Generalized SMOTE: A universal generation oversampling technique for all data types in imbalanced learning

Conference Paper · September 2021

CITATIONS
0

READS
130

6 authors, including:

Carlos Cernuda
Mondragon Unibertsitatea
23 PUBLICATIONS **304** CITATIONS
SEE PROFILE

Daniel Reguera-Bakhache
Mondragon Unibertsitatea
12 PUBLICATIONS **12** CITATIONS
SEE PROFILE

Aitor Aguirre
Mondragon Unibertsitatea
8 PUBLICATIONS **13** CITATIONS
SEE PROFILE

Mikel Iturbe
Mondragon Unibertsitatea
12 PUBLICATIONS **115** CITATIONS
SEE PROFILE

Generalized SMOTE: A universal generation oversampling technique for all data types in imbalanced learning

Carlos Cernuda
Electronics and Computing Dept.
Mondragon University
Arrasate-Mondragon, Spain
ccernuda@mondragon.edu

Daniel Reguera-Bakhache
Electronics and Computing Dept.
Mondragon University
Arrasate-Mondragon, Spain
dreguera@mondragon.edu

Aitor Aguirre
Electronics and Computing Dept.
Mondragon University
Arrasate-Mondragon, Spain
aaguirre@mondragon.edu

Mikel Iturbe
Electronics and Computing Dept.
Mondragon University
Arrasate-Mondragon, Spain
miturbe@mondragon.edu

Iñaki Garitano
Electronics and Computing Dept.
Mondragon University
Arrasate-Mondragon, Spain
igaritano@mondragon.edu

Urko Zurutuza
Electronics and Computing Dept.
Mondragon University
Arrasate-Mondragon, Spain
uzurutuza@mondragon.edu

***Abstract*— A common problem that arises when facing classification tasks is the class imbalance problem, which happens when one or more classes are heavily underrepresented compared to the rest, being usually those minority classes the ones of interest. A natural solution consists of correcting the imbalance by sampling methods, being Synthetic Minority Oversampling TEchnique (SMOTE) the most widely used method. In the same way as all other oversampling techniques, it relies on using distances/similarities in order to focus on the neighborhoods of minority samples in the synthetic samples generation procedure, thus it is meant for pure numerical data. Nevertheless, it is really common to collect categorical data or to discretize numeric attributes as a preprocessing step, being limited to random sampling approaches to correct imbalance. Some approaches have been proposed to deal with mixed-type data or pure categorical data, but they ignore part of the information of the samples or end up being almost random approaches.**

We propose GSMOTE, a generalization of SMOTE method, suitable for any data type. For the neighborhoods determination, the distance between samples is obtained by means of a transformation of Gower's General Similarity Coefficient into a novel General Distance Coefficient, in which the part corresponding to the way of measuring similarities between categories in categorical variables has been replaced by a recently presented similarity measure called Variable Entropy measure, inspired by Shannon's Entropy.

GSMOTE has been tested on six public imbalanced datasets, with different characteristics and imbalance levels.

***Index Terms*—Imbalanced Learning, Categorical Data, Mixed-Type Data, Oversampling Techniques, SMOTE**

I. Introduction

When facing a classification task is quite common to deal with datasets in which one or several classes are clearly underrepresented when compared to the rest. Such scenario is known as *imbalanced learning*. For instance, in real-world industrial or medical machine learning applications, such as fault detection or disease diagnosis, the classes of interest, respectively faulty and disease states, are pretty infrequent.

The underlying reasons for the imbalance are diverse. Industrial manufacturing processes [1] are highly optimized because no company could afford even moderate defective levels, thus usually less than 2% of the observations correspond to faulty states. In biomedical applications, the types of diseases that could benefit from machine learning models, for instance, are severe ones that are hard to detect by traditional methods with acceptable accuracy, which are not frequently happening [2] [3] [4] [5] [6]. In other cases, like in fraud detection in monetary transactions [7], the imbalance is due to the nature of the process itself, since most of the transactions are legal ones.

Independently of the root cause for the imbalance, the actions to prevent a classifier from predicting always the majority class (naïve classifier) are quite limited. Despite of the existence of certain taxonomies with more than two possibilities, we could include all into two:

(i) *Cost-sensitive learning* [8]. A cost-matrix is defined for the different errors that a classifier could make. Then, the classifier is trained in such a way that its goal is to minimize the global cost of all its errors. In this way, assigning a high cost to failing in classifying the minority class (or classes) would lead to models that are more accurate on predicting it (resp. them). The counterpart is that small improvements on the undesired errors would provoke an increase on the others that is much higher [9]. The trade-off between errors provided by our control over the cost-matrix is tricky to adjust, due to overfitting problems because of forcing the algorithms to focus on minority class samples during training [10].

