Neurocomputing ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Neurocomputing

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

A method for resampling imbalanced datasets in binary classification tasks for real-world problems

Silvia Cateni*, Valentina Colla, Marco Vannucci

TeCIP Institute, Scuola Superiore Sant'Anna, Pisa, Italy

ARTICLE INFO

Article history:
Received 25 October 2012
Received in revised form
19 February 2013
Accepted 30 May 2013

Keywords:
Oversampling
Undersampling
Imbalanced dataset

ABSTRACT

The paper presents a novel resampling method for binary classification problems on imbalanced datasets. Imbalanced datasets are frequently found in many industrial applications: for instance, the occurrence of particular product defects, the diagnosis of severe diseases in a series of patients or machine faults are rare events whose detection is of utmost importance. In this paper a new resampling method is proposed combining an oversampling and an undersampling technique. Several tests have been developed aiming at assessing the efficiency of the proposed method. Four classifiers based, respectively, on Support Vector Machine, Decision Tree, labelled Self-Organizing Map and Bayesian Classifiers have been developed and applied for binary classification on the following four datasets: a synthetic dataset, a widely used public dataset and two datasets coming from industrial applications. The results that have been obtained in the tests are presented and discussed in the paper; in particular, the performances that are achieved by the four classifiers through the proposed novel resampling approach have been compared to the ones that are obtained, without any resampling, through a widely applied and well known resampling technique, i.e. the classical SMOTE approach, and through another approach coupling informed SMOTE-based oversampling and informed clustering-based undersampling.

1. Introduction

In the literature and in real-world problems, binary classification often needs to be performed on a so-called *imbalanced dataset*, where the samples belonging to one of the two classes to detect are far less numerous than the other ones. Often, the rare patterns are also difficult to separate from the most frequent ones. However, in many cases, rare events within a dataset are the most important ones to detect, such as, for instance, in the medical field for disease diagnosis or in the industrial field for fault diagnosis.

The class imbalance negatively affects the performances of all the common classifiers, such as support vector machines, decision trees and neural networks. Actually, the standard versions of these classifiers are designed to optimize the overall performance [1] on the whole dataset. As a consequence, they correctly classify most of the patterns belonging to the frequent class but they obtain a very poor performance on the other ones. This undesirable result is caused by the rarity of the minority class, which compromises the correct characterization of its patterns and prevents the correct separation from the majority class. As shown in many review works in the literature [2,3], any classifier is sensitive to the effect of imbalance in datasets, which affects not only the correct identification of rare patterns but also the rate of overall correct classifications.

Due to its significance, the problem of classifying imbalanced datasets has been faced in many literature works. The methods developed for coping with class imbalance can be divided into two main groups on the basis of the different approaches to the problem: the *internal methods* deal with the development of new algorithms expressly designed to cope with uneven datasets, while the *external methods* operate on the dataset to be used for the training of the classifier in order to attenuate the imbalance of the dataset.

Among internal methods it is worth to mention a class of approaches based on the use of different misclassification weights during the training of the classifiers [4]. In this framework the costs of misclassifying a rare pattern are higher with respect to other kinds of errors in order to encourage their correct detection. Support Vector Machines (SVMs) have also been employed for facing imbalanced datasets: in [5] an ensemble of SVM is used, while in [6] a variation of traditional SVM architecture (the so-called v-SVM) is used for the detection of rare samples: in this latter case the v-SVM is trained by using only rare patterns so as to allow their sole identification in the classification stage. A particular kind of radial basis function has been developed and tested in [7] where hyper-rectangular activation function neurons are used in the hidden layer in order to achieve more precision in the

* Corresponding author at: Valentina Colla, PERCRO laboratory, Via Alamanni 13D, 56010 San Giuliano Terme, Pisa, Italy. Tel.: +39 050 882507.
E-mail addresses: s.cateni@sssup.it (S. Cateni), colla@sssup.it (V. Colla), mvannucci@sssup.it (M. Vannucci).

0925-2312/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.05.059

detection of the boundary of the input space regions reserved to each class. Finally in [8] the approach based on the use of an uneven cost matrix is combined to a self-organizing map and a fuzzy inference system obtaining interesting results in the literature and industrial datasets.

The external methods modify the distribution of rare and frequent patterns within the database in order to favour the detection of the rare ones. This operation is called *resampling* and aims at increasing the rate of rare samples by creating a new dataset from the original one. This rebalance can be obtained by both removing samples belonging to the frequent class and adding samples to the rare one. These two techniques, which can be eventually combined such as in [9], are called *undersampling* and *oversampling*. Through undersampling, some frequent samples are removed from the original dataset until a predetermined balance ratio is reached. Data to be removed can be selected in a random way or, more fruitfully, according to particular criteria such as, for instance, the removal of those patterns lying on the external regions of the input space. This latter technique effectively reduces the input space area assigned to the frequent class by the classifier and, as a consequence, favours the correct classification of rare patterns [10].

On the other hand, the oversampling methods counteract the original database unbalance by adding new samples of the rare class. This operation can be done by both replicating existing rare patterns and creating new ones in a particular region of the input space. The replication can be performed in a random way or by selecting those patterns lying on the boundary between rare and frequent samples in order to force the classifier to allocate such regions of the space to the rare class. The main criticality of the replication-based oversampling is that it does not add any new informative content to the dataset but it simply rebalances the ratio between rare and frequent samples. In order to overcome this drawback, several techniques creating new synthetic rare samples in those regions where they likely could be have been developed.

In this paper a new resampling method combining oversampling and undersampling is presented which is named *Similarity-based UnderSampling and Normal Distribution-based Oversampling* (SUNDO) methods. This method combines the two approaches: for the oversampling phase, it places new samples where they likely could be and avoids to place them close to frequent samples; moreover it employs an innovative undersampling technique.

The paper is organized as follows. Section 2 presents an overview of the literature survey, the general method is described in Section 3 and the details of oversampling and undersampling are provided in Sections 3.1 and 3.2, respectively, while Section 3.3 depicts through a simple example the advantages of SUNDO over some widely applied traditional approaches. The results achieved by SUNDO on several public and industrial imbalanced datasets are reported and discussed in Section 4, where they are compared to the ones achieved by other standard resampling techniques. Finally, in Section 5 some conclusions are drawn.

This paper is an extended version of the paper entitled “Novel resampling method for the classification of imbalanced datasets for industrial and other real-world problems” which was presented by the same authors at the 11th International Conference on Intelligent Systems Design and Applications ISDA 2011.

2. State of art

Imbalanced datasets occur in several real-world applications where the decision system is aimed to detect rare cases such as telecommunication customers, detection of oil spills in radar images, text classification and many others [11–13]. There are many different forms of resampling such as random oversampling, random undersampling, directed oversampling, directed undersampling and combinations of the two techniques.

Random undersampling [14] balances the minority class through the random elimination of some samples belonging to the majority class. This approach has been criticized in some papers because potentially useful data could be removed leading to a huge information loss [15–19]. Another disadvantage with this method is that often the aim of machine learning is for the classifier to assess the probability distribution of the target.

Kubat and Matwin [20] proposed a selective undersampling method keeping the original samples belonging to the minority class. They used a geometric mean as a performance test for the classifier. In particular the samples belonging to the minority class were categorized into four parts: noise overlapping the minority class, borderline samples, redundant samples and safe samples. Borderline samples are selected through the *Tomek link* notion [21]. Given two samples s_i and s_j belonging to different classes, a pair (s_i, s_j) is called a Tomek link if there is not a sample s_w such that $d(s_i, s_w) < d(s_i, s_j)$ or $d(s_j, s_w) < d(s_i, s_j)$ where $d(\cdot,\cdot)$ is the *distance* between the two considered samples. When two samples form a Tomek link, one of these samples is noise or both samples are borderline. The idea of this undersampling method is to delete samples belonging to the majority class that are distant from the decision border in order to eliminate samples less relevant for learning.

Random oversampling consists in balancing the minority class by randomly replicating its samples. The main disadvantage of this approach is the enhancement of occurring overfitting, since it creates coincident samples. Moreover the overfitting due to the addition of new samples increases the computational requirements when the dataset is already large and imbalanced.

