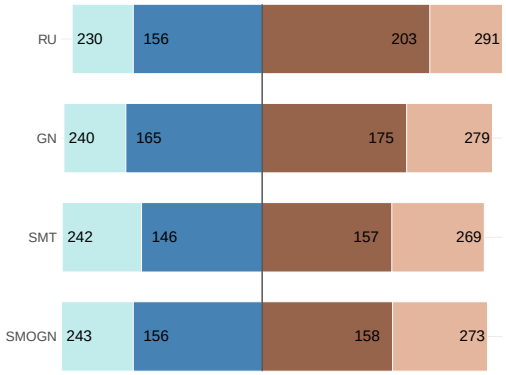

Figure 8: Overall wins (left) and losses (right) against the baseline of using the original imbalanced data sets considering the $prec^{\phi}$ metric.

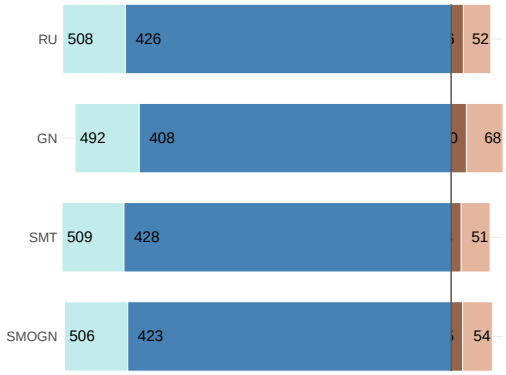

Figure 9: Overall wins (left) and losses (right) against the baseline of using original imbalanced data sets considering the rec^{ϕ} metric.

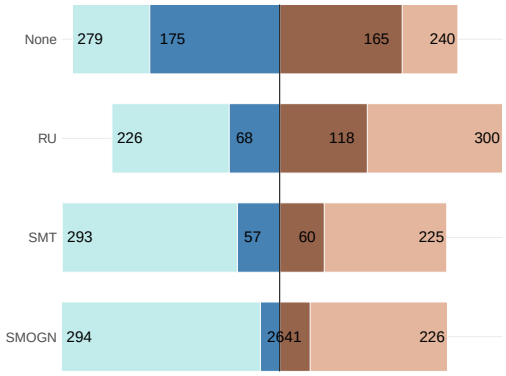

Figure 10: Overall wins (left) and losses (right) against the baseline of using introduction of Gaussian Noise strategy, considering the $prec^{\phi}$ metric.

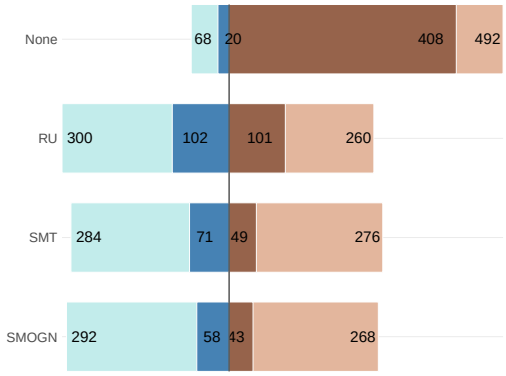

Figure 11: Overall wins (left) and losses (right) against the baseline of using introduction of Gaussian Noise strategy, considering the rec^{ϕ} metric.

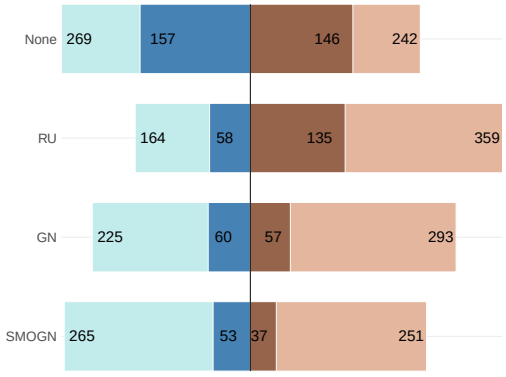

Figure 12: Overall wins (left) and losses (right) against the baseline of using SmoteR strategy, considering the $prec^{\phi}$ metric.