(ii) *Sampling strategies*. The differences in the amount of samples from majority and minority classes is compensated

by increasing the amount minority class samples (oversampling) and/or decreasing the amount of majority class ones (undersampling). Because both have pros and cons [11] [12] [13], it is also common to opt for a hybrid approach (*mixed sampling*) [14] [15] [16] combining them.

In the case of *oversampling*, apart from randomly repeating some minority class samples, most of the approaches *generate* new synthetic minority class samples in the neighborhoods of the existing original ones. The most widely used algorithm is *synthetic minority oversampling technique (SMOTE)* [17]. In Smote, each synthetic sample, x_{new}, is randomly selected in the segment that connects one of the original minority class samples, x_i, with another minority class sample from its neighborhood, x_j. Unless we opt for random-based approaches, which usually perform poorly [15], we will focus on neighborhoods, thus distance-based approaches, meaning that we are limited to numerical features. Moreover, we could not discretize numerical features and then correct imbalance, but the other way around, leading to bins that are influenced by the synthetic data we have artificially introduced.

In the original SMOTE paper, the authors propose one adaptation for mixed type data, called *SMOTE Nominal Continuous* (SMOTE-NC), which assigns to every non-matching category of a categorical attribute a fixed distance equal to the median of the standard deviations of the numerical features for the minority class samples. Therefore, it uses only the information provided by the numeric part of the samples, leading to dangerous distances assumptions in case only a few of the attributes are numerical. Moreover, it ignores any information of the categorical attributes such as the amount of categories and their distribution.

Another adaptation is proposed for pure nominal features, called *SMOTE Nominal* (SMOTE-N), which uses a metric called Value Difference Metric [18] for characterizing distances/dissimilitudes. The way the synthetic samples are generated promotes keeping parts that are common with any of the neighbors'. This makes that synthetic samples coming from attributes with few values would be mainly copies, reducing to *random oversampling*.

In this paper, we propose an oversampling strategy that is suitable for all data types, i.e. pure categorical or numerical and mixed-type data, meant for binary problems. We have called it Generalized Synthetic Minority Oversampling TEchnique (GSMOTE) because we get original SMOTE when applied to pure numerical data, thus it could be considered as a generalization of SMOTE. The reason for considering the original SMOTE as an starting point is to see more clearly the potential of the approach. Nevertheless, the proposed generalization could be straightforwardly made from almost any variant.

In GSMOTE, the distances are calculated by means of a transformation of Gower's General Similarity Coefficient [19] into a distance coefficient. In this new General Distance Coefficient, the part related to categorical features has been replaced by a distance derived from the Variable Entropy (VE) measure [20], a similarity measure originally developed for hierarchical clustering of categorical data.

The creation of synthetic minority class samples from pairs of existing ones is decoupled for categorical and numerical parts of the samples. In both parts the same idea employed by SMOTE is maintained, i.e. randomly selecting one point in the segment that connects both original points by means of a convex combination, being adapted for the categorical parts keeping the concept of proximity to the extremes. Final synthetic samples are obtained by joining both parts.

The rest of the paper is organized as follow. In Section II, we introduce all previous works that are related to our method. In Section III, we present our contributions, consisting of a novel General Distance Coefficient (GDC), and the generalization of SMOTE algorithm, which uses the GDC. Section IV describes the datasets, experimental scheme and results achieved. Finally, in Section V we present conclusions and possible future works.

II. Premises

A. General Similarity Coefficient

In order to measure the similarity between two mixed-type data points, $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$, Gower [19] defined the General Similarity Coefficient, given by

$$s_G\left(\boldsymbol{x_i}, \boldsymbol{x_j}\right) = \frac{1}{\sum_{k=1}^{d} w(x_{ik}, x_{jk})} \sum_{k=1}^{d} w(x_{ik}, x_{jk})\, s(x_{ik}, x_{jk}) \qquad (1)$$

where $s(x_{ik}, x_{jk})$ is a similarity component for the k-th attribute and $w(x_{ik}, x_{jk})$ is a binary constant with value 1 if the comparison between values for the k-th attribute is valid, and 0 otherwise.