SMOTE [22] is the most famous oversampling method. The basic idea of this algorithm is to place the new samples along the lines connecting existing rare samples. SMOTE has been widely used in literature works and often combined with other techniques such as in [23], where it is coupled with an ad hoc version of Support Vector Machine (SVM). Moreover several other methods follow the basic idea of recreating new rare samples in those regions of the space where they would be more likely to be found. For instance in [24] new samples are created in the surrounding areas of rare observations.

3. The proposed resampling method

The method that is proposed here combines undersampling and oversampling techniques in order to obtain a balanced dataset without significant loss of information and without the addition of a great number of synthetic patterns.

Firstly the imbalanced dataset is divided into a training set (75%) and a validation set (25%) maintaining the same proportion between the two classes. Then the training set is divided into two subsets containing the samples belonging to the minority class and to the majority class. Let us suppose that n_1 is the number of samples belonging to the minority class subset of the training dataset and n_0 is the number of the samples included in the majority class subset of the training dataset. Moreover let us suppose that k_0 and k_1 are, respectively, the target percentages of samples to be achieved by the majority and minority classes: obviously $k_0 + k_1 = 1$. These parameters may vary from their initial value of imbalance until to obtain a perfect equilibrium between the two classes, which means that they have the same number of samples. The target of 50% of samples has been chosen as a limit value for the minority class, as the assumption has been made that the majority class should not lose its primacy, which reflects

Please cite this article as: S. Cateni, et al., A method for resampling imbalanced datasets in binary classification tasks for real-world problems, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2013.05.059

the actual situation that the corresponding event is more likely. Actually the proposed approach can also be applied to overturn the proportion between the two classes (i.e. to obtain a final database where the minority class has more than 50% of the samples), which is however not considered useful in this work. On the other hand, one might want to keep a higher importance of the majority class as it is considered more important and this is allowed by the proposed method. In fact k_0 and k_1 are parameters to be set by the user and the number N of samples to be removed or added is calculated as follows:

$$N = \text{round}[(k_1 \cdot n_0) - (k_0 \cdot n_1)] \tag{1}$$

Once the majority class has been undersampled and the minority class has been oversampled, two subsets are built, which are afterwards merged in order to create the new balanced training set. This new training set is used to train the chosen classifier while the (imbalanced) validation set, that has been initially created, is used to evaluate the performance of the classifier. A scheme of SUNDO is depicted in Fig. 1.

3.1. Oversampling

In this paper a novel oversampling approach is proposed, where the minority class is oversampled by adding synthetic samples in the neighborhood of each "original" sample $\mathbf{x}_i^R (1 \le i \le n_1)$ (the superscript R stands for "Rare"). Let us suppose that the patterns to classify are m-dimensional. The new patterns are generated as samples of n_1 stochastic processes, each one characterized by a Gaussian Probability Density Function $f_i(\mathbf{x})$, whose centers coincide with the corresponding original samples $\mathbf{x}_i^R$ and whose covariance matrices $\mathbf{\Sigma}_i$ are diagonal. The non-null entries of $\mathbf{\Sigma}_i$ can be chosen as proportional to the corresponding entry of $\mathbf{x}_i^R$, namely

$$\mathbf{\Sigma}_{i(p,q)} = \begin{cases} 0 & \text{for} \quad p \neq q \\ \beta \cdot (x_{i,p}^R)^2 & \text{for} \quad p = q \end{cases} 1 \le p, q \le m \tag{2}$$

where β is a proportionality factor. However, this choice is not always suitable for all the cases (let us consider, for instance, a database where the null pattern (or a pattern with very small entries with respect to the other rare ones) is rare. An alternative choice could be $\mathbf{\Sigma}_{i(p,p)} = \beta \cdot \sigma_p^R$, where σ_p^R is the standard deviation of the p-th component of the rare patterns.

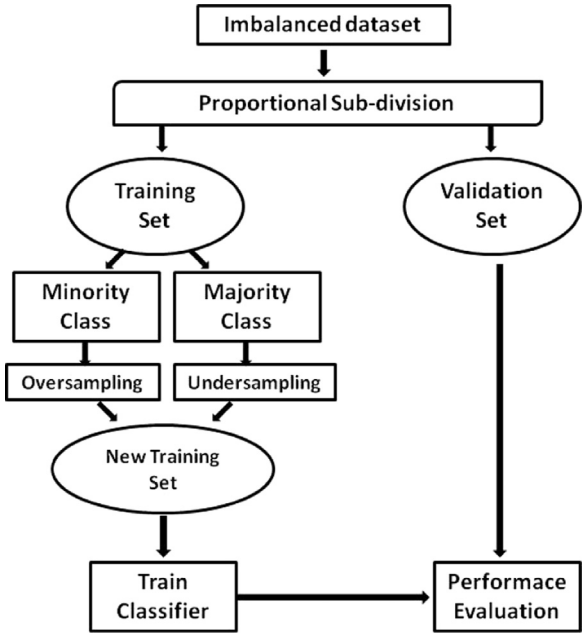

Fig. 1. Scheme of the SUNDO method.

The number K_R of new patterns to generate for each original sample $\mathbf{x}_i^R$ should result from a trade-off between the computational burden and the probability to generate suitable synthetic patterns, which both increase with K_R. Obviously the relation $K_R \cdot n_1 > N$ must hold. A constant value for K_R can be used or it can be set as $K_R \simeq \alpha N/n1$, where α is a proportionality factor to fix.

For all the new synthetic samples $\mathbf{x}_j^S (1 \le j \le n_1 \cdot K_R)$ the following index is evaluated:

$$I_j^S = \min_{1 \le k \le n_0} \sqrt{\| \mathbf{x}_j^S - \mathbf{x}_k^O \|^2} \tag{3}$$

where $\mathbf{x}_k^O$ is the k-th sample of the majority class. The N synthetic samples which have the highest values of I_j^S (namely which are considered to be farthest from the majority class according to the metric defined in Eq. (3)) are selected to oversample the minority class, paying attention that no rare patterns remain isolated if $N > n_1$.

3.2. Undersampling

We propose an undersampling approach based on a novel methodology that is capable to detect anomalous data, i.e. non-coherent associations between input and output patterns [25]. Our approach is based on the comparison between two distance matrices associated to the input and output patterns and it has been originally applied to detect samples in a dataset where similar values of the input features are associated to different output values. We used this approach in a different way, namely to point out the similarity between inputs which belong to the same class (that is the majority class).

Let us consider only the subset of the classifier training dataset, which contains the n_0 data belonging to the majority class. These data can be represented in a matrix formulation, i.e. through a matrix $\mathbf{X}^O$ that has n_0 rows and m columns. The different steps required by the algorithm (which are described in details in [25]) can be summarized as follows:

1. Firstly a static normalization of the entries of $\mathbf{X}^O$ with respect to their own maximum values is performed and a transformed matrix $\boldsymbol{\pi}$ is obtained, where the entry at row i and column k is evaluated as follows:
$$\pi_{i,k} = \frac{x_{i,k}^O}{\max_{1 \le j \le n_0}\{x_{j,k}^O\}} \tag{4}$$
2. An *Euclidean distance* is computed between each pair of columns of $\boldsymbol{\pi}$ by obtaining a symmetrical square *distance matrix* $\mathbf{D_I}$. The entry of such matrix at row p and column q corresponds to the distance between the p-th and q-th columns of $\boldsymbol{\pi}$ (i.e. π^p and π^q) and is evaluated as follows:
$$\mathbf{D}_{\mathbf{I}_{p,q}} = \frac{1}{n_0}(\pi^p - \pi^q) \cdot (\pi^p - \pi^q)^T \tag{5}$$
Clearly the entries located on the main diagonal are null. Due to the symmetry, the following step 3 can be performed considering only the upper triangle of the matrix. $\mathbf{D_I}$ is named *dissimilarity matrix*.
3. The smaller the $\mathbf{D}_{\mathbf{I}_{p,q}}$, the more "similar" the two patterns X_p and X_q. The couples of patterns are ranked according to this similarity index and in the N most similar couples one of the patterns (the choice is randomly performed) is eliminated without significant loss of information on the original class distribution.