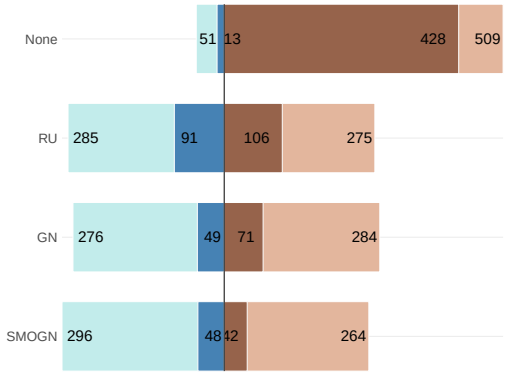

Figure 13: Overall wins (left) and losses (right) against the baseline of using SmoteR strategy, considering the rec^{ϕ} metric.

We show that our method has advantages in comparison to the SmoteR and the introduction of Gaussian Noise strategies. We also show that the change in the data distribution achieved with SMOGN has a different impact on the learners tested, displaying more advantages in RF and MARS learners and less in SVM and NNET. The key contributions of this paper are: i) the proposal of a new pre-processing method that is able to incorporate two over-sampling strategies; ii) test and compare our proposal against the baseline of using the original data and against the strategies SmoteR and introduction of Gaussian Noise.

As future work we plan to extend these approaches to imbalanced classification tasks, comparing the impact of SMOGN strategy in different domains. We would also like to explore: why this strategy has a different impact on the different learners, and which data characteristics may be related with a better performance in SMOGN strategy. Another important aspect to consider is the application of SMOGN proposal on data sets with different percentages of rare cases. To achieve this, different thresholds on the relevance values could be used on the data sets.

Acknowledgments

This work is financed by the ERDF – European Regional Development Fund through the COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, and by National Funds through the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) as part of project UID/EEA/50014/2013. The work of P. Branco is supported by a PhD scholarship of FCT (PD/BD/105788/2014). Prof. L. Torgo would also like to thank the support of Projects NORTE-01-0145-FEDER-000036 and UTAP-ICDT/CTM-NAN/0025/2014.

References

P. Branco. Re-sampling approaches for regression tasks under imbalanced domains. Master's thesis, Dep. Computer Science, Faculty of Sciences - University of Porto, 2014.

P. Branco, R. P. Ribeiro, and L. Torgo. UBL: an R package for utility-based learning. *arXiv preprint arXiv:1604.08079*, 2016a.

P. Branco, L. Torgo, and R. P. Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, 49(2):31, 2016b.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *JAIR*, 16:321–357, 2002.

E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2011.

H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

B. Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, pages 1–12, 2016.

S. S. Lee. Regularization in skewed binary classification. *Computational Statistics*, 14(2): 277, 1999.

S. S. Lee. Noisy replication in skewed binary classification. *Computational statistics & data analysis*, 34(2):165–191, 2000.

A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

V. López, A. Fernández, S. García, V. Palade, and F. Herrera. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250:113–141, 2013.

S. Milborrow. *earth: Multivariate Adaptive Regression Spline Models. Derived from mda:mars by Trevor Hastie and Rob Tibshirani.*, 2012.

R. P. Ribeiro. *Utility-based Regression*. PhD thesis, Dep. Computer Science, Faculty of Sciences - University of Porto, 2011.

L. Torgo. An infra-structure for performance estimation and experimental comparison of predictive models in r. *CoRR*, abs/1412.0436, 2014.

L. Torgo and R. P. Ribeiro. Utility-based regression. In *PKDD'07*, pages 597–604. Springer, 2007.

L. Torgo and R. P. Ribeiro. Precision and recall in regression. In *DS'09: 12th Int. Conf. on Discovery Science*, pages 332–346. Springer, 2009.

L. Torgo, R. P. Ribeiro, B. Pfahringer, and P. Branco. Smote for regression. In *Progress in Artificial Intelligence*, pages 378–389. Springer, 2013.

L. Torgo, P. Branco, R. P. Ribeiro, and B. Pfahringer. Resampling strategies for regression. *Expert Systems*, 32(3):465–476, 2015.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0.