The similarity components are defined differently depending on the types of attributes as follows

- For *quantitative* attributes, the similarity $s(x_{ik}, x_{jk})$ is defined as
$$s(x_{ik}, x_{jk}) = 1 - \frac{|x_{ik} - x_{jk}|}{R_k},$$
with R_k the range of the attribute.
- For *binary* attributes, the similarity is defined as
$$s_b(x_{ik}, x_{jk}) = \begin{cases} 1 & , \quad \text{if } x_{ik} = x_{jk} = true \\ 0 & , \quad \text{otherwise} \end{cases}$$
- For *nominal* or *categorical* attributes, the similarity $s(x_{ik}, x_{jk})$ is defined as
$$s(x_{ik}, x_{jk}) = \begin{cases} 1 & , \quad \text{if } x_{ik} = x_{jk} \\ 0 & , \quad \text{if } x_{ik} \neq x_{jk} \end{cases}$$

B. Variable Entropy similarity measure

Given a random variable X, taking values $x_1, \ldots, x_n$, *Shannon's Entropy* of X is defined as

$$H(X) = -\sum_{k=1}^{n} P(x_k) \log_b P(x_k)$$

where b is the base of the logarithm (usually taking values 2, e or 10). It could be interpreted as the uncertainty or the informativeness inherent to the possible outcomes of X, and it is

deeply related to the variability of the random variable. In fact, the maximum entropy happens when all values of the random variable are equally probable, i.e. maximum uncertainty.
With this interpretation in mind, we think of two different categorical variables, one with large variability and all values almost equally likely, and the other one with small variability, having one predominant category and the rest quite sparse. In such case, it is desirable that a match in any category of the first variable is interpreted as higher similarity than a match in the predominant category of the second, but lower similarity that a match in any of the sparse ones.
Accordingly, a similarity measure following the aforementioned philosophy should give higher weight to matches in the case of high variability, thus entropy.
Considering the sample relative frequencies of the categories as their probabilities, the similarity between categories is defined as [20]

$$S_k(x_{ik}, x_{jk}) = \begin{cases} -\dfrac{1}{\ln N_k} \displaystyle\sum_{u=1}^{N_k} p_u \ln p_u & , \quad \text{if } x_{ik} = x_{jk} \\ 0 & , \quad \text{if } x_{ik} \neq x_{jk} \end{cases} \tag{2}$$

Notice that in case of a match, it could theoretically take values in the unit interval $[0, 1]$, including 0. That null value could only be attained if one of the variables is constant. Nevertheless, in such case that category is not relevant for the similarity, thus null value seems reasonable. On the contrary, a value 1 is only possible if all categories in one of the variables are equally likely.
In order to determine the similarity between two samples $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$, we define the *Variable Entropy* (VE) similarity measure as

$$S_{ve}\left(\boldsymbol{x_i}, \boldsymbol{x_j}\right) = \frac{\displaystyle\sum_{k=1}^{d} S_k(x_{ik}, x_{jk})}{d} \tag{3}$$

Again, the similarity measure takes values in $[0, 1]$, being the extreme values accessible if and only if all d averaged values take that exact extreme value.

C. Synthetic Minority Oversampling TEchnique

Synthetic Minority Oversampling TEchnique [17], briefly described above, is an oversampling technique that, in its original version, generates synthetic minority class samples inside randomly selected minority class samples' neighborhoods, which are obtained by *k*-Nearest Neighbors algorithm with a predefined *k*. The new synthetic minority class samples are determined by randomly sampling points in the segment that connects the initial sample with one of its *k* neighbors.
If we denote by $\boldsymbol{x_i}$ the central original minority class sample, and by $\boldsymbol{x_{n_1}}, \ldots, \boldsymbol{x_{n_k}}$ its K minority class nearest neighbors, then one of the K neighbors is randomly selected (let us denote it by $\boldsymbol{x_j}$). Mathematically, the segment joining two points is given by their convex linear combination. In this case, it corresponds to

$$\boldsymbol{x_{new}} = \lambda \boldsymbol{x_j} + (1 - \lambda)\boldsymbol{x_i}, \quad \text{with } \lambda \in [0, 1] \tag{4}$$

The selection of one single point is obtained by sampling one λ value. Notice that the two boundary λ values would replicate the extreme points of the segment.