Please cite this article as: S. Cateni, et al., A method for resampling imbalanced datasets in binary classification tasks for real-world problems, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2013.05.059

3.3. Advantages over traditional approaches

SUNDO is firstly compared to the SMOTE-based approach that combines informed oversampling of the minority class with random undersampling of the majority class. SMOTE randomly selects some of the k-nearest minority samples and the selected neighbours. The number of the neighbours depends on the amount of samples to be added to reach a certain dimension of the minority class. An important limitation of SMOTE is that synthetic samples are created without considering the majority class. In order to avoid this problem, SUNDO adds new samples with regard to the majority class. Moreover undersampling-based approaches can be dangerous because the random elimination of samples can lead to the loss of relevant information about the original data. Therefore an under-sampling algorithm is proposed, which selects the most redundant samples.

Figs. 2 and 3 depict through an example the advantage of SUNDO with respect to SMOTE. The oversampling procedure adds 3 samples: A–B–C. All the new samples that are created by our method lie near the minority class but quite far from the majority class. As far as SMOTE is concerned, sample A is almost equally distant from samples belonging to both classes, while samples B and C lie quite near to the majority class. On the other hand, our undersampling procedure deletes samples with the highest density of neighbours belonging to the same class, as it tends to minimize the information loss. The random oversampling can eliminate some samples that are very important with respect to the original distribution data, i.e. whose elimination causes a significant change in the distribution of the data belonging to the majority class. In this example, sample E is isolated and should be kept, but the random undersampling eliminates it.

Moreover, SUNDO has also been compared to a resampling procedure (in the following referred as *SCB approach*) having similar features, namely where informed undersampling is coupled to informed SMOTE-based oversampling. In particular, the selected informed undersampling algorithm is a standard clustering-based approach that is described in [26] and it obviously requires a preliminary clustering of all the available data into a given number K of clusters. Afterwards, for each cluster i ($1 \le i \le K$) the algorithm computes the number of samples belonging to the majority and the minority class (indicated, respectively, as N_i^O and N_i^R) and calculates the number of samples of the majority class N_i to be eliminated in the i-th cluster according to the following formula:

$$N_i = m \cdot n_1 \cdot \frac{\frac{N_i^O}{N_i^R}}{\sum_{i=1}^{K} \frac{N_i^O}{N_i^R}} \tag{6}$$

where m is the integer approximation of the ratio n_0/n_1. The parameter K heavily affects the performance of the algorithm: in many literature works this parameter is heuristically chosen, i.e. many attempts are made with different values of K and the best result is chosen. This is a very time consuming procedure, which makes this approach far more computationally expensive with respect to SUNDO. In order to overcome this problem and allow a fair comparison with the proposed approach, here K has been automatically determined according to a procedure firstly introduced in [27] and widely adopted (see for instance [28]). According to this method, two parameters are introduced, which are named *intra-cluster distance* D_{IAC} and *inter-cluster distance* D_{IRC}, that are defined, respectively, as the mean value of distances among the samples in each cluster and their respective clustering centers and the minimum distance between the clusters centers. The optimal number of clusters is thus determined as the integer number which maximizes the so-called *validity measure* that is defined as the ratio D_{IAC}/D_{IRC}.

Fig. 4 depicts how the SBC approach works in the previously introduced exemplar case: the undersampling procedure deletes three samples D, E and F while oversampling procedure adds three samples A, B and C. Sample F is deleted also by SUNDO, sample C is close to one of the samples eliminated by SUNDO, while sample D is isolated and thus should be kept: this is due to the fact that the

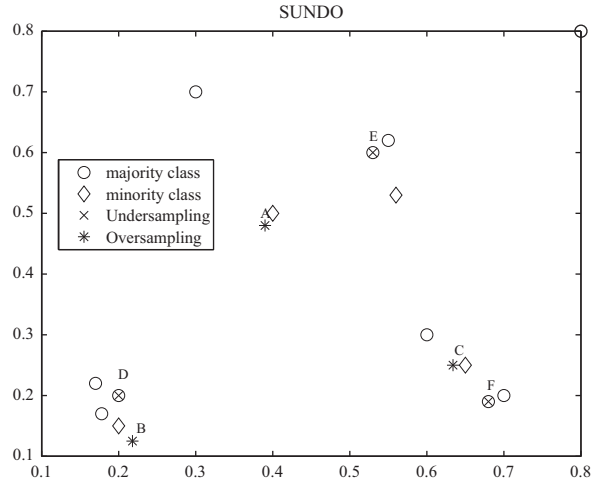

Fig. 2. Example of application of the SUNDO method.

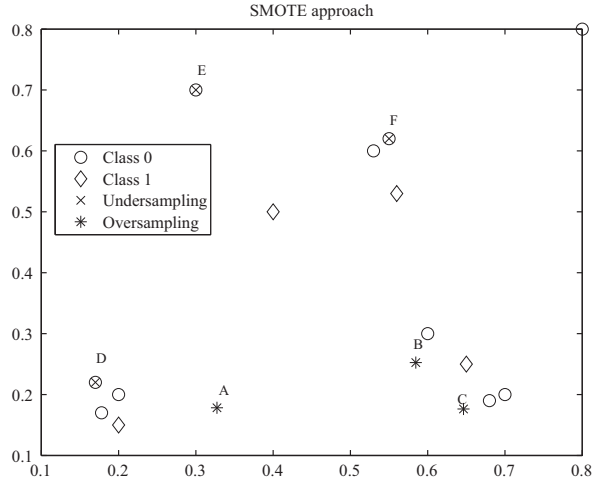

Fig. 3. SMOTE approach and random undersampling.

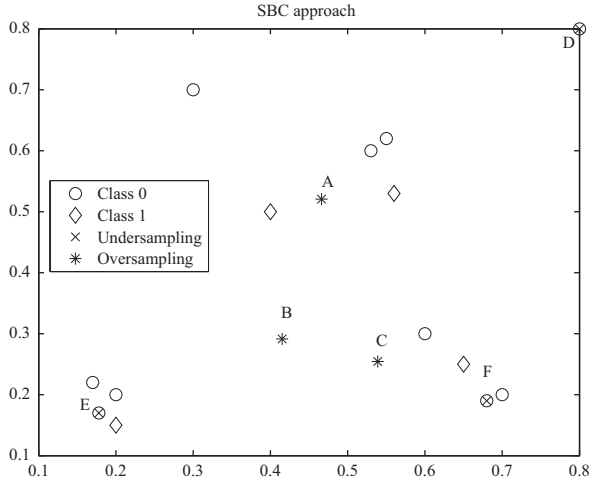

Fig. 4. Example of application of the SBC approach.

Please cite this article as: S. Cateni, et al., A method for resampling imbalanced datasets in binary classification tasks for real-world problems, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2013.05.059

SBC approach computes the number of the samples to be eliminated for each cluster but afterwards randomly chooses the patterns to eliminate by not taking into account their information contents. Conversely, SUNDO recognizes that the elimination of an isolated sample like D can eliminate useful information and heavily alter the data distribution and thus keeps it. On the other hand, among the samples added by the SBC approach, sample C is closer to the majority class than to the minority one, while A and especially B are quite isolated and their presence changes the data distribution.

4. Test of the method

In order to prove the effectiveness of the proposed method we developed experiments on four datasets where a binary classification must be performed. These datasets are different in their sizes as well as in the proportion between the two classes, so as to give several domains for the resampling method. Moreover four machine learning-based algorithms have been used for the classification, i.e. a Support Vector Machine (SVM) [29], a Decision Tree (DT) [30], a labelled Self-Organizing Map (SOM) [31] and a Bayesian classifier [32].

In particular, the adopted SVM has a linear kernel function and quadratic programming is applied to find the separating hyperplane, while the DT is constructed through the CART algorithm.

The Bayesian Classifier estimates the parameters of a probability distribution assuming that inputs are independent. Subsequently the method calculates the posterior probability of validation samples belonging to each class. The samples are classified according to the largest posterior probability.

The labelled SOM, compared to the standard SOM, contains class information concerning each neuron it is composed of. Firstly the standard unsupervised training of SOM is achieved so as to produce a cluster map maintaining topology and distribution of the training data. In a second step the training data are fed again as inputs to the net with the class to which they are associated in order to allocate a cluster to each corresponding class from the majority of the samples which have excited each neuron.