III. Our approach

A. General Distance Coefficient

On the basis of Gower's General Similarity Coefficient, and taking into account the relationship between dissimilarity and distance, we can derive the following *General Distance Coefficient*, given by

$$d_{gdc}\left(\boldsymbol{x_i}, \boldsymbol{x_j}\right) = \left(\frac{1}{\displaystyle\sum_{k=1}^{d} w(x_{ik}, x_{jk})} \sum_{k=1}^{d} w(x_{ik}, x_{jk})\, d^2(x_{ik}, x_{jk}) \right)^{1/2} \tag{5}$$

where $d^2(x_{ik}, x_{jk})$ is a squared distance component for the k-th attribute and $w(x_{ik}, x_{jk})$ is a binary constant with value 1 if the comparison between values for the k-th attribute is valid, and 0 otherwise.
The distance components are defined differently depending on the types of attributes as follows

- For *continuous* and *ordinal* attributes, the distance $d(x_{ik}, x_{jk})$ is defined as

  $$d(x_{ik}, x_{jk}) = \frac{|x_{ik} - x_{jk}|}{R_k},$$

  with R_k the range of the attribute.
- For *quantitative* attributes, $d(x_{ik}, x_{jk})$ is defined as

  $$d(x_{ik}, x_{jk}) = |x_{ik} - x_{jk}|$$

  If we standardize $x_{\bullet k}$ as

  $$x^*_{\bullet k} = \frac{x_{\bullet k} - \mu_k}{\sigma_k},$$

  where μ_k and $sigma_k$ are, respectively, the mean and standard deviation of the k-th attribute, it could also be normalized, such that

  $$d(x_{ik}, x_{jk}) = \frac{\left|x^2_{ik} - x^2_{jk}\right|}{\sigma_k}.$$

- For *binary* attributes, the distance is defined, also following VE philosophy, as

  $$d(x_{ik}, x_{jk}) = \begin{cases} 1 - \dfrac{1}{\ln 2} \displaystyle\sum_{u=1}^{2} p_u \ln p_u & , \quad \text{if } x_{ik} = x_{jk} = true \\ 1 & , \quad \text{otherwise} \end{cases}$$

  Notice that null distance is reachable if and only if

  $$p_1 = P(true) = 1/2 = P(false) = p_2$$

- For *nominal* or *categorical* (not *binary*) attributes, the distance $d(x_{ik}, x_{jk})$ is defined as the dissimilarity obtained from the VE similarity measure

  $$d(x_{ik}, x_{jk}) = \begin{cases} 1 - \dfrac{1}{\ln N_k} \displaystyle\sum_{u=1}^{N_k} p_u \ln p_u & , \quad \text{if } x_{ik} = x_{jk} \\ 1 & , \quad \text{if } x_{ik} \neq x_{jk} \end{cases}$$

  where $N_k \geq 3$.

The reason for the separation into *continuous and ordinal* and the *rest of quantitative* attributes in the way of normalizing, avoiding the range for the latter, is the lack of determination of the variability using only the information provided by the maximum and minimum values in such case. Instead, the standard deviation is more reasonable [21].

B. Generalized Synthetic Minority Oversampling TEchnique

The Generalized Synthetic Minority Oversampling TEchnique (GSMOTE) algorithm comprises the following steps:

1) Input:

(i) A dataset D, with dimensionality d (i.e. number of input features) and a binary target class vector, given by

$$\{(\boldsymbol{x_i} = (x_{i1}, \ldots, x_{id}), y_i), i = 1, \ldots, m\}$$

where m_p samples correspond to the positive (minority) class and m_n to the negative (majority) class, with $m = m_p + m_n$; and where $d_c \geq 0$ features are categorical/binary and $d_n \geq 0$ are numerical, so that $d = d_c + d_n$.

(ii) A predefined *number of neighbors* to be considered for synthetic samples generation (default $K = 5$)

(iii) A predefined *maximum permitted imbalance ratio* (default $\beta = 0.1$).

2) Procedure:

(i) Calculation of the *degree of imbalance*, which belongs to $(0, 1]$,

$$b = {}^{m_p}/_{m_n} \tag{6}$$

(ii) Calculation of the *amount of synthetic minority samples*, m_s, needed to fulfill the β limitation, that is

$$m_s = \lceil m_n(1 - \beta) - m_p \rceil \tag{7}$$

where $\lceil z \rceil$ denotes the ceiling function, i.e. the smallest integer bigger or equal than z. Notice that, if no imbalance is permitted, then $\beta = 0$ and, therefore, $m_s = m_n - m_p$, as expected.

(iii) For each sample $\boldsymbol{x_i}$ we calculate its K nearest neighbors, using the *general distance coefficient* in Equation 5.

(iv) Assuming, without loss of generality, that the first d_c samples are categorical/binary and the last d_n are numerical, and denoting by P the subset of minority class samples

$$P = \{\boldsymbol{x_i} = (x_{i1}, \ldots, x_{id}), i = 1, \ldots, m_p\} \tag{8}$$

For each sample

$$\boldsymbol{x_i} = (\underbrace{x_{i1}, \ldots, x_{id_c}}_{\text{Cat/Bin}}, \underbrace{x_{i(d_c+1)}, \ldots, x_{id}}_{\text{Numerical}}) \tag{9}$$

For operative reasons in the following path calculation step, we have kept together categorical and binary attributes in the separation. Nevertheless, they have been treated differently in the distance calculation.

(v) We repeat G times the following process

1) Randomly select one sample $\boldsymbol{x_i}$ from P.
2) Consider its neighborhood and randomly select one of its neighbors, $\boldsymbol{x_j}$. Split both samples in their categorical/binary and numerical parts. We denote them by $\boldsymbol{x_i^c}$ and $\boldsymbol{x_j^c}$, for categorical/binary parts, and by $\boldsymbol{x_i^n}$ and $\boldsymbol{x_j^n}$ for numérical.
3) Calculate the path from the categorical/binary part of $\boldsymbol{x_i}$

$$\boldsymbol{x_i^c} = (x_{i1}, \ldots, x_{id_c})$$

to the one of $\boldsymbol{x_j}$

$$\boldsymbol{x_j^c} = (x_{j1}, \ldots, x_{jd_c})$$

ordering the samples from closest to farthest from $\boldsymbol{x_i}$, according to the *Manhattan distance*. In this case, with our notation, that distance is defined as

$$d_M(\boldsymbol{x_i^c}, \boldsymbol{x_j^c}) = \sum_{k=1}^{d_c} |x_{ik} - x_{jk}|$$

As an example, consider that $\boldsymbol{x_i^c} = (1, 1, 3, 2)$ and $\boldsymbol{x_j^c} = (1, 3, 3, 1)$. The steps in the path, including all Manhattan distances to $\boldsymbol{x_i^c}$, would be

$$\begin{array}{rcrcll} \boldsymbol{x_i^c} & = & \boldsymbol{x_{s_0}^c} & = & (1,1,3,2) & \Rightarrow \; d_M = 0 \\ & & \boldsymbol{x_{s_1}^c} & = & (1,2,3,2) & \Rightarrow \; d_M = 1 \\ & & \boldsymbol{x_{s_2}^c} & = & (1,1,3,1) & \Rightarrow \; d_M = 1 \\ & & \boldsymbol{x_{s_3}^c} & = & (1,2,3,1) & \Rightarrow \; d_M = 2 \\ & & \boldsymbol{x_{s_4}^c} & = & (1,3,3,2) & \Rightarrow \; d_M = 2 \\ \boldsymbol{x_j^c} & = & \boldsymbol{x_{s_5}^c} & = & (1,3,3,1) & \Rightarrow \; d_M = 3 \end{array}$$

4) Sample a random λ in $[0, 1]$, and then
4.1. Apply Eq. 4 to the numeric parts, obtaining the new synthetic numerical part $\boldsymbol{x_{new}^n}$, as in original SMOTE algorithm.
4.2. If we denote by n_s the amount of intermediate steps in the path, and we define $\rho = \frac{1}{n_s+1}$, we can define the following partition of the unit interval

$$[0,1] = \underbrace{\left[0, \frac{1}{2}\rho\right]}_{I_0} \cup \bigcup_{v=1}^{n_s} \underbrace{\left(\frac{2v-1}{2}\rho, \frac{2v+1}{2}\rho\right]}_{I_v} \cup \underbrace{\left(\frac{2n_s+1}{2}\rho, 1\right]}_{I_{n_s+1}} \tag{10}$$

Then, we obtain the new synthetic categorical/binary part by selecting $\boldsymbol{x_{new}^c} = \boldsymbol{x_{s_w}^c}$, where w is so that $\lambda \in I_w$. In the previous example, $n_s = 4$, $\rho = {}^1/_5$, thus the partition of the unit interval would be

$$\underbrace{\left[0, \frac{1}{10}\right]}_{I_0} \cup \underbrace{\left(\frac{1}{10}, \frac{3}{10}\right]}_{I_1} \cup \underbrace{\left(\frac{3}{10}, \frac{1}{2}\right]}_{I_2} \cup \underbrace{\left(\frac{1}{2}, \frac{7}{10}\right]}_{I_3} \cup \underbrace{\left(\frac{7}{10}, \frac{9}{10}\right]}_{I_4} \cup \underbrace{\left(\frac{9}{10}, 1\right]}_{I_5}$$

For instance, if $\lambda = 0.47 \in \left(\frac{3}{10}, \frac{1}{2}\right] = I_2$, then $\boldsymbol{x_{new}^c} = \boldsymbol{x_{s_2}^c} = (1, 1, 3, 1)$.