Other classifiers could be more suitable depending on the features of the treated database, but, as the focus here is on the efficiency of the resampling techniques, we preferred to use standard and widely adopted classifiers trained through standard procedures, which can provide reasonable results also with moderately imbalanced datasets.

In order to demonstrate the efficiency of the proposed method, the classification performance of all the classifiers has been compared in four cases: without any resampling, by using SMOTE as over-sampling method and a random undersampling, by using SMOTE as oversampling and a clustering-based undersampling approach and by adopting the SUNDO method.

In all the tests, the datasets have been balanced, decreasing the final imbalance until the same number of samples for both classes is obtained. For SUNDO and for the resampling method which exploits SMOTE-based oversampling method and random undersampling this target has been achieved by removing from the majority class and adding to the minority one the number of patterns calculated according to Eq. (1). Moreover in the SUNDO oversampling procedure the covariance matrices are computed according to Eq. (2) with $\beta=0.05$ and the number of candidate synthetic data to generate (before the final selection) has been set as $K_R \simeq 10N/n1$. On the other hand, for the SCB approach, the total number of data of the majority class that are removed is represented by $\sum_{i=1}^{K} N_i$ where N_i is calculated from Eq. (6). Thus the number of samples which are added to the minority class through the SMOTE oversampling is calculated according to the desired final imbalance between classes as $(k_1/k_0)(n_0-\sum_{i=1}^{K} N_i)$.

The training of the classifiers has been performed by using the obtained resampled training datasets. Moreover, according to standard procedure, part of the training data (20%) has been used as *test set* in order to avoid overfitting. In particular, when SOM and SVM based classifiers have been used, test data have been exploited for early stopping while in the case of DT such data have been employed for the pruning phase.

For each dataset, 100 trials have been conducted and the average value of the performance achieved on the validation dataset (which contains the 25% of the available data) is shown in the following subsections. The performances of the classifiers are measured in terms of the following indexes:

- *True Positive* (*TP*): percentage of correctly classified rare patterns.
- *True Negative* (*TN*): percentage of correctly classified frequent patterns.
- *False Positive* (*FP*): percentage of frequent patterns incorrectly classified as rare patterns.
- *False Negative* (*FN*): percentage of rare patterns incorrectly classified as belonging to the frequent class.
- *Global Accuracy* (*ACC*): percentage of correctly classified patterns which is defined as the ratio between the number of correctly classified patterns (belonging to both classes) and the total number of the patterns.

Moreover a novel index [33] is calculated in order to encourage the detection of the salient events rather than the correctness of the classification and, at the same time, it is useful to avoid false positive patterns. The index is evaluated as follows:

$$\mathrm{IDX}=\frac{\gamma\, TP-FP}{ACC+\mu\, FP} \tag{7}$$

where γ and μ are two empirical parameters set in our experiments to 0.8 and 0.2, respectively. In formula (7), *TP* on the numerator specifies the direct linear relation between number of rare events detected and performance goodness while the $-FP$ term penalizes the generation of false alarms. This latter figure is used on the denominator of the fraction as well with the same aim (the higher, the worse); in addition the *ACC* rate on the denominator is used not to favour performances based on the overall classification correctness. The *IDX* index has been expressly designed for the assessment of classifiers performance when coping with uneven datasets, therefore its main focus is not on the overall rate of correct classifications but particular attention is paid to the detection of rare events and, at the same time, to the avoidance of a critical rate of false alarms. The IDX index is used to evaluate the classifiers performance considering the particular framework determined by the peculiar nature of uneven datasets and the use of this index, obviously, does not influence the classifiers performance as it is used *a posteriori* for its assessment.

In the following, according to this latter consideration, the results achieved by the employed classifiers will be compared – independently, for each one of the tested unbalance rates – mostly in terms of the introduced *IDX* index; nevertheless, for completeness, other figures such as the overall accuracy, the false alarms and the rare patterns detection rates will be reported as well.

4.1. Synthetic dataset

A dataset including 134 samples and two independent variables has been created. The minority class contains 14 samples, with an unbalance of about 10%. Fig. 5 depicts the synthetic dataset that is used in the experiments.

Please cite this article as: S. Cateni, et al., A method for resampling imbalanced datasets in binary classification tasks for real-world problems, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2013.05.059

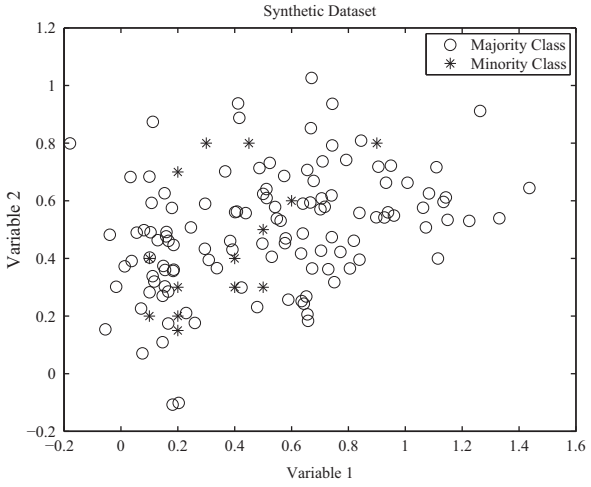

Fig. 5. Synthetic dataset.

The 3rd, 17th, 31st and 46th rows of Table 1 show the classification results achieved by each classifier on the synthetic dataset without applying any re-sampling method: as it was expected, without any re-sampling techniques, all the classifiers attribute the validation samples to the majority class. Table 1 also shows the classification results using the three different re-sampling methods for several indices of imbalance and applying the four selected classifiers.

The results obtained on the synthetic dataset demonstrate that SUNDO provides the best results with respect to both the SMOTE and SBC approaches in all considered cases. The best performance is obtained using the SVM classifier when the two classes are equally balanced. It must be noticed that, considering only the global accuracy, one could erroneously believe that the best performance is achieved by adopting the original dataset without any re-sampling operations. Obviously global accuracy is not a correct performance index, as, in most practical applications (e.g. fault diagnosis), the classification of rare patterns is far more important than the correct classification of the patterns belonging to the majority class. This fact underlines the importance of introducing a novel index for evaluating the effectiveness of the re-sampling techniques in very imbalanced datasets, which takes into account both the correct classifications and the classification errors, but with a different weight of such errors (i.e. depending on the classes that are erroneously selected).

4.2. Classical problem: Wisconsin Breast Cancer Dataset

The Wisconsin Breast Cancer Dataset (WBCD) is extracted from the UCI repository [34]. The data (699 samples) refer to biopsies conducted on patients that are suspected of breast cancer. Each sample includes nine attributes related to analysis of the biopsy image. The target indicates the benign or malignant nature of the tumor. The minority class includes 241 samples with an imbalance of about 34% with respect to the majority class. As shown in the 3rd, 11th, 19th and 27st rows of Table 2, the four classifiers achieve an overall accuracy, which is comparable to the methods that are applied in the literature to face the same problem (see for instance [35,36]). Table 2 also shows the results provided by the several classifiers. The performance of different classifiers does not improve significantly after a re-sampling, probably due to the fact that the considered dataset is not very unbalanced. The best performance is achieved by using SVM and SOM classifiers while the worst performance is obtained through the DT classifier.

In order to put into evidence the potential of the proposed approach, a further test has been performed by increasing the unbalance of the original dataset through a random deletion of some samples belonging to the majority class. Tables 3 and 4 show the results that are achieved by the different classifiers with a classes unbalance of about 20% and 10%, respectively, without any resampling technique and by applying the SMOTE, SBC and SUNDO re-sampling techniques with a different final imbalance

Table 1
Classification results on the synthetic dataset without any resampling and by applying SMOTE, SBC and SUNDO re-sampling techniques with the four different classifiers.

Method	k_1	TP (%)	TN (%)	FP (%)	FN (%)	ACC (%)	IDX
SVM							
None	–	0	100	0	100	90.9	0
SMOTE	0.2	0	100	0	100	90.9	0
	0.3	18.7	76.1	23.9	81.3	70.9	−0.12
	0.4	56.1	55.8	44.2	43.9	55.8	0.01
	0.5	57	66.6	33.4	43	65.7	0.17
SBC	0.2	0	100	0	100	90.9	0
	0.3	0	90	10	100	81.8	−0.12
	0.4	33.3	56.7	43.3	56.7	54.5	−0.42
	0.5	33.3	66.7	33.3	66.7	63.6	−0.21
SUNDO	0.2	0	100	0	100	90.9	0
	0.3	22.8	74.1	25.9	77.2	69.4	−0.1
	0.4	68	69.4	30.6	32	69.3	0.32
	0.5	96.7	75.3	24.7	3.3	76.6	0.64
DT							
None	–	33.3	95.2	4.8	66.7	89.6	0.24
SMOTE	0.2	32.1	85.5	14.5	67.9	80.6	0.13
	0.3	46.5	66.7	33.3	53.5	64.8	0.05
	0.4	42.2	72.8	27.2	57.8	70.0	0.08
	0.5	45.6	70.5	29.5	54.4	80.5	0.08
SBC	0.2	0	73.3	26.7	100	66.7	−0.37
	0.3	33.3	76.7	23.3	66.7	72.7	−0.03
	0.4	33.3	60	40	66.7	57.8	−0.34
	0.5	66.7	63.3	36.7	33.3	63.6	0.02
SUNDO	0.2	45.3	84.1	15.9	54.7	80.6	0.24
	0.3	38	78.7	21.3	62	75	0.11
	0.4	70.6	74.1	25.9	29.4	73.8	0.39
	0.5	95	86.3	13.7	5	87.1	0.69
SOM							
None	–	0	100	0	100	90.9	0
SMOTE	0.2	3.5	96.4	3.6	96.5	88	0
	0.3	15.9	87.3	12.7	84.1	80.8	0
	0.4	36.5	75.3	24.7	63.5	71.7	0.05
	0.5	54.3	63.4	36.6	45.7	62.6	0.1
SBC	0.2	0	90	10	100	81.8	−0.12
	0.3	0	90	10	100	81.8	−0.12
	0.4	0	93.3	6.7	100	84.8	−0.07
	0.5	66.7	43.3	56.7	33.3	45.4	−0.46
SUNDO	0.2	14.4	88	12	85.6	81.3	0
	0.3	33.2	73.9	26.1	66.8	70.2	0
	0.4	63.3	81.3	18.7	36.7	79.7	0.38
	0.5	65.9	65.9	34.1	34	65.9	0.26
BAYESIAN							
None	–	0	100	0	100	90.9	0
SMOTE	0.2	3.8	95.2	4.7	96.2	86.9	0
	0.3	26.4	85.9	14.1	73.6	80.5	0.08
	0.4	59.3	59.9	40.1	40.7	59.9	0.1
	0.5	71	56	44	29	57.4	0.19
SBC	0.2	0	86.7	13.3	100	78.8	−0.16
	0.3	33.3	63.3	36.7	66.7	60.61	−0.27
	0.4	33.3	43.3	56.7	66.7	42.42	−0.79
	0.5	66.7	70	30	33.3	69.7	0.14
SUNDO	0.2	13.7	89.5	10.5	86.3	82.6	0
	0.3	34.6	78.1	21.9	65.4	74.1	0.08
	0.4	53	72.1	27.9	47	70.4	0.19
	0.5	78	58.1	41.9	22	59.9	0.3

Please cite this article as: S. Cateni, et al., A method for resampling imbalanced datasets in binary classification tasks for real-world problems, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2013.05.059

Table 2
Classification results on the UCI WBCDB dataset without any resampling, by applying SMOTE, SBC and SUNDO re-sampling techniques with the four different classifiers.

Method	k_1	TP (%)	TN (%)	FP (%)	FN (%)	ACC (%)	IDX
SVM							
None	–	96.2	97.2	2.8	3.8	96.8	0.76
SMOTE	0.4	95.4	97.3	2.7	4.6	96.5	0.76
	0.5	96.8	96.6	3.4	3.2	96.7	0.76
SBC	0.4	95	98.2	1.8	5	97.1	0.74
	0.5	100	95.6	4.4	0	97.1	0.75
SUNDO	0.4	95	97.5	2.5	5	96.7	0.76
	0.5	97.6	97	3	2.4	97.2	0.77
DT							
None	–	93	95	5	7	94.3	0.73
SMOTE	0.4	92.7	95.2	4.8	7.3	94.3	0.73
	0.5	92.3	94.5	5.5	7.7	93.7	0.72
SBC	0.4	90	97.4	2.6	10	94.8	0.69
	0.5	95	93.9	6.1	5	94.2	0.70
SUNDO	0.4	92.9	95	5	7.1	94.3	0.73
	0.5	94	94.3	5.7	6	94.2	0.73
SOM							
None	–	97.9	95.7	4.3	2	96.5	0.76
SMOTE	0.4	97.1	96.1	3.9	2.9	96.4	0.76
	0.5	97.9	94.5	5.5	2.1	95.7	0.75
SBC	0.4	100	95.6	4.4	0	97.1	0.75
	0.5	98.3	94.7	5.3	1.7	96	0.73
SUNDO	0.4	97.8	95.9	4.1	2.1	96.6	0.76
	0.5	98.1	96	4	1.9	96.7	0.76
BAYESIAN							
None	–	95	97.5	2.5	5	96.6	0.76
SMOTE	0.4	95.4	97	3	4.6	96.4	0.75
	0.5	95.9	97	3	4.1	96.6	0.76
SBC	0.4	93.3	98.2	1.8	6.7	96.5	0.73
	0.5	100	92.1	7.9	0	94.8	0.72
SUNDO	0.4	96.2	96.5	3.5	3.8	96.4	0.75
	0.5	96.7	96.3	3.7	3.3	96.4	0.76

Table 3
Classification results on the modified (20%) UCI WBCDB dataset without any resampling and by applying SMOTE, SBC and SUNDO re-sampling techniques with the four different classifiers.

Method	k_1	TP (%)	TN (%)	FP (%)	FN (%)	ACC (%)	IDX
SVM							
None	–	90	97	3	10	96	0.71
SMOTE	0.3	91.9	97.4	2.5	8.1	96.3	0.73
	0.4	92	97.6	2.4	8	96.5	0.73
	0.5	93.6	96.8	3.2	6.4	96.2	0.74
SBC	0.3	93.3	94.7	5.3	6.7	94.4	0.69
	0.4	83.3	98.2	1.8	16.7	95.1	0.65
	0.5	96.7	94.7	5.3	3.3	95.1	0.72
SUNDO	0.3	96.1	97.6	2.4	8.9	96.3	0.73
	0.4	93.8	97.3	2.7	6.2	96.6	0.74
	0.5	94.7	97	3	5.3	96.6	0.75
DT							
None	–	85.7	96.2	3.8	14.3	94	0.68
SMOTE	0.3	87.3	96	4	12.6	94.2	0.69
	0.4	88.7	95.5	4.5	11.3	94.1	0.7
	0.5	86.6	94	6	13.4	92.9	0.67
SBC	0.3	90	98.2	1.8	10	96.5	0.7
	0.4	83.3	99.1	0.9	16.7	95.8	0.66
	0.5	83.3	95.6	4.4	16.7	93	0.62
SUNDO	0.3	88	96.5	3.5	12	94.7	0.7
	0.4	89.1	96.4	3.6	10.9	94.9	0.71
	0.5	88.6	96.2	3.8	11.4	94.9	0.7
SOM							
None	–	95.2	96.6	3.4	4.8	96.3	0.75
SMOTE	0.3	94.5	96.4	3.5	5.5	96	0.74
	0.4	94.4	96.5	3.5	5.6	96.1	0.74
	0.5	96.2	96.2	3.8	3.8	96.2	0.75
SBC	0.3	96.7	93	7	3.3	93.7	0.70
	0.4	96.7	98.2	1.8	3.3	97.9	0.75
	0.5	96.7	98.2	1.8	3.3	97.9	0.75
SUNDO	0.3	96.5	96.2	3.8	3.5	96.2	0.76
	0.4	97.4	95	5	2.6	95.5	0.76
	0.5	98.5	95.9	4.1	1.5	96.5	0.77
BAYESIAN							
None	–	93.5	97.4	2.6	6.5	96.5	0.74
SMOTE	0.3	93.9	97.3	2.7	6.1	96.6	0.75
	0.4	94.8	96.7	3.3	5.2	96.3	0.75
	0.5	94.1	97.2	2.8	5.9	96.6	0.74
SBC	0.3	96.7	98.3	1.7	3.3	97.9	0.75
	0.4	93.3	99.1	0.9	3.3	97.9	0.74
	0.5	96.7	95.6	4.4	3.3	95.8	0.73
SUNDO	0.3	94.3	97.1	2.9	5.7	96.5	0.77
	0.4	96	96.4	3.6	4	96.3	0.75
	0.5	96.4	95.9	4.1	3.6	96	0.75

rate. It clearly appears that the re-sampling improves the performance of classifiers, especially when the dataset is strongly imbalanced. The best performances are obtained with the labelled SOM and SVM classifiers balancing perfectly the minority class with the majority class ($K_1 = 0.5$). In any case, however, the proposed method SUNDO overcomes the SMOTE and SBC approaches.