5) Get the new synthetic sample $\boldsymbol{x_{new}}$ by joining both new parts $\boldsymbol{x_{new}^c}$ and $\boldsymbol{x_{new}^n}$. Under our assumptions

$$\boldsymbol{x_{new}} = (\boldsymbol{x_{new}^c}, \boldsymbol{x_{new}^n}) \tag{11}$$

(vi) We add the new G synthetic minority samples to the dataset.

TABLE I
SUMMARY INFORMATION OF THE UCI DATASETS USED.

Name	Target	Imb. ratio	# Samples	# Feat.
Ecoli	imU	8.6 : 1	336	7 num
Pen Dig.	Number 5	9.4 : 1	10992	16 num
Wine Qual.	Quality ≤ 4	26 : 1	4898	11 num
Sick Euth.	sick_euthyroid	9.8 : 1	3163	6 num, 36 cat
Car Eval.	good, vgood	12 : 1	1728	6 cat
Solar Fl.	Moderate = 0	19 : 1	1389	10 cat

TABLE II
RESULTS, MEAN $\pm$ STD WITH TEN REPETITIONS, FOR DIFFERENT PERCENTAGES OF CATEGORICAL/BINARY DATA IN ECOLI, PEN DIGITS AND WINE QUALITY DATASETS.

Perc.	Method	Ecoli	Pen Dig.	Wine Qual.
25%	No	.9123 $\pm$ .0140 †	.99984 $\pm$ 10^{-5}	.8427 $\pm$ .0027 †
	SM-NC	.9926 $\pm$ .0019	.99981 $\pm$ 10^{-5}	.9843 $\pm$ .0008 †
	GSMOTE	**.9970 $\pm$ 0.0033**	**.99993** $\pm$ 10^{-6}	**.9942 $\pm$ .0008**
50%	No	.9176 $\pm$ .0107 †	.99984 $\pm$ 10^{-5}	.8559 $\pm$ .0029 †
	SM-NC	.9872 $\pm$ .0116	.99988 $\pm$ 10^{-5}	.9865 $\pm$ .0017
	GSMOTE	**.9931 $\pm$ .0014**	**.99991** $\pm$ 10^{-5}	**.9946 $\pm$ .0007**
75%	No	.9354 $\pm$ .0098 †	.99984 $\pm$ 10^{-5}	.8514 $\pm$ .0032 †
	SM-NC	.9822 $\pm$ .0087 †	**.99990** $\pm$ 10^{-5}	.9933 $\pm$ .0007
	GSMOTE	**.9934 $\pm$ .0021**	**.99990** $\pm$ 10^{-5}	**.9960 $\pm$ .0004**

IV. EXPERIMENTS

A. Experimental scheme

We have considered six datasets from UCI repository [22], in the way they have been proposed in [23]. Table I summarizes the information about them.

For validation purposes we will compare the performance of GSMOTE with the original imbalanced sets, i.e. no imbalance correction, in the case of categorical/binary data, and also with SMOTE-NC in mixed-type data.

In the latter, in order to check the behavior of the algorithms for different ratios of categorical vs numerical attributes, we have considered the three pure numerical datasets (Ecoli, Pen Digits, and Wine Quality [24]). For each of them, we have randomly selected 25%, 50% and 75% of their attributes, and we have discretized them by means of the Chi2 discretization algorithm [25]. In order to mitigate possible deviations because of the attribute selection by chance, we have performed each ten times. In the former, we have considered all six datasets described in Table I, in which all numerical attributes have also been discretized by Chi2 algorithm.

The algorithm employed for the tests is random forest (RF) [26], because it is a powerful algorithm that is pretty efficient for mixed-type data, due to be based on decision trees. For the selection of the hyperparameters of the model, we have performed an exhaustive grid search, varying the number of trees (in $\{50, 100, 200\}$), the maximum depth of each tree (in $\{5, 10, 20, \infty\}$) and the amount of attributes visible in each splitting node (in $\{d, \sqrt{d}\}$), coupled with a 10-folds cross validation strategy, using the area under the receiver operator curve (AUC) for scoring because it is robust against the effect of imbalance in absence of a cost-sensitive approach [27]. Moreover, as random forest is a stochastic method, so we have repeated each procedure 10 times. Therefore, for each imbalance learning method, we have made 720 trials per dataset, including the original ones.