4.3. Industrial datasets

The proposed resampling method has also been tested on two real world datasets coming from the metal industry.

The first one concerns an automatic inspection systems which examines by means of a set of cameras the metal sheets during their production. The inspection system records all the defects found on the products surface. Afterwards, on the basis of the defect number, shape and seriousness, the quality of the examined sheet is assessed. This latter operation is extremely time consuming if performed by a human operator, thus the automatic classification system has been developed. Data belonging to defective products are far less with respect to the good ones and the unbalance rate is 24% which makes the classification task difficult. For this reason, resampling techniques have been adopted to improve the performance of the classifiers.

The four classifiers coupled with different resampling techniques have been used for the identification of defective sheets. A database containing 1915 samples (460 of which correspond to defective sheets), recording 6 features of the examined sheets, has been exploited for the tests. The results obtained in all the tests performed on this database are summarized in Table 5.

Different from the previous cases, in this application the best performance of all the considered classifiers is obtained when $k_1 = 0.4$, while the performance decreases (such as for the SVM-based or the for the Bayesian classifiers) or remains substantially unchanged when $k_1 = 0.5$ (e.g. for labelled SOM and the DT-based classifiers). Also in this case, SUNDO outperforms both the SBC and SMOTE approaches.

Also the second industrial dataset that has been used for testing the proposed method comes from the steel-making field and, in particular, to the casting process. In this process the liquid steel flows from one container, called tundish, to the subsequent

Please cite this article as: S. Cateni, et al., A method for resampling imbalanced datasets in binary classification tasks for real-world problems, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2013.05.059

Table 4
Classification results on the modified (10%) UCI WBCDB dataset without any resampling and by applying SMOTE, SBC and SUNDO re-sampling techniques with the four different classifiers.

Method	k_1	TP (%)	TN (%)	FP (%)	FN (%)	ACC (%)	IDX
SVM							
None	–	85.5	98.4	1.6	14.5	97.2	0.68
SMOTE	0.2	84.7	98.4	1.6	15.3	97.1	0.68
	0.3	88.1	97.7	2.3	11.9	96.8	0.7
	0.4	90	97.4	2.6	9.9	96.7	0.71
	0.5	87.6	97	3	12.4	96.1	0.69
SBC	0.2	83.3	97.4	2.6	16.7	96	0.64
	0.3	91.7	96.5	3.5	8.3	96	0.7
	0.4	91.7	95.6	4.4	8.3	95.2	0.69
	0.5	100	94.7	5.3	0	95.2	0.74
SUNDO	0.2	96.6	98	2	13.4	97.9	0.76
	0.3	88	98.1	1.9	12	97.2	0.7
	0.4	93.7	97.5	2.5	6.2	97.1	0.74
	0.5	96.5	97.4	2.6	3.5	97.3	0.76
DT							
None	–	77.4	97.2	2.8	22.6	95.3	0.62
SMOTE	0.2	79.2	96.9	3.1	20.8	95.2	0.63
	0.3	80	96.5	3.5	20	95	0.63
	0.4	81	96.3	3.7	19.9	94.8	0.64
	0.5	77.9	96.9	3.1	22.14	95.1	0.62
SBC	0.2	58.3	97.4	2.6	41.7	93.6	0.44
	0.3	83.3	95.6	4.4	16.7	94.4	0.62
	0.4	91.7	93	7	8.3	92.9	0.66
	0.5	83.3	95.6	4.4	16.6	94.4	0.62
SUNDO	0.2	80.3	97.3	2.7	19.7	95.7	0.64
	0.3	81.3	97.1	2.9	18.7	95.6	0.65
	0.4	83.2	96.5	3.5	16.8	95.2	0.66
	0.5	83.4	95.6	4.4	16.6	94.5	0.65
SOM							
None	–	88.3	98.1	1.9	11.7	97.1	0.7
SMOTE	0.2	90.5	97.8	2.2	9.5	97.1	0.72
	0.3	92.1	97.2	2.8	7.9	96.7	0.73
	0.4	94	96.6	3.4	5.8	96.3	0.74
	0.5	95.2	96.4	3.6	4.8	96.3	0.75
SBC	0.2	75	99.1	0.9	25	96.8	0.59
	0.3	91.7	96.5	3.5	8.3	96	0.70
	0.4	100	95.6	4.4	0	96	0.75
	0.5	100	93.9	6.1	0	94.4	0.74
SUNDO	0.2	92.3	97.2	2.8	7.7	96.7	0.73
	0.3	95	96.7	3.3	5	96.6	0.75
	0.4	95.6	95.7	4.3	4.4	95.6	0.75
	0.5	96.4	96.3	3.6	3.6	96.3	0.76
BAYESIAN							
None	–	92.3	98.3	1.7	7.7	97.7	0.73
SMOTE	0.2	91.7	98.4	1.6	8.3	97.8	0.73
	0.3	91.5	98.4	1.6	8.4	97.8	0.73
	0.4	93.2	98.1	2	6.8	97.6	0.74
	0.5	94.7	96.8	3.1	5.3	96.6	0.75
SBC	0.2	83.3	98.2	1.8	16.7	96.8	0.65
	0.3	91.7	96.5	3.5	8.3	96	0.70
	0.4	91.7	95.6	4.4	8.3	95.2	0.69
	0.5	100	94.7	5.3	0	95.2	0.75
SUNDO	0.2	93.9	97.2	2.8	6.1	96.9	0.74
	0.3	94.1	98	2	5.9	97.7	0.74
	0.4	94.8	97.8	3.1	5.2	97.7	0.74
	0.5	96.2	98.1	1.9	3.8	98.7	0.76

Table 5
Results of the metal sheet quality assessment without any resampling and by applying SMOTE, SBC and SUNDO re-sampling techniques with the four different classifiers.

Method	k_1	TP (%)	TN (%)	FP (%)	FN (%)	ACC (%)	IDX
SVM							
None	–	61.1	98.1	1.9	38.9	89	0.53
SMOTE	0.3	66.8	98	2	33.2	90.5	0.57
	0.4	69.2	97.3	2.7	30.8	90.5	0.58
	0.5	62.6	97.4	2.6	37.4	89	0.53
SBC	0.3	60	97.8	2.2	40	88.7	0.45
	0.4	65.2	97	3	34.8	89.3	0.48
	0.5	64.3	97.2	2.8	35.7	89.3	0.48
SUNDO	0.3	67	98	2	33	90.5	0.57
	0.4	69	97.7	2.3	31	90.8	0.58
	0.5	66.6	97.2	2.8	33.7	89.9	0.56
DT							
None	–	67	90	10	33	84	0.52
SMOTE	0.3	63.8	86.3	13.7	36.2	80.9	0.45
	0.4	66.4	83.4	16.6	33.6	79.3	0.44
	0.5	70.8	83.6	16.4	29.2	80.5	0.48
SBC	0.3	69.6	83.7	16.3	30.4	80.3	0.36
	0.4	72.2	86.2	13.8	27.8	82.4	0.42
	0.5	70.4	81.5	18.5	29.6	78.9	0.34
SUNDO	0.3	64.1	86.8	13.2	35.9	81.3	0.45
	0.4	71.6	83.6	16.4	28.4	80.7	0.49
	0.5	75.2	80	20	24.8	78.8	0.49
SOM							
None	–	0.9	99.2	0.8	99.1	75.5	0
SMOTE	0.3	56.4	97	2.9	43.3	87.3	0.48
	0.4	58.3	97	3.5	41.7	87.3	0.49
	0.5	59.5	96.1	3.9	4.5	87.3	0.5
SBC	0.3	0	100	0	100	75.9	0
	0.4	17.4	89.3	10.7	82.6	72	0
	0.5	53.9	44.6	55.4	46.1	46.9	−0.52
SUNDO	0.3	62.9	97.9	2.1	37.1	89.5	0.54
	0.4	64.8	97.8	2.2	35.2	89.9	0.55
	0.5	67.7	97.4	2.6	34.2	89.7	0.55
BAYESIAN							
None	–	25.6	98.2	1.8	74.4	80.7	0.23
SMOTE	0.3	30.6	97	3	69.4	80.9	0.26
	0.4	39.1	97.6	2.3	6.9	83.5	0.34
	0.5	60.6	49.6	50.4	39.3	52.3	−0.03
SBC	0.3	40	85.9	14.1	60	74.9	0.14
	0.4	57.4	99.2	0.8	42.6	89.1	0.45
	0.5	87	8.8	91.2	13	27.6	−1.29
SUNDO	0.3	47	88	12	52.9	78.1	0.32
	0.4	58.5	96	4	41.5	86.9	0.48
	0.5	67.7	41.5	58.5	32.3	47.8	−0.07

phases of manufacturing through a series of nozzles that are located on the bottom of the tundish itself. Unfortunately, depending on the liquid steel properties and on some manufacturing parameters, these nozzles can get blocked by solid inclusions. This latter event is detrimental for the final product quality and should be predicted before its occurrence.

A classifier for the early detection of nozzle occlusion has been developed by exploiting a set of available data coming from the industrial plant. This database contains about 3800 observations gathering information on the steel chemistry and on the process parameters. The rate of unbalance of this dataset is 1.25% thus resampling techniques need to be applied in order to maximize the overall performance and the detection of occlusions.

The results of these tests are shown in Table 6 and confirm the efficacy of SUNDO with respect to the SMOTE and SBC approaches as well as the need to apply some resampling techniques when the database is strongly imbalanced. In fact, when no resampling is applied, all classifiers attribute the validation samples to the majority class because the initial dataset is highly unbalanced. The

Please cite this article as: S. Cateni, et al., A method for resampling imbalanced datasets in binary classification tasks for real-world problems, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2013.05.059

Table 6
Results obtained on the occlusion detection dataset without any resampling and by applying SMOTE, SCB and SUNDO re-sampling techniques with the four different classifiers.

Method	k_1	TP (%)	TN (%)	FP (%)	FN (%)	ACC (%)	IDX
SVM							
None	–	0	100	0	100	98.7	0
SMOTE	0.2	51.2	94.7	5.3	48.7	94.2	0.37
	0.3	55.9	92.3	7.7	44.1	91.9	0.4
	0.4	59.5	90	10	40.5	89.7	0.41
	0.5	83.3	76.3	23.7	16.7	76.3	0.53
SBC	0.2	25	95	5	75	94.1	0.14
	0.3	75	86.5	13.5	25	86.4	0.45
	0.4	58.3	80.1	19.9	41.7	79.9	0.23
	0.5	58.3	76.3	23.7	41.7	76	0.17
SUNDO	0.2	64	94	6	36	93.7	0.48
	0.3	70	91.5	8.5	30	91.2	0.51
	0.4	73.5	86	14	26.5	85.9	0.51
	0.5	83.3	78.9	21	16.7	79	0.55
DT							
None	–	0	100	0	100	98.7	0
SMOTE	0.2	14.1	97.5	2.5	85.9	96.6	0.09
	0.3	15.1	97.3	2.7	84.9	96.4	0.1
	0.4	15.1	97.2	2.8	84.8	96.3	0.1
	0.5	16.7	94	6	83.4	93.2	0.08
SBC	0.2	16.7	94.4	5.6	83.3	93.4	0.07
	0.3	25	91.1	8.9	75	90.3	0.10
	0.4	8.3	92.9	7.1	91.7	91.8	0
	0.5	50	89.9	10.1	50	89.3	0.29
SUNDO	0.2	18	96.6	3.4	81.9	95.7	0.11
	0.3	18.6	98	2	81.4	97.2	0.13
	0.4	19.4	98	1.8	80.6	97.2	0.14
	0.5	25	97	3	75	96.2	0.18
SOM							
None	–	0	100	0	100	98.7	0
SMOTE	0.2	44.4	92.2	7.8	55.5	91.6	0.30
	0.3	56.4	90.6	9.4	43.5	90.2	0.39
	0.4	62.5	88	12	37.5	87.7	0.42
	0.5	67.8	85.3	14.7	32.1	85.1	0.45
SBC	0.2	8.3	98.4	1.6	91.7	97.2	0.05
	0.3	75	78.9	21.1	25	78.8	0.34
	0.4	50	80.1	19.9	50	79.8	0.16
	0.5	75	77.8	22.2	25	77.7	0.33
SUNDO	0.2	54	91	9	46	90.6	0.37
	0.3	73	86.1	13.9	27	86	0.50
	0.4	73.5	88.8	11.2	26.5	88.7	0.52
	0.5	80	83.5	16.5	20	83.5	0.55
BAYESIAN							
None	–	6	98.9	1.1	94	97.9	0.04
SMOTE	0.2	25.4	93.3	6.5	74.5	92.5	0.15
	0.3	25.5	90.4	9.6	74.5	89.7	0.12
	0.4	38.5	91.4	8.9	61.5	90.6	0.24
	0.5	42.5	97.8	2.1	57.5	97.2	0.33
SBC	0.2	25	90.9	9.1	75	90.1	0.10
	0.3	50	90.5	9.5	50	90	0.30
	0.4	58.3	80.7	19.3	41.7	80.4	0.23
	0.5	58.3	75.5	24.5	41.7	75.3	0.16
SUNDO	0.2	29.5	92.6	7.1	70.5	91.9	0.18
	0.3	40.5	97.8	2.2	59.5	97.2	0.31
	0.4	43.5	98	2	56.5	97.4	0.34
	0.5	51.5	96	4	48.5	95.5	0.39

performances of the classifiers increase for increasing values of the parameter k_1 and the best performance is achieved when $k_1=0.5$, namely when the two classes are perfectly balanced. The classifiers based on SVM and labelled SOM outperform the other ones. Also for this application SUNDO outperforms both the SMOTE and SBC approaches.

Noticeably, for all the proposed datasets and all the tested classifiers, when comparing the results obtained for the same unbalance degree (i.e. for a fixed value of the k_1 parameter) on the basis of the *IDX* index, SUNDO never achieves a lower performance with respect to the other tested resampling methods. However there are cases where the *ACC* index is higher for SMOTE or SCB with respect to SUNDO, which means that by using the SMOTE-based approach a higher percentage of patterns belonging to the frequent class is correctly classified. This fact does not affect the efficiency of SUNDO, as the improvement of the global accuracy is not the main purpose of the proposed resampling approach, which mostly aims at improving the classification of the rare patterns while keeping the classification of the frequent ones within acceptable limits.

5. Conclusions and future work

The problem of binary classification of imbalanced dataset has been treated. Through a combination of novel oversampling and undersampling procedures, it is possible to balance the original dataset and to improve the classification accuracy. The proposed method, that is named SUNDO, has been tested on different exemplar cases, where the obtained results have been compared to those achievable by another widely adopted resampling method. Considering the obtained results concerning the different classifiers, different datasets and different indices of unbalance we can conclude that

- When the class unbalance is significant, it could happen that classifiers associate the validation samples to the majority class demonstrating the need of the re-sampling technique.
- The best results, in terms of a novel index that take into account both corrected classified and false alarms, have been obtained using the classifiers SVM and labelled SOM with the two classes equally balanced.
- The proposed approach outperforms the widely adopted combination of SMOTE oversampling and random undersampling.