B. Results and discussion

The results discussed are presented as follows. First, in the case of mixed-type data, we report the results of the average score of the 10 repetitions of RF, for the best hyperparameters of the grid in each case in Table II. Last, in the case of pure categorical/binary data, the results are presented in Table III. In both cases, statistical significance tests have been carried out by means of Wilcoxon signed-rank test [28], being indicated in the tables in case the null hypothesis has been rejected (meaning a significant difference) when compared with the best (in bold).

In the case of Ecoli, we can appreciate that the performance of both approaches is around 10% better in average than the imbalanced dataset. Despite it cannot be said that GSMOTE behaves much better, it is slightly better most of the times, independently of the percentage of categorical attributes. One should notice that the dimensionality of this dataset is low, so there is not much margin to appreciate the influence of the type of estimation made by SMOTE-NC based on the standard deviations, except perhaps for 75%. In fact, this is the case in which the difference of the average AUCs is higher (2.5 times higher than for 25%). For pure categorical data, we can appreciate that the AUC score is also high for 100% categorical attributes, but lower than the averages of the rest of lower percentages.

When it comes to mixed data in Pen Digits dataset, the margin for improvement is really low because the AUC is very high even without correcting the imbalance. Nevertheless, this could be seen as a challenge both approaches have passed. Again, GSMOTE overtakes SMOTE-NC in single wins, but this time the highest difference happens with the lowest proportion of categorical features. Without an exhaustive study of the nature and characteristics of the datasets, out of the scope of this paper, it is hard to find a reason for it. The tendency remains the same for full categorical data, including the tight margin for improvement (see Table III).

Now, on mixed-type data for Wine Quality dataset, we could say that GSMOTE is clearly the best algorithm, winning every single time, and rising up the AUC score around 15% with respect to the original imbalanced data. It is also noticeable that, in all three cases, the higher the percentage of categorical features the better the performance. This is not maintained in the case of pure categorical data, as can be seen in Table III.

Finally, Table III presents the results on pure categorical/binary data for all datasets. In all cases, it is beneficial to perform the imbalance correction. Apart from the three numerical datasets, commented before, Eurothyroids is a clear example of the possibility of getting a mixed-type set and discretize all numerical attributes (6 up to 36 in this case) before correcting the imbalance, leading to successful results. Finally, CarEval

TABLE III
RESULTS FOR PURE CATEGORICAL/BINARY DATA.

Method	Ecoli	Pen	Wine	Sick	Car	Solar
No	.9477 †	.9998	.8602 †	.9835 †	.9871	.7919 †
GSMOTE	**.9926**	**.9999**	**.9517**	**.9950**	**.9966**	**.9030**

and SolarFlare, pure categorical originally, benefit as well from the correction (14% in the case of SolarFlare).

V. CONCLUSIONS

We have presented Generalized SMOTE, a prototype-generation oversampling method that could be employed in pure numerical and categorical, as well as in mixed-type data. In the particular case of pure numerical data, it reduces to the original SMOTE.

The main advantage over other existing techniques is the better use of the inherent information, because it takes into account the amount of categories and their distribution in the minority class subspace, in order to assign the distances between samples in the neighborhood search. Moreover, it decouples the prototype generation in categorical and numerical parts, being able to keep the notion of proximity implicit in the way SMOTE generates synthetic samples in the segments connecting original ones.

The results obtained in six public diverse imbalanced datasets support GSMOTE as a powerfull method.

ACKNOWLEDGMENT

This work has been partially funded by the project Semantic Knowledge Integration for Content-Based Spam Filtering (TIN2017-84658-C2-2-R) from the Spanish Ministry of Economy, Industry and Competitiveness (SMEIC), State Research Agency (SRA) and European Regional Development Fund (ERDF), and by the project REMEDY - REal tiME control and embeddeD securitY from the Department of Economic Development and Infrastructures of the Basque Government under the grant agreement KK-2021/00091. It has been developed by the intelligent systems for industrial systems group supported by the Department of Education, Language policy and Culture of the Basque Government.