References

[1] A. Estabrooks, A combination scheme for inductive learning from imbalanced datasets (M.Sc. thesis), Faculty of Computer Science, Dalhousie University, 2000.
[2] A. Estabrooks, N. Japkowicz, A multiple resampling method for learning from imbalanced dataset, Comput. Intell. 20 (1) (2004).
[3] N. Japkowicz, The class imbalance problem: significance and strategies, in: Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI'2000): Special Track on Inductive Learning, Las Vegas, Nevada, 2000.
[4] M. Pazzani, C. Marz, P. Murphy, K. Ali, T. Hume, C. Brunk, Reducing misclassification cost, in: Proceedings of the 11th International Conference on Machine Learning, 1994.
[5] P. Li, K.L. Chan, W. Fang, Hybrid Kernel Machine Ensemble for Imbalanced Data Sets, in: 18th International Conference on Pattern Recognition, IEEE, 2006.
[6] B. Scholkopf, New support vector algorithms, Neural Comput. 12 (2000) 1207–1245.
[7] V. Soler, M. Prim, Rectangualr basis functions applied to imblanced datasets, Lecture Notes in Computer Science Vol. 4668 LNCS, Issue PART 1, 2007, Springer Verlag, Berlin Heidelberg, pp. 511–519.
[8] M. Vannucci, V. Colla, Novel classification methods for sensitive problems and uneven datasets based on neural networks and fuzzy logic, Appl. Soft Comput. 11 (2011) 2383–2390.
[9] C. Seiffert, T.M. Khoshgoftaar, J. Van Hulse, Hybrid sampling for imbalanced data, in: IEEE International Conference on Information Reuse and Integration IRI-08, Las Vegas (USA), 13–15 July 2008, pp. 202–207.
[10] N.V. Chawla, C4.5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure, in: Workshop on Learning from Imbalanced Dataset II, ICML, Washington, DC, 2003.

Please cite this article as: S. Cateni, et al., A method for resampling imbalanced datasets in binary classification tasks for real-world problems, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2013.05.059

[11] B. Raskutti, A. Kowalczyk, Extreme rebalancing for SVMs: a case study, SIGKDD Explor. 6 (1) (2004) 60–69.
[12] G. Wu, E. Chang, Class boundary alignment for imbalanced dataset learning, in: ICML 2003 Workshop on Learning from Imbalanced Data Sets II, Washington, D.C.
[13] R. Yan, Y. Liu, R. Jin, A. Hauptmann, On predicting rare classes with SVM ensembles in scene classification, in: IEEE International Conference on Acoustics Speech and Signal Processing, 2003.
[14] S. Kotsiantis, P. Pintelas, Mixture of expert agents for handling imbalanced datasets, Ann. Math. Comput. Teleinform. 1 (1) (2003) 46–55.
[15] S. Kotsiantis, D. Kanellopoulos, P. Pintelas, Handling imbalanced datasets: a review, GESTS Int. Trans. Comput. Sci. Eng. 30 (2006).
[16] R. Barandela, J.S. Snchez, V. Garca, E. Rangel, Strategies for learning in class imbalance problems, Pattern Recognit. 36 (3) (2003) 849–851.
[17] W. Cohen, Fast effective rule induction, in: Proceedings of the 12th International Conference on Machine Learning, 1995, pp. 115–123.
[18] P. Domingos, MetaCost: a general method for making classifiers cost-sensitive, in: Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining, ACM Press, pp. 155–164.
[19] R. Batuwita, V. Palade, Efficient resampling methods for training support vector machines with imbalanced datasets, in: The 2010 International Joint Conference on Neural Networks (IJCNN), 2010.
[20] M. Kubat, S. Matwin, Addressing the curse of imbalanced training sets: one sided selection, in: Proceeding of the 14th International Conference on Machine Learning, Nashville, Tennessee, Morgan Kaufmann, 1997, pp. 179–186.
[21] F. Provost, T. Fawcett, Robust classification for imprecise environments, Mach. Learn. 42 (2001) 203–231.
[22] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[23] W. He-Yong, Combination approach of SMOTE and biased-SVM for imbalanced datasets, in: Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN 2008), Hong Kong (PRC), 1–8 June 2008, pp. 22–31.
[24] C. Leichen, C. Zhihua, C. Lu, G. Qiong, A novel differential evolution-clustering hybrid resampling algorithm on imbalanced datasets, in: International Conference on Digital Object Identifier, 2010, pp. 81–85.
[25] N. Matarese, L.M. Reyneri, V. Colla, A method to point out anomalous input-output patterns in a database for training neuro-fuzzy system with a supervised learning rule, 9th International Conference on Intelligent Systems Design and Applications (ISDA 2009), Pisa; Italy, 30 November - 2 December 2009, pp.1307–1311, http://dx.doi.org/10.1109/ISDA.2009.202.
[26] S.J. Yen, Y.S. Lee, Cluster-based under-sampling approaches for imbalanced data distributions, Expert Syst. Appl. 36 (2009) 5718–5727.
[27] S. Ray, R.H. Turi, Determination of number of clusters in k-means clustering and application in colour image segmentation, in: Proceedings of 4th International Conference on Advances in Pattern Recognition and Digital Techniques (ICAPRDT 99), Calcutta, India, 27–29 December 1999, pp. 137–143.
[28] S. Cateni, V. Colla, G. Nastasi, A multivariate fuzzy system applied for outliers detection, J. Intell. and Fuzzy Syst. 24 (4), 2013, pp. 809–903, http://dx.doi.org/10.3233/IFS-2012-0607.
[29] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995).
[30] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth and Brooks/Cole Advanced Books and Software, Monterey, CA, 1984.
[31] V. Colla, M. Vannucci, S. Fera, S. Valentini, Ca-treatment of Al-Killed Steels: inclusion modification and application of Artificial Neural Networks for the prediction of clogging, in: Proceedings of 5th European Oxygen Steelmaking Conference (EOSC'06), Aachen, Germany, 26–28 June 2006, pp. 387–394.
[32] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, John Wiley and Sons, New York, 1973.
[33] M. Vannucci, V. Colla, M. Sgarbi, O. Toscanelli, Thresholded neural networks for sensitive industrial classification tasks, in: Bio-Inspired Systems: Computational and Ambient Intelligence, Lecture Notes in Computer Science, vol. 5517, 2009, pp. 1320–1327.
[34] W.H. Wolberg, O.L. Mangasarian, Multisurface method of pattern separation for medical diagnosis applied to breast cytology, Proc. Natl. Acad. Sci. 87 (1990) 9193–9196.
[35] P. Meesad, G.G. Yen, Combined numerical and linguistic knowledge representation and its application to medical diagnosis, IEEE Trans. Syst. Man Cybern. A: Syst. Hum. 33 (March (2)) (2003).
[36] R. Setiono, L. Chi Kwong Hui, Use of quasi-Newton method in a feed-forward neural network construction algorithm, IEEE Trans. Neural Netw. 6 (January (1)) (1995).

Silvia Cateni got her master degree in telecommunication engineering, in 2005, from Pisa University. Her research activity includes mathematical modeling and data analysis through statistical and artificial intelligence-based techniques. She collaborated in several research projects particularly dealing with steelmaking industry. She is presently a research assistant at Scuola Superiore Sant'Anna.

Valentina Colla received her Ph.D. (Dr. Eng.) in Robotics, in 1998. She became an associate professor at Scuola Superiore Sant'Anna in 2000 and She is presently a technical research manager at the PERceptual RObotic Laboratory of the Centre of Excellence for Information, Communication and Perception Engineering (CEIICP) of Scuola Superiore Sant'Anna in Pisa. Her research fields include the application of neural networks and other Artificial Intelligence techniques to data analysis as well as to the simulation and control of industrial processes.

Marco Vannucci got the master degree in Computer Science, in 2001, from Pisa University and the Ph.D. in Engineering from Scuola Superiore S. Anna, in 2006. His research activity includes simulation, mathematical modeling, artificial intelligence and robotics. He collaborates in several research projects focusing his interest in the exploitation of simulation techniques in the industrial framework. He is presently a research assistant at Scuola Superiore Sant'Anna.

Please cite this article as: S. Cateni, et al., A method for resampling imbalanced datasets in binary classification tasks for real-world problems, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2013.05.059