REFERENCES

[1] Cernuda C.: On the Relevance of Preprocessing in Predictive Maintenance for Dynamic Systems. In: Lughofer E., Sayed-Mouchaweh M. (eds) Predictive Maintenance in Dynamic Systems. Springer, **Chapter 3**, 53–93 (2019)

[2] Davis D.N., Nguyen T.T.T.: Generating and Verifying Risk Prediction Models Using Data Mining (A Case Study from Cardiovascular Medicine). 57th Annual Congress of the European Society for Cardiovascular Surgery (ESCVS), Barcelona, Spain (2008)

[3] Mar J., Gorostiza A., Arrospide A. et al.: Estimation of the epidemiology of dementia and associated neuropsychiatric symptoms by applying machine learning to real-world data. Journal of Psychiatry and Mental Health (2021). DOI: *https://doi.org/10.1016/j.rpsm.2021.03.001*

[4] Laza R., Pavon R., Reboiro-Jato M., Fdez-Riverola F.: Evaluating the effect of unbalanced data in biomedical document classification. Journal of Integrative Bioinformatics **8**(3), 1–13 (2011)

[5] Rahman M.M., Davis D.N.: Addressing the Class Imbalance Problem in Medical Datasets. International Journal of Machine Learning and Computing **3**(2), 224–228 (2013)

[6] Mar J., Gorostiza A., Ibarrondo O. et al.: Validation of random forest machine learning models to predict dementia-related neuropsychiatric symptoms in real-world data. Journal of Alzheimer's Disease **77**(2), 855–864 (2020)

[7] Phua, C., Alahakoon, D.: Minority report in fraud detection: Classification of skewed data. ACM SIGKDD Explorations Newsletter **6**(1), 50–59 (2004)

[8] Thai-Nghe N., Gantner Z., Schmidt-Thieme L.: Cost-sensitive learning methods for imbalanced data. The 2010 International Joint Conference on Neural Networks (IJCNN), Barcelona Spain, 1–8 (2010)

[9] Attenberg J., Ertekin S.: Class Imbalance and Active Learning. In: He H., Ma Y. (eds) Imbalanced Learning: Foundations, Algorithms, and Applications. John Wiley & Sons, Chapter 6, 101–149 (2013)

[10] Fernández A., García S., Galar M., Prati R.C., Krawczyk B., Herrera F.: Cost-Sensitive Learning. In: Learning from Imbalanced Data Sets. Springer, Cham, Chapter 4, 63–78 (2018)

[11] Chawla, N.V.: C4.5 and Imbalanced Data sets: Investigating the Effect of Sampling Method, Probabilistic Estimate, and Decision Tree Structure. Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data sets, Washington, DC, USA (2003)

[12] Drummond, C., Holte, R.: C4.5, class imbalance, and cost sensitivity: Why undersampling beats over-sampling. Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets, Washington, DC, USA (2003)

[13] Maloof, M.: Learning when data sets are imbalanced and when costs are unequal and unknown. Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets, Washington, DC, USA (2003)

[14] Kubat, M., Matwin, S.: Addressing the Curse of Imbalanced Training Sets: One Sided Selection. Proceedings of the Fourteenth International Conference on Machine Learning, 179–186, Nashville, Tennesse. Morgan Kaufmann (1997)

[15] Japkowicz, N.:The Class Imbalance Problem: Significance and Strategies. Proceedings of the International Conference on Artificial Intelligence (IC-AI'2000): Special Track on Inductive Learning, 111–117, Las Vegas, Nevada, USA (2000)

[16] Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter **6**(1), 20–29 (2004)

[17] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Oversampling TEchnique. Journal of Artificial Intelligence Research **16**, 321–357 (2002)

[18] Cost S., Salzberg S.: A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features. Machine Learning **10**(1), 57–78 (1993)

[19] Gower, J.C.: A General Coefficient of Similarity and Some of Its Properties. Biometrics **27**(4), 857–871 (1971)

[20] Sulc Z., Rezankova H.: Comparison of Similarity Measures for Categorical Data in Hierarchical Clustering. Journal of Classification **35**(1), 58–72 (2019)

[21] Wishart D.: k-Means Clustering with Outlier Detection, Mixed Variables and Missing Values. In: Schwaiger M. and Opitz O. (edts) Exploratory Data Analysis in Empirical Research, Springer, 216–226 (2003)

[22] Dua D., Graff C.: UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science (2019). *https://archive.ics.uci.edu/ml/*

[23] Zejin D.: Diversified Ensemble Classifiers for Highly Imbalanced Data Learning and their Application in Bioinformatics. Dissertation, Georgia State University (2011)

[24] Cortez P., Cerdeira A., Almeida F., Matos T., Reis J.: Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems **47**(4), 547–553 (2009)

[25] Liu H., Setiono R.: Feature selection via discretization. IEEE Transactions on Knowledge and Data Engineering **9**(4), 642–645 (1997)

[26] Breiman, L.: Random forests. Machine Learning, vol. 45(1), pp. 5 – 32 (2001)

[27] Ferri C., Flach P., Orallo J., Lachice N. (edts). ECAI'2004 First Workshop on ROC Analysis in Artificial Intelligence (2004)